音声言語情報処理 19－9
(1997. 12. 12)

対立音素間の尤度差に基づく信頼度尺度によるリジェクション

實廣 貴敏　高橋 敏　相川 清明

NTT ヒューマンインタフェース研究所

〒239 神奈川県横須賀市光の丘 1-1
{jitsu,taka,aik}@nttspch.hil.ntt.co.jp

あらまし　音声認識の実際的な使用においては、語彙外発声に対する適切な処理が重要である。代表的なリジェクト方法として、anti-keyword モデルを用いるものがある。これにはタスク依存性があり、大語彙タスクや語彙が切り替わるタスクでは実現が困難である。これに対し、タスクに独立な方法は、語彙制約のない音声認識系を並列に動作させ、尤度正規化を行うものである。しかし、語彙制約なし認識系の精度が一般に低いため、リジェクト性能にも限界がある。そこで、本研究では、タスク独立なリジェクト方法を目標に、語彙制約なし認識系による尤度正規化に加え、各音素における尤度の分布差を利用し、部分区間での照合を導入する。探索時にその差に応じて累積音響尤度に重みづけすることで、各音素での照合精度を向上させる。実験により、等誤り率において音素信頼度尤度を用いない場合に比べ、認識率を4% 改善できた。さらに、誤り最小化学習 (MCE) により 3% 認識率を改善できた。また、演算量削減のため、照合用音響モデルを小規模化し、精度低下はあるものの大きくないことが分かった。
キーワード　音声認識、リジェクション, 未知語, 信頼度尺度, 識別誤り最小化学習

Rejection by Confidence Measure Based on Likelihood Difference Between Competing Phonemes

Takatoshi JITSUHIRO　Satoshi TAKAHASHI　Kiyoaki AIKAWA

NTT Human Interface Laboratories

1-1, Hikarinooka, Yokosuka-Shi, Kanagawa, 239 Japan

Abstract　The rejection of unknown words is important in improving the performance of speech recognition. The anti-keyword model method can reject unknown words with high accuracy in small vocabulary and specified task. Unfortunately, it is either inconvenient or impossible to apply if vocabulary words in the vocabulary change frequently. We propose a new method for task independent rejection of unknown words, where a new phoneme confidence measure is used to verify partial utterances. It is used to verify each phoneme while locating candidates, and the whole utterance is verified by a phonetic typewriter. This method can improve the accuracy of verification in each phoneme. Tests show that the proposed method improves the recognition rate by 4% compared to the conventional algorithm at equal error rates. Furthermore, a 3% improvement is obtained by training acoustic models with the MCE algorithm.

key words　speech recognition, rejection, unknown word, confidence measure, MCE

1 はじめに

音声認識の実際的な使用においては、認識精度も重要であるが、それ以上に、語彙外発声に対する適切な処理が重要である。このため、対象外の発声を検出し、リジェクトするための方法がいくつか提案されている。

代表的なリジェクト方法としては、文献[1][2]などがある。これらは、anti-keywordモデルをタスクに応じて作成し、尤度正規化に用いる。さらに、識別誤り最小化学習(MCE)を用い、対象カテゴリと非対象カテゴリの分布間距離をなるべく引き離し、識別精度を上げるものである。枠組みとしてはタスク依存のリジェクト方法といえる。これらは大語彙のタスクや頻繁に語彙が入れ替わるタスクでは不向きである。

これに対し、タスクに依存しない主な方法としては、文献[6][7]などに代表されるように、語彙制約のない音声認識系を語彙に基づく音声認識系と並列に動作させ、尤度正規化を行い、その正規化尤度の大きさでリジェクションの判定をするものである。語彙制約なし認識系によって、語彙にしばられることなく、入力音声を認識できる。この結果と語彙による認識結果を比較することで、入力が未知語であれば、尤度の差が大きくなり、語彙内か語彙外かの判定に使える。

しかし、一般に現時点での語彙制約のない認識系による精度は低く、確実に入力音声どおりの結果を得ることは期待できない。いくつかの音素や音節では一致しても、挿入、脱落、置換誤りが存在するため、厳密な文字列としての比較はできない。したがって、語彙制約なし認識系による尤度正規化は、全体的な尤度を比較しているに過ぎない。つまり、語彙制約なし認識系の尤度で正規化した場合、語彙内単語と全体的に音素系列が異なるものはリジェクトできるが、部分的に異なるものに対しては、効果的には働かなくなる。

