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

2004－SLP－54（32）
2004／12／21

複数の信頼度尺度を統合した音声認識

小林彰夫† 尾上和穂† 佐藤庄衛† 今井亨†

† NHK 放送技術研究所
〒157-8510 東京都世田谷区砧 1-10-11
E-mail: †{kobayashi.a-fs,onoe.k-ec,sato.s-gu,imai.t-mq}@nhk.or.jp

あらまし 複数の信頼度尺度を統合した音声認識について報告する．信頼度尺度は，一般に音声認識システムが出力する単語仮説の正誤の識別に利用されている．信頼度尺度が仮説の識別に対して高い精度を持つのであれば，システムの単語誤り率の改善に寄与すると考えられる．そこで，システムの出力する単語仮説の識別精度を向上させるために，複数の信頼度尺度を統合して新たな信頼度尺度を構成する．信頼度尺度の統合は，最大エントロピー（ME）法に基づいた統計的手法を用いる．さらに，統合した信頼度尺度を用いて単語仮説の音響・言語スコアを補正し，単語誤り率の改善を試みるリスコアリング手法を提案する．提案手法をサポートベクターマシン (SVM) に基づく統合手法と比較したところ，ME に基づく手法の方が識別誤り率，単語誤り率の改善の両者で改善効果が大きかった．

キーワード 信頼度尺度, 最大エントロピー法, サポートベクターマシン, n-best リスコアリング

Speech Recognition with Integration of Confidence Measures

Akio KOBAYASHI†, Kazuo ONOE†, Shoe SATO†, and Toru IMAI†

† NHK Science & Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan
E-mail: †{kobayashi.a-fs,onoe.k-ec,sato.s-gu,imai.t-mq}@nhk.or.jp

Abstract This paper describes speech recognition using integration of confidence measures. Confidence measures are widely used in classifying word hypotehses into correct or incorrect classes. The measures will make a large contiribution to the improvement of the word error rates (WERs) when they give high classification performances. In this paper we design an integrated confidence measure by maximum entropy (ME) modeling for the classifier. We propose an n-best rescoring method using output scores of the classifier to improve the WERs of the system by correcting the acoustic and language scores. Our proposed method was compared with the classifier based on support vector machines (SVMs), which also integrates the confidence measures. The results of classification and n-best rescoring showed the ME-based classifier performed better than the SVM-based classifier.

Key words confidence measure, maximum entropy models, support vector machines, n-best rescoring

1. はじめに

コーパスに基づく音声言語処理は，数々のアプリケーションで成果を挙げている．例えば，大語彙音声認識システムはテレビ放送のクローズドキャプション（字幕）の作成のためにすでに利用されている [1]．このようなアプリケーションでは，統計的音響・言語モデルが重要な役割を果たしているが，いわゆる読み上げニュースを除けば，満足できるような認識率が得られているとは言いがたい．とりわけ，屋外など雑音環境下での発話や，対談などのややくだけた発話の認識率改善が求められている．

音声認識システムでは，音響・言語モデルによって単語仮説が評価され，もっとも正解らしい単語仮説列が認識結果として出力される．しかし，雑音下や対談といった環境では認識率が低いため，音響・言語スコアの積（和）による評価では不十分である．単語仮説の正解らしさを推定できる他の指標を併用すれば，認識率の改善に役立つと考えられる．

信頼度尺度は単語仮説の正解らしさを表す指標であり，一般には，認識率の高い単語仮説列を音響・言語モデルの学習データとして使う際に用いられる [2] [3]．本稿では，信頼度尺度を音響・言語モデルと併用することによって音声認識システムの認識率の改善を試みる．この際，単語仮説の正解・不正解を高精度に識別する信頼尺度を構成すれば，認識率の改善も大きくなることが期待される．

そこで，識別精度の高い信頼度尺度の構成および認識率の向上を目的として，最大エントロピー（ME）モデルに基づいた信頼度尺度統合による音声認識を提案する．信頼度は識別器の出力スコアとして与えられ，このスコアをリスコアリング時にペナルティとして与えることにより，単語誤り率の改善を図る．一方，強力な識別器として近年よく用いられているのが，SVM（Support Vector Machines）である．本稿では，提案手法をSVMに基づく信頼度統合と比較し，識別精度・単語誤り率による評価を行う．

