Vol.2012-SLP-93 No.3
2012/10/26

音声データの隠れ属性を利用した異種音響モデル群の構築

福田 隆[1,a)]　立花 隆輝[1]　西村 雅史[1]　Upendra Chaudhari[2]　Bhuvana Ramabhadran[2]
Puming Zhan[3]

概要：音声の多様な変化を高精度にモデル化する方法は，音声認識の分野で長らく重要課題の一つに位置づけられてきた．近年では，大規模コーパスの整備に伴い，音響的に類似したサブセットを用いて個々にユニークな特性を持つ音響モデル集合を作成し，システム統合を介してさならる高精度化を図る手法が増えている．本報告では，SNR や話速といった音声に内在する隠れ属性を利用して学習データを分割し，システム統合法のための効果的な音響モデル集合を構築する方法を提案する．提案法では，各発話を事後確率に基づく単一ベクトルで表現した後，コサイン類似度に由来する目的関数を用いて音声データクラスタの独立性を評価する．その後，生成されたデータクラスタ毎に音響モデルを構築し，n-best ROVER によるシステム統合を行う．提案手法は音声検索タスクに特化した大語彙連続音声認識で，単一モデルの音声認識システムと比較して相対的に 4%の性能改善を達成した．

キーワード：音声認識，音響モデル，システム統合，大規模コーパス

Constructing Ensembles of Dissimilar Acoustic Models Using Hidden Attributes of Large Speech Corpus

Takashi Fukuda[1,a)]　Ryuki Tachibana[1]　Masafumi Nishimura[1]　Upendra Chaudhari[2]
Bhuvana Ramabhadran[2]　Puming Zhan[3]

Abstract: One of the objectives in acoustic modeling is to realize robust statistical models against the wide variety of acoustic conditions that are present in real world environments. As large amounts of training data become available, modeling subsets of the data with similar acoustic qualities can be done accurately and multiple acoustic models are jointly used as a form of system combination or model selection. In this paper, we propose a method to partition the training data for constructing ensembles of acoustic models using meta-data attributes such as SNR, speaking rate, and duration via a binary tree. The metadata attribute used at each binary split in the decision tree is obtained using a metric proposed in this paper that is cosine-similarity based. The resulting multiple models are combined using voting techniques such as n-best ROVER. The proposed method improved the recognition accuracy by up to 4% relative over the state-of-the-art system on a large vocabulary continuous speech recognition voice search task.

Keywords: Automatic speech recognition, multiple acoustic modeling, system combination, large corpora

1. はじめに

多様に変化する音声時系列データのモデル化には長らく HMM（Hidden Markov Model）が用いられてきた．音声信号は雑音が少ない環境であってもスペクトル変動がとても大きく，雑音下で収集された音声のモデル化ともなると，多くの状態や出力確率分布を持つ複雑な構成が必要とさ

[1] 日本アイ・ビー・エム株式会社　東京基礎研究所
IBM Research – Tokyo, IBM Japan Ltd., Toyosu, Tokyo 135–8511, Japan
[2] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
[3] Nuance Communications Inc., Boston, MA, USA
[a)] fukuda1@jp.ibm.com

れてきた．近年では，多様なスペクトル変形に対応するため，単語間の依存関係を考慮したクロスワード音素コンテキストモデルが広く利用されており，また，様々な環境・雑音への適用も踏まえて，数十万以上のパラメータを内包する音響モデルが利用されることも少なくない．しかしながら，実環境においてはそれでもなお，単一の音響モデルでは対処できないスペクトル変動が数多く観測される．

音声変動の複雑さに対応するアプローチの一つとして，音声に含まれる隠れ属性（隠れ変数）を利用して音響環境を限定し，モデルの音響的許容範囲を細分化することによってシステム全体としての性能を高める方法が提案されている．ここで隠れ属性とは，性別，話者（もしくは話者グループ），話速，伝達特性，雑音，SNR など音声信号に含まれる二次的な性質を指す．これら隠れ属性を効果的に扱う先駆的手法として，音素決定木に属性情報を付与し，環境に応じた複数の状態を同一音素に割り当てる方法が提案されている [1]．また，音声データの隠れ属性を利用して学習データを分割し，音響モデルを環境ごとに独立に構築する方式の検討も盛んである [2], [3]．