文献[3]では、タスクに依存しない照合方法としてanti-subwordモデルを用いている。一旦、認識した後、認識候補に対し、各subword毎に照合するものである。これは、2段階式の照合方法といえる。これに対し、文献[4]では、1段階式の照合方法を提案しているが、これはHMMの各状態でanti-subwordモデルを用いて尤度正規化し、その値を累積尤度として用いて認識処理を行っている。

本研究では、タスクに依存しないリジェクト方法を目標に、語彙制約なし認識系による尤度正規化に加え、部分的な照合を行う仕組みを導入して、リジェクト精度向上を図る。ここでは、部分的な照合として、音素単位での照合を用いる。他の音素との相対的な分布の差を表す音素信頼度尤度を定義する。この音素信頼度尤度を用いて、音声認識処理での照合途中において、現在の経路における累積尤度に重みづけし、1段階式の照合方法とする。これにより、よりもっともらしい経路は残し、そうでない経路は、認識途中で枝刈りしていくことができる。また、語彙制約なし認識系の累積尤度で尤度正規化を行い、音声終端に達したと同時に、尤度正規化を行うだけで語彙内かどうか判定できる。さらに、部分区間に対する照合により、信頼できる区間を重みづけできるので、認識精度も向上できるという利点もある。

2 音素信頼度尤度を用いたリジェクション

2.1 音素信頼度尤度を用いた認識計算

部分的な尤度分布を反映させることで、リジェクトの精度向上を図る。具体的には、音素単位で他の音素に対し相対的な尤度を求め、その対数尤度を各経路に対する音響モデルの累積対数尤度に加えることで、各音素の確からしさに応じて重みづけする。あらかじめ統計的にこの相対的な尤度分布を求めておき、その分布から認識時に尤度を得る。ここでは、音素単位の相対的な尤度を音素信頼度尤度と呼ぶことにする。

これにより、認識処理の過程で、音素信頼度尤度の小さい音素は枝刈りされる可能性が大きくなる。また、最終的にその音素を含む候補が残った場合でもその候補全体の尤度を下げることになり、誤認識が減る。さらに、未知語の場合でも、音声タイプライタによる尤度正規化で、リジェクトしやすくなると考えられる。したがって、語彙に独立で、より精度良くリジェクションが可能になる。

音素信頼度尤度の定義は次節で行うとし、ここでは、認識処理の流れについて説明しておく。図1に概念図を示す。それぞれの経路において各音素終端で、音素信頼度尤度 $p_i(X_1^2)$ の対数を計算し、定数倍したあと、その時点での累積対数尤度に加算する。累積対数尤度は、前音素までの累積対数尤度 $L_{i-1}(X_0^1)$ と現音素での音響累積対数尤度 $L_i^{(a)}(X_1^2)$ である。

$$L_i(X_0^2) = L_{i-1}(X_0^1) + L_i^{(a)}(X_1^2) + \alpha \times \log\{p_i(X_1^2)\}$$

ここで、X_1^2: 時刻 t_1 から t_2 までの音声特徴量、α は定数。この $L_i(X_0^2)$ をその経路の累積対数尤度とすることで、その音素の信頼度に応じ、重みづけすることになる。

