TR-IT-0033

音節バランス無意味単語を用いた合成音声主観評価結果と動的ケプストラムを用いた接続歪みの関係

Relationship between subjective evaluation results using syllable-blanced meaningless word sets and spectral discontinuity measured with dynamic cepstral distances

三村 克彦　樋口 宜男　匂坂 芳典

Katsuhiko Mimura　Norio Higuchi　Yoshinori Sagisaka

1993.12

概要

我々は、客観評価尺度構築の手始めとして、単語のなじみ度の影響を軽減するために音節連鎖の統計的性質は日本語一般の文章と酷似しているが、個々は無意味語からなる単語セットの作成方法を提案し、これを用いて音節明瞭度試験・音節良否評価試験を行ない、従来の有意味単語による評価試験結果との比較した。一方、客観尺度としてはこれまで用いてきたケプストラム距離に加えて、スペクトルの動的特性を良く表現できる動的ケプストラム距離を用いて、音声単位間の接続歪みを取り上げ、両者の相関を調べた。

あらまし

合成音声の主観評価試験結果と客観評価尺度の関係を明らかにすることは効率良く合成音声の品質向上を図るという観点から極めて重要な課題である。そこで、主観評価試験方法と客観尺度の両面から検討を行なった。まず、主観評価試験の方法として（1）単語のなじみ度の影響を軽減するために音節連鎖の統計的性質は日本語一般の文章と酷似しているが、個々の単語は基本的に無意味語からなる単語セットの作成方法を提案し、（2）この単語セットを用いて単語中の各音節の明瞭度を求めると共に、（3）明瞭度の低下にまで至らない軽度の問題点を把握するために、正解を被験者に示した上でその音らしさを判定する音節良否試験を行なった。一方、客観尺度としてはこれまで用いてきたケプストラム距離に加えて、スペクトルの動的特性を良く表現できる動的ケプストラムを用いた距離を用いて、接続部の接続歪みを求めた。その結果、（1）明瞭度試験の無意味単語を用いた明瞭度は従来の有意味語を用いたものの平均的な値におさまること、（2）有意味語で瀕出した他の有意味語への転移がほぼ半減したこと、（3）他の有意味語への転移は単語長が長い程起こり易いこと、（4）通常の接続方法では接続歪みが比較的小さいため、通常のケプストラムでも動的ケプストラムでも明瞭度との相関が現れないこと、（5）動的ケプストラムを用いた場合の接続歪みと音節良否度(その音らしさ)との間にはある程度の相関があること、が明らかになった。

1　はじめに

音声の評価方法には客観評価法と主観評価法とがある。従来、規則合成音声の場合には、適当なリファレンスが存在しないために主観評価法により評価を行なってきた。しかしながら、主観評価法には1) 試験条件、被験者による評価結果の揺らぎ、2) 試験の準備、試験結果の集計作業に多大な労力がかかる、などの問題点がある。また、このようにして求めた主観評価結果(了解性、自然性)と品質劣化要因との関係についても、あまり分析がされていないため、主観評価結果に基づいて効率良く合成音声の品質を向上させることが困難であった。

我々は、主観評価尺度と強い相関のある品質劣化要因の物理量に基づいた尺度(客観評価尺度)を導入することで、これらの問題点に対処し、合成音声の品質向上を効率よく行なうことを考えている。このためには、品質劣化要因と主観評価尺度の関係を明らかにすることが必要であると思われる。

品質劣化要因と主観評価尺度との関係を調べるにあたり、注意しなければならない点として、規則合成音声の了解性評価試験に用いる単語リストの問題がある。従来は、(1)モーラ数やアクセント型、なじみ度などを制御することにより、日本語自立語辞書と統計的な性質の近い有意味単語セットを作成し、これを評価に用いる方法[2]や(2)各種の先行音韻（無音、5母音、撥音など）と日本語の全音節を組合せた無意味単語を用いる方法[3]等が行なわれて来たが、両者ともそれぞれ問題点があった。すなわち、前者では単語のなじみ度を扱わなければならないが、単語のなじみ度は被験者によって異なると共に、それが了解性に与える影響を定量的にモデル化することは極めて難しい。一方、後者ではすべての音節の組合せを等確率で用い、しかも、極めて短い不自然な単語を用いて評価を行なうため、実環境で使われる場合の音節連

