情報処理学会論文誌 Vol.54 No.2 443–451 (Feb. 2013)

音声情報案内システムにおけるBag-of-Wordsを用いた無効入力の棄却

真嶋 温佳[1,a)] 藤田 洋子[1] トーレス ラファエル[1,b)] 川波 弘道[1,c)] 原 直[2,d)]
松井 知子[3,e)] 猿渡 洋[1,f)] 鹿野 清宏[1,g)]

受付日 2012年5月31日, 採録日 2012年11月2日

概要：実環境における音声認識を用いた情報案内システムでは，雑音等の非音声やユーザ同士の背景会話など，システムへの入力として不適切な入力が存在する．これらの入力はシステムの誤作動・誤認識の原因となるので，システムへの入力として適切な入力（有効入力）と不適切な入力（無効入力）の識別を行い，無効入力を棄却することにより，無効入力に対する応答処理を行わないことが重要である．従来，有効入力と無効入力との識別には，メル周波数ケプストラム係数などの音響的特徴量によるGMM（Gaussian Mixture Model）が用いられる．しかし，入力データの音声認識結果から得られる言語的な情報を使うことにより，システムのタスクを考えたうえで有効入力と無効入力の識別が可能になると考えられる．そこで本論文では，音響特徴量にBag-of-Words（BOW）を言語的特徴量として併用した無効入力の識別を検討した．識別手法としては，サポートベクタマシン（SVM）および最大エントロピー法を用いた．実験には実環境音声情報案内システム「たけまるくん」の入力データを用いた．SVMによる識別結果では，GMMによる音響尤度のみを用いた場合に比べて，BOWを用いた場合，F尺度を82.19%から85.41%に改善することができた．さらに，GMMによる音響尤度，発話時間，SNRを組み合わせた特徴量にBOWを追加することで，F尺度を86.58%まで改善することができた．詳細な分析の結果，BOWは特に無効入力の誤受理を減らす効果があることが示された．

キーワード：音声情報案内システム，無効入力棄却，Bag-of-Words

Invalid Input Rejection Using Bag-of-Words for Speech-oriented Guidance System

Haruka Majima[1,a)] Yoko Fujita[1] Rafael Torres[1,b)]
Hiromichi Kawanami[1,c)] Sunao Hara[2,d)] Tomoko Matsui[3,e)]
Hiroshi Saruwatari[1,f)] Kiyohiro Shikano[1,g)]

Received: May 31, 2012, Accepted: November 2, 2012

Abstract: On a real environment speech-oriented information guidance system, a valid and invalid input discrimination is important as invalid inputs such as noise, laugh, cough and utterances between users lead to unpredictable system responses. Generally, acoustic features such as MFCC (Mel-Frequency Cepstral Coefficient) are used for discrimination. Comparing acoustic likelihoods of GMMs (Gaussian Mixture Models) from speech data and noise data is one of the typical methods. In addition to that, using linguistic features, such as speech recognition result, is considered to improve discrimination accuracy as it reflects the task-domain of invalid inputs and meaningless recognition results from noise inputs. In this paper, we introduce Bag-of-Words (BOW) as a feature to discriminate between valid and invalid inputs. Support Vector Machine (SVM) and Maximum Entropy method (ME) are also employed to realize robust classification. We experimented the methods using real environment data obtained from the guidance system "Takemaru-kun." By applying BOW on SVM, the F-measure is improved to 85.09%, from 82.19% when using GMMs. In addition, experiments using features combining BOW with acoustic likelihoods from GMMs, Duration and SNR were conducted, improving the F-measure to 86.58%.

Keywords: spoken dialogue system, invalid input rejection, Bag-of-Words

1. はじめに

音声認識システムを応用した音声情報案内システムは主に，音声区間検出，音声認識，応答生成により構成され，これらの処理は入力音声に対して順次処理されていく．実環境における音声情報案内システムへの入力には，システムとして適切な発話（有効入力）以外の様々な入力が存在し，このような入力はシステムの誤認識や誤作動の原因となる．特に Push-to-talk などの機構を持たず，つねに音声入力を受け付けるような音声情報案内システムでは，システムが応答すべきではない入力音声は非常に多い [1]．特に，応答生成処理はシステムが対応するドメインによって肥大化する可能性があり，すべての入力に対して応答処理をすることはシステム負荷の観点から避けるべきである．したがって，システムにとって不適切な入力（無効入力）はできる限り応答生成処理に送られる前に棄却することが望ましい．

これまでにも，音声認識結果に対する統計的仮説検定による発話照合手法が研究されている [2]．この手法では，ある音声認識結果を含む音声発話区間が入力された場合にその発話を受理すべきであるという帰無仮説と棄却すべきであるという対立仮説を立てる．そして，それぞれの仮説を条件とした入力音声の条件付き確率分布を事前に学習することができれば，それらの尤度比を検定統計量とした仮説検定を行うことができる．つまり，未知の入力発話に対して帰無仮説が棄却された場合，その入力発話は高い確率で受理すべきではないと判定される．この手法は，音声の確率分布を考えるということで一種の生成的なアプローチと考えられるが，より直接的な識別的なアプローチも考えられる [3]．