2. 信頼度尺度

信頼度尺度は，音声認識システムの出力する単語仮説の正解らしさを表す指標である．本稿で採用する尺度は以下の通りである．

- 単語事後確率[4]
- 音響安定度[5]
- 単語仮説密度[6]
- 音響・言語スコア
- バックオフケース
- 探索中にアクティブとなったHMMの数
- 1音素あたりの平均フレーム長
- 直前単語の正解・不正解ラベル

単語事後確率[4]は，音声認識システムの出力する単語lattice上の単語仮説についての事後確率である．音響安定度[5]は，言語重みを変更してリスコアリングしたときの単語仮説列に対し，リファレンス（ある言語重みでリスコアリングしたときの単語仮説列）とマッチする単語仮説の数である．単語仮説密度[6]は，特定フレームにおける単語lattice中の単語仮説の割合である．また，単語仮説の直前単語仮説の正解・不正解ラベルを信頼度尺度として用いる．これは，「不正解単語は連続しやすい」という観察に基づいている．例えば，本稿で用いる評価データ（表2に示すニュース文）では，不正解単語に続く単語の44.5%はやはり不正解単語である．認識誤りの原因として単語境界のエラーがあり，これが連続した単語の不正解につながっていると考えられる．

表1に評価データから求めた各信頼度尺度と正解単語との相関係数を示す．いずれの尺度も弱い相関しかないが，音響安定度の相関係数が最も大きく，次に直前単語ラベル，単語事後確率が大きい．図1のROC(Receiver Operator Characteristic)曲線をみると(注1)，相関係数の順序とは異なり，単語事後確率による信頼度尺度が最も高精度（誤棄却率と誤受理率がともに最小）であることが分かる．

しかし，事後確率や単語仮説密度など複数の信頼度尺度を組み合わせることで，個々の信頼度尺度よりも精度が向上すると期待される．

表1 信頼度尺度と正解・不正解との相関係数

尺度	相関係数
単語事後確率	0.30
音響安定度	0.47
仮説密度	0.24
音響スコア	0.13
言語スコア	0.25
バックオフケース	0.23
HMM数	-0.27
平均フレーム長	0.02
直前単語ラベル	0.37

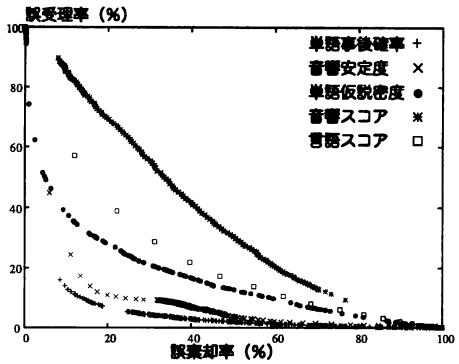

図1 信頼度尺度のROC曲線

3. 最大エントロピーモデルによる信頼度尺度統合

3.1 最大エントロピーモデル

本節では，単語仮説の正解・不正解の推定を確率モデルを用いて定式化し，最大エントロピー（ME）モデルに基づいた複数の信頼度尺度の統合手法を提案する．

データ $(\boldsymbol{x}_t, y_t)$ $(t = 0, 1, \ldots)$ を考える．ここで $\boldsymbol{x}_t$ は，単語仮説 w_t についての単語事後確率，単語仮説密度などの信頼度からなるベクトルであり，$y_t \in -1, 1$ は w_t に対する正解 $(y_t = 1)$ または不正解 $(y_t = -1)$ ラベルである．音声認識システムの出力する単語仮説列 $\boldsymbol{w} = w_0 w_1 \ldots w_t \ldots$ に対し，正解・不正解ラベル系列を求めることは

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} \prod_t P(y_t|\boldsymbol{x}_t) \tag{1}$$

にしたがって最適ラベル系列 $\boldsymbol{y}^*$ を求める問題として定式化され，信頼度ベクトル $\boldsymbol{x}_t$ が与えられたときのラベル y_t の生起確率 $P(y_t|\boldsymbol{x}_t)$ が求めるモデルとなる．

MEモデルではデータの特徴を

$$\mathcal{F} = \{f_i : (\boldsymbol{x}_t, y_t) \longmapsto \{0,1\}, i \in \{1, \ldots, I\}\} \tag{2}$$