鎖の頻度とは異なった環境で了解性を評価していることになる。合成音声の場合、了解性の評価だけでは十分な評価が困難であり、併せて自然性の評価を行なうことが診断的評価の見地から重要な場合がある。ここで注意しなければならないのは、了解性と自然性の評価結果の質的な差である。すなわち、了解性を評価する場合には、範疇判断を伴うため評価結果は離散的であるのに対して、自然性の評価結果は連続的である。このため、了解性については評価試験時の条件さえ確実に押えれば定量的な評価が比較的容易であるが、自然性についてはこの点がかなり難しい。従って、了解性の評価の場合には、聴取の対象となる単語リストを規定することにより、評価試験の枠組みが明確に表現できるが、逆に誤認識の対象と成り得る候補の単語がない限り、相当品質の悪い音声でも誤認識が起こらず、評価結果に差がでないという特徴を持つ。これに対して、自然性の評価の場合には、わずかな品質の劣化も評価結果に現れるが、評価試験時の指示により評価結果に大幅な差が生じ易いという特徴を持つ[1]。我々は、これらの点を考慮し、客観評価尺度構築の手始めとして、以下のことを行なった。音節連鎖の統計的な性質は日本語の一般的な文章に極めて近く、しかも無意味単語からなる評価用単語セットを作成し、これを用いて音節明瞭度試験を行ない、従来の有意味単語による評価試験結果[4]との比較を行なった。明瞭度の低下にまで至らない軽度の問題点を把握するために、正解を被験者に示した上でその音らしさを判定する音節良否試験[3]も合わせて行ない音声単位間の接続歪みとその音らしさ(以下、音節良否度という)を表すオピニオン値の関係を調べた。一方、客観尺度としてはこれまで用いてきたケプストラム距離に加えて、スペクトルの動的特性を良く表現できる動的ケプストラムを用いた距離を用いて、接続部の接続歪みを求め、了解性との相関を調べた。

本稿では、これらの結果について報告する。

2　日本語音節連鎖の統計的性質に基づく無意味単語セットの作成

2.1　日本語音節連鎖の統計処理

音節連鎖の統計的な性質が日本語の一般的な文章に極めて近い無意味単語セットを作成するために、まず2音節連鎖および3音節連鎖の種類とそれらの出現率を調査した。各音節連鎖の出現率は、新聞、雑誌および小説から無作為に抽出した9609文章(約10万語)のテキストデータ[5]より、算出した。なお、引き音・促音は同一音節、二重母音・撥音は別音節として扱った。出現率の高い2音節連鎖および3音節連鎖の種類と出現率を表1に示す。

2.2　無意味単語セットの作成方法

無意味単語セットの2音節連鎖および3音節連鎖の統計的性質を、日本語文章のそれに近付ける方法としては種々の方法が考えられるが、ここでは単語セット作成の過程を次の2段階に分けて行なった。

(1) 評価単語セット中の2音節連鎖数の決定
(2)3音節連鎖統計に基づく評価単語の生成

2.2.1　評価単語セット中の2音節連鎖数の決定

ここで用いる無意味単語セットは有意味単語セットと比較するためのものであるから、総単語数、単語長などはなるべく有意味単語セットのそれと合わせる必要がある。このように総単語数と総音節数を与えると、単語セット内に含めることができる2音節連鎖の総数が自動的に決まり、この値をQとする。ここで、テキストデータより算出した2音節連鎖の出現率リスト中の2音節連鎖の

ラベルを$L2(i)$、その出現率を$T2(i)$とすると、評価単語セット中に含めるべき$L2(i)$の個数$N2(i)$は、次のように求めることができる。

$$N2(i) = int(T2(i) * Q) \tag{1}$$

$T2(i)*Q$の小数部を切捨てているため、$\sum N2(i)$はQより少なくなる。そこで、2音節連鎖の不足分D個($= Q - \sum N2(i)$)を補うために、

$$M2(i) = T2(i) * Q - N2(i) \tag{2}$$

とし、$M2(i)$の大きい順にD個の2音節連鎖を補う。

2.2.2 3音節連鎖統計に基づいた評価単語の生成

評価単語作成はテキストデータより算出した3音節連鎖の出現率を参照する。ここで、3音節連鎖のラベルを$L3(i)$、その出現率を$T3(i)$とし、3音節連鎖毎接続試行回数を$N3(i)$ ($N3(i)$の初期値はすべて0)とする。

ここで、すべての3音節連鎖$L3(i)$について、

$$M3(i) = T3(i)/(N3(i) + 1) \tag{3}$$

を最大とする$L3(i)$について、評価単語セット内での接続可能性を確かめ、可能な場合には接続を行なう。なお、$N3(i)$は接続の可否に関係なく、1増やすものとする。

なお、むやみに長い評価単語を生成させないために、接続音節数には上限を設ける。基本的には上記の接続処理を総単語数が所望の数に減少するまで繰り返すが接続音節数の上限および接続可能性の関係から、必ずしも所望の単語数まで減少しないことがある。接続率を(接続した回数)/(所望の接続回数)として定義すると、本評価単語セットにおける接続率は約85％であった。