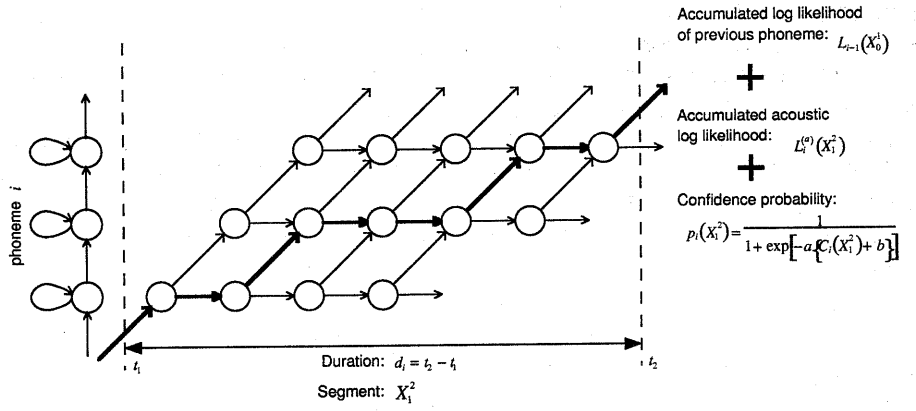


図 1　信頼度尤度の計算

さらに音声終端では、語彙制約なし音声認識系から得られる累積対数尤度、および音声長によって、認識候補の尤度を正規化する。この正規化尤度の大きさにより、リジェクト判定を行う。

2.2　音素信頼度尤度の定義

音素信頼度を以下のように定義する。

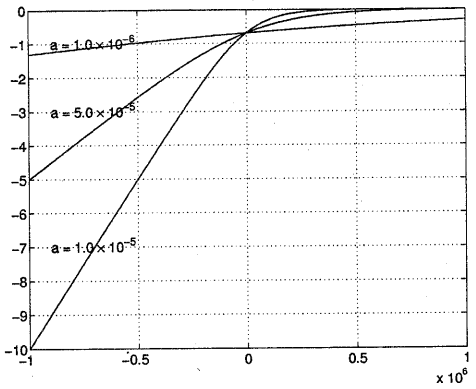
$$C_i(X_1^2) = \frac{1}{d_i}\sum_{t=t_1}^{t_2}\left[g_i(X_t) - \frac{1}{N-1}\sum_{j,j\neq i} g_j(X_t)\right]$$

$g_i(X_t)$: 時刻 t の音声特徴量 X_t に対する、現在注目している候補の第 i 音素モデルの対数尤度、N: 音素モデルの総数、d_i: 継続時間で $d_i = t_2 - t_1$ である。このような定義により、注目している区間が対象音素であれば、高い値を得ることができ、逆に、対象音素とは異なる区間であれば、低い値を得ることができる。ここで、右辺第 2 項は非対象音素すべてに対する平均尤度値としている。全音素中での最大尤度値を用いることも考えられるが、予備的な実験から上記の定義式で安定していたので、これを用いることにした。

この値を確率値として用いるため、シグモイド関数を利用して、音素信頼度尤度 $p_i(X_1^2)$ を定義する。

$$p_i(X_1^2) = \frac{1}{1+\exp[-a\{C_i(X_1^2)+b\}]}$$

ここで、a, b: 定数。$p_i(X_1^2)$ は 0 ～ 1 の間の値を取ることになり、今注目している音素モデルが他の音素モデルに対し、相対的に尤度が大きい場合には、1 に近づき、そうでない場合は、0 に近づくことになる。また、シグモイド関数中の定数 a は傾きを表し、これは実験から設定する。実際には、このシグモイド関数の対数を取って累積対数尤度に加える。図 2に、いくつかの定数 a に対する対数シグモイド関数を示す。横軸が正の領域では、どの a に対しても 0 に近い値を取るが、負の領域では、a の値が大きいほど関数は小さな値を取る。このことから、a の値によって、音素信頼度尤度の重みを制御できることがわかる。定数 b については、実際の音声から信頼度の統計を取り、その最小値を各音素信頼度尺度モデルごとに設定する。



$$y = -\log[1+\exp\{-a(x+b)\}]$$

図 2　対数シグモイド関数

2.3　音素信頼度尤度履歴

発声自体は全体的には了解可能であっても、発声上のゆれや周囲雑音などで部分的には不明瞭な音素が存在する場合もある。そのため、音素信頼度尤度

は必ずしも実際に該当する音素において他の候補に対し、優位な値を得られないときもある。したがって、該当する音素の信頼度だけで重みづけすることは危険なので、信頼度尤度の履歴情報を用いる。

音素単位で得られた信頼度尤度を保持しておき、それを累積対数尤度と同時に伝搬していくことで履歴を残す。各音素終端では、履歴を用いてその経路の累積対数尤度に重みづけする。履歴の平均値を信頼度尤度として再定義することで、極端な信頼度尤度の変動を平滑化できる。

$$L_i(X_0^2) = L_{i-1}(X_0^1) + L_i^{(a)}(X_1^2) + \alpha \times \frac{1}{M+1} \sum_{j=0}^{M} l_{i-j}$$

l_{i-j}: 第 i 音素信頼度対数尤度の j 個前の履歴、M: 履歴の数で、$M = 0$ のときは履歴情報を用いない場合になる。

3 識別誤り最小化学習 (MCE)

