日産科学振興財団研究報告書, 14 (1991)

確率的手法に基づく不特定話者単語音声の認識

Speaker-independent word recognition based on stochastic models

代表研究者　東北大学工学部情報工学科助手　下平　博
Res. Assoc., Dept. of Information Eng., Fac. of Eng., Tohoku Univ.
Hiroshi SHIMODAIRA

A speaker-independent isolated wored recognition system is described, which is based on the use of intra-word local and global structural features. The local features are incorporated into the system with usnig matrix quantization of segment patterns obtained by dividing speech word pattern into partial patterns of equal frame length of about 4. Then global features are incorporated by making use of the stochastic correlations between segments. On a database with a vocabulary of 212 words, the system shows higher performance in recognition accuracy and processing speed than the system using whole-word template based dynamic time warping (DTW) algorithm.

研究目的

近年の音声認識技術は計算機技術の発達に助けられた確率統計的手法および最適化手法によって著しい発展を遂げている。特に連続音声認識の分野においては曖昧性の多い音素認識の結果を言語情報ならびに知識の利用によって補うさまざまな手法が研究されている。しかし，言語知識による音声理解は計算量の爆発という問題に直面し，実用化の面ではいかに言語処理以前の段階で候補を削減するかが重要な課題となっている。このような状況の中で本研究は統語情報を必要としない単語認識技術に対象を絞り，この処理段階における認識率の向上を研究の目的とする。単語認識技術の研究課題として，話者の不特定化および語彙数の増加による認識率低下の問題がある。本研究ではこれらの問題の解決にあたって認識の基本単位として信頼性の低い音素認識を使用することをできるだけ避け，代わりに物理的な単位である単語中の部分パターンを用いている。そして部分パターン間の関係を確率的に表現することによって，単語の大局的な構造と局所的な特徴を捉え，話者の変動ならびにパターンの変形に強い単語認識の手法を開発する。

研究経過

不特定話者環境における評価実験は膨大なデータと計算時間を必要とするが，効率的な研究を行うために，研究の当初にさまざまなモデルを考案し，少人数音声資料による予備実験を行いモデルの選択と種々のパラメータの検討を行った。ついで話者数を増加し総合的な評価を行うとともにモデルの拡張について検討した。

研究成果

1. 前処理

フレーム長 T からなる単語音声の入力特徴ベクトル系列（本研究では10次のメルケプストラム係数）を $X=(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ とし，線形伸縮によって L フレームに変換したものを $Y=(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L)$ と表す。以後，Y を単語音声パターンあるいは単に単語音声と呼ぶ。

次に，単語音声 Y を時間方向に M 個の区間（セグメント）に物理的に等分割し，このとき m 番目 ($1 \leq m \leq M$) の区間（以後区間 m と呼ぶ）の部分パターンを $\boldsymbol{u}^{(m)}$ で表す。各区間に重複がない場合，区間長 J は $J=L/M$ となる（ただし，J が整数となるように L，M を選ぶものとする）。したがって，単語音声パターン Y は以下のように表

せる。

$$Y=(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L)=(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}, \ldots, \boldsymbol{u}^{(M)}) \tag{1}$$

2. モデルの定義

各セグメント $S_i, i=1, \ldots, M$ における部分パターンを表す確率変数を $\mathrm{U}^{(i)}$ とし，その観測値を $\boldsymbol{u}^{(i)}$ とする。任意の二つのセグメント S_i, S_j において部分パターン $\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)}$ が同時に発生する確率 $p(\mathrm{U}^{(i)}=\boldsymbol{u}^{(i)} | \mathrm{U}^{(j)}=\boldsymbol{u}^{(j)})$ を簡単のために，$p(\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)})$ と記す。観測系列

$$Y=(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L)=(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}, \ldots, \boldsymbol{u}^{(M)})$$

が単語 W_r から生成される確率は，$p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r)$ と書ける。単語の認識は，以下のように事後確率最大の単語 W^*を求める問題として定義される。

$$W^*=\arg\max_{W_r} (W_r | \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}) \tag{2}$$

ここで，Bayes の法則を適用すれば，

$$p(W_r | \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}) = \frac{p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r) p(W_r)}{p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)})} \tag{3}$$

なる関係があり，単語の発生確率が単語に関係なく一定であると仮定すれば，上式の $p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r)$ のみ計算すればよく，

$$W^*=\arg\max_{W_r} p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r)$$

なる単語 W^*を求めればよいことになる。

さて，実際に $p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r)$ を求めることは現実的ではないので，一般には $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}$ に何等かの従属関係を仮定することになる。まず，各セグメントが互いに独立であると仮定すると，

$$p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r) = \prod_{i=1}^{M} p(\boldsymbol{u}^{(i)} | W_r) \tag{4}$$

となる。

次に，単純マルコス性を仮定し，状態と観測シンボルを 1 対 1 に対応させれば，

$$p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r) = p(\boldsymbol{u}^{(1)} | W_r) = p(\boldsymbol{u}^{(1)} | W_r) \prod_{i=2}^{M} p(\boldsymbol{u}^{(i)} | \boldsymbol{u}^{(i-1)}, W_r) \tag{5}$$

