修士論文

聴覚的スペクトル表現に基づいた音響ゼロ電子透かしと音声改ざん検出への応用

市川 敦暉

主指導教員 鵜木 祐史

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和５年３月

Abstract

In recent years, all kinds of information have been digitized, collected, analyzed, and used. In such cyber-physical space, a system to ensure the authenticity of information is required for appropriate information utilization. In particular, audio zero-watermarking technology has been proposed as a mechanism to ensure the authenticity of audio signals. Audio zero-watermark technique creates a detection key from watermark and binary pattern generated from features of the audio signal instead of embedding watermark directly into the signal. This technique apparently embeds the watermark into the signal and uses this key to detect the watermark from the audio signal.

However, this technique has the problem of incorrectly detecting the watermark when the audio signal is subjected to sound information processing such as information compression and speech coding. These sound information processing are unavoidable in the process of communication and storage of digital audio signals. Therefore, a high robustness against these processing is essential. In information compression and speech coding, information is thinned out to the extent that it does not affect human hearing. Therefore, the digital signal itself will be very different, even though the human ear does not notice the change. Audio zero-watermarking generates a binary pattern from the features of the audio signal to create a detection key. This is then used to detect the watermark from the audio signal. If audio zero-watermarking method can generate a binary pattern from an audio signal that has undergone sound information processing using the same signal features, it is expected to be robust to various types of sound information processing. In order to be unaffected by changes in the signal due to sound information processing, it is desirable to use as signal features only those essential elements of the signal that are not subject to information compression (important for human hearing).

Conventional methods extract features unique to a signal from the time and frequency information of an audio signal and generate a binary pattern using these features. However, these methods still have issues, especially in terms of robustness against speech coding. Therefore, this paper decided to use Spikegram, an auditory spectral representation, for feature representation of audio signals. Spikegram represents information on human auditory nerve firings and is derived by sparse coding. Spikegram is a sparse representation of audio signals with a small number of firing points (spikes). The horizontal axis of each spike represents the timing of firing, the vertical axis represents the location of firing, and the shading represents the intensity of the firing. The small number of spikes in the Spikegram are not only important for the signal, but are also essential for human hearing. In particular, this paper investigated the derivation process of each Spike

in the computation of Spikegram with iterative processing, and found that the Spikes are derived in the order of the Spikes with the highest intensity. Therefore, this paper decided to use the Spikegram consisting of Spikes in the initial iteration (up to 100 iterations) as the signal feature. Spikes obtained in the early iterations are the most important of all the spikes and are the essential elements of the signal. Therefore, these spikes are considered to be robust features that cannot be subtracted by information compression or other processes.

The proposed method in this study generates a binary pattern using a Spikegram (consisting of 100 Spikes) derived from an audio signal. The conversion from Spikegram to binary pattern is realized by a binarization process. If the binary pattern has the same size as the watermark, the element is set to 1 if there is even a single Spike in the Spikegram corresponding to the element, and is set to 0 if there is no Spike. The watermark embedding process creates a detection key by calculating the exclusive or of the binary pattern generated by the binarization process and the watermark. In the process of detecting watermark, the exclusive or of the binary pattern generated from the signal to be detected in the same way as in the embedding process and the detection key.

Simulations were conducted to investigate the robustness against the proposed method to sound information processing. In the simulation, a detection key is created from the watermark (W) and the audio signal (original signal) and, and the watermark (W') is detected using the detection key from the audio signal (target signal) in which sound information processing is applied to the original signal. BER (Bit Error Rate) is calculated from the embedded watermark (W) and the detected watermark (W'), and robustness is evaluated using BER as an indicator. First, this simulation used 1-10 seconds of Japanese speech as the audio signal. The watermark information is a random binary matrix of various sizes. Low-pass filtering, re-quantization, re-sampling, information compression (MP3 and AAC), and speech coding (G.711 and G.729) were used for sound information processing. Simulation results showed that the proposed method can detect watermark information without error in the absence of sound information processing. The proposed method also achieves high robustness with BER less than 1% for all sound information processing except for some speech coding. In general, the proposed method achieves the same high robustness against sound information processing as the conventional method, and also achieves sufficient robustness against re-quantization and speech coding, which the conventional method is not good at. In addition, the relationship between the payload and the robustness of the proposed method was investigated. The results showed that the robustness of the proposed method tended to decrease slightly as the payload was increased, but the increase was small, indicating that the method remained robust enough

even when the payload was increased. In addition, to confirm the effectiveness of the proposed method for audio signals other than speech signals, simulations were conducted using music signals. The results showed a similar trend to the results for speech signals, suggesting that the proposed method can be used for all audio signals, both speech and music.

Finally, an application of the audio zero watermarking method based on auditory spectral representation is examined for detection of speech tampering. The determination of tampering will be based on BER calculated by the proposed method. If there is a sufficient difference between BER of the audio information processing and BER of the tampering attack, a binary decision (tampered/no tampered) can be made using a certain BER as a threshold value. First, this paper investigated BER of the proposed method when a tampering attack was applied. The result showed that BER for tampering attacks such as adding noise that drowns out the original speech signal, silencing of the speech signal, or changing the pitch of the speech signal was approximately 2.4%. The BER for sound information processing was about 1.6% for G.729 speech coding, which is the highest BER, indicating that there is a sufficient difference between BER with tampering attacks and BER with sound information processing.

Next, to find the appropriate threshold (BER), this paper investigated the ROC curve and relationship between FRR (False Rejected Rate) and FAR (False Acceptance Rate) when the threshold BER was varied from 1.0% to 5.0% in 0.1% increments for the noise addition, zero interpolation (silence), and pitch shift tampering attacks. The ROC curve showed that the appropriate threshold was 1.2%, while the FRR/FAR relationship showed that the appropriate threshold was 1.1%. Therefore, in this paper, the threshold BER was set at 1.15%, and areas exceeding this threshold were judged to be tampered areas, while areas below this threshold were judged to be non-tampered areas.

The evaluation of speech tampering detection methods against tampering attacks showed that the method has a certain detection capability for zero interpolation, pitch shift, and sample replacement. On the other hand, when sound information processing was applied at the same time as the tampering attack, the detection capability was found to deteriorate.

As a result, it is showed that the audio zero-watermarking method based on auditory spectral representation proposed in this study can be used as a mechanism to guarantee the authenticity of audio signals. It is also suggested that this method can be applied to detect tampering with speech signals.