一般に，複数の音響モデルを並列に用いる場合，デコード時にモデルを切り替えるモデル選択や，各音響モデル（認識システム）からの音声認識出力を結合するシステム統合法が利用される．システム統合法では，互いに性質の異なる音響モデル群を構築し，各音響モデルの独立性・専門性を高めることで，全体として単一モデルの音声認識システムよりも高い性能を得ている．そのため，システム統合に利用する音響モデル集合は互いに無相関（非類似）であることが望ましい*1．

本報告では，システム統合法の利用を前提として，その枠組みの中で最大限の効果が得られるような音響モデル集合の構築方法を提案する．提案手法は，各音響モデルの独立性を高めるための効果的な学習データ分割法に関するものである．評価では，5000 時間以上の検索語彙発話からなる音声データについて，隠れ変数を考慮した複数の音響モデルを構築し，そのモデル集合を利用したシステム統合が単一モデルの音声認識システムと比較して大きく性能を改善することを示す．また，デコード時に最適な音響モデルが選択できる理想的状況を想定した場合の実験結果についても紹介する．

本報告は次のように構成される．2 章では，大語彙連続音声認識（LVCSR：Large Vocaburary Continuous Speech Recognition）におけるアンサンブルモデル手法に関する従来手法についてまとめる．続いて，3 章で提案手法であるデータ分割手法について述べた後，4 章，5 章で実験結果と考察を示し，最後に 6 章で結論を述べる．

*1　モデル選択法でも同様のことが言える

2. 先行研究

特定の環境や性質に特化した音響モデルに焦点を当てた研究は数多く行われている．それらを大別すると，ユニバーサルモデル [4]，アンサンブルモデル（音響モデル集合）[5], [6], [7], [8]，性別・話者依存モデル，環境・話者適応などが代表的であり，研究の幅は広い．

大規模データによる LVCSR のための音響モデル集合構築法として，文献 [9] では音声データを教師なしクラスタリングにより階層的に分割し，各クラスタから個別に音響モデルを構築する方法を試みている．大規模データを何らかの音響的基準で分割し，各データクラスタの特性を反映させた音響モデルを利用する方法は，単独の LVCSR システムでは捉える事ができなかった言わば音響的にロングテールに位置する発話に対して，対象範囲を広げていることに相当する．複数のモデルを利用するシステム統合法において効果的な音響モデル集合を構築するためには，各データクラスタが互いに独立（最大限に非類似）な性質を示すことが望ましい．文献 [9] の方法では，各発話をスパースな事後確率ベクトルで表現し，その発話ベクトル群に対してカルバック・ライブラー距離によるデータ空間分割を行うことによって独立性を高めている．カルバック・ライブラー距離に基づくデータクラスターは，結果として，ラウドネス，SNR，性別，ピッチ，そしてパープレキシティーを反映したものになっていた．

Beaufays らの方法 [9] では，ルートノードに対応するモデル（細分化前の汎用的なモデル）の学習には人手による書き起こしデータのみを用い，非書き起こしデータの追加は汎用モデルの性能改善につながらなかったことを示した．一方で，大量に存在する非書き起こしデータは，独立性が必要な音響モデル群の構築に対してこそ大きな利用価値があることを論じている．文献では，システム全体の性能比較について具体的な数値が示されていないが，テスト発話に対して音響上最も近い性質を示すモデルが認識誤りの削減に大きな効果をもたらす可能性について言及した．本報告で提案する方法は，大規模コーパスを用いた LVCSR において，これまであまり利用されてこなかったデータの隠れ属性を基にしたクラスタリング法に関するものであり，単一モデルの LVCSR システムとの比較結果を認識誤り率の観点から明確に議論する．

3. 提案手法

提案法では，まず音声データに内在する隠れ属性を用いて音声コーパスを分割する．その後，各発話を事後確率によって単一ベクトルで表現し，分割対象の音声データセットについて，どの隠れ属性が最も独立性の高いユニークな性質を持つクラスタを生成し得るかを評価する．以下に各