識別誤り最小化学習 (MCE)[5] により、モデル間の識別力をあげることで認識精度を上げることができる。また、音素信頼度尤度は、モデル間の識別力が高ければ精度も高くなる。そこで、音響モデルの識別力を高めるため、MCE 学習を導入する。

ここで、MCE 学習について簡単に述べておく。音響モデルのパラメータセット Λ において、観測ベクトル系列 X のクラス k に対する対数尤度を識別関数 $g_k(X, \Lambda)$ として、識別誤り関数を

$$d_k(X, \Lambda) = -g_k(X, \Lambda) + G_k(X, \Lambda)$$

とする。ここで、

$$G_k(X, \Lambda) = \log \left[\frac{1}{K-1} \sum_{j, j \neq k} \exp\{\eta g_j(X, \Lambda)\} \right]^{1/\eta}$$

であり、クラス k に対する対立候補の尤度に相当する。ここで、K は対立候補の数、η は定数である。損失関数をシグモイド関数を用いて以下のように定義する。

$$l_k(X; \Lambda) = \ell(d_k) = 1/\{1 + e^{-a(d_k + b)}\}$$

ここで、a, b は定数である。音響モデルのパラメータは次式で更新される。

$$\Lambda_{t+1} = \Lambda_t - \epsilon_t V_t \nabla l_k(X; \Lambda)|_{\Lambda = \Lambda_t}$$

ϵ_t は学習ステップサイズで小さな正の実数、V_t は正定値行列。各パラメータの微小変動分 $\nabla l_k(X; \Lambda)|_{\Lambda = \Lambda_t}$ を ϵ_t と V_t で制御して、パラメータを更新する。

実際的な学習では、語彙制約なし認識系の複数候補を対立候補として用いる [9]。正解の音素系列で学習音声を音素単位で分割、対立候補において対応するモデルを用いて、新たなパラメータを求める。このような学習により、音素単位で識別学習が可能であり、分布の近い音素同士の分布間距離を引き離すように学習できる。さらに、挿入、脱落、置換誤りを学習することになる。そのため、学習データが音素バランスの整ったデータであれば、汎用的でより精度の高い音素モデルを得ることができる。

4 実験 1

4.1 実験条件

分析条件をサンプリング周波数 12 kHz、フレーム長 32 ms、フレーム周期 8 ms とし、特徴量として 16 次選択線形予測ケプストラム、16 次 Δ ケプストラム、Δ パワーを用いた。音響モデルとして 27 音素 450 状態 4 混合分布の音素環境依存モデルを使用した。総モデル数は 1534 である。

Baum-Welch アルゴリズムによる学習では、学習データとして、ATR データベース A セット音素バランス 216 単語，重要語 5240 単語の男女各 10 名分、日本音響学会データベース 503 文の男性 30 名、女性 34 名分を用いた。そのモデルをさらに MCE 学習した。MCE 学習用のデータとしては、ATR データベース A および C セットの 216 単語を男女それぞれ 20 名分を用いた。

評価は、100 都市名および駅名を含む 1202 単語での単語認識をタスクとした。語彙内の発声として男性 5 名、女性 4 名による 100 都市の発声を用いた。未知語としては、ATR データベース C セットから男女各 10 名の音素バランス 216 単語を用いた。

また、簡単のため、$g_i(X_t)$ については、3 状態音素モデルの中心状態を用いて計算した。

4.2 実験結果

尤度正規化して最終的に得られた候補の正規化尤度を閾値をもとに、リジェクトの判定を行った。この閾値を変えたときの実験結果として、図 3に誤棄却率 (False Rejection Rates) に対する誤受理率 (False Acceptance Rates) を、図 4に誤棄却率に対する単語認識率 (Word Recognition Rates) を示す。図中、"No phoneme confidence likelihood (baseline)" は、信頼度尤度を用いないで語彙制約なし認識系の結果で正規化する場合であり、これが従来法になる。