最後に、禁則処理として、評価単語の先頭に促音・撥音が存在する場合に語頭にダミーの母音を補った。

3　客観評価尺度としての接続歪み

隣接音声単位間s_p、s_fのスペクトル接続歪$C_s(s_p, s_f)$を次式で定義する。

$$C_s(s_p, s_f) = \sum_{i=-1}^{1} h_c(i) \sum_{j=1}^{m} (x_{p_{i,j}} - x_{f_{i,j}})^2, \qquad (4)$$

ただし、

$h_c(0) = 1$, $h_c(1) = h_c(-1) = 0.5$.

ここで、$x_{p_{i,j}}$ と $x_{f_{i,j}}$ は、隣接音声単位 s_p と s_f の接続フレームからのi番目のフレームのj番目のケプストラムパラメータである。ケプストラムパラメータとして、30次ケプストラムを用いる。

さらに、接続歪みの尺度として、ケプストラム係数よりも聴知覚系に近い音響的特徴を抽出していると思われる動的ケプストラム[6]を加えた。ここで、時点iにおけるk次のケプストラム係数を$C_k(i)$、マスキングパターンのケプストラム展開係数を$M_k(i)$とすると、動的ケプストラム$B_k(i)$は次式で与えられる。

$$B_k(i) = C_k(i) - M_k(i) \qquad (5)$$

ただし、

$$M_k(i) = \sum_{n=1}^{N} C_k(i-n) L_k(n) \qquad (6)$$

$L_k(n)$はn時点前のスペクトルを平滑化するリフタのk次の係数である。Nはマスキングの影響の及ぶ最大時間長を表している。図4に示すように、隣接音声単位間の接続点近傍の3フレーム分の動的ケプストラムの自乗平均距離D_cを算出し、各接続点での和をその音節の接続歪みとした。

4 明瞭度試験

4.1 試験方法

音声サンプルは、各評価単語セット間の使用音声単位数を同等にするため、音節より細かな音韻を1単位とした[7][8]。また、音声単位間の接続歪みとの関係を調べるため、幾つかの音声単位候補列から無作為に1つの音声単位列を選び、各単語での接続歪みの総計に偏りをなくすようにして規則合成した。明瞭度試験には、155単語、548音節からなる有意味単語セット[2]と今回作成した181単語600音節からなる無意味単語セットを用いた。試験は、あまり合成音声を聞きなれていない10名の被験者に対して行い、防音室でヘッドホーンを用い1回音声を呈示した後に書き取らせた。

4.2 試験結果

音声サンプルのなじみ度と音節正解率の関係を図1に示す。無意味単語セットの音節正解率は、従来の有意味単語セットの比較的なじみのある単語の音節正解率ほど高くない。これは、言語的冗長性により見かけ上高くなっていたなじみのある有意味単語における評価結果を補正したものと言える。図2は各評価単語セットの有意味語への転移率である。図2の示すように、有意味語セットで頻出した他の有意味語への転移は、無意味語セットでは、ほぼ半減している。図3に無意味単語セットでの有意味単語へ転移した単語が誤り単語総数に占める割合を示す。音節数が多くなるほど、有意味単語への転移率は高く、音節数の多い無意味単語ほど他の有意味単語との共通部分が長くなりやすいことを示唆している。この無意味語から有意味語への転移には、/oNmeiji/(おんめいじ)を/omei/(おめい)、/takaraki/(たからき)を/katayaki/

(かたやき)、/itagana/ (いたがな) を /hiragana/ (ひらがな)、/kotoda/ (ことだ) を /kotoba/ (ことば)、/teta/ (てた) を /geta/ (げた) 等がみられた。

5 音節明瞭度と接続歪み

先に有意味単語を用いて行なった音節明瞭度試験では、単語になじみのない場合に音節明瞭度とスペクトル接続歪みに相関がみられた[9]。これは、スペクトル接続歪みの総計が最小または最大になるように音声単位を選択して規則合成した音声サンプルを対象としたので、音声サンプル間で極端なスペクトル接続歪みの差が生じており、接続歪み最大の場合には、音節明瞭度の低下を招いていたためと思われる。一方、今回用いた幾つかの音声単位候補列から無作為に1つの音声単位列を選びスペクトル接続歪みの総計に偏りをなくすようにして規則合成した音声サンプルでは、スペクトル接続歪みの差があまり大きくなく、両者の間には明確な相関は見られなかった。この傾向は今回作成した無意味単語を用いた音節明瞭度試験においても、接続歪みの尺度に動的ケプストラム[6]を用いても同様であった。これは、接続歪みが大きい音声サンプルでも、被験者にはかなり違和感があるが正解にしか聞こえない場合が多いことに起因していると思われる[1]。

