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

2005－ＳＬＰ－59（23）
2005／12／22

[招待講演] 複数音声コーパスの俯瞰的分析

庄境 誠†

†旭化成株式会社情報技術研究所 〒243-0021 神奈川県厚木市岡田 3050 厚木ｱｸｽﾄﾒｲﾝﾀﾜｰ 22F
E-mail: †shozakai.mb@om.asahi-kasei.co.jp

あらまし　市場からの要請を受けて，認識性能保証の方法論を確立することは，音声認識ベンダーとして重要な技術課題の１つである．認識性能保証ができることは，音声認識アプリケーションの仕様に対して，認識性能分布を予め予測できるベンチマーク技術を有することを意味する．そのためには，実環境における様々な日本語音声の変動要因を把握することが出発点である．筆者らは，音声コーパスから学習された統計モデル間の距離尺度と多次元尺度法を組み合わせて，コーパス空間の俯瞰的分析手法の研究を進めている．既に，公開されている日本語の複数音声コーパスを本手法により分析した予備的検討結果を紹介し，ベンチマーク方法論確立に向けた今後の技術課題について述べる．

キーワード　音声，コーパス，隠れマルコフモデル，多次元尺度法

[Invited Talk] Comprehensive Analysis of Multiple Speech Corpora

Makoto SHOZAKAI†

† Information Technology Laboratory, Asahi Kasei Corporation Atsugi AXT Main Tower 22F, Okada 3050, Atsugi, Kanagawa, 243-0021 Japan
E-mail: †shozakai.mb@om.asahi-kasei.co.jp

Abstract We are now carrying researches of comprehensive analysis method of corpus space by combining a distance measure among statistical models trained from speech corpora and a multidimensional scaling technique. Some preliminary results obtained from the analysis of open Japanese speech corpora by the method are shown. Technical issues to be solved in near future to establish the benchmarking methodology are discussed.

Keyword Speech, Corpus, HMM, Multidimensional Scaling

1. はじめに

音声認識技術の実用化が停滞している根本原因の問題の１つは，市場が期待する音声認識性能と音声認識ベンダーが提供する音声認識ソフトウェア製品の性能との埋めがたい乖離にある．一般に，前者の方が後者の方より大幅に高い．場合によっては，前者は体感１００％であることを求められる．それに対して，実環境における後者の大幅な性能劣化が市場を落胆させることがしばしば起こっている．どうして，このような乖離が起こるのであろうか？

音声認識性能に関する評価（以下，ベンチマークと呼ぶ）フローの一般的な手法は，以下の通りである．まず，指定された認識タスクに固有の語彙に対する記述（以下，認識タスク記述と呼ぶ）に基づいて作成された語彙リストを，ある金銭的かつ時間的コストの範囲で，限定的な被験者に発声して貰い，音声サンプル（以下，認識タスク依存評価音声サンプルと呼ぶ）を収集する．次に，認識タスク依存評価音声サンプルを用いて，所定の音声認識システムの認識率を評価して，平均値，最高値，最低値などを算出する．

従来のこの手法における問題点として，以下のことが挙げられる．（１）認識タスクが変わる度に，限定的な被験者に依頼して，認識タスク依存評価音声サンプルを収集するために，金銭的及び時間的コストが嵩む．（２）被験者の網羅的な選択に関して確立された方法論に乏しく，限定的な被験者の選択の正当性の根拠が希薄である．（３）被験者の不可避な交代が起こった場合，評価基準が経時的に揺らぐ．（４）音声認識ベンダー毎に独自に認識タスク依存評価音声サンプルを用意するため，業界標準が不在である．

市場は，認識タスク記述を与えただけで，実環境での認識性能分布を見積もり，速やかに提供するよう，音声認識ベンダーに求めている．アプリケーションにおける認識性能分布を評価させられることに対し，市場は苛立ちを禁じ得ない．現状を放置していては，音声認識技術自体が，市場から省みられなく危機すら感じる．

