Vol.2013-MUS-99 No.13
2013/5/11

擬似正弦波成分を用いた残響・雑音にロバストなオーディオフィンガープリンティング

澁谷 崇[1,a)]　安部 素嗣[1,b)]　西口 正之[1,c)]

概要：テレビ番組に関連する情報を視聴者のモバイル端末に表示・提供するセカンドスクリーンサービスを運用するためには，ユーザの視聴している番組を把握する必要がある．その一つの実現方法として，ユーザのモバイル端末で録音した視聴環境の音を参照データベースと照合する方法がある．しかしながら，残響や外来雑音の影響により，正確な照合・検索が非常に困難であるという問題がある．本稿では，残響や雑音に頑健なコンテンツ検索を実現する，新たなフィンガープリンティング手法を提案する．我々は，微小時間区間において正弦波と見なせる “擬似正弦波成分” に着目し，擬似正弦波性分の時間周波数分布を表現するフィンガープリントを提案する．実験では，提案手法によって，1 台の PC で 5 sec の入力信号を 792 h 分の参照信号と 1.29 sec でマッチングを行うことができ，実環境において 92% 以上の再現率，100% の適合率で検索が可能であることを示す．

キーワード：情報検索，放送，ストリーミング配信，テレビ同期，セカンドスクリーン

Audio Fingerprinting Robust Against Reverberation and Noise using Pseudo-Sinusoidal Components

Shibuya Takashi[1,a)]　Abe Mototsugu[1,b)]　Nishiguchi Masayuki[1,c)]

Abstract: The implementation of second-screen service requires a technology for quick, accurate content identification. This enables the service to trace the channel of a broadcast program that a user is watching or listening to. One approach is to record an audio signal from the user's mobile device, and match it with one in a reference database. However, reverberation and exogenous noise distort a recorded audio signal, making accurate identification more difficult. This paper presents a new fingerprinting method for content identification that is robust against reverberation and noise. It employs *pseudo-sinusoidal components*, which are components that can be regarded as sinusoidal over a short period of time. The method generates a fingerprint that represents the distribution of pseudo-sinusoidal components in the time-frequency domain. Experimental results show that the method can match a 5-s-long input signal against 792 hours of reference signals in 1.56 s on a single PC, and can identify the correct program with a recall of over 92% and a precision of 100% in a realistic setting.

Keywords: Information Retrieval, Broadcast, Streaming Media, TV Synchronization, Second Screen

1. はじめに

近年，スマートフォンやタブレット PC 等のモバイル端末の普及に伴い，放送番組の視聴時に視聴者のモバイル端末に番組関連情報を提示するセカンドスクリーンと呼ばれるサービスが注目を集めている [1]．端末に番組関連情報を提示することによって，テレビでスポーツ観戦をしている視聴者は選手やチームの情報を確認でき，ドラマや映画を見ている視聴者は俳優の着ている服に関する情報を調べることができる．また，放送局や番組のスポンサーにとっ

[1] ソニー株式会社
Sony Corporation
[a)] shibuyat@jp.sony.com
[b)] Mototsugu.Abe@jp.sony.com
[c)] Masayuki.Nishiguchi@jp.sony.com

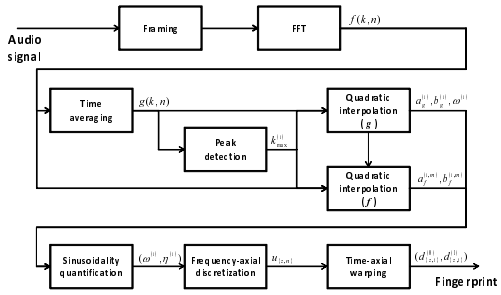


図 1　提案フィンガープリンティング手法のブロック図
Fig. 1　Block diagram of our fingerprinting

ても自社の製品やサービスを視聴者に知ってもらう機会を増やすことができる．

このようなサービスを実現するためには，視聴者がいま何の番組を見ているかについての情報が必要となるが，その情報は視聴者の手を患わせない方法で取得することが望ましい．その観点から，モバイル端末のマイクで視聴環境の音を録音して，その音から視聴している番組を推定するという方法が考えられる．この方法は Query-by-Example オーディオ検索と呼ばれ，高速な検索を行うためにマッチング処理には録音された音声信号から生成されたフィンガープリント（特徴量データ）が用いられる．