また，Lane らは有効入力を対象とした音声対話システムへのドメイン外発話の検出手法を提案している [4]．彼らは，N-best 認識仮説から得られた Bag-of-Words（BOW）を特徴量として発話のトピック推定を行い，そのトピック推定結果を用いて有効入力を対象としたドメイン外発話の検出を行っている．ただし，入力はすべて有効入力であることを想定しており，本研究が対象とする大量の無効入力を含むような対話システムは考慮されていない．これらの手法は，特徴量として音声認識結果を利用しており，認識結果の後処理として位置づけられる．

一方で，音声認識の前処理として音響的特徴量を用いた手法も検討されており，たとえば，メル周波数ケプストラム係数（Mel-Frequency Cepstral Coefficient: MFCC）特徴量に対する混合ガウス分布モデル（Gaussian Mixture Model: GMM）から算出される音響尤度による最尤判別に基づく音声と雑音の識別 [5], [6] などがあげられる．また，音声区間検出においても，音声認識結果の言語的制約を用いる手法が提案されており [7]，これは言語的な情報も音声と雑音の識別に有効であることを示唆している．

以上をふまえて，本研究では音声認識の後処理で用いられる音声認識結果に基づく言語的特徴量を，音声認識の前処理で用いられる音響的特徴量と併用した，音声情報案内システムへの無効入力の検出手法を提案する．本論文では音声認識結果から得られた BOW を言語的特徴量として扱う．BOW から得られる特徴量が有効入力のドメイン外発話検出において有効であることは示されている [4] が，本論文では音声認識結果に含まれる有効入力に出現しやすい単語や無効入力の認識結果の傾向を利用することができれば，システムのタスクを考えたうえで有効入力と無効入力を識別することも可能になると考える．識別手法として，サポートベクタマシン（Support Vector Machine: SVM）[8] および最大エントロピー法（Maximum Entropy method: ME）[9], [10] を用いる．

本論文の構成は以下のとおりである．2 章では実験に用いた音声情報案内システム「たけまるくん」と，この「たけまるくん」によって収集された音声コーパスについて述べる．3 章では本論文で扱う特徴量および SVM や ME を用いた無効入力の識別手法を述べる．4 章では本論文の提案する特徴量と識別手法を用いた識別実験の結果を示し，5 章で結論を述べる．

2. 音声情報案内システム「たけまるくん」

2.1 システムの概要

「たけまるくん」[1] は，生駒市北コミュニティセンター内に設置された音声情報案内システムである．2002 年 11 月より運用を開始し，現在までの約 10 年にわたり運用を継続している．「たけまるくん」に対してユーザが発話によって質問すると，合成音声とアニメーションを用いて，エージェントが応答する．「たけまるくん」の主な応答内容は，コミュニティセンター内の施設案内，周辺の観光案内，エージェント（たけまるくん）自身に対する質問への応答，現在時刻・天気・ニュースなどである．これらとは別に，ユーザが発話した単語を Web で検索する「Web 検索モード」も利用できる．「たけまるくん」は，音声対話システムの実環境対話データ収集および対話システムのフィー

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[2] 岡山大学
Okayama University, Okayama 700–8530, Japan
[3] 統計数理研究所
The Institute of Statistical Mathematics, Tachikawa, Tokyo 190–0014, Japan
a) haruka-m@is.naist.jp
b) rafael-t@is.naist.jp
c) kawanami@is.naist.jp
d) hara@cs.okayama-u.ac.jp
e) tmatsui@ism.ac.jp
f) sawatari@is.naist.jp
g) shikano@is.naist.jp

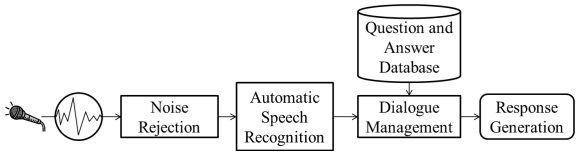

図 1 「たけまるくん」[1] における応答処理の流れ
Fig. 1 Processing flow in “Takemaru-kun”.

ルドテストを兼ねて運用されている．本システムの構成を**図 1** に示す．「たけまるくん」は，音声のみを入力インタフェースとしている．そのため，ユーザはシステムに対して話しかけるという単純な行動のみでシステムから情報を得ることができ，自由度の高いシステムとなっている．しかし，無効入力によってシステムが誤作動すると，ユーザにとって非常に扱いにくいシステムとなる．図 1 の “Noise Rejection” では，Lee ら [11] による 5 クラスの GMM を用いた音響尤度に基づく棄却が行われており，単純に有効と無効の 2 クラスのモデルを作るよりも雑音それぞれのクラスのモデルを作成することが有効であると示されている．しかし，本論文で無効入力としている「無効発話」は実験データには含まれていない．したがって，「無効発話」も含めた無効入力の識別については，ここで調査する必要がある．