世界のいずれの音声認識ベンダーも，自動車内，屋内，屋外などの多様な実環境における認識性能分布を客観的に評価する技術を確立できておらず，客観的評価手法に基づく認識性能分布情報を市場に提供できて

いない．ベンチマークに余りに多くの時間とお金をかけ過ぎると，音声認識製品の市場投入が遅れ，採算性も危うくなる．今こそ，効率的なベンチマーク方法論の確立が必要な所以である．

認識タスク記述を与えただけで，認識性能分布を客観的に評価できる技術を確立し，その結果を提供することにより，音声認識ベンダーと市場は，性能評価に関する判断基準を共有することが出来る．これにより，上記の乖離の問題は次第に解消され，音声認識市場の健全なる成長を促すと期待される．

認識性能保証ができることは，音声認識アプリケーションの認識タスク記述に対して，認識性能分布が予め予測できるベンチマーク技術を有することを意味する．そのためには，実環境における様々な日本語音声のバリエーションを把握することが出発点となる．

日本語音声のバリエーションを把握するためには，そのバリエーションを表現する音声コーパスの獲得が必須である．上述のように，音声認識ベンダーにとって，音声コーパスの収集コストは無視できない要素であり，経営判断の中で大きなウエイトを占める．音声コーパスの収集コストが経営を圧迫する音声認識ベンダーにとって，網羅的な音声コーパスの迅速な確保は最重要課題といえる．実環境フィールドに設置した音声認識システムを介して，自然発生的な発声を待ち受けて収集するアプローチも有効であるが，それでは間に合わないほど事態は切迫している．

日本では，音声コーパスの規模が認識性能に与えるインパクトの大きさが以前から深く認識され，既に非常に大量で多種の日本語音声コーパスが収集され，公開されている．しかしながら，これらの音声コーパス間の横断的な分析がされたことはなかった．これらの豊富な資源を俯瞰的に分析し，それぞれの日本語音声コーパスの利用価値を判断することができれば，音声認識ベンダーにとっての経費の節約につながる．

筆者らは，統計モデル間の距離尺度と多次元尺度法を組み合わせたコーパス空間の俯瞰的分析手法(以下，COSMOS 法と呼ぶ) [2]を提案し，その有効な活用方法の研究を進めている．その研究の一環として，このCOSMOS 法を用いて，既存の日本語音声コーパスを1つのコーパス空間地図(「以下，日本語音声コーパス空間地図と呼ぶ」)の上に写像するプロジェクトを進めている．

以下では，まず，COSMOS 法の概要を紹介する．次いで，日本語音声空間地図の予備的作成例を紹介し，分析結果を述べる．また，ベンチマーク方法論確立に向けて，今後取り組まなければならない技術課題について述べる．

2. 音響空間俯瞰技術

最初に，ベクトル情報を入力とする多次元尺度法の古典的手法として提案されたSammon法[1]を紹介する．そして，筆者らが，Sammon 法の入力信号をベクトルから統計モデルに拡張した，コーパス空間俯瞰技術COSMOS 法[2]に関して概説する．

2.1. Sammon 法

L 次元空間内の N 個のベクトルデータを $P(i)(i=1,\cdots,N)$ で表す．$P(i)$ に一対一で対応する2次元空間の N 個のベクトルを $Z(i)=[x_1(i)\quad x_2(i)]^T\ (i=1,\cdots,N)$ と表す．$x_q(i)\ (q=1,2)$ の初期値は乱数を与える．L 次元空間内での $P(i)$ と $P(j)$ の相互距離を $D(i,j)$ で表す．L 次元空間から2次元空間への最急降下法による非線形写像の m 回目の繰り返しにおける座標を $Z_m(i)=[x_{m,1}(i)\quad x_{m,2}(i)]^T\ (i=1,\cdots,N)$ で表す．また，$Z_m(i)$ と $Z_m(j)$ のユークリッド距離を $\tilde{D}_m(i,j)$ で表す．

