精度の異なる分布の混合による頑健な音響モデル

高野 優

ATR 音声翻訳通信研究所
〒619-0288 京都府相楽郡精華町光台 2-2
Tel.: 0774-95-1319 e-mail: mtakano@itl.atr.co.jp

あらまし 本稿では、自由発話の発声変形によって生ずる精細な音素カテゴリ間の混同を防ぎ、なおかつ周辺音素環境による発声変形に関する音素環境依存モデルの利点を保持するモデルとして、精度の低い分布を混入した音響モデルの使用を提案する。本モデルでは HMM の各状態を表現する分布として、通常の音素環境依存モデルに使用する精細なモデルから得られる分布に音素環境非依存モデル等の粗いモデルから得られる分布を加えた混合分布を使用する。粗いモデルを併用することで、自由発話の発声変形によって生ずる、精細なモデルに適合しない音声の吸収を図る。本モデルを用いた、ホテル予約タスク自由発話認識実験では、同分布数の音素環境依存モデルに比べて一割程度少ない誤認識率を示すことを確認した。

キーワード 自由発話音声, 音素環境依存モデル, 音素環境非依存モデル, 頑健性, 混合分布, 精度, 音声認識

Robust acoustic models by mixing distributions with various precision

Masaru Takano

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288
Tel.: +81-774-95-1319 e-mail: mtakano@itl.atr.co.jp

Abstract In this report, new acoustic models made by mixing probabilistic distributions with various precision are proposed. The proposed models can prevent phonetic confusion among precise phonetic models caused by acoustic variations in spontaneous speech, while keeping the advantage of precise model that can deal with acoustic variations caused by phonetic environment. Distributions in the proposed models are taken from both precise context-dependent(CD) model and another rough model, such as context-independent model. By using distributions from rough model, a model can match acoustic features which don't fit the precise model because of variations due to spontaneity. Experiments on spontaneous speech for hotel reservation task indicated that some of the proposed models can reduce error rate with CD model by around 1/10.

key words spontaneous speech, context-dependent model, context-independent model, robustness, continuous mixture density, precision, speech recognition

1 はじめに

音声認識に使用する単語モデルを構成する音素モデルとして、「音素環境依存モデル」(図1) が使われることが多い。音素環境依存モデルは、周辺音素の影響による発声変形を考慮に入れ、音素モデルを周辺音素によってさらに細分類した精細なモデルである。一般に、音素環境依存モデルは、音素環境非依存の音素モデルよりも良い性能を出すとされており、最近では標準的な音素モデルとなっている。

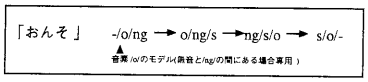

図 1: 音素環境依存音素モデル「おんそ」

しかしながら、自由発話音声認識においては音素環境依存モデルの精細さが逆効果となる場面がある。自由発話では頻繁に発声変形 (語尾の母音無声化、発声途中の母音の曖昧化、等) が起こるからである。自由発話認識を目的とした音響モデルは自由発話音声を学習サンプルとして作成されることが多く、自由発話中に起こるこのような顕著な発声変形は概ね学習されていると思われるが、自由発話音声における変形はさまざまであり、変形のすべての型を学習することは非常に困難であると推測される。

また、自由発話音声には、「～をお願いします」→「～お願いします」のような識別困難な発声が存在する。この発声は、音響的には先頭部分が */o/n(後続音素が /n/ である /o/) であるか */o/o(後続音素が /o/ である /o/) であるかによって識別される。しかしながら、このような音素環境による母音の違いは話者の違いや自由発話による発声変形を上回るほど大きいとは言えないことから、この発声は音響的には非常に識別困難である。このような音声を無理に分類することは誤識別につながりかねない。

これらの観点から本研究では、自由発話における予想不可能な変形に対処するため、以下の方策を考える。

- 一部の音素間に関して意図的に周辺音素環境等の識別を行なわないことにより、学習サンプル中に現れない音響的特徴に対する頑健性を増す
- 学習サンプル中に現れる、特定音素の組合せに特徴的な音響は、そのまま識別に利用する

