Vol.2010-SLP-84 No.25
2010/12/21

異なる認識単位の認識器から得られた信頼度を素性に用いた音声認識

田中智之[†]　西田昌史[†]　山本誠一[†]

本稿では，異なる認識単位の認識器から得られた信頼度を素性に用いた音声認識手法を提案する．本手法は，複数の認識器から得られた認識結果に対して Support Vector Machine（SVM）により認識結果の正否を判別し，複数の認識結果のうち正しい認識結果を選択する．さらに，SVM により複数の認識結果の正否を判別することにより，音声認識精度を改善するだけでなく，正しい認識結果が含まれていない場合に認識結果を棄却することができる．カーナビの目的地設定を想定したタスクにおいて，単一の認識器および ROVER 法との比較実験を行ったところ，提案手法では従来手法に比べてより高い認識結果の判別精度および音声認識精度が得られ，本手法の有効性を示した．

Automatic Speech Recognition with Confidence Measures Obtained by Multiple Recognizers of Various Recognition Units

Tomoyuki Tanaka[†]　Masafumi Nishida[†]
and Seiichi Yamamoto[†]

We propose a novel speech recognition method that improves word accuracy by using confidence measures obtained from multiple recognizers of various recognition units. We use Support Vector Machine (SVM) to select a single recognition result from multiple recognition results. Our method can also identify when the correct recognition is not included in the multiple recognition results. Experimental results show that our method gives higher word accuracy and classification accuracy in comparison with a single recognizer and ROVER.

1. はじめに

現在の音声認識技術において，誤認識は避けられない．そのため，得られた認識結果が正しいか否かを判別することが音声認識を行う上で重要である．認識結果の正否を判別する尺度として，信頼度が用いられた研究が盛んに行われている．

信頼度に関する研究として，単一の認識器から得られる単語ラティスやN-bestの情報を用いた手法がある[1) 2)]．さらに，近年では複数の認識器を用いて認識結果を選択する手法が考案されてきた．複数の認識器を用いる研究として，複数の認識結果を多数決で判別するROVER法（Recognizer Output Voting Error Reduction: ROVER）がある[3) 4)]．そして，ROVER法を音声理解におけるコンセプト誤り率（Concept Error Rate: CER）の改善に応用した研究がある[5)]．また，ROVER法に基づき，複数の異なる音響モデルや言語モデル，デコーダを組み合わせて認識結果の共通部分を用いる手法[6)]，複数の言語モデルと言語理解モデルを用い，得られた信頼度に基づきロジスティック回帰により認識結果を選択する手法がある[7)]．しかし，これらの手法は，複数の認識結果のうち，正しい認識結果が含まれていない場合も正しい認識結果として 1 つを選択している．そのため，従来研究では，誤認識の考慮がなされていない．

我々は以前，カーナビの目的地設定を想定したタスクにおいて，一般的に用いられる認識単位である形態素単位や単語単位に加えて，認識対象である目的地名を手動で“名称”と“属性”に分割する“部分単語単位”を新たに導入した．そして，それら認識単位の異なる複数の認識器から得られた信頼度に基づき，Support Vector Machine（SVM）によって得られた複数の認識結果の正否を判別し，複数の認識結果から正しい認識結果を選択する手法を提案した[8)]．本稿では，語彙数を増加させ，我々が提案した認識単位である“部分単語単位”を形態素間の連接尤度に基づき自動的に生成する手法について検討し，さらにROVER法との比較を行った．本手法により，単一の認識器やROVER法に基づく手法と比較して音声認識精度が改善するだけでなく，複数の認識結果に正しい認識結果が含まれていない場合もその旨を判別することができ，認識結果を棄却することができる．また，新しい手法を導入した部分単語単位について，その有効性を示す．

以降，2 章で認識器に用いた認識単位と言語モデルについて述べ，3 章で SVM により複数の認識結果の正否を判別して正しい認識結果を選択する手法について述べる．そして，4 章で評価実験の内容とその結果について述べ，5 章で本稿をまとめる．

