音声/非音声区間検出のための自動モデル学習法の評価

竹内伸一[†] 杉山雅英[†]

近年映画やTV等に基づく大量のマルチメディアコンテンツが作成されており，それに基づくマルチメディアデータベースの構築が可能となっている．マルチメディアデータの一例であるテレビ番組やラジオ番組等の音響データには音声以外の音が含まれていることが多く，音声認識の際にはそれらを音声と誤認識することによる性能低下が生じる．本論文の目的はマルチメディアデータに含まれる音声区間を検出することである．音声認識を行う際に音響信号からあらかじめ音楽のみの非音声区間等の認識対象外の区間を除外することで，それらの区間による音声認識システムの誤認識を避けることができる．一般的な音声認識手法として，認識対象の音声データを用いてモデルを事前に学習する手法があり，それらが事前学習に使用するデータは教師信号を手動で作成している．本論文はモデル作成のための教師信号を自動的に生成し，学習により性能を向上させる手法を提案する．認識対象とする未学習のデータから教師信号を自動で作成することにより，事前の学習が難しいデータに対しても適用できる．また，学習済みと未学習のデータの特性が異なることによる判別性能の低下を抑えることができる．提案手法により未学習のデータに対して最大4.2%の判別誤り率で判別を行うことができた．これはデータの40%から50%を手動で教師信号を作成したものと同程度の性能であり，提案手法により作業量を大幅に削減することが可能となった．

Evaluation of Automatic Model Training Method for Voice/Non-voice Classification

Shin'ichi Takeuchi[†] and Masahide Sugiyama[†]

The purpose of this paper is to detect voice section of these multimedia contents. By removing unnecessary section (noise, music, and so on) from sound data, speech recognition techniques can prevent miss recognition. This section detection technique can use pre-process of speech recognition techniques. This paper investigates voice/non-voice discrimination method and its automatic training method. By creating teaching signal from target data, it is able to apply recognition techniques to some data that it is difficult to do previous training. Also it is able to prevent reduction of effectiveness by the difference of training data and testing data. As a result, proposed method shows 4.2% of classification error rate. The efficiency of voice/non-voice model which is created by proposed method equals to model which is created by about 40% to 50% of manually created teaching signal and proposed method can reduce manual creating work of teaching signal.

1. はじめに

近年大量のマルチメディアコンテンツが作成されており，それらを集めたマルチメディアデータベースの構築が可能となっている．マルチメディアデータの一例であるテレビ番組やラジオ番組等の長時間データには音声以外の音が含まれることが多いため，音声認識の際にはそれらを音声と誤認識することによる性能低下が生じる．この問題に対応するための1つの方法として，音響信号内の非音声区間を認識対象から除外する，ということが考えられる．あらかじめ音声認識に不必要な区間を除くことで，従来の認識手法の障害となっている対象以外の音による誤認識を避けることができる．このような着目点から音響信号中の音声区間検出は近年活発に研究されている[1)~3)]．

本論文の目的はマルチメディアコンテンツに含まれる音声区間と非音声区間とを判別する手法を提案しその評価を行うことである．我々は音声/音楽区間やそれを拡張した音声/非音声区間の判別を目的とする研究を行っており，これまでに音声/音楽の判別に効果的な特徴量Block Cepstrum Flux（BCF）を提案し[4)]，その有効性を示してきた[5)]．本論文では，これまで提案してきたBCFを用いたモデルの自動作成および自動学習法について述べる．

文献1)や3)において提案されている手法は音響的

† 会津大学大学院コンピュータ理工学研究科
Graduate School of Computer Science and Engineering, University of Aizu

なモデルを作成するために事前に教師信号を与える．テレビ番組等に含まれる音声はつねに一定の条件下で発声されるわけではなく，学習済みデータと未学習のデータの特性が異なることによる判別性能の低下は避けられない．本論文で用いる BCF は事前学習なしで音声/非音声の判別がある程度可能である．本論文では未学習のデータに対して BCF を用いて音声/非音声の予備選択を行い，その結果の信頼区間を教師信号としてモデルを学習する．さらに判別結果の信頼区間を用いてモデルの再学習を行う．これにより学習時と評価時のデータの特性が異なることによる性能の低下を抑えることができる．また，事前の学習が難しいデータに対しても適用可能となる．本論文ではモデル化手法として VQ 識別器および GMM（混合ガウス分布モデル）を用いたが，他手法でも同様の学習が可能である．学習が容易な特徴量 BCF と性能が高いモデルに基づく手法を併用することで両手法の特徴を有効に活用できる．