6 音節良否評価試験

6.1 試験方法

音節良否度と接続歪みの関係を調べるため、正答を示した上で音節の良否の判断を5段階で評価する音節良否試験[3]を行なった。表2に音節良否評価の基準を示す。音声サンプルは、3節の無意味単語セットの音声を用い

た。被験者は、2名である。音声サンプルは音節良否評価が決定されるまで、繰り返し呈示した。

6.2 試験結果

図5に音節良否評価試験の結果を示す。図5において、横軸は表2で示した5段階評価値、縦軸は動的ケプストラムを用いた接続歪みのである。また、ひげは標準範囲、箱の切れ込み部分は中央値の95%信頼区間、箱の中の破線は中央値、箱の横幅は各カテゴリのデータサンプル数の平方根に比例した値、箱の縦幅は50%のデータがその範疇にあることを示している。図5に示すように、音節良否度と接続歪み間には、接続歪みが大きくなるほど音節良否度が低くなる傾向がみられた。しかしながら、接続歪みが大きくてもその音らしさがよく出ていると判定されたサンプルや、逆に接続歪みが小さくてもその音に聞こえないと判定されたサンプルもあった。これは、選択された音声単位と、同一音声単位の平均的な音声単位との類似度の違いに起因するものと思われる。従って、音節明瞭度、音節良否度を客観的に説明する尺度として、音声単位間の接続歪みだけでなく、用いた音声単位と、同一音声単位の平均的な音声単位との類似度も考慮にいれる必要があるものと考えられる。

7 おわりに

合成音声の主観評価試験結果と客観評価尺度の関係を明らかにすることは効率良く合成音声の品質向上を図るという観点から極めて重要な課題である。そこで、主観評価試験方法と客観尺度の両面から検討を行なった。まず、主観評価試験の方法として（1）単語のなじみ度の影響を軽減するために音節連鎖の統計的性質は日本語一般の文章と酷似しているが、個々の単語は基本的に無意味語からなる単語セットの作成方法を提案し、（2）こ

の単語セットを用いて単語中の各音節の明瞭度を求めると共に、（3）明瞭度の軽減にまで至らない軽度の問題点を把握するために、正解を被験者に示した上でその音らしさを判定する音節良否試験を行なった。一方、客観尺度としてはこれまで用いてきたケプストラム距離に加えて、スペクトルの動的特性を良く表現できる動的ケプストラムを用いた距離を用いて、接続部の接続歪みを求めた。その結果、（1）明瞭度試験の無意味単語を用いた明瞭度は従来の有意味語を用いたものの平均的な値におさまること、（2）有意味語で瀕出した他の有意味語への転移がほぼ半減したこと、（3）他の有意味語への転移は単語長が長い程起こり易いこと、（4）通常の接続方法では接続歪みが比較的小さいため、通常のケプストラムでも動的ケプストラムでも明瞭度との相関が現れないこと、（5）動的ケプストラムを用いた場合の接続歪みと音節良否度との間にはある程度の相関があること、が明らかになった。しかしながら、今回客観尺度として用いた接続部における接続歪みのみでは評価試験結果との対応が難しく、用いた音声単位と、同一音声単位の平均的な音声単位との類似度も考慮する必要が考えられ、今後この点について検討を行なう予定である。

謝辞　研究の機会を与えていただいた、ＡＴＲ音声翻訳通信研究所山崎泰弘社長に感謝いたします。また、熱心に討論いただくATRの皆様に感謝いたします。

参考文献

[1] 樋口、山本、松崎：”規則合成音の了解性・自然性の評価方法に関する検討”, , 音講論集,3-4-6 (1986-10).

[2] 渡辺、長渕、北脇：”規則合成音声の了解性評価に用いる単語リストの構成法”, 信学論A Vol. J71-A, No.3, pp.616-623 (1988-03).

[3] 樋口、山本、清水：”パラメータ導出型日本語音声規

則合成装置の評価", 信学論D-II Vol. J72-D-II, No.8, pp.1133-1140 (1989-08).

[4] 三村、海木、匂坂："ATR υ -Talk合成音声の明瞭度評価", 音講論集,1-P-13 (1992-10).

[5] 磯、渡辺、桑原："音声データベース用文セットの設計", 音講論集,2-2-19 (1988-03).

[6] 相川、河原、東倉:"順向マスキングの時間周波数特性を模擬した動的ケプストラムを用いた音韻認識", 信学論A,Vol. J76-A, No.11, pp.1514-1521 (1993-11)

[7] 武田、安部、匂坂："選択的に合成単位を用いる規則音声合成", 信学論D-2, J73-D-2, No.12, pp.1945-1951 (1990-12).

