

論文 / 著書情報
Article / Book Information

論題(和文)	音声認識における個人差の学習法について
Title	
著者(和文)	古井 貞熙
Author	SADAOKI FURUI
出典(和文)	資料番号S75-25, , No. S75-25, pp. 1-8
Journal/Book name	, , No. S75-25, pp. 1-8
発行日 / Issue date	1975, 11

昭和50年度科研費
音声総合研究委員会 資料
音響学会研究委員会
資料番号 S75-25(1975-11)

音声認識における個人差の学習 法について

古 井 貞 熙

(日本電信電話公社・武蔵野電気通信研究所)

1975年11月22日

昭和50年度 音声の生成・知覚過程およびその
科 研 費 自動化に関する総合的研究委員会

日本音響学会 音 声 研 究 委 員 会

音声認識における個人差の学習法について

古井 貞熙

(日本電信電話公社 武蔵野電気通信研究所)

昭和50年 11月22日 於 東北大学電気通信研究所

内容梗概： 不特定多数の話者が発声した多数の語いを高い認識率で認識するには、話者の個人差をいかにして学習し正規化するかという問題が重要なかぎとなる。本報告では、声帯波スペクトルと声道長の効果を正規化する方法とその効果について検討した結果について述べている。前者の効果を音声スペクトルの概形（傾き）としてとらえ、後者の効果を相関係数パタンの時間軸方向の伸縮としてとらえて正規化する方法を検討した。未知音声を区分毎に正規化する方法と、標準パターンを話者に適応化する方法を用いて、男・女性60名が発声した10数字の認識実験を行なった結果、この正規化によって、個人差による誤認識の $1/2$ は除去できることが確認された。

Learning and Normalization of the Talker Differences in the Recognition of Spoken Words

Sadaoki FURUI

(Musashino Electrical Communication Laboratory, N.T.T.)

November 22, 1975 at R.I.E.C., Tohoku Univ.

Abstract: It is important to solve the problem of how to learn and normalize the talker differences in order to achieve high accuracy in the recognition of spoken words, when the vocabulary is very large and the words are uttered by unspecified talkers. In this paper, methods of the normalization of the effects of the vocal-cord wave spectrum and the vocal-tract length are considered. As the measure of these effects, global pattern (inclination) of the speech spectrum and the expansion-contraction rate of the auto-correlation pattern are estimated. Input speech is normalized in every frame, or reference patterns are adapted for the talker. Recognition experiments were done for spoken digits uttered by sixty male and female talkers, and half of the recognition error due to the talker differences are removed by the normalization.

化は、この音韻境界に平行しているものが多く、しかも $I \leftrightarrow F$ の距離は異なる音韻間の距離よりも大きい場合もある。このことは、高い音韻認識率を得るためには話者に適応した標準パターンを用いるか、入力音声の個人差を正規化する必要のあることを示している。

音韻に依存し、かつ音韻に共通の構造を有する個人差をモデル化する最も簡単な方法は、それを一次元的に表現することである。これはたとえは、各音韻の個人差による変化を $I \leftrightarrow F$ の直線上の化例的な動きとして考えることに相当する。筆者は、このような原理にもとづいて個人差に自動的に適応するシステムを構成し、実験的検討を行ったが、まだ必ずしも十分な結果は得られていない。^{6,7)} 好田ら^{2,3)}も同様な原理にもとづいて、標準ボタンの一部から不足分を推定する方法を採用し、検討を行っている。個人差を一次元ではなくて、多次元で表現するためのアルゴリズムには、正準相関分析を用いる方法があり、現在検討が進められている。⁷⁾

一方、音声生成のメカニズムにもとって考えると、音声スペクトル上の、異なる音韻について共通の構造を有する個人差の原因として、声帯波スペクトルと声道長の効果が上げられる。これらは厳密には音韻によってその効果が異なるであろうが、近似的には共通な効果を有すると考えても差支えないであろう。また声帯波スペクトルの寄与は、主として音声スペクトルの概形(全体的な傾斜)として、声道長の寄与は音声スペクトルの周波数軸方向の伸縮として知っていると考えられる。本報告では、個人差をこのような意味での声帯波スペクトルの声道長の効果と考え、この正規化を行う方法と、単語認識におけるこの正規化の有効性について検討した結果について述べる。

3. 認識系

3.1. 認識系の構成 ここでは、最大ス
ペクトルパラメータを用いた認識系によ
って10数字の単語認識を行う。認識系
の基本的な構成は文献1)と同じであり、
そのブロック図は図2の通りである。

認識に必要な音響のスペクトルを標準パターンとして最尤スペクトルパラメータの形で登録し、入力音声の自己相関係数 $\{\rho_T^{(i)}\}_{T=0}^P$ と各音響の最尤スペクトルパラメータ $\{A_T^{(i)}\}_{T=0}^P$ とから、次式で定義す

れる尤度を計算し、これを類似度として用いる。 x は標準スペクトルの音韻名、 j は入力音声の区分(フレーム)番号、 p は極の数で、 $p=9$ とした。標準ボタンとして登録する音

$$l_j(x) = -\log \left(\sum_{\tau=0}^p A_{\tau}^{(x)} \rho_{\tau}^{(j)} \right) \quad (1)$$

Figure 1 is a block diagram of the word recognition system. It shows the flow from input signal to word decision. The process involves correlation analysis, likelihood calculation using phonetic standard patterns, and then summing likelihoods for each word in the dictionary to make a final decision.