で定義される複数の素性関数によって表現する．$f_i(\boldsymbol{x}_t, y_t)$ は，$(\boldsymbol{x}_t, y_t)$ がある制約にマッチするときに1を返すような関数である．一連の学習データ $(\boldsymbol{x}_t, y_t)$ が与えられたとき，モデル

(注1)：誤棄却率と誤受理率はそれぞれ

$$\text{誤棄却率 (\%)} = \frac{\text{誤って不正解に識別された正解ラベルの数}}{\text{正解ラベルの数}} \times 100$$

$$\text{誤受理率 (\%)} = \frac{\text{誤って正解に識別された不正解ラベルの数}}{\text{不正解ラベルの数}} \times 100$$

$P(y_t|\boldsymbol{x}_t)$ は，$f_i(\boldsymbol{x}_t, y_t)$ の P に関する期待値 E_P について

$$E_P[f_i(\boldsymbol{x}_t, y_t)] = E_{\tilde{P}}[f_i(\boldsymbol{x}_t, y_t)] \tag{3}$$

の制約に従う．ここで $E_{\tilde{P}}$ は経験的確率 $\tilde{P}$ による素性 $f_i(\boldsymbol{x}_t, y_t)$ の期待値である．また，$P(y_t|\boldsymbol{x}_t)$ は確率の条件 $P(y_t|\boldsymbol{x}_t) \geqq 0$ および $\sum_{y_t=-1,1} P(y_t|\boldsymbol{x}_t) = 1$ に従う．求めるモデルは以上の条件を満たすもののうち，最大エントロピー原理

$$P_{ME}(y_t|\boldsymbol{x}_t) = \arg\max_P H(P) \tag{4}$$

を満たすものである．ただし，

$$H(P) = -\sum_t P(\boldsymbol{x}_t, y_t) \log P(y_t|\boldsymbol{x}_t). \tag{5}$$

P_{ME} は

$$P_{ME}(y_t|\boldsymbol{x}_t) = \frac{\exp \sum_i \lambda_i f_i(\boldsymbol{x}_t, y_t)}{\sum_{y=-1,1} \exp \sum_j \lambda_j f_j(\boldsymbol{x}_t, y_t)} \tag{6}$$

で表され (λ_i $(i = 1, 2, \ldots)$ は各素性に関する重み)，式 (3) および確率の条件のもと，Generalized Iterative Scaling アルゴリズム [7] により λ_i が繰り返し推定される．

3.2 素性関数の決定法

ここで問題となるのは，各信頼度（またはその時系列）を素性関数を用いてどのように表すかということである．単語事後確率や音響安定度などの信頼度は連続値を取るため，(2) 式のような「計算可能な制約（ある条件を満たしたときに 1 を返すという制約)」を表現した素性関数では工夫が必要である．

ここでは，信頼度に閾値を設け，「閾値以上（以下）である」という制約を用いて，連続値をもつデータを素性関数で表現することを考える．以下では，学習データの信頼度を詳細に表現可能な素性関数の定義について，順を追って説明していく．

閾値を使った素性関数の最も単純なものは，単語仮説 w_t，対応する信頼度ベクトル $\boldsymbol{x}_t$ のある成分 x_t，ラベルを y_t として，例えば

$$f^{simple}(x_t, y_t) = \begin{cases} 1 & if\ x_t > c_T \wedge y_t = 1 \\ 0 & else \end{cases} \tag{7}$$

などと決める．ここで c_T は信頼度に対する閾値である．関数 f^{simple} の学習データにおける期待値 $E_{\tilde{P}}[f^{simple}]$ は，閾値 c_T で単語 w_t を正解・不正解に識別したときに，正解単語となる期待値を示す．この素性から得たモデル（式 (6)）は，$x_t > c_T$ となる入力に対する正解（不正解）の確率を与える．また，条件を $x_t < c_T \wedge y_t = -1$ として素性関数を定義すれば，不正解単語に関する素性となる．