非線形写像の m 回目の繰り返しにおけるベクトルデータ $P(i)$ の局所的写像誤差 $E_m(i)$ を相互距離 $D(i,j)$ とユークリッド距離 $\tilde{D}_m(i,j)$ の正規化ユークリッド距離の総和で表すとすると，$E_m(i)$ は(1)式と(2)式で表される．

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

$$c\equiv\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}D(i,j) \quad (2)$$

非線形写像により最小化される大局的写像誤差値 E_m を局所的写像誤差値 $E_m(i)$ の総和で表すとすると，E_m は(3)式で与えられる．

$$E_m\equiv\sum_{i=1}^{N}E_m(i) \quad (3)$$

非線形写像の $m+1$ 番目の繰り返しにおける座標 $Z_{m+1}(i)$ と $Z_{m+1}(j)$ は，最急降下法により，(4)式と(5)式で算出される．

$$x_{m+1,q}(i)=x_{m,q}(i)-\alpha\cdot\Delta x_{m,q}(i) \quad (4)$$

$$\Delta x_{m,q}(i)=\frac{\partial E_m}{\partial x_{m,q}(i)}\Big/\left|\frac{\partial^2 E_m}{\partial x_{m,q}(i)^2}\right| \quad (5)$$

これにより，元の L 次元空間内で近くに（遠くに）位置するベクトルは，２次元空間内でも相対的に近くに（遠くに）写像される．最小の E_m を与える繰り返しの N 個の全ての座標 $Z_m(i)$ を記録し，この座標を用いて地図を作成する．

2.2. COSMOS 法

一般に，音声認識システムの音響モデルセットは，音素(monophone)，二つ組音素(biphone)，三つ組音素(triphone)に対応する音声単位モデル毎の HMM の集合体である．従って，音響モデルセット i と音響モデルセット j の相互距離 $D(i,j)$ は，(6)式で定義される．

$$D(i,j)=\sum_{k=1}^{K} d(i,j,k)\cdot w(k) \Big/ \sum_{k=1}^{K} w(k) \qquad (6)$$

ここで，$d(i,j,k)$ は，音響モデルセット i の中の音声単位モデル k の HMM と音響モデルセット j の中の音声単位モデル k の HMM の相互距離を表す．$w(k)$ は，音声単位 k の HMM の重み（例えば，出現頻度）を表す．K は，音響モデルセットにおける音声単位モデルの総数を表す．

音響モデルセット i と音響モデルセット j の全ての音声単位モデルが同じトポロジーを有する HMM でモデル化され，状態対応が一対一であると仮定すると，$d(i,j,k)$ は，(7)-(9)式で表される．

$$d(i,j,k)\equiv \frac{1}{S(k)}\sum_{s=0}^{S(k)-1}\frac{1}{L}\sum_{l=0}^{L-1}\frac{dd(i,j,k,s,l)}{pp(i,j,k,s,l)} \qquad (7)$$

$$dd(i,j,k,s,l)\equiv \sum_{m_i=0}^{M_i}\sum_{m_j=0}^{M_j} p(i,k,s,l,m_i)\cdot p(j,k,s,l,m_j)\cdot c(i,j,k,s,l,m_i,m_j) \qquad (8)$$

$$pp(i,j,k,s,l)\equiv \sum_{m_i=0}^{M_i}\sum_{m_j=0}^{M_j} p(i,k,s,l,m_i)\cdot p(j,k,s,l,m_j) \qquad (9)$$

ここで，$p(i,k,s,l,m)$ は，音響モデルセット i の音声単位モデル k の状態 s の次元 l の m 番目の正規分布の重みを表す．$S(k)$ は音声単位モデル k の状態数を表す．M_i は，音声単位モデル i の正規分布の混合数を表す．また，$c(i,j,k,s,l,m_i,m_j)$ は，音響モデルセット i の音声単位モデル k の状態 s の次元 l の m_i 番目の正規分布と音響モデルセット j の音声単位モデル k の状態 s の次元 l の m_j 番目の正規分布間のバタチャリア距離を表す．

