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

2007－ＳＬＰ－69（23）
2007／12／20

パス数削減や平滑化法を用いたSSS-freeによる音素認識の高精度化

本間大輔† 大河雄一† 鈴木基之† 伊藤彰則† 牧野正三†

† 東北大学大学院工学研究科
〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-5
E-mail: †{daisuke,moto,aito,makino}@makino.ecei.tohoku.ac.jp, ††kuri@ei.tohoku.ac.jp

あらまし SSS-free で構築された HMnet に基づく音響モデルの各パスは，音素環境，話速等の何らかの環境を表している．しかし SSS-free は構築後の各パスがどの環境に対応したモデルかが分からないという問題点があるため，認識実験に用いる場合には何らかの工夫が必要となる．そこで学習データにおけるパスの接続からパス間に接続確率を与え実験を行ったところ，接続確率の学習データへの特化が原因で認識率にあまり改善が見られなかった．そこで本報告では，クラス N-gram による接続確率の平滑化法，SSS-mix によるパス数削減法を提案した．特定話者における音素認識実験において，提案した両手法は接続確率分布の特化をある程度防ぐことができ，従来法よりも良い結果を出すことができた．

キーワード SSS-free, HMnet におけるパス間の接続確率，接続確率平滑化，SSS-free HMnet パス数削減

Phoneme Recognition with SSS-free HMnet using, Cutting number of paths Method and Smoothing Method

Daisuke HONMA†, Ohkawa YUICHI†, Motoyuki SUZUKI†, Akinori ITO†, and Shozo MAKINO†

† Graduate school of engineering, Tohoku University
6-6-05, Aoba, Aramaki-aza, Aoba-ku, Sendai, Miyagi, 980-8579 Japan
E-mail: †{daisuke,moto,aito,makino}@makino.ecei.tohoku.ac.jp, ††kuri@ei.tohoku.ac.jp

Abstract When carrying out phoneme recognition with SSS-free HMnet's path connection probability, as probability is specialization for training data, phoneme accuracy don't improve. In this paper, We propose smoorhing method and cutting number of paths Method. In phoneme recognition for specific speaker, as a result both of methods prevent connection probability's specialization, phoneme accuracy improve better than conventonal method.

Key words SSS-free, HMnet's path connection probability, path connection probability Smoothing, SSS-free HMnet's cutting number of paths

1. はじめに

音声認識の精度を向上のためには，音響モデルの改善も有効とされている．

従来の高精度な音響モデルである Triphone モデルは，前後の音素環境のみを考慮したモデルであり，任意の音素環境における学習サンプル間の音響的特徴量の違いは，出力確率分布のパラメータで吸収されている．

しかし音素環境が同じであれば，その他の環境 (遠い前後の音素，音素位置，話速等) の影響による，学習サンプル間の音響的特徴量にどれだけ大きな違いがあっても一つのモデルで表現される点が，十分がどうかという問題ははっきり分かっていない．

そこで提案された手法が SSS-free であり，我々は SSS-free[4] で構築された HMnet に基づく音響モデルの高精度化を大きな目的とする．

2. SSS-free

2.1 SSS-free アルゴリズム

SSS-free は音素環境とその他の環境をモデル化できる手法である．SSS-free は初期モデルとして音素モデルを与え，状態を音素環境とは関係なく学習サンプルの音響的特徴量の違いのみを考慮して，HMnet の状態を分割していく手法である．
そのため任意の音素環境において，学習サンプル間の音響的特徴量が大きく異なる場合は，それぞれのサンプルが別々のモデルとして表現され複数のパスが存在する場合や，任意のパスが複数の音素環境を表す場合もある．

よって構築後の HMnet に基づく音響モデルである任意のパスが, 'ある音素環境で, 話速が速いモデル' などのように, 各々のパスは音素環境やその他の環境 (遠い前後の音素, 音素位置, 話速等) 等の何らかの複雑な環境を表現したモデルである.

2.2 認識時における問題点

SSS-free で生成したより詳細に環境を表現した音響モデルを用いて認識を行う場合, SSS-free は構築後の各パスがどの環境に対応したモデルかが分からないという問題点があるため, 何らかの工夫が必要となる.