しかし，上記のようにただ一つの閾値で信頼度を二分するような制約では，できあがったモデルはあまりにも単純であり，信頼度に対する正解（不正解）の出現確率を十分に表現できないだろう．そこで，$\boldsymbol{c}_T = \{c_{T_0}, c_{T_1}, \ldots c_{T_k} \ldots\}$ として閾値を複数個決め，式 (7) のようにそれぞれ素性関数を定めれば，入力 x_t に対して正解（不正解）ラベルの確率をより詳細に与えることができる．閾値の組 $\boldsymbol{c}_T$ は実験的に与えることになるが，ここでは，ある点を中心に幅 Δ ずつ離れた点を閾値として採用する．まず最初に，ある信頼度に対して識別誤りが最小になる点を閾値 c_{T_0} とする．c_{T_0} に対して，$\pm\Delta$ だけ離れた点を $c_{T_1} = c_{T_0} + \Delta, c_{T_2} = c_{T_0} - \Delta$ として新たな閾値を 2 点定める．新たに定めた閾値から，式 (7) により素性を定め，モデルを作成する．閾値を追加する際は，さらに c_{T_0} から $\pm 2\Delta, \pm 3\Delta, \ldots$ だけ離れた点を閾値としていく．モデルによる識別誤り率が変わらなくなるまで閾値（素性）の追加を繰り返すことにより，信頼度尺度を表現するのに必要な閾値を順次求めていく．ただし，式 (7) にしたがった場合，$c_{T_i} < c_{T_{i'}}, c_{T_i}, c_{T_{i'}} \in \boldsymbol{c}_T$ に対して $f_{i'}^{simple} = 1$ となれば，必ず $f_i^{simple} = 1$ となるので，$f_{i'}^{simple} = 1$ となるデータ $(\boldsymbol{x}_t, y_t)$ に対しては，f_i^{simple} を定義しない．

単語仮説 w_t に関する信頼度のみではなく，前後の単語仮説から得られる信頼度も，正解・不正解ラベルを得る上で重要と考えられる．注目している単語 w_t の前後の単語仮説で信頼度が低ければ，仮に w_t の信頼度が十分大きくても正解であることは疑わしくなるだろう．そこで，式 (7) の x_t の代わりに x_{t-1}, x_t とし，任意の閾値のペア $c_{T_k}, c_{T_{k'}} \in \boldsymbol{c}_T$ を用いて

$$f^{seq}(x_{t-1}, x_t, y_t) = \begin{cases} 1 & if\ x_{t-1} > c_{T_k} \wedge x_t > c_{T_{k'}} \wedge y_t = 1 \\ 0 & else \end{cases} \tag{8}$$

のようにする．これにより，注目している単語仮説 w_t の直前単語仮説 w_{t-1} の信頼度が，正解・不正解に及ぼす影響を考慮できる．上記により，実験的に定められた閾値を使って f^{seq} を定義していく．前後の単語仮説のうち，いくつまでの単語仮説の信頼度を使うかは，実験的に求める．また，素性関数に用いる信頼度の個数と閾値の数により，上で定義される素性関数の総数は組み合わせ的に増加するが，学習データに含まれない組み合わせは（$E_{\tilde{P}}[f^{seq}] = 0$ となることより）不要な素性となりモデルには含まれない．

複数の信頼度尺度については，第 1 の信頼度尺度による信頼度系列を x_{t-1}^1, x_t^1，第 2 の信頼度尺度による信頼度系列を x_{t-1}^2, x_t^2 として，$\boldsymbol{z}_t = (x_{t-1}^1, x_t^1, x_{t-1}^2, x_t^2)$ とすると，$f^{seq1}(\boldsymbol{z}_t, y_t), f^{seq2}(\boldsymbol{z}_t, y_t)$ などとして個々の信頼度尺度について素性関数を定めれば，式 (6) によりすべての信頼度の系列を考慮したモデルが作成できる．

素性関数の決定スキームのポイントは，信頼度尺度のそれぞれに対して複数の閾値を決定することと，素性関数に採用する信頼度の個数を増やしていくことである．

閾値を使うことにより，連続値を取る信頼度が「閾値より大きい（小さい)」という基準でパターン化される．閾値を複数採用したり，信頼度の系列を使ったりすることは，これらの単純なパターンを組み合わせてより複雑なパターンを構成することを意味する．そして，これらのパターンにマッチする単語仮説が正解（不正解）となる確率は，ME モデルにより与えられる．

