ドメイン選択による複数音声認識システムの効果的活用法*

◎金川裕紀、太刀岡勇気、成田知宏 (三菱電機)

1 はじめに

特定のドメインにおける音声認識の性能を上げるには、そのタスクに特化したデータを用いて学習した音響モデル、言語モデルを使用することが有効である。しかし音声認識システムの学習ドメインと認識ドメインにミスマッチがある場合、性能が低下する傾向がある。したがって、複数ドメインにまたがる発話に対応するためには、マルチドメインのデータをすべて使用した汎用モデルによる認識システムを構築するか、ドメイン毎に複数の認識システムを構築し、各認識システムから得た認識結果を統合する方法 [1, 2] のいずれかの方法が一般的である。後者の方法は、対象ドメインの変更にも容易に対応できるほか、認識結果のスコアをもとにドメインを特定することができるため、言語識別にも応用されている [3, 4]。

これらのメリットから、本報では後者の方法に着目する。このドメイン選択に基づく方法では、従来は認識結果の尤度に応じて最大のものを選択するのが主流であったが、類似するドメイン間では尤度差が小さくなり、ドメインの選択が難しくなるケースが考えられる。そこで学習データの認識時に得られる尤度とドメインの関係を統計的に学習する方法を提案する。

2 最尤ドメイン選択によるマルチドメイン音声認識システム

本節では従来の、D 個のドメインに対応した音声認識システムから、認識結果の尤度に応じて最大のドメインを最適ドメインとして選択する手法について述べる。ドメイン d のデコーダーにおいて得られるフレーム平均された音響尤度と言語尤度をそれぞれ $\mathcal{L}_d^{\mathrm{AM}}$、$\mathcal{L}_d^{\mathrm{LM}}$ とするとき、次式でスコアを定義する。

$$\mathcal{L}_d = \alpha^{\mathrm{AM}}\mathcal{L}_d^{\mathrm{AM}} + \alpha^{\mathrm{LM}}\mathcal{L}_d^{\mathrm{LM}} \tag{1}$$

ここで α と β は、それぞれ $\mathcal{L}_{d,\mathrm{AM}}$、$\mathcal{L}_{d,\mathrm{LM}}$ に対するスケーリング係数である。尤度最大に基づく方法では次式のように、式 (1) のスコアが最大となるようなドメイン $\bar{d}$ を最適なドメインとして選択する。

$$\bar{d} = \operatorname*{argmax}_d \mathcal{L}_d \tag{2}$$

しかしこの方法では、ドメイン間の尤度の差が小さいときに適切なドメインを決定できないおそれがある。

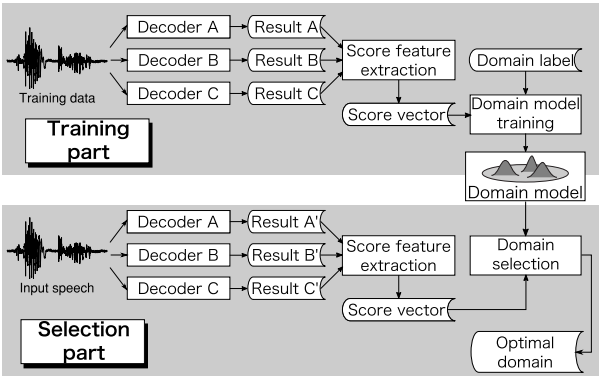

Fig. 1 An outline of the proposed multiple domain ASR systems with statistical domain selection.

また α^{AM} と α^{LM} は最適化できるパラメータではないため、$\mathcal{L}_d^{\mathrm{AM}}$ と $\mathcal{L}_d^{\mathrm{LM}}$ のスケールに応じて実験的にパラメータを調整する必要がある。

3 統計的モデルを用いたドメイン選択法

本節では、提案するマルチドメイン音声認識システムについて述べる。2 節で述べた方法とは異なり、最適ドメインの決定に統計的モデルを用いる。デコーダーから得られる尤度を特徴量とし、正解ドメインとの対応をモデル化するため、各デコーダーの尤度の差に依存せずにドメインの選択ができる。

