社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＨＩ－109　(23)
2004－SLP－ 52　(23)
2004／7／17

多次元尺度構成法による音響空間の２次元可視化

庄境　誠,　奈木野　豪秀

アブストラクト

MLLR などの話者適応手法により有効な効果を得るためには，十分な量の音声サンプルをユーザから取得する必要があるが，実用化の現場では困難な場合が多い．音声認識システムを初めて使う瞬間から高い認識性能を確保するには，十分高い認識性能を提供する，精密な音響モデルライブラリの事前開発が重要である．一般に，多次元ベクトルの正規分布で表現される HMM 音響モデルを分析することは，困難である．音響モデルを２次元平面上に可視化し，人間の視覚分析能力を利用した，精細な音響モデルライブラリの開発を支援する手法として，COSMOS(aCOustic Space Map Of Sound)法を提案する．性別，信号雑音比，タスク，発話様式などの分析を例に取り，音響空間の分析手法としての提案法の有効性を示す．

Two-dimensional Visualization of Acoustic Space by Multidimensional Scaling

Makoto Shozakai and Goshu Nagino

Abstract

In order to achieve sufficient results in speaker-adaptive techniques as represented by the MLLR method,　it is essential to obtain adequate voice samples of the user,　rendering the application of the method difficult in practical applications. To ensure recognition performance matching up to the level of required practicality from the outset,　prior development of highly precise acoustic model libraries for the voice recognition systems are necessary. The analysis of HMM acoustic models expressed as Gaussian distributions of multi-dimensional vectors is typically a difficult task. The COSMOS (aCOustic Space Map Of Sound) method featuring the visualization of the distribution of the acoustic models in a two dimensional diagram by use of multi-dimensional linear measurement is proposed as a technique to support the analysis through the utilization of human visual perception. The effectiveness of the proposed technique as a method of analyzing the acoustic distribution is reviewed based on examples of differences in sample gender,　signal-to-noise ratio,　tasks and styles of speech.

１．はじめに

カーナビゲーションシステムなどの車載機器，情報入力が困難な業務用ＰＤＡ，ヒューマノイドロボット，身体障碍者・高齢者支援のＩＴ機器など，組み込み機器での音声認識機能の需要は日増しに高まっている．HMM をベースとした不特定話者認識技術の実用化が加速しているが，現状では，全てのユーザに対して，平等に高精度の音声認識性能を提供しているとは言い難い．

音声認識性能は，ユーザの個人性（声帯，声道），タスク（語彙の難易度，語彙に含まれる音素バランス），発話様式，雑音環境（種類，大きさ）などの要因によってばらつきが起こる．実環境では，これらの変動要因が複雑に絡み合い，音声認識性能を全てのユーザに対して，予測，保証することは容易ではない．全ての状況に共通に用いることが可能な高精度の音響モデルを開発することがこ

旭化成株式会社　情報技術研究所

Asahi Kasei Corporation, Information Technology Laboratory

れまでの音声認識研究の目的であったが，その限界も見えてきた．いくら多量の音声コーパスを収集し，いくら複雑度の高い音響モデルを作成しても，全てのユーザに均一の音声認識性能を提供することは出来ていない．

そこで，研究が深められたのは，話者適応技術である．MAP(Maximum A Posteriori probability)法［1］，MLLR(Maximum Likelihood Linear Regression)[2]などの手法が研究されているが，話者適応の効果を得るためには十分な量の適応用サンプル音声が必要である．

1）ユーザの視点

しかしながら，実用化の現場では，十分な量の適応用サンプルを確保することが困難な場合が多い．なぜなら，大抵のユーザは十分な量の適応用サンプルを自分の時間を消費して提供することは面倒であると考えており，適応用サンプルを提供しなければならないことが妥当であるとは認めていないからである．全てのユーザは音声認識システムを初めて使う瞬間から高い認識性能を享受できることを当然のように期待している．音声認識ベンダーは，全てのユーザに最初の瞬間から高い認識性能を提供する技術を有していなければならない．十分な量の適応用サンプルを自分の時間を消費して提供しても良いから，自分に適応化された音響モデルを欲しいという動機を備えたユーザに対してのみ，現状の話者適応技術は有効であるといえる．

十分な量の適応用サンプル音声を取得することが困難な実用化の現場では，上記の変動要因に応じた音響モデルを予め複数用意しておき，ユーザの利用状況に応じて，音響モデルを選択的に利用する手法が有用であると思われる．例えば，服は，体型やファッションの好みに合わせて，様々なタイプの服が予め用意されており，お店でユーザが試着して，適切なサイズや好みの色の服を選択し手頃な価格で購入している．音声認識システムに用いられる音響モデルもユーザの使用状況に応じて，選択的に使用する仕組みの構築が望まれる．様々な話者の音声を複数の話者グループに分類して，音響モデルを作成する方法として，話者クラスタリング法[3]，クラスタ適応学習法[4]，EigenVoice(EV)法[5]などのクラスタリング手法が提案されているが，専門家にしか利用できない．

