# 音種別の識別に関する一検討

渡辺 政義 †， 林 貴宏 ††， 尾内 理紀夫 † ††

† 電気通信大学大学院電気通信学研究科情報工学専攻 東京都調布市調布ヶ丘 1-5-1
†† 電気通信大学電気通信学部情報工学科 東京都調布市調布ヶ丘 1-5-1

# A study on classification of sounds

MasayoshiWATANABE † TakahiroHAYASHI †† RikioONAI † ††

† Department of Computer Science,Graduate School of Electro-Communications,The University of Electro-Communications 1-5-1 Chofugaoka,Chofu-shi,Tokyo, Japan
†† Department of Computer Science,The University of Electro-Communications 1-5-1 Chofugaoka,Chofu-shi,Tokyo, Japan

一定の時間長の音データを会話，音楽，会話+BGM，歌声+音楽の 4 種類の音種別に識別する手法を検討する．従来手法には，識別対象の音データとして，音声，音楽などの単一の音種別の音データを用いた研究や，それらが重畳した音データを用いた研究がある．しかし，識別対象とする音種別の種類が十分でない問題や精度の問題などがあり，複数の音種別が重畳している場合も，音量比を正規化して人工的に重畳した音データを使用するという制約がある．そこで，本研究では単一の音種別の音データと，重畳音の音量比が不確定な音データの識別手法について検討した．各音種別は，短時間スペクトルの時間的な変化の特徴がそれぞれ異なるため，各音種別ごとのソナグラムの特徴を反映した特徴量を考案して識別に利用した．考案した特徴量を用いて学習により識別器を作成し，識別実験を行って識別手法の有効性を検証した．さらに，音データの時間長を変化させて，音データの時間長と識別精度の関係を調査した．音データの時間長を 2 秒とした識別実験では，各音種別に対する識別結果の平均 F 尺度は 74.8%となった．

## 1 はじめに

近年，ストレージ容量の増大や通信回線の高速化に伴い，動画投稿サイトが普及し，我々が視聴できる動画コンテンツは急速に増加してきた．そのため，動画コンテンツを効率的に検索するための手法が必要とされている．

現在は，動画コンテンツに人手でキーワードやタグを付与するテキストベースの検索が実用化されているが，キーワードやタグのみで動画コンテンツの特徴を記述するには限界があり，人手によるキーワード付与のコストの問題もある．そのため，動画コンテンツの検索を効率的に行うために，動画コンテンツの画像データや音データに含まれる多くの情報を検索に利用する研究 [1] が進められている．

そこで，本研究では動画コンテンツ内の音データに着目し，音データに含まれる音の種類 (音種別) を識別することを目的とする．動画コンテンツの音データには音声，音楽，音響といった様々な種類の音が含まれ，それらの種類の音が重畳している区間も存在する．音種別の識別に関する研究は，音声と音楽の識別を行うもの [2] [3]，それに加え歌声の識別を行うもの [4]，歌声と朗読音声の識別を行うもの [5]，ソナグラムを用いて音声，音楽，ノイズを識別するもの [6]，2 つの種類の音が重畳した音データに対して識別を行うもの [7] などがある．しかし，これらの研究には識別の精度が低い問題や，識別する音種別が少ない問題などがある．さらに，各種の音データの音量を正規化して，人手で重畳した音データを対象としているため，重畳音の音量比が不確定である一般的な動画コンテンツに単純にそのまま適用することは難しい．

そこで本研究では，単一の音種別である会話，音楽と，重畳音である会話+BGM，歌声+音楽の 4

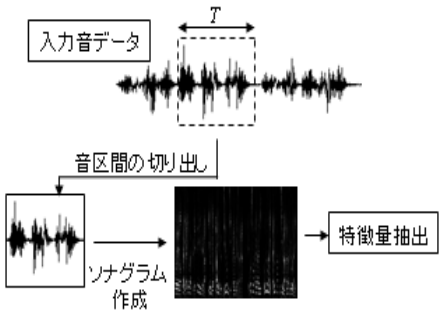


図 1　特徴量抽出までのプロセス

種類の音種別の識別を行う．識別に用いる特徴量は，各音種別のソナグラムの特徴をもとに考案する．重畳音の音種別の音データは音量比が不確定なものを使用し，音量比が不確定な重畳音の識別を目指す．以下，2 章で識別対象とする音種別とその特徴について，3 章で音種別の識別に用いる特徴量について，4 章で評価実験について，5 章で関連研究について述べ，6 章でまとめる．