まとめると，素性関数は以下の手順で決定される．

Step 1. ある信頼度尺度について，閾値 c_{T_0} を単語の識別誤

りが最小となるように決める．

Step 2. 閾値 c_{T_0} を中心に上下に刻み幅 Δ で新たな閾値を決め，素性関数（例えば (7)）を定義し，ME モデルを学習する．

Step 3. 識別誤り率が変わらなくなるまで，Step 2 を繰り返す．

Step 4. 信頼度の系列の個数を当該単語の前後 1 単語ずつ増やし，素性関数（例えば式 (8)）を定義して ME モデルを学習する．

Step 5. 識別誤り率が変わらなくなるまで Step 4 を繰り返す．

Step 6. すべての信頼度尺度を ME モデルにて統合する．

3.3 関連研究

信頼度尺度の統合については Hazen [8] らの研究がある．これは，複数の信頼度からなる信頼度ベクトルを事後確率最大化の観点から適当な重みで線形に足し合わせたものであり，識別器とは異なる．識別器を使った手法としては，SVM (Support Vector Machines) による信頼度統合 [9] が提案されている．識別器どうしの比較としては，提案手法と SVM との比較が妥当と考えられる．本稿では ME に基づく統合手法を SVM に基づく手法と比較する．

4. SVM による信頼度尺度統合

近年，強力な識別器として，SVM が数々のアプリケーションで利用されている [10]．SVM では，観測データ $\boldsymbol{x}$ を 2 つのクラスに識別するために，識別面（超平面）を用いることが特徴である．$\boldsymbol{x}_t$ に対し，識別面からの距離を

$$f(\boldsymbol{x}_t) = \sum_k y_k \alpha_k \cdot k(\boldsymbol{x}_t, \boldsymbol{x}_k) + b \tag{9}$$

とし，

$$y_t = sign(f(\boldsymbol{x}_t)) = \begin{cases} 1 & if\ f(\boldsymbol{x}_t) > 0 \\ -1 & else \end{cases} \tag{10}$$

で識別を行う．ここで，$\boldsymbol{x}_t, \boldsymbol{x}_k \in \mathrm{R}^m$，$\boldsymbol{\alpha} \in \mathrm{R}^m, 0 \leqq \alpha_i \leqq C$，$b \in \mathrm{R}^1$ とする [10] [11]．$k(\boldsymbol{x}_t, \boldsymbol{x}_k)$ はカーネルとよばれ様々なものが提案されているが，ここでは多項式カーネル

$$k(\boldsymbol{x}_t, \boldsymbol{x}_k) = (\boldsymbol{x}_t \cdot \boldsymbol{x}_k + c)^d \tag{11}$$

を考える．

SVM の出力 $f(\boldsymbol{x}_t)$ を用いて，式 (1) で定式化された問題を解く上で重要なのは，$f(\boldsymbol{x}_t)$ は確率ではなく超平面からの距離になるということである．これを直接式 (1) で用いることはできない．Platt [12] は，超平面からの距離をシグモイド関数

$$P_{SVM}(y_t = 1|\boldsymbol{x}_t) = \frac{1}{1 + \exp(Af(\boldsymbol{x}_t) + B)} \tag{12}$$

によってマッピングし，確率のように扱う手法を提案している．ここで A および B は学習データから求められる定数である．式 (12) は，確率を与えるものではないが，$0 \leqq P_{SVM} \leqq 1$ を満たし，事後確率のように使うことができる．

5. リスコアリングによる音声認識

前述した ME および SVM に基づく識別器の出力スコアは，単語の正解らしさを示す．これを用いて，正解らしい単語仮説を含む文仮説を選ぶことにより，認識率の改善を図る．音声認識システムの出力として得られた n-best 文仮説リストの n 番目の仮説 $\boldsymbol{w}^{(n)}$ に対して，識別器の出力スコアの対数 $\log P(y_t^{(n)} = 1|\boldsymbol{x}_t^{(n)})$ を重み付けして加えて補正し，リスコアリングスコア $S(\boldsymbol{w}^{(n)})$ とする．

$$S(\boldsymbol{w}^{(n)}) = \sum_t \{ac(w_t^{(n)}) + gw \times lm(w_t^{(n)}) + cw \times \log P(y_t^{(n)} = 1|\boldsymbol{x}_t^{(n)})\} \tag{13}$$

ここで，$ac(w_t^{(n)})$，$lm(w_t^{(n)})$ は音響スコアおよび言語スコア（いずれも対数），gw は言語重み，cw は識別器スコアに対する重みである．