(4)式の値 α は，非線形写像の収束速度を制御するパラメータであり，推奨値として 0.3-0.4 が示唆されている[5]が，COSMOS 法においては，0.1 以下の値が適当であることが分かった．COSMOS 法によって得られる地図を COSMOS 地図と呼ぶ．

3. 日本語音声コーパス空間地図

ここでは，COSMOS 法を用いて，旭化成収録の話者・発話様式コーパス[2]，日本音響学会収録の JNAS（新聞記事読み上げ音声コーパス）[3]，奈良先端大収録の S-JNAS（高齢者の音声認識用大規模データベース）[4]，ATR 収録の APP-BLA（多数話者音声データベース音素バランス文）[5]，名古屋大学収録の CIAIR-HCC（車内対話音声データベース）[6]の各音声コーパスから作成した空間地図の例を示す．全て，接話マイクで収録されたコーパスを用いた．

3.1. 話者・発話様式

ATR5240 の音素バランス単語セットを分割して作成された 30 種類の 175 単語リストを，126 人の日本人女性が，表 1 に示されたいくつかの発話様式で発声した．

表 1　発話様式

	被験者に対する収録時の指示	記号
normal	発声リストを通常の速度で読みなさい	□
fast	発声リストを通常の速度よりも早口で読みなさい	◆
high	発声リストを通常の速度よりも高い声で読みなさい	▲
whisper	発声リストを近くの人に聞かれないように読みなさい	○
loud	発声リストを少し離れた人に聞こえるように読みなさい	●
Lombard	発声リストを自動車雑音をヘッドフォンで聴きながら読みなさい	■
syllable-enhanced	発声リストを仮名を強調して読みなさい	◇

457 個の話者・発話様式の組み合わせの音声コーパスを収録した．音声データには，SNR20dB で展示会雑音を重畳した．サンプリング周波数は 11.025kHz，音響パラメータは，10 次の MFCC，10 次のデルタ MFCC，デルタ対数パワーである．音声単位モデルは，３状態の biphone レベルの HMM（１混合）である．ノイズキャンセル処理とイコライザ処理を施した．

図１に，この音声コーパスの話者・発話様式の組み合わせ毎の HMM から作成した COSMOS 地図を示す．発話様式として対極にある syllable-enhanced と fast は，原点を挟んで反対側に位置することが見て取れる．whisper と loud の関係も同様である．

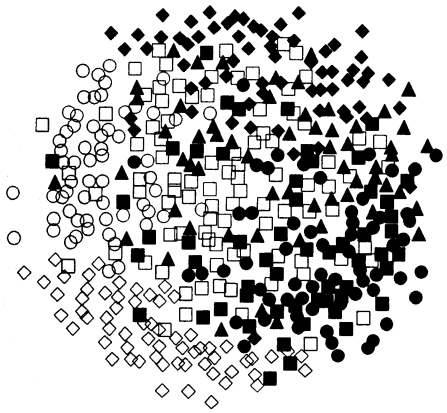


図1　話者・発話様式 COSMOS 地図

3.2. JNAS

学習用セットの男性話者 120 人について，2つのタスク（新聞記事(newspaper)，バランス文(balanced)）毎に作成した特定話者 HMM（IPA,monophone, 4 混合）から作成した COSMOS 地図を図2に示す．この地図からは，JNAS コーパスのタスク依存性はほとんどないように見える．

