Vol.2010-SLP-80 No.3
2010/2/12

単語の頻度と音響の特徴を利用したSVMによる無効入力の棄却

藤田洋子[†1] 竹内翔大[†1] 川波弘道[†1]
松井知子[†2] 猿渡洋[†1] 鹿野清宏[†1]

実環境で，音声認識を用いた音声情報案内システムを稼動させる場合には，雑音などの音声以外の入力やユーザ同士の背景会話などが混入されてくることがある．これらの入力はシステムの誤作動・誤認識を引き起こし，システムの応答性能を低下させる原因となる．そのため，システムへの入力として適当な入力（有効入力）と不適当な入力（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となる．

一般的に有効入力と無効入力を識別には音響的特徴が用いられる．しかし，入力音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，無意味な認識結果が出力される雑音の識別に加え，システムのタスクの言語的な特徴を反映させた有効入力，無効入力を識別することが可能になると考えられる．そこで本稿では，Bag-of-Words (BOW) を特徴量とした Support Vector Machine (SVM) による無効入力の識別を検討した．実環境音声認識システム「たけまるくん」の入力データを用いた実験では，GMM に基づく無効入力の識別と比べ，分類誤り率を 23.30% から 15.90% に削減することができた．また，BOW に GMM から得られる音響尤度，発話時間や SNR を組み合わせた手法についても検討した．その結果，分類誤り率を 13.60% まで削減することができた．

Rejection of Invalid Input Using SVM based on BOW and acoustic features

Yoko FUJITA,[†1] Shota TAKEUCHI,[†1]
Hiromichi KAWANAMI,[†1] Tomoko MATSUI,[†2]
Hiroshi SARUWATARI[†1] and Kiyohiro SHIKANO[†1]

On a real environment speech-oriented information guidance system, a valid and invalid input discrimination process is important as invalid inputs such as noise, laugh, cough and meaningless utterances lead to unpredictable system responses.
Generally, acoustic features such as MFCC 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 report, the authors propose to introduce Bag-of-Words (BOW) as a feature to discriminate between valid and invalid inputs. Support Vector Machine (SVM) is also employed to realize robust classification. Experiments using real environment data from the guidance system "Takemaru-kun" were conducted. By applying BOW and SVM, the classification error rate (CER) is reduced to 15.90% , from 23.30% when using GMMs. In addition, experiments using features combining BOW with acoustic likelihoods from GMMs, SNR and duration were conducted, improving the CER to 13.6% .

1. はじめに

実環境で音声認識を行う場合には，非音声も含めたさまざまな音の入力を想定しておく必要がある．音声認識を用いた音声情報案内システムを稼動させる場合には，システムの入力として適当である発話（有効入力）以外の入力があることになり，このような入力がシステムの誤作動・誤認識の原因になる．このため，システムへの入力としては不適当な入力（無効入力）は棄却し，無効入力に対する応答は行わせないことが望ましい．

無効入力の識別には一般的に Mel-Frequency Cepstrum Coefficient (MFCC) 特徴量などの音響的特徴が使われており，混合ガウス分布モデル (Gaussian Mixture Model : GMM) による音響尤度を用いた識別手法が代表的である[1)]．また，音声と雑音との識別においては，音声区間検出の手法として音声認識結果から得られる言語的な制約を用いる手法が提案されており[2)]，言語的な情報を用いた雑音と音声の識別がある程度可能であることが分かっている．そこで，本稿では音声認識結果から得られた Bag-of-Words (BOW) を特徴量とした Support Vector Machine (SVM) による無効入力の識別手法を検討する．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

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of science and Technology
†2 統計数理研究所
The Institute of Statistical Mathematics

Vol.2010-SLP-80 No.3
2010/2/12

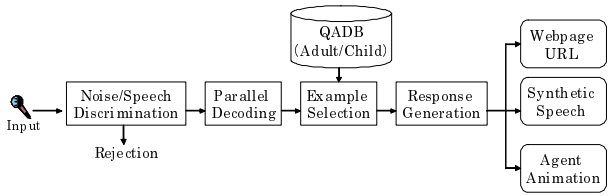

図 1 「たけまるくん」の応答処理の流れ
Fig. 1 Processing flow of "Takemaru-kun."