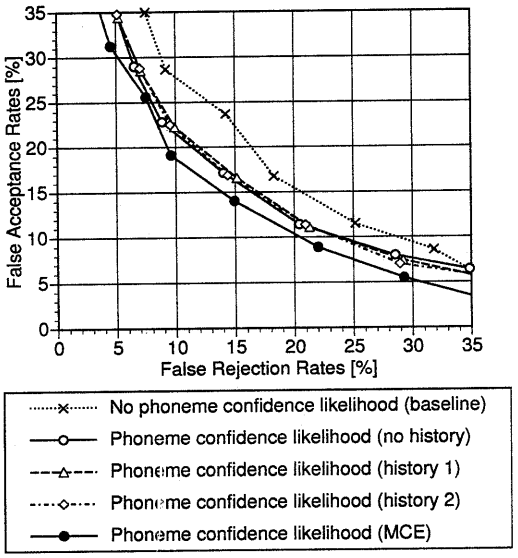


図 3　誤棄却率に対する誤受理率

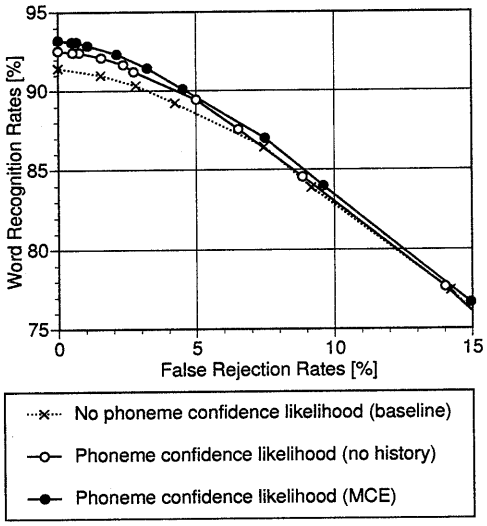


図 4　誤棄却率に対する単語認識率

図中、"Phoneme confidence likelihood (no history)" は音素信頼度尤度を履歴なしで用いる場合、"history 1" や "history 2" は履歴を音素 1 つ前あるいは 2 つ前まで利用する場合である。"MCE" は MCE アルゴリズムにより、音響モデルを学習した場合である。また、シグモイド関数の係数 a としては、MCE 学習した場合は 1.0×10^{-4}、それ以外は 5.0×10^{-5} のときの結果を図に示している。ここで、信頼度尤度を加える際の係数は $\alpha = 1.0$ とした。

表 1　等誤り率

	等誤り率 [%]
音素信頼度尤度なし	18.0
音素信頼度尤度あり (履歴なし)	16.0
音素信頼度尤度あり (履歴なし,MCE)	14.0

表 2　等誤り率での単語認識率

	認識率 [%]
音素信頼度尤度なし	71.0
音素信頼度尤度あり (履歴なし)	75.0
音素信頼度尤度あり (履歴なし,MCE)	78.0

表 3　リジェクトなしの場合の単語認識率

	認識率 [%]
音素信頼度尤度なし	91.4
音素信頼度尤度あり (履歴なし)	92.6
音素信頼度尤度あり (履歴なし,MCE)	93.2

図 3では、曲線が原点に近づくほど精度がよいことを示しており、信頼度尤度を用いることで精度の改善が得られたのがわかる。誤受理率と誤棄却率が等確率になる点では 2% 改善した (表 1)。MCE 学習を用いた場合では、4% の改善になった。その時の単語認識率は表 2に示すように 4% 向上した。MCE 学習により、さらに 3% 向上できた。

また、図 4に示すように、リジェクト性能を高めた場合でも語彙内発声に対する認識率は従来法とほとんど変らないか、精度が高くなっている。表 3にリジェクトを全くしない場合の単語認識結果を示すように、14.0% の誤り改善率が得られた。MCE 学習を用いた場合では、18.6% の誤り改善率となった。これは、信頼度尤度を用いることで認識処理内で各音素の確からしさに応じて重みづけでき、それまで誤認識していた場合でも部分的な精度改善により、正しく認識できるようになっているといえる。MCE に関しては、音素信頼度尤度を用いない場合、MCE 学習により、単語認識率は 93.0% に上昇するものの、図に示してはいないが、FA vs. FR 曲線は baseline の場合とさほど差はなかった。しかし、音素信頼度尤度を用いることでリジェクション精度に改善が見られた。これにより、MCE 学習により、音素信頼度尤度の信頼度が向上しているといえる。