そこでまず SSS-free を用いた認識実験を行うため, 各パスがどの音素をモデル化したものかは分かっているので, 構築後の HMnet を音素が同じパスへは等確率に遷移するというマルチパスモデルで表現し音素認識実験を行った.

2.3 SSS-free 音素認識実験

実験条件を表 1, 認識率の結果を表 2 に表す.

'Triphone' が状態共有 Triphone.

'SSS-free Same' が SSS-free HMnet において音素が同じパスへ等確率に遷移するマルチパスモデルの結果である.

今回, 言語モデルは音素認識実験のため, 一般的な単語に確率のついた単語 N-gram ではなく, 音素やパスが語彙であり, 確率の付いた音素 N-gram パス N-gram 等を用いた. そのため接続確率モデルと表記することとする. 共に学習データから生成した音素 3-gram を用いた.

結果から 'Triphone' と 'Same' を比較すると, 認識率に改善が見られないことが分かる. 原因として, SSS-free HMnet における各パスは何らかの環境を表現したモデルであるので, 本来であれば環境が対応するパス同士を連結して認識に用いたい. しかしマルチパスモデルではその点を考慮していないため結果が良くなかったと考えられる. そのため環境毎にパスの接続を制御するため, パス間に接続確率を導入することを提案した.

3. パス接続確率

音素環境, 話速等の環境が似通ったパス同士は接続しやすく, 異なる環境のパス同士は接続しづらいなどのような, パスの接続を考慮して認識を行うため, パス接続確率を提案した.[1]

パス接続確率算出の手法を以下に説明する.

接続確率算出手順

(1) SSS-free HMnet 生成

(2) HMnet と学習データとのアライメントを取り, 各学習サンプルがどのパスを遷移しているのか調べる

(3) N-gram の確率計算を用いて, パスに確率を与える

3.1 音素認識実験結果

実験条件を表 1, 認識率を表 2 に表す.

'SSS-free Path' がパス接続確率 (パス 3-gram) を導入した結果である.

表 2 から 'Path' が 'Same' よりも 認識率が良いことが分かる. パスの接続を考慮した結果であると考えられるが, 'Triphone' よりは劣っており, 期待していたほどの改善は見られなかった.

表 1 実験条件

音声コーパス	ATR B set 503 文
	学習 : 400 文, 評価 : 103 文
話者	6 名 (特定話者)
状態共有 Triphone	500 状態 2 混合
SSS-free HMnet	500 状態 2 混合
接続確率モデル	Tied-State Triphone : 音素 3-gram
	SSS-free Same : 音素 3-gram
	SSS-free Path : パス 3-gram
back-off 平滑化	witten-bell
認識エンジン	HTK

表 2 6 名話者平均 音素認識率

Triphone	SSS-free Same	SSS-free Path
91.46 [%]	89.78 [%]	90.70 [%]

3.2 パス接続確率の問題点

そこで学習データから作成したパス接続確率がテストセットに対してどの程度有効かを Perplexity で評価することとする. 結果が表 3 である.

表 3 6 名話者平均 Perplexity

音素 3-gram	SSS-free パス 3-gram
音素数 : 43 音素	パス数 : 3655 本
学習セット : 8.23	学習セット : 3.29
テストセット : 9.12	テストセット : 404.26

表 3 から '音素 3-gram' の学習セット, テストセットの Perplexity がほぼ同じ値をとっているの対して, 'パス 3-gram' は学習セットの Perplexity がテストセットと比較して非常に低い値であることが分かる.

このことから 'パス 3-gram' は, パス数に比べて学習サンプルが不足しているため過学習が起こり, 学習データに特化してしまったため, テストセットに対して有効に機能しなかったと考えられる. よって頑健な接続確率分布を得るため, 今回大きく分けて 2 つの手法を提案する.

1 つは既に得ているパス接続確率分布について平滑化を行う手法. もう 1 つは認識に邪魔をする余分なパス表現がない HMnet を構築し, パス数を減らすことで学習データに特化していないパス接続確率分布を得る手法である.

まず平滑化手法について紹介する.