さらに，音声認識の対象は，人間の声ばかりとは限らない．実環境においては，人間の声だけではなく，自然発生的に雑音が生じる．SS(Spectral Subtraction)法［6］，CSS(Continuous Spectral Subtraction)法［7］[8]，MMSE(Minimum Mean Square Error)法[9][10]などの雑音抑制技術が進歩し，自動車内などでの音声認識機能の実用化が進んでいるものの，突発性の雑音や非定常な雑音を１００％抑圧することは出来ていない．従って，抑圧し切れない雑音の音響モデルを作成し，明示的に認識するという手法も有効である．人間の音声の音響モデルばかりでなく，実環境に存在する雑音の音響モデルの整備が今後不可欠になると考えられる．

2）メーカの視点

一方で，組み込み機器の製造メーカの視点に立ってみると，認識性能の予測，保証が困難な現状の音声認識手段は，キーボード，マウス，スイッチなどの入力手段に比べて，使いづらいという問題がある．

全てのユーザにとって，より信頼できる音声認識システムの実現を目指すには，各ユーザに好適な音響モデルを迅速に提供・維持する技術の確立が求められている．

3）音響モデルの需給バランスの視点

現在，音響モデルの開発は，音声認識ベンダーの音響モデル開発者の手に委ねられているが，音響モデル開発には，HMM や上記のクラスタリング手法に関する高度な知識が要求されるため，従事者の数は限られる．音声認識ベンダーによる音響モデルの供給能力は大幅に不足しており，音響モデルに対する市場の需要規模とのバランスは明らかに適正ではない．音響モデルの生産人口を増やすという観点は，今後ますます重要になると思われる．

本稿の目的は，上記の視点を踏まえ，音響モデルライブラリの開発に有効な音響空間の数学的分析手法を提案することにある．第２章で音響モデルの分布を２次元可視化し，音響空間の拡がりを把握する方法（COSMOS 法）を提案する．第３～８節では，提案法を利用して，性別，信号雑音比，タスク，音素，発話様式，生活雑音の異なりによる音響空間の違いについて論じる．最後に，第９節で，本稿についてまとめる．

２．COSMOS 法

カーナビゲーションシステムなどの組み込み機器の開発者から，音声認識ミドルウェア製品のベンダーに対して要求されることは，誰でもいつでもどこでも一定以上の認識性能が得られる製品の提供である．認識性能を高めるためには，特徴抽出処理や照合処理の改良が必要であるが，本稿では，後者の中の音響モデルに着目し，より精細な音響モデルを効率よく開発する手法を提案する．

音響モデルの精細化を効率良く行うためには，音声認識システムが対象とする入力（音声及び雑音）の音響空間の全貌を把握することがまず何よりも必要である．その

ためには，多次元情報を低次元空間に可視化する方法，すなわち，多次元尺度構成法（MDS:MultiDimensional Scaling）法[11][12]が極めて有効である．一般に，多次元情報を低次元平面に写像して可視化するMDSとして，以下の手法がある．

（１）クラスラベルなどの識別情報を用いる（教師あり）手法としては，判別分析法[13]，Aladjem 法[14]，ニューラルネットワークによる手法[15]，グラフを利用した手法[16]が提案されている．

（２）クラスラベルなどの識別情報を用いない（教師なし）手法としては，射影追跡法[17]，Sammon 法[18]，ＳＯＭ(Self-Organizing Maps)法[19]，主成分分析による手法[20]が提案されている．

これらの手法は全て多次元ベクトルを２次元平面に写像する手法であり，多次元正規分布を有する情報を２次元平面に写像することは出来ない．特に，主成分分析による手法[20]は，オランダ語の連続数字のタスク評価から，HMM による音響モデルの第１主成分と第２主成分を利用して，音響モデルを２次元平面上に写像する方法を提案している．主成分分析の結果として，第１主成分として性別，第２主成分として調音結合の異なりが対応していることを示している．しかしながら，第２主成分までの累積寄与率は，9.4%であり，主成分分析の累積寄与率の目安とされる 80%に比べて著しく小さく，得られた２次元散布図は, 元の多次元正規分布を有する情報の空間を忠実に表現しているとは言い難い．拮抗する寄与率を有する主成分が３個以上存在する音響モデルを対象とする場合，主成分分析による手法では，２次元可視化が困難である．多次元正規分布を有する情報を情報欠落なくそのまま２次元平面に写像できる手法が必要である．