[†] 同志社大学大学院 工学研究科 情報工学専攻
Department of Information Engineering, Graduate School of Engineering, Doshisha University

Vol.2010-SLP-84 No.25
2010/12/21

2. 認識単位と言語モデル

本研究において，タスクはカーナビの目的地設定を想定した“目的地は京都信用金庫です”といった目的地名入力発話であり，認識対象は“京都信用金庫”といった目的地名である．音声認識において，一般的に認識単位は，認識対象を形態素解析によって分割した認識単位（形態素単位）や認識対象をそのまま用いた認識単位（単語単位）を適用する場合が多い．しかし，前者では認識単位が短くなり音響的に類似した認識候補が増加し，後者では語彙数が増加するため音声認識精度が低下する．そのため，音声認識に最適な認識単位について検討する必要がある．そこで，本研究では，全ての認識対象における形態素間の連接尤度を考慮し，連接尤度が高ければ形態素間を結合させることにより，形態素単位より長く，かつ単語単位より短い認識単位である“部分単語単位”を生成した．

図 1 に，形態素間の連接尤度に基づく部分単語単位の生成法を示す．例として，“京都信用金庫”という認識対象を挙げる．まず，認識対象を形態素解析によって，“京都”，“信用”，“金庫”という形態素単位に分割できる．形態素間の連接尤度は，以下の式で表される．

$$\log P(w_i|w_{i-1}) \geq \theta \qquad (1)$$

ここで，w は形態素であり，θ は閾値である．式 (1) より，形態素 w 間の連接尤度 $\log P(w_i|w_{i-1})$ が閾値 θ を超えた場合，形態素間を結合させる．閾値 θ は，実験的に求め，今回の実験においては $\theta = -1.4$ とした．

言語モデルには，統計的モデルの N-gram と有限状態文法モデルの Finite State Automaton（FSA）を用いた．N-gram は，1 つの目的地名あたり 72 種類の言い回しを含んだ 10 万目的地名からなるテキストデータによって学習させ，形態素単位および部分単語単位において，それぞれ前向き bi-gram および後ろ向き tri-gram を作成した．FSA は，72 種類の言い回しに対応するような文法規則を人手で作成した．図 2 に，言い回しを含んだ目的地名の発話例を示す．以上の認識単位と言語モデルから以下の 3 つを認識器として用意する．そして，その 3 つの認識器のうち 2 つの認識器を併用し，それぞれ認識器 A，認識器 B とする．

(1) **N-gram（形態素単位）**
(2) **N-gram（部分単語単位）**
(3) **FSA（単語単位）**

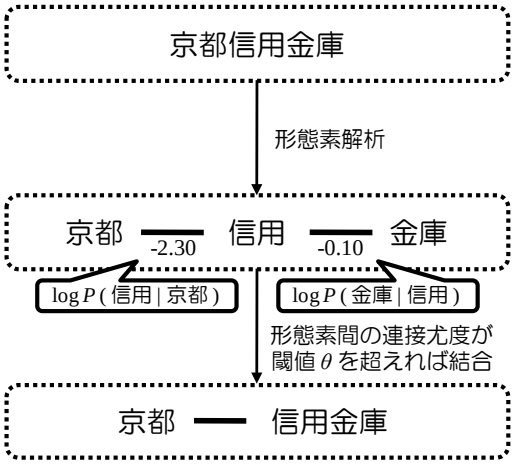

図 1　形態素間の連接尤度に基づく部分単語単位の生成法
Figure 1　Scheme of making “partial word unit” based on connection likelihood between morphemes.

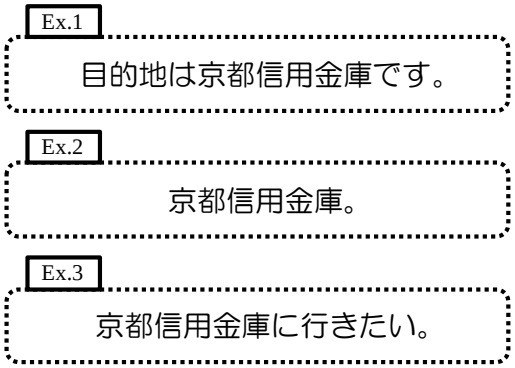

図 2　言い回しを含んだ発話例
Figure 2　Example of utterance included expressions.

