音声の構造的表象に基づく音響的照合に関する実験的検討 *

◎丸山和孝 (東大・工) 村上隆夫 (東大・情報理工) 峯松信明 (東大・新領域) 広瀬啓吉 (東大・情報理工)

1 はじめに

計算機技術の進展及び大規模データベースの構築により、音声認識技術は格段に進歩した。しかし技術の安定性という観点から見た場合、まだその頑健性は十分満足できるレベルに達していない。従来の音響モデリングは、対象とする音素の音響的実体（スペクトル包絡）を統計的にモデル化する手法のみが検討されてきた。音響的実体をそのままモデル化した場合、話者性・マイク特性・環境特性などの乗算性歪みや線形変換性歪みが不可避的に混入する。従来手法では、これらの歪みを「集める」ことで解決を図ってきた。しかし、不特定話者モデルと言えども話者適応技術を必要とする現状を考えると、「集める」こと以外の解決策を模索する必要があると考えられる。

近年、音響的普遍構造という、線形変換性・乗算性の歪みを原理的に持ち得ない音声表象が提案されている [1]。音声事象そのものをモデル化するのではなく、対象とする事象群の関係だけをモデル化する本手法は、構造音韻論の物理実装として、あるいは、音声言語ゲシュタルトとして解釈される [1]。本論文では、音声の構造的表象を用い、音素の音響的実体を明示的に用いない音声認識に関して基礎的検討を行なった。

2 音響的普遍構造とそれに基づく音響的照合

任意の音声事象をケプストラムから成る混合ガウス分布とし、2 事象間の距離をバタチャリア距離の平方根で表す。バタチャリア距離は分布 u、v に対し

$$BD(u,v) = -\ln \int_{-\infty}^{\infty} \sqrt{p_u(x)p_v(x)}dx \qquad (1)$$

で与えられる。この距離は共通の一次変換 ($c' = Ac + b$) に対し不変である。A は声道長や聴覚特性差異などを表し、構造の回転となり、b は収録機器・収録環境の差異などに相当し、構造の遷移となる。n 事象から成る構造は全事象間距離（${}_nC_2$ 個）を求めることで（距離行列）一意に決まり、この構造も不変である。

二つの「n 事象から成る構造」に対して、その構造間距離は、一方の構造を回転及び遷移させて他方の構造に近づけ「対応する点間距離の総和」の最小値で定義される。ユークリッド空間における構造間距離尺度は [1] において解析的に求められているが、(**1**) 式は非ユークリッド空間を張るため、ここでは近似的な距離尺度を用いる。文献 [2, 3] において、距離行列

表 1. 分析条件

サンプリング	16bit / 16kHz
窓	窓長 25msec、シフト長 10msec
分析区間	母音中心部の 1 モーラ相当区間 (140msec)
パラメータ	MFCC(1〜12 次元)

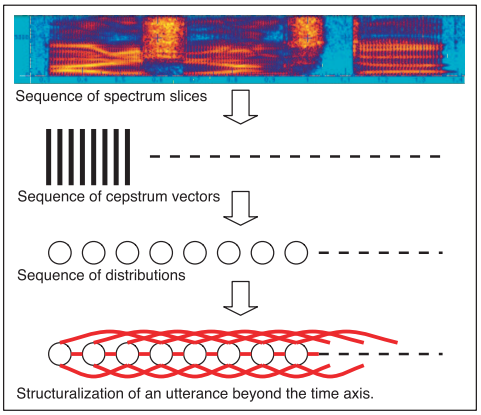

図 1. 一発声の構造化

をベクトルと見なしたときのユークリッド距離が近似的な尺度になることが示されており、本研究でも、距離行列を特徴ベクトルとして使用することとする。

3 音声の構造的表象を用いた音声認識実験

3.1 使用した音声資料

日本人男性 4 名女性 4 名の計 8 名が、日本語 5 母音を 5 回ずつ孤立発声した音声を用いた。分析条件を表 **1** に示す。各母音に対し、その中心部の一モーラ相当部分の音声事象を単一ガウス分布でモデル化する（事象モデル）。なお、分散行列は対角とした。