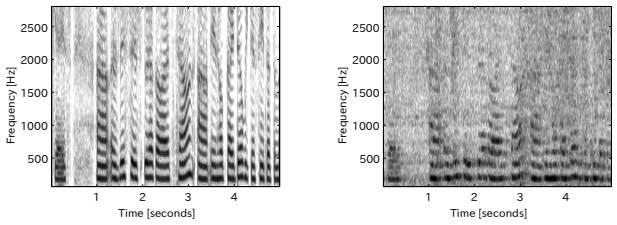
Query-by-Example オーディオ検索を用いたサービスとしては，SoundHound[*1] や Gracenote 社[*2] の MusicID 等の音楽をターゲットとしたもの (Query-by-Example 音楽検索) がすでに普及している．また，Query-by-Example 音楽検索に関する研究も盛んに行われている [2], [3], [4], [5], [6]．一方で，音楽以外も含めた放送コンテンツを対象とした検索技術については，IntoNow[*3] や Shazam[*4] がサービスを展開しているものの，音楽検索技術ほど多くの研究がなされていない [7], [8]．これは，検索対象が音楽だけでないため，音楽特有の性質を利用した手法や技術が使えないケースも発生し，フィンガープリントの設計が難しいからと考えられる．

そこで，本稿では残響・雑音下で，スピーカから離れた位置で録音された音声信号からコンテンツを検索するため手法を提案する．このような検索手法には下記の項目が要求される．

- 視聴環境の残響や外来雑音に頑健であること
- 録音される音声信号はできるだけ短いこと
- 一致検索を高速に行うこと

視聴環境については予備調査に基づき，テレビから 3 m 離れた位置で視聴しているとし，そこでの S/N 比が 10 dB であることを想定する．録音される音声信号の長さは運用されているサービスは 10 sec 以上の長さを必要とするが [6]，本稿では 5 sec を目標とする．高速性については，放送中の 100 チャンネル程度と一致検索をする必要があり，サーバがそれぞれの番組について過去 30 sec 分のデータ， 100 チャンネルで合計 50 min 分のデータを保持しているとし，1 台の PC で 100 クエリを 1 秒以内に処理することを目標とする．

*1 http://www.soundhound.com
*2 http://www.gracenote.com
*3 http://www.intonow.com
*4 http://www.shazam.com



(a) Original signal (Reference)　(b) Recorded signal (Input)
図 2　残響環境下でのスペクトログラム
Fig. 2　Spectrogram under a reverberant environment

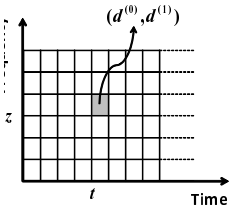


図 3　フィンガープリントの概要
Fig. 3　Concept of fingerprint

2. アプローチ

2.1 擬似正弦波成分の利用

図 2 に二つのスペクトログラムを示す．図 2 (a) は放送番組の音源から生成したスペクトログラム，図 2 (b) は (a) の音源をスピーカーで再生し，3 m 離れた位置で録音した音声のスペクトログラムである．録音時には，録音位置から約 2 m 離れた位置で洗い物をしており，その音（外来雑音）も重畳されている．図 2 (b) には外来雑音と空間残響の影響でスペクトルエンベロープに歪みが生じている．また，エネルギーの集中している時間周波数領域について，周波数の変化が速い成分は，元音源のスペクトルピークに比べ形状が鈍ってしまっていて，原型をとどめていない．一方で，周波数変化の遅い成分は，原型が比較的保たれていると言え，目視によって図 2 (b) の音源には図 2 (a) の音源が含まれていることが確認できる．我々はそのような成分を，微小な時間区間内で正弦波と見なせることから，“擬似正弦波成分” と名付け，この擬似正弦波成分を抽出し，フィンガープリントに用いることを考える．

2.2 フィンガープリントの設計

図 3 は擬似正弦波成分の分布と状態を表現する，我々のフィンガープリントの概略である．我々のフィンガープリントは周波数軸（行成分） z と時間軸（列成分）t を持っ