[8] 岩橋、海木、匂坂:"音響的尺度に基づく複合音声単位選択法", 信学技報, SP91-5 (1991-05).

[9] 三村、岩橋、匂坂："単位接続歪みが合成音声明瞭度に与える影響について", 音講論集,1-8-11 (1993-03).

表1. 出現率の高い音節連鎖

2音節連鎖	出現率	3音節連鎖	出現率
ない	0.006726	わたし	0.005130
こと	0.006251	してい	0.002645
てい	0.006249	じぶん	0.002172
して	0.005494	たしは	0.002096
よう	0.005168	ように	0.002093
した	0.005031	ような	0.001887
いる	0.004745	ことが	0.001551
ある	0.004632	こころ	0.001447
たし	0.004338	のであ	0.001371
かん	0.004263	かんが	0.001310
から	0.004225	んげん	0.001299
わた	0.003950	ではな	0.001284
んの	0.003869	るよう	0.001270
いた	0.003856	にんげ	0.001263
せい	0.003789	のよう	0.001245
とい	0.003764	ことを	0.001219
であ	0.003748	こども	0.001212
もの	0.003597	おもい	0.001212
かい	0.003427	れてい	0.001209
たい	0.003419	んがえ	0.001072

表2. 音節良否評価の基準

基準値	意味
4	その音（正解）らしさがよく出ている。
3	やや違和感はあるが、その音（正解）として十分聞こえる。
2	かなり違和感があるが、強いてどの音かと言われればその音（正解）にしか聞こえない。
1	正解を示されればその音に聞こえるが、正解が示されなければ、別の音として聞いてしまう。
0	正解を示されても、その音には聞こえない。