多次元情報として，音声データの特徴パラメータベクトルを直接用いる方法もあるが，データ数が膨大であり，現実的ではない．ここでは，音声データから学習された音響モデルを音響空間の近似表現とみなし，音響モデルの分布を２次元平面上に可視化し，人間の視覚分析能力を利用して，精細な音響モデルを効率的に開発する MDS として，COSMOS (aCOustic Space Map Of Sound)法を提案する．本提案法は，HMM による音響モデルの集合を２次元平面上に非線形写像するように，Sammon 法[18]を拡張したものであり，統計的 MDS と位置付けられる．作成される２次元散布図を COSMOS と呼び，COSMOS に写像された音響モデルを STAR と呼ぶ．

Sammon 法は，高次元空間上の高次元情報の相互距離の総和と低次元空間上の写像位置座標の相互ユークリッド距離の総和の差が最小となるように，最急降下法により低次元空間上の写像位置座標を最適化する非線形写像手法である．相互距離が小さい２つの高次元情報は低次元空間上でも互いに近くに，相互距離が大きい２つの高次元情報は低次元空間上でも互いに遠くに位置するように全ての高次元情報を低次元空間に射影する．多次元情報を２次元平面に非線形に写像する場合の Sammon 法の定式化を付録に示す．

音響モデルは一般に，複数の音響単位のモデルの集合の総称である．そこで，音響モデル i と音響モデル j の相互距離 $D(i,j)$ を次式で定義する．

$$D(i,j) \equiv \sum_{K=1}^{K} d(i,j,k) * w(k) \Big/ \sum_{K=1}^{K} w(k) \quad (1)$$

ここで，$w(k)$ は音響単位 k の出現頻度を表す．K は音響単位の総数を表す．$d(i,j,k)$ は，音響モデル i に含まれる音響単位 k のモデルと音響モデル j に含まれる音響単位 k のモデルの相互距離である．

$d(i,j,k)$ としては，正規分布の平均ベクトルのユークリッド距離，バタチャリア距離，カルバック情報量などの公知の距離尺度を用いることが可能であるが，ここでは，認識性能に対する効果がカルバック情報量と同程度であることが明らかにされた[22]，正規分布の標準偏差の積で正規化された平均値ベクトルのユークリッド距離を利用する．音響単位の音響モデルは，混合正規分布を持ち，全て同一のトポロジーを共有するとする．音響モデルの状態アライメントが，１：１であると仮定すると，$d(i,j,k)$ は，次式で表現される．

$$d(i,j,k) = \frac{1}{S(k)} \sum_{s=0}^{S(k)-1} \frac{1}{L} \sum_{l=0}^{L-1} \frac{dd(i,j,k,s,l)}{pp(i,j,k,s,l)} \quad (2)$$

$$dd(i,j,k,s,l) = \sum_{m_i=0}^{M_i} \sum_{m_j=0}^{M_j} p(i,k,s,l,m_i) \cdot p(j,k,s,l,m_j) \cdot \frac{\left\{\mu(i,k,s,l,m_i) - \mu(j,k,s,l,m_j)\right\}^2}{\sigma(i,k,s,l,m_i) * \sigma(j,k,s,l,m_j)} \quad (3) \qquad \downarrow (4)$$

$$pp(i,j,k,s,l) = \sum_{m_i=0}^{M_i} \sum_{m_j=0}^{M_j} p(i,k,s,l,m_i) \cdot p(j,k,s,l,m_j)$$

ここで，$\mu(i,k,s,l,m)$，$\sigma(i,k,s,l,m)$，$p(i,k,s,l,m)$ はそれぞれ音響モデル i の音響単位 k の状態 s の次元 l の m 番目の正規分布の平均値，標準偏差値，重みである．$S(k)$ は音響単位 k の音響モデルの状態数を表す．D は音響モデルの次元数を表す．

以下では，組み込み機器向け音声認識ミドルウェア

VORERO（「ボレロ」）［21］に対し，COSMOS 法を適用した場合の例を紹介する．ＶＯＲＥＲＯは，組み込み機器向けに計算量・メモリサイズをコンパクトにするために，単一正規分布による diphone HMM を実装している．従って，(2)-(4)式は以下のように簡単化される．　　↓(5)

$$d(i,j,k) \equiv \frac{1}{S(k)} \sum_{s=0}^{S(k)-1} \frac{1}{L} \sum_{l=0}^{L-1} \frac{\{\mu(i,k,s,l) - \mu(j,k,s,l)\}^2}{\sigma(i,k,s,l) * \sigma(j,k,s,l)}$$

音響パラメータは，１０次元のＭＦＣＣ，１０次元のデルタＭＦＣＣ，１次元のデルタエネルギーである(D=21)..