処理の詳細を示す．

3.1 事後確率に基づく発話ベクトル

音声データセットの発話集合を $\chi = \{x_1, x_2, ..., x_n, ..., x_N\}$ とする．ここで x_n は発話を，N は総発話数を表す．時刻 t における発話 n の特徴ベクトルを $\boldsymbol{y}_{nt}$ とすると，発話の音響的性質を表す単一ベクトルは，フレーム毎に求まる事後確率を発話単位で平均することによって導出される．

$$p(g_i|\boldsymbol{y}_n) = \frac{1}{T_n}\sum_{t=1}^{T_n}\frac{p(\boldsymbol{y}_{nt}|g_i)}{\sum_{k\in\mathcal{G}}p(\boldsymbol{y}_{nt}|g_k)}, \quad (1)$$

ここで $p(\boldsymbol{y}_{nt}|g_i)$ はモデル $\mathcal{G}$ における i 番目のガウス分布による尤度，T_n は発話 n の総フレーム数であり，事後確率の導出には $p(g_i)$ が全て等しいという仮定を入れている．また，$\boldsymbol{y}_n = \{\boldsymbol{y}_{n1}, \boldsymbol{y}_{n2}, ..., \boldsymbol{y}_{nt}, ..., \boldsymbol{y}_{nT_n}\}$ である．D を音響モデル $\mathcal{G}$ のガウス分布数とすると，発話 n の事後確率ベクトルは $\boldsymbol{p}_n = [p(g_1|\boldsymbol{y}_n), p(g_2|\boldsymbol{y}_n), \cdots, p(g_D|\boldsymbol{y}_n)]^T$ と表すことができる．本報告では，式 (1) の事後確率で構成されるベクトルを発話ベクトルと呼び，後段の処理でデータ分割の良し悪しを評価するために用いる．

先行研究において，Beaufays らは我々のアプローチと同様に，音響モデルから事後確率を計算することによって発話ベクトルを構成している [9]．ただし，Beaufays らの方法では，事後確率ベクトルを認識処理に用いる大きなサイズの音響モデルから生成しているため，発話ベクトルが高次元であり，またスパースな構成となっている．これに対し，我々の提案する事後確率ベクトルは，認識用の音響モデルから直接生成するのではなく，認識用の音響モデルについてクラスタリングを行い，音素情報を一般化したガウス分布集合から求めている．データクラスタの音響的独立性を正しく評価するためには，事後確率ベクトルは発話内容に依存せず，音響環境のみを反映させることが望ましい．音響モデルからガウス分布集合を一般化する方法は，特徴空間の識別学習にも用いられており，そこから生成される発話ベクトルは発話内容に依存せず，音響環境のみを忠実に反映させることができる [10]．本報告では，発話ベクトルの次元数を 512 とした（$D = 512$）．

3.2 コサイン類似度によるデータクラスタの生成法

クラスタリングによってデータクラスタの独立性を高めるためには，同一クラスタ内で音響的類似度が高く，一方で異なるクラスタ間ではなるべく異質な特性を持つようなクラスタを生成することが望ましい．ここでは，前節で示した発話ベクトルを用いてデータクラスタの独立性（非類似性）をスコア化する方法を示すと共に，クラスタリングの流れについて説明する．以下では，このスコアをデータクラスタ生成に対するスプリットスコアと呼び，隠れ属性によるデータ分割後のクラスターペアから計算する．隠れ属性によって生成されるデータクラスタのペアを全てスコア化し，最も高いスコアを示す属性を現在対象としているデータセットの分割に用いる．具体的な処理ステップを次に示す．

(a) 音声データセット χ を隠れ属性（性別，話速など）の内の一つを用いて，二つのクラスター χ_1 と χ_2 に分割する．隠れ属性が SNR のような連続値を取る場合，分割後のクラスタがほぼ同サイズになるように分割する．

(b) 3.1 節で示した発話ベクトルを用いてスプリットスコアを計算する．まずクラスタ内の音響的類似性について言えば，発話集合が同じような音響性質を示す場合，事後確率ベクトルはほぼ一定方向を指し示すことが予想される．この性質を利用してクラスタ内コサイン類似度 c_W を次のように定式化する．

$$c_W = \frac{1}{N}\sum_{i=1}^{2}\sum_{p_n\in\chi_i}\frac{\boldsymbol{p}_n\cdot\boldsymbol{m}_i}{|\boldsymbol{p}_n||\boldsymbol{m}_i|}, \quad (2)$$