これらを実現するために、通常の音素環境依存音素モデルに含まれる分布を保持したままで、音素環境非依存モデル等から得られる精度の低い分布を付加、共有するモデルを作成した。また、共有する分布 (緩衝分布) の選択について検討した。

2 識別性能の限界

2.1 自由発話に見られる逆識別

自由発話においては、表1に示すような発声変形により音素間の混同が頻繁に起こる。

そうです	→	そうどぅす、そうれす、そうっす
お直ししします	→	おなおっします、おなおします

表 1: 自由発話における各種発声変形

自由発話音声認識用の音響モデルを学習する際には、自由発話音声そのものが学習サンプルとして用いられることが多い。よって、表1に示すような顕著な発声変形は自由発話音響モデルには既に含まれていると考えられる。しかしながら、発声変形そのものは表1に示すような文字列として明確に現れるとは限らない。変形した音声は、実際にはこれらの音響の周辺に比較的広く分布しているものと推察される。

したがって、たとえ自由発話音声を用いて学習した音響モデルといえども、自由発話における発声変形をすべて網羅しているとは考えづらい。また、精細すぎる分割はかえって逆効果となるおそれがある。

例えば、状態分割 / 混合分布分割によって作成されたモデルで、図2に示されるような認識結果がよく見られる。これは、「ビザ」とい

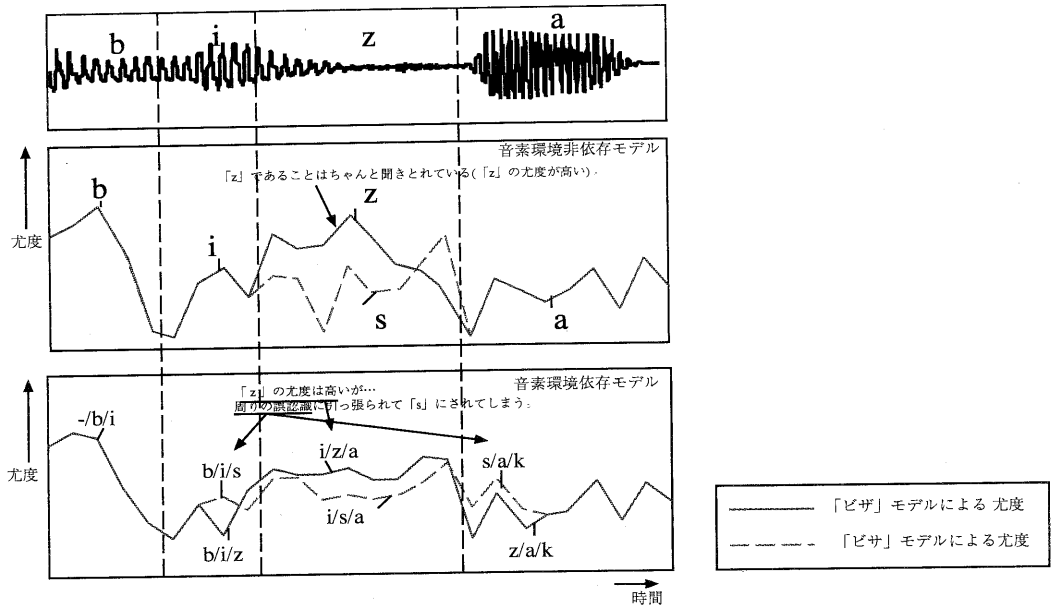

図 2: 周辺音素による尤度逆転に起因する誤認識例

う音声を認識させた例である。音素環境非依存の音響モデルでは正しく認識されているにもかかわらず、音素環境依存の音響モデルでは「ビサ」と誤認識されている。音素環境依存モデルを使用した時の各音素区間の尤度を詳しく調べると、すべての音素が、中心音素という観点から見た場合には正しく認識 (*/b/*, */i/*, */z/*, */a/*) されている。しかしながら、/i/ および /a/ の区間の、音素環境の違いによる尤度の差が、/z/ の区間の「正しい」尤度差を相殺し、結果的に誤認識となる。