本研究では，Simulated Annealing の手法を導入し，(A-4)式の α の値を繰り返し計算の途中で時折大きくすることにより，(A-2)式で定義される写像誤差 E_m が局所的最小値に捕捉されることのないように工夫している．

また，本提案法で作成される COSMOS にプロットされるのは，音響モデルの平均ベクトルの２次元空間上の位置である．正規分布の拡がりは，陽には表現されていない．COSMOS 上で２つの STAR が離れてプロットされているとしても，正規分布の拡がりが十分に大きい場合は，２つの STAR はほぼ同一の音響空間を表現していると見なすべきである．ある STAR の空間的な拡がりを近似的に表現する方法として，正規分布の平均ベクトルから標準偏差の整数分だけ±方向に変位したベクトルを生成し，それを平均ベクトルとし，元の正規分布の分散ベクトルとする STAR(n σ STAR と呼ぶ．但し，n は自然数）を作成して，本提案法により同一の COSMOS 上に重ね合わせる方法を採用する．D=21 の場合，次元毎に独立に標準偏差の３倍分だけ±方向に変位させると，2**21=2,097，152 通りの３ σ STAR が生成されることになり現実的ではない．そこで，ある１つの次元のみ標準偏差の３倍分だけ±方向に変位させることにすると，2*21=42 通りの３ σ STAR を生成される．これらの３ σ STAR を COSMOS 上にプロットすることにより，ある STAR の３ σ 空間を近似的に視察することが出来る．日本語５母音の定常部の男性不特定話者音響モデルについて，３ σ STAR 空間をプロットした例を図１に示す．

３．性別の分析

女性１５０名及び男性１３６名が，表１の通常の発話様式で発声した日本語１７６単語から作成された特定話者音響モデルから作成された COSMOS を図２に示す．男性と女性がそれぞれ明瞭なクラスタを形成していることから，一般に不特定話者音響モデルを男性用と女性用に分けて作成する方法の妥当性を裏付けている．

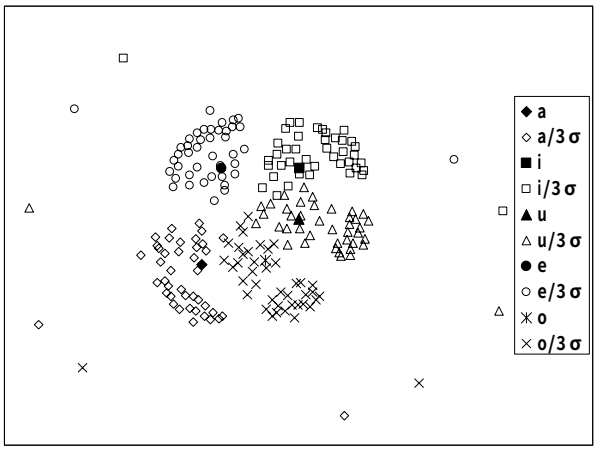

図１　日本語５母音の３ σ 空間

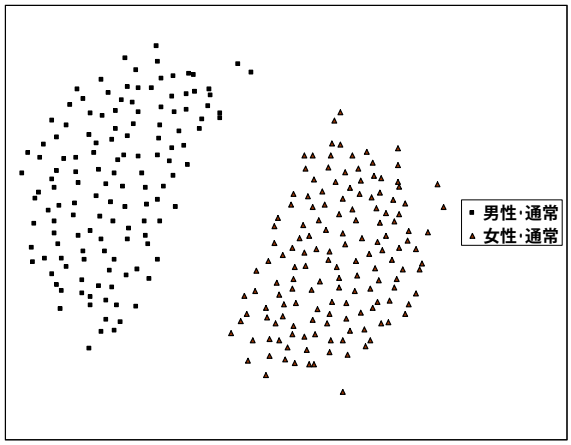

図２　性別 COSMOS

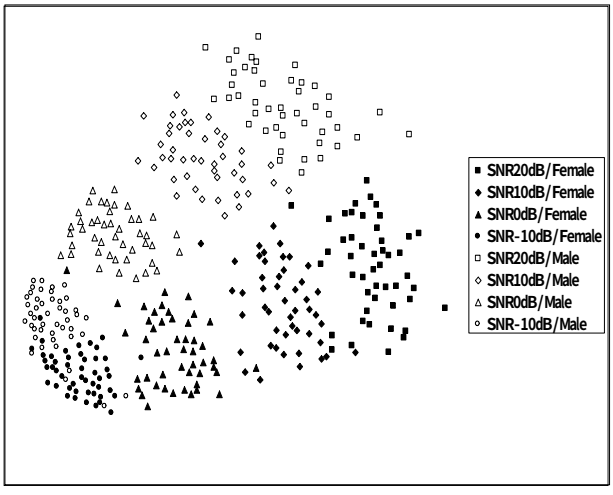

図３　信号雑音比 COSMOS

４．信号雑音比の分析

女性４８名及び男性４５名に日本語１２８単語を表１の通常の発話様式を指定して発声して貰った．この音声に展示会雑音を信号雑音比（SNR:Signal-to-Noise Ratio）