また、履歴情報を用いた場合を比較すると、誤棄却率の高い領域で履歴を考慮しない場合と若干精度がよくなっているが、大きな改善は見られなかった。

5 実験2

実験1では語彙制約なし認識系で用いる音響モデルとして、語彙に基づく認識系と同じモデルを用いた。これでは演算量が多いため、小規模のモデルを使う必要がある。文献[8]などのように、簡易モデル、簡易計算法を使用しても性能はそれほど低下しないという報告がこれまでになされている。しかし、精度を保つため、ある程度音素環境を考慮すべきだと考えられる。そこで、環境依存性は保ったまま、小規模なモデルを作成することにした。実験1で用いた音素環境依存モデルの中心状態を抽出し、規模の小さい音素環境依存モデルへ構成し直した。このモデルはMCEによる再学習は行っていない。このモデルを語彙制約なし認識系と音素信頼度尤度で用いた。語彙ネットワークでは、元のモデルと音素信頼度尤度用の小規模モデルとを関連づけておく。探索時にそれぞれを参照しながら、尤度計算を行う。

5.1 実験条件

前節で用いた音素環境依存モデルからそれぞれの中心状態を取り出し、総モデル数としてはそのままで、総150状態の小規模モデルとした。その他の条件は実験1と同様である。

5.2 実験結果

小規模モデルを使用した場合の誤棄却率に対する誤受理率を図5に示す。図中、"No phoneme confidence likelihood (baseline)"は、信頼度尤度を用いないで語彙制約なし認識系の結果で正規化する場合であり、前節の元の音響モデルを用いた場合の従来法である。"Small model (no confidence likelihood)"とあるのは、音素信頼度尤度なしで語彙制約なし認識系に小規模モデルを用いた場合である。これは小規模モデルを用いた場合の従来法になる。"Small model"は音素信頼度尤度、語彙制約なし認識系ともに小規模モデルを用いた場合である。小規模モデルでの従来法は、オリジナルモデルでの従来法より若干リジェクト精度が下がった。それを音素信頼度尤度を用いることで小規模モデルにおいて、精度を改善できているのがわかる。等誤り率を表4に示す。小規模モデルにおいて音素信頼尤度を用いることで等誤り率は18.8%から16.6%へ改善できている。このときの単語認識率を表5に示す。オリジナルモデルでMCE学習していない場合に近い値となった。

小規模モデルでの誤棄却率に対する単語認識率を図6に示す。オリジナルモデルの場合より、小規模モデルの従来法が若干良いところもあるが、認識精度としてはほぼ同等もしくはやや高い性能が得られた。また、リジェクションなしでの単語認識率を表6に示す。認識率も近い値となった。

小規模モデルを用いることで、前節のオリジナルモデルに比べ、処理速度は約2倍になった。しかし、リジェト性能としては若干低下する。効率よく精度の高い音素信頼度尤度を得られるモデルを設計する必要がある。

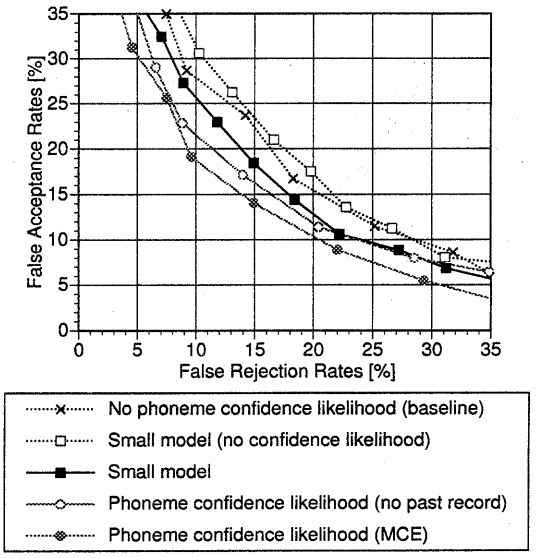


図5 誤棄却率に対する誤受理率(小規模モデル)