この例の場合、音素環境非依存のモデルで正しく認識されていることから、音素そのものが変形して発声されたものではなく、該当部分の音響の微妙な違いによる誤認識と見なすのが妥当である。すなわち、この場合の /i/ および /a/ は自由発話時の細かい発声変形による未学習音声である。

このような未学習音声は、周辺音素環境等によって細かい分類のなされた精細なモデルにとってはほぼ未知音声といえるため、むしろ音素環境非依存モデルのような粗いモデルの方が、よく適合することがあると推定される。

2.2 分布数の増大と性能向上

一般に、音響モデルの状態数または混合数 (ガウス分布数) を増やすと、認識性能は向上する。とはいえ、同一学習サンプルを用いて状態数を次第に増加させると、認識性能は飽和した後、最後には悪化していく (過分割による性能低下 - 図 3)。

原因として、カテゴリ数の増大に伴い、以下の現象が起こることが挙げられる。

- 1 モデルあたり学習サンプル数の減少によるモデルの信頼性の低下
- モデルを表現する分布の尖鋭化、未学習の特徴を持つ音声への適合性の低下
- カテゴリ数の増大に伴うカテゴリ間競合の激化

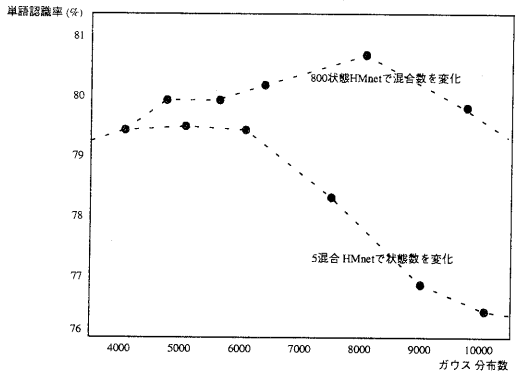

図 3: 過分割による性能低下

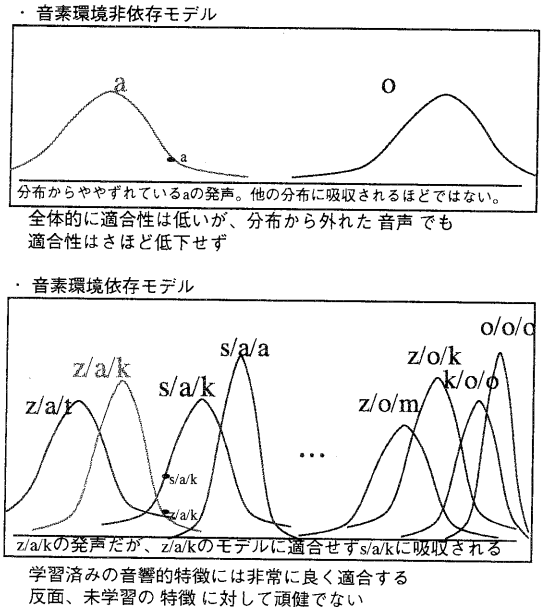

図 4: カテゴリ数の増大による競合激化

モデルあたり学習サンプル数の減少によるモデルの信頼性の低下については、AIC[3] 等の情報量基準を用いて分割数を制御する方法が考えられるが、実際にモデルを作成する場合には有効でないことがある。図 5に見られるように、単純な状態分割 / 混合分布分割により作成したモデルは、AIC の最適値を出すパラメータ数よりも小さいパラメータ数で性能の飽和を見せている。情報量基準を用いるためにはモデル化が有効に行なわれている必要があるが、このような単純なモデル化では、十分有効なモデルが作成できていないことを示唆している。