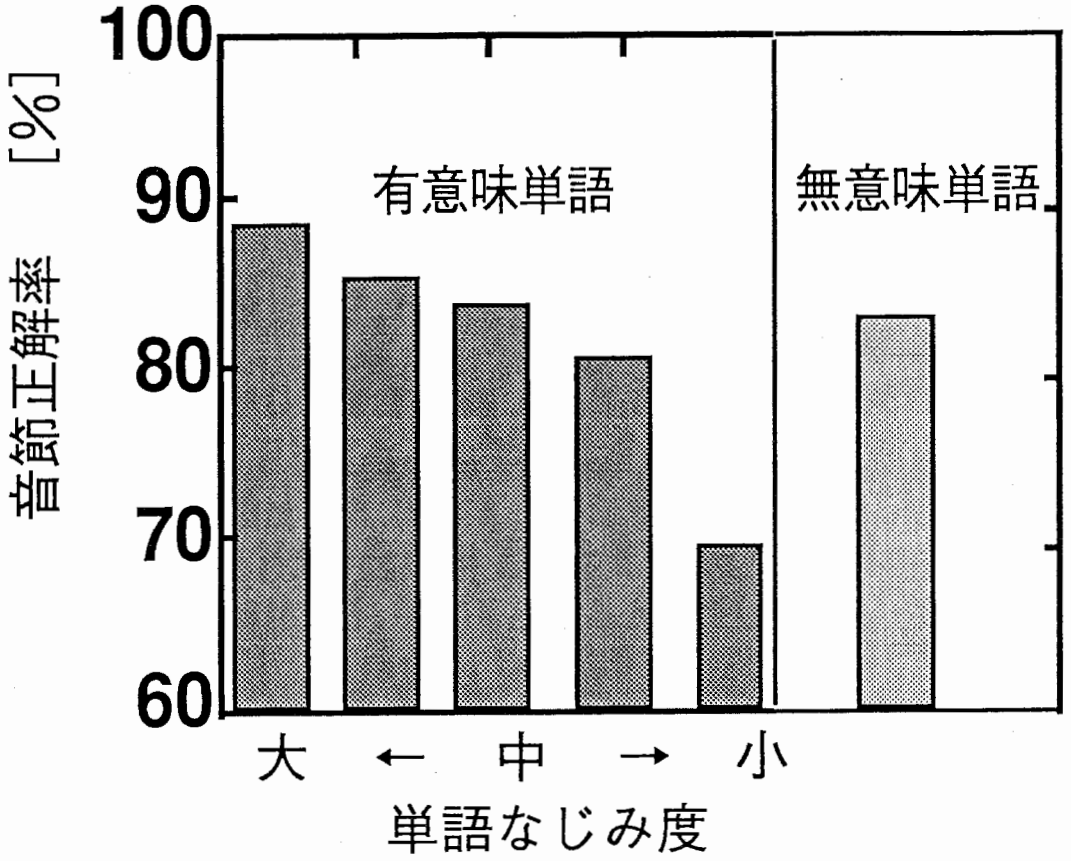


図1. 音節正解率

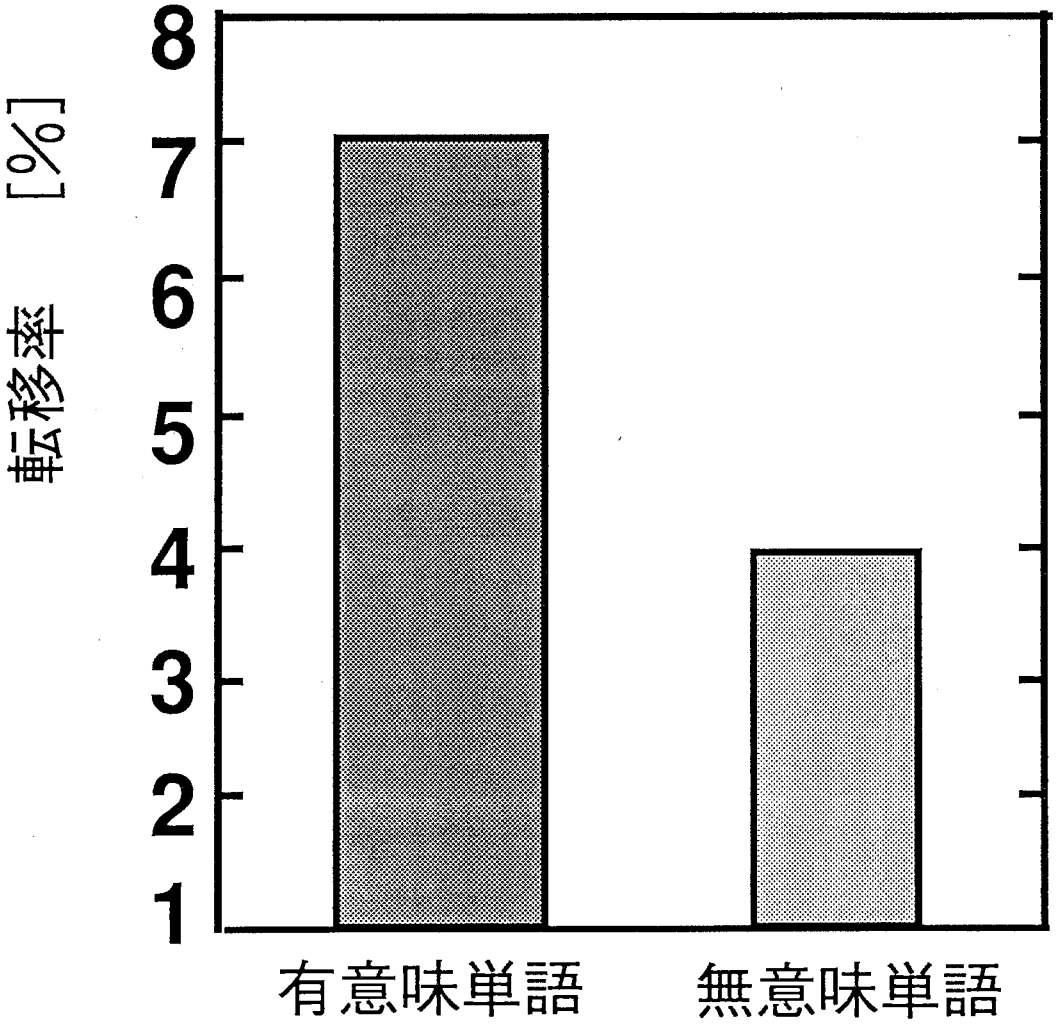