2.2 収集データ

「たけまるくん」は 2002 年 11 月より運用を開始し，以降現在までのすべての入力データを収録している．このうち，最初の 2 年間分のデータは聴取による書き起こしと，有効入力または無効入力のラベル，年齢層，性別，雑音などのタグが付与され，データベースとして整備が進められている．本論文における有効入力，無効入力の分類はこのラベルに従ったものである．ここで，有効入力はシステムがユーザに対して何らかの応答を行うことが適切と判断された入力のことであり，システムが想定したタスク内発話のみでなくタスク外の質問も含まれる．オーバフローやレベル不足の入力は発話内容の判断が困難なものを含むため，本研究では無効入力として扱った．詳細な分類を**表 1** に示す．表 1 において，無効入力は，人の音声による「無効発話」，「咳」，「笑い声」，およびそれ以外の非音声の「雑音」に大きく分類される．「無効発話」にはさらに詳細なタグが付けられている．「背景会話」とは発話者の背後で他人の会話が重なって聞こえるものおよび明らかにシステムに対する発話ではなくマイクの周辺で会話されている発話，「発話不明瞭」とは音声が聞き取りにくく客観的に判別できないもの，「意味のない発話」とはフィラや「マイクテスト」などのシステムからの情報取得が目的ではない発話，「音声区間検出ミス」とは文頭もしくは文末が欠損している発話，「オーバフロー」は発話者の声が大きすぎて音割れを起こしている発話，「レベル不足」は入力音が小さすぎて発話内容が聴取できない発話を指す．なお，これらのタグは重複を許している．また，雑音タグが与えられていても，発話内容がシステムの入力として有効である場合は有効入力として分類しており，表 1 の無効入力には集計していない．

表 1 「たけまるくん」の入力データの分類結果（2002 年 11 月から 2004 年 10 月まで）
Table 1 Classification result on input data of “Takemaru-kun” (from Nov. 2002 till Oct. 2004).

カテゴリ			発話数	合計
有効入力	大人発話		20,436	
	子供発話		85,889	106,325
無効入力	無効発話	背景会話	26,319	122,939
		発話不明瞭	13,348	
		意味のない発話	11,991	
		音声区間検出ミス	12,937	
		オーバフロー	1,417	
		レベル不足	7,347	
	咳		727	
	笑い声		6,232	
	雑音		50,756	

3. BOW 特徴量を用いた無効入力の識別

3.1 特徴量

音声認識の前処理で得られる特徴量 3 種類と，音声認識結果も使用した特徴量 1 種類からなる，以下の 4 種類の特徴量を検討した．

- **GMM による音響尤度（GMM）**
  入力音の 6 クラス，「大人発話」，「子供発話」，「無効発話」，「咳」，「笑い声」，「雑音」の各 GMM に対する尤度の時間平均値を要素とする 6 次元のベクトル．GMM は音響的特徴量が大きく異なると考えられるクラスごとに作成した．「大人発話」，「子供発話」を有効入力とし，「無効発話」，「咳」，「笑い声」，「雑音」を無効入力として用いた．GMM 作成に使用したデータは「たけまるくん」によって実環境で収集されたデータである．
- **発話時間（Duration）**
  音声認識エンジン Julius による振幅と零交差法に基づく音声区間検出により，1 発話と見なされた入力音の時間長．
- **信号対雑音比（Signal to Noise Ratio: SNR）**
  1 発話ごとに算出した SNR の値．入力音をフレームに分割し，便宜的にそのフレームの中で平均パワーの大きいフレームの上位 10%を信号区間，平均パワーの小さいフレーム下位 10%を雑音区間と考え，式 (1) により求める．

$$\mathrm{SNR} = 10 \log_{10} \frac{P_S - P_N}{P_N} \tag{1}$$

ここで，P_S は信号区間の平均パワー，P_N は雑音区間の平均パワーを表す．

- **Bag-of-Words（BOW）**
  音声認識結果の N-best に含まれている単語の出現頻度を要素とした特徴量ベクトル．数え上げる単語は，学習データの音声認識結果から作られた単語辞書中にあるものに限る．

3.2 識別手法

3.2.1 SVM による識別手法

SVM [8] は教師あり学習機械であり，2 クラス分類問題を対象とする．SVM は，与えられたデータを，カーネル関数によって高次元へと写像し，写像した空間において 2 クラスに分類する．その際に 2 クラス間のマージンが最大となる識別境界を求める．今，n 次元の特徴量ベクトル $\mathbf{x}_i \in R^n$，$i = 1, \ldots, l$（l：サンプル数）とラベル $y_i \in \{+1, -1\}$ のペア集合が与えられたとすると，次の式 (2) に従って，2 クラス分類のための識別境界が求められる．ここで $\mathbf{w}$，b は識別関数のパラメータ，C はコストパラメータ，$\phi(\mathbf{x}_i)$ は非線形関数である．

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \frac{1}{2} \quad & \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ & \text{subject to } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{2}$$