れば，システムのタスクを考えた上で有効入力と無効入力を識別することが可能になると考える．また，BOW に音響尤度などを組み合わせ，識別精度を向上させる検討も行う．

以下，2 節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「たけまるくん」から得られた無効入力について説明し，3 節では従来手法として GMM を用いた環境雑音の識別手法を述べる．4 節では本稿で提案する SVM を用いた無効入力の識別手法を説明し，5 節にてその性能評価実験の結果を示し，6 節でまとめを行う．

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」[3] とは，奈良県生駒市の北コミュニティセンターに常設されている音声情報案内システムである．「たけまるくん」は，質問と正しい応答のペアをデータベース化した質問応答データベース（Question and Answer Database: QADB）を用い，データベース検索による一問一答方式の応答を行っている．「たけまるくん」の主な応答内容としては，施設利用者に対する施設・観光案内，エージェント（たけまるくん）に対する質問，現在の時間，天気，ニュースなどがある．この時の「たけまるくん」における処理の流れを図 1 に示す．

2.2 使用データ

「たけまるくん」は 2002 年 11 月の運用開始以降，すべての入力データを収集している．この内，2004 年 10 月までの 2 年間分のデータは聴取による年齢層，性別，有効入力／無効入力のラベル，雑音などタグの付与及び書き起こし作業が終了している．本稿における有効入力，無効入力の分類はこのラベルに従ったものである．これをより詳細に分類した結果を表 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」および「咳」，

表 1 「たけまるくん」の入力データの分類結果 (2002.11〜2004.10)
Table 1 Input data of "Takemaru-kun."

カテゴリ			発話数	合計
有効入力		大人発話	20436	
		子供発話	85889	106325
無効入力	無効発話	背景会話	26319	
		発話不明瞭	13348	
		意味のない発話	11991	
		音声区間検出ミス	12937	
		オーバーフロー	1417	122939
		レベル不足	7347	
	咳		727	
	笑い声		6232	
	雑音		50756	

※無効入力のタグは重複を許している．
※雑音タグが付与されていても発話内容がシステムに対する入力として有効な発話は有効入力に分類されている．

「笑い声」，それ以外の非音声の「雑音」に大きく分類している．「無効発話」におけるタグの「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及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づらく客観的に判別できないもの，「意味のない発話」とはフィラーや「マイクテスト」などのシステムからの情報提供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けている発話，「オーバーフロー」は発話者の声が大きすぎ，音割れを起こしている発話であり，「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示している．ただし，これらのタグは重複している．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いる．

3. GMM を用いた無効入力の識別方法

無効入力を識別する先行研究としては，GMM や主成分分析を用いる手法がある[1)4)]．今回は，代表的な手法として GMM を用いた識別手法をとりあげる．一般的に，この手法では MFCC 特徴量から識別したい音のクラスごとに GMM を作成し，新たな音が入力されてきた場合，この入力音の各 GMM に対する時間平均の尤度を計算し，入力音がどの音のクラスにもっとも類似しているかにより識別を行う (図 2)．

今回はこの GMM を用いた手法を従来手法と位置づける．本実験では，GMM は音響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．有効入力として「大人」発話，「子

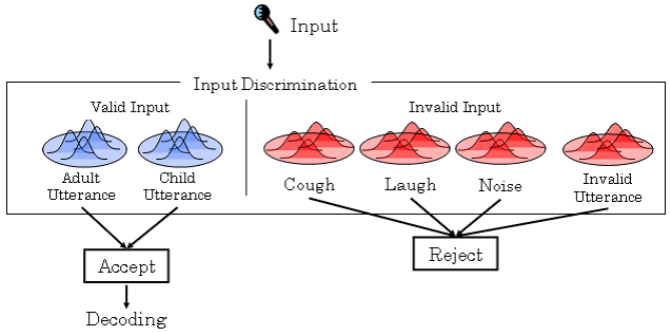

図 2 GMM による入力識別
Fig. 2 Input discrimination by GMM

供」発話，無効入力として「無効発話」,「咳」,「笑い声」,「雑音」の計 6 クラスがある．

4. SVM を用いた無効入力の識別方法

本稿では，BOW を特徴量とした無効入力の識別を検討する．しかし，BOW は次元が非常に大きい特徴量であるため，学習手法によっては過学習の可能性がある．そこで，過学習を起こすことが少ないと言われている SVM を用いる．

4.1 SVM

SVM 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 はカーネル関数を用いて，与えられたデータを高次元へと写像し，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 $\mathbf{x_i} \in R^n, i = 1, ..., l$ (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 (1) に従って，2 クラス分類のための識別境界が求められる．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 (1) を式 (2) へと拡張することにより，正例と負例の数がアンバランスな問題に対処できる．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．コストパラメータの値は事後的に設定した．

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \qquad (1)$$
$$\text{Subject to } y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, ..., l.$$