3.1 提案法の流れ

Fig. 1 に示すように、提案法は学習パートと選択パートの 2 つに分けられる。まず学習パートでは、学習データを各ドメインのデコーダーで音声認識し、それぞれ認識結果 (Result A〜C) とそれらの音響尤度、言語尤度を得る。次に音響尤度、言語尤度を用いてスコアベクトルを抽出する。音響尤度と言語尤度からのスコアベクトル抽出については、次節にて説明する。正解ドメインを教師データとして与え、スコアベクトルとの対応をドメインモデルとしてガウス混合分布モデル (Gaussian mixture model : GMM) により表現する。

選択パートでは、学習パートと同じデコーダーを用いて、入力音声をデコードする。認識結果（Result A'〜C'）からスコアベクトルを抽出したのち、学習パートにて生成したドメインモデルを用いて最適ドメインを選択する。

*An effective exploitation of multiple ASR systems by domain selection. by KANAGAWA, Hiroki and TACHIOKA, Yuuki and NARITA, Tomohiro (Mitsubishi Electric Corporation)

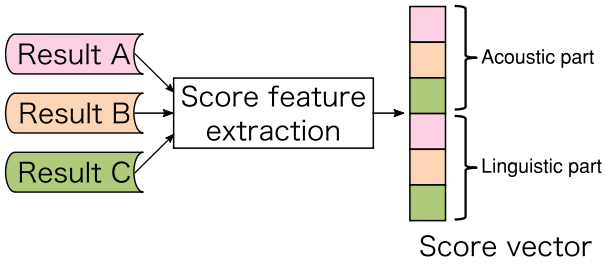

Fig. 2　Concrete example of the proposed score vector extraction by using 1-best ASR results.

3.2　スコアベクトル抽出およびドメインモデルの学習

Fig. 2 に 3 ドメイン、1-best の認識結果からのスコアベクトル抽出例を示す。音響尤度 $\mathcal{L}_{d,\mathrm{AM}}$、言語尤度 $\mathcal{L}_{d,\mathrm{LM}}$ ごとに各デコーダーの尤度を縦に連結する。ベクトルの次元数は音響尤度、言語尤度分に対してドメイン数だけあるため、$D \times 2$ となる。このようにスコアベクトルを設計し、これを全共分散 GMM でモデル化することで、ベクトルの次元間相関により各ドメイン間のスコアの相関を表現できる。

モデルには GMM を用い、学習には、$2D$ 次元のベクトルをそのまま入力とする学習と、音響尤度部分と言語尤度部分を分けたマルチストリームでの学習 [5] の 2 通りを実施する。マルチストリーム学習では音響尤度と言語尤度の相関を共分散行列で表現できない一方で、GMM の次元数を D に抑えられるという利点がある。

まずマルチストリーム無 GMM は、抽出したスコアベクトルのうち、次式で定義される出力確率を最大化するよう学習される。

$$b_j(\boldsymbol{o}) = \sum_{m=1}^{M_j} c_{jm} \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right) \tag{3}$$

ここで $\boldsymbol{o}$, M_j はそれぞれ、スコアベクトル、状態 j の GMM 混合数、c_{jm}, $\boldsymbol{\mu}_{jm}$, $\boldsymbol{\Sigma}_{jm}$ はそれぞれ状態 j、混合 m の GMM 混合重み、平均ベクトル、共分散行列である。

次にマルチストリーム有 GMM では、抽出したスコアベクトルのうち音響尤度部分と言語尤度部分をそれぞれ別ストリームとして扱う。学習は、次式の出力確率を最大化するよう行われる。

$$b_j(\boldsymbol{o}) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_{sj}} c_{sjm} \mathcal{N}\left(\boldsymbol{o}_s; \boldsymbol{\mu}_{sjm}, \boldsymbol{\Sigma}_{sjm}\right) \right]^{\gamma_s} \tag{4}$$