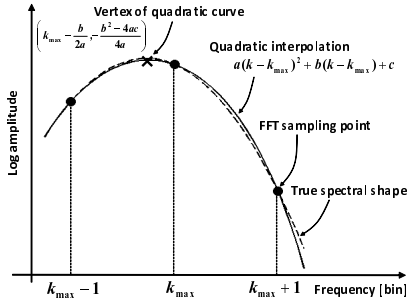


図 4 Quadratically interpolated FFT 法
Fig. 4 Quadratically interpolated FFT method

た 2 次元配列で，配列の各要素はベクトル $(d^{(0)}, d^{(1)})$ を持つ．$d^{(0)}$ は該当する時間周波数領域に擬似正弦波成分が存在するか否かを示す 2 値の変数，$d^{(1)}$ は該当する時間周波数領域に擬似正弦波成分が存在した場合のその状態を示す 3 値の変数である．具体的にはベクトル $(d^{(0)}, d^{(1)})$ は以下の 4 種類の値をとる．

- $(0, -)$: 擬似正弦波成分がない状態
  （“ − ” は不定値を表す.）
- $(1, +1)$: 擬似正弦波成分の立ち上がり点
- $(1, 0)$: 擬似正弦波成分の持続
- $(1, -1)$: 擬似正弦波成分の立ち下がり点

2 次元配列のうち，第 z 行，第 t 列の要素を $(d^{(0)}_{\{z,t\}}, d^{(1)}_{\{z,t\}})$ と表記する．また，フィンガープリントの行数を Z とする（列数はオーディオ信号の長さに依存する）.

3. フィンガープリンティング手法

3.1 正弦波らしさの定量化

擬似正弦波成分の時間周波数分布を表現するために，まずスペクトログラム上の個々のピークについて，それがどれほど正弦波らしいかを表す “正弦波らしさ” の評価を行う.

正弦波らしさの定量化は，QIFFT 法（Quadratically Interpolated FFT 法）[9], [10] を応用して行う．QIFFT 法は対数振幅スペクトルから局所ピークを抽出し，ピーク近傍 3 点を用いて 2 次補間を行うことで，正弦波パラメータ（周波数と振幅）を推定する手法である．2 次補間で得られた 2 次関数の頂点の座標は推定周波数と推定対数振幅となる（図 **4**）．また，得られた 2 次関数の 2 次係数はピーク近傍の曲率を表すが，これには窓関数の形状により定まる理論値 $\bar{a}$ が存在する．この QIFFT 法は単一の FFT フレームにおいて正弦波パラメータを推定する手法であるが，擬似正弦波成分を抽出し，その正弦波らしさを定量化するために，我々は QIFFT 法を複数の FFT フレームをまたいだ形で応用する．擬似正弦波成分は，その対数振幅スペクトルのピーク近傍の形状について下記の 2 点の特徴を持つ.

- 微小な時間区間内の連続するフレームでは形状がほぼ一定
- ピーク付近の曲率が理論値 $\bar{a}$ にほぼ等しい

これらの性質を利用して，正弦波らしさの定量化を行う．図 **1** に従って，フィンガープリント生成方法を説明する.

Time Averaging

まず，FFT によって得られた第 n フレームの近傍 $(2N_\mathrm{a}+1)$ フレームについて，対数スペクトルの時間平均 $g(k,n)$ を求める.

$$g(k,n) = \frac{1}{2N_\mathrm{a}+1} \sum_{m=n-N_\mathrm{a}}^{n+N_\mathrm{a}} f(k,m). \qquad (1)$$

ここで， $f(k,m)$ は第 k bin，第 m フレームの対数スペクトルである.

Peak Detection

次に, 第 n フレームの平均対数スペクトル $g(k,n)$ のうち，局所ピークとなる bin 番号群 $(k^{(1)}_{\max}, \cdots, k^{(i)}_{\max}, \cdots, k^{(I_n)}_{\max})$ を検出する．I_n は $g(k,n)$ から検出されたピークの数，i は検出されたピークの番号である．ピークは $(2K_\mathrm{p}+1)$ - 近傍の最大で検出する[*5].

Quadratic Interpolation (g)

そして，各ピーク $k^{(i)}_{\max}$ の 3 近傍の対数振幅値の 2 次補間を行い，2 次の係数 $a^{(i)}_g$ および 1 次の係数 $b^{(i)}_g$ ，擬似正弦波成分の周波数 $\omega^{(i)}$ を推定する.