$\mathbf{w}$: 識別関数のパラメータ
C : コストパラメータ
ϕ : 非線形関数

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+\sum_{\{i:y_i=1\}}\xi_i + C_-\sum_{\{i:y_i=-1\}}\xi_i \qquad (2)$$
$$\text{Subject to } y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, ..., l.$$

4.2 特 徴 量

特徴量の選択は，SVM の識別性能に大きな影響を与える．本実験では以下の特徴量を検討した．

・GMM による音響尤度 (GMM)

3 節において述べた従来手法においてクラスごとに作成した 6 つの GMM の音響尤度を要素とする 6 次元のベクトル．

・Bag-of-Words (BOW)

音声認識結果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．数え上げる単語は単語辞書 (Wordlist) 中にあるものだけに限る．Wordlist は学習データより作られるが，この時全データを用いず，有効入力（もしくは無効入力）のデータから Wordlist を作ることもできる．どのような単語がよりよい特徴量となるかを調べるため，今回は以下のような Wordlist を用いた．

- すべての音声認識結果から作った Wordlist (4482 単語)
- 有効入力の音声認識結果から作った Wordlist (3838 単語)
- 無効入力の音声認識結果から作った Wordlist (2850 単語)

・発話時間

音声認識エンジン Julius[5] による振幅と零交差に基づく音声区間検出により一発話と見なされた入力音の時間．

・信号対雑音比（Signal to Noise Ratio: SNR）

一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の下位 10% を雑音区間と考え，式 (3) より求める．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．

$$SNR = 10 \log_{10} \frac{P_S - P_N}{P_N} \tag{3}$$

4.3 マルチカーネル法

一般的に SVM ではすべての特徴量ベクトルから各データのカーネル値を計算し，識別境界を求める．しかし，傾向の異なる複数の特徴量を用いる場合，各特徴量間の値の大きさの違いや次元数の違いを考慮する必要がある．そこで本稿ではマルチカーネル法を用いる．

データ $\mathbf{x_i}$，$\mathbf{x_j}$ のカーネル値 $k(\mathbf{x_i}, \mathbf{x_j})$ を i, j 成分とする行列のことをグラム行列と呼ぶ．このグラム行列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 個の時，マルチカーネル法は式 (4) により表すことができる．

$$K(\mathbf{x_i}, \mathbf{x_j}) = \sum_{c=1}^{M} k_c(\mathbf{x_i}, \mathbf{x_j}) \tag{4}$$

5. 評 価 実 験

本実験における目的は，次の 2 つである．

- 4 節で挙げた BOW などの特徴量を用いた SVM による有効入力と無効入力の識別精度を評価すること
- 複数の特徴量を組み合わせて，識別精度の向上をはかること

そこで以降のような実験を行った．

5.1 実験データ

本実験において使用するデータを表 2 に示す．なお，表 2 の学習データと GMM の学習データは一致させているが，そのより詳細なデータを表 3 に示す．

5.2 評 価 方 法

本実験における評価尺度に分類誤り率 (Classification Error Rate : CER) を用いる．また，

表 2 実験データ
Table 2 Experiment data

	有効入力	無効入力	計
学習データ	7607	7274	14881
テストデータ	3782	3902	7684

表 3 GMM の学習条件
Table 3 Training conditions of GMM

学習データ数	有効入力	大人	1053
		子供	6554
	無効入力	無効発話	3640
		咳	29
		笑い声	287
		雑音	3318
標本化/量子化	16kHz/16bit		
窓長	25msec		
窓シフト長	10msec		
特徴量	MFCC (12 次元), Δ MFCC, Δパワー		
混合数	128		

表 4 実験条件
Table 4 Experiment condition

音声認識エンジン	Julius4.0.2[5]
形態素解析器	Chasen2.3.3
SVM ツール	LIBSVM[6]
カーネル関数	Radial Basis Function（RBF）
パラメータ C	0.01〜10000（10 倍刻み）

誤り傾向などを調べるための尺度として，無効入力を有効入力に誤分類する誤り率（False Acceptance Rate: FAR）と有効入力を無効入力と誤分類する誤り率（False Rejection Rate: FRR）を用いる（式 (5)-(7)）．