ここで S、$\boldsymbol{o}_s$、γ_s はそれぞれ、ストリーム数、およびストリーム s における特徴量ベクトル、ストリーム重みを示す。また M_{sj} はストリーム s、状態 j の GMM 混合数であり、c_{sjm}, $\boldsymbol{\mu}_{sjm}$, $\boldsymbol{\Sigma}_{sjm}$ はそれぞれストリーム s、状態 j、混合 m における GMM 混合重み、平均ベクトル、共分散行列を示す。

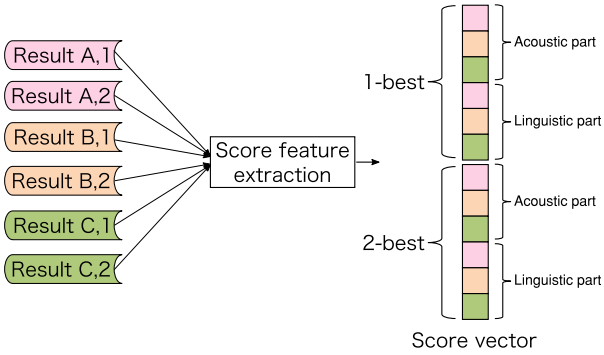

Fig. 3　Concrete example of the proposed score vector extraction by using 2-best ASR results.

3.3　N-best を用いたドメインモデル学習

3.2 節では、認識結果の 1-best から特徴量を生成し、モデルを学習する方法について述べた。さらにシステムからの尤度の出方を詳細にモデル化するために、下位の認識結果も利用してスコアベクトルを抽出する。Fig. 3 に 3 ドメイン、2-best の認識結果からのスコアベクトル抽出例を示す。図中の、認識結果 “Result” に付随するインデックス A〜C はドメインに、1、2 は N-best に対応する。

なお N-best 特徴量のモデル化においてはモデルの次元数増大を抑制するため、N-best 毎にストリームを設定する。また前節同様、音響尤度部分と言語尤度部分に分けてストリームを設定することも可能である。したがって実験では、1-best の音響尤度部分、言語尤度部分をそれぞれストリーム 1、2 に、2-best の音響尤度部分、言語尤度部分をそれぞれストリーム 3、4 にとしてモデル化する場合についても検討する。

4　マルチドメイン音声認識実験

4.1　実験条件

4.1.1　各ドメインの音声認識システム

音声認識実験には、日本語話し言葉コーパス (CSJ)、日本音響学会 新聞記事読み上げ音声コーパス (JNAS)、ATR 自然発話音声・言語データベース (SLDB) の 3 つのデータセットを使用した。各データセットの内訳を Table 1 に示す。SLDB は CSJ と JNAS に比べて学習データ量が少ないため、学習データ量の偏りを減らすべく、フレーム数を基準として CSJ と JNAS の学習データ量を SLDB に合わせて調整した。

これら学習データを用いて、それぞれデータセット毎に音響モデルと言語モデルを学習することで、各

Table 1 The detail of both training and testing datasets.

Subset	Database	Time[hour]	# of utterance	# of speaker
Train	CSJ	17.0	25,298	967
	JNAS	17.6	10,119	268
	SLDB	15.8	12,649	1,097
Test	CSJ	1.7	2,443	10
	JNAS	6.6	3,707	28
	SLDB	1.7	1,421	121

Table 2 A comparison of domain error rate (DER) between the conventional ML domain selection and the proposed method with 1-best ASR result.

Method	Multi stream	CSJ	JNAS	SLDB
ML domain selection	-	17.23	11.79	16.68
Proposed	No	**0.04**	**2.86**	**0**
Proposed	Yes	0.29	5.61	**0**

ドメインに対応する認識システムを構築した。音響モデルは 3,500 状態の GMM-HMM で、ガウス分布の数は 96,000 である。言語モデルはトライグラムで, 各ドメインごとに対応するコーパスを使用し作成した[1]。全評価データをこれら 3 つのシステムで音声認識し、発話毎に、各システムから得られる認識結果のうちシステムを選択し、ドメイン誤り率 (domain error rate : DER) を算出する。その後、選択されたシステムの認識結果を用いて音声認識性能として単語誤り率 (word error rate : WER) を算出する。