ここで $\boldsymbol{m}_i$ は i 番目のクラスターにおける発話ベクトル集合 $\boldsymbol{p}_n$ の平均ベクトル，N は全データ数を表す．データ分割後のクラスタ χ_1 および χ_2 がクラスター内で同じような性質を示す場合，クラスタ内コサイン類似度 c_W は 1 に近づく．続いて，クラスタ間の類似性は各クラスタの平均発話ベクトルを用いて次のようにスコア化する．

$$c_B = \frac{\boldsymbol{m}_1\cdot\boldsymbol{m}_2}{|\boldsymbol{m}_1||\boldsymbol{m}_2|}, \quad (3)$$

ここで c_B はクラスタ間コサイン類似度である．このスコアはデータクラスタ χ_1 と χ_2 の性質が異なるほど小さな値をとる．最終的に，分割されたデータ集合 $\{\chi_1, \chi_2\}$ のスプリットスコア J_c を，クラスタ内・クラスタ間コサイン類似度を用いて次のように定義する．

$$J_c(\chi_1, \chi_2) = c_W - \alpha c_B, \quad (4)$$

ここで α スケーリング係数である．このスプリットスコアが大きいほど，各クラスターは独立性（非類似性）が高いと見なす．

(c) 残りの隠れ属性についてもスプリットスコアを算出し，最も高いスコアを示す属性を，そのノードで最良のクラスタを生成する属性であると判断する．

(d) ステップ (a)〜(c) で得られた最大非類似データ集合を $\hat{\chi_1}$，$\hat{\chi_2}$ とし，続いて $\hat{\chi_1}$，$\hat{\chi_2}$ をそれぞれ新たな χ として，所望のクラス数になるまでデータ分割を繰り返す．

提案法では，データ分割に音声の隠れ属性を利用しているが，Beaufays らの方法のような教師なしクラスタリング [9] についても，分割後のクラスタを式 (4) で評価することができる．そのため，教師あり／教師なしの両方を用いるハ

イブリッド・クラスタリングのような形式を取ることも可能である．

他方，スプリットスコアの算出にはコサイン類似度尺度に代えて分散に基づく尺度も考えることができるが，分散尺度は発話の音響環境だけでなく，発話内容そのものにも影響をされてしまうので，提案手法からは除外することとした．これについては，5.1 節で具体例を示す．

4. 実験の概要

4.1 音声データ

英語の検索語彙発声（音声検索）に特化した大語彙連続音声認識システムを用いて提案手法の評価を行った．現在，英語の音声検索タスクを対象としたベンチマークテストは存在しないため，データセットは独自に収集した英語音声を用いた．音声データは，数百名の話者からなる検索語彙発話であり，話者ごとのデータサイズは数秒から数時間と幅広いものとなっている．音声検索を対象とした研究は近年盛んになりつつあり，ベースラインとして用いられているシステムの性能は単語誤り率（WER：Word Error Rate）で概ね 16%〜25%程度の範囲に分布している [11], [12]．

4.2 ベースラインモデル

学習・評価データは共にサンプリング周波数 8kHz で収録されており，音声データから，256 点 FFT および 23 チャンネルのバンドパスフィルタ処理を経由して，13 次元の PLP（Perceptual Linear Predictive）特徴量を抽出する．PLP は発話区間のみで求めたケプストラム平均から発話単位の CMN を行っている．その後，PLP 特徴量について前後 4 フレームを結合して合計 9 フレームの音声セグメントを構成し，LDA（Linear Discriminant Analysis）変換を通じて 40 次元の特徴量（LDA 特徴量）に圧縮する．LDA 特徴量は STC（Semi-Tied Covariance）によって近似的に対角化している [13]．