Quadratic Interpolation (f)

一方で，n フレームの近傍 $(2N_\mathrm{a}+1)$ フレームの各々の対数振幅スペクトル $f(k,m)$ についても $k^{(i)}_{\max}$ の 3 近傍から 2 次の係数 $a^{(i,m)}_f$ および 1 次の係数 $b^{(i,m)}_f$ を求める.

Quantification of Sinusoidality

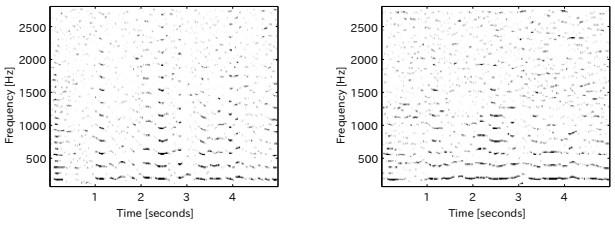
平均対数スペクトルの i 番目のピークの正弦波らしさ $\eta^{(i)}$ を算出する. 正弦波らしさは先に述べた 2 つの項目について評価を行うことで定量化される．ピーク近傍のスペクトル形状の連続フレーム間での安定性は，時間平均スペクトルと各フレームのスペクトルの 2 次の係数および 1 次の係数の誤差で評価する．また，ピーク近傍の曲率の理論値 $\bar{a}$ との近さについては，$a^{(i)}_g$ と $\bar{a}$ との誤差で評価を行う．最終的な正弦波らしさの値は次の式で算出される.

$$\eta^{(i)} = \max\left(1 - \sqrt{\sum_{m=n-N_\mathrm{a}}^{n+N_\mathrm{a}} \varepsilon^{(i,m)}}, 0\right), \qquad (2)$$

$$\text{where } \varepsilon^{(i,m)} = \pi_1 \left(a^{(i,m)}_f - a^{(i)}_g\right)^2 + \pi_2 \left(b^{(i,m)}_f - b^{(i)}_g\right)^2 + \pi_3 \left(a^{(i)}_g - \bar{a}\right)^2. \qquad (3)$$

ここで， $\varepsilon^{(i,m)}$ は 3 つの二乗誤差の加重和で，π_1, π_2, π_3 はそれぞれの重みである．$\eta^{(i)}$ の取りうる値の範囲は $0 \le \eta^{(i)} \le 1$ で， 理想的な正弦波のピークについては $\eta^{(i)} = 1$ となる.

[*5] K_p の値は FFT で用いた窓関数のサイドローブ性ピークを検出しないように設定する.



(a) Original signal (Reference)　(b) Recorded signal (Input)

図 5　Sinusoidality

Fig. 5　正弦波らしさ

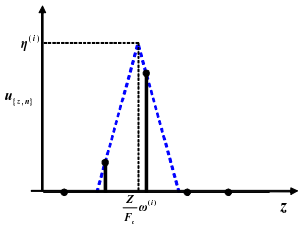


図 6　周波数軸方向の離散化

Fig. 6　Frequency-axial discretization

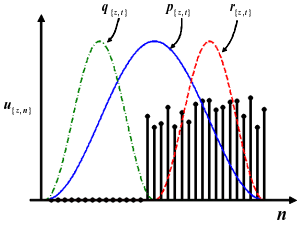


図 7　時間軸方向の平滑化

Fig. 7　Time-axial warping

3.2　正弦波らしさに基づくフィンガープリンティング

Frequency Discretization

第 n フレームから得られたスペクトラルピークの推定周波数 $\omega^{(i)}$ を連続値から離散化する．I_n 個のピークの推定周波数と正弦波らしさのペア $((\omega^{(1)},\eta^{(1)}),\cdots,(\omega^{(i)},\eta^{(i)}),\cdots,(\omega^{(I_n)},\eta^{(I_n)}))$ を，正弦波らしさの周波数分布を表すベクトル $\boldsymbol{u}_n = (u_{\{1,n\}},\cdots,u_{\{z,n\}},\cdots,u_{\{Z,n\}})^T$ に変換する．ベクトル $\boldsymbol{u}_n$ の要素数 Z は最終的に得られるフィンガープリントの行数 Z である（2.2 節）．連続値の推定周波数 $\omega^{(i)}$ と組になっていた正弦波らしさ $\eta^{(i)}$ について，三角窓を用いて周波数の離散化を行う（図 **6**）．