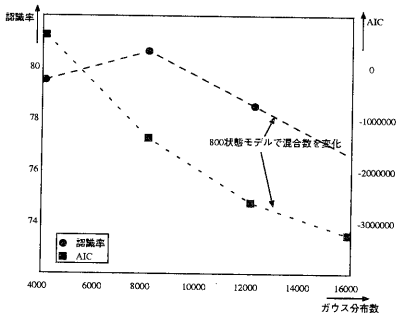

図 5: AIC 最適点以前での性能飽和

3 緩衝分布付加法

以上の観察から、単純な状態分割 / 混合分布分割では対応しづらい例が自由発話中に存在すると言える。本稿では、過分割による誤識別の悪影響を減殺することを目的とし、「識別しきれない音声を識別することを強制しない」モデルを作成する。方式として、カテゴリ間に共有分布 (緩衝分布) を設け、異カテゴリ間に跨る曖昧な音声は緩衝分布で吸収する、すなわち、識別できない音声を該当カテゴリ群のすべてがほぼ同尤度で出力するように混合分布を構成することにした。

3.1 緩衝分布

学習サンプル中に現れない音響的特徴は、HMM の学習によってモデル化されない。した

がってそのような特徴が入力音声中に出現した場合、該当音素の分布にうまく当てはまらず、低い尤度を出すことが多い。そのような音声は誤認識される確率が高いが、誤認識された音素としての尤度が高いというわけではない。すなわち、該当音素は「どのカテゴリにも適合しない」(図6)と判定されているということである。

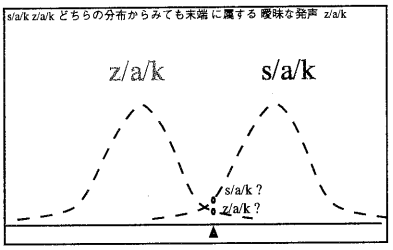

図6: 「どのカテゴリにも適合しない」音声

このような音声に対し、どちらかの音素に分類する(尤度差をつける)ことは無意味であり、誤認識の原因にもなる。図6において、該当音声は「/a/ である」とは言えても、「s/a/k である」、または「z/a/k である」とは言えない。このような場合には、s/a/k や z/a/k でなく「s z/a/k である」、または単に「/a/ である」と判定するのが適切であると思われる。この考え方に基づき、該当音声には曖昧な判定を下すことにする。それには、該当音声に対して両モデルからの出力確率の差が小さくなるように各混合分布を設計するのが望ましい。そこで、両カテゴリの中間的な音響を表す分布を両モデルの緩衝分布として付加・共有する。緩衝分布からの出力確率は両者に同時に寄与することから、両者の中間的な特徴を持つ(緩衝分布の尤度寄与分が大きい)発声に対しては両モデルの出力確率の差が小さくなるような混合分布が構成される(図7)。ここでは、状態集合 $S = \{s_1, s_2, s_3, \ldots, s_N\}$ の全ての状態に、緩衝分布集合 $G = \{q_1, q_2, q_3, \ldots, q_L\}$ を加える方式を採る。

元のモデルにおいて、入力音声 y に対する状態集合 S の各要素 s_i の出力確率 $b_{s_i}(y)$ は、s_i の出力分布を構成する M_{s_i} 個の分布 $P_{s_i} = \{p_{s_i1}, p_{s_i2}, p_{s_i3}, \ldots, p_{s_iM_{s_i}}\}$ および分布重み $\Lambda = \{\lambda_{s_i1}, \lambda_{s_i2}, \lambda_{s_i3}, \ldots, \lambda_{s_iM_{s_i}}\}(\Sigma_{m=1}^{M_{s_i}} \lambda_{s_im} = 1)$ から、以下の式に従って計算される。

$$
\begin{aligned}
b_{s_1}(y) &= \Sigma_{m=1}^{M_{s_1}} \lambda_{s_1m} p_{s_1m}(y) \\
b_{s_2}(y) &= \Sigma_{m=1}^{M_{s_2}} \lambda_{s_2m} p_{s_2m}(y) \\
b_{s_3}(y) &= \Sigma_{m=1}^{M_{s_3}} \lambda_{s_3m} p_{s_3m}(y) \\
&\vdots \\
b_{s_N}(y) &= \Sigma_{m=1}^{M_{s_N}} \lambda_{s_Nm} p_{s_Nm}(y)
\end{aligned}
$$