音響モデルは話者独立であり，最尤基準でモデルを学習した後，Boosted MMI 基準の識別学習によってモデルの識別性能を高めている [14]．音素コンテキストは前後 2 フレームを考慮した quinphone モデルとした．実験では，学習データのサイズを変えて 3 種類の音響モデルを構築し，それぞれ性能を比較している．表 1 に学習データサイズおよび，各データセットに対するベースラインシステムの性能を示す．Baseline A は 150 時間から，Baseline B はおよそ 1000 時間から，Baseline C は 5000 時間以上の音声データから音響モデルを学習している．学習データには別途用意した音声認識システムによってトランスクリプションを付与している．モデルサイズとして Baseline A はガウス分布数 150K，状態数 5000，Baseline B はガウス分布数 200K，状態数 7000，そして Baseline C はガウス分布数 600K，状態数 20000 である．評価には Dev セットと Eval セットの 2 種類を用意し，Dev セットは 4000 発話から，Eval セットは 70000 発話から構成されている．先行研究では，学習データを約 5 倍にすると 5%〜7%の WER を削減する効果があり，またオーダー単位での学習データの増加は相対的に 14%〜16.5%の改善をもたらすことが示されている．我々の実験環境におけるベースラインシステムの性能推移も概ねこの知見に合致している．

表 1 識別学習を導入したベースラインモデルの性能
Table 1 Baseline performance with discriminatively trained models.

System	Data size (hour)	*WER*%	
		Dev	Eval
Baseline A	150	24.0	25.0
Baseline B	1000	22.5	23.1
Baseline C	5000	20.8	21.4

5. 実験結果

5.1 コサイン類似度によるデータクラスタ生成結果

本節では，コサイン類似度に基づくスプリットスコアを用いたデータ分割結果について議論する．図 1 は Baseline C の学習データ（5000 時間）について，データ分割の結果を 2 分木の形で図示したものである．ここで，比較した音声データの隠れ属性は，音声データ長（ファイル長），発話前の無音区間長，発話後の無音区間長，ラティスのサイズ（密度），発話単位の平均尤度，SNR，話速，ユーザーの熟練度，発話区間長，ランダムである．ここでユーザの熟練度とは単一話者の発声回数を閾値処理でスコア化したものである．図 1 において，各ノードは分割に用いた属性に，アークはデータクラスターに対応している．生成された木を見ると，ルートノードでは SNR が選択されていることがわかる．これは SNR が互いに無相関なモデルを構築するにあたって最も重要な要素であることを示している．図 2 はルートノードにおけるデータ属性のスプリットスコアをグラフ化したものであるが，ルートノードでは，SNR のスコアが際立って高いスコアを示していることがわかる．

次に木の第二層を見ると，N_0 ノードでは再び SNR が選択されており，一方で N_1 ノードでは話速が優性となった．図 3, 4 はそれぞれ N_0，N_1 ノードでのスプリットスコアを比較したものである．N_1 ノードのスプリットスコアを見ると，ルートノードおよび N_0 ノードで優性だった SNR のスコアがそれほど高くないことがわかる．これは，高 SNR の音声データは低 SNR のデータと比較して，クラスタ間の独立性を出すためにそれほど重要ではないことを意味している．

図 1 はコサイン類似度によるスプリットスコアでデータ分割を行ったものであるが，ここで分割基準として分散尺度を用いた場合，すなわち式 (2)，(3) の c_W，c_B をそれぞれクラスタ内分散，クラスタ間分散に置き換えた場合には，

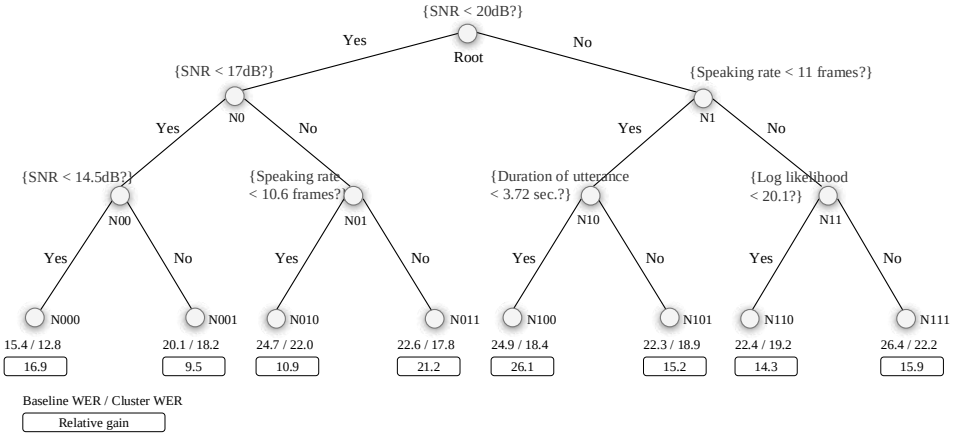

図 1 コサイン類似度指標に基づくデータ分割
Fig. 1 Tree illustrating splits based on cosine similarity-based metric.