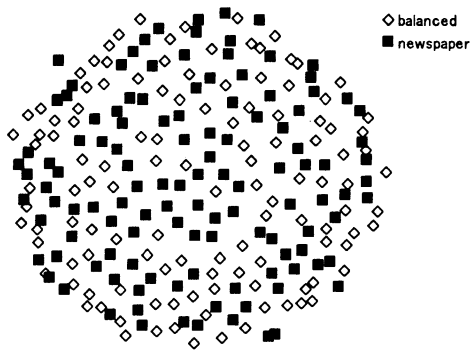


図2　JNAS COSMOS 地図

3.3. S-JNAS

音韻モデル用男性話者 151 人の2つのタスク（音素バランス文(balanced)，新聞記事文(newspaper)）と評価実験用男性話者 51 人の2つのタスク（情報検索タスク文(inquiry)，新聞記事文(newspaper)）から，404 個の HMM（IPA,monophone, 4 混合）を作成し，音韻モデル用話者(training)と評価実験用話者(evaluation)から作成した COSMOS 地図を図3に示す．この地図から，評価用話者の網羅性に偏りがあるように見える．特に，情報検索タスク文の HMM は外縁に散逸しており，音素バランス文，新聞記事文から学習した音響モデルでは高い性能が得られない可能性を示唆している．

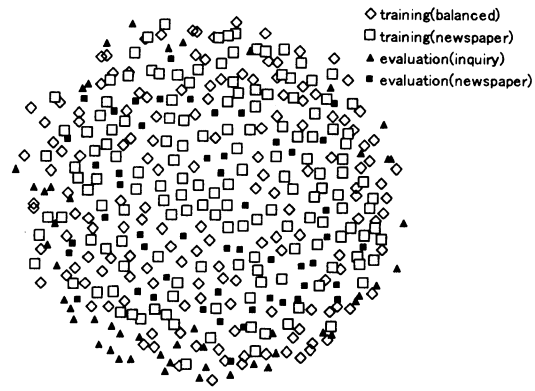
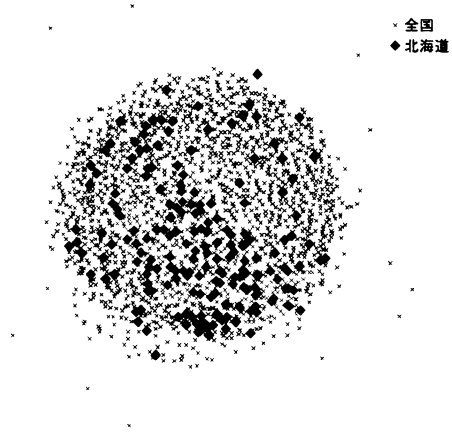


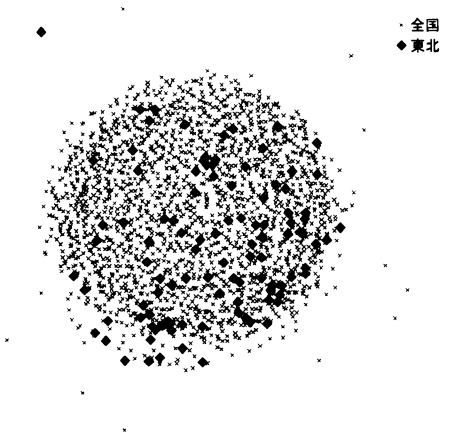
図3　S-JNAS COSMOS 地図

3.4. APP-BLA

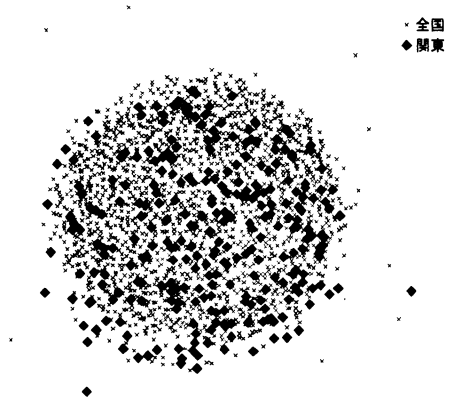
男声話者 1379 人が音素バランス(ATR503)文の 50 文のセットを読み上げた音声から作成した特定話者 HMM（IPA,monophone,1 混合）1826 個（複数セットを読み上げている話者あり）から作成した COSMOS 地図を図4に示す．この地図から，北海道，東北，九州は明らかな地域性が認められるが，その他については地域性が観察されなかった．地域別の音響モデルの構築[7]に加えて，COSMOS 地図上での位置を考慮したコーパス空間の分割による音響モデルライブラリの構築も有効であろう．COSMOS 地図の粗な領域（外縁等）に位置する日本語音声の特徴を同定し，その音声を効率的に収集する方法の研究も重要になろう．