4. パス接続確率平滑化法

4.1 クラス N-gram による平滑化

SSS-free で構築した HMnet には, 環境が似通ったパスが複数存在すると考えられる. これらのパス同士は, 同じような接続確率分布を持つと考えられるが, 学習サンプルが足りないと, 異なった分布になってしまう. そこで, 似たパスを 1 つのクラスとし, クラス N-gram を導入することでこうした過学習を防ぎ, 頑健な接続確率分布を得る.[2]

クラス N-gram 作成の手順を以下に説明する.

クラス N-gram 作成手順

（1） 何らかのパス間距離を定義し, HMnet の各モデル化している音素が同じパスについて求める.

（2） 距離尺度に基づき LBG クラスタリングを行うことで, 環境が似通ったパスをクラス毎にまとめる.

（3） 得られた各クラスタをそれぞれ 1 クラスとし, 学習サンプルからクラス N-gram を学習する.

4.2 パス間距離

クラス N-gram ではどうクラスを設計するかが大きな問題であり, 距離尺度であるパス間距離をどう定義するかも重要な問題である. 今回は 2 種類の距離定義を用いた.

4.2.1 分割履歴を考慮したパス間距離

SSS-free では, 出力分布が最も拡がった状態を分割すべき状態と判断し, 2 つに分割していく. そのため, パスも 2 つに分割されていく. その結果, パスの分割履歴は 2 分木で表現することができる (図 1). ここで, 各ノードはパスを表し, 葉ノードにあるパスが, 現在の HMnet 内にあるパスとなる.

この木において, より下部で分割されたパス同士は似た環境に対応している. そこで, パス間距離を「共通する親ノードまでの距離」と定義する. しかし, 同じ「1 回の分割」であっても, 2 分木の上部での分割と下部での分割では, 上部の分割の方がより距離が遠いと思われる. そこで, 1 世代離れる距離 (NodeDis) を根ノードから 1, 1/2, 1/4 $\cdots$ と定義する. このようにすると, ある親ノード p から, 葉ノード l までの距離は式 (1) で計算でき, 2 つのパス間距離は対応する葉ノード l_1, l_2 間距離 $Dis(l_1, l_2)$ として計算できる

$$\begin{aligned} D(p,l) &= \Sigma_{i=p}^{l} 2^{-(i-1)} \\ &= \Sigma_{i=1}^{l} 2^{-(i-1)} - \Sigma_{i=1}^{p} 2^{-(i-1)} \\ &= 2^{1-p} * \frac{(1-2^{p-l})}{0.5} \end{aligned} \tag{1}$$

$$Dis(l_1, l_2) = D(p(l_1, l_2), l_1) + D(p(l_1, l_2), l_2) \tag{2}$$

$p(l_1, l_2)$: l_1, l_2 に共通する親ノード.

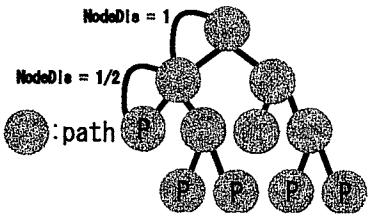

図 1 SSS-free パス分割履歴

4.2.2 bhattachryya 距離を用いたパス間距離

もう一つのパス間距離として, サンプル集合が n 次元正規分布で与えられている場合のパターンクラス間同士の距離である bhattachryya 距離を用いた距離定義とする. 以下に算出過程を説明する.

（1） 各状態は 2 混合正規分布を持つ. bhattachryya 距離を求めるため, 2 混合正規分布の分散と平均を単一正規分布へと近似する.

$$\sigma_{ik}^2 = \lambda_{i1}\sigma_{i1k}^2 + \lambda_{i2}\sigma_{i2k}^2 + \lambda_{i1}\lambda_{i2}(\mu_{i1k} - \mu_{i2k})^2 \tag{3}$$

$$\mu_{ik} = \lambda_{i1}\mu_{i1k} + \lambda_{i2}\mu_{i2k} \tag{4}$$

$\lambda_{i1}, \lambda_{i2}$: 状態 i が持つ 2 つの分布の重み係数

μ_{i1k}, μ_{i2k} : 状態 i が持つ 2 つの分布の k 次元目の平均ベクトル

$\sigma_{i1k}^2, \sigma_{i2k}^2$: 状態 i が持つ 2 つの分布の k 次元目の分散

（2） 状態間の bhattachryya 距離 $u_{12}(\frac{1}{2})$ を求める.