図2. 認識系の構成

綴は、/a/、/i/、/u/、/e/、/o/、/n/、/N/、/g/、/r/
の9種類である。ただし調音結合の影響を

考慮して, /a/, /i/, /u/ については2種類の標準ボタンを用意している。

3.2. 個人差の正規化 個人差の正規化を行う方法としては、図3に示すように、入力音声の相関係数 $\{\rho_{\tau}^{(i)}\}_{\tau=0}^P$ について行なう方法(方法1)と、各音韻の標準パターンの最尤スペクトルパラメータ $\{A_{\tau}^{\alpha}\}_{\tau=0}^P$ について行なう方法(方法2)が考えられる。

3.2.1. 方法1—入力音声の正規化— 学習サンプルを用いて、認識すべき話者の掩蔽の音幅の区間を抽出し、各音幅の平均相関係数を求める。まずスペクトルの概形(傾き)の正規化を行うために、入力音声の各区分毎あるいは上記の平均相関係数を用いて、スペクトルの概形を抽出する。簡単のために、ここでは1次系または2次臨界制動系(2次CD系)の零点を用いた有理関数で近似し、この特性の逆フィルタを構成する。逆フィルタの伝達関数

め、次に各音韻のスペクトルの傾きを付与した標準パターン(相関係数)を時間軸方向に伸縮する。

このようにして話者に適応化した標準パターンと入力音声の各区分との類似度を求め、認識を行なう。方法1と比べて、この方法には個人差の正規化の程度を音韻毎に変えられるという長所があるが、スペクトルの概形を零点をとりたばい有理関数で近似する場合には、平坦なスペクトルを有する標準パターンに傾きを付与するフィルタのインパルスレスポンスは有限長とならないため、近似をよくするには極めて大きな遅れまでの相関係数と、やや複雑な計算が必要になるという欠点がある。簡単のためにフィルタの特性を1次系に限るとすると、その伝送関数は $1/(1+\beta Z)$ 、 $\beta = -\rho_1$ であらわされ、インパルスレスポンスを b_s であらわすと

$$1/(1-\rho_1 Z) = \sum_{s=0}^{\infty} b_s Z^s, \quad b_s = \rho_1^s \quad (3)$$

となる。従って、フィルタに通したのちの信号の自己相関関数を得るためには、厳密には、入力信号とインパルスレスポンスのそれぞれの無限の遅れまでの自己相関関数が必零であるが、ここでは便宜上インパルスレスポンスを $s=20$ までで「ハミングウィンドウ」をかけあわせて打ち切り、その自己相関関数は、遅れて $s=10$ までを用いた。

4. 予備実験

4.1. 声道長伸縮比と基本周波数の関係

男性30名、女性30名および子供8名について、上で述べた方法により、5母音と子音 /n/, /N/, /g/, /r/ の9種類の各音韻の声道長伸縮比を求めた。その結果、子音の伸縮比は同じ話者でも音韻によって大きくばらつくが、母音の伸縮比のばらつきは比較的小さく、その標準偏差は、男性、女性、子供の各群の平均声道長の5~8%で、音韻による偏りは特になかった。また分散分析を行なったところ、音韻間の分散に比べて、話者間の分散が大きいことが確認された。次に各話者の伸縮比の値を5母音について平均し、各話者の平均基本周波数との関係を調べた。結果は図5に示す通りで、両者には強い負の相関があり、相関係数は-0.92であった。この図には、男性、女性、子供の各群の平均値を結ぶ直線と、その±0.1の範囲を破線で示してあるが、各サンプル点はこの範囲内に入っており、伸縮比を L 、基本周波数を f_0 (kHz) であらわすと、この両者にはほぼ、 $1.2-2f_0 \leq L \leq 1.4-2f_0$ の関係がある。入力音声は男性、女性、子供のいずれによるものかが不明の場合には、過度の伸縮を防ぐのにこの関係が役に立つと考えられる。

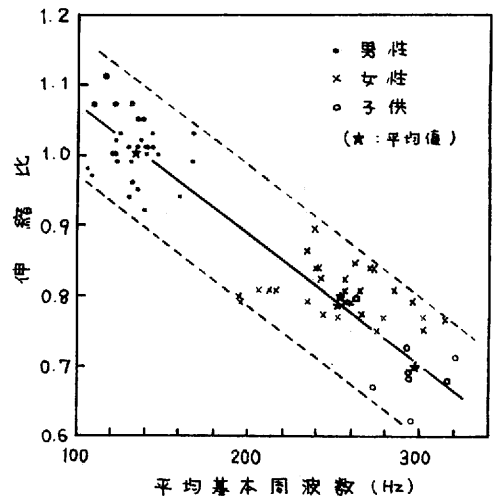


図5. 声道長伸縮比と基本周波数の関係

であった。この図には、男性、女性、子供の各群の平均値を結ぶ直線と、その±0.1の範囲を破線で示してあるが、各サンプル点はこの範囲内に入っており、伸縮比を L 、基本周波数を f_0 (kHz) であらわすと、この両者にはほぼ、 $1.2-2f_0 \leq L \leq 1.4-2f_0$ の関係がある。入力音声は男性、女性、子供のいずれによるものかが不明の場合には、過度の伸縮を防ぐのにこの関係が役に立つと考えられる。

