少数の音声データによる音声標準パタン作成方法

5V－7　　○佐野　常世　　山本　稔　　和田　誓一
（東京電力株式会社）　（沖電気工業株式会社）

１．はじめに

離散単語音声認識において高性能な認識を実現させるためには、単語ごとの音声波形の特徴を十分に表現させた、良好な音声標準パタンを作成する必要がある。音声標準パタンは、実際に収集した多数の音声データをクラスタリング化・平均化することにより作成される。しかし従来の音声標準パタン作成方式では、数百～千人程度もの多数の音声データを収集しなければ単語ごとの発音･発声スピード･ゆらぎ等の発声個人差を的確に表現させることができず、非常に多大な労力を費やしていた。以下本論ではこのような問題点を解決させる一手法として、音声データの正規性の検定結果に基づき、統計的な推定手法を用いることによって数十人の少数の音声データから良好な音声標準パタンを作成させる方式を開発したので、その方式と評価結果について述べる。

２．音声認識方式概要

図１に音声認識方式を示す[1]。音声分析部では、サンプリング周波数12kHz、フレーム周期10.67ms、チャネル数２１によるＢＰＦ分析を行う。特徴量抽出部では、音声データの特徴点を強調するためにローカルピーク、スペクトル変化、子音性などの６種類の特徴量を抽出し、各々の特徴量を周波数軸７分割、時間軸８分割の５６領域に分割することにより、6×56＝336次元のデータを得る。マッチング部では、あらかじめ単語ごとに作成された音声標準パタンと、前述の入力音声データとの距離計算を行い、認識結果を得る。

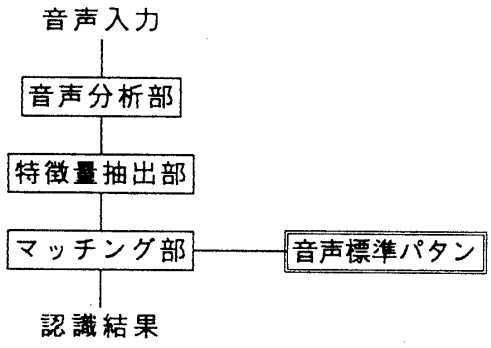

図１　認識処理手順

このようにして３２単語を認識対象とした不特定話者用離散単語音声認識を実現している。

３．音声データ分布の正規性

本論では音声標準パタンの作成に統計的な推定手法を用いているが、そのためには音声データの分布状態が正規性を有することが前提条件となる。そこで認識対象語である３２単語それぞれについて、今回収集した男性８３人分の音声データの分布状態を調べた。具体的には、８３人の男性に発声してもらった認識対象語ｘ(x=1~32)の実際の音声データを、前章で示した方法により全て３３６次元の数値データｘ[i,j](i=1~336,j=1~83)へ変換し、任意の単語の任意の次元ｘ[i]についてｘ[j](=1~83)を抽出し、その分布（すなわち、32単語×336次元＝10,752種類それぞれにおける83人分の音声データ分布）の正規性を調べた。以下に正規分布の検定に用いたSkewnessとKurtosisの式を示す。

[Skewness]

$$\text{Skewness} = \frac{1}{n}\sum\left(\frac{x_j-\overline{x}}{s}\right)^3$$

[Kurtosis]

$$\text{Kurtosis} = \sum\left(\frac{x_j-\overline{x}}{s}\right)^4 \times \frac{n(n+1)}{(n-1)(n-2)(n-3)} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

ここで、ｎはデータ数、$\overline{x}$はデータの平均、ｓは標準偏差、x_jはデータの実態(j=1～83)とする。一般的には、SkewnessとKurtosisの値の絶対値が双方ともに小さい値を示す場合に、その分布は標準正規分布に近いと言われている。表１に今回の計算結果を示す。なお、全要素が"０"となった４４０種類のデータについては区間推定の対象とならないため、計算結果から除いた。

表１　SkewnessおよびKurtosisの計算結果

u	1.0	1.5	2.0
対象データ数	4,426	5,597	6,202
データ比率[%]	42.9	54.3	60.1

(注1)　|Skewness & Kurtosis| ≦ u
(注2) 調査ﾃﾞｰﾀ数 ＝ 10,312

A Study of Making Templates for Word Recognition from Small Speech Database
Tsuneyo Sano　Minoru Yamamoto (Tokyo Electric Power Company)
Seiichi Wada (OKI electric ind. Co., Ltd)

この結果から、全体の約６割のデータがSkewnessとKurtosisともに２以下の値を示しており、正規性が高いことが分かる。また、残りのデータについても、１カテゴリの分布は正規性の高い各クラスタごとの分布で構成された混合分布と見なすことができる(詳しくは他の機会に委ねる)。したがって、区間推定法による効果の期待が持てることが予想される。