を変化させて重畳した．この雑音重畳音声から作成された特定話者音響モデルから作成された COSMOS を図３に示す．SNR が下がるにつれて，音響空間が縮小し，男性の音響空間と女声の音声空間が接近している．

５．タスクの分析

北米英語について，文章タスクの音声データ(ResourceManagement, TIMIT)，単語タスクの音声データ(single word)から作成した特定話者音響モデル及び３つの音声データを混合(Task Mixed)して作成した不特定話者音響モデルから作成した COSMOS を図４に示す．図４から，２つの文章タスクの音響空間は類似しており，文章タスクと単語タスクの音響空間は相違している様子が分かる．

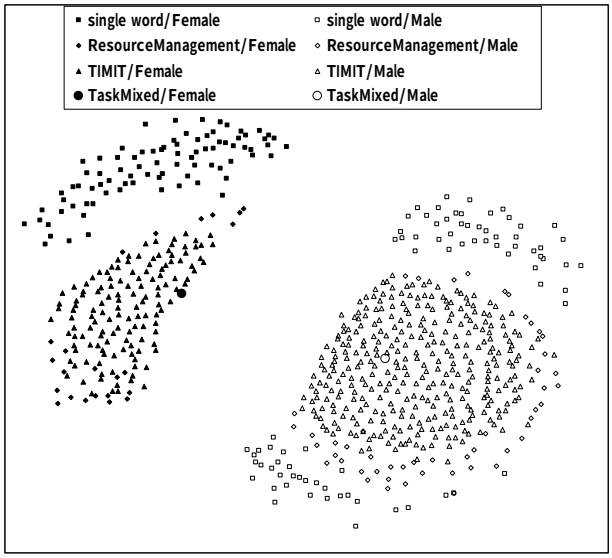

図４　タスク COSMOS

６．音素の分析

不特定話者日本語男性用音素モデルの CV および VC 接続の diphone から作成した COSMOS を図５に示す．diphone 全体は，ハート型形状に分布していることが分かる．ハート型形状の上半分には VC 接続が，下半分には，CV 接続が凝集している．ハート型形状の左部には，diphone が存在しない窪みがある．この窪みは，日本語では発音しない音素の領域に対応しているものと思われる．

７．発話様式の分析

男性１４５名に，表２に示す発話様式の中から複数の発話様式を指定して，ATR5240 単語の中の１７６単語から成る複数の単語リストを発声して貰った．まず，全てのデータを利用して，男性用不特定話者音響モデルを作

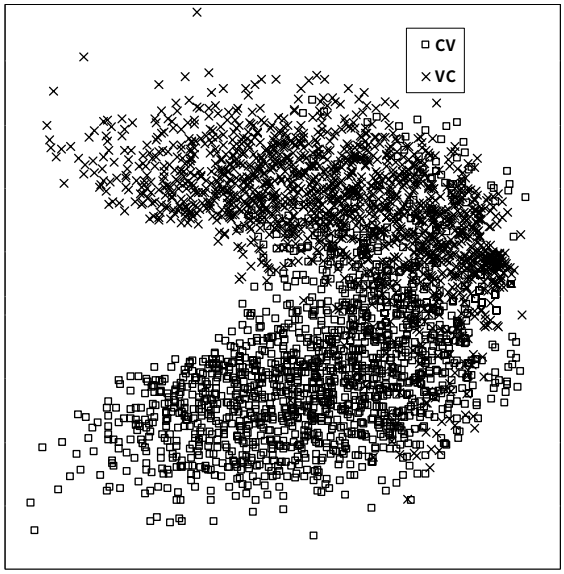

図５　Diphone COSMOS

成した．次に，この音響モデルを初期モデルとして，連結学習により，話者と収録時に指示された発話様式の組み合わせ毎に音響モデル（話者・発話様式音響モデルと呼ぶ）を作成した．こうして作成された話者・発話様式音響モデルの COSMOS を図６に示す．(1)式の $w(k)$ として，ATR5240 単語の音響単位の頻度を利用した．中心付近にプロットされている Star▲は，初期モデルとして使用した男性用不特定話者音響モデルを表す．