が得られる (Model 1)。

さらに，マルコフ・モデルを一般化したものとしてデンドロイド (dendroid) 分布を用いることができる (Model 2)。デンドロイド分布とは，2 次元周辺分布がすべて既知であるとき，各セグメントをグラフ上の頂点（ノード）とみなしたときの最小木 (minimum spanning tree) に対応する。すなわち，木の枝 (edge) の組合せの列を

$$\boldsymbol{k}=((k_{11}, k_{12}), (k_{21}, k_{22}), \ldots, (k_{n1}, k_{n2}))$$

としたとき，以下の式を満たす $\boldsymbol{k}$ を求める問題に帰着する。

$$p(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)} | W_r) = \max_{\boldsymbol{k}} p(\boldsymbol{u}^{(k_{11})} | W_r) \prod_{i=1}^{n} p(\boldsymbol{u}^{(k_{i2})} | \boldsymbol{u}^{(k_{i1})}, W_r) \tag{6}$$

ただしここで，$n=M-1$ である。

これらの二つのモデルは状態と観測シンボルを 1 対 1 に対応させているため，パターンの時間的，空間的な変動に弱いと予想される。そこで，この問題に対して以下のような同時確率を利用したモデルを設定する (Model 3)。入力パターンと単語 W_r を記述するモデルとの一致の程度を評価関数 R_{W_r} で表し，

$$R_{W_r}(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}) = \prod_{(i,j)\in\Psi} p(\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)} | W_r) \tag{7}$$

と定義する。従って，認識単語 W^*は，

$$W^*=\arg\max_{W_r} R_{W_r}(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(M)}) = \arg\max_{W_r} \prod_{(i,j)\in\Psi} p(\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)} | W_r) \tag{8}$$

として定義される。ここで，Ψ は考慮するセグメント組の集合を表しており，最大で全セグメントの組合せを表す。すなわち，上式は二つのセメグメントに同時に観測されるシンボル組の同時確率を Ψ に含まれるすべての組合せについて積を取ることを示している。なお，上式の右辺の，$p(\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)} | W_r)$ は事後確率 $p(W_r | \boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)})$ としても一般性を失わない。

上記の確率分布の計算にあたっては各セグメント単位のベクトル量子化に基づいた離散型モデルを用い，Fuzzy クラスタリングで用いられるマルチ・ラベリング法に基づく平滑化法を適用する。

3. 認識実験結果

212 単語音声資料を用いて単語認識実験を行っ

表 1. 認識率の比較（話者 20 名）

L: 量子化数

Recognition methods	L	Percent correct				
		Top	2	3	5	10
MDM	16	89.7	96.0	97.5	98.6	99.3
MDM	32	92.3	97.2	98.2	98.9	99.5
MDM	(∞)	93.6	97.4	98.3	99.1	99.5
DTW	(∞)	94.8	98.1	99.0	99.4	99.9
Model 1 (Markov)	16	93.9	97.3	98.5	99.0	99.5
Model 2 (Dendroid)	16	94.8	97.8	98.5	99.2	99.6
Model 3 (Ψ_0)	16	92.7	97.3	98.5	99.0	99.6
Model 3 (Ψ_1)	16	94.4	97.7	98.4	99.0	99.5
Model 3 (Ψ_*)	16	95.4	98.4	99.0	99.4	99.7

た。比較実験として，確率を用いずに単純に参照パターンとのベクトル量子化歪が最小となる単語を選択する手法 (MDM) ならびに非線形時間軸伸縮 (DTW) による単語マッチング法を用いる。話者 20 名（男女各 1 名）による認確実験結果を表 1 に示す。量子化歪最小法 (MDM) と比較して確率モデル (Model 1〜Model 3) は認識率が高く，なかでも同時確率を利用した Model 3 が最も高い認識率を示している。特に考慮するセグメントの組合せが多いほど認識率は高くなる傾向がある。また，この時 DTW 法の認識率を若干上回っている。

さらに，Model 3 (Ψ_*) において話者数を 58 名に増加させたとき，認識率は量子化数 $L=16, 32, 64$ についてそれぞれ 94.7, 95.6, 96.2% が得られた。このことから，話者数の増加に対しても本手法が有効であるとともに，量子化数を増加することによって認識精度を向上させることができることが分かった。

今後の課題と発展

本研究で用いた単語内部におけるセグメント間確率モデルは量子化数を増加するほど認識性能が向上する傾向にあるが，計算処理に必要とされる記憶メモリ容量は量子化数の 2 乗にほぼ比例して増加する問題がある。また，量子化数の増加は確率分布の推定精度を低下させる問題もある。そのため，さらに高精度のモデルを開発するには離散的な確率モデルを連続分布型のモデルに拡張する必要がある。

なお本研究で開発された手法は単語認識のみならず，音素等の小さな単位の認識にも適用可能であり，現在音素認識における同様の手法を開発中である。

発表論文

下平　博，掘内芳雄，黄翠凌，木村正行："不特定話者単語音声認識における構造情報の利用"，第 16 回応用情報学研究センター・シンポジウム (1990-05).

H. Shimodaira and M. Kimura: "Speaker-Independent Isolated Word Recognition Using Local and Global Structural Features", 1990 International Conf. on Spoken Language Processing, 13-12 (1990-11).

下平　博，掘内芳雄，木村正行："構造的特徴を利用した音声認識"，第 27 回東北大学電気通信研究所シンポジウム論文集，2-4, pp. 101〜110 (1991-03).