本論文が提案する手法の応用例としては音声認識の前処理だけではなく，さらにマルチメディアデータ内の音声区間だけを高速で再生する時間圧縮音声再生法[6] が考えられる．HDD レコーダ等の普及にともない，大量のマルチメディアデータの蓄積が可能となっている．データの蓄積は専用機械によって自動的に行われる一方，データの視聴は人間が行わざるをえないため大量のデータを視聴する場合には視聴速度の向上が必要となる．映像の高速再生は一定間隔でフレームを飛ばす等の手法で比較的容易に行えるが，内容の理解を第 1 として視聴する場合には音声区間が重要となる．提案手法により音声区間のみを抽出することで，効果的に高速視聴を行うことが可能となる．

本論文は以下のように構成されている．2 章では音声/非音声の判別手法とモデルの自動学習について述べる．3 章では評価実験について述べ，4 章をむすびとする．

2. 音声/非音声の判別手法

この章では音声/非音声の判別手法について述べる．モデル作成に必要な教師信号を自動作成する方法を提案する．すでに提案した BCF を述べ，それを用いた教師信号の作成法について述べる．また，モデルに基づく音声/非音声の判別法と自らの結果に基づく自動学習について述べる．

2.1 教師信号の自動作成

Block Cepstrum Flux（BCF）は Cepstrum Flux から計算される．Cepstrum Flux は Spectral Flux[7] を時間軸にそって拡張したものであり，式 (1) で表される．対象となるフレームと過去 J フレームそれぞれとのケプストラムベクトルのユークリッド距離の 2 乗値を平均化したものであり，$J = 1$ のとき Spectral Flux と等しくなる．Cepstrum Flux はフレームごとに計算され，n はフレーム番号を示している．

$$D_n(J) = \frac{1}{J}\sum_{j=1}^{J} d^2(\boldsymbol{c}_n, \boldsymbol{c}_{n-j}). \tag{1}$$

ここで $\boldsymbol{c}_n$ は n フレーム目のケプストラムベクトル，$d^2(\boldsymbol{c}_n, \boldsymbol{c}_{n-j})$ はケプストラムベクトル間のユークリッド距離の 2 乗値を表す．

BCF は Cepstrum Flux をさらに一定の時間窓内で平均化して計算され，時間窓を F としたとき，m 番目の区間に対する BCF は式 (2) によって表される．

$$B_m(F) = \frac{1}{F-J}\sum_{i=mF+J}^{(m+1)F-1} D_i(J). \tag{2}$$

$J = 3$，$F = 7$ のときの Cepstrum Flux と BCF の関係を**図 1** に示す．丸は各時刻ごとのケプストラムベクトルである．丸と丸を結ぶ曲線は 2 つのベクトル間距離を求めることを示している．図では $F - J = 4$ 個の D_n の値を平均化することになる．

一般的に音声区間では音楽等の区間に比べてスペクトルの変化が大きいため，区間中の平均変化量を表す BCF の値は非音声区間での値に比べて大きな値になる．したがって適切な閾値 T を設定することで，$B_m(F)$ の値が閾値よりも大きいときに音声，小さいときに非音声区間と判別することができる．**図 2** に音声/非音声区間から計算された BCF の分布を示す（用いたデータは 3.1 節で後述するデータ B である）．音声であっても長母音のように定常音が長く持続する場合には BCF は小さな値となるので，図からも分か

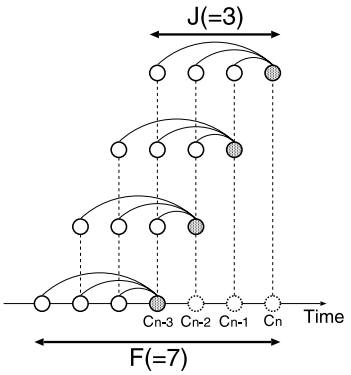

図 1 Cepstrum Flux と Block Ceptrum Flux
Fig. 1 Cepstrum Flux and Block Cepstrum Flux.

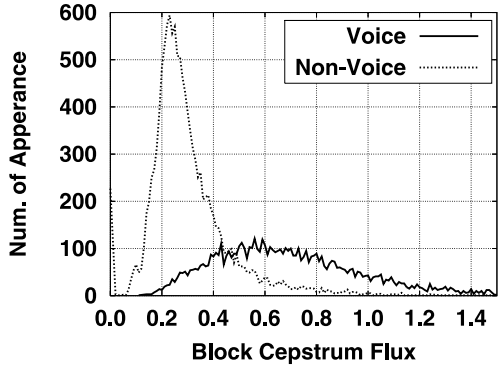

図 2 音声/非音声区間から計算される BCF の値
Fig. 2 Distribution of BCF calculated by voice/non-voice.

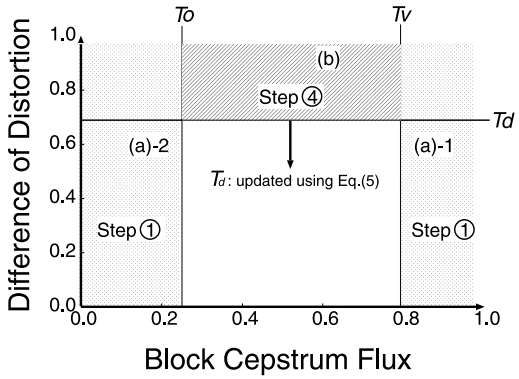

図 3 教師信号に用いる領域
Fig. 3 Data section used to teaching signal.