表１　発話様式

発話様式名	収録時の指示	Star 記号
通常	普段の速度で単語リストを読め.	・
早口	通常より早口で単語リストを読め.	○
高い声	通常より高い声で単語リストを読め.	●
小声	近くの人に聞こえないように単語リストを読め.	□
大声	離れた人にも聞こえるように大きい声で単語リストを読め.	□
ロンバード	自動車雑音を聞きながら単語リストを読め.	■
モーラ強調	読み仮名それぞれを強調するように単語リストを読め.	×

図６から，以下のことが観察された。

1）同一の発話様式を指定したとしても，実際の発話様式は話者によって異なることが分かる．予め付与された発話様式ラベルを鵜呑みにして，発話様式毎の不特定話者音響モデルを作成する場合のデメリットを示唆している。

2）発声速度および発声音量が対極の発話様式（例えば，モーラ強調 vs 早口，ささやき声 vs 大声／ロンバード)が，COSMOS 上でも原点対称の場所に位置している．

3）図６の発話様式ＣＯＳMＯＳを同心円上のゾーンに分割し，ゾーン毎に音響モデルを再作成すると，全ゾーンのデータから作成した不特定話者音響モデルと比較して，周辺部に位置する話者に関し，顕著な性能改善が得られることが分かった[23].

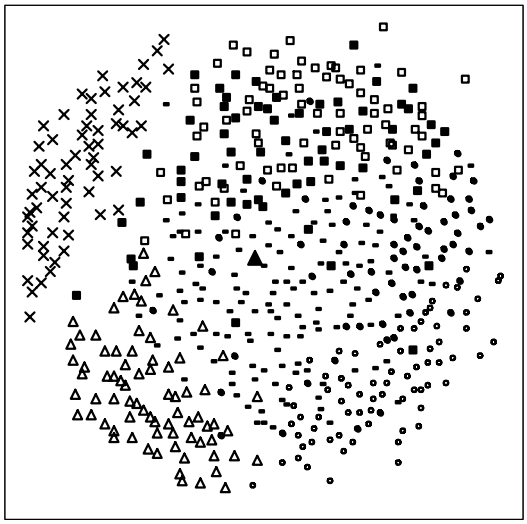

図６　発話様式 COSMOS

８．サウンドの分析

実環境には，自然発生的に様々な雑音が存在する．実環境において，社会生活に役立つ音声認識技術を実現するためには，雑音の存在を受け入れて，雑音に対処する技術を確立しなければならない．雑音には，比較的ゆっくりと変化する定常性雑音と急激に変化する非定常性雑音に区分できる．自動車の走行雑音のような定常性雑音については，先述の技術などにより，克服されつつある．今後の課題は，先述の技術では除去しきれない，音源未知の非定常な加法性雑音への対処である．そのためには，実環境での音声・雑音の総体であるサウンド全般の音響モデルを用意し，音声と雑音を区別することなく，認識することが重要である．音声の音素の男性・女性不特定話者音響モデル(phoneme.male, phoneme.female))，野鳥の鳴き声の音響モデル(bird)および表２に示す住宅内での生活雑音の音響モデル(house noise)とから作成した音響モデルの COSMOS を図７に示す．図７から，表２に示す生活雑音の音響空間と人間音声の音響空間は明確に分離しており，生活雑音と人間音声の識別は，容易であることが期待できる．また，野鳥の鳴き声は，人間音声と生活雑音の中間に位置しており，興味深い．

表２　住宅内の生活雑音

スリッパ音	カーテン開閉音	雨戸開閉音
ボールペン落下音	冷蔵庫ドア開閉音	カップ接触音

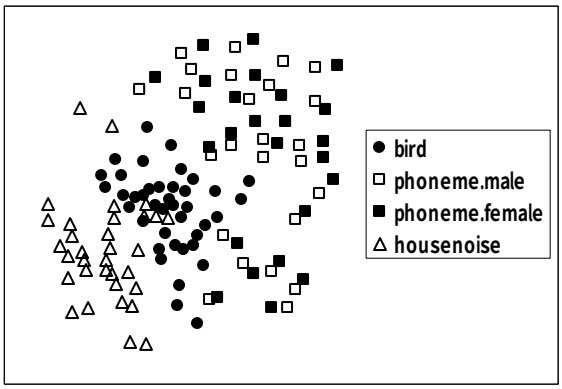

図７　サウンド COSMOS

９．おわりに

音響モデルの分布を可視化する COSMOS 法を提案し，音響空間の分析手法としての有効性を明らかにした．現在，COSMOS 法を利用した音響モデルライブラリの研究を行っている．その基本的な考え方は，以下の通りである．