Vol.2010-SLP-84 No.25
2010/12/21

3. 信頼度を素性としたSVMによる認識結果の判別

表 1 に，認識結果の正否を判別するために用いた信頼度尺度を示す．2 つの認識器を併用した認識結果の正否判別では，2 つの認識結果が次の 4 パターンのいずれであるか判別する必要がある．

(1) **両方認識正解**
(2) **認識器 A が認識正解かつ認識器 B が誤認識**
(3) **認識器 A が誤認識かつ認識器 B が認識正解**
(4) **両方誤認識**

本稿では，表 1 に示した信頼度尺度を素性とし，Support Vector Machine（SVM）により 2 つの認識結果を上記の 4 パターンに判別する．そして，SVM による判別結果から正しい認識結果が含まれていれば 2 つの認識結果の正しい認識結果を選択し，そうでなければ認識結果を棄却することにより，音声認識精度を改善でき，かつ 2 つの認識結果が両方誤認識であってもその旨を判別することができる．

図 3 に，本手法の流れを示す．まず，入力音声に対して認識器 A と認識器 B で音声認識を行い，2 つの認識結果を得る．そして，それらの認識結果から表 1 に示した単一の認識器および認識器の併用から得られる信頼度尺度を抽出し，それらを素性として SVM により認識結果の判別を行う．SVM のカーネル関数には Radial Basis Function（RBF）を用い，ハイパーパラメータは実験的に求めた．

4. 評価実験

4.1 実験条件

カーナビの目的地設定を想定したタスクにおいて，認識単位の異なる 2 つの認識器を併用した音声認識, および SVM による 2 つの認識結果の正否判別と選択を行った．

特徴量には，サンプリング周波数 16 kHz，フレーム長 25 ms，シフト幅 10 msによって抽出された 12 次元Mel Frequency Cepstral Coefficients（MFCCs），パワー，およびそれらの 1 次差分から構成される 25 次元のパラメータを用いた．音響モデルには，状態数 3000，混合分布数 64，性別非依存のPhonetic Tied-Mixture（PTM）triphoneを用いた．言語モデルには，N-gramとして前向きbi-gramと後ろ向きtri-gram，およびFSAを用いた．デコーダには，Julius（ver.4.1.5）を用いた[9].

単語辞書には, 10 万目的地名を登録し, 語彙数は, N-gram（形態素単位）では 3,328, N-gram（部分単語単位）では 8,994，FSA（単語単位）では 100,000 であった．図 4 に，単語辞書に登録した 10 万目的地名におけるモーラ数ごとの頻度数を示している．

表 1　SVM の素性に用いた信頼度尺度
Table 1　Confidence measures for SVM.

CMscore	認識単位ごとの単語事後確率
N-best	認識候補 1 位の目的地名が 10 位までに出現する割合
音響尤度	各認識器の音響スコア
言語尤度	各認識器（N-gram）の言語スコア
一致度	2 つの認識結果が一致したか否か
音素数差	2 つの認識結果の音素数の差