4.1.2 ドメインモデルの学習について

提案するマルチドメイン音声認識では、3 節で述べたように、GMM を用いてドメインを選択する。ドメインモデルの学習データは、Table 1 の学習セットをすべてのシステムで認識し、それにより得た尤度を用いてスコアベクトルを抽出することで得た。GMM の混合数は 8[個/ドメイン] で、共分散行列は全分散とした。またマルチストリーム学習における各ストリーム s に対する重み γ_s はすべて 1.0 とした。

4.2 最尤ベースの手法と提案法によるドメイン誤り率による比較

2 節で述べた従来の最尤基準でドメインを選択法と、3.2 節で述べた提案法をドメイン誤り率にて比較する。Table 2 に各データセットに対する両手法のドメイン誤り率を示す。

表中の “ML domain selection” は従来法を示し、“Proposed” は 1-best の音声認識結果を用いた提案法を示す。“Multi stream” はそれぞれ音響尤度と言語尤度のマルチストリーム学習の有 (Yes)、無 (No) に対応している。従来法では音響尤度と言語尤度のスケールが大きく異なるため、音響尤度と言語尤度が同スケールとして扱われるよう事前実験により、式 (1) におけるスケーリング係数を $\alpha^{\mathrm{AM}} = 4 \times 10^{-3}$、$\alpha^{\mathrm{LM}} = 1.0$ とした。

従来法では CSJ や SLDB のドメイン誤りが多く、CSJ を SLDB に、SLDB を CSJ に誤るケースが目立った。この原因として両データセットとも読み上げ文である JNAS とは異なり自由発話調であるため、音響的に類似しており、両者の尤度差が小さかったことが挙げられる。一方で提案法はマルチストリーム学習の有無の両方で全データセットにわたり、従来法よりも大きくドメイン誤り率を改善した。また全データセットのドメイン誤り率において、マルチストリーム学習無の方が優れた。このことから音響尤度と言語尤度間の相関の考慮が有効であるといえる。

Table 3 WER [%] for each decoder with corresponding domain datasets.

Testset	CSJ	JNAS	SLDB
WER [%]	21.80	11.45	16.22

Table 4 WER [%] for each dataset with generic ASR system, which is trained by three domain datasets.

Testset	CSJ	JNAS	SLDB
WER [%]	22.53	13.41	16.49

4.3 単語誤り率による比較

4.3.1 タスク専用システムでの単語誤り率

次にマルチドメイン音声認識実験のための予備検討として、ドメイン毎に構築したシステムにて、対象ドメインのデータセットで音声認識したときの性能を確認する。Table 3 に各データセットと対応するドメインのデコーダーによる単語誤り率を示す。本表に示す値は言語重みを 9 から 20 まで 1 刻みで振り、3 ドメインの平均 WER が最小であった 13 のものである。以降の実験における WER も言語重みを 13 とする。

4.3.2 汎用システムでの単語誤り率

Table 1 のうち、CSJ、JNAS、SLDB の学習セットを全て使用し、音響モデルおよび言語モデルを構築した汎用システムにおける単語誤り率を Table 4 に示す。前節にて検討したタスク専用システムの結果

[1]音響モデルの学習データは Table 1 に示したように間引いたが、テストセットに対してカバー率を下げないことを目的として、言語モデルの学習には学習データ全ての文例を用いた。

Table 5　A comparison of WER between the conventional ML domain selection and the proposed method with 1-best ASR result.

	Multi stream	CSJ	JNAS	SLDB
ML domain selection	-	24.65	12.03	16.37
Proposed	No	**21.80**	**11.72**	**16.22**
Proposed	Yes	**21.80**	12.02	**16.22**

（Table 3）と比較すると、全テストセットにおいてWERが悪い。このことから特定のタスクを認識する場合、汎用のシステムよりもタスクに特化したシステムを使用することが有効であるといえる。

以降のマルチドメインの認識実験ではCSJ、JNAS、SLDBのシステムにおいて音声認識し、ドメイン選択後のWERによりTable 3との比較を行う。

4.3.3　ドメイン選択後の単語誤り率

Table 5に、4.2節にてドメイン選択した後の認識結果に対して単語誤り率を示す。表の見方はTable 2と同様で、従来法である最尤ベースの手法と提案法のWERを比較している。