$$CER = \frac{\text{誤分類したデータ数}}{\text{全データ数}} \times 100 \tag{5}$$

$$FAR = \frac{\text{無効入力を有効入力として誤分類した数}}{\text{無効入力のデータ数}} \times 100 \tag{6}$$

$$FRR = \frac{\text{有効入力を無効入力として誤分類した数}}{\text{有効入力のデータ数}} \times 100 \tag{7}$$

5.3 実験 1：各特徴量の性能

5.3.1 実 験 方 法

実験 1 では 4 節で述べた特徴量を用い，SVM を用いて有効入力と無効入力の識別を試み

表 5 SVM による有効入力と無効入力の識別結果
Table 5 Result of invalid input identification by SVM

		CER(%)	FRR(%)	FAR(%)
従来手法（GMM の最大音響尤度による識別）		23.30	13.96	32.34
SVM による識別	GMM	**19.59**	10.97	27.93
	BOW(Wordlist:すべての入力)	**15.94**	18.14	13.81
	BOW(Wordlist:有効入力のみ)	**15.90**	18.11	13.76
	BOW(Wordlist:無効入力のみ)	**16.20**	18.38	14.10
	発話時間	**22.44**	17.93	26.81
	SNR	32.20	25.62	38.57

る．3 節で述べた「大人」発話,「子供」発話,「無効発話」,「咳」,「笑い声」,「雑音」の 6 クラスの GMM を用いた無効入力の識別手法を従来手法として考え，SVM による識別手法の結果と比較する．実験条件は表 4 の通りである．

5.3.2 実 験 結 果

実験 1 の結果を表 5 に示す．SVM を用いた識別においては，SNR を特徴量とした SVM を除くすべての場合で CER を従来手法よりも低く抑えられた．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 CER を 23.30% から 15.90% に改善できた．また GMM の音響尤度を特徴量とした SVM と比べても CER を 19.59% から 15.90% に改善できている．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．

また BOW に着目すると，3 種類の Wordlist の内，有効入力から作成した Wordlist による BOW の CER がもっとも低かった．ただし，これらの差は小さく，各 Wordlist に含まれている単語も類似していたことから，どれを用いても問題はなかったと思われる．なお，以降の BOW では有効入力から作られた Wordlist を用いている．

5.4 実験 2：複数の特徴量の組み合わせ

5.4.1 実 験 方 法

マルチカーネル法を用いて複数の特徴量を組み合わせた無効入力の識別を行う．今回は特徴量を加えるごとに，識別精度がどのように変動していくのかを見るため，実験 1 において有効であった特徴量を順番に加えていく実験方法を採用した．なお，実験条件は実験 1 と同じく表 4 の通りである．

5.4.2 実 験 結 果

実験 2 の結果を図 3 に示す．本実験では「BOW，GMM，発話時間，SNR」のすべての特徴量を用いた結果が最良であり，CER を 13.60% に抑えることができた．これは従来手法と比較すると 9.7 ポイント削減できている．また, 表 6 に学習データ，テストデータの CER を比較した結果を示す．GMM の尤度を用いた場合は 5% 程度だった学習データとテストデータの CER の差が BOW を用いると 8% 程度に上がっているため，過学習が起こっている可能性が考えられる．この影響を減らすためには学習データを増やすことが効果的である．

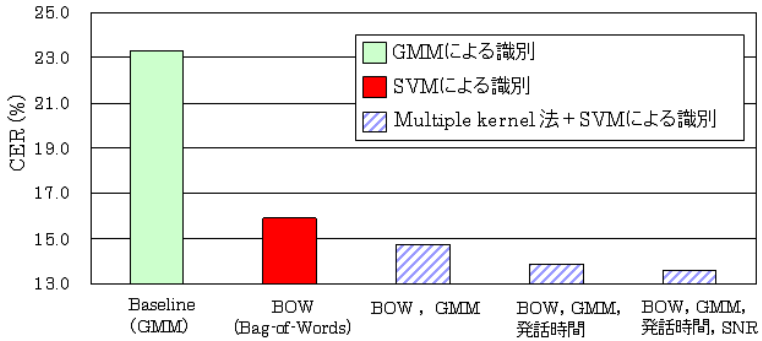

図 3 マルチカーネル法を用いた場合の識別性能
Fig. 3 Result of invalid input identification by Multiple kernel

表 6 学習データとテストデータの CER
Table 6 CER of training data and test data

	学習データ	テストデータ
従来手法（GMM を用いた識別）	18.83	23.30
SVM（GMM）	14.29	19.59
SVM（BOW）	7.71	15.90
SVM（BOW, GMM, 発話時間, SNR）	5.14	13.60

5.5 考　　察

本実験では CER による識別精度の評価を行ってきたが，FAR と FRR にばらつきがあり，表 5 における従来手法と BOW を特徴量にした SVM の FRR を比較すると BOW の方が劣化している．そこで，4 節で説明したコストパラメータ C_+, C_- を調整し，この FAR と FRR のバランスを変動させ，FAR と FRR の両方を削減することができないか調べる．コストパラメータ C と C_+, C_- を式 (8) によって表し，パラメータを変動させることによって得られた曲線を図 4 に示す．