## 2　識別対象とする音種別とその特徴

入力音データから特徴量抽出までのプロセスを図 1 に示す．本章では，音区間の切り出しとソナグラムについて述べ，識別対象とする音種別とその特徴についても述べる．

### 2.1　入力音データの音区間への分割

本研究では入力音データから一定の長さ ($T$[sec]) の時間区間の音データを切り出し，その時間区間の音データの音種別を識別する．以降，切り出した音データを音区間と呼ぶ．識別対象とする音種別は会話，音楽，会話+BGM，歌声+音楽の 4 種類である．歌声のみや歌声と会話が重畳している音種別は対象としない．音種別が会話である音区間を会話区間と呼び，同様にその他の音種別の音区間も音楽区間，会話+BGM 区間，歌声+音楽区間と呼ぶ．

### 2.2　ソナグラム

本研究では，切り出された音区間のソナグラムを作成し，特徴量の抽出に利用する．ソナグラムとは，音データに短時間フーリエ変換 (STFT) を行って得られる短時間スペクトルの時系列データである．ソナグラムを可視化した画像 (以降単にソナグラムと呼ぶ) は，横軸に時間，縦軸に周波数をとり，ある時刻のある周波数のパワーを色の濃度で表す．ソナグラム作成の詳細については 3 章で述べる．

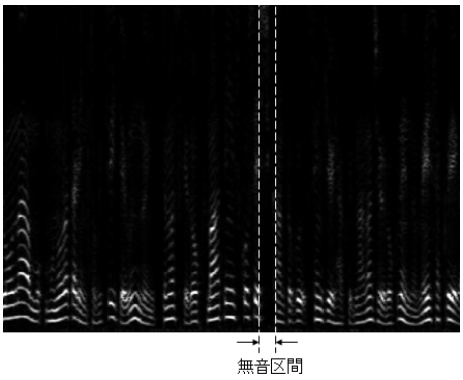


図 2　会話区間のソナグラム

ソナグラムの例として，図 2 の会話区間のソナグラムを見ると，白色の曲線を観測することができる．白色の画素は濃度が高いので，その画素に対応する時刻の短時間スペクトルの，対応する周波数成分のパワーが大きいことを示している．

### 2.3　会話区間の特徴

会話区間は人間の発話音声のみが含まれる音区間である．ニュースキャスターの朗読音声や会話などがこの音区間にあたる．人間の音声には悲鳴，歓声，笑い声，泣き声といった音響に近いものもあるが，これらの種類の音声は識別対象としない．また，3 人以上の複数話者の会話も対象外とする．会話区間は，イントネーションや音節の変化により音の大きさ，高さ，音色が短時間で変化するという特徴や，無音区間が音節間に発生するという特徴がある．

図 2 に約 5 秒の会話区間のソナグラムを示す．会話区間は音高などが短時間で変化するため，音楽に比べて短時間スペクトルの変化が大きい．その結果，図 2 のように会話区間のソナグラムでは，短時間スペクトルのスペクトルピーク (白線) が緩やかなカーブを示す．また，図 2 の点線で囲まれた区間のように，音節間の無音区間も確認できる．

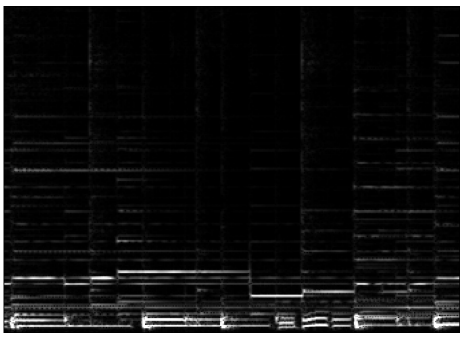

図 3　音楽区間のソナグラム

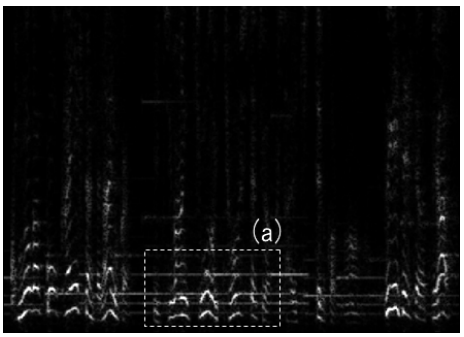


図 4　会話+BGM 区間のソナグラム

### 2.4　音楽区間の特徴

音楽区間は楽器音のみが含まれる音区間である．単独楽器の演奏のみではなく，複数楽器による演奏も含まれる．ただし，打楽器のみで構成されるものや歌声は対象としない．音楽区間は，同じ音高の音が一定時間持続するという特徴や会話区間に比べて無音区間が少ないという特徴がある．