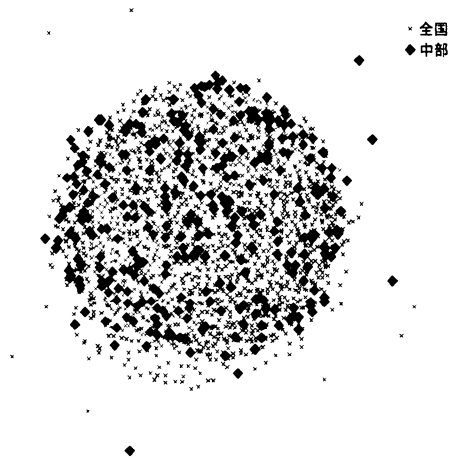
(a)北海道



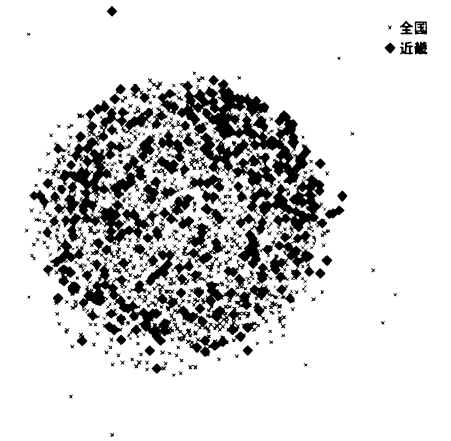
(b)東北



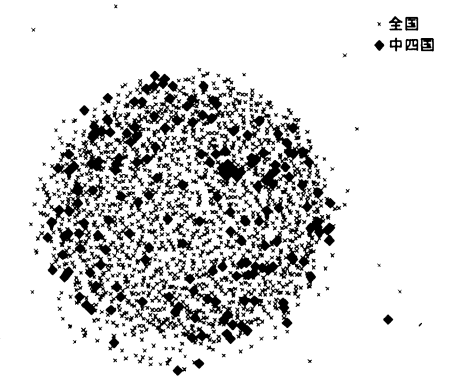
(c)関東



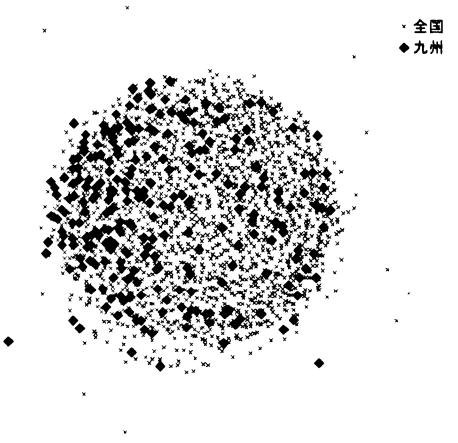
(d)中部



(e)近畿



(f)中四国



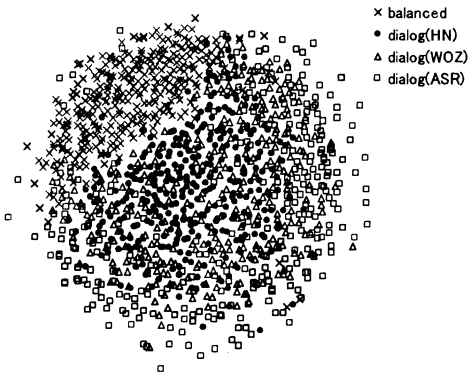
(g)九州

図 4　ATR APP-BLA COSMOS 地図

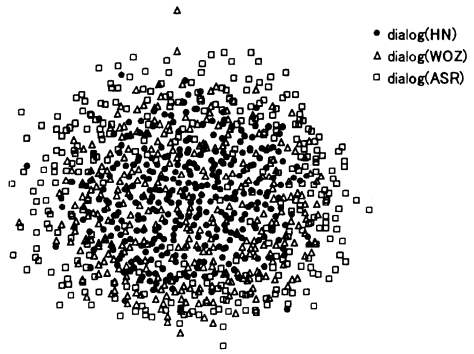
3.5. CIAIR-HCC

男性 320 名が発声した２つのタスク（バランス文 balanced)，運転中対話セッション(dialog)）の音声データを接話マイクで収録したコーパスから作成した特定話者 HMM（IPA,monophone,1 混合）から得た COSMOS 地図を図５(a)に示す．balanced と dialog でタスク間の音響的な異なりが明確であることが分かる．

同じ男性 320 名が異なる対話モード（運転中対話（人）セッション(HN)，運転中対話（WOZ システム）セッション(WOZ)，運転中対話（音声対話システム）セッション(ASR)）で発声した音声を接話マイクで収録したコーパスから作成した特定話者 HMM（IPA,monophone,1 混合）のみから得た COSMOS 地図を図５(b)に示す．対話モードと発話様式の間に強い相関性があることが分かる．WOZ の発話様式は，HN 及び ASR に類似する二面性がある様子が見て取れる．



(a)バランス文＋運転中対話セッション



(b)運転中対話セッションのみ

図５　CIAIR-HCC COSMOS 地図

3.6. 今後の予定

今後は，これらの日本語音声コーパスに AURORA-2J や CENSREC-3 などを加えて，同様の俯瞰的分析を進め，多数の日本語音声コーパスを網羅する１枚の日本語音声コーパス空間地図に仕上げる予定である．その観察結果から，それぞれの日本語音声コーパスの利用価値や，効果的な利用方法を判断することが出来ると思われる．また，従来，謎解きが十分に進んでいない音響モデルのタスク依存性の正体も解き明かされていくものと期待している．

