

論文 / 著書情報
Article / Book Information

論題(和文)	単語音声認識における時間軸正規化方法の検討
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会1976年講演論文集, , , pp. 341-342
Citation(English)	, , , pp. 341-342
発行日 / Pub. date	1976, 5

3-2-16 単語音声認識における時間軸正規化方法の検討*

古井 貞昭 (日本電信電話公社 武蔵野電気通信研究所)

1. まえがき 音声の機械認識の研究には、単語から文章を対象とするものまで種々のレベルのものがあるが、単語音声の認識が現在最も実現の可能性が大きいと考えられる。単語音声の認識に関して重要な課題の一つに、時間伸縮の正規化の問題があり、入力サンプルと標準パターンとの間で伸縮の度合いに一定の制約をもたせながら、最もよく整合するように、DPの手法を用いて非線形に対応づける方法が広く用いられている^{(1)~(3)}。この伸縮の度合いに関する制約の与え方としては、種々の方法が試みられているが、これらの有効性に関する統一的な比較はまだ行われていないようである。ここでは認識実験により、これまでにとり上げられていない新しい方法を含めた種々の方法の比較を行う。

2. 音声スペクトルと類似度の表現 単語音声波を3.2 kHzの低域通過フィルタに通したのち、8 kHz, 11ビットでサンプリングおよび量子化を行い、15 msの区分毎に10次までの相関係数を抽出する。さらに線形予測分析法により予測係数を求める。各区分における入力サンプル(x)と標準パターン(予測係数 a)の距離(類似度)を、坂倉⁽³⁾の最小予測残差原理(Minimum Prediction Residual Principle)にもとづいて、次のように正規化対数予測残差によって表現する

$$d(x/a) = \log(aVa'/\hat{a}V\hat{a}') \\ = \log(aa) + \log[(br)/(\hat{a}r)]$$

ただし(yz)は内積を表わし、 V は相関行列、 r は相関係数、 $\hat{a}$ は入力サンプルから推定した予測係数で、 $b = (1, b_1, \dots, b_p)$ の要素は次の通りである

$$b_i = 2 \sum_{j=0}^{p-i} a_j a_{j+i} / (aa)$$

3. 時間軸の正規化方法 入力サンプルと k 番目の標準パターンの距離(類似度)を次のように定義する。

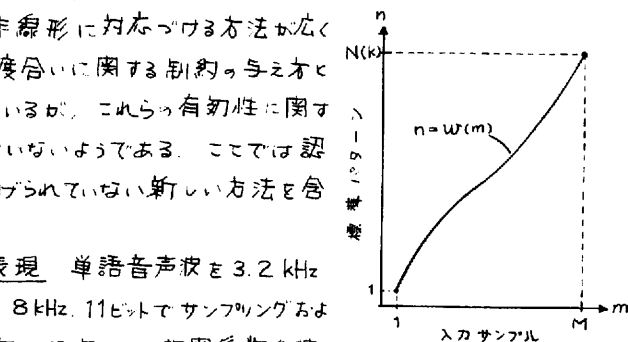


図1. 入力サンプルと標準パターンの時間軸の対応づけ

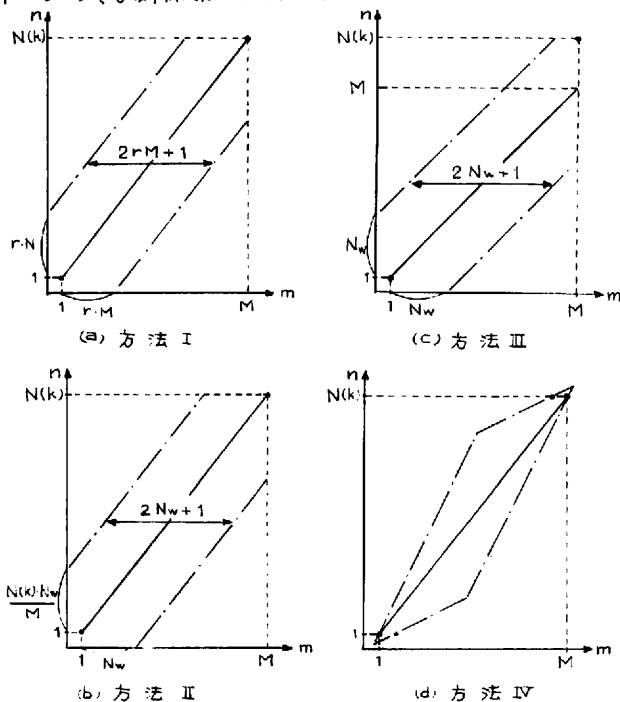


図2. 種々の整合容

* Evaluation of Time-Warping Methods in Isolated Word Recognition.