$$\begin{aligned} u_{12}(\frac{1}{2}) &= \frac{1}{8}(M_1 - M_2)^T(\frac{\Sigma_1 + \Sigma_2}{2})^{-1}(M_1 - M_2) \\ &\quad + \frac{1}{2}\ln\frac{|\,(\Sigma_1 + \Sigma_2)/2\,|}{|\,\Sigma_1^{1/2}\,||\,\Sigma_2^{1/2}\,|} \end{aligned} \tag{5}$$

平均 : $\vec{M}$ = (25 次元), 共分散行列 : $\Sigma = diag\{\rho_1 \cdots \rho_{25}\}$

（3） bhattachryya 距離を用いたパターンクラス間の類似度を表すベイズの誤り率を任意の状態間の距離とする ((6) 式). (6) 式中の定義における生起確率は, 各状態の遷移確率をして計算する. ϵ_1 が大きいほど任意の状態同士の持つパラメータが似ていることとなる.

$$\epsilon_1 = (\frac{P(S_2)}{P(S_1)})^{1/2}\exp^{-u_{12}(\frac{1}{2})} \tag{6}$$

$P(S_i)$: 状態の生起確率

（4） 状態間の距離 ϵ_1 を用いて, パス間の距離を DP を用いて求める.

4.3 クラス数の決定

パス間距離を尺度した LBG クラスタリングを以下に説明する.

（1） 代表点

クラス内の全てのパス間距離の総和が最も小さいパス

（2） クラスタリング

代表点に決定したパス間距離の小さいクラスへと分類

（3） クラス数の決定

クラス内における全てのパスの学習データ内の出現頻度の総和を閾値とし, 閾値よりもクラス内のパスの総出現頻度が小さくなったらクラスタリングを止める.

4.4 平滑化手法を用いた音素認識実験

4.4.1 実験条件

2 つの距離定義を用いて, 学習データからクラス 3-gram を作成し音素認識実験を行った. 表 4 が実験条件である.

表 4 実験条件

音声コーパス	ATR B set 503 文
	学習：400 文, 評価：103 文
話者	6 名 (特定話者)
状態共有 Triphone	[200 - 700] 状態 2 混合
SSS-free HMnet	[200 - 700] 状態 2 混合
接続確率モデル	Tied-State Triphone：音素 3-gram
	SSS-free Same：音素 3-gram
	SSS-free Path：パス 3-gram
	Class Path：クラス 3-gram
back-off 平滑化	witten-bell
認識エンジン	HTK

4.4.2 実験結果 (接続確率性能比較)

まず Perplexity の値で評価を行う. (表 5)

'PP' が各接続確率モデルを用いた場合のテストセットに対する Perplextiy である.

'クラス 3-gram Node' が状態分割履歴を考慮した距離定義を尺度としてクラスタリングを行った結果.

'クラス 3-gram Bhattacharyya' が bhattacharyya 距離を利用した距離定義を用いた結果である.

表 5 6 名話者平均 接続確率モデル情報

状態数	500	600	700
パス数	3655	4850	5927
パス 3-gram PP	404.26	739.02	1154.48
クラス数 Node	289	304	330
クラス 3-gram Node PP	203.43	289.00	372.75
クラス数 Bhattacharyy	1921	1649	1694
クラス 3-gram Bhattacharyya PP	320.08	471.19	641.94

クラス 3-gram での結果は, 閾値 (パスの出現頻度の総和) を適当に割り振り, クラス数を決定し, 最も Perplexity が低下した時の値である.

まずどちらの距離定義を用いたとしてもパス 3-gram に比べて値の低下に成功している. 平滑化を行うことで, テストセットの接続確率分布にある程度近いものが得られたと考えられる. 距離定義 'Node' の方が 'Bhattacharyya' よりも良い結果となった.

4.4.3 実験結果 (音素認識率)

Perplexity が最も低い値をとったクラス 3-gram を用いて音素認識実験を行った. 結果を図 2 に示す.