るように音声/非音声の 2 つの分布を 1 つの閾値で完全に分離できるということはない．

後述するモデルによる音声/非音声判別手法で用いる教師信号を BCF を用いて作成する．窓長 F ごとに求められた BCF の信頼性の高い区間のみを用いて音声/非音声用の教師信号とする．**図 3** に教師信号の作成に用いる BCF の値の範囲を示す．横軸が BCF の値，縦軸が式 (4) で示される d_m である．縦軸および領域 (b) については教師信号の更新の際に用いるので 2.3 節で後述する．領域 (a) は BCF の値のみによって決定され，音声検出用の閾値 T_v および非音声検出用の閾値 T_o を用いて，入力フレームを音声（領域 $(a)-1$），非音声（領域 $(a)-2$），その他（T_v と T_o の中間領域）の領域に分ける．図 2 に示したように BCF の値だけでは音声/非音声が混在する場合があるため，T_v より大きい領域 $(a)-1$ を信頼できる音声区間，T_o より小さい領域 $(a)-2$ を信頼できる非音声区間と見なし，次のモデル作成の際の教師信号として用いる．

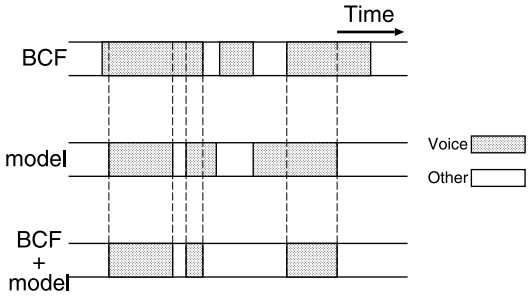

図 4 音声と判別される区間
Fig. 4 The section considered as voice section.

2.2 音声/非音声のモデルに基づく判別手法

式 (3) の 2 つの条件を同時に満たした区間を音声と判別する．

$$\begin{cases} B_m(F) > T \\ \dfrac{1}{F}\displaystyle\sum_{i=mF}^{(m+1)F-1}\{\min_{\boldsymbol{v_0}\in\boldsymbol{V_0}} d^2(\boldsymbol{c}_i,\boldsymbol{v}_0) \\ \qquad\qquad -\min_{\boldsymbol{v_1}\in\boldsymbol{V_1}} d^2(\boldsymbol{c}_i,\boldsymbol{v}_1)\}<0 \end{cases} \tag{3}$$

ここで $\boldsymbol{V}_0$ は $\boldsymbol{v}_0$ を要素とする音声用モデルを，$\boldsymbol{V}_1$ は $\boldsymbol{v}_1$ を要素とする非音声用モデルを示す．式 (3) の閾値 T は拒否誤りを防ぐため予備実験から $T=0.3$ とした．この条件によって大部分の非音声区間が除外される．$T=0.3$ は環境に依存するためすべてのデータに対して最適な値ではないが，予備実験の結果から一定の値を用いる場合としては良好な性能を示すことが分かっている．最適値の事前推定に関しては今後の検討課題とする．

残った区間に対して下式でモデルを用いて判別を行う．音声用および非音声用のモデルを，BCF の結果から得られた教師信号によって学習する．BCF と同じ窓長 F ごとに区切られた入力データに対して，窓内の各点とそれぞれのモデルとの歪みを計算し，窓内での平均歪みを求める．歪みがより小さいモデルを判別結果とするので，入力データと音声用モデル $\boldsymbol{V}_0$ との歪みが非音声用モデル $\boldsymbol{V}_1$ との歪みよりも小さい区間を音声とする．

BCF およびモデルに基づく判別手法の組合せを**図 4** に示す．図中の斜線部はそれぞれの手法において音声と判別された区間を示す．BCF およびモデルによる判別手法の両方で音声とされた区間を最終的な判別結果とする．学習が容易な特徴量 BCF と性能が高いモデルによる手法を併用することで両手法の特徴を有効に活用できる．

なお，本論文では判別の際にスムージングを行って

いない．対象の音響データ中の音声長はあいづち等の短い区間が多く出現し，また音声/非音声の切り替わりも頻繁に発生する．スムージング中の多数決によって音声区間が削除されてしまう，もしくは多数決自体が意味を持たないことが考えられるためである．

2.3 モデルの自動学習

2.2 節で得られた結果を用いてモデル作成のための教師信号を更新し，モデルの自動学習を行う．本論文ではモデル化手法として VQ 識別器および GMM を用いるが，他手法でも同様に学習することが可能である．以下に処理の流れを示す．**図 5** は手順を示しており，手順の番号は図中の番号に対応している．