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

近年，著しい情報技術の進歩に伴い，ありとあらゆる情報がディジタル化され，データとしてやり取りされるようになった．このような世界は，サイバーフィジカルシステム（CPS）[1] によって実現されている．実世界（フィジカル空間）における情報は様々なセンサーで収集され，サイバー空間で収集，分析，利用される．これにより，物理的にその場に居なくても，他人とコミュニケーションを取るだけでなく，生体情報を共有しあたかも「その場」に自分がいるように感じることができる未来が訪れようとしている [2]．特に音声コミュニケーションは，電話の発明により古くから実現され広く普及していたが，COVID-19 の流行によって大きく様態が変わり，我々の生活を根底から変えた．インターネット環境が整っていればどこでもできる仕事を皮切りに，リモートワークが一つの働き方として一般的になった．このような変化に伴い，以前までオフィスで開催されていた会議なども，VoIP のようなインターネット通話技術を用いたものが増え，物理的に居る場所を問わず，オンライン上でコミュニケーションを取ることが当たり前になった．さらに，スマートデバイスの普及を通じて，音声はコンピュータとの主要なインターフェースの一つになった．音声のみを用いたコンピュータの操作は，体を動かすことなく可能であるため大変便利であるが，従来は精度が良くないものも多く，批判的な声も多かった．ところが，精度が著しい向上してきたことで，現在では多くの人が日常で使用するようになった．

これらの変化は我々の日常生活を便利にし，数多のメリットをもたらしたが，一方で見過ごすことのできない問題ももたらした．音声情報は，書き言葉などのテキストのみの情報とは性質が異なり，言語によって伝えようとしているメッセージだけでなく，話者の心理的な状態に関わる情報や，話者の個人性に関する情報も含んでいる [3,4]．そのため，音声情報のプライバシー保護の仕組みが欠かせない．加えて，今日ではこれらの音声情報をあらゆる人物が簡単にアクセス・取得できるため，音声改ざんとその悪用を防ぐ仕組みも必要不可欠である．図 1.1 に示すように，音声情報が不正取得され改ざん・悪用される例として，これまではテレビやラジオのような規模の大きなメディアが発信する音声情報が標的になることが多かった．ところが，今日ではインターネット通話の爆発的な普及により，より手軽に，より簡単に音声情報を不正に取得することが可能となったため，これらの音声全てが改ざん・悪用される危険性を持つことを想定すると，非常に身近な問題に

発展したといえる．例えば，VoIP 技術を基盤とするオンライン会議は，各参加者のデバイス（PC やスマホ等）で録音していても，他者には気付かれない．つまり，このときに無断で録音した人物が，その音声情報に改ざんを加え悪用することも容易にできてしまう．改ざん内容が些細なものでならば大きな問題には発展しないかもしれないが，企業や国を担うような立場にある人物の音声が改ざん・悪用される場合，甚大な影響を与える可能性を秘めているといえる．また，スマートデバイスを通じて収集された音声情報も同様に悪用の危険性がある．これらの音声は各種サービスを提供している企業のサーバーに収集されているため，ここに集められた音声が不正取得され，改ざん・悪用される可能性は十分あるといえる．

さらに悪いことに，情報技術の発達に伴い，音声を改ざんする技術もバリエーションが増え，能力も向上してきている．例えば，本人になりすました音声を合成する技術の発展などが例に挙げられる．音声分析合成技術を利用したボコーダ [5] の発展のみでなく，ディープラーニングを活用した音声変換技術 [6] に基づくディープフェイクも台頭しはじめている．これによりディープフェイク音声の検出に関わる研究 [7,8] も増加傾向が見られる．これらの最新技術を使った音声改ざんでは，改ざん対象となる音声の話者に類似するような特徴を持つ音声を合成し発話内容の一部を変更したり，異なる話者による音声を加工・編集し，あたかもその人物が発言しているように聴こえる音声を作り出す攻撃などが施される．また，既にある音声を加工・編集する方法とは異なり，Text-to-Speech のようにテキスト情報から特定話者の個人性を有する音声を，新しく生み出すような改ざん攻撃も可能である．これらの最新技術を利用した改ざん音声は，聴感では違和感なく，ヒトの耳では見分け（聴き分け）ることができないクオリティまで来ている．今後，さらに自然で違和感のない音声改ざんが容易に出来るようになると予想される．したがって，音声情報の悪用やディープフェイクによる改ざんは，避けることのできない社会問題であり，適切な情報活用のためには，音響信号（音声を含むあらゆる音の信号）の真正性（原本性）を担保する仕組みが欠かせない．

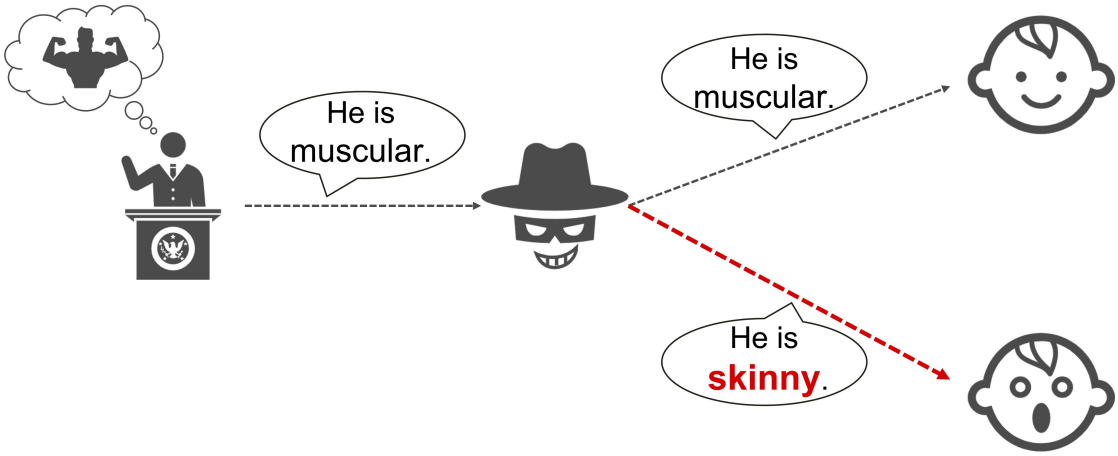

図 1.1: 音声情報の改ざん.