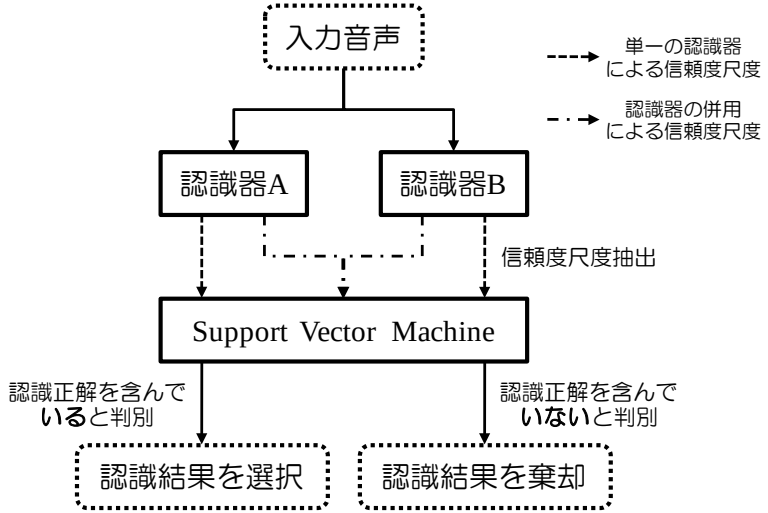

図 3　提案手法の流れ
Figure 3　Flow of proposed method.

音声データには，男性 10 名による目的地名 100 種，言い回し 4 種の計 4000 発話を用いた．そして，比較対象である単一の認識器による手法，および提案手法である 2 つの認識結果から得られた信頼度尺度に対して 10-fold cross-validation により認識結果の正否を判別した．

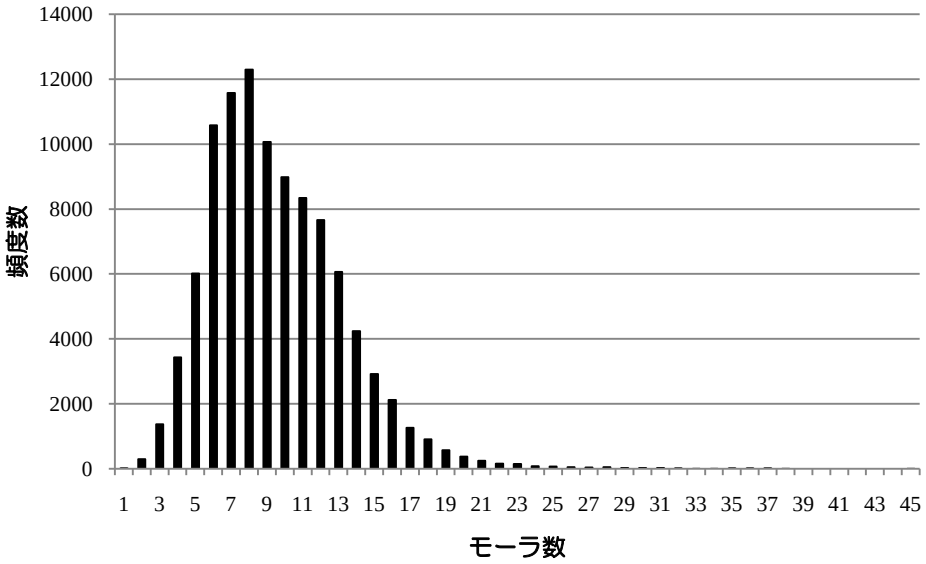

図 4　10 万目的地名におけるモーラ数ごとの頻度分布
Figure 4　Frequency distribution of each number of mora in 100k destination names.