4. おわりに

音声認識ベンダーは，ベンチマーク方法論の確立に一体どのくらい時間と費用をかけてきたのだろうか？市場に対し，信頼される音声認識製品を提供する責務を負う音声認識ベンダーは，今まで以上にベンチマークを優先することが求められる．それが，音声認識市場の拡大のために何より必要ではないだろうか？一刻も早いベンチマーク方法論の確立が求められているが，その基本にあるのは，日本語音声の音響空間の把握である．そのためには，既に存在する日本語音声コーパスの分析が最も効果的である．その結果，不足する日本語音声コーパスの仕様の同定が可能となる．不足する日本語音声コーパスの収集を如何に効率的に行うかの方法論の研究がますます重要になるであろう．一刻も早く，方言，音響環境，行動・動作が話し言葉の発話様式に与える影響の解明などの実用上解決すべき課題の真相に踏み込んで行かなければならない．

5. 謝辞

日頃から，共に COSMOS 法に関し研究，議論を重ね，かつ，本講演のために COSMOS 地図のデータを提供してくれた弊社奈木野豪秀氏に感謝する．

文　献

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

[2] M. Shozakai and G. Nagino, "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," Proc. ICSLP-04, vol.I, pp.717-720, Jeju, Korea, 2004.

[3] http://www.milab.is.tsukuba.ac.jp/jnas/

[4] http://db.ciair.coe.nagoya-u.ac.jp/dbciair/koureisha_files/index.htm

[5] http://www.red.atr.jp/product/02/pro_02_04.html

[6] http://db.ciair.coe/nagoya-u.ac.jp/dbciair/dbciair2/shanai_taiwa.htm

[7] 奥田浩三，松井知子，内藤正樹，匂坂芳典，中村哲，“大規模日本語音声データベースの構築と評価，” 音学誌，vol.58, no.9, pp.569-578, 2002.