1）先験的知識により音響的に類似していると思われる特定条件のクラスラベルが付与された音声データを集約し，HMMを学習する．

2）学習した複数個のHMMから COSMOS を作成する．

3）COSMOS の表示を参考にして，大きなクラスタに手動で分割する．

4）分割したクラスタ毎に COSMOS を再作成する．

5）再作成した COSMOS を同心円状の複数のゾーンに手動で分割する．

6）分割したゾーン毎に音声データを再集約する．（クラスラベルの再付与に相当する．）

7）分割したゾーン毎にHMMを再学習する．

8）再学習したHMMの中から，最も適合した音響モデルを選択し，認識に使用する．少ない発声で適合した音響モデルを選択する方法については，現在研究中である．

COSMOS 法は，声質及び発話様式の変化による音響空間の拡がりばかりでなく，雑音の種類・大きさ，雑音除去・歪み補正処理の種類，語彙の種類など，音響モデルに影響を与える様々な要因の影響を把握するのに便利なツ

ールである． COSMOS 法による音響モデル開発ツールを整備することにより，HMM やクラスタリング手法に関する高度な専門知識がなくとも音響モデルの開発に従事できる人口を増やすことができ，音響モデルの需給バランスの適正化に役立つと期待できる．

文献

(1) J. L. Gauvain et al.,“Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291-298, 1994.

(2) C. J. Leggretter et al.,“Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models,” Computer Speech and Language, vol.9, pp.171-185, 1995.

(3) T. Kosaka et al.,“Tree-structured speaker clustering for speaker-independent continuous speech recognition,” ICSLP-94, pp.1375-1378,1994.

(4) M. Gales,“Cluster adaptive training for speech recognition,” ICSLP-98, pp.1783-1786,1998.

(5) R. Kuhn et al.,“Rapid speaker adaptation in eigenvoice space,” IEEE Trans. Speech Audio Process., vol.8, no.6, pp.695-707, 2000.

(6) S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. ASSP, Vol.ASSP-27, No.2, pp.113-120, 1979.

(7) J. A. Nolazco Flores et al., ”Continuous speech recognition in noise using spectral subtraction and HMM adaptation,” Proc. ICASSP, pp.I-409-412, Adelaide, Australia, 1994.

(8) M. Shozakai et al., “A speech enhancement approach E-CMN/CSS for speech recognition in car environments,” Proc. IEEE ASRU97 Workshop, Santa Barbara, USA, pp.450-457, 1997.

(9) Y. Ephraim et al., ”Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” IEEE Trans. Vol.ASSP-32, No.6, pp.1109-1121, 1984.

(10) Y. Ephraim et al., ”Speech enhancement using a minimum mean square error log-spectral amplitude resonator,” IEEE Trans. Vol.ASSP-33, No.2, pp.443-445, 1985.

(11) A. K. Jain et al., "Statistical pattern recognition: a review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp.4-37, 2000.

(12) I. Borg et al., Modern multidimensional scaling, Berlin, Springer-Verlag, 1997.

(13) R. A. Fisher, "The use of multiple measurements in taxonomic problems," Ann. Eugenics, vol.7, no.part II, pp.179-188, 1936.

(14) M.Aladjem,“Multiclass discriminant mappings,” Signal Process.,vol.35, pp.1-18, 1994.

(15) J.Mao et al.,“Artificial neural networks for feature extraction and multivariate data projection," IEEE Trans. Neural Networks, vol.6, no.2, pp.296-317, 1995.

(16) Y. Mori et al., “Comparison of low-dimensional mapping techniques based on discriminatory information," Proc.2nd International ICSC Symposium on Advances in Intelligent Data Analysis(AIDA'2001), CD-ROM Paper-no.1724-166, Bangor, United Kingdom, 2001.

(17) J. H. Freidman et al., “A projection pursuit algorithm for exploratory data analysis," IEEE Trans. Comput., vol.C-18, no.5, pp.401-409, 1969.

(18) J. W. Sammon, “A nonlinear mapping for data structure analysis,” IEEE Trans. Computers, vol.C-18, no.5, pp.401-409, May 1969.

(19)T. Kohonen, "Self-Organizing Maps," Springer Series in Information Sciences, vol. 30, Berlin, 1995.

(20) A. Nagorski et al., “Optimal selection of speech data for automatic speech recognition system,” Proc. ICSLP, vol.4, pp.2473-2476, 2002.

(21) http://www.vorero.com

(22)岡登他, “クラスタリングによるHMM間の距離尺度の検討,” 信学技法, SP94-16, pp.15-20, 1997.