図2. 有意味単語への転移率

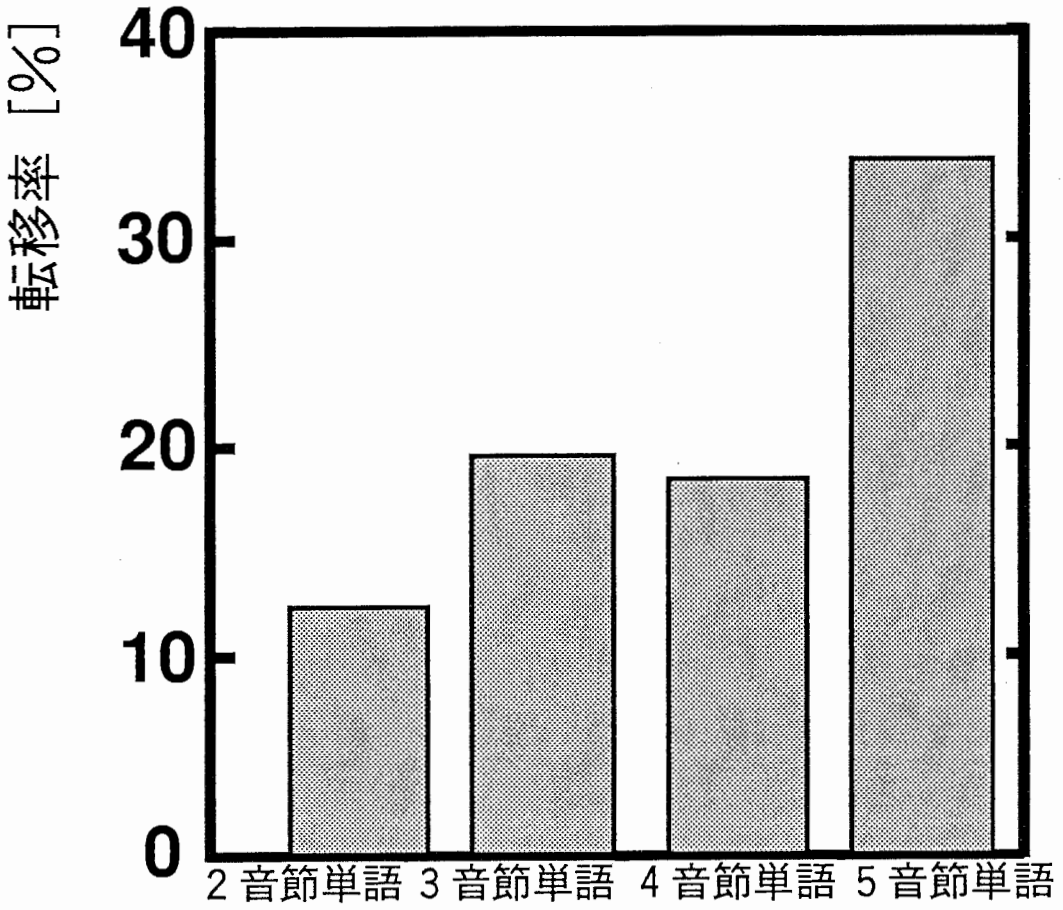
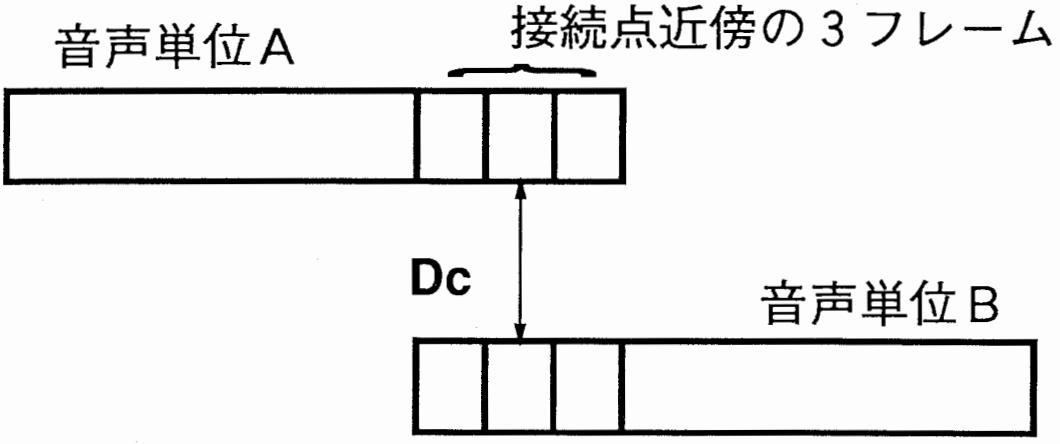