4.2. 方法1による認識実験 男性30名がそれぞれ10数字を10回ずつ発声した計3000サンプルを用いて、方法1による認識実験を行なった。学習サンプルとしては、便宜上各話者が別に10数字を1回ずつ発声したものを用了。スペクトルの概形を次の4通りの方法で近似して平坦化するとともに、声道長の正規化を行なった。逆フィルタの特性は、a)の方法では入力音声の区分毎に変化し、b)の方法では単語内で一定、c) d)の方法では発声者について固定される。

- a) 入力音声の区分毎にスペクトルを1次系で近似する
- b) 各単語の平均スペクトルを1次系で近似する
- c) 9音韻の平均スペクトルを1次系で近似する

表1. 誤認識行列

(a) 男性・正規化なし

入	出	1	2	3	4	5	6	7	8	9	0	認識率
1	298										1	99.7
2	1	289			1	1				9		96.0
3			290	6					4			96.7
4				296	1			3				98.7
5					2	290	8					96.7
6					2	5	291	1	1			97.0
7			5			3	1	291				97.0
8							4		295			98.7
9					14	5	3			278		92.7
0		1				1				5	293	97.7
平均認識率												97.1

(b) 男性・傾き・声道長正規化

入	出	1	2	3	4	5	6	7	8	9	0	認識率
1	297	1									1	99.3
2	1	294				1				5		97.7
3			292	6					2			97.3
4				299						1		99.7
5					2	291	7					97.0
6					2	8	288	1	1			96.0
7			3			1	2	294				98.0
8			3				1		295			98.7
9					2		3			295		98.3
0		1				1	1				297	99.0
平均認識率												98.1

方が認識率が良く、特に /i/、/u/ の改善が著しいが、子音については正規化によって逆に認識率が低下している。このことは、ここで用

いた正規化の方法が母音に主体をなした方法であるために、子音にそれを適用することによって悪影響が生じたものと思われる。これは、単語認識における誤認識の一つの原因になっていると考えられる。

6. 認識実験2 (方法2)

前節の結果から、母音についてのみ正規化を行ない、子音については多数の話者の平均値をそのまま標準パターンとして用いた方がよいのではないかと考えられる。前述のように、方法1では未知サンプルについて処理を行なうためにこのようなことはできないが、方法2ではこのような処理が可能である。そこで男性30名の3000サンプルを用いて、方法2によって単語認識実験を行なった。母音の標準パターンについてはスペクトルの傾きと声道長の効果を話者に適応させ、子音の標準パターンについては多数話者の平均値を用いた。その結果、スペクトルの傾きの正規化のみの場合の認識率は方法1と変わらず、声道長の効果まで正規化した場合の認識率は、方法1に比べて0.1%の改善になった。

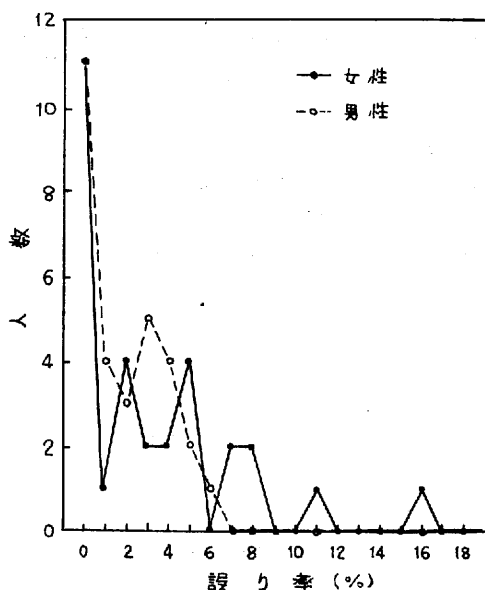


図8. 話者による誤り率の分布 (傾き・声道長正規化)

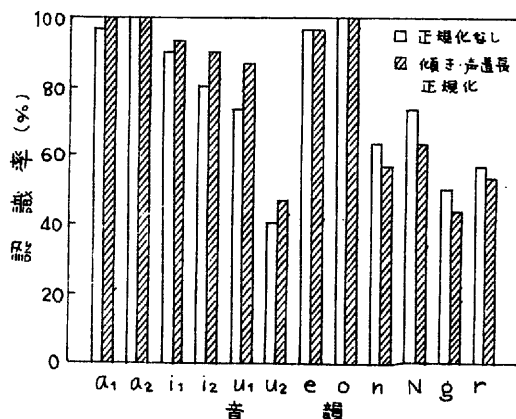


図9. 音韻認識実験の結果