(1) **BCF 計算**：検出窓 F ごとに式 (2) で示した BCF を計算し，図 3 の領域 $(a)-1$ で示された範囲を音声，領域 $(a)-2$ で示された範囲を非音声とする初期教師信号を作成する．

(2) **モデル作成**：作成した教師信号に基づいて音声/非音声のモデル $\boldsymbol{V}_0$ および $\boldsymbol{V}_1$ を作成する．

(3) **音声/非音声判別**：式 (3) を満たす区間を音声区間と見なす．

(4) **教師信号更新**：手順 (3) の判別結果をもとに教師信号を更新し，手順 (2) に戻る．

手順 (3) での判別結果が BCF 単独よりも良好な性能を示すとすると，手順 (4) で作成したより精度の高い教師信号を用いることによりモデルの精度が向上し，この過程を繰り返すことにより最終的な判別結果が向上すると考えられる．手順 (2) から手順 (4) までは以下の条件で繰り返される．図 3 において，領域 (a) は更新後の教師信号でも同じように用いられる．それ以外の区間について，式 (4) に示された入力と各モデル間ごとの歪みの差 d_m を求める．

$$d_m = \frac{1}{F}\sum_{i=mF}^{(m+1)F-1}\left|\min_{\boldsymbol{v_0}\in\boldsymbol{V_0}} d^2(\boldsymbol{c}_i,\boldsymbol{v}_0) - \min_{\boldsymbol{v_1}\in\boldsymbol{V_1}} d^2(\boldsymbol{c}_i,\boldsymbol{v}_1)\right| \quad (4)$$

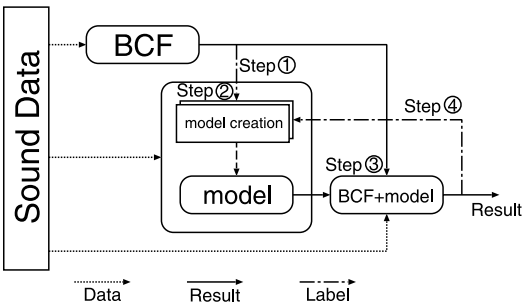

図 5 モデル学習法の流れ
Fig. 5 Model training flow.

図 3 の縦軸は d_m を示す．入力がどちらかのモデルに非常に近い場合（すなわち d_m が大きい場合），その区間は音声/非音声のどちらかに属している可能性が高い．よって閾値 T_d よりも d_m が大きい領域 (b) を追加の教師信号とする．対象の区間が音声/非音声のどちらに属するかは入力との歪みが小さい方とする．

閾値 T_d は d_m に対する閾値で，式 (5) によって定める．k 回目の更新時の閾値 $T_d^{(k)}$ は

$$T_d^{(k)} = \frac{N-k}{N}T_d \quad (0 \le k \le N) \quad (5)$$

とする．ここで N は最大学習回数である．この式に基づくと $T_d^{(k)}$ は更新ごとに減少し，モデルの作成に用いられるデータは増えることになる．モデル作成に用いる領域を更新ごとに広げることによりモデルの精度向上が期待される．

3. 評価実験

3.1 実験条件

評価には音声/非音声区間を含むデータとして 2 種類のデータベースを用いた．**表 1** にその構成を示す．データ A は会津若松市の観光案内ビデオ中の音声で，複数人の声，音楽，無音区間，その他背景情景音が含まれる．話者数は男性 5 名，女性 2 名であり，発話内容は演劇の台本に基づくセリフによるもので区切りが明瞭である．データ B は FM ラジオ番組から作成された CampusWave Database[8] で，2 名の女性パーソナリティの会話，リクエスト曲，CM，無音区間音が含まれる．本論文ではデータベース 15 セット中の 3 セットを用いた．データ長はそれぞれ約 60 分，話者数は女性パーソナリティ 2 名および CM 区間の男女話者若干名，発話内容はパーソナリティの会話および市内企業や FM 番組の CM である．データ A，データ B ともに音声と BGM が重畳している区間は音声区間と見なした．これらのデータには音声/非音声のラベルが波形の目視および実際の試聴により付与されている．

データに対する音声/非音声の定義として，実際の受聴によって音声/非音声の切り替わり点を手動で付与し，音声/非音声状態が 160 ms 以上継続している

表 1 評価用データベース
Table 1 Experimental database.

	data A	data B
データ数	1	3
データ長	19 分	各 60 分
音声区間の比率	31.3 %	40.0 %
音声区間の平均継続時間長	1.2 秒	2.4 秒

表 2 音声分析条件
Table 2 Analysis condition.