図3. 有意味単語へ転移した単語数の
誤り単語総数に対する割合



Dc ： ３フレームの動的ケプストラムの距離

図4. 動的ケプストラムを用いた接続歪み

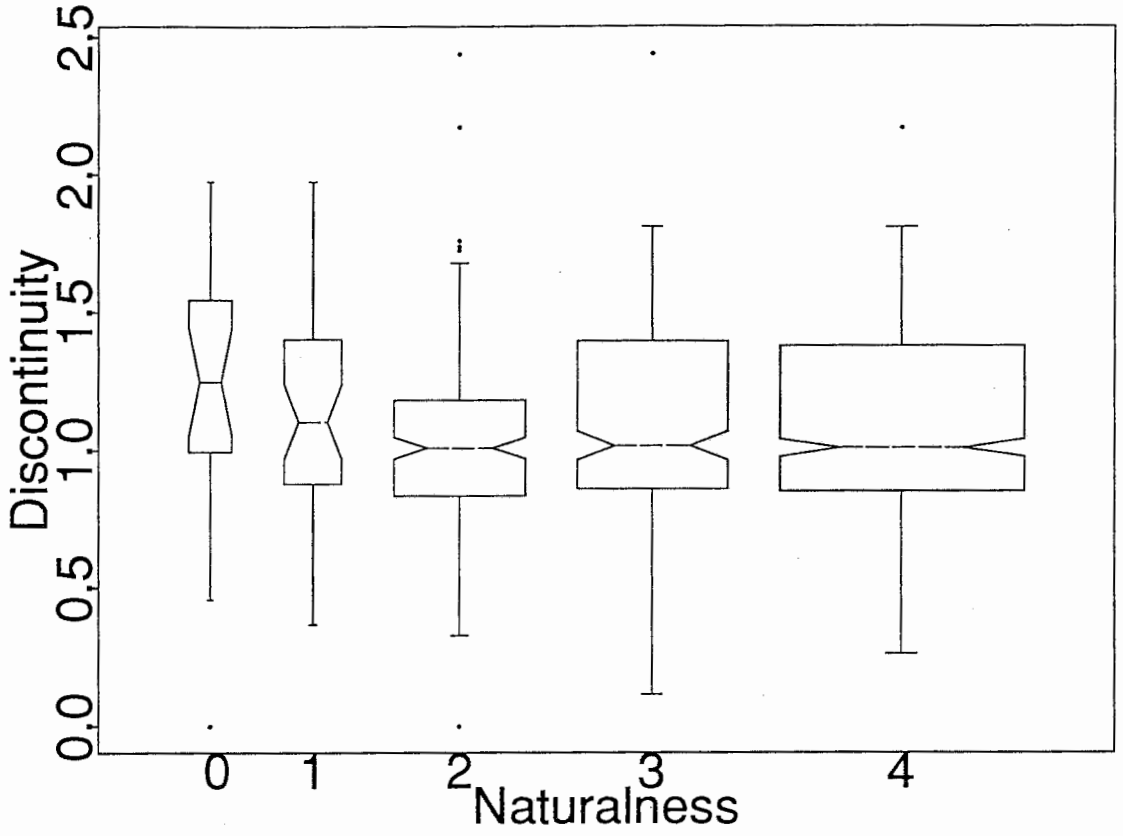


図5. 音節良否度と接続歪み

付録1. 無意味単語リスト

ないとか	できなく	どこにあ	はらをみ
こととがで	かくにと	ては	くやは
こととがな	おはい	いいさ	ろは
ことだ	ところに	まちのな	きみのな
してきに	あいだん	ばかりと	がち
したかさ	のは	または	えっかり
いるかど	されてい	なけれど	すばらく
あるとそ	しかたが	のく	しゃに
あるう	なるとと	どんなこ	あってこ
かんがい	だった	ううの	つれてい
からない	くるのが	がきがし	てす
わたしに	ならずに	みちをす	すなおに
いたがな	うであ	にひ	あをて
せいじに	たからき	にほんじゅ	けは
ものであ	しないで	いっても	えったに
なかにい	ちょう	しあいだ	ねつきあ
いうのも	じゅう	えられる	てう
こうとい	なのです	おはは	ひのつ
おもわれ	こどもを	とって	あせをな
うってい	なった	りながれ	ぐち
えってい	めいにな	さかのこ	おおし
しんけい	いんせいで	おをふ	きするこ
ひとりあ	おとなた	いんざい	えは
しょう	いはあ	わって	むいたこ
がいのか	のはなし	おんめいじ	ううす
ろうじの	うじんぼ	りたいし	うちゅ
このひ	おはか	あなた	えったひ
うしんら	いえない	がっこ	どれ
じょう	もって	せかいお	ひたすら
しい	くともに	てつだ	やは
あはな	えをう	とてもこ	えっせい
かれがこ	にはなし	うはさ	ごいたみ
かった	にこ	つは	きっと
たちをお	によりお	ひかりを	みどり
では	うするた	らがわに	くぼ
くないこ	いさん	ううを	らふ
にんのか	うせいを	くようにに	えはしゅう
ながらも	しょく	べんちょ	よごれ
それでま	かおをみ	のば	もぼ
おんだいな	くがひ	うめいじゅ	ざつ
いきるや	しきがし	なわれる	きざ
しまし	もとめな	とどま	えっとの
きょう	ちは	だすとぼ	なて
のような	ふたりし	てた	
さんがえ	だからと	あったく	
です	うはじ	きりのこ	