図 3 に約 5 秒の音楽区間 (楽器はピアノ) のソナグラムの例を示す．音楽区間は同じ音高の音が一定時間持続するため，音楽区間のソナグラムでは，スペクトルピークが時間軸と水平な線分として現れていることが確認できる．また，会話区間と比べて，無音区間が少ないことも確認できる．

### 2.5　会話+BGM 区間の特徴

会話+BGM 区間は人間の発話音声と背景音楽 (BGM) が重畳している音区間である．ドラマや映画の 1 シーンやトーク番組などに多く現れる．重畳している発話音声と背景音楽は 2.3 節，2.4 節で述べた条件を満たすものとする．発話音声と背景音楽の音量比は固定しない．

図 4 に約 5 秒の会話+BGM 区間のソナグラムの例を示す．会話+BGM 区間は発話音声と背景音楽が重畳している音区間であるため，会話+BGM 区間のソナグラムは 2.4 節で述べたような水平線分の上に，2.3 節で述べた会話区間のソナグラムが重なった図となる．図 4 の (a) 部分を見ると，会話区間の成分であるスペクトルピークの緩やかなカーブと音楽区間の成分である水平線分が共に含まれていることが確認できる．また，音楽区間の特徴が含まれるため，無音区間が少ないことも確認できる．

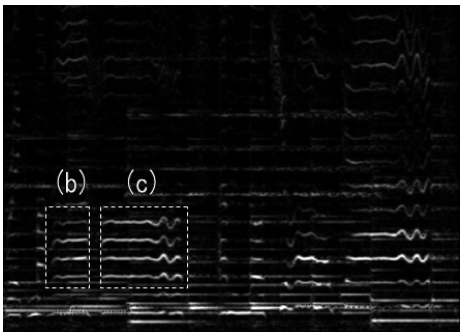


図 5　歌声+音楽区間のソナグラム

### 2.6　歌声+音楽区間の特徴

歌声+音楽区間は人間の歌唱音声と音楽が重畳している音区間である．ポップスやロックといった一般的なアーティストのヴォーカル付の楽曲などがある．歌唱音声と音楽は，同じ音高の音が一定時間持続する特徴や，発話音声に比べて無音区間が少ないという特徴が一致する．ただし，歌唱音声は，発声時に微小な音高の変化が生じたり，ビブラートによる周期的な音高の変化が生じることがある．

図 5 に約 5 秒の歌声+音楽区間のソナグラムの例を示す．歌声+音楽区間では歌声，音楽共に同じ音高の音が一定時間持続することが多い．したがって，同様に重畳音の音区間である会話+BGM 区間に比べ，スペクトルピークの変動は小さい．図 5 の (b) の部分は歌声の成分を示しており，発声時

の微小な音高の変化によるスペクトルピークの微小な変化を観測できる．図 5 の (c) の部分も歌声の成分を示しており，右端のスペクトルピークの周期的な変化はビブラートによるものである．会話+BGM 区間と同様に音楽区間の特徴である水平線分も存在していることが確認できる．

## 3 音種別の識別に用いる特徴量

本章では，音種別の識別に用いる特徴量について述べる．

### 3.1 ソナグラムの作成

本研究におけるソナグラムの作成の詳細について述べる．

対象とする音データのサンプリング周波数 $f_s$ は 16000[Hz] とする．この音データに対して，解析窓 (フレーム) 長 $T_w$ を 1024 点，シフト幅 $T_s$ を 256 点として STFT を行う．窓関数にはハミング窓を用いた．

STFT により，時刻 $(T_s/f_s)n$[sec] $(n = 0, 1, 2, .., N)$ における，周波数方向の分解能が 512 段階の短時間スペクトル

$$\boldsymbol{s}^{(n)} = (s_0^{(n)}, s_1^{(n)}, \cdots, s_{511}^{(n)})$$

を得る．以降，STFT の解析時刻を表す $n$ をフレーム番号と呼び，$n$ の最大値 $N$ をフレーム数と呼ぶ．$s_m^{(n)}$ はフレーム番号 $n$ における周波数 $(f_s/T_w)m$[Hz] $(m = 0, 1, 2, .., 511)$ の成分のパワーである．