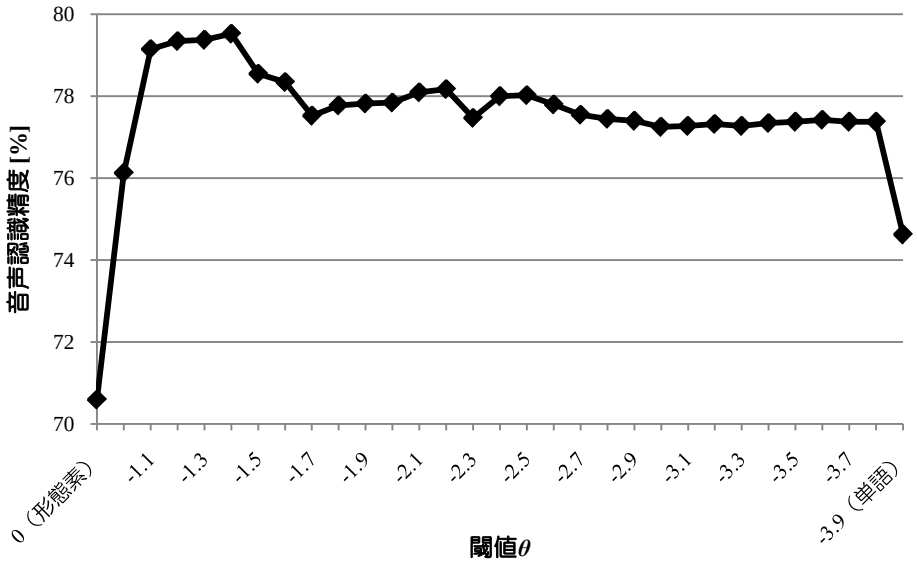

図 5　部分単語単位における閾値ごとの音声認識精度
Figure 5　Word accuracy with partial word unit for each threshold.

4.2　実験結果と考察

4.2.1 認識単位の違いによる音声認識結果

前述の実験条件下において，音声認識精度は，N-gram（形態素単位）では 70.6%, N-gram（部分単語単位）では 79.5%，FSA（単語単位）では 84.1%であった．図 5 に，部分単語単位における閾値ごとの音声認識精度を示す．図 5 では、形態素間の連接尤度に対する閾値を-1.0 から-3.8 まで変動させた際の音声認識精度を算出した．なお，閾値 0 では形態素単位に，閾値-3.9 では単語単位に一致する．今回の実験では，最も音声認識精度が高かった閾値-1.4 を採用している. 図 5 より, N-gram（部分単語単位）では，N-gram（形態素単位）と N-gram（単語単位）と比較して高い音声認識精度が得られ，形態素間の連接尤度に基づく認識単位による音声認識が有効であることが分かった．

4.2.2 異なる認識単位の認識器ごとの判別精度および音声認識結果

表 2 に単一の認識器による認識正解の判別精度を, 表 3 に ROVER 法による認識正解の判別精度を，表 4 に提案手法による認識正解の判別精度を示す．また，表 5 表 2 に単一の認識器による誤認識の判別精度を，表 6 に ROVER 法による誤認識の判別精度を，表 7 に提案手法による誤認識の判別精度を示す．ここで，表 4 および表 7 は，認識器 A に対して認識器 B を組み合わせた際の認識器 A の判別精度を示している．本手法において，認識正解および誤認識の再現率と適合率は以下の式で表される．

$$\text{再現率}=\frac{\text{認識正解（誤認識）と正しく判別した数}}{\text{認識正解（誤認識）した数}}\times 100 \quad (2)$$

$$\text{適合率}=\frac{\text{認識正解（誤認識）と正しく判別した数}}{\text{認識正解（誤認識）と判別した数}}\times 100 \quad (3)$$

単一の認識器による再現率と適合率は，CMscore，N-best，音響尤度，および言語尤度を素性として，SVM により認識結果が認識正解，もしくは誤認識であるかを正しく判別した割合で算出した．また，ROVER 法では，認識単位の異なる 3 つの認識器を全て併用して得られる認識結果から多数決で 1 つ選択し，その結果を正しい認識結果として判別している．ここで，3 つの認識結果が全て異なった場合，正しい認識結果が含まれていないと判別した．

ROVER 法では，単一の認識器による手法と比較して認識正解に対する判別精度が高かった．しかし，誤認識に対する判別精度が低いことより，正しい認識結果が含まれていない場合における判別が困難であることが分かった．提案手法では，他の手法と比較して認識正解，および誤認識に対する判別において高い精度を示しており，異なる認識単位の認識器を併用した SVM による認識結果の正否判別手法，および SVM の素性に用いた信頼度尺度が有効であったことが分かった．特に，認識正解の判別では，N-gram（部分単語単位）と FSA（単語単位）との併用が，誤認識の判別では，N-gram（形態素単位）と FSA（単語単位）との併用が高い精度を示しており，SVM による認識結果の正否判別では N-gram ベースの認識器と FSA ベースの認識器の併用が有効であることが分かった．

表 2　単一の認識器による認識正解の判別精度
Table 2　Classification accuracy of correct recognition result of each single recognizer.