$$u_{\{z,n\}} = \sum_{i=1}^{I_n} v\left(z - \frac{Z}{F_\mathrm{c}}\omega^{(i)}\right)\eta^{(i)}, \tag{4}$$

$$\text{where } v(x) = \begin{cases} 1-|x| & \text{if } |x|<1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$\boldsymbol{u}_n$ の各要素 $u_{\{z,n\}}$ は周波数 $(F_\mathrm{c}/Z)z$ における正弦波らしさを表す．F_c は $z=Z$ のときの周波数で，ここで F_c 以下の周波数に帯域制限を行っている．図 2 (a), (b) で示した信号から得られた $u_{\{z,n\}}$ をそれぞれ図 **5** (a), (b) に示す．色が濃いほど正弦波らしさの値が高いことを意味する．外来雑音の影響は残っているものの，正弦波成分を抽出できていることがわかる．

Time Warping

最後に，この $\boldsymbol{u}_n$ からフィンガープリントの配列要素である $d^{(0)}_{\{z,t\}}$ と $d^{(1)}_{\{z,t\}}$ を算出する．

2.2 節で述べたように，$d^{(0)}_{\{z,t\}}$ は該当する時間周波数領域に擬似正弦波成分が存在するか否かを表す 2 値の変数である．$d^{(0)}_{\{z,t\}}$ を求めるために $u_{\{z,n\}}$ を Hann 窓を用いて時間軸方向に平滑化し（図 **7**），ダウンサンプリングを行う．

$$p_{\{z,t\}} = \sum_{n=1}^{N_\mathrm{w}} w_0(n) u_{\{z,n+N_\mathrm{h}(t-1)\}}, \tag{6}$$

$$\text{where } w_0(n) = 0.5 - 0.5\cos(2\pi n/N_\mathrm{w}). \tag{7}$$

ここで，N_w は Hann 窓のサイズ，N_h はホップサイズである．そして，連続値の $p_{\{z,t\}}$ に対して閾値処理を行い，2 値化することで $d^{(0)}_{\{z,t\}}$ を得る．

$$d^{(0)}_{\{z,t\}} = \begin{cases} 1 & \text{if } p_{\{z,t\}} > \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

ここでは 2 値化の閾値 α は固定の値ではなく，$p_{\{z,t\}}(1 \le z \le Z)$ の値に応じた適応的な値を用いる．これは，残響・雑音環境下では正弦波らしさの値が全体的に小さくなる傾向があるためである．閾値は次の式により定める．

$$\alpha = \beta \frac{\sum_{z=1}^{Z}\left(p_{\{z,t\}}\right)^2}{\sum_{z=1}^{Z} p_{\{z,t\}}}. \tag{9}$$

$d^{(1)}_{\{z,t\}}$ は該当する時間周波数領域の擬似正弦波成分の状態を表す 3 値の変数である．$d^{(1)}_{\{z,t\}}$ の算出では，サイズが $N_\mathrm{w}/2$ の Hann 窓を用いて 2 つの平滑化を行う（図 7）．

$$q_{\{z,t\}} = \sum_{n=1}^{N_\mathrm{w}/2} w_1(n) u_{\{z,n+N_\mathrm{h}(t-1)\}}, \tag{10}$$

$$r_{\{z,t\}} = \sum_{n=1}^{N_\mathrm{w}/2} w_1(n) u_{\{z,n+N_\mathrm{h}(t-1)+(N_\mathrm{w}/2)\}}, \tag{11}$$

$$\text{where } w_1(n) = 0.5 - 0.5\cos(4\pi n/N_\mathrm{w}). \tag{12}$$

$r_{\{z,t\}}$ は $q_{\{z,t\}}$ の $N_\mathrm{w}/2$ フレーム分先を平滑化したものである．これら 2 つの値は，擬似正弦波成分の状態によって次のような関係性になることが期待される．

- 擬似正弦波成分の開始 $(d^{(1)}_{\{z,t\}} = +1)$
  : $r_{\{z,t\}}$ が $q_{\{z,t\}}$ より有意に大きい
- 擬似正弦波成分の持続 $(d^{(1)}_{\{z,t\}} = \ \ 0)$
  : $q_{\{z,t\}}$ と $r_{\{z,t\}}$ に有意な差がない
- 擬似正弦波成分の終了 $(d^{(1)}_{\{z,t\}} = -1)$
  : $q_{\{z,t\}}$ が $r_{\{z,t\}}$ より有意に大きい