Vol.2010-SLP-80 No.3
2010/2/12

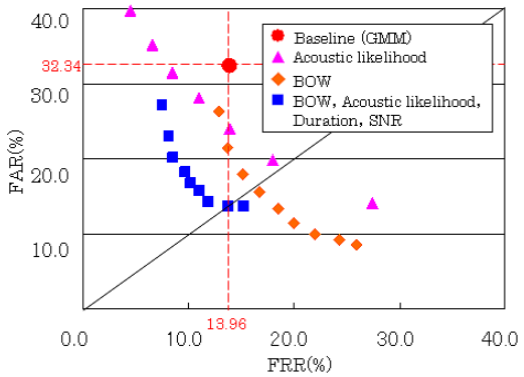

図 4 重みをかけた場合における FAR と FRR の変化
Fig. 4 Transition of FAR and FRR using cost-parameter

$$C = C_+ + C_- \tag{8}$$

今回は従来手法と GMM を特徴量にした SVM，BOW を特徴量にした SVM，「BOW, GMM，発話時間，SNR」を特徴量にした SVM の比較を行う．図 4 により FRR を従来手法と同程度に固定した場合であっても，他の手法の FAR は改善できていることが確認された．また，FAR と FRR が等しくなる点においても，「BOW, GMM，発話時間，SNR」を特徴量にした SVM の結果がもっとも良い．

次に各カテゴリ別に見た CER を表 7 に示す．GMM の音響尤度を用いた従来手法と SVM では，「意味のない発話」や「音声区間検出ミス」における CER が 50% 前後だったが，BOW を用いることにより 20% 程度に改善できている．また，従来手法と BOW を用いた SVM を比較すると，BOW を用いることにより「有効入力」，「オーバーフロー」，「咳」の誤りが増えていたが，これは音声認識誤りによるものだと考えられる．なお，従来手法と「BOW, GMM，発話時間，SNR」を特徴量にした SVM を比較すると，CER が悪化しているのは「オーバーフロー」のみであり，残りはすべて改善されていた．

6. ま と め

無効入力を識別する手法として BOW を特徴量とした SVM を提案し，GMM による識別手法と比較して CER を 23.30% から 15.90% に改善できた．また，複数の特徴量を組み合わせることにより識別性能を高め，CER を 13.6% まで減少させることができた．

表 7 各カテゴリごとの分類誤り率の傾向調査
Table 7 Classification error rate of each category

			テストに含まれるデータ数	従来手法 (GMM)	SVM GMM	SVM BOW	Multiple Kernel+SVM BOW, GMM, 発話時間, SNR
有効入力			3782	13.96	10.97	18.11	10.60
無効入力	無効発話	背景会話	926	25.49	20.30	16.41	12.85
		発話不明瞭	329	38.60	41.64	24.32	28.88
		意味のない言葉	907	50.28	47.30	19.18	28.67
		音声区間検出ミス	547	58.32	58.68	22.12	27.24
		オーバーフロー	63	19.05	38.10	42.86	46.03
		レベル不足	215	35.81	6.98	0	0
	咳		30	10.00	0	3.33	3.33
	笑い声		160	7.50	5.63	8.75	5.63
	雑音		1395	12.90	8.67	3.66	5.02

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) および文部科学省科学研究費基盤研究 (A)“新しい音声メディアによるユニバーサルコミュニケーションの研究”(研究課題番号 : 19200009) の援助を受けて行われた．

参 考 文 献

1) 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 SP2003-172, pp13-18, 2004.
2) 酒井啓行, ツィンツァレク トビアス, 川波弘道, 猿渡洋, 鹿野清宏, 李晃伸, “実環境ハンズフリー音声認識のための音響モデルと言語モデルに基づく音声区間検出と認識アルゴリズム，” 電子情報通信学会技術研究報告 SP2007-17, vol.107, pp.55-60, 2007.
3) R. Nisimura, A. Lee, H. Saruwatari, K. Shikano, “Public Speech-oriented Guidance System with Adult and Child Discrimination Capability,” *In Proc. ICASSP* 2004, pp.433-436, 2004.
4) 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, Vol.2004, No.154-163, 2005.
5) A. Lee, T. kawahara and K. Shikano, “Julius - an open source real-time large vocabulary recognition engine,” Proc. of*Eurospeech*, pp1691-1694, 2001.
6) Chih-Chung Chang, Chih-Jen Lin: LIBSVM : a library for support vector machines, http://www.csie.ntu.edu.tw/˜cjlin/libsvm.