3.2 認識タスクと構造特徴ベクトル

認識語彙としては、系列長 5 の孤立母音発声とした。同じ母音は二度出現しないとの制約を加えて語彙を構成した（語彙数=${}_5P_5$=120）。各語彙発声は距離行列で表現され，対角以外の上三角成分を特徴ベクトルとした。各語彙に対して、構造特徴ベクトル群から構造統計モデルを構築した。学習話者は評価話者を除く 7 名で、5 母音 5 回ずつの孤立発声があるため、各語彙に対し、学習データは学習話者毎に $5^5 = 3,125$ 個、計 21,875 個になる。以上を用いて、混合数 1、2、4 のガウス分布（全角分散行列を使用）を推定した。

3.3 音声事象分布のパラメータ推定法

提案手法は、認識器に与える入力データも構造として表象される必要があるため（一発声の構造化。図 **1** 参照）、例えば、一単音を分布として捉える必要がある。分布推定に用いるデータ量が絶対的に少ないため、ML 推定の他に、MAP 推定も検討した [4]。

MAP 推定の事前確率密度分布は、評価話者の音声事象に対する分布推定時は、学習話者（計 7 名）の全母音音声から作成した。学習話者音声の構造化の際に必要となる事前確率密度分布は、評価話者と当該学習話者以外の話者（計 6 名）の全母音音声を用いた。実験時は、MAP 推定時の重み w として種々の値を使用して検討した（$w = \infty$ で ML 推定）。

---

* Experimental study of acoustic matching based on the structural representation of speech
By Kazutaka Maruyama, Takao Murakami, Nobuaki Minematsu and Keikichi Hirose (University of Tokyo)

表 2. 音声の構造的表象のみによる認識結果 (%)

構造統計モデル＝単一ガウス分布

推定法＼順位	1	2	5	10
ML	19.7	39.5	69.3	89.2
MAP(w=10)	29.7	55.7	84.0	96.8
MAP(w=1)	35.9	82.1	96.6	100.0
MAP(w=0.1)	43.2	88.5	98.2	100.0
MAP(w=0.01)	53.0	98.9	100.0	100.0

構造統計モデル＝混合ガウス分布 (混合数 2)

推定法＼順位	1	2	5	10
ML	21.1	39.3	69.4	87.3
MAP(w=10)	29.1	52.2	80.8	96.0
MAP(w=1)	53.8	84.0	95.6	100.0
MAP(w=0.1)	59.1	87.9	99.3	100.0
MAP(w=0.01)	58.5	97.2	100.0	100.0

構造統計モデル＝混合ガウス分布 (混合数 4)

推定法＼順位	1	2	5	10
ML	19.1	33.8	62.4	85.1
MAP(w=10)	23.4	39.9	74.8	93.5
MAP(w=1)	43.7	72.9	93.5	99.3
MAP(w=0.1)	57.8	92.1	100.0	100.0
MAP(w=0.01)	60.9	93.9	100.0	100.0

3.4 音声の構造的表象のみを用いた認識実験

入力音声は 25,000 個の「あいうえお」のみである。それ以外の母音列は「あいうえお」構造を表現する特徴ベクトルの要素を置換することで得られるため、実験に用いる必要は無い。構造統計モデルのみによる認識実験の結果を表 **2** に示す。第一位正解率以外に、ある順位までに正解が含まれる率も示している。ML 推定の場合、認識率は約 20%程度と非常に低い。しかし、語彙数が 120 であることを考えると、ML 推定による事象分布群が成す構造にも語彙同定のための情報が含まれていることが分かる。