1.2 研究背景

音響信号の真正性を担保するために，これまで様々な方法が提案されてきた．代表的なものに，音響信号を含むディジタルデータ（音データ）のヘッダ領域に，真正性情報を格納する方法がある．音データは，音響信号そのものをディジタル化したデータに加え，音響信号の量子化ビット数やサンプリング周波数，チャンネル数といった信号に関わる属性情報を，ヘッダと呼ばれる領域に格納している．真正性情報をこのヘッダ領域に格納することで，音響信号の真正性を担保する仕組みである．この方法はシンプルで実用的であるが，ヘッダ領域に格納されている真正性情報のみを削除すれば，容易に不正が可能である．

これを踏まえ，ヘッダ領域ではなく，音響信号自体に真正性情報を付加する音響電子透かし技術が提案され，注目されている [9]．音響電子透かしの概要を図 1.2(a) に示す．音響電子透かしでは，はじめに，真正性を担保したい音響信号に透かし情報（真正性情報）を埋め込む．このとき透かし情報には，著作権情報や個人を特定する ID，タイムスタンプ情報など，利用用途に応じた情報を用いることができる．つぎに，透かし情報が埋め込まれた音響信号から，透かし情報を検出する．ここで，音響信号に何らかの攻撃が施された場合，透かし情報は壊れ正しく検出できなくなる．また，逆に，音響信号に攻撃が施されていない場合は，正しく検出することができる．つまり，音響信号から透かし情報を正しく検出できれば，その音響信号は攻撃が施されていないことになり，これにより信号の真正性は担保される．

従来の電子透かしを基盤としながらも，異なる考え方を採用したゼロ電子透かし技術 [10] が提案された．これを音響信号に対して適用させた音響ゼロ電子透かし [11] も，音響電子透かしと同様に注目を集めている．音響ゼロ電子透かしの概要を図 1.2(b) に示す．音響ゼロ電子透かしでは，音響信号に透かし情報を直接埋め込む代わりに，信号の特徴から得られるバイナリパターンと透かし情報から検出鍵を作成することで，見かけ上，音響信号に影響をまったく与えない形（知覚不可能な形）で透かしを埋め込む（ゼロ電子透かし）ことができる．また，この検出鍵を用いて，音響信号から透かし情報を正確で簡単に検出することができる．音響ゼロ電子透かしでは，バイナリパターンの表現の方法を検討することで，秘匿情報量を制御可能である．これにより，音響電子透かしが抱えていた，知覚不可能性と頑健性，秘匿情報量に関わるトレードオフの問題を解消している．

その他の取り組みとしては，音響指紋技術 [12] に基づく信号の照合によって，同一性を確認する方法がある．音響指紋技術を用いた方法では，音響信号から信号特徴を抽出し，これを指紋とみなし大量の指紋情報を学習させることで，目的の信号と一致する信号をデータベース上の膨大な信号から検索する．この方法による信号の検索は，音響ゼロ電子透かしによる信号の真正性の確認過程と類似しているが，学習させる信号にバリエーションを持たせることで，微細な信号の変化に影響を受けにくい方法が実現できる．また，信号の内容が異なるものでも検索

できる可能性もある．例えば，同一話者の異なる内容の発話に対して，話者の音声特徴を指紋として学習済みであれば，発話内容に影響されず本人性の確認により，真正性を担保できる可能性も有している．一方で，音響指紋技術では，さまざまな指紋を学習させることが重要となるため，大量の信号を用意する必要がある．また，照合においても，大量の信号から検索する必要があるため，音響ゼロ電子透かしと比べ必要な計算量が増大する．

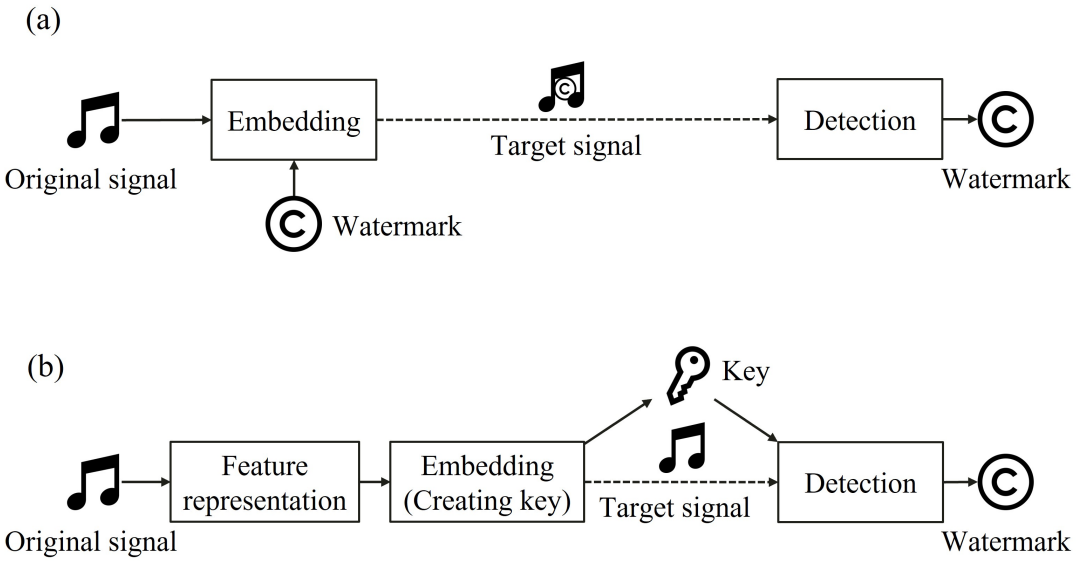

図 1.2: 透かし情報の埋め込み／検出処理のブロックダイアグラム：(a) 音響電子透かしを用いた方法と (b) 音響ゼロ電子透かしを用いた方法.

1.3 研究目的

音響信号の真正性担保に利用できるような音響ゼロ電子透かし法を確立することを，本研究の目的とする．これまでの音響ゼロ電子透かし法にはいくつかの課題がある．音響ゼロ電子透かしを利用した，透かし情報の埋め込みと検出は，信号と検出鍵が一対の関係である性質上，信号長の変化や音情報処理による影響を受けやすい．特に，情報圧縮／音声符号化のような，ディジタル信号自体が大きく変わってしまう音情報処理に対する頑健性には，課題が残されている．しかし，音響信号の真正性を担保するような音響ゼロ電子透かし法は，これらの処理に対して頑健であることが求められる．さらに，音情報処理以外の改ざん攻撃などの処理には脆弱である必要がある．

1.4 論文構成

本論文は，7 章で構成される．図 1.3 に本論文の構成を示す．第 1 章の序論では，本研究の研究背景および研究目的について述べた．第 2 章では，音響電子透かし技術と，音響ゼロ電子透かし技術に着目し，それらの代表的な方法を紹介する．第 3 章では，はじめに音響ゼロ電子透かし技術の課題を説明し，その課題を解決するような提案法の着想を述べる．第 4 章では，聴覚的スペクトル表現に基づく音響ゼロ電子透かし法の具体的な説明を行う．第 5 章では，聴覚的スペクトル表現に基づく音響ゼロ電子透かし法を，頑健性および秘匿情報量の側面から評価する．第 6 章では，聴覚的スペクトル表現に基づく音響ゼロ電子透かし法を利用した，音声改ざん検出を試みる．第 7 章では，聴覚的スペクトル表現に基づく音響ゼロ電子透かし法と，音声改ざん検出の結果を踏まえ，全体のまとめを述べ，残された課題を整理する．