つまり，$p_{\{z,t\}}$ と $r_{\{z,t\}}$ を比較することで，擬似正弦波成分の状態を推定できると考えられる[*6]．ここでは $q_{\{z,t\}}$ と $r_{\{z,t\}}$（ともに非負値）の比を用いる．

$$d^{(1)}_{\{z,t\}} = \begin{cases} +1 & \text{if } \frac{r_{\{z,t\}}}{q_{\{z,t\}}} > \gamma_+ \\ -1 & \text{if } \frac{r_{\{z,t\}}}{q_{\{z,t\}}} < \gamma_- \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

[*6] $q_{\{z,t\}}$ と $r_{\{z,t\}}$ の比較は，正弦波らしさの時間変化量の評価を行っていることを意味する．

なお，$d^{(0)}_{\{z,t\}}=0$ となった配列要素 $\{z,t\}$ については，$d^{(1)}_{\{z,t\}}$ の値を不定値 “–” とする．

4. 類似度計算

本節では，残響や雑音の混入した入力信号から得られたフィンガープリントと，データベース中の参照信号から得られたフィンガープリントの類似度を算出する方法について述べる．

入力信号と参照信号から，それぞれ Z 行 $\times$ T 列のフィンガープリントが得られたとする．フィンガープリントは時間周波数領域における擬似正弦波成分の分布とその状態を表現するが，比較の際には擬似正弦波成分の分布の近さについての評価と，擬似正弦波成分が存在した場合のその状態の一致度についての評価を行う．これら 2 つの評価を総合した評価値をフィンガープリントの列ごとに算出した上で，T 列分の評価値の平均値を求め，フィンガープリント全体の類似度とする．

第 t 列における擬似正弦波成分の分布の近さの評価には，相関指標を用いる．

$$S^{(0)}_t=\frac{\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}}d^{(0)}_{\mathrm{R}\{z,t\}}}{\lambda\left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}}\right)+(1-\lambda)\left(\sum_{z=1}^{Z} d^{(0)}_{\mathrm{R}\{z,t\}}\right)}. \qquad (14)$$

ここで，d_{Q} は入力信号のフィンガープリントの要素，d_{R} は参照信号のフィンガープリントの要素である．$S^{(0)}_t$ の分母は入力信号に存在する擬似正弦波成分の数と参照信号に存在する擬似正弦波成分の数の加重平均を，分子は両者に共通して存在する擬似正弦波成分の数を表している．λ は $0\le\lambda\le1$ の値の範囲をとるパラメータで，入力信号の擬似正弦波成分と参照信号の擬似正弦波成分のどちらに重きを置くかを表す．$\lambda=0$ のとき，$S^{(0)}_t$ は参照信号に存在する擬似正弦波成分のうちのどれほどが両信号に共通して存在するかの指標となる．この場合，参照信号に擬似正弦波成分が存在しない周波数領域に，入力信号が擬似正弦波成分を持っていても $S^{(0)}_t$ の値には影響しない．$\lambda=1$ のときはその逆となる．この λ の調整により，入力信号に乗る外来雑音の影響を軽減できる．

一方で，第 t 列における擬似正弦波成分の状態の一致度は，次の式で評価する．

$$S^{(1)}_t=\frac{\sum_{z=1}^{Z} I\left(d^{(1)}_{\mathrm{Q}\{z,t\}}=d^{(1)}_{\mathrm{R}\{z,t\}}\right)}{\sum_{z=1}^{Z} d^{(0)}_{\mathrm{Q}\{z,t\}}d^{(0)}_{\mathrm{R}\{z,t\}}}. \qquad (15)$$

$I(\mathrm{cond.})$ は条件式 cond. が真であるときに 1 の値を返し，偽であるときに 0 の値を返す指示関数である．$S^{(1)}_t$ の分母は入力信号と参照信号の両方に共通して存在する擬似正弦波成分の数を，分子はそれらの状態が一致した数を表している．