本方式では、y に対する出力確率 $b'_{s_i}(y)$ は、各状態 s_i の分布集合 P_{s_i} それぞれに、先に挙げた共通の緩衝分布集合 $G(|G| = L)$ を加えた $M_{s_i}+L$ 個の分布の集合 $G'_{s_i} = \{p_{s_i1}, p_{s_i2}, p_{s_i3}, \ldots, p_{s_iM_{s_i}}, q_1, q_2, q_3, \ldots, q_L\}$ および再学習された重み $\Lambda' = \{\lambda'_{s_i1}, \lambda'_{s_i2}, \lambda'_{s_i3}, \ldots, \lambda'_{s_iM_{s_i}}, \lambda'_{s_iM_{s_i}+1}, \ldots, \lambda'_{s_iM_{s_i}+L}\}(\Sigma_{m=1}^{M_{s_i}+L} \lambda'_{s_im} = 1)$ から、以下の式で計算されることになる。

$$
\begin{aligned}
b'_{s_1}(y) &= \Sigma_{m=1}^{M_{s_1}} \lambda'_{s_1m} p_{s_1m}(y) + \Sigma_{m=1}^{L} \lambda'_{s_1,M_{s_1}+m} q_m(y) \\
b'_{s_2}(y) &= \Sigma_{m=1}^{M_{s_2}} \lambda'_{s_2m} p_{s_2m}(y) + \Sigma_{m=1}^{L} \lambda'_{s_2,M_{s_2}+m} q_m(y) \\
b'_{s_3}(y) &= \Sigma_{m=1}^{M_{s_3}} \lambda'_{s_3m} p_{s_3m}(y) + \Sigma_{m=1}^{L} \lambda'_{s_3,M_{s_3}+m} q_m(y) \\
&\vdots \\
b'_{s_N}(y) &= \Sigma_{m=1}^{M_{s_N}} \lambda'_{s_Nm} p_{s_Nm}(y) + \Sigma_{m=1}^{L} \lambda'_{s_N,M_{s_1}+m} q_m(y)
\end{aligned}
$$

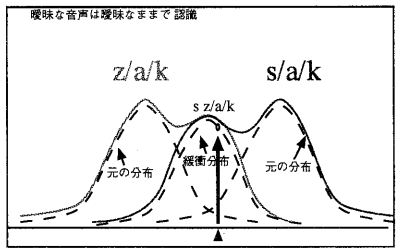

図7: 緩衝分布付加により生成される混合分布

このような混合分布を作成することで、明らかにどちらかのカテゴリに属する音声には従来の分布集合 G_{s_i}、どちらともつかない音声には付加した緩衝分布集合 G で定義される中間カテゴリの出力確率が大きく寄与した尤度を算出することになり、識別し難い音声に関しては曖昧な判定を下す結果となる。

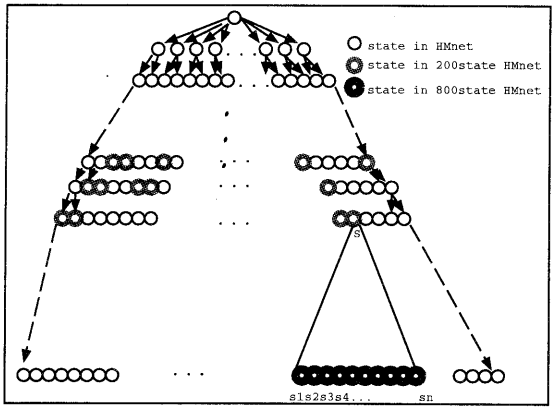

図 9: ML-SSS 状態分割木における先祖と子孫

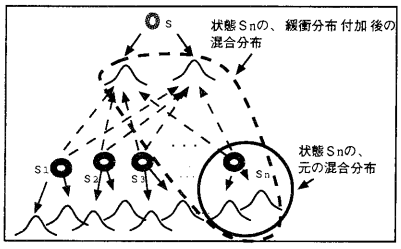

図 8: 状態からの緩衝分布の付加

3.2 緩衝分布の選択

音素環境依存音素モデルは、ML-SSS[4] 等を用いて作成した状態共有構造を持つ最適なモデルでも数百状態程度の状態数を持つ。上記の緩衝分布を全「状態の組」に対して付加することは分布数の非常な増大を招くため、緩衝分布を付加する状態組 (必ずしも 2 つ組に限らない) は必然的に限定せざるを得ない。ここでは、その状態組として、以下を用いることにした。なお、各状態とも対応する分布はガウス分布の混合分布である。