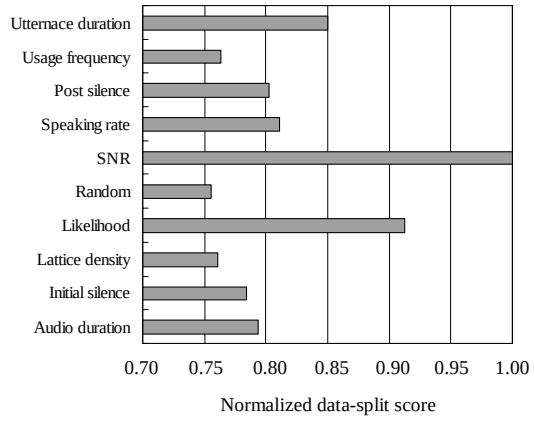

図 2 ルートノードのスプリットスコア
Fig. 2 Split score on root node.

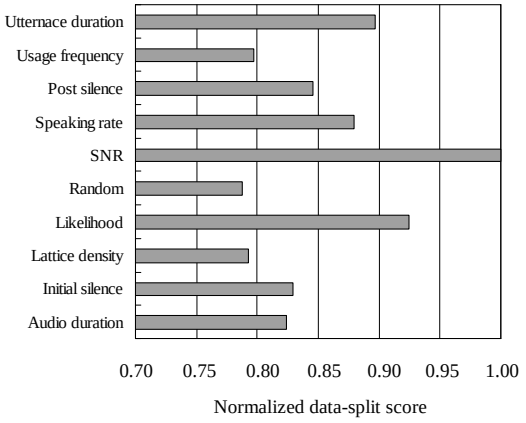

図 3 第二層 N_0 ノードのスプリットスコア
Fig. 3 Split score on N_0 node at second level.

N_1 ノードで発話区間長が優性となった．分散尺度を基準とすると，例えば発話に音素的なバラエティが少ないとき(「青々（あおあお)」,「居合い（いあい)」など)，発話内でのスペクトル変化が少なくなるため，音響環境に関係なく分散が小さくなる[*2]．言い換えれば，分散尺度は音声データの発話内容に影響されてしまうため，N_1 ノードではコサイン類似度尺度とは異なる属性が選択される結果となった．これに対して，コサイン類似度基準によるデータ分割は発話内の音素の偏りに頑健であるため，N_1 ノードでは話速が選択された．

最後に第三層を見ると，N_{00} ノードには SNR が再び現れた．この事実は，低 SNR のデータが強い個性を持っていることを表しており，SNR は無相関性を必要とするシステム統合法の枠組みで最も重要なデータクラスタを生成する属性であることを示唆している．その他，第三層では尤度や発話区間長が選択され，さらには話速も再び優性となった．

[*2] 発話に音素的なバラエティが少ないケースでは，CMN も悪影響を及ぼすことが知られている [15].

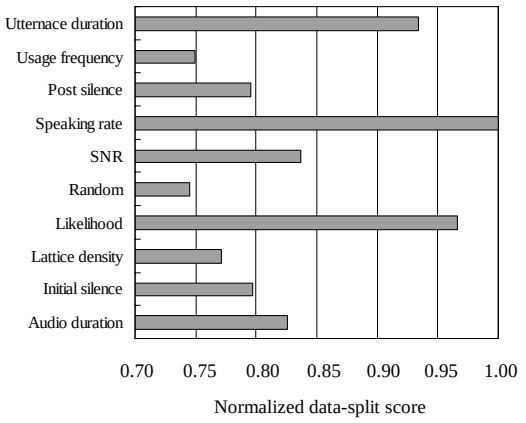

図 4 第二層 N_1 ノードのスプリットスコア
Fig. 4 Split score on N_1 node at second level.