'Class Path Node' が 'クラス 3-gram Node' を用いた結果, 'Class Path Bhattacharyya' が 'クラス 3-gram Bhattacharyya' を用いた結果である.

各手法において最も認識率が高い値で評価する.

まず'Class Path Bhattacharyya' は パス 3-gram を用いた 'SSS-free Path' とほぼ同じ結果となったが, 'Class Path Node' は 'SSS-free Path' よりも 0.33 [%] の改善に成功した. しかし 'Triphone' と比較すると 0.34 [%] 及ばなかった.

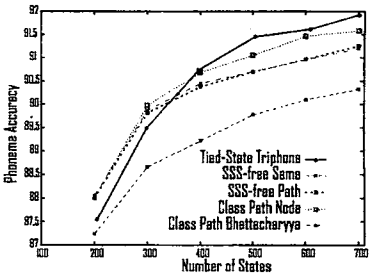

図 2 6 名話者平均音素認識率

4.5 平滑化手法まとめ

距離定義としては 'Node' の方が 'Bhattacharyya' よりも良い結果となった.

' クラス 3-gram Node' の結果は, パス 3-gram に比べて Perplexity の低下に成功し, 更に認識率も 0.33 [%] 改善した.

この実験からパス接続確率に対して適切なクラスを設計し, クラス N-gram を作成することで平滑化が可能なこと, Perplexity の低下が音素認識率の改善にある程度結び付くことが確認できた. そのためクラス N-gram を作成するにあたってパス間距離定義, クラス数の決定方法についてはまだ検討する余地があると考えられる.

平滑化手法を用いて認識率はある程度改善ができたが, 'Triphone' に対してはまだ及んでいない. パス接続確率のパス数に比べて, 学習サンプルが不足しているための特化の影響が非常に大きいと考えられる. しかし状態分割が多いほど認識率の向上の傾向がある程度見られるため, 何らかの方法で音響モデルの精度を保ちつつ, 認識の妨げになるような余分なパス表現を削減する必要があると考えられる.

そこで次にパス数削減法について検討する.

5. パス数削減法

5.1 SSS-mix によるパス数削減法

SSS-free は音素環境とは関係なく音響の類似性に従って状態分割を行っていくので, 特異な特徴を持ったサンプルを独立したパスとして表現する問題があると考えられる.

そこで先行研究でされていた Triphone モデルを初期状態として, SSS-free で状態分割を行う SSS-mix[3] において構築した HMnet にパス接続確率を導入する手法を提案する.

Triphone モデルは音素環境毎にパスが一本である. そのため SSS-mix では初期状態を音素環境毎にある程度分割した Triphone にすることで, SSS-free による早い段階からの特異な特徴をもつサンプルのパス表現を防ぎ, 音響モデルの精度は保ったまま, 認識の悪影響を及ぼす余分なパス表現の少ない HMnet を構築することができる.

5.2 SSS-mix 音素認識実験

5.2.1 実験条件

SSS-mix を用いる場合, 問題となるのがどれぐらいの状態数の Triphone を初期状態とするかである. SSS-free で状態分割

を行うことで, 音素環境でも複数のパスがあったほうが良い場合や, その逆の場合があると考えられるためである.

しかしどの初期状態数が良いかははっきりと分からないため, 今回は Triphone での認識率が最も良かった状態数から 100, 200 ,300 状態前からそれぞれ SSS-free による分割を行い実験を行った. 表 6 が実験条件である.

表 6 実験条件

音声コーパス	ATR B set 503 文
	学習 : 400 文, 評価 : 103 文
話者	6 名 (特定話者)
状態共有 Triphone	[800 - 1000] 状態 2 混合
SSS-free HMnet	[800 - 1000] 状態 2 混合
SSS-mix HMnet	[800 - 1000] 状態 2 混合
接続確率モデル	Tied-State Triphone : 音素 3-gram
	SSS-free Same : 音素 3-gram
	SSS-free Path : パス 3-gram
	Class Path : クラス 3-gram
	SSS-mix Path : パス 3-gram
back-off 平滑化	witten-bell
認識エンジン	HTK

5.2.2 実験結果 (接続確率性能比較)

同様にまず Perplexity の値で評価する. (表 7)

'(Mix)' が SSS-mix による結果である.