認識器	再現率	適合率	F 値
N-gram（形態素単位）	92.5	86.1	89.2
N-gram（部分単語単位）	95.9	88.2	91.9
FSA（単語単位）	96.8	89.7	93.2

表 3　ROVER 法による認識正解の判別精度
Table 3　Classification accuracy of correct recognition result of ROVER.

認識器	再現率	適合率	F 値
N-gram（形態素単位） N-gram（部分単語単位） FSA（単語単位）	95.0	91.9	93.4

表 4　提案手法による認識正解の判別精度
Table 4　Classification accuracy of correct recognition result of proposed method.

認識器 A	認識器 B	再現率	適合率	F 値
N-gram（形態素単位）	N-gram（部分単語単位）	96.7	89.8	93.1
	FSA（単語単位）	97.5	95.0	96.3
N-gram（部分単語単位）	FSA（単語単位）	98.3	95.5	96.9
	N-gram（形態素単位）	97.1	89.8	93.3
FSA（単語単位）	N-gram（形態素単位）	98.6	94.2	96.4
	N-gram（部分単語単位）	98.6	94.7	96.6

表 5　単一の認識器による誤認識の判別精度
Table 5　Classification accuracy of false recognition result of each single recognizer.

認識器	再現率	適合率	F 値
N-gram（形態素単位）	64.1	78.1	70.4
N-gram（部分単語単位）	50.2	75.8	60.4
FSA（単語単位）	41.7	71.5	52.7

表 6　ROVER 法による誤認識の判別精度
Table 6　Classification accuracy of false recognition result of ROVER.

認識器	再現率	適合率	F 値
N-gram（形態素単位） N-gram（部分単語単位） FSA（単語単位）	33.8	46.5	39.2

表 7　提案手法による誤認識の判別精度
Table 7　Classification accuracy of false recognition result of proposed method.

認識器 A	認識器 B	再現率	適合率	F 値
N-gram（形態素単位）	N-gram（部分単語単位）	73.7	90.3	81.2
	FSA（単語単位）	87.8	93.6	90.6
N-gram（部分単語単位）	FSA（単語単位）	81.9	92.6	86.9
	N-gram（形態素単位）	57.1	83.7	67.9
FSA（単語単位）	N-gram（形態素単位）	68.2	90.4	77.7
	N-gram（部分単語単位）	70.8	90.8	79.6

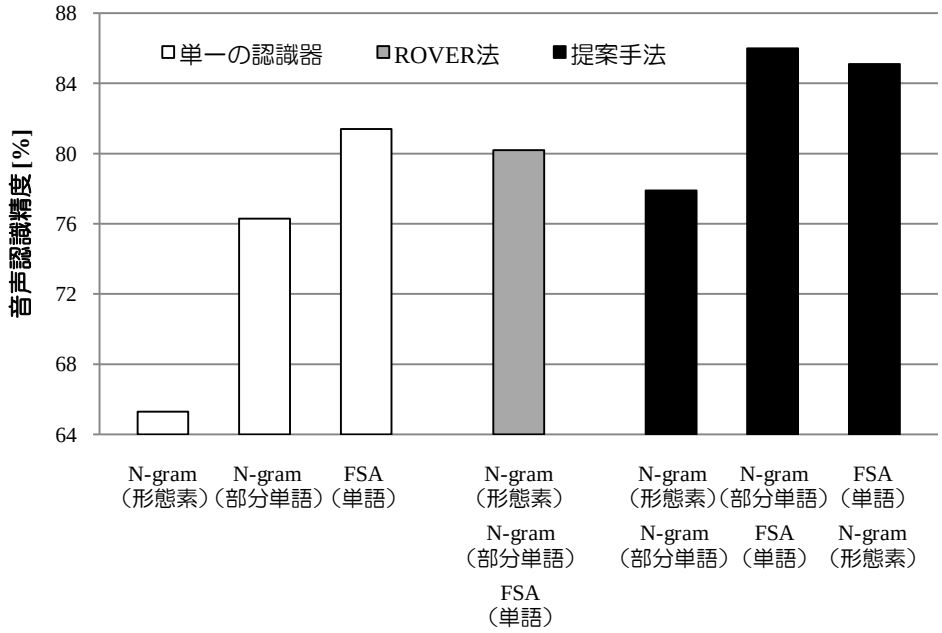

図 6　各手法による認識結果の判別に基づく音声認識精度
Figure 6　Word accuracy based on classification of recognition results by each method.