ここで，ξ_i はスラック変数であり，これによりある程度の誤分類を許容しつつマージンの最大化を行う．また，この式 (2) を式 (3) へと拡張することにより，正例（$y_i = +1$）と負例（$y_i = -1$）の数がアンバランスな問題に対処できる．

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \frac{1}{2} \quad & \mathbf{w}^T\mathbf{w} + C_+\sum_{\{i:y_i=+1\}}\xi_i + C_-\sum_{\{i:y_i=-1\}}\xi_i \\ & \text{subject to } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{3}$$

具体的には，分類誤りに対するコストパラメータ C を，正例を負例とする誤りのコストパラメータ C_+ と負例を正例とする誤りのコストパラメータ C_- に分けて，2 つの誤り率のバランスをとる．本論文では，いくつかのコストパラメータ C によって評価データの識別性能を求めて，最適値を求めた．C_+ と C_- は次式により与える．

$$C = C_+ + C_-$$
$$C_+ = \frac{N_-}{N_+ + N_-} \times C, \quad C_- = \frac{N_+}{N_+ + N_-} \times C$$

ここで，N_+ は正例のデータ数，N_- は負例のデータ数である．

本論文では，各特徴量間の値の大きさの違いや次元数を考慮するために，マルチカーネル法 [12] を用いる．データ x_i，x_j のカーネル関数の値 $k(x_i, x_j)$ を i，j 成分とする行列（グラム行列）[13] を特徴量ごとに算出し，足し合わせてから識別境界を求める手法をマルチカーネル法と呼ぶ．特徴量の種類が M 個のとき，マルチカーネル法は式 (4) により表すことができる．

$$K(x_i, x_j) = \sum_{c=1}^{M} a_c k_c(x_i, x_j), \quad \left(\sum_{c=1}^{M} a_c = 1,\ a_c \geq 0\right) \tag{4}$$

ただし，$k_c(x_i, y_j)$ は $\phi(x_i)$ と $\phi(x_j)$ の内積であり，a_c は特徴量 c の重み係数である．

3.2.2 ME による識別手法

ME [9], [10] は，分類問題によく用いられる一般的な機械学習手法である．ME では，特徴量はモデルの制約に対応しており，複数の特徴量を統合することができる．クラスのラベル集合 E と特徴量の集合 D について，学習データのセット (E, D) が与えられたとき，以下の対数尤度を最大化することにより，$\Lambda = \{\lambda_i, i = 1, \ldots, l\}$ を学習する．

$$\log P(E|D, \Lambda) = \sum_{(e,d)\in(E,D)} \log \frac{\exp \sum_i \lambda_i f_i(e, d)}{\sum_{e'} \exp \sum_i \lambda_i f_i(e', d)} \tag{5}$$

ここで，f_i は特徴量に対する素性関数であり，λ_i は各素性関数に対する重みである．そして，学習された Λ を用いて，事後確率が最大となるクラスにデータを識別することができる．なお，ME で複数の特徴量を用いる場合，単純に特徴量の集合 D に新たな特徴量を追加する．

4. 有効入力と無効入力の識別実験

本実験の目的は，以下の 2 つである．3 章であげた特徴量を用いた SVM および ME による有効入力と無効入力の識別精度を評価することおよび，複数の特徴量を組み合わせて識別精度を評価することである．

実験条件

本実験において使用するデータを**表 2** に示す．この学習データ，テストデータはそれぞれ「たけまるくん」によって得られた 1 カ月分の入力データである[*1]．なお，表 2 の学習データと GMM の学習データは同一である．GMM の学習において，標本化/量子化は 16 kHz/16 bit，分析窓長は 25 msec，窓シフト長は 10 msec とし，MFCC（12 次元），Δ MFCC，Δパワーを用いた．混合数は 128 とし，学習に用いたデータは**表 3** のとおりである．実験条件を**表 4** に示す．

本実験における評価尺度には，F 尺度 [14] を用いる．F 尺度は，適合率（P）と再現率（R）という，正確性と網羅性の総合的な評価尺度であり，以下の式で定義される．

[*1] 学習データは 2003 年 8 月分，テストデータは 2003 年 10 月分．夏期休暇中は「たけまるくん」の利用者が大幅に増加するため，8 月分は 10 月分のおよそ 2 倍の数になっている．

$$\mathrm{F} = \frac{2 \cdot P \cdot R}{P + R}$$

ただし，

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}, \quad R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}$$

であり，N_{TP} は無効入力を無効入力と識別した数，N_{FP} は有効入力を無効入力と識別した数，N_{FN} は無効入力を有効入力と識別した数である．

4.1 予備実験：各特徴量の識別性能の評価実験

3 章で述べた特徴量を個別に用いて，SVM および ME によって有効入力と無効入力の識別を試みる．3 章で述べた 6 クラスの GMM を用いた SVM による無効入力の識別手法を従来手法として考え，その他の特徴量を用いたときの SVM および ME による識別手法の結果と比較する．