単純な補正は，重み cw を固定してリスコアリングスコアへ識別器スコアを加えることである．もうひとつの方法は，識別器スコアを文仮説の認識率に応じて補正することである．つまり，各 n-best 文仮説に対して ME モデルによりラベル付けを行い，正誤のラベル y_t とスコア $\log P(y_t^{(n)}|\boldsymbol{x}_t^{(n)})$ を求め，式 (13) を用いて n-best 文仮説をリスコアリングする．このとき，cw を以下の関数で置き換える．

$$\hat{cw} = \frac{1}{1 + C_a \exp(C_b - e(\boldsymbol{w}^{(n)}))} cw \tag{14}$$

C_a, C_b はそれぞれ定数である．$e(\boldsymbol{w}^{(n)})$ $(0 \leqq e(\boldsymbol{w}^{(n)}) \leqq 1)$ は単語誤り率 (WER) だが，これは不明である．そこで，識別結果から

$$\tilde{e}(\boldsymbol{w}^{(n)}) = \frac{\sum_{t:y_t^{(n)}=-1} |y_t^{(n)}|}{\sum_t |y_t^{(n)}|} \tag{15}$$

を求め，単語誤り率を近似する．単語誤り率に応じたペナルティをリスコアリングスコアに与えることにより，単語誤り率の高い文仮説をいっそう n-best 下位に押し下げることが期待される．

6. 実　　験

6.1 実験条件

ME モデル，SVM モデルの学習データとシステムの評価データを表 2 に示す．学習データは 2001 年 5 月 1 日から 10 日までの放送ニュースから，評価データは同年 6 月 1 日から 14 日までのニュース番組（1 日あたり 4 番組）から選んだ男性発話である．評価データは雑音環境（屋外でのアナウンサー・記者によるリポート），対談環境（2 人の発話者による自由度の高い発話）から構成される．また，学習データには，2 節で述べた各信頼度が付与されている．ME モデルの学習は，データから 8,210 個の素性を選択して行った（表 3）．SVM モデルは SVM^{light} [13] を利用して 4 次の多項式カーネルを用いて学習を行った．

表2 学習・システム評価データ

	文章数	単語数	PP	%OOV	%WER
学習	5,696	139,651	73.1	1.74	13.4
評価	993	23,308	15.7	0.63	8.6

表3 素性関数

信頼度尺度	閾値の数	系列個数	素性の総数
単語事後確率	13	5	7164
音響安定度	11	3	602
仮説密度	1	1	4
音響スコア	7	3	272
言語スコア	1	3	98
バックオフケース	-	3	42
HMM 数	3	3	8
平均通過フレーム数	1	3	20

学習データのうち 18,106 語は，ME モデルの素性関数の閾値，各信頼度尺度に基づく識別器の閾値，デコーダパラメータ (言語重み，挿入ペナルティ) などの決定，および SVM モデルのシグモイド関数 (式 (12)) のパラメータ決定に用いた．

音声認識は第 1 パスで tree lexicon および bigram でデコードしたのち，trigram lattice を生成する．第 2 パスは lattice 上を trigram モデルでリスコアリングし，300 の文仮説を生成する．その後，識別器によるリスコアリングを行い結果とする．

言語モデルの語彙サイズは 60k 単語とし，Good Turing によるスムージングを行った．cut off は bigram に対して 1，trigram に対して 2 とした．評価データは 45 のニュース番組から構成されているため，言語モデルは各ニュースごとに直近のニュース原稿を用いて適応化した [14]．

音響モデルは，118 時間の放送ニュース音声から RASTA [15] を行ったうえで，39 次元 (12 次元 MFCC+対数パワーおよび 1 次・2 次の回帰係数) のパラメータを求めて triphone HMM を学習した．

6.2 識別実験結果

表 4 に識別結果を示す．比較は識別誤り率 (CER)，誤棄却率 (FRR)，誤受理率 (FAR) で行った．図 1 の ROC 曲線を拡大すると，ME および SVM の結果は，それぞれ図 2 に示すように単語事後確率の描く ROC 曲線の下部にプロットされる．ME は識別誤り率が単語事後確率に対し 5.8%ほど改善しているが，誤棄却率が小さくなったかわりに誤受理率が大きくなっており，単語仮説を正解に識別する傾向がある．この原因としては，ラベルバイアスの問題 [16] が考えられる．評価データの 90%以上は正解ラベルであり，ラベルの分布に偏りがある．したがって，前単語が正解ラベルの場合，後続の単語仮説が正解の方に偏って判定されるケースが増えたものと考えられる．