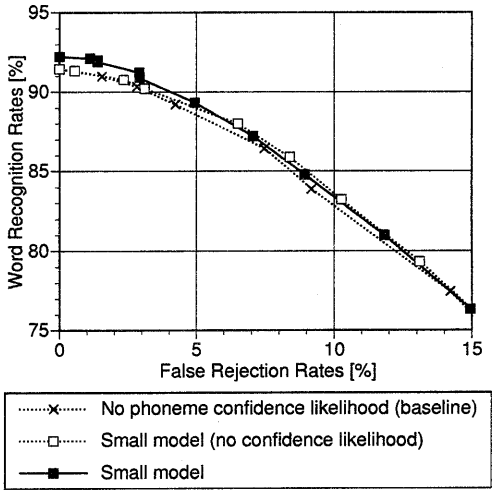


図6 誤棄却率に対する単語認識率(小規模モデル)

表4 等誤り率

	等誤り率 [%]
音素信頼度尤度なし	18.0
音素信頼度尤度なし(小規模モデル)	18.8
音素信頼度尤度あり(小規模モデル)	16.6

表5 等誤り率での単語認識率

	認識率 [%]
音素信頼度尤度なし	71.0
音素信頼度尤度なし(小規模モデル)	70.0
音素信頼度尤度あり(小規模モデル)	74.0

表6 リジェクトなしの場合の単語認識率

	認識率 [%]
音素信頼度尤度なし	91.4
音素信頼度尤度あり(小規模モデル)	92.2

6 まとめ

音素単位での信頼度を定義し、これを尤度として用い、認識処理時に重みづけすることで、未知語のリジェクト精度向上を図った。また、認識精度自体もそれほど変わらないか、条件によっては認識精度を向上させることができた。これは認識候補の部分区間に対し、音素信頼度に応じて重み付けすることができているといえる。さらに、識別誤り最小化学習を用いることで、音素信頼度の精度を改善でき、リジェクト精度も向上できた。状態数を削減した小規模音響モデルについての検討では、オリジナルモデルに近い性能が得られた。しかし、より精度向上を図るために、音素信頼度尤度にとって効率の良いモデルの構築方法を工夫する必要がある。

謝辞

本研究の機会を与えて下さったヒューマンインタフェース研究所音声情報研究部西野部長ならびに日頃から貴重なご意見を頂く音声情報研究部の皆様に感謝いたします。

参考文献

[1] R. A. Sukkar, "Rejection for Connected Digit Recognition Based on GPD Segmental Discrimination," Proc. ICASSP94, pp. 393–396 (1994)

[2] M. G. Rahim, C. -H. Lee, and B. -H. Juang, "Discriminative Utterance Verification for Connected Digits Recognition," IEEE Trans. Speech and Audio Processing, vol. 5, no. 3, pp. 266–277 (1997)

[3] R. A. Sukkar and C. -H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition," IEEE Trans. Speech and Audio Processing, vol. 4, no. 6, pp. 420–429 (1996)

[4] E. Lleida, R. C. Rose, "Efficient Decoding and Training Procedures for Utterance Verificationn in Continuous Speech Recognition," Proc. ICASSP96, pp. 507–510 (1996)

[5] C.-S. Liu, C.-H. Lee, W. Chou, B.-H. Juang and A. E. Rosenberg, "A study on minimum error discriminative training for speaker recognition," J. Acoust. Soc. Am. 97 (1), pp.637–648 (1995)

[6] K. Kita, T. Ehara and T. Morimoto, "Processing Unknown Words in Continuous Speech Recognition," IEICE, Trans. vol. E74, no. 7, pp.1811–1816, 1991.

[7] 伊藤 克亘、速水 悟、田中 穂積: "連続音声認識における未知語の扱い," 信学技報, SP91-96 (1991)

[8] 渡辺 隆夫、塚田 聡: "音節認識を用いたゆう度補正による未知発話のリジェクション," 信学論, J75-D-II, 12, pp. 2002–2009 (1992)

[9] 實廣 貴敏、松永 昭一、嵯峨山 茂樹: "語彙制約なし音声認識における識別誤り最小化学習の検討," 音響学会講演論文集, pp. 107–108 (1997.3)