$s_m^{(n)}$ を用いて，ソナグラムの座標 $(x, y)$ の画素の濃度 $P(x, y)$ を次式により定義する．

$$P(x, y) = s_y^{(x)} (0 \leq x \leq N, 0 \leq y \leq 511) \quad (1)$$

フレーム数 $N$ は，識別対象とする音データのサンプル点の総数を $S$ とすると

$$N = \left\lfloor \frac{S}{T_s} - \frac{T_w}{T_s} \right\rfloor$$

として算出できる．ただし，$\lfloor v \rfloor$ は実数 $v$ を超えない最大の整数を表す．

0.5 秒の音区間を対象として STFT を行う場合，音データのサンプル点の総数は $f_s$[点/sec] $\times$ 0.5[sec] = 8000[点] なので

$$N_{0.5} = \left\lfloor \frac{8000}{256} - \frac{1024}{256} \right\rfloor = 27$$

となり，0.5 秒の音区間のソナグラムは，フレーム番号 $0 \leq n \leq 27$ における短時間スペクトル $\boldsymbol{s}^{(n)}$ を時系列に表示した画像となる．

### 3.2 Spectral Flux を用いた特徴量 $f_{sf}$

SpectralFlux[8] は音声と音楽の識別に用いられる既存の特徴量であり，時間軸上で隣接するフレーム間の短時間スペクトルのユークリッド距離で定義される．本研究では，得られた SpectralFlux に対し，短時間スペクトルのエネルギーによる正規化を行う．フレーム番号 $n$ における SpectralFlux を $a(n)$ とすると，$a(n)$ は次式によって定義される．

$$a(n) = \sum_{m=0}^{511} |s_m^{(n)} - s_m^{(n-1)}| / e(n)$$

ただし，$e(n)$ はフレーム番号 $n$ の短時間スペクトルのエネルギーで

$$e(n) = \frac{1}{2} \sum_{m=0}^{511} (|\boldsymbol{s}_m^{(n)}| + |\boldsymbol{s}_m^{(n-1)}|)$$

で定義される．全ての $n$ に対し，$a(n)$ を総和し特徴量 $f_{sf}$ とする．すなわち

$$f_{sf} = \sum_{n=1}^{N} a(n) \quad (2)$$

である．

### 3.3 音楽成分の画素数 $f_m$

2.4 節で述べたように，音楽のソナグラムは水平な線分を多く含む．そこで，水平線分を検出するために，ソナグラムのある行 $y$ に対して，左端 $(x = 0)$ から右端 $(x = N)$ に向かって走査する．そして，濃度 $P(x, y)$ (式 (1)) がソナグラム内の最大濃度の 1/50 以上となる最初の画素を，水平線分の開始点 $l$ とする．開始点 $l$ からある点 $x$(ただし $l \leq x$) までを水平線分の範囲 $[l, x]$ としたとき，$[l, x]$ 間の画素の平均濃度を $\mu(l, x)$ とし，以下の条件

$$|\mu(l, x) - \mu(l, x + 1)| < 0.1 \times \mu(l, x)$$

が成立すれば，水平線分の範囲を $[l, x + 1]$ に伸ばす．この操作を上記の条件が成立しなくなるまで繰り返す．上記の条件が成立しなければ，位置 $(x+1)$ から $(x = N)$ に向かって再び開始点の探索を行い，

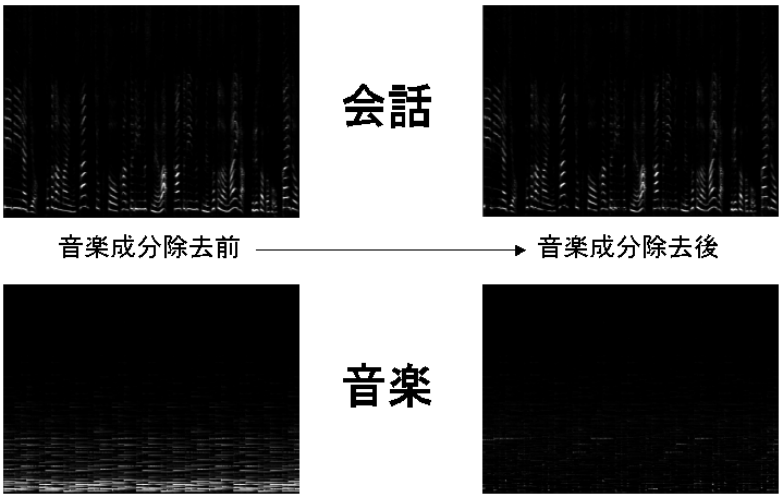


図 6　ソナグラムの音楽成分除去

同様に水平線分の検出を繰り返す．検出された水平線分のうち，長さが 4 未満のものを取り除く．

水平線分として検出された画素数を全行で総和し，特徴量 $f_m$ とする．重畳音の音量比の変化に対応可能にするため，水平線分として検出された画素の濃度ではなく，画素数を用いて特徴量を抽出する．すなわち