標本化周波数	16 kHz
窓長	256 点（16 ms）
更新周期	256 点（16 ms）
窓関数	Hamming 窓
高域強調	$(1-0.97z^{-1})$
特徴量	LPC ケプストラム
LPC 分析	14 次
ケプストラム分析	16 次
検出窓長 F	63 フレーム（$\simeq$ 1 秒）

区間をそれぞれの区間とした．これはデータ B において音声であると考えられる話者のあいづちが収まる最短の区間長に基づいている．160 ms 未満の区間に関しては以前の状態が継続していると見なした．

表 2 に音響分析条件を示す．ケプストラム距離計算時に一次の差分フィルタ $1-\alpha z^{-1}(\alpha=-0.8)$ を用いて低域周波数に重み付けを行った[9]．

音声/非音声のモデルとして，本論文では VQ 識別器および GMM を用いて性能比較を行った．VQ 符号帳は LGB アルゴリズムを用いて作成し，GMM は対角化を行った．

GMM による実験ではモデルの比較尺度として対数尤度[10] を用いた．モデル $\boldsymbol{V}$ を G 次元のガウス分布 M 個の混合として表したとき，要素 $\boldsymbol{v}_m=\{\lambda_m,\mu_m,\sigma_m^2\}$（混合重み，平均，共分散）で表される．尤度が大きいモデルを判別結果とするため，式 (3) のモデルに関する不等式は

$$\frac{1}{F}\sum_{i=mF}^{(m+1)F-1} d_p(\boldsymbol{c}_i,\boldsymbol{V}_0)-d_p(\boldsymbol{c}_i,\boldsymbol{V}_1)>0 \quad (6)$$

となる．このとき

$$d_p(\boldsymbol{c}_i,\boldsymbol{V})=\log\left\{\sum_{m=1}^{M}\lambda_m\prod_{g=1}^{G}\frac{1}{\sqrt{2\pi}\sigma_{m,g}}\exp\left(-\frac{1}{2}\frac{(c_{i,g}-\mu_{m,g})^2}{\sigma_{m,g}^2}\right)\right\} \quad (7)$$

である．判別は窓内がすべて音声/非音声である区間を対象にして行った．すなわち窓内で音声/ 非音声が混在する区間は評価から省いた．判別結果と正解ラベルとの比較を行い，拒否誤りと受理誤りの平均を判別誤り率として評価基準に用いた．比較のための closed なモデルは以下の手順で作成した．データ A に対しては，データを前半と後半に分割し，closed な場合は評価用と同じデータで，open な場合は異なるデータでモデルを作成した．データ B に対しては，全 3 セット中，1 セットでモデル作成を行い，closed な場合は同一セットで評価を，open な場合は他 2 セットで評価を行った．

表 3 モデル化手法の違いによる性能差（$N=1$）(%)
Table 3 Comparison of model creating method ($N=1$) (%).

	サイズ	データ A		データ B	
		VQ	GMM	VQ	GMM
	0	6.2		13.3	
B	8	10.6	5.6	13.4	13.5
C	16	5.7	5.2	11.9	12.6
F	32	5.5	4.8	10.9	12.3
	64	4.7	5.3	10.8	12.1
モデルのみ (64)		8.2	9.6	12.0	20.0

3.2 モデルの自動学習の評価

この節では提案手法により自動学習された音声/非音声のモデルの性能評価を行う．

表 3 に音声/非音声のモデル化手法の違いによる判別の性能差を示す．サイズが 0 のときの結果は BCF 単独での性能を示し，以降の各行はモデルの大きさを変化させて BCF と組み合わせた場合の判別結果を示す．最下段は式 (3) でモデルのみ（サイズ 64）を用いた場合の判別結果を示す．各サイズは VQ 符号帳サイズまたは GMM の混合数を意味する．両手法ともほとんどの場合で BCF 単独よりも良好な判別結果が得られ，その結果を用いてさらに精度の高いモデルを再作成することが可能であることが分かる．モデルのサイズが 64 の場合において，BCF とモデルを組み合わせた提案手法と VQ および GMM をそれぞれ単独で用いたときの性能を比較した場合，提案手法はモデル単独の性能よりも 3.5% から 7.9% 良好な結果を示した．一度の学習にかかる時間は VQ 識別器の約 30 秒に対して GMM では約 28 分となる．GMM を用いた場合 VQ よりも判別性能が低くなる理由としては，音楽区間中の歌声を音声として判別しているケースが VQ の場合よりも多いことがあり，歌声が会話の音声と近いようなラップ等の曲で顕著に現れる．この原因としては，音声用モデルの作成時に音声のみの区間だけでなく音声+BGM の区間も使用しているので上記の区間も音声と誤認識しているためと考えているが，同じ枠組みで行っている VQ との性能差が大きいのでこの原因については今後検討する．以降の実験では VQ 識別器を用いる．

図 6 は各データに対してそれぞれ closed な VQ 符号帳を用いた場合の VQ 識別器による誤り率，open な VQ 符号帳を用いた場合の VQ 識別器による誤り率，提案手法の初期誤り率，学習回数 $N=20$ としたときの自動学習後の提案手法の誤り率および学習し

表 4 各データに対する拒否誤りおよび受理誤り（%）
Table 4 Reject error and accept error (%).