実験の結果を**図 2** に示す．BOW を特徴量とした場合，SVM，ME のどちらも従来手法より F 尺度が向上している．BOW を特徴量とした SVM が最良の結果であり，従来手法と比較し，F 尺度が 82.19%から 85.09%に改善された．また，BOW を特徴量とした ME でも，83.19%に改善された．このことから，無効入力の識別における BOW の有効性が示された．

4.2 複数特徴量の組合せによる識別性能の評価実験

複数の特徴量を組み合わせて無効入力の識別を試みる．本実験では，特徴量の組合せによる BOW 特徴量の効果を検証するため，「GMM」，「GMM + Duration + SNR」，「GMM + BOW」，「GMM + Duration + SNR + BOW」の 4 つの組合せについて無効入力の識別を行った．

実験の結果を**図 3** に示す．「GMM」と「GMM + BOW」を比較すると，後者の方が F 尺度が改善されている．また，「GMM + Duration + SNR」と「GMM + Duration + SNR + BOW」を比較すると，同様に後者の方が F 尺度が改善されている．最良の結果を示したのは，「GMM + Duration + SNR + BOW」の 4 つすべての特徴量を用いた SVM による識別手法であり，F 尺度は 86.58%に改善された．これは，「GMM」を用いた従来手法と比較すると 4.39 ポイントの改善となる．このことから，BOW 特徴量は複数特徴量の組合せにおいても，無効入力の識別において有効であることが示された．

表 2 実験データ
Table 2 Experiment data.

	有効入力（負例）	無効入力（正例）	計
学習データ	7,607	7,274	14,881
テストデータ	3,782	3,902	7,684

表 3 GMM の学習に用いたデータ数
Table 3 Training data used in GMM training.

有効入力	大人発話	1,053
	子供発話	6,554
無効入力	無効発話	3,640
	咳	29
	笑い声	287
	雑音	3,318

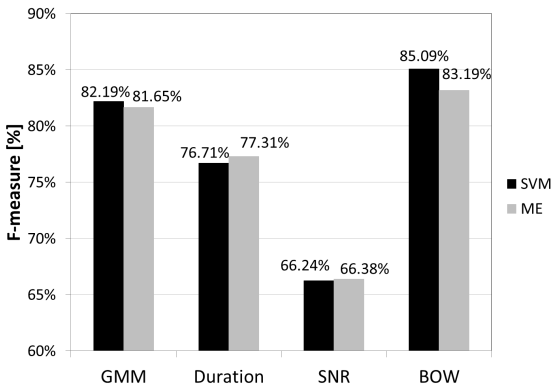

図 2 単一特徴量による無効入力の識別性能
Fig. 2 Result of invalid input discrimination using a single feature.

表 4 実験条件
Table 4 Experimental condition.

音声認識	音声認識エンジン	Julius 4.0.2 [15]
	言語モデル	「たけまるくん」の 2 年間分の書き起こし文から作ったモデル [7]
	音響モデル	JNAS [16] モデルを「たけまるくん」のデータで適応したトライフォンモデル
	出力	10-best
形態素解析器		Chasen 2.3.3 [17]
BOW の単語辞書の大きさ		4,488
SVM	SVM ツール	LIBSVM [18]
	カーネル関数	Radial Basis Function（RBF）
	パラメータ C	$10^{-2}, 10^{-1}, \ldots, 10^{4}$（10 倍刻み）
ME	ME ツール	Stanford Classifier 2.1.3 [19]

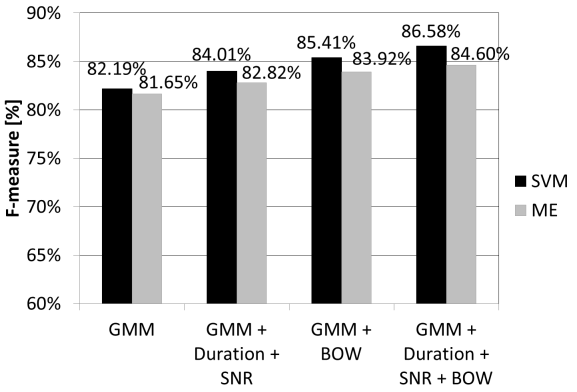

図 3 複数特徴量による無効入力の識別性能

Fig. 3 Result of invalid input discrimination using multiple features.

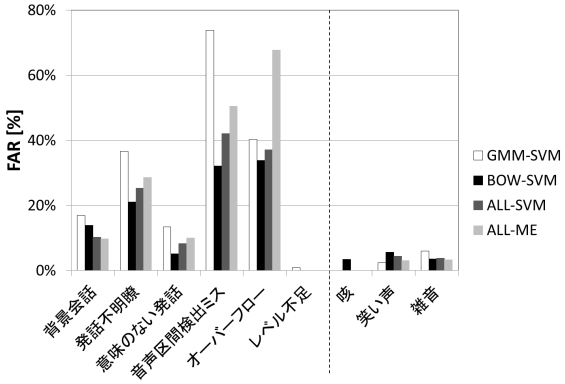

図 4 無効入力のカテゴリごとの FAR

Fig. 4 False Acceptance Rate of invalid inputs for each category.