SVM による認識結果の正否判別，および ROVER 法に基づき，認識結果を選択した音声認識精度を図 6 に示す．提案手法では，単一の認識器，および ROVER 法による音声認識精度より高い結果が得られ，本手法が有効であることが分かった．提案手法の中でも，N-gram ベースの認識器と FSA ベースの認識器の併用による手法が音声認識精度の改善に大きく貢献している．特に，N-gram（部分単語単位）と FSA（単語単位）の併用が有効であった．

以上の結果より，認識正解と誤認識に対する認識結果の判別精度，および認識結果の選択に基づく音声認識精度から，N-gram（部分単語単位）と FSA（単語単位）の併用による音声認識が有効であると考えられる．

5. おわりに

本稿では，カーナビの目的地設定を想定したタスクにおいて，認識単位の異なる複数の認識器から得られた信頼度を素性とし, SVM により複数の認識結果の正否を判別して正しい認識結果を選択する手法を提案した．また，新たな認識単位として，形態素間の連接尤度に基づく部分単語単位を導入し，形態素単位よりも高い音声認識精度が得られ，部分単語単位の有効性を示した．本手法において，従来手法である単一の認識器や ROVER 法による手法と比較して，認識結果の判別精度において高い結果が得られた. そして, SVM による判別結果に基づき認識結果を選択する本手法において，従来手法と比較して音声認識精度が改善し，その有効性を示した．

今後の課題として，SVM によって認識結果が棄却された場合，認識候補の第 2 位以降の認識結果を正否判別して認識結果を選択する手法について検討する．また，新たな信頼度尺度についても検討する．

謝辞　本研究は，科研費若手研究（B）（21700184）の助成を受けたものである．

参考文献

1) Kemp, T. and Schaaf, T.: Estimating Confidence using Word Lattices, *Proc. Eurospeech*, pp.827-830 (1997).
2) Wessel, F., Macherey, K. and Ney, H.: A Comparison of Word Graph and N-best List Based Confidence Measures, *Proc. Eurospeech*, pp.315-318 (1999).
3) Fiscus, J. G.: A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), *Proc. ASRU*, pp.347-354 (1997).
4) Schwenk, H. and Gauvain, J. -L.: Combining Multiple Speech Recognizers using Voting and Language Model Information, *Proc. ICSLP*, pp.915-918 (2000).
5) Hahn, S., Lehnen, P. and Ney, H.: System Combination for Spoken Language Understanding, *Proc. Interspeech*, pp.236-239 (2008).
6) 宇津呂武仁，西崎博光，小玉康広，中川聖一: 複数の大語彙連続音声認識モデルの出力の共通部分を用いた高信頼度部分の推定，電子情報通信学会論文誌, Vol.J86-D-II, No.7, pp.974-987 (2003).
7) 勝丸真樹，中野幹生，駒谷和範，船越孝太郎，辻野広司，尾形哲也，奥乃博: 複数の言語モデルと言語理解モデルによる音声理解の高精度化，電子情報通信学会論文誌, Vol.J93-D, No.6, pp.879-888 (2010).
8) 田中智之，西田昌史，山本誠一: 認識単位の異なる認識器から得られた信頼度に基づく音声認識，日本音響学会秋季研究発表会講演論文集, 2-9-4 (2010).
9) Lee, A. and Kawahara, T.: Recent Development of Open-Source Speech Recognition Engine Julius, *Proc. APSIPA*, pp.131-137 (2009).