		VQ		BCF+VQ		
		closed	open	0 回学習	20 回学習	20 学習 VQ のみ
A	受理誤り	4.6	13.6	3.4	3.4	7.0
	拒否誤り	0.8	2.7	6.0	5.0	5.0
B	受理誤り	6.8	20.1	10.6	11.9	14.3
	拒否誤り	1.6	3.2	11.0	5.9	4.0

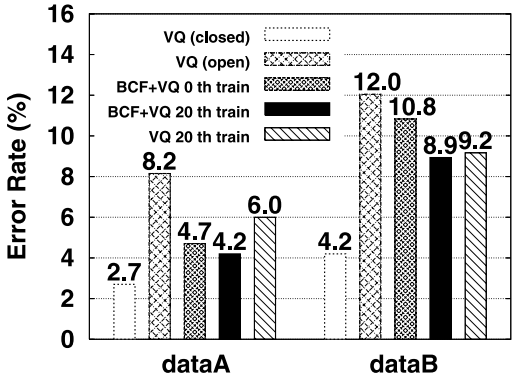

図 6 VQ 識別器による手法と提案手法の評価
Fig. 6 Comparison of VQ classifier and the proposed method.

たモデルのみを用いた誤り率を表している．符号帳のサイズは 64 とし，提案手法に用いられる各閾値は予備実験の結果から誤識別を起こさないと考えられる安全な値（$T_o = 0.4$，$T_v = 0.8$，$T_d = 0.2$）を用いた．学習は自動で行われるので，提案手法の性能は学習後の値で比較を行う．データ B は音声区間に BGM として音楽が重畳している場合が多いためデータ A に比べて誤り率が高いが，どちらのデータに対しても提案手法は学習なしでも open な符号帳を用いるより判別誤り率が低く，さらに自動学習することで性能が向上している．学習後のモデルを単独で用いた場合ではデータ A および B に対してそれぞれ 6.0%，9.2% の判別誤り率であるのに対して，BCF と組み合わせた場合はそれぞれ 4.2%，8.9% と性能が向上した．**表 4** に図 6 中の各誤り率に対する拒否誤りおよび受理誤りを示す．

表 5 は closed な VQ 符号帳と自動学習前後の VQ 符号帳との距離を示す．VQ 符号帳間の距離尺度として式 (8) に示す Hausdorff 距離を用いた．

$$d_h(\boldsymbol{V}, \boldsymbol{V'}) = \frac{1}{2}\{h(\boldsymbol{V}, \boldsymbol{V'}) + h(\boldsymbol{V'}, \boldsymbol{V})\}. \quad (8)$$

ここで $\boldsymbol{V}$ および $\boldsymbol{V'}$ はそれぞれ v および v' を符号とするサイズ M の VQ 符号帳であり

表 5 closed な符号帳と学習前後の符号帳との Hausdorff 距離
Table 5 Hausdorff distance between closed codebook and before/after training codebook.

符号帳		0 回学習時	20 回学習時
A	音声	8.9	5.0
	非音声	1.2	1.1
B	音声	0.8	0.5
	非音声	1.4	0.9

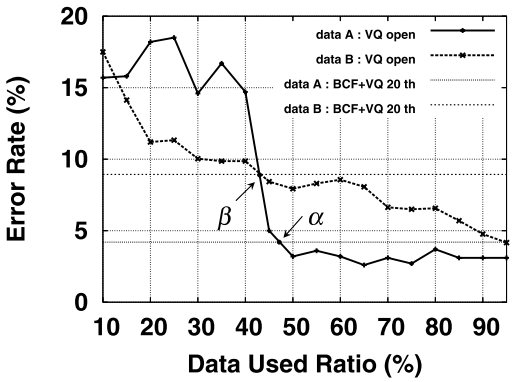

図 7 事前学習に用いるデータ量と提案手法の比較
Fig. 7 Effectiveness of VQ codebook created by part of training data.

$$h(\boldsymbol{V}, \boldsymbol{V'}) = \frac{1}{M}\sum_{v \in \boldsymbol{V}} \min_{v' \in \boldsymbol{V'}} d^2(v, v'), \quad (9)$$

である．学習によって VQ 符号帳が closed なものに近付いていることが分かる．

音声と判別された区間についてデータの切り出しを行い実際に試聴した結果，音声/非音声の判別はほぼ正しく行われていた．判別に失敗する箇所としては，音楽区間，とりわけヒップホップ等のように歌い方が発話と変わらないジャンルの音楽は判別に失敗することが多かった．