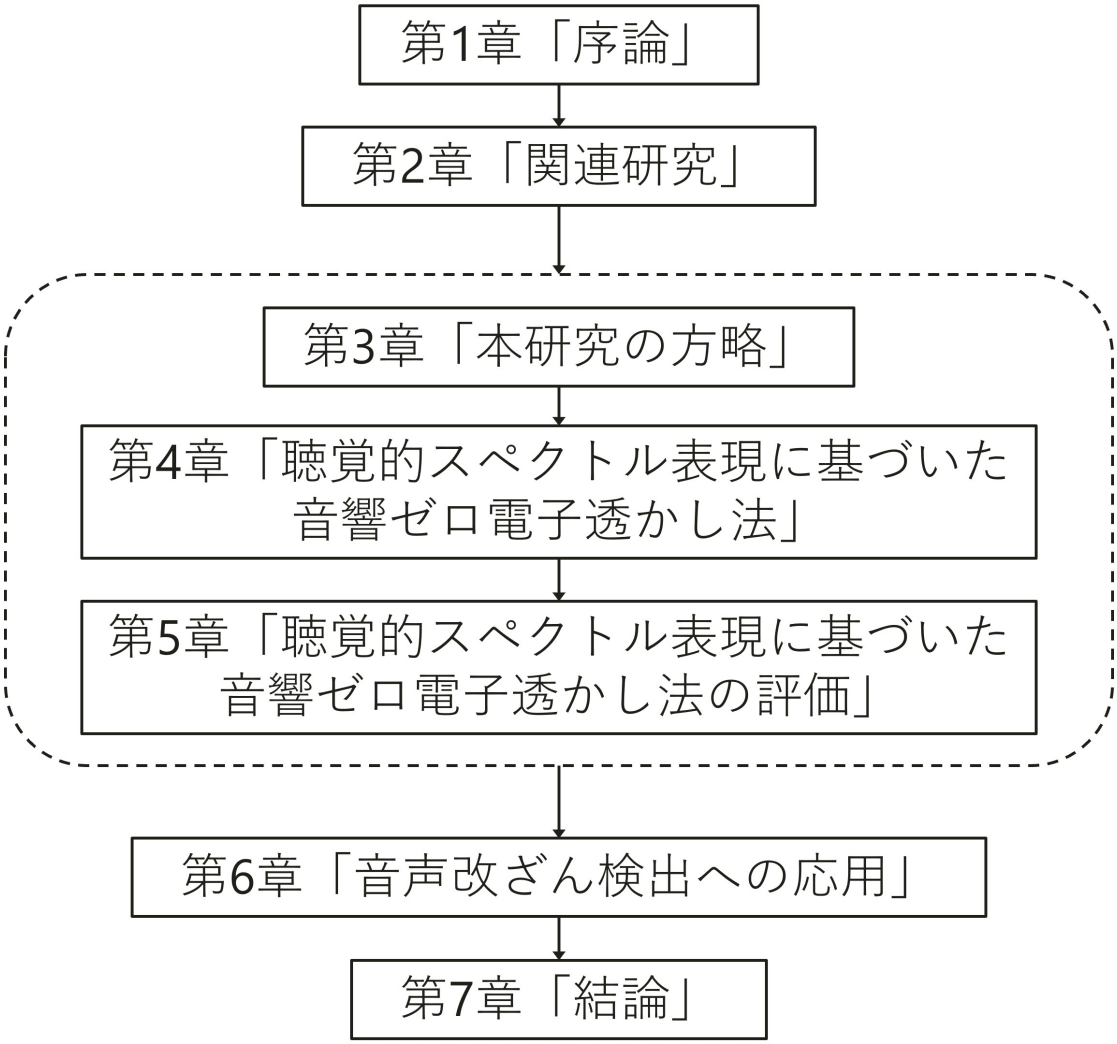
第1章「序論」
第2章「関連研究」
第3章「本研究の方略」
第4章「聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法」
第5章「聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法の評価」
第6章「音声改ざん検出への応用」
第7章「結論」

図 1.3: 論文の構成.

第2章　関連研究

2.1　音響電子透かしに関する研究

音響電子透かしは，透かし情報を音響信号そのものに埋め込み，これを検出する一連の技術である．この方法では，透かし情報が埋め込まれている信号が流通されることになる．そのため，信号に透かし情報を埋め込む際には，埋め込むことで生じる信号の変化がヒトに知覚されてはならない（知覚不可能性）という制限が生じる．この制限下で，埋め込み情報量（秘匿情報量）を増やす工夫や，音情報処理に対する耐性（頑健性）を上げる工夫を凝らした方法が提案されてきた．

代表的な音響電子透かしの方法として，最下位ビット置換法（LSB 置換法）[13,14] が挙げられる．アナログの音響信号をディジタル変換する場合，サンプリング周波数の逆数の時間間隔で，振幅値に応じた量子化が行われる．LSB 置換法では，このときの各サンプルの量子化ビットの最下位ビットに情報を埋め込む．また，LSB 置換法には，全てのサンプルの最下位ビットを埋め込み対象とする方法や，信号の品質変化を知覚不可能性の側面から考慮して，埋め込むサンプルを限定する方法などがある．

その他には，ディジタル変調方式の一つであるスペクトル拡張通信の原理を利用した，直接拡散方式（Direct-sequence Spread Spectrum，DSS）に基づく方法がある [14]．この方法では，疑似乱数系列を使ってスペクトル拡散変調した透かし情報を原信号に加算し，埋め込みを実現する．また，埋め込に利用した疑似乱数系列により透かし情報を検出する．

これらの音響電子透かし法はシンプルで実用的であるが，知覚不可能性の面で課題が残されている．そこで，ヒトの聴覚特性を活用し，透かし情報を埋め込むことで，知覚不可能性を向上させる方法なども提案されている [15]．ヒトは，音源から発生した音（直接音）だけでなく，壁などで反射された音（反射音）も知覚する．これはエコー知覚と呼ばれる．直接音と反射音の二音は，時間的なずれが短いとき一つの音として知覚され，ずれが大きいとき二つの音に分離して知覚される．この特性を利用した方法がエコーハイディング法 [16,17] である．エコーハイディング法では，エコー時間とエコーの振幅を調整することにより，音響電子透かしを実現している．また，その他に，聴覚特性を利用した音響電子透かし法として，蝸牛遅延を利用した方法 [18,19] が挙げられる．ヒトは音を蝸牛内の基底膜振動（進行波）を通じて知覚する．このとき，基底膜はその場所ごとに共振しやすい特徴周波数を持ち，蝸牛入口付近では高い周波数成分，蝸牛先端部では