識別誤り率をみると，SVM に対して ME の結果が上回っている．この理由は，ME モデルでは連続値を持つ信頼度が素性関数の形でパターン化されるため，結果としてスムージングと同等の効果を果たしたのではないかと考えられる．また，同量の学習データを用いた場合，ME に比べて SVM の方が学習に

表4 識別結果

	CER(%)	FRR(%)	FAR(%)
単語事後確率	5.2	2.1	49.1
音響安定度	6.0	1.5	69.6
単語仮説密度	6.6	0.0	99.7
ME	4.9	1.3	54.7
SVM	5.2	0.9	64.9

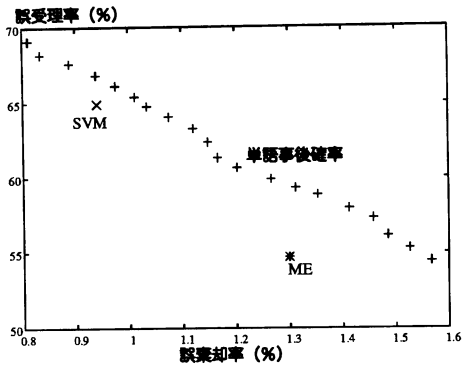

図 2 ME と SVM 識別器の ROC グラフ上でのプロット

時間がかかるため(注2)，識別精度に関わるパラメータ (多項式カーネルの次数，soft margin パラメータ) の決定が困難になりやすい．提案手法は，多数の閾値をパラメータとして決める必要がある一方，3.2 節で述べた各 Step での 1 回あたりのモデルの作成が短時間ですむため，効率的にモデルが作成可能である．

6.3 リスコアリング実験結果

リスコアリング結果を表 5 に示す．表中，"可変"とあるのは，可変重み (式 (14)) を用いたときの結果である．ベースライン (リスコアリング時のペナルティなしの場合) と比較すると，補正を行った場合は，いずれも単語誤り率を改善している．(ME (可変) でベースラインに対して 4.7%改善した)．単語事後確率を用いた場合と比較すると，2.4%の改善となった．

重みを可変にした効果を詳細に見るために，評価データをベースラインの単語誤り率に応じて Set1 (単語誤り率 20%以上の 3,223 単語)，Set2 (同 30%以上の 1,871 単語)，Set3 (同 40%以上の 1,081 単語) に分けて評価した．ベースラインに対する改善率を見ると，ME (可変) のケースで Set1 から順に 4.3%，5.0%，4.9%であった．単語誤り率の大小によらず，ほぼ同程度の改善効果であった．

ME による手法で，重みを固定とした場合と可変とした場合を改善率で比較すると，Set1 から順に 1.7%，1.8%，0.8%であった．重みを可変にした場合，単語誤り率の改善がほぼ評価データ全体に及んでいることが分かる．

6.4 環境ごとの評価

最後に評価データの環境ごとのリスコアリング結果について

(注2)：SVM では，超平面を決定するデータを探索するため，学習データが大きくなるほど超平面となるサポートの候補が増加し，学習に時間がかかるようになる．

表 5　リスコアリング結果：WER(%)

	全体	Set1	Set2	Set3
ベースライン	8.6	36.8	46.1	54.8
単語事後確率	8.4	35.9	45.1	53.4
単語事後確率（可変）	8.4	36.0	45.1	53.3
ME	8.4	35.8	44.6	52.5
ME（可変）	**8.2**	**35.2**	**43.8**	**52.1**
SVM	8.4	36.1	45.2	53.8
SVM（可変）	8.4	35.9	44.7	52.9

表 6　リスコアリング結果 (雑音環境)：WER(%)

	全体	Set1	Set2	Set3
ベースライン	7.0	35.2	45.5	56.2
単語事後確率	6.8	34.1	44.3	55.4
単語事後確率（可変）	6.8	34.2	44.3	54.2
ME	**6.7**	33.6	43.3	53.8
ME（可変）	**6.7**	**33.3**	**42.3**	**51.8**
SVM	6.8	34.1	44.2	55.3
SVM（可変）	**6.7**	33.8	44.5	53.2

表 7　リスコアリング結果 (対談環境)：WER(%)