図 7 は提案手法によって得られた VQ 符号帳がどの程度の量で事前学習を行うことに対応するかを示している．図の縦軸は判別誤り率を，横軸は closed な VQ 符号帳を作る際に使用するデータ量を示し，100% のときに図 6 の closed に相当する．2 本の横線は，デー

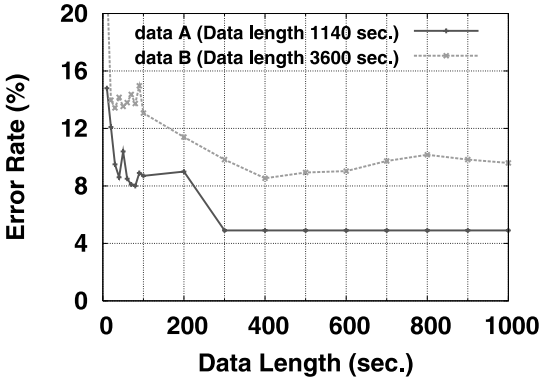

図 8 短時間学習データを用いた場合の性能
Fig. 8 Effectiveness with small training data.

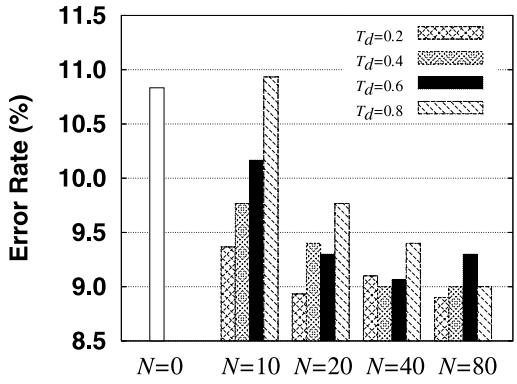

図 9 T_d と学習効果との関係
Fig. 9 Robustness for T_d.

タ A と B の $N = 20$ 時の判別誤り率である．図から，データ A では約 48% （図中の点 α），データ B で約 43% 程度（図中の点 β）のときの誤り率に一致する．これは，データ A で約 8.2 分，データ B では約 25 分のラベルを手動で付与して VQ 符号帳を学習することに相当する．モデルを事前に作成する手法では学習のために教師信号を与える必要があるが，本論文で用いたデータベースのように背景音等が混在しているデータでは目視や試聴で教師信号を付与するのは労力がかかる．提案手法を用いることによりそのようなデータに対する教師信号を人手で作成する労力を大幅に削減できることが分かった．

図 8 にデータ全体ではなく一部のみ（開始から指定の長さまで）を用いた場合の提案手法の性能を示す．図の横軸はそれぞれのモデルの学習に用いたデータ長を，縦軸は判別誤り率を示す．各実験条件は図 6 と同一のものを用いた．データ全体を用いた場合の性能は図 6 に示すとおりそれぞれ 4.2% と 8.9% であることから，音声/非音声区間がそれぞれおよそ 400 秒あれば全体で学習するのと同等の性能が得られることが分かる．

3.3 閾値と学習効果の関係

3.2 節では，閾値を予備実験に基づいて誤判別を起こさない範囲の値に設定したが，それが未学習のデータに対して有効かどうかは不明である．本節では，提案手法が用いる閾値と学習効果の関係について述べる．以下の実験はデータ量がより多いデータ B を用いて行う．

提案手法で用いる各閾値はデータに依存するために，未知のデータに対して初期の閾値として適切な値（データ B に対しては $T_d = 0.2$ 程度が適値）をあらかじめ推定することは困難である．そこでいかなる閾値に対しても提案手法が安定して動作することを確認

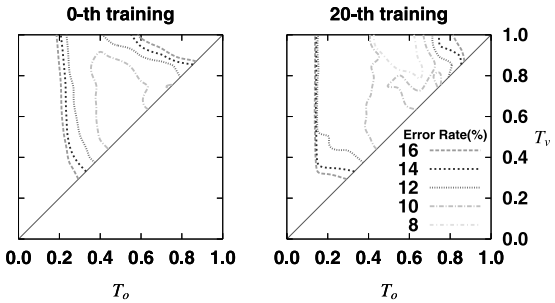

図 10 T_v および T_o と学習効果との関係 ($T_d = 0.3$)
Fig. 10 Robustness for T_v and T_o ($T_d = 0.3$).