低い周波数成分に対して反応する．したがって，音の信号の周波数成分ごとに進行波の伝搬遅延が生じる．蝸牛遅延を用いた音響電子透かし法は，この特性を利用しヒトの聴知覚に影響しない範囲で埋め込みを行う方法である．また，かつてはヒトの知覚に与える影響が大きくないと考えられていた位相情報に着目した方法 [20] なども提案されている．

音響電子透かしを利用することで，透かし情報の検出結果から信号の改ざん有無を判定するような方法も提案されている [21–23]．これらの方法では，改ざん攻撃に対する音響電子透かしの脆弱性（非頑健性）を活用し，改ざん攻撃に起因する検出誤りの情報から改ざんを検出する仕組みを実現している．

2.2 音響ゼロ電子透かしに関する研究

音響ゼロ電子透かしによる透かし情報の埋め込み処理と検出処理を，図 2.1 のブロックダイアグラムで示す．図 2.1(a) の埋め込み処理では，はじめに，透かし情報 $W(k,m)$ を埋め込みたい対象となる原信号 $x(n)$ からバイナリパターン $B(k,m)$ を生成する．バイナリパターン B は，透かし情報 W と同じ大きさで，信号 x の特徴を二値表現したものである．つぎに，バイナリパターン B と透かし情報 W の排他的論理和（XOR）を計算し，検出鍵 $Y(k,m)$ を作成する．音響ゼロ電子透かしでは，この一連の処理により，透かし情報を埋め込む．信号（ターゲット信号）から透かし情報を検出するために，埋め込み処理同様，はじめに信号 x からバイナリパターン B を生成する．つぎに，埋め込み処理で作成した検出鍵 Y と，ターゲット信号から生成されるバイナリパターン B の XOR を計算し，透かし情報の検出を行う．

埋め込みおよび検出処理からわかるように，音響ゼロ電子透かしでは透かし情報を信号そのものに埋め込まない．したがって，埋め込むことで信号が変化する心配がなく，知覚不可能性を考慮する必要がない．そのため，音響ゼロ電子透かしにおける秘匿情報量や頑健性を向上させる工夫は，音響電子透かしに比べ高い自由度で行うことができる．音響ゼロ電子透かしの代表的な方法には，音響信号の時間情報から得られる信号特徴を利用する方法と，周波数情報から得られる信号特徴を利用する方法がある．

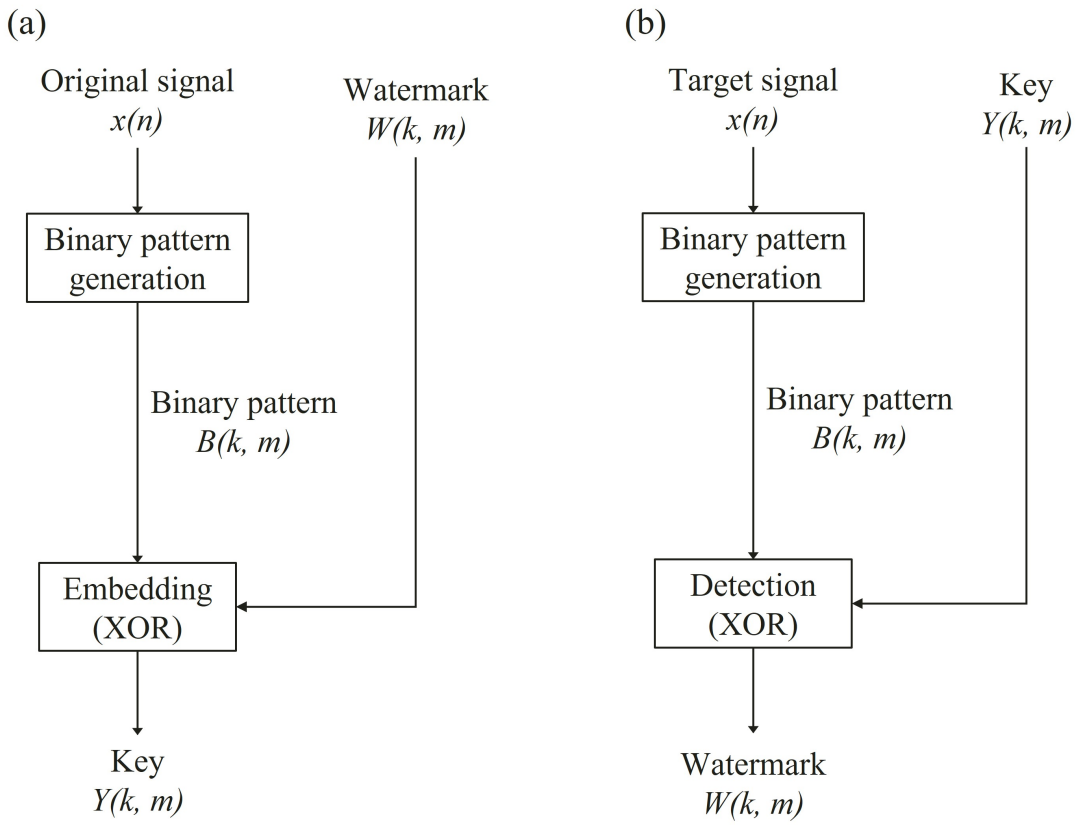

図 2.1: 音響ゼロ電子透かしのブロックダイアグラム：透かし情報の (a) 埋め込み処理と (b) 検出処理.

2.2.1　音響信号の時間情報に着目した方法

信号の振幅情報の特徴を利用する音響ゼロ電子透かし法として，Ali ら [24] の方法が挙げられる．Ali らは，隣接する時間サンプルの振幅値に着目した．この方法では，5 つの時間サンプル点をまとめて 1 つのセグメントとした．各セグメント内の中央（3 番目）の振幅値を基準の値とし，隣接した左右 2 つ（計 4 つ）の振幅値が基準値を上回る場合は 1，下回る場合は 0 とすることで，5 つの時間サンプルから 4 ビットのバイナリパターンを生成する．これを信号の全ての時間サンプルに適用し，バイナリパターンを生成し，透かし情報との XOR を計算することで検出鍵を作成し，この鍵を透かし情報の検出に用いる．