$$f_m = \sum_{y=0}^{511} \sum_{x=0}^{N} \sum_{j=1}^{L(y)} b(x,y,j) \qquad (3)$$

とする．ただし，$b(x,y,j)$ は

$$b(x,y,j) = \begin{cases} 1 & (l_y^{(j)} \leq x \leq r_y^{(j)}) \\ 0 & (l_y^{(j)} > x \ , \ r_y^{(j)} < x) \end{cases}$$

である．$L(y)$ はソナグラムのある行 $y$ に存在する水平線分の数を表し，$[l_y^{(j)} \ , \ r_y^{(j)}]$ は $y$ 行の左から $j$ 番目の水平線分の範囲を表す．

上述した手法を用いて水平線分を検出し，ソナグラムから音楽成分を除去したものを図 6 に示す．図 6 より，音楽区間のソナグラム (下) の白色画素数は減少しているが，会話区間のソナグラム (上) の白色画素数はほぼ変化していないことが確認できる．したがって，本手法による水平線分の検出は，音楽成分の抽出に有効であると期待できる．

### 3.4　2kHz 以上の高濃度画素数 $f_h$

ソナグラムの高周波領域は，音声や楽器音などの倍音成分を多く含む．倍音のエネルギーは SpectralFlux などの他の特徴量にも影響を与える．その影響を考慮するため，2000[Hz] 以上の周波数成分のうち，ソナグラム内の最大濃度の 1/20 以上の濃度を示す画素の総数を特徴量 $f_h$ とする．すなわち

$$f_h = \sum_{x=0}^{N} \sum_{y=128}^{511} c(x,y) \qquad (4)$$

とする．ただし，$c(x,y)$ は

$$\begin{aligned} c(x,y) &= \begin{cases} 1 & (P(x,y) \geq P_{max}/20) \\ 0 & (P(x,y) < P_{max}/20) \end{cases} \\ P_{max} &= \max\{P(x,y) \,|\, 0 \leq x \leq N \ , \ 0 \leq y \leq 511\} \end{aligned}$$

である．$P(x,y)$ (式 (1)) はソナグラムの座標 (x,y) の画素の濃度である．

### 3.5　隣接スペクトルピークの差分の総和 $f_p$

短時間スペクトルの対数をとった対数短時間スペクトルに対し，周波数軸方向にガウスフィルタをかけて平滑化したスペクトルを作成する．フレーム番号 $n$ における平滑化スペクトルを $\boldsymbol{g}^{(n)}$ とする．

$$\boldsymbol{g}^{(n)} = (g_0^{(n)}, g_1^{(n)}, \cdots, g_{511}^{(n)})$$

$\boldsymbol{g}^{(n)}$ において，パワーが最大値を示す周波数を $p(n)$ とし，次式で特徴量 $f_p$ を定義する．

$$f_p = d_{sum}/k_{sum} \quad (5)$$

ただし，$d_{sum}$，$k_{sum}$ は

$$\begin{aligned} d_{sum} &= \sum_{n=1}^{N} d(n) \\ d(n) &= \begin{cases} |p(n)-p(n-1)| & (|p(n)-p(n-1)| \leq 3) \\ 0 & (|p(n)-p(n-1)| > 3) \end{cases} \\ k_{sum} &= \sum_{n=1}^{N} k(n) \\ k(n) &= \begin{cases} 1 & (|p(n)-p(n-1)| \leq 3) \\ 0 & (|p(n)-p(n-1)| > 3) \end{cases} \\ p(n) &= \underset{m}{\operatorname{argmax}}\, g_m^{(n)} \end{aligned}$$

である．ここで，$f_p$ は，連続したスペクトルピークにおける，隣接フレーム間の $p(n)$ の差の絶対値の平均である．$p(n)$ と $p(n-1)$ の差が 3 ピクセルより大きい時はスペクトルピークは連続していないとみなし，カウントしない．短時間スペクトルの時間変化が大きな会話区間において，高い値を示すと考えられる．