式 14 と式 15 の積 $S^{(0)}_tS^{(1)}_t$ を 2 つのフィンガープリントの第 t 列の類似度とし，フィンガープリント全体の類似度には T 列分の類似度の平均値を用いる．

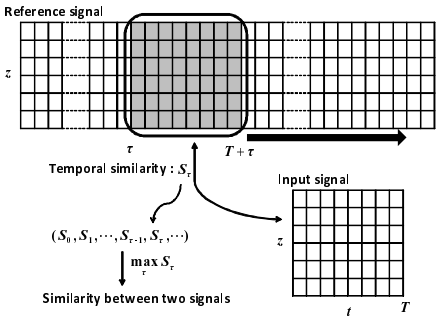


図 8 マッチング処理
Fig. 8 Matching process

$$S=\frac{1}{T}\sum_{t=1}^{T} S^{(0)}_tS^{(1)}_t \qquad (16)$$

5. 実験

本節では，検索の精度と速度の観点で提案手法の評価を行う．本節の実験では，ステレオ音源は 2 ch の信号の和信号を入力信号として扱った．FFT は 16 kHz の信号に対して，32 msec の Hann 窓を用いて，8 msec シフトで行った．周波数分解能は 15.625 Hz となるよう FFT を行った．その他のパラメータの値は $N_{\mathrm{a}}=4$，$K_{\mathrm{p}}=2$，$F_{\mathrm{c}}=2800$，$Z=128$，$N_{\mathrm{w}}=64$，$N_{\mathrm{h}}=16$，$\beta=0.24$，$\gamma_+=1.27$，$\gamma_-=0.79$，$\lambda=0.32$ とした．

5.1 実験 1: 残響に対するロバスト性

テレビ番組の放送音源に，2 件のアパートのリビングルームで計測した室内インパルス応答を畳み込み，空間残響に対するロバスト性を調べた．実験には，参照信号のデータベースとして 1374 番組（792 h 分）の放送音源を用意した．これら参照信号からランダムに 5 sec の信号を 823 本切り出し，2 件のアパートでそれぞれ 3 地点（スピーカから正面 1 m, 3 m, 5 m）で計測した室内インパルス応答を畳み込んだ．これらを畳み込むことによって，実際にそれらのアパートでテレビを視聴しているときの残響を再現することができる．

検索精度の評価をするために，まず上記のデータセットから（入力信号 823 本）×（参照信号 1374 本）の全ての類似度を計算した．実際には参照信号は 5 sec より長いため，入力信号とマッチングを行う参照信号のフィンガープリントを時間方向にスライドさせながら，順次類似度計算を行った（図 **8**）．そして，得られた類似度系列のうち，その最大値を入力信号とその参照信号の最終的な類似度とした．全ての類似度を計算したのちに，コンテンツの一致判定をするための閾値 θ を導入したときの Recall と Precision，および F-measure（Recall と Precision の調和平均）を算出した．閾値 θ の値は予備実験に基づき $\theta=0.26$ とした．

表 **1** は実験結果である．スピーカーから 3 m 離れてい

表 1　検索精度 (実験 1, $\theta = 0.26$)
Table 1　Accuracy (Experiment 1, $\theta = 0.26$)

Distance	F-measure	Recall	Precision
1 m	0.997	0.997	0.996
3 m	0.996	0.994	0.998
5 m	0.995	0.991	0.999

表 2　検索精度 (実験 2, $\theta = 0.26$)
Table 2　Accuracy (Experiment 2, $\theta = 0.26$)

S/N Ratio	F-measure	Recall	Precision
10 dB	0.993	0.994	0.993
5 dB	0.987	0.979	0.995
0 dB	0.955	0.919	0.993
−5 dB	0.839	0.729	0.987

表 3　検索精度 (実験 3, $\theta = 0.26$)
Table 3　Accuracy (Experiment 3, $\theta = 0.26$)

F-measure	Recall	Precision
0.959	0.921	1.000

る位置においても Recall，Precision 共に 0.99 以上の値を示しており，これは提案手法が残響に対して頑健であることを示している．検出漏れを起こした入力信号については，出演者が細切れに話している等，5 sec の信号の中で無音区間の割合が大きい区間であった．