４．音声標準パタンの作成

図２に音声標準パタンの作成処理手順を示す。まず全データを所定のマルチテンプレート数と同数になるようにクラスタリングを行い、各クラスタごとのデータを平均化し、仮の標準パタンとする。データ数が多ければこれを標準パタンとしても充分な結果を得られることが報告されている[2]。今回は、次にテンプレートごとの位置関係を壊さぬように仮の標準パタンを区間推定法によって拡張させ、その結果を真の標準パタンとして登録する。

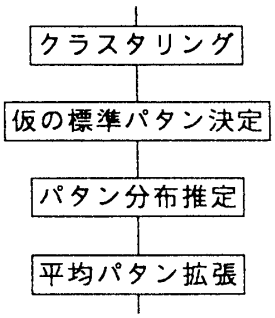

図２ 音声標準パタン作成手順

図３に区間推定法によるパタン分布推定方式の概念図を示す。ここで、分布１は今回使用した実データの分布、分布２は分布１を区間推定法によって拡張した分布、分布３は多数の学習データを用いた場合に予想される分布、t_1, t_2は信頼区間、ａ,ｂ,ｃは各クラスタを代表する仮の標準パタン、Ａ,Ｂ,Ｃは仮の標準パタンの位置関係を壊さぬように拡張して作成した真の標準パタンとする。なお図３は説明の便宜上、次元数１、ﾏﾙﾁﾃﾝﾌﾟﾚｰﾄ数３で示してある。

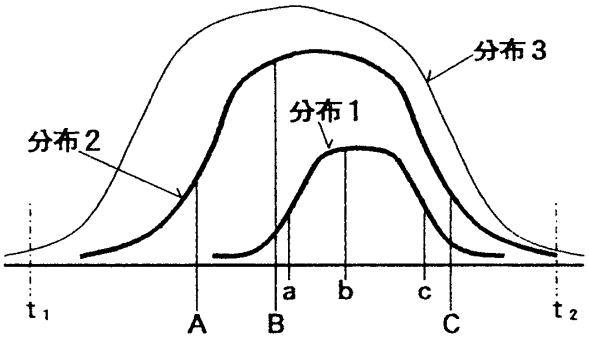

図３ 標準パタンの分布推定

具体的な計算方法は、学習データx_j(j=1~83)を正規分布と仮定した母集団からの任意標本集団と見なせば、母集団分布ｐ(ｘ:θ)から抽出した母平均μの(1-α)×100%の信頼区間を、以下に示す区間推定法の式によって求める。

$$\bar{x} - t_0 \frac{S}{\sqrt{n-1}} < \mu < \bar{x} + t_0 \frac{S}{\sqrt{n-1}}$$

ここで、t_0は推定の信頼度を示し、自由度(n-1)のｔ分布に従う係数とする。ここで求めた母平均の推定区間の範囲内で学習データの位置関係を崩さぬように学習データ分布を拡張し、パタンａ,ｂ,ｃをそれぞれＡ,Ｂ,Ｃに置き換え、それを標準パタンとして登録する。

５．認識評価

第２章で述べた認識方式を用いて、各クラスタごとのデータを平均化して作成した標準パタン(図3のａ,ｂ,ｃ)と、統計的推定手法を用いて作成した標準パタン(図3のＡ,Ｂ,Ｃ)の認識評価実験を行った。実験条件は、認識カテゴリを10数字,ﾊｲ,ｲｲｴ,制御語等の３２単語、マルチテンプレート数を１カテゴリあたり８パタンとし、男性８３名の音声データを用いたオープンおよびクローズによる評価である。その結果を表２に示す。

表２ 認識評価実験結果

標準パタン種類	ｵｰﾌﾟﾝ	ｸﾛｰｽﾞ
従来方式で作成	89.69 %	98.09 %
推定法を利用して作成	92.77 %	98.84 %

いずれも統計的推定手法を用いて学習データ分布を拡張させた場合のほうが良好な認識結果となっていることが分かる。

６．まとめ

音声データが正規性の高い分布状態を示していることに基づき、少数の音声データから統計的推定手法を利用して不特定話者離散単語音声認識用の音声標準パタンを作成する方法を提案し、その有効性を示した。今後は、この方式を組み込んで開発したアプリケーションシステムを利用して、実フィールドにおける本方式の有効性を検証し、更なる認識率の向上を目指したい。

参考文献

［１］ 和田,梅澤,山本："少数の学習データによる不特定話者用標準パタン作成方法", 日本音響学会講演論文集,2-Q-13,pp.183-184,1992

［２］ 山田他："複数マッチング特徴による不特定話者音声認識方式の検討", 音響学会講演論文集,1987