その他の方法としては，信号のパワー（振幅二乗値）を信号特徴として利用する Tsai の方法 [25] が挙げられる．この方法では，はじめに信号を時間方向に M 個のフレーム（セグメント）に分ける．そして各フレームのパワー FE を以下で計算する．

$$FE(m) = \sum_{n=m}^{m+L}, m = 1, \ldots, M \tag{2.1}$$

このとき $x(n)$ は信号，L は各フレームのサンプル数である．つぎに，信号全体のパワーを以下で計算する．

$$E_{all} = \sum [x(n)]^2 \tag{2.2}$$

ここから，フレームのパワー平均を求める．

$$E_{ave} = \frac{E_{all}}{M} \tag{2.3}$$

ここで，以下のように，全体の信号から求めたフレームのパワー平均と，各フレームのパワー平均を比較しバイナリパターン B を生成する．

$$B(m) = \begin{cases} 1 & if\ FE(m) > E_{ave}, \\ 0 & otherwise, \end{cases} \tag{2.4}$$

最後に，このバイナリパターンと透かし情報の XOR を計算することにより検出鍵を作成し，これを透かし情報の検出に用いる．

2.2.2　音響信号の周波数情報に着目した方法

周波数情報から得られる信号の特徴を利用した方法としては，高速フーリエ変換の出力から得られる特徴を利用する方法 [26,27] や離散ウェーブレット変換の出力から得られる特徴を利用する方法 [28, 29] などが挙げられる．また，離散コサイン変換の出力に着目した方法 [30] や修正離散コサイン変換の出力に着目した方法 [31] などもある．これらの方法では，時間フレームごとの周波数スペクトルから信号固有の（頑健な）特徴を見つけ出し，2 を法とする合同式や閾値との大小比

較，ニクラス分類などによって，バイナリパターンを生成し検出鍵を作成する．そしてこの鍵を透かし情報の検出を行う．

その他の方法としては，離散ウェーブレット変換に加え離散コサイン変換を用いることで，周波数スペクトル上の強固な信号特徴からバイナリパターンを生成し検出鍵を作成する方法 [32] が挙げられる．この方法では，はじめに，信号を時間方向に M 個のフレーム（セグメント）に分割する．つぎに，各フレームの信号に対して離散ウェーブレット変換を行い，その出力である近似係数の離散コサイン変換を行うことで変換係数列を求める．係数列を二乗し得られるパワー情報の平均値を各フレームの代表値とし，全フレームの代表値から計算されるフレームの平均値と大小比較を行うことで，バイナリパターンを生成する．

音響電子ゼロ電子透かし技術は，信号に情報を直接埋め込まないことで，知覚不可能性の問題を解消している．これにより，音響ゼロ電子透かしには，頑健性を向上させる工夫がしやすい，秘匿情報量を従来の音響電子透かしよりも大きくすることができるというメリットがある．そこで，本研究では音響ゼロ電子透かし技術を利用した，透かし情報の埋め込みと検出の方法を検討する．また，音響ゼロ電子透かしにおける頑健性と秘匿情報量の関係性にも着目し，調査を行う．

第3章　本研究の方略

3.1　問題設定

音響ゼロ電子透かしでは，音響信号に直接透かし情報を埋め込む代わりに，信号に対応する検出鍵を作成し，この鍵を用いることで透かし情報の検出を実現する．そのため，信号と検出鍵は一対の関係であり，信号長の変化や音情報処理による信号の変化などに影響を受けやすい．特に，情報圧縮／音声符号化のような，ディジタル信号自体が大きく変化する音情報処理に対する頑健性には，課題が残されている．しかし，音響ゼロ電子透かし技術を利用して，音響信号の真正性を担保するには，音情報の通信や保存の過程で施される信号処理の影響を受けるような方法は望ましくない．したがって，音響ゼロ電子透かしの課題である，情報圧縮／音声符号化を含む，あらゆる音情報処理（非攻撃的な信号処理）に対して，高い頑健性が求められる．一般に情報圧縮／音声符号化のような音情報処理では，ヒトの聴覚の特性（最小可聴値やマスキング特性）を考慮し，聴感に冗長である情報が間引かれる [33]．このことから，ヒトの聴感にとって本質的な信号の要素は情報圧縮の影響を受けにくいと考えられるため，この部分を信号特徴として活用しバイナリパターンを生成するような音響ゼロ電子透かし法であれば，あらゆる音情報処理に対して高い頑健性を有することが期待できる．

本稿では，スパースコーディングによって得られる聴覚的スペクトル表現に着目し，ヒトの聴感にとって本質的な信号要素を活用した音響ゼロ電子透かし法を提案する．

3.2　Spikegramによる聴覚的スペクトル表現

Spikegramは，入力信号を複数の基底の組み合わせで表現するスパースコーディングによって得られる信号表現の一つである [34]．Spikegramの導出に用いる基底には，ヒトの蝸牛における基底膜振動のインパルス応答を近似した聴覚フィルタ [35–37] を用いているため，蝸牛基底膜上の聴神経発火のタイミングと場所，ならびに強度の情報をスパースに（全要素数に対し少数の非ゼロ要素で）表現している．そのため，Spikegramは時間・周波数領域における聴覚的スペクトル表現と解釈できる．

図 3.1 は，音声信号/kon-banwa/の波形の一部と，それに対応する Spikegram である．Spikegram の白い要素は全てゼロであり，いくつかの黒くプロットされた点が非ゼロ要素（Spike，発火点）である．Spikegram の横軸（時間）は発火のタイミング，縦軸（周波数）は発火の場所を表現している．また，Spikegram の時間分解能は，信号の時間分解能と一致しているため，信号のサンプリング周波数に依存する．Spikegram の周波数分解能は，周波数分割数に一致しているため，基底関数の周波数帯域幅に依存する．ここでは，Gammatone フィルタ [35, 38] を基底関数としているため，周波数分解能は，サンプリング周波数とその帯域幅 $\mathrm{ERB_N}$ によって決まるチャンネル数に依存する．