まず実験条件から SSS-mix での初期状態が約 500, 600 状態となるため, [800 - 1000] 状態での結果で比較する. '(Mix)' での結果は, それぞれの初期状態数での実験で最も Perplexity が低下した時の値である.

結果からまず SSS-mix のパス数が SSS-free よりも少ないこと, Perplexity の値もより低下していることが確認できる. 余分なパス表現を減らすことである程度頑健な接続確率モデルが得られたと考えられる.

表 7 6 名話者平均 接続確率モデル情報

状態数	800	900	1000
パス数	6949	7867	8745
パス 3-gram	1648.91	2207.51	2864.60
クラス数 Node	284	298	322
クラス 3-gram Node	555.54	661.26	775.00
パス数 (Mix)	2363	3020	3607
パス 3-gram (Mix)	202.90	309.34	423.27

5.2.3 実験結果 (音素認識率)

表 7 の Perplexity の値をとった SSS-mix HMnet において, パス 3-gram (Mix) を作成し音素認識実験を行った. (図 3)

'SSS-mix Path' が提案法を表す.

同様に各手法において最も認識率が高い値で評価する.

結果から提案法は, SSS-free パス 3-gram を用いた 'SSS-free Path' よりも 0.80 [%] の改善に成功し, また 'Triphone' と比較しても 0.13 [%] の改善が見られた.

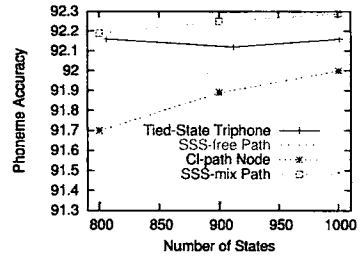

図 3 6 名話者平均音素認識率

5.3 SSS-mix まとめ

SSS-mix における提案手法は, 従来法よりも良い結果となった. 認識の邪魔をするような余分なパス表現を少なくすることができたため, 接続確率分布の学習が SSS-free の接続確率分布よりも十分に行うことができた結果と考えられる.

また従来の'Triphone' よりも僅かではあるが結果が良かったため, 今回は初期状態数を適当に選んだが, 何らかの指標で決定し, SSS-mix での提案法の有効性を確認する必要がある.

6. 総まとめと今後の課題

パス接続確率の学習データへの特化の問題を解決するため, クラス N-gram による平滑化手法, パス数削減手法として SSS-mix を用い, 構築した HMnet においてパス接続確率を導入する手法を提案した.

両手法において, 従来のパス 3-gram よりもテストセットに対する Perplexity の低下に成功し認識率もある程度改善することができた. またパス数削減手法においては, Triphone との比較を行っても認識性能が僅かにだが改善した.

今後の課題としては, クラス N-gram 平滑化手法において, パス間距離定義, クラス数の決定方法の有効性の検討, パス数削減手法において SSS-mix における初期状態決定の問題の解決, また SSS-mix HMnet においてクラス N-gram 平滑化手法を用いての実験を行いたいと思う.

またパス接続確率の一番の問題として, 学習データに存在しないパス間の接続確率を現在は back-off によってのみ与えているが, この確率をどう推定するかについて検討していきたいと思う.

文 献

[1] 本間大輔, 大河雄一, 鈴木基之, 伊藤彰則, 牧野正三 "HMNet のパス接続確率を利用した音素認識の検討" 2-9-9, 春季音響学会, 2007.

[2] 本間大輔, 大河雄一, 鈴木基之, 伊藤彰則, 牧野正三 "複数パスを有する音素モデル連結のためパス間接続確率の平滑化の検討" 3-3-2, 秋期音響学会, 2007.

[3] 坂本創, 鈴木基之, 伊藤彰則, 牧野正三 "SSS-free を併用した音素環境依存 HMnet 学習法の検討" 秋期音響学会, 2005.

[4] M.SUZUKI, S.MAKINO, A.ITO, H.ASO, H.SHIMODAIRA. "A New HMnet Construction Algorithm Requiring No Contexual Factors" IEICE Trans. Syst.& Info., Vol.E78-D, No.6, pp.662-668.(1995)

[5] 鹿野清宏: 「音声認識システム」, 情報処理学会, 1998.