5.2 中規模コーパスによる音声認識結果

提案手法を用いて構築した異種音響モデル群を評価した．ここでは，Baseline B の学習データを 8 クラスタに分割し，それぞれのクラスタで音響モデルを構築したものを

利用した．本実験で，異種音響モデル群は Baseline B の ML 基準によるモデルを出発点として，識別学習の段階で学習データをクラスタ依存のものに変更することによって作成した．すなわち，モデル構造や特徴量，音素決定木などは共通で，音響モデルのパラメータのみが各クラスタに特化したものに再構築される．この学習プロセスは識別学習だけでモデル間の非類似性を生成するため，教師なし識別的環境適応と見なすことができる．本節の実験では，学習データサイズが元の Baseline B のものから 8 分の 1 のサイズになることから，モデルサイズは Baseline A と同じガウス分布数 150K，状態数 5000 とした．

Dev セットによる実験結果を表 2 に示す．実験では，システム統合に n-best ROVER を用いた [3]．表において，“ROVER with random split” は Baseline B の学習データをランダムに 8 分割したものから，それぞれ音響モデルを構築したものである．ROVER によるシステム統合には，Baseline B のモデルも含めている．表を見ると，提案手法は Baseline B と比較して着実な改善を示していることがわかる．別途，ランダムな分割による個々のモデルの性能を調べてみたところ，WER は 24.8%〜25.5%の範囲に分布していることがわかり，一方でコサイン類似度尺度のデータ分割による各モデルの性能は 24.6%〜28.6%に分布していた．ランダムな分割による音響モデル群は専門性に乏しいものの汎用性には富むので，コサイン類似度尺度による音響モデルと比較すると平均的に性能が高い．一方，コサイン類似度による音響モデル群でモデル間の性能差が大きいのは，各モデルの得手不得手が明確になった結果であり，汎用性を捨てて専門性を高めることでシステム統合の枠組みで大きな効果をもたらした．結果として提案手法は Baseline B の単体音声認識システムと比較して，相対的に 2.7%の改善を得ることができた．

5.3 大規模コーパスによる音声認識結果

データサイズに関するスケーラビリティの検証のため，ここでは 5000 時間以上の学習データからなる Baseline C を 8 クラスタに分割して実験を行った．個々のクラスタは Baseline B と同じモデルサイズであるガウス分布数 200K，状態数 7000 とした．前節では ML 基準による単一モデルを出発点として識別学習により 8 種類の異種音響モデルを構成したが，本節では，個々のモデルの非類似性をさらに強化するため，異なる決定木を持つ音響モデル群を構築した．すなわち，8 種類の音響モデルの設計を，音素決定木の構築を含む ML 基準によるモデル推定から始めている．表 3 に実験結果を示す．表から，提案手法は Eval セットにおいて WER が 21.4%から 20.5%に改善しており，認識誤りを相対的に 4%削減していることがわかる．

次に，n-best ROVER によるシステム統合に代えて，モデル選択による理想的な状況下での認識性能を算出した．

表 2 中規模コーパスにおけるデータ分割法の性能比較（Dev セット）
Table 2 Comparison of models built with different partitioning criteria on middle size corpus (Dev. Set)

System	WER%
Baseline B (Single model)	22.5
ROVER with random split	22.3
Proposed: ROVER with cosine split	21.9

表 3 大規模コーパスにおける提案法の性能
Table 3 Results with system combination on large size corpus.