	全体	Set1	Set2	Set3
ベースライン	19.9	39.4	48.2	54.9
単語事後確率	19.5	38.7	47.1	55.4
単語事後確率（可変）	19.5	38.8	47.0	54.1
ME	19.4	38.6	46.6	53.6
ME（可変）	**19.2**	**38.0**	**45.9**	**53.3**
SVM	19.5	38.7	46.8	54.0
SVM（可変）	19.7	38.7	46.7	53.6

示す（表 6，7）．評価データ全体のうち，雑音環境は 870 文 (20,858 単語)，対談環境は 286 文 (5,617 単語) であった（両方の環境に含まれるものは評価データ中に 163 文）．対談環境の方が，単語誤り率が大きいが，これは主に言語的な難しさ (雑音環境のパープレキシティが 14.1 であったのに対し，対談環境では 48.7) が原因と考えられる．

ME（可変）は，雑音環境，対談環境のいずれの環境に対しても，単語誤り率が最小となった．SVM の結果と比べると，統合された尺度がどちらの環境に対しても有効なペナルティを与えていることが予想される．ベースラインに対する ME（可変）の改善率は，雑音環境で 4.3%，対談環境で 3.5%と，雑音環境の方が改善の効果がやや高かった．雑音環境では，ME（可変）と SVM（可変）の単語誤り率がともに最小だが，単語事後確率と比べると差はわずかであった．対談環境をみると，ME（可変）での単語誤り率が単語事後確率や SVM よりも小さくなっており，この結果が表 5 の評価データ全体の単語誤り率の改善に貢献している．

7. おわりに

複数の信頼度尺度を組み合わせた音声認識手法について報告した．最大エントロピー法に基づく統合手法を提案し，従来法に基づくリスコアリングに比べ 4.7%の単語誤り率の削減，単語事後確率に基づく手法に比べ 2.4%の削減となった．提案手法を SVM に基づく統合手法と比較し，識別誤り率および単語誤り率のいずれも ME に基づく提案手法の方が良い結果となった．

今後は，提案手法について新たな信頼度尺度を追加した場合の識別誤りや認識誤りの改善を検討していく．

文　献

[1] T. Imai, A. Kobayashi, S. Sato, S. Homma, K. Onoe, and T.S. Kobayakawa, "Speech recognition for subtitling Japanese live broadcasts," The 18th International Congress on Acoustics, pp.I-165–168, April 2004.

[2] Gökhan Tür, M. Rahim, and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," Eurospeech, Geneva, Switzerland, 2003.

[3] M. Nakano, "Using untranscribed user utterances for imporving language models based on confidence scoring," Eurospeech, Geneva, Switzerland, 2003.

[4] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence meaure for large vocabulary continuous speech recognition," IEEE Transactions Speech and Audio Processing, vol.9, pp.288–298, March 2001.

[5] T. Zeppenfeld, M. Finke, and K. Ries, "Recognition of conversational telephone speech using the Janus speech engine," IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.1815–1818, 1997.

[6] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," Eurospeech, Rhodes, Greece, pp.827–830, 1997.

[7] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," The Annals of Mathematical Statistics, pp.1470–1480, 1972.

[8] T. J.Hazen, S. Seneff, and J. Polifromi, "Recognition confidence scoring and its use in speech understanding systems," January 2002.

[9] P. J.Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," Eurospeech, Aalborg, Denmark, 2001.

[10] T. Joachims, Learning to classify text using support vector machines, Kluwer Academic Publishers, 2002.

[11] T. Joachims, "Introduction to support vector learning," in Advances in Kernel Methods, ed. B. Scho-lkopf, C.J. Burges, and A.J. Smola, The MIT Press, 1999.

[12] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," Advances in Large Margin Classifiers, ed. A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, pp.61–74, 2000.

[13] T. Joachims, "Making large-scale SVM learning practical," in Advances in Kernel Methods, ed. B. Scho-lkopf, C.J. Burges, and A.J. Smola, The MIT Press, 1999.

[14] A. Kobayashi, K. Onoe, T. Imai, and A. Ando, "Time dependent language model for broadcast news transcription and its post-correction," Int. Conf. Spoken Lanugage Processing, pp.2435–2438, 1998.

[15] H. Helmansky and N. Morgan, "RASTA processing of speech," IEEE Transactions Speech and Audio Processing, vol.2, pp.587–589, October 1994.

[16] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. 18th International Conf. on Machine Learning, pp.282–289, Morgan Kaufmann, 2001.