する．**図 9** は閾値 T_d に関して学習回数 N を変化させたときの様子である．$N = 0$ が学習前の判別誤り率であり，残りはそれぞれ N を 10 から 80 としたときの学習後の判別誤り率である．例えば $T_d = 0.8$, $N = 10$ では誤り率は増加しているが，N を大きくするにつれて判別誤り率は減少することが分かる．ほとんどの場合において，学習後の誤り率は初期誤り率よりも下回っている．この図から最適値ではない T_d を用いても，十分な学習回数を用いることによって判別性能が向上することが分かる．

図 10 は閾値 T_o と T_v に関しての安定性を確認した実験結果である．縦軸を T_v, 横軸を T_o とした 2 次元の平面であり，座標が 2 つの閾値の組合せとなる．左図は学習回数 0 回，右図は 20 回目の誤り率を等高線を用いて示している．図の右下半分は $T_o > T_v$ となるので存在しない．図の中央部で最も低い誤り率を示し，端の部分の誤り率は高くなる．端の部分は T_o と T_v の組合せが極端な場合を意味し，この場合は性能が悪いが，中央部の組合せでは良い性能が得られている．学習前後の図を比較すると全体的に学習後に誤り率が減少していることが分かる．T_o および T_v に関しても，最適以外の組合せの閾値を用いてもほとんど

の場合において学習の効果が得られることが分かる．

4. むすび

本論文は音声と非音声の判別を目的として，BCFを使った教師信号の自動作成方法およびモデルの自動学習法を提案した．判別対象とするデータから自動的に教師信号を作成することで，事前学習なしでモデルに基づく判別手法を用いることを可能とした．これにより判別手法の精度向上やその活用度を高めることができた．

実験結果として，データに対して手動で 40% から 50% 程度のラベルを付与したものと同じ程度の性能である最大 4.2% の判別誤り率で判別を行うことができた．また，手法内で用いられる各閾値は，最適値以外に設定されても有効に判別できることを示した．

提案手法は判別の際に窓長を固定している．そのため，窓内に音声/非音声の区切りがある場合には対象外とした．今後の課題として雑音下での性能評価，特徴量として音声認識に有効なメルケプストラムを用いた場合の性能評価，判別結果を利用した時間圧縮音声再生法[6] 等の研究を進めていく．

謝辞 Cepstrum Flux および BCF に関してご教示いただいた浅野利明および山下昌毅両氏に感謝します．また日頃有益な討論をいただく本学ヒューマンインタフェース学講座の諸氏に感謝します．

参考文献

1) 天野明雄：耐環境騒音音声認識のための音声/非音声識別の高精度化に関する検討，音学講論，3-5-1, pp.77–78 (2005-03).
2) 山本幸一，益子貴史，田中信一：発話間の類似度を用いた教師なし話者インデキシングの検討，音学講論，2-4-2, pp.89–90 (2004-09).
3) 谷口 徹，安達了慈，大川茂樹，誉田雅彰，白井克彦：Sinusoidal Segment の時間的特徴を用いた音声・楽器音・歌声が混在した音響信号中のカテゴリ検出，音学講論，2-6-5, pp.267–268 (2005-09).
4) 浅野利明，山下昌毅，杉山雅英：Cepstrum Flux を用いた音声区間の検出，音学講論，3-Q-2, pp.121–122 (1999-09).
5) 竹内伸一，杉山雅英：音響信号に含まれる音声区間の検出，音学講論，1-P-4, pp.169–170 (2004-09).
6) 竹内伸一，杉山雅英：音声/非音声判別法を用いた時間圧縮音声再生法，音学講論，1-Q-24, pp.421–422 (2005-09).
7) Scheirer, E. and Slaney, M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, *Proc. ICASSP*, pp.1331–1334 (1997-04).
8) 内田貴之，杉山雅英：CampusWave 音声データベースの作成，電気関係学会東北支部連合大会，2A-6 (2000-08).
9) 杉山雅英，鹿野清宏：周波数軸重み付け LPC スペクトルマッチング尺度，信学論，Vol.J65-A, No.9, pp.965–972 (1982-09).
10) 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄（編）：音声認識システム，オーム社 (2001-05).

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)

竹内 伸一（学生会員）
1977 年生．2003 年会津大学大学院コンピュータ理工学研究科修士課程修了．現在，同研究科博士課程に在学中．音声認識の研究に従事．

杉山 雅英（正会員）
1954 年生．1977 年東北大学理学部数学科卒業．1979 年同大学院理学研究科数学専攻修士課程修了．同年日本電信電話公社武蔵野電気通信研究所（現 NTT 武蔵野研究センター）入所．1985 年東北大学より工学博士号を取得．1986 年から米国 AT&T Bell 研究所滞在研究員，1987 年から NTT 基礎研究所主任研究員，1990 年から ATR 自動翻訳電話研究所主幹研究員の後，1993 年から会津大学コンピュータ理工学部ヒューマンインタフェース学講座教授．現在まで，LPC スペクトル距離尺度（歪み尺度），ベクトル量子化による音声認識，特徴ベースによる音声認識，教師なし話者適応，テキスト独立話者認識，音声スペクトル推定，情報幾何学（微分幾何学）による音声分析，音響特徴量による言語識別，音声特徴キーによる音声検索等の音声認識処理の研究に従事．日本音響学会，電子情報通信学会，IEEE 各会員．