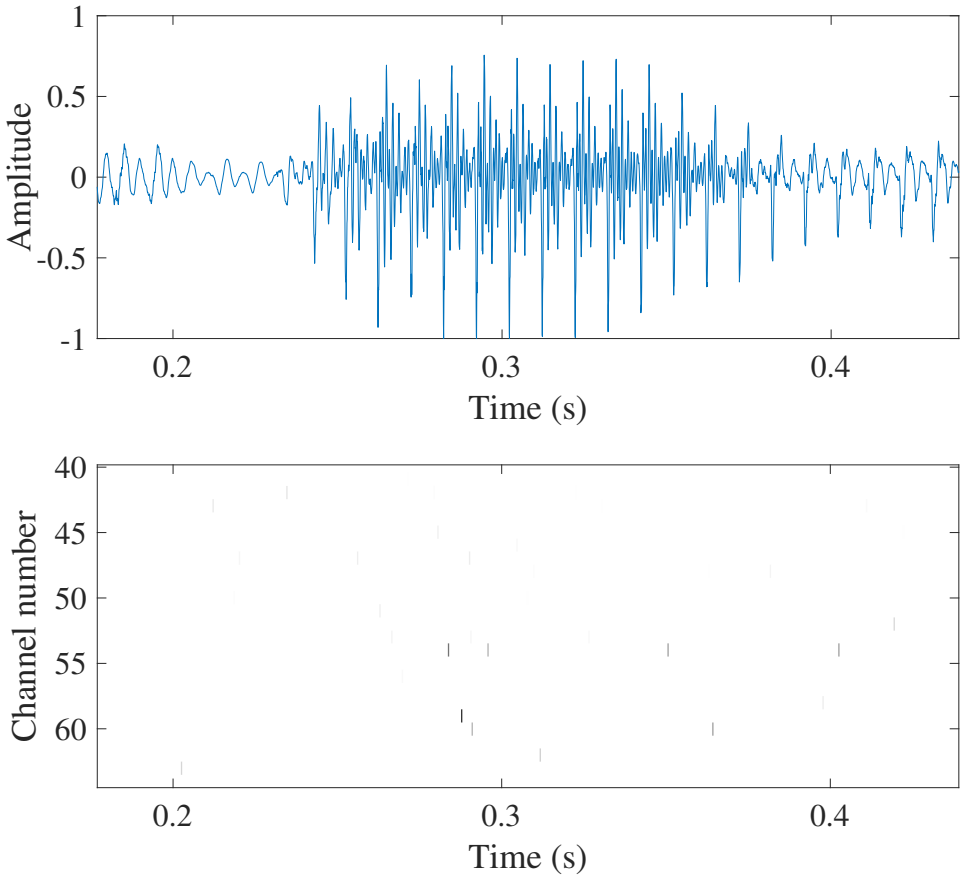

図 3.1: 音声信号に対する Spikegram の例.

3.3 Spikegramの計算方法

Spikegramは，式3.1のように，基底関数$g_k(t)$を利用して音響信号$x(t)$をスパースコーディングにより表したものである [39].

$$x(t) = \sum_{k=0}^{\infty} \langle R^k x(t),\ g_k(t) \rangle\ g_k(t) \tag{3.1}$$

このとき，$g_k(t)$は辞書Dを構成するカーネル（基底）である．また，係数$\langle \cdot \rangle$は，基底$g_k(t)$への差分信号$R^k x(t)$の射影であり，これがSpikegramである．基底$g_k(t)$は，$\mathrm{ERB_N}$の帯域幅を持つGammatoneフィルタで定義される．帯域幅$\mathrm{ERB_N}$（Hz）は，中心周波数f_c (Hz)によって定まる．

$$\mathrm{ERB_N}(f_c) = 24.7\left(\frac{4.37 f_c}{1000} + 1\right) \tag{3.2}$$

中心周波数f_cを持つGammatoneフィルタは，以下の式3.3で計算される．

$$g(t) = a t^{l-1} \exp(-2\pi b \mathrm{ERB_N}(f_c) t) \cos(2\pi f_c t + \phi) \tag{3.3}$$

このとき，t $(t > 0)$は時間，aは振幅である．また，lはフィルタ次数，bは帯域幅係数，ϕは位相である．

また，Spikegramは図3.2に示す計算処理フローチャートを利用して導出される．信号長（サンプル数）をN，サンプリング周波数をFsとする．辞書Dを構成する基底は，中心周波数が異なる帯域幅$\mathrm{ERB_N}$を持ったK個のGammatoneフィルタとする．このとき，Spikegramは時間方向にNサンプル，周波数方向にKチャンネルを持ち，全ての要素に対して初期値0を持つ．まず，入力信号（はじめは原信号）から一点のSpikeを導出するために，辞書Dの基底関数を時間軸にシフトし，最大の発火強度（$Spike\ value$）を示す場所（$maxk$）とタイミング（$maxn$）を見つける．この場所とタイミングに対応するSpikegram上の点にSpike強度を加算し，Spikegramを更新する．

$$S(maxk, maxn) \leftarrow S(maxk, maxn) + Spike\ value \tag{3.4}$$

つぎに，導出されたSpikeを起点に信号を合成し，これを再合成信号に加算することで，再合成信号を更新する．最後に，再合成信号と原信号から差分信号を計算する．反復処理を継続する場合は，この差分信号が次の反復処理における入力信号となり，再びSpikeを導出する．反復処理の終了判定には，音響信号の客観品質推定方法であるPEMO-Q [40]を用いる．ここでは，再合成信号の推定値が0.9以上であればSpikeの導出を終了し，0.9を下回る場合は，再度Spikeの導出とSpikegramの更新を行う．