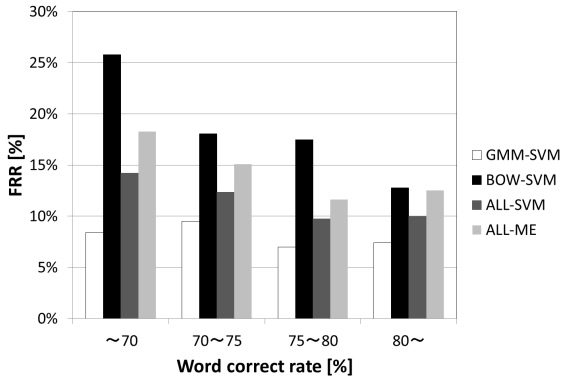

図 5 有効入力の音声認識結果の単語正解率に対する FRR

Fig. 5 False Rejection Rate of valid inputs for corection of ASR result.

4.3 BOW の有効性に関する考察

前節までの実験によって，BOW の有効性は F 尺度の上昇として示された．本節ではさらに無効入力の誤受理と有効入力の誤棄却に注目して考察を行う．ここでは，SVM について，「GMM（GMM-SVM）」，「BOW（BOW-SVM）」，「GMM + Duration + SNR + BOW（ALL-SVM）」の 3 手法と，ME について，「GMM + Duration + SNR + BOW（ALL-ME）」の，計 4 手法について示す．

まず，無効入力について，カテゴリ別の識別誤りの傾向を調査するため，カテゴリごとの FAR（False Acceptance Rate；誤受理率）を算出し，**図 4** に示す．FAR は以下の式で定義される．

$$\mathrm{FAR} = \frac{N_{accept}}{N_d}$$

ただし，N_d はテストデータ中のそのカテゴリに含まれるデータの数，N_{accept} は N_d のうち誤って有効入力と識別されたデータの数である．従来手法の GMM-SVM と BOW-SVM を比較すると，左 6 個の無効発話のカテゴリにおいて，BOW-SVM の方が FAR が小さい．たとえば，GMM の音響尤度を用いた SVM による従来手法では「音声区間検出ミス」の FAR が 80%近くあったが，BOW を用いた SVM による識別では，30%台に改善できている．これは，認識精度が問われない無効発話音声であっても，音声認識結果の BOW を用いることで単語単位でのベクトル量子化が行われ，有効な識別器が構成されたと考えられる．ただし，咳や笑い声では GMM-SVM が優位であることから，まったく言語的特徴を含まない無効入力に対しては BOW は有効ではないと考えられる．BOW-SVM と ALL-SVM を比較すると，無効発話のカテゴリでは，BOW-SVM の性能が良い．評価実験で述べたように，全体の無効入力識別性能は全特徴量の組合せが良いが，無効発話のカテゴリに限定すると BOW 単体の方が良いことが分かる．特徴量の組み合わせ方によっては，より性能を向上させることができる可能性があると考えられる．

続いて，有効入力について，音声認識率による識別誤り傾向を調査するため，テストデータの音声認識結果の単語正解率ごとに FRR（誤棄却率；False Rejection Rate）を算出し，**図 5** に示す．FRR は以下の式で定義される．

$$\mathrm{FRR} = \frac{N_{reject}}{N_{valid}}$$

ただし，N_{valid} はテストデータ中の有効入力のデータ数，N_{reject} は N_{valid} のうち誤って無効入力と識別されたデータの数である．テストデータ中の有効入力の音声認識結果から，日付ごとに単語正解率を算出し，単語正解率が 70%未満，70%以上 75%未満，75%以上 80%未満，80%以上の 4 グループに分割して FRR を算出した．従来手法の GMM-SVM と他の 3 手法を比較すると，すべての単語正解率で GMM-SVM の FRR が小さい．また，GMM-SVM 以外では，おおよそ単語正解率が大きくなるにつれて，FRR は小さくなるという傾向がある．BOW-SVM では，単語正解率 70%未満のときは 26%程度と FRR が大きくなっているが，単語正解率 80%以上のときにはおよそ 13%となっており，音声認識率が高い方が識別誤りも起こりにくいということが分かる．ALL-SVM では BOW-SVM に見られた低い認識率での FRR を低くしており，結果として F 尺度

表 5 SVM と ME による識別結果の差異
Table 5 Difference between Classification results of SVM and ME.