(23)庄境他,“２次元視覚化手法を利用した音響モデルの高精度化,” 音講論集, 3-Q21, pp.185-186, March, 2003.

付録　多次元空間上の情報を２次元平面へ非線形写像する場合のSammon法の定式化

L 次元空間上の N 個の情報を $P(i)(i=1,\cdots,N)$ と表す． $P(i)$ に１対１で対応する，２次元平面上の N 個のベクトル

（点）を $Z(i)=\begin{bmatrix} x(i) \\ y(i) \end{bmatrix}(i=1,\cdots,N)$ と表し，$x(i)$，$y(i)$ の初期値を乱数で与える．$P(i)$，$P(j)$ の L 次元空間上での相互距離を $D(i,j)$ と表す．（$P(i)$ が HMM による音響モデルの場合の $D(i,j)$ の定義を(1)式，(2)式に示す．）L 次元空間から 2 次元平面への非線形写像の m 番目の繰り返し時の N 個のベクトル（点）の座標を $Z_m(i)=\begin{bmatrix} x_m(i) \\ y_m(i) \end{bmatrix}(i=1,\cdots,N)$ と表す．L 次元空間から 2 次元平面への非線形写像の m 番目の繰り返し時の $Z_m(i)$，$Z_m(j)$ のユークリッド距離 $\tilde{D}_m(i,j)$ を次式で定義する．

$$\tilde{D}_m(i,j)=\{x_m(i)-x_m(j)\}^2+\{y_m(i)-y_m(j)\}^2 \tag{A-1}$$

m 番目の繰り返し時の L 次元空間から 2 次元平面への写像誤差 E_m を次式で定義する．

$$E_m \equiv \frac{1}{c}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left[\{D(i,j)-\tilde{D}_m(i,j)\}^2 \Big/ D(i,j)\right] \tag{A-2}$$

$$c=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}D(i,j) \tag{A-3}$$

この時，$m+1$ 番目の繰り返し時の $Z_{m+1}(i)$，$Z_{m+1}(j)$ の 2 次元座標を次式で計算する．

$$x_{m+1}(i)=x_m(i)-\alpha\cdot\Delta x_m(i),\quad y_{m+1}(i)=y_m(i)-\alpha\cdot\Delta y_m(i) \tag{A-4}$$

$$\Delta x_m(i)=\frac{\partial E_m}{\partial x_m(i)}\Bigg/\left|\frac{\partial^2 E_m}{\partial x_m(i)^2}\right|,\quad \Delta y_m(i)=\frac{\partial E_m}{\partial y_m(i)}\Bigg/\left|\frac{\partial^2 E_m}{\partial y_m(i)^2}\right| \tag{A-5}$$

$$\frac{\partial E_m}{\partial x_m(i)}=\frac{-2}{c}\sum_{\substack{l=1\\ l\neq i}}^{N}\left\{\frac{D(i,l)-\tilde{D}_m(i,l)}{D(i,l)\cdot\tilde{D}_m(i,l)}\right\}\{x_m(i)-x_m(l)\} \tag{A-6}$$

$$\frac{\partial E_m}{\partial y_m(i)}=\frac{-2}{c}\sum_{\substack{l=1\\ l\neq i}}^{N}\left\{\frac{D(i,l)-\tilde{D}_m(i,l)}{D(i,l)\cdot\tilde{D}_m(i,l)}\right\}\{y_m(i)-y_m(l)\} \tag{A-7}$$

$$\frac{\partial^2 E_m}{\partial x_m(i)^2}=\frac{-2}{c}\sum_{\substack{l=1\\ l\neq i}}^{N}\frac{1}{D(i,l)\cdot\tilde{D}_m(i,l)}\left[\{\tilde{D}_m(i,l)-D(i,l)\}-\frac{\{x_m(i)-x_m(l)\}^2}{D(i,l)}\left\{1+\frac{\tilde{D}_m(i,l)-D(i,l)}{D(i,l)}\right\}\right] \tag{A-8}$$

$$\frac{\partial^2 E_m}{\partial y_m(i)^2}=\frac{-2}{c}\sum_{\substack{l=1\\ l\neq i}}^{N}\frac{1}{D(i,l)\cdot\tilde{D}_m(i,l)}\left[\{\tilde{D}_m(i,l)-D(i,l)\}-\frac{\{y_m(i)-y_m(l)\}^2}{D(i,l)}\left\{1+\frac{\tilde{D}_m(i,l)-D(i,l)}{D(i,l)}\right\}\right] \tag{A-9}$$

ここで，α は，収束速度を制御するパラメータであり，文献[18]では 0.3-0.4 を推奨している．