一方、MAP 推定による事象分布群が成す構造を用いた場合、ML 推定に比べて認識精度は格段に向上する。特に事前分布に対する重みを小さくするほど認識率が上昇する（但し、重み w を 0.0 にすると、事象分布が事前分布と等しくなるため、距離行列の成分は全て 0.0 となり、認識不能となる）。これは、各音声事象の少数サンプルデータを用いて、事前分布を僅かに修正することで推定される分布群が張る構造には、語彙同定に必要な情報が含まれていることを意味する。注目すべきは、提案照合手法は、音声事象の実体を直接的には用いていない点である。音声事象の実体を用いた場合、不可避な歪みによる影響を直接的に被るが、本照合では、原理的に影響を受け得ない。

構造統計モデルの混合数は結果に影響を与えていない。全角の分散共分散行列を用いたため、混合数 1 でも十分なモデル化能力が得られていると考えられる。

3.5 1 母音を既知とした場合の認識実験

構造的表象のみを用いた認識実験では、音響的照合において、音声の物理実体は全く参照していない。そこで、5 母音の中で 1 つの母音が既知とした場合の認識率について検討した。即ち、1 母音を既知とすることで、構造の回転・遷移の自由度を削減し、認識率の向上を検討した。結果（第一位正解率）を、表 **3** に示す。「う」または「え」を既知とした場合の認識率が高いが、これは誤認識結果の多くが「あいえうお」であったことに因る。そこで、「あいえうお」も正解に含めた場合の認識率を求めた。表 **4** に示す。音声の実体を直接用いなくても、ほぼ 100%の性能で語彙を 2/120 まで削減することができている。

表 3. 一母音を既知とした時の認識結果 (%)

構造統計モデル＝単一ガウス分布

推定法＼既知母音	a	i	u	e	o
ML	33.8	27.5	58.6	55.6	35.9
MAP(w=10)	39.6	35.8	72.0	72.0	41.6
MAP(w=1)	40.1	36.7	94.0	94.9	41.4
MAP(w=0.1)	44.3	43.3	98.2	98.1	45.3
MAP(w=0.01)	53.9	53.2	99.4	99.6	54.0

構造統計モデル＝混合ガウス分布 (混合数 2)

推定法＼既知母音	a	i	u	e	o
ML	35.6	30.8	56.6	55.5	37.2
MAP(w=10)	39.0	35.7	67.6	70.9	41.4
MAP(w=1)	59.0	55.7	94.1	95.0	60.4
MAP(w=0.1)	60.7	59.1	98.2	97.7	61.7
MAP(w=0.01)	58.9	58.6	99.7	99.8	58.9

構造統計モデル＝混合ガウス分布 (混合数 4)

推定法＼既知母音	a	i	u	e	o
ML	38.7	30.7	49.3	46.2	40.3
MAP(w=10)	41.1	34.0	58.6	60.1	45.1
MAP(w=1)	53.5	48.8	86.8	90.5	56.5
MAP(w=0.1)	60.0	58.5	98.1	97.2	62.0
MAP(w=0.01)	61.4	61.2	99.0	99.9	61.4

表 4. 「あいえうお」を含む認識結果 (%)

推定法＼混合数	1	2	4
ML	47.9	46.5	38.7
MAP(w=10)	66.0	59.8	43.8
MAP(w=1)	93.3	91.3	79.2
MAP(w=0.1)	97.9	97.3	94.5
MAP(w=0.01)	99.1	99.6	99.2

4 まとめ

不可避な音響歪みを保有しない音声表象である音響的普遍構造を用いた音声認識を検討した。その結果、音声の実体を直接的に音響的照合に用いなくても、十分に音声を識別する性能が得られることを示した。今後、子音を含む連続音声に対する構造化、加算性の雑音が混入した音声、電話音声、子供音声など、様々な音声を対象として本手法の性能を検討したい。

参 考 文 献

[1] 峯松他、音講論、1-5-13、pp.25–26 (2005-3)

[2] 峯松，信学技報，SP2003-180，pp.31–36 (2004)

[3] 村上他，音講論，2-P-9，pp.379–380 (2004-9)

[4] C.H. Lee et al, IEEE Trans. Signal Processing, vol.39, no.4, pp.806-814 (1991)