5.2　実験 2: 雑音に対するロバスト性

入力信号に掃除機や流し台などの生活雑音を重畳し，外来雑音に対するロバスト性を調べた．参照信号のデータベースには実験 1 と同じものを用いた．入力信号は実験 1 で切り出した 823 本の信号に生活雑音を 4 段階の S/N 比 (10 dB, 5 dB, 0 dB, −5 dB) で加算したものを用いた．評価方法は実験 1 と同じである．

実験結果を表 **2** に示す．S/N 比が 5 dB のときに Recall，Precision 共に 0.97 以上，S/N 比が 0 dB の状況でも Recall，Precision 共に 0.91 以上の値を示している．これは提案手法が外来雑音に頑健であることを示している．検出漏れの主な原因は実験 1 と同様，無音区間の存在であった．

5.3　実験 3: 実環境のシミュレーション

実際の視聴環境を模擬して実験を行った．入力信号に実験 1 で用いた 3 m の室内インパルス応答を畳み込み，さらに 10 dB の S/N 比で生活雑音を加算したものを用いた．それ他の実験方法は実験 1, 2 と同じである．

Recall が 0.921 ，Precision が 1.000 の値を示しており (表 **3**)，実用的な性能であると言える．

5.4　検索速度

実験 1, 2, 3 は，1 台の PC （Windows 7 (64bit)，CPU クロック周波数 3.20 GHz，クアッドコア）のみで行った．1 つの入力信号（5 sec 分）を 792 h 分の参照信号とマッチングするのに要した平均時間は 1.29 sec であった．これらの値から，1 節で提示した 50 min 分の参照信号とマッチングするのに要する時間は約 1.36 msec である．

6.　おわりに

本稿では，残響・雑音環境下で録音された音声から視聴番組を検索する手法を提案した．本手法のポイントは，微小な時間区間内で正弦波と見なせる成分は残響の影響をあまり受けないことに着目し，利用したことである．実験では，提案手法が残響や外来雑音に頑健であることを示した．

現在の課題としては，本稿の実験では 5 sec の入力信号から生成されたフィンガープリントのみを用いてマッチング処理を行ったが，無音の区間が多く含まれていた場合に検出漏れを起こしていた．しかしながら，そのような場合でも，その前後では正解コンテンツを検出でできていた．今後は，平均検索速度をできるだけ短く保ちながらも，検出漏れを起こした場合に，より長い信号を利用して検索精度を向上させる枠組みを考える．

参考文献

[1] Howson, C., Gautier, E., Gilberton, P., Laurent, A. and Legallais, Y.: Second Screen TV Synchronization, *IEEE ICCE-Berlin*, pp. 361–365 (2011).
[2] Haitsma, J. and Kalker, T.: A Highly Robust Audio Fingerprinting System, *ISMIR*, pp. 107–115 (2002).
[3] Ke, Y., Hoiem, D. and Sukthankar, R.: Computer Vision for Music Identification, *IEEE CVPR*, pp. 597–604 (2005).
[4] Wang, A.: The Shazam Music Recognition Service, *Communications of the ACM*, Vol. 49, No. 8, pp. 44–48 (2006).
[5] Kashino, K., Kimura, A., Nagano, H. and Kurozumi, T.: Robust Search Methods for Music Signals based on Simple Representation, *IEEE ICASSP*, Vol. 4, pp. 1421–1424 (2007).
[6] Chandrasekhar, V., Sharifi, M. and Ross, D. A.: Survey and Evaluation of Audio Fingerprinting Schemes for Mobile Query-by-Example Applications, *ISMIR*, pp. 801–806 (2011).
[7] Anguera, X., Garzon, A. and Adamek, T.: MASK: Robust Local Features for Audio Fingerprinting, *IEEE ICME*, pp. 455–460 (2012).
[8] Duong, N. Q. K., Howson, C. and Legallais, Y.: Fast second screen TV synchronization combining audio fingerprinting technique and generalized cross correlation, *IEEE ICCE-Berlin*, pp. 241–244 (2012).
[9] III, J. O. S. and Serra, X.: PARSHL: A Program for the Analysis/Synthesis of Inharmonic Sounds based on a Sinusoidal Representation, *ICMC*, pp. 290–297 (1987).
[10] Abe, M. and III, J. O. S.: AM/FM Rate Estimation for Time-Varying Sinusoidal Modeling, *IEEE ICASSP*, Vol. 3, pp. 201–204 (2005).