		ME	
		無効	有効
SVM	無効	3,073	229
	有効	168	432

の低下につながったものと考えられる．以上より，GMM などの音響特徴量に BOW を追加することで，FAR の減少が期待できるが，一方で低認識率のデータが多い場合には，FRR が悪化する可能性がある．このことは異なったデータベースを用いた実験により検証する必要があるが，本論文での議論の範囲を超えるため今後の課題とする．

最後に，SVM と ME の識別誤りの傾向を調査するため，テストデータ中の無効入力のうち，SVM および ME で無効入力・有効入力と識別されたものをそれぞれ計上し，**表 5** に示す．特徴量は，すべての特徴量「GMM + Duration + SNR + BOW」を用いた．SVM および ME によって，無効入力と識別されたデータが 3,073 個と最も多いが，約 30%のデータについては，SVM と ME の識別結果に入れ替わりが起こっている．これより，SVM と ME との識別誤りの傾向が異なっていることが分かる．SVM では特徴量を非線形に扱っているのに対して，ME では特徴量を線形に扱っているため，異なる分類傾向を持った識別器が構成されたと考えられる．この事実は 2 つの識別器の出力を合成することで，さらに高精度な識別器が構成できることを示唆している．

5. まとめ

無効入力の識別手法として，音響的特徴量に言語的特徴量である BOW を加えた，SVM および ME による識別を提案した．GMM を用いた SVM による従来手法と比較して，GMM，Duration，SNR および BOW を特徴量とした SVM では，F 尺度が 82.19%から 86.58%に改善された．BOW は無効入力の識別に有効であり，特に無効発話の識別性能の向上には寄与しており，音響的特徴量と組み合わせると高い識別性能が得られることが示された．ただし，特徴量の組合せによっては，必ずしも性能が向上するとは限らないため，組合せの手法には検討の余地がある．

以上の実験により，BOW 特徴量は無効入力の識別に有効であることが示されたが，いくつかの課題が残されている．BOW の次元数は対話ドメインに依存するため，別ドメインの対話コーパスに適用した場合の比較実験により，本手法の有効性を示す必要がある．また，本論文で用いた「たけまるくん」データベースには大量のラベルなしデータがあるため，それらを活用した半教師あり学習（semi-supervised learning）[20] による識別性能向上が考えられる．

謝辞 本研究の一部は，戦略的創造研究推進事業「共生社会に向けた人間調和型情報技術の構築」（JST/CREST）の援助を受けて行われた．

参考文献

[1] Nisimura, R., Lee, A., Saruwatari, H. and Shikano, K.: Public Speech-oriented Guidance System with Adult and Child Discrimination Capability, *Proc. ICASSP*, pp.433–436 (2004).
[2] Sukka, R.A. and Lee, C.H.: 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).
[3] Matsui, T., Soong, F. and Juang, B.H.: Verification of multiple class recognition: A classification approach, *IEICE Trans. Information and Systems*, Vol.E88-D, No.3, pp.455–462 (2005).
[4] Lane, I.R., Kawahara, T., Matsui, T. and Nakamura, S.: Out-of-Domain utterance detection using classification confidences of multiple topics, *IEEE Trans. Acoustis, Speech, and Language Processing*, Vol.15, No.1, pp.150–161 (2007).
[5] 中村敬介，西村竜一，李　晃伸，猿渡　洋，鹿野清宏：実環境音声情報案内システムにおける環境雑音および不要発話の識別，電子情報通信学会技術研究報告，Vol.SP2003-172, pp.13–18 (2004).
[6] 鈴木智詞，竹内義則，松本哲也，工藤博章，大西　昇：聴覚障害者のための警告音の識別，電子情報通信学会技術研究報告 SP2004-156，pp.154–163 (2005).
[7] Sakai, H., Cincarek, T., Kawanami, H., Saruwatari, H., Shikano, K. and Lee, A.: Voice activity detection applied to hands-free spoken dialogue robot based on decoding using acoustic and language model, *Proc. 1st International Conference on Robot Communication and Coordination* (*ROBOCOMM2007*), No.16 (2007).
[8] Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer (1995).
[9] Berger, A.L., Pietra, S. and Pietra, V.: A Maximum entropy approach to natural language processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).
[10] Manning, C. and Klein, D.: Optimization, Maxent Models, and Conditional Estimation without Magic, *Tutorial at HLT-NAACL and ACL* (2003).
[11] Lee, A., Nakamura, K., Nishimura, R., Saruwatari, H. and Shikano, K.: Noise robust real world spoken dialog system using GMM based rejection of unintended inputs, *Proc. ICSLP*, pp.173–176 (2004).
[12] Lanckriet, G.R.G., Cristianini, N., Bartlett, P., Ghaoui, L.E. and Jordan, M.I.: Learning the Kernel Matrix with Semidefinite Programming, *Journal of Machine Learning Research*, Vol.5, pp.27–72 (2004).
[13] Bishop, C.M.: *Pattern recognition and machine learning*, Springer (2006).
[14] Manning, C.D., Raghavan, P. and Schuetze, H.: *Introduction to Information Retrieval*, Cambridge University Press Anniversary (2008).
[15] Lee, A., Kawahara, T. and Shikano, K.: Julius – An open source real-time large vocabulary recognition engine, *Proc. Eurospeech*, pp.1691–1694 (2001).
[16] 国立情報学研究所音声資源コンソーシアム：新聞記事読み上げ音声コーパス JNAS，入手先 ⟨http://www.mibel.