提案法は最尤ベースの手法よりWERが良く、Table 2の結果と総合して、ドメイン誤り率が減少することで単語誤り率も削減できることが確かめられた。Table 3と本結果を比較すると、従来法ではドメインを誤ることで全データセットにおいてWERが悪化していることがわかるが、提案法では特にCSJとSLDBではドメイン誤りが非常に小さく、WERがTable 3の値と同じとなった。また全データセットで提案法のWERはTable 4に示した汎用システムのものよりも良いことから、タスク専用のシステムからの認識結果を適切に選択することが有効であることがわかった。

なお本実験では、ドメイン誤りが単語の誤りに大きく影響するようなドメイン同士でなかったため単語誤り率に及ぼす影響が小さかったが、カバーする語彙が全く異なるドメインを同時に認識する場合では、ドメインの誤りが単語誤りに大きく影響すると考えられる。

4.4　N-best 認識結果に基づくドメイン選択

ここまで提案法の1-bestの認識結果に基づいたドメイン選択について述べてきた。本節では3.3節で述べた方法により、音声認識結果の1〜10-bestを用いてそれぞれに対しドメインモデルを生成した。音響尤度部分と言語尤度部分のマルチストリーム学習の有無について、N-best毎にモデル化したときの、3ドメインの平均ドメイン誤り率をFig. 4に示す。横軸はドメインモデルに使用したN-best、縦軸はドメイン誤り率である。図よりマルチストリーム学習有の

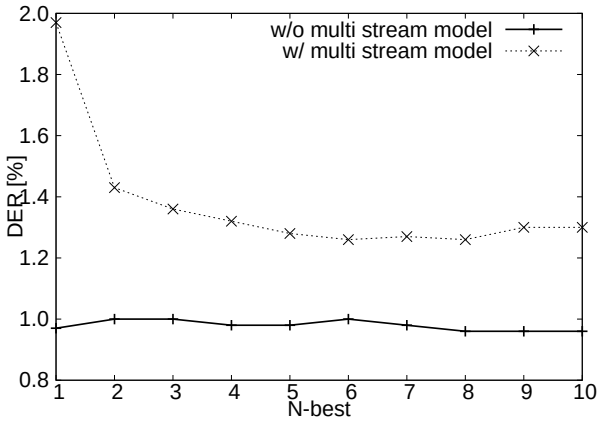

Fig. 4　Average DER [%] in terms of N-best.

場合、N-best数の増加に伴いドメイン誤りが小さくなり、N-bestの使用の有効性を確認した。一方でマルチストリーム学習無では、N-best数を増加させても性能に大きな変化はなかったが、これはGMMにおけるモデル化での性能限界と考えられる。

5　おわりに

複数の音声認識システムより得られる音響尤度と言語尤度を用いて、尤度の出方とドメインの対応関係を統計的にモデル化する手法を提案した。提案法は、従来のスコア最大に基づくドメイン選択よりもドメイン誤り率と単語誤り率の両方の面で優れた。今後は、N-best毎にストリーム重みを変えた場合の検証や、発話長を考慮したドメインコンテキストラベルの導入、またGMMの代わりにニューラルネットの使用を検討する。

参考文献

[1] 神田直之，駒谷和範，中野幹生，中臺一博，辻野広司，尾形哲也，奥野博，“マルチドメイン音声対話システムにおける対話履歴を利用したドメイン選択，”情報処理学会論文誌，**48**，1980–1989 (2007).

[2] 磯健一，“複数の話題言語モデルによる音声認識結果の事後統合，”日本音響学会研究発表会講演論文集 (秋季)，205–206 (2008).

[3] B.P. Lim, H. Li, and Y. Chen, “Language identification through large vocablary continuous speech,” Proc. ISCSLP, 49–52 (2004).

[4] 岡本拓磨，廣江厚夫，河井恒，“尤度ベース言語識別における待ち時間短縮法，”日本音響学会研究発表会講演論文集 (秋季)，23–26 (2016).

[5] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” IEEE Trans. Multimedia, **2**, 141–151 (2000).