Combination	WER%	
	Dev	Eval
Set C-based baseline single model	20.8	21.4
ROVER with proposed cosine split	20.0	20.5
Model selection (Oracle)	16.6	18.2

ここでは，テスト発話に対して話者単位の WER が最も小さくなるクラスタを最良のクラスタ（オラクルモデル）として選択することとした．図 1 にテスト発話が属するクラスタ単位での WER と，Baseline C に対する改善率を付記している．また，モデル選択によるシステム全体としての認識性能を表 3 の “Model selection (Oracle)” の行に示している．図 1 を見ると，提案手法は N_{000}，N_{011}，N_{100} クラスタで大きな改善を与えていることがわかる．この結果は，低 SNR，話速が遅い発話，そして短い発話は特に認識が難しく，個別に音響モデルを用意することで大きな改善が得られることを示しており，結果としてシステム全体の WER 削減にも大きく貢献する可能性を示唆した．今回の実験では，最適なクラスターを誤りなく選択した場合，最大で 15%の性能改善が得られることを示した．

6. おわりに

本報告では，システム統合法のための効果的な音響モデル集合の構築に焦点を当てたデータ分割法について述べた．提案法では，隠れ属性に応じて音声データの分割を行い，その分割結果をコサイン類似度に基づく目的関数で評価することによって，クラスタ間の独立性を向上させた．提案手法で構築した音響モデル群によるシステム統合は，識別学習を組み込んだ最新の音声認識システムと比較して，相対的に 4%の改善を得ることができることを示した．今後は，モデル選択法についての検討を行っていきたい．

参考文献

[1] W. Reichl and W. Chou, “A unified approach of incorporating general features in decision-tree based acoustic modeling,” *Proc. ICASSP*, pp. 573-576, 1999.
[2] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus among words: Lattice-based word error minimization,” *Proc. EuroSpeech*, pp. 495-498, 1999.
[3] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. Gadde, M. Plauche, C. Richey, E. Shriberg, K. Son-

mez, F. Weng, J. Zheng, “The SRI March 2000 HUB-5 Conversational Speech Transcription System,” *Proc. NIST Speech Transcription Workshop*, 2000.

[4] D. Povey, S. M. Chu, and B. Varadarajan, “Universal Background Model Based Speech Recognition,” *Proc. ICASSP*, pp. 4561-4564, 2008.

[5] G .Cook and T. Robinson, “Boosting the performance of connectionist large vocabulary speech recognition,” *Proc. ICSLP*, pp. 1305-1308, 1996.

[6] Y. Tsao, C. H. Lee, “An ensemble speaker and speaking environment modeling approach to robust speech recognition,” *IEEE Trans., Audio, Speech, and Language Processing*, Vol. 17, No. 5, pp. 1025-1037, 2009.

[7] J. Xue and Y. Zhao, “Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition,” *IEEE Trans., Audio, Speech, and Language Processing*, Vol. 16, No. 3, pp. 519-528, 2008.

[8] R.Tachibana, T. Fukuda, U. Chaudhari, B. Ramabhadran, and P. Zhan, ”Frame-level AnyBoost for LVCSR with the MMI Criterion,” Proc. ASRU, pp.12-17, 2011.

[9] F. Beaufays, V. Vanhoucke, B. Strope, “Unsupervised discovery and training of maximally dissimilar cluster models,” *Proc. Interspeech*, pp. 66-69, 2010.

[10] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, G. Zweig, “fMPE: Discriminatively trained features for speech recognition,” *Proc. ICASSP*, pp. 961-964, 2005.

[11] C. Chelba, J. Schalkwyk, T. Brants, V. Ha, B. Harb, W. Neveitt, C. Parada, and P. Xu, “Query language modeling for voice search,” *Proc. 2010 IEEE Workshop on Spoken Language Technology*, 2010.

[12] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition,” *IEEE Trans. Audio, Speech, and Language Processing*, 2011.

[13] M. J. F. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Trans., Speech and Audio Processing*, Vol. 7, No. 3, pp. 272-281, 1999.

[14] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, K. Visweswariah, “Boosted MMI for model and feature-space discriminative training,” *Proc. ICASSP*, pp. 4057-4060, 2008.

[15] 福田 隆, 新田恒雄, “音韻的偏りに対する推定信頼度を用いた CMN 制御, ” 音講論（春）, Vol. I, 1-5-1, pp.1-2, 2002.

[16] H. Soltau, G. Saon, B. Kingsbury, H. K. J. Kuo, L. Mangu, D. Povey, A. Emami, “Advances in Arabic speech transcription at IBM Under the DARPA GALE program,” *IEEE Trans. Audio, Speech, and Language processing*, Vol. 17, No. 5, pp. 884-894, 2009.