cs.tsukuba.ac.jp/_090624/jnas/〉（参照 2012-11-12）.
[17] 奈良先端科学技術大学院大学：形態素解析器 Chasen，入手先〈http://chasen-legacy.sourceforge.jp/〉（参照 2012-11-12）.
[18] Chang, C. and Lin, C.: LIBSVM: A library for support vector machines, available from 〈http://www.csie.ntu.edu.tw/˜cjlin/libsvm〉 (accessed 2012-11-12).
[19] The Stanford Natural Language Processing Group: Stanford Classifier, available from 〈http://nlp.stanford.edu/software/classifier.shtml〉 (accessed 2012-11-12).
[20] Zhu, X.: Semi-supervised learning literature survey, available from 〈http://pages.cs.wisc.edu/˜jerryzhu/research/ssl/semireview.html〉 (accessed 2012-11-12).

真嶋 温佳

平成 23 年関西大学システム理工学部電気電子情報工学科卒業．同年より，奈良先端科学技術大学院大学情報科学研究科博士前期課程在学中．音声対話システムのための音声認識の研究に従事．日本音響学会学生会員．

藤田 洋子

平成 20 年京都府立大学人間環境学部環境情報学科卒業．平成 22 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年 4 月株式会社プロアシスト入社．現在，組み込み，脳波関連技術の業務に従事．

トーレス ラファエル

平成 17 年パナマ工科大学計算機システム工学科卒業．平成 22 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大学院博士後期課程在学中．文部科学省国費留学生．音声対話システムのための音声認識と自然言語処理の研究に従事．日本音響学会，IEEE 各学生会員．

川波 弘道

平成 6 年東京大学工学部電気工学科卒業．平成 12 年同大学大学院工学系研究科博士課程修了．博士（工学）．同年電子技術総合研究所入所．平成 13 年奈良先端科学技術大学院大学情報科学研究科助手，平成 19 年同助教．現在，音声分析，音声対話の研究に従事．電子情報通信学会，日本音響学会各会員．

原 直 （正会員）

平成 15 年名古屋大学工学部卒業．平成 17 年名古屋大学大学院情報科学研究科博士前期課程修了．平成 23 年同博士後期課程修了．博士（情報科学）．同年奈良先端科学技術大学院大学情報科学研究科助教．平成 24 年岡山大学大学院自然科学研究科助教，現在に至る．音声認識システムの実用化に関する研究に従事．日本音響学会，電子情報通信学会，ヒューマンインタフェース学会各会員．

松井 知子

昭和 63 年東京工業大学大学院修士課程修了．同年 NTT（株）入社．話者・音声認識の研究に従事．平成 10 年 ATR 音声翻訳通信研究所，平成 12 年 ATR 音声言語通信研究所および音声言語コミュニケーション研究所に出向．平成 13 年 1 月〜6 月米ルーセント・テクノロジー社ベル研究所客員研究員．平成 15 年情報・システム研究機構統計数理研究所准教授．平成 20 年同研究所教授．統計数理の研究に従事．東京工業大学博士（工学）．IEEE senior member，日本音響学会，日本統計学会各会員．1993 年電子情報通信学会論文賞受賞．

猿渡 洋

平成 3 年名古屋大学工学部電気工学科卒業．平成 5 年同大学大学院修士課程修了．平成 12 年同大学院博士課程修了．工学博士．平成 5 年セコム（株）入社．セコム IS 研究所音声情報処理研究室において，音響アレー信号処理に関する研究に従事．平成 12 年奈良科学技術大学院大学助教授．平成 19 年同准教授．音声信号処理，統計的信号処理等に関する研究に従事．平成 13，18 年電子情報通信学会論文賞受賞．平成 15，20，23 年電気通信普及財団テレコムシステム技術論文賞受賞．平成 23 年ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞受賞．日本音響学会，日本 VR 学会，IEEE 各会員．

鹿野 清宏 （フェロー）

昭和 45 年名古屋大学工学部電気工学科卒業．昭和 47 年同大学大学院修士課程修了．同年電電公社武蔵野電気通信研究所入所．昭和 59～61 年カーネギーメロン大客員研究員．昭和 61～平成 2 年 ATR 自動翻訳電話研究所音声情報処理研究室長．平成 4 年 NTT ヒューマンインタフェース研究所主席研究員．平成 6 年より奈良先端科学技術大学院大学情報科学研究科教授．音情報処理学講座を担当．工学博士．音声・音情報処理の研究および研究指導に従事．昭和 50 年電子通信学会米沢賞，平成 3 年 IEEE SP 1990 Senior Award，平成 6 年日本音響学会技術開発賞，平成 12 年情報処理学会山下記念研究賞，平成 13 年 VR 学会論文賞，平成 17，18 年電子情報通信学会論文賞，平成 17 年猪瀬賞．電子情報通信学会フェロー，IEEE フェロー，ISCA，音響学会各会員．