1. 中心音素を共有する音素環境依存音素
2. ML-SSS の分割途中状態およびその子孫にあたる状態

1. は、先に述べたように音素環境から来る細分割の悪影響を緩和するためのものである。2. は、音素環境依存 / 非依存の両音素モデルの中間的な精度を持つ緩衝分布を得るためのものである。モデルは、ML-SSS により状態の逐次分割を行なうことで漸次高精度になってゆくことから、ML-SSS の分割途中のモデルを利用することで、音素環境依存から非依存までの種々の精度の緩衝分布が得られる。

1. も 2. も、状態の組の要素すべてに、付加される状態に含まれるガウス分布をすべて付加する形式になる。例えば、図 8のように、もともと各状態の持つ 2 つの分布に緩衝分布 2 つを共有分布として加え、各状態が緩衝分布を含む見かけ上 4 つのガウス分布からなる混合分布を保有することになる。この状態の組と付加する状態をそれぞれ以下のように選んだ。

1. 中心音素を共有する音素環境依存音素

   音素環境非依存モデルを構成する状態の全分布を、同音素を中心音素とする全音素環境依存モデルの対応する状態に付加

2. ML-SSS の分割途中状態およびその子孫にあたる状態

   ML-SSS による分割途中の HMnet から別に音響モデルを学習し、その各状態を構成する全分布を、子孫にあたる状態すべてに付加 (図 9)

表 2: 実験条件

学習サンプル	日本語旅行対話音声データベース [5] から話者 230 名 (マイク録音)
認識サンプル	同データベースから話者 41 名 (学習サンプルの話者とは完全に異なる。マイク録音)
サンプリング周波数	16000Hz
特徴量	パワーと 12 次元の MFCC、およびそれらの一次回帰係数 (計 26 次元)
フレーム周期	10ms
音響モデル (従来法)	ML-SSS[4] を用いて所定の状態数まで分割、 さらに k-means 法を用いて混合ガウス分布を所定数まで分割ののち、 Baum-Welch 法で再学習 (ただし、各音素の状態数は 3 に固定)
音響モデル (提案法)	元のモデルとして従来法のものを使用 付加分布は、 (1) 音素環境非依存モデル (2) ML-SSS による分割数の小さい分布で作成した従来法によるモデル 緩衝分布付加ののち Baum-Welch 法で再学習
言語モデル (共通)	多重クラス複合 n-gram[6][7]

4 実験

前節で挙げた各選択により緩衝分布を付加したモデルを用い、表 2の条件下で音声認識実験を行なった。実験結果は、表 3に示すようなものとなった。

表 3: 実験結果

従来法 (ガウス分布数)	認識率
800 状態 5 混合 (4010)	79.5%
1000 状態 5 混合 (5010)	79.6%
800 状態 10 混合 (8010)	80.4%
2000 状態 5 混合 (10010)	76.3%
800 状態 15 混合 (12010)	78.6%
提案法 (ガウス分布数)	**認識率**
800 状態 5 混合モデルを、	
CI 5 混合モデルで緩衝 (4395)	81.4%
100 状態 5 混合モデルで緩衝 (4510)	79.8%
200 状態 5 混合モデルで緩衝 (5010)	80.7%
CI 20 混合モデルで緩衝 (5520)	82.0%
400 状態 5 混合モデルで緩衝 (6010)	79.4%

※ CI は音素環境非依存を表す

同ガウス分布数で比較した場合、緩衝分布を用いた音響モデルは概ね、従来法による音響モデルより高い性能を示した。今回の実験の範囲では緩衝分布は粗いモデルから取るのが効果的であると観測される。2.1節に示した誤認識例 (図 2) も、図 10に示すように改善された (提案法のうち、CI 20 混合モデルで緩衝したモデルを使用)。