## 4 評価実験

識別に用いる特徴量の有効性の検証，音区間の時間長の変化に伴う識別精度の変化の調査，誤識別が多く発生する音種別の調査を目的として評価実験を行った．

### 4.1 実験方法

実験に用いる音データとして，会話，音楽，会話+BGM，歌声+音楽の各音種別が単独で含まれる 5 秒間の音データを各種 100 ファイルずつ用意した．音源は Web 上の動画やラジオから収集した．表 1 に，収集した音データの例を示す．

実験は，5 秒間の音データの中から時間長 $T$[sec] の音区間をランダムに切り出して行った．識別結果の評価には情報検索の評価尺度として使用されることがある F 尺度 [9] を用いた．$T$ は 0.5 秒から 4 秒まで 0.5 秒刻みで変化させる．切り出した音区間の特徴量を抽出し，得られた各種 100 個ずつ全 400 個の特徴量データを 1 セットとする．

表 1　実験用音データ例

| 音種別 | 内容 |
|---|---|
| 会話 | podcast , ニュース, ラジオ |
| 音楽 | ギター, ピアノ, バイオリン, バンド演奏 |
| 会話+BGM | ドラマ・アニメの 1 シーン, トーク |
| 歌声+音楽 | J-POP, J-ROCK |

各 $T$ に対し，決定木の識別器による識別実験を行った．識別器の学習には C4.5 アルゴリズムを用い，交差検定法により識別結果の F 尺度を算出した．交差検定法のデータ分割数は 10 とした．すなわち，ランダムに選ばれた各種 90 個の特徴量データを識別器の学習に用い，残りの 10 個の特徴量データを識別するという操作を 10 回繰り返した．

以上の識別実験を 10 セットの特徴量データに対して行い，その平均 F 尺度を求めた．

### 4.2 結果と考察

#### 4.2.1 $T$ の変化に伴う F 尺度の変化に関する結果と考察

$T$ と識別結果の F 尺度の関係を図 7 に示す．$T$ の増加に伴い F 尺度が向上することが確認できる．特に，$T = 0.5$[sec] と $T = 1$[sec] の間では，F 尺度の平均値が 5.4%増加しており，他に比べより大きく上昇している．これは，0.5 秒という長さが十分でなかったためと考えられる．特に，音楽区間と歌声+音楽区間の識別が困難であった．

F 尺度が最も高い値を示した音区間は会話区間であり，次いで会話+BGM 区間，歌声+音楽区間，音楽区間の順となっている．音楽区間では $T = 2$[sec] から $T = 3.5$[sec] にかけて F 尺度はほぼ横這いになっており，$T = 4$[sec] で 2.7% 上昇した．

$T$ を大きくすれば，識別の精度は向上するが，時間分解能が低くなってしまう．さらに，同じ音種別の音が持続する時間も不確定であるので，$T$ はなるべく短い方が良い．したがって，$T = 0.5$[sec] から $T = 1$[sec] において F 尺度が急激に向上したことと，$T = 2$[sec] から $T = 3.5$[sec] にかけて音楽区間の識別結果の F 尺度がほぼ横這いであることから，$T$ は 1 秒〜2 秒の時間長とすることが妥当であると考えられる．

#### 4.2.2 誤識別を引き起こす音種別に関する考察

$T = 2$[sec] における識別結果を表 2 に示す．各行は音データの実際の音種別，各列は本研究の識別手法によって識別された音種別を表す．例えば，「会話」の行における「会話+BGM」の列の値 168 は 100 ファイル× 10 セット=1000 個の会話区間の

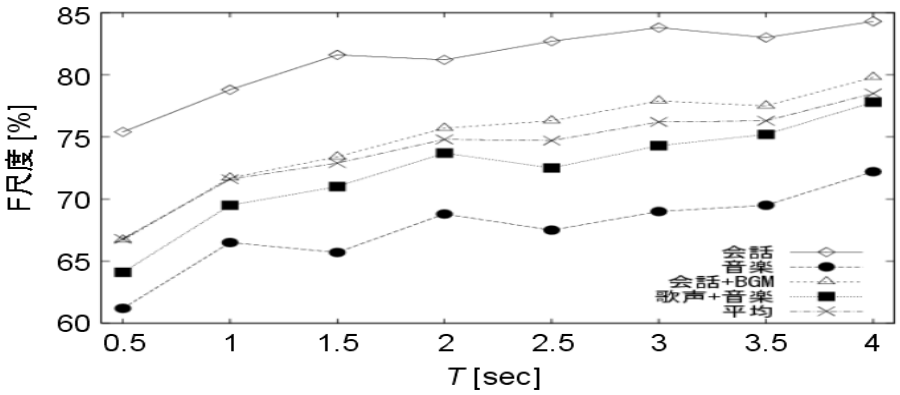


図 7　$T$ の変化に伴う F 尺度の変化

表 2　$T = 2$[sec] における識別結果

| | 会話 | 音楽 | 会話+BGM | 歌声+音楽 |
|---|---|---|---|---|
| 会話 | 820 | 6 | 168 | 6 |
| 音楽 | 7 | 606 | 23 | 364 |
| 会話+BGM | 182 | 8 | 758 | 52 |
| 歌声+音楽 | 1 | 128 | 33 | 838 |