By Sadaaki FURUI (Musashino ECL, NTT)

$$D(k) = \min_c \left[\left\{ \sum_{l=1}^L d(c_l; k) \cdot \Delta_l \right\} / \sum_{l=1}^L \Delta_l \right], \quad C = \{c_l\} = \{(m_l, n_l)\}, \quad l=1, 2, \dots, L$$

$$c_1 = (1, 1), \quad c_L = (M, N(k)); \quad \text{境界条件}$$

ここに、 Δ_l は c_l と c_{l-1} の関係により定まる。この操作は、図1に示すように、入力サンプル(長さ M)と標準パターン(長さ $N(k)$)の時間軸に対応する伸縮関数 $n = w(m)$ を考え、両者の全体としての距離が最小となるものを選ぶことに相当する。 $w(m)$ のとりうる範囲に、次のような種々の制約を加える。

[方法 I]: $m \frac{N(k)}{M} - r N(k) \leq w(m) \leq m \frac{N(k)}{M} + r N(k); \quad 0 \leq r \leq 1$

[方法 II]⁽⁴⁾: $m \frac{N(k)}{M} - N_w \leq w(m) \leq m \frac{N(k)}{M} + N_w; \quad 0 \leq N_w$

[方法 III]⁽⁵⁾: $m - N_w \leq w(m) \leq m + N_w$

[方法 IV]⁽³⁾: $\begin{cases} w(m+1) - w(m) = 0, 1, 2 & (w(m) \neq w(m-1)) \\ = 1, 2 & (w(m) = w(m-1)) \end{cases}$

各方法における $w(m)$ のとりうる範囲(整合窓)を図2に示す。方法 III において、 $M - N_w \leq N(k) \leq M + N_w$ の条件が満たされないとすば、 $(M, N(k))$ の点が $w(m)$ のとりうる範囲内に入らないうので、その場合のみ境界条件を変更した。方法 IV は、入力サンプルと標準パターンの各時点の相対速度を $1/2$ と 2 の間に制限することと意味し、傾斜制限 DP マッチング ($P=1$)⁽⁵⁾ にほぼ等しい。

4. 認識実験 男性10名が区切って発声した10数字音声を用いて、認識実験を行った。各話者について、各数字をはじめに1回ずつ発声したものを標準パターンに用いた。未知サンプルは計965サンプルである。(標準パターンと無声化傾向の異なるサンプルは実験の対象から除いた)。全体としての伸縮率に関連して、 $N(k) \geq 2M$ または $N(k) \leq \frac{1}{2}M$ の場合は $D(k) = \infty$ とおくという条件を付加した。

認識結果を図3に示す。この結果から次のことがわかる。①認識

率のピーク値は99.8%で、方法 I~III で共通である。②整合窓を用いることにより、用いない場合よりも誤認識が $1/2$ に減少する。③方法 I と II は類似した結果を示す。④方法 III よりも I あるいは II の方が安定した結果を示し、整合窓の幅も狭くてすむ。⑤方法 IV による認識率は99.7%で、他の方法のピーク値よりも低い。⑥好田ら⁽²⁾による認識方法を適用したところ、無声化サンプルを除いた場合(好田-1)に100%、除かない場合(好田-2)に99.8%の認識率が得られた。

[謝辞] 日頃御指導頂く斎藤博伸室長の方々に、熱心に討論を頂く好田調達室長をはじめとする研究室の諸氏に感謝する。
[文献] 1) 迫江他、音字誌, 22, 9, p.483, (1971), 2) 好田他、研果報, 21, 7, p.1361, (1972), 3) E. Itakura, IEEE Trans. ASSP-23, 1, p.67, (1975), 4) 長島他、本手稿集, 5) 迫江、春季音楽学論, 1-2, 15, (1974)

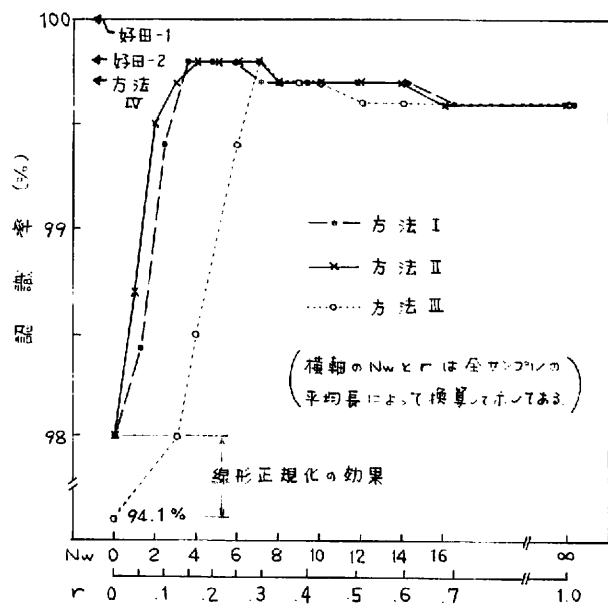


図3. 認識実験の結果