図 10を見ると、本稿でエラーの主因と見なした /i/ および /a/ 部分での周辺音素環境による尤度差が縮小している。提案法で付加した音素環境非依存モデルの分布による尤度の寄与分が該当音声に対する b/i/z および z/a/k の尤度を底上げした効果と思われる。

すなわち、緩衝分布を導入した目的である、自由発話の発声変形と精細なカテゴリ同士の競合に起因するエラーが実際に改善されている例であると言える。

緩衝分布の種類に関しては、音素環境非依存モデルを使用する方法が最も高性能であった。単に緩衝分布を取得したモデルの分布数を精度として見ると、CI 5 混合モデル (385 分布)<ML-SSS 100 状態 (500 分布)<ML-SSS 200 状態 (1000 分布)<CI 20 混合モデル (1500 分布) であり、最適な緩衝分布は単に分布数からでは予測できないように見える。

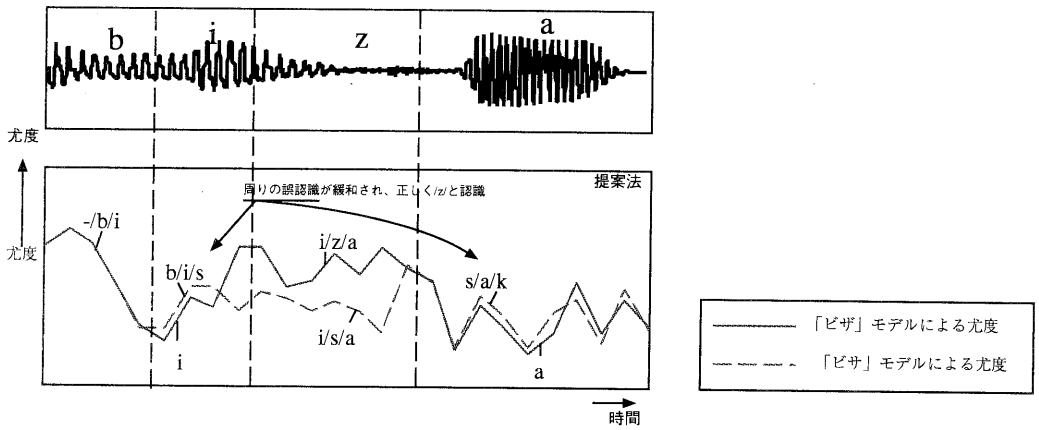

図 10: 周辺音素による細分割に起因する誤認識: 修正後

5 まとめ

自由発話における発声変形が精細な音響モデルに合わずに起こる誤認識を修正するため、緩衝分布の付加による音響モデルの頑健化を図った。実験の結果、自由発話音声に対する認識性能の向上を得るとともに、自由発話の発声変形に起因すると思われる誤認識を実際に修正していることを確認した。

今後、以下の項目について検討したい。

- 音素環境依存音素 – その中心音素、ML-SSS における子孫状態 – その先祖状態、等の人為的な緩衝分布選択法でない、識別性能等の尺度による選択方法
- 緩衝分布の適切な精度および使用する精度の種類と数

参考文献

[1] A. Nakamura. Restructuring gaussian mixture density functions in speaker-independent acoustic models. In *Proc. ICASSP*, pp. 559–562, 1998.

[2] M.-Y. Hwang and X. Huang. Dynamically configurable acoustic models for speech recognition. In *Proc. ICASSP*, pp. 569–572, 1998.

[3] 赤池弘次. 情報量基準とは何か. 数理科学, No. 153, pp. 5–11, 1976.

[4] M. Ostendorf and H. Singer. HMM topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, Vol. 11, No. 1, pp. 17–41, 1997.

[5] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka. Japanese speech databases for robust speech recognition. In *Proc. ICSLP*, pp. 2199–2202, Philadelphia, 1996.

[6] 山本博史, 匂坂芳典. 接続の方向性を考慮した多重クラス N-gram モデル. 音講論, pp. 75–76, September 1998.

[7] 山本博史. 多重クラス N-gram による効率的言語モデル表現. 音講論, pp. 77–78, September 1998.