表 3　音種別を 2 種類にまとめた識別結果

| | 会話, 会話+BGM | 音楽, 歌声+音楽 |
|---|---|---|
| 会話, 会話+BGM | 1928 | 72 |
| 音楽, 歌声+音楽 | 64 | 1936 |

音データの内，会話+BGM 区間であると誤識別された数を表している．

表 2 より，会話区間を会話+BGM 区間と識別してしまう誤りと，音楽区間を歌声+音楽区間と識別してしまう誤りが多く発生していることが確認できる．特に，音楽区間を歌声+音楽区間と識別してしまう誤りが多く，一方で歌声+音楽区間を音楽区間と識別してしまう誤りは少ない．これは，音楽区間に含まれるノイズやドラム音 (ドラム単独音ではなくバンド演奏などのドラム音) によって，音楽区間の同じ音高の音が一定時間持続するという特徴が現れにくくなったためであると考えられる．

ここで，会話と会話+BGM を一つの音種別とし，音楽と歌声+音楽を一つの音種別とすると，表 2 は表 3 のようになる．表 3 のように音種別を 2 種類にまとめると，識別精度は 96%程度となることが確認できる．

表 4　各特徴量の貢献度

| | $c_{sf}$ | $c_m$ | $c_h$ | $c_p$ |
|---|---|---|---|---|
| 会話 | -0.1 | 6.5 | -0.3 | -0.6 |
| 音楽 | 0.9 | -2.7 | -0.2 | 1.3 |
| 会話+BGM | 4.9 | 4.7 | -0.5 | -1.0 |
| 歌声+音楽 | 3.0 | 0.3 | 0.7 | 0 |
| 平均 | 2.1 | 2.2 | -0.1 | -0.1 |

したがって,「会話, 会話+BGM」と「音楽, 歌声+音楽」という大雑把な識別を行った後に，会話区間と会話+BGM 区間，音楽区間と歌声+音楽区間の各々の識別に特化した特徴量を用いて識別を行うことで，識別手法の有効性の向上が期待できる．

### 4.3　各特徴量の有効性に関する検証

特徴量 $f_{sf}$(式 (2))，$f_m$(式 (3))，$f_h$(式 (4))，$f_p$(式 (5)) が識別精度 (F 尺度) に与える影響を貢献度 $c_{sf}$，$c_m$，$c_h$，$c_p$ として次式

$$\begin{aligned} c_{sf} &= F_{all} - F_{sf} \\ c_m &= F_{all} - F_m \\ c_h &= F_{all} - F_h \\ c_p &= F_{all} - F_p \end{aligned}$$

で定義する．ここで，$F_{sf}$ は全特徴量から $f_{sf}$ を除いて，$f_m$，$f_h$，$f_p$ の 3 つの特徴量のみを用いて識別を行った結果の F 尺度である．同様に，$F_m$，$F_h$，$F_p$ は，それぞれ全特徴量から $f_m$，$f_h$，$f_p$ を除いて識別を行った結果の F 尺度である．$F_{all}$ は全特

徴量を用いた識別結果のF尺度である．$T = 2$[sec]における各音種別に対する各特徴量の貢献度を表4に示す．

表4の貢献度の平均値を見ると，$f_{sf}$ と $f_m$ は貢献度が高いことが確認できる．特に，$f_{sf}$ は会話+BGM区間の識別と歌声+音楽区間の識別に対して貢献度が高く，$f_m$ は会話区間の識別と会話+BGM区間の識別に対して貢献度が高い．

一方，$f_h$ と $f_p$ の貢献度の平均値はほぼ0と低い．しかし，音種別ごとの貢献度を見ると，$f_h$ は歌声+音楽区間の識別に対して貢献度が高く，$f_p$ は音楽区間の識別に対して貢献度が高いことが確認できる．特に，音楽区間の識別に対しては $f_p$ が最も貢献度が高い．

$f_m$ は音楽区間の識別の貢献度が，$f_h$ と $f_p$ は会話+BGM区間の識別の貢献度が比較的大きなマイナスの値を示す．すなわち，それらの特徴量を用いることにより識別精度が低下している．

上述したように，各特徴量は特定の音種別の識別に対してのみ有効である．したがって，識別する音種別によって，特徴量を取捨選択することにより，識別精度を向上できると考えられる．

例えば，4.2.2節で述べたように，まず「会話, 会話+BGM」と「音楽, 歌声+音楽」の大雑把な識別を行う．そして，会話区間と会話+BGM区間の識別に $f_{sf}$，$f_m$ の2種類の特徴量を使用し，音楽区間と歌声+音楽区間の識別に $f_{sf}$，$f_h$，$f_p$ の3種類の特徴量を使用することにより，識別精度が向上すると推察される．

## 5　関連研究

高柳らは，本研究と同様に，ソナグラムの特徴から特徴量を抽出して音種別の識別をする研究[6]を行っている．高柳らは，音声，音楽，ノイズの識別を91%の精度で実現しているが，本研究と異なり，重畳音への対応はなされていない．

また，重畳音を含む音種別を識別する研究として谷口らの研究[7]がある．谷口らは，sinusoidal segmentという識別単位とそれに関連する特徴量を利用して，重畳音の時間単位での識別を行っている．しかし，重畳音の音種別は音声+音楽，音声+歌声，歌声+音楽の3種類であり本研究とは音種別が異なる．さらに，重畳音の音量比を正規化しているので，重畳音の音量比の不確定さに対しロバストに識別することを目標とする本研究とは異なる．

## 6　終わりに

本研究では，音データの音種別を，既存の特徴量とソナグラムの特徴から考案した特徴量を組み合わせて，会話，音楽，会話+BGM，歌声+音楽に識別する手法を提案した．音区間の時間長 $T$ を2秒とした時の，4種類の音種別の識別結果の平均F尺度は74.8%であった．さらに，$T$ を大きくすることで，精度が向上することを確認した．誤識別に関しては，会話区間と会話+BGM区間の識別と，音楽区間と歌声+音楽区間の識別が困難であった．したがって，精度の向上にはこれらの困難な識別に特化した特徴量を考案する必要がある．識別に用いた特徴量については，どの音種別の識別に対して有効なのかを貢献度という尺度を用いて考察した．今後は，精度の向上を目指すと共に，対象とする音種別をさらに増やしたいと考えている．

## 参考文献


1) 木村彰吾, 林貴宏, 尾内理紀夫. 類似理由の提示機能を具備した類似動画検索システムの構築. 情報処理学会 第49回プログラミングシンポジウム, pp. 99–106, 2008.

2) Eric Scheirer and Malcolm Slaney. construction and evaluation of a robust multifeature speech/music discriminator. *Proc. 1997 ICASSP*, Vol. 2, pp. 1331–1334, 1997.

3) 谷口徹, 安達了慈, 大川茂樹, 白井克彦. Hmmを用いた音声・音楽識別. 信学技法 SP2003-92, Vol. 103, No. 331, pp. 47–51, 2003.

4) Wu Chou and Liang Gu. robust singing detection in speech/music discriminator design. *Proc. 2001 ICASSP*, Vol. 2, pp. 865–868, 2001.

5) 大石康智, 後藤真孝, 伊藤克亘, 武田一哉. スペクトル包絡と基本周波数の時間変化を利用した歌声と朗読音声の識別. 情報処理学会論文誌, Vol. 47, No. 6, pp. 1822–1830, 2006.

6) 高柳直, 林貴宏, 尾内理紀夫. ソナグラムの画像特徴に着目した音声・音楽・ノイズ区間識別手法の提案. 信学技法 PRMU2006-209, Vol. 106, No. 538, pp. 17–22, 2006.

7) 谷口徹, 安達了慈, 大川茂樹, 誉田雅彰, 白井克彦. 音声・楽器音・歌声が重畳した音響信号中のカテゴリ識別. 信学技法 SP2004-153, Vol. 104, No. 631, pp. 49–54, 2004.

8) 谷口徹, 大川茂樹, 白井克彦. 音声・音楽識別を目的とした特徴量の検討. 信学技法 SP2002-135, Vol. 102, No. 529, pp. 87–91, 2002.

9) 尾内理紀夫. マルチメディアコンピューティング, p. 204. コロナ社, 2008.