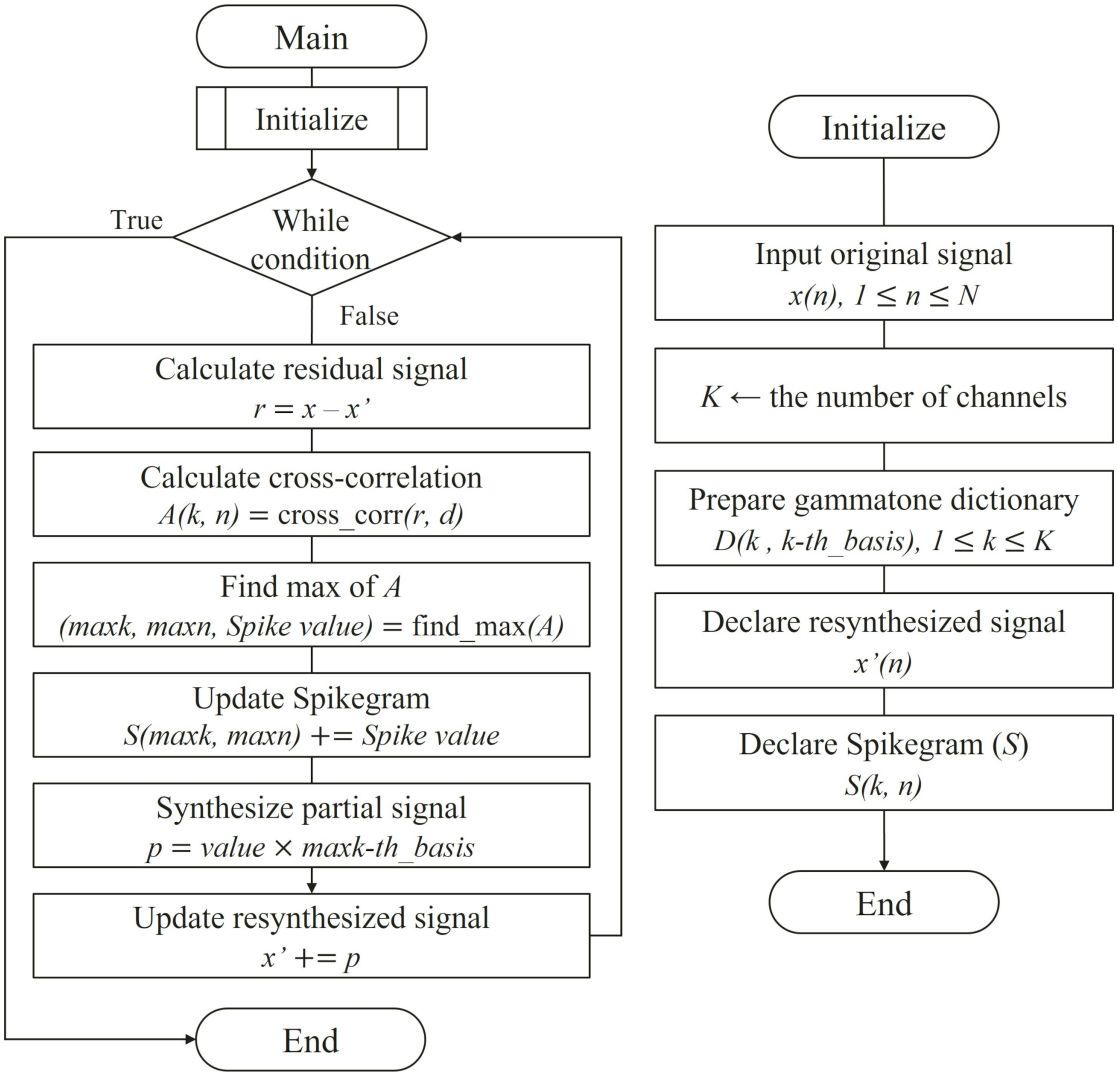

図 3.2: Spikegram の計算処理フローチャート.

3.4 反復処理の回数とSpikegramの関係

Spikegramを計算する際の反復処理では，一回の処理ごとに一点のSpikeが導出される．このときの反復回数とSpike強度の関係を図3.3に示す．横軸を反復回数，縦軸をSpike強度とすると，図3.3から，Spike強度は，反復回数が増すごとに指数関数的に小さくなることがわかる．この結果から，反復初期に導出されるSpikeは信号にとって重要な要素であり，信号が情報圧縮／音声符号化のような音情報処理を受けた場合でも，これらのSpikeのタイミングや場所は変化しない（頑健である）ことが期待できる．

そこで，Spikegramの計算において，Spike数（反復処理回数）を変化させたとき，Spikeのタイミングや場所の頑健性（一致率）がどう変化するかを調査した．4秒の音声信号を原信号とし，原信号にMP3圧縮符号化を施した信号をターゲット信号とする．原信号から生成されるバイナリパターン（50×50）と，ターゲット信号から生成されるバイナリパターン（50×50）の不一致率を調べる．不一致率（MisMatch Rate, MMR）は，原信号のバイナリパターンの1の位置（Spikeの位置）とターゲット信号のバイナリパターンの1の位置において，不一致の可能性のあったビットの中で実際に不一致であったビットの数の割合と定義し，以下で計算する．

$$\mathrm{MMR} = \frac{\text{the number of mismatch bits}}{2 \times \text{the number of Spikes}} \times 100 \tag{3.5}$$

実際に不一致であったビットの数は，二つのバイナリパターンのXOR結果の1の数である．また，原信号のバイナリパターンの1の数とターゲット信号のバイナリパターンの1の数は，今回Spike数を統一しているため等しい．ここから，不一致が生じる可能性があるビットの数はSpike数の二倍で計算される．Spike数を横軸，不一致率を縦軸としたときの結果を図3.4に示す．この結果から，Spike数を増やすほど，不一致率は大きくなることがわかる．これは，反復回数が増えるほど，導出されるSpike強度が小さくなることと関係していると考えられ，強度が小さいSpikeは音情報処理に対し脆弱であり，Spikeの位置（タイミングと場所）が変化しやすく，逆に強度が大きいSpikeは処理に頑健であり，Spikeの位置も変化しにくいことを示唆している．

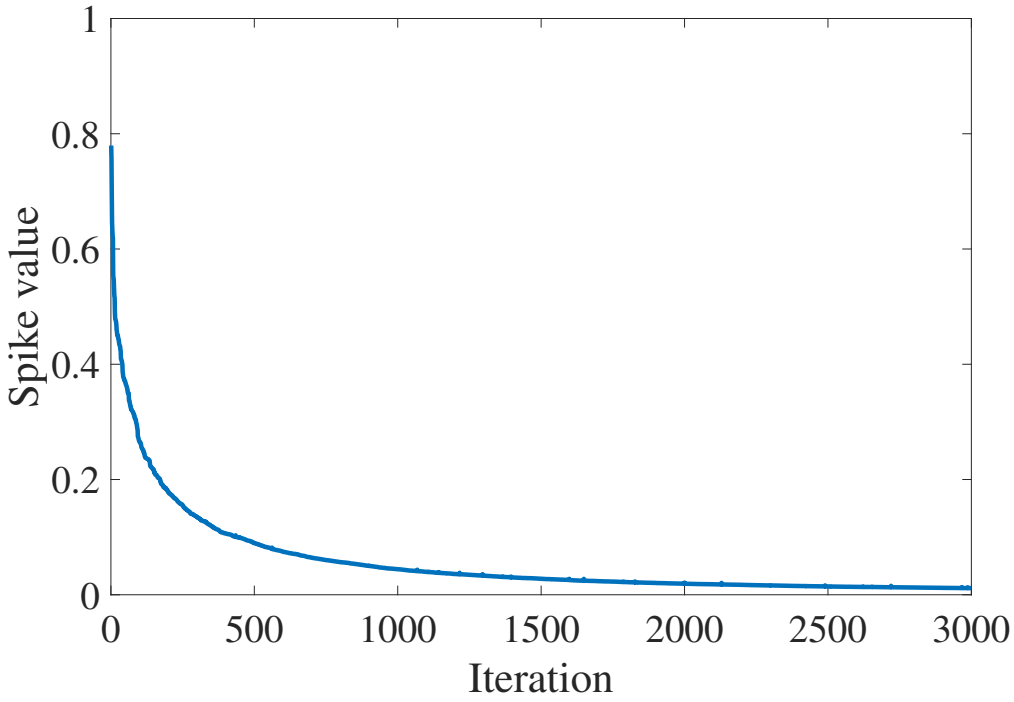

図 3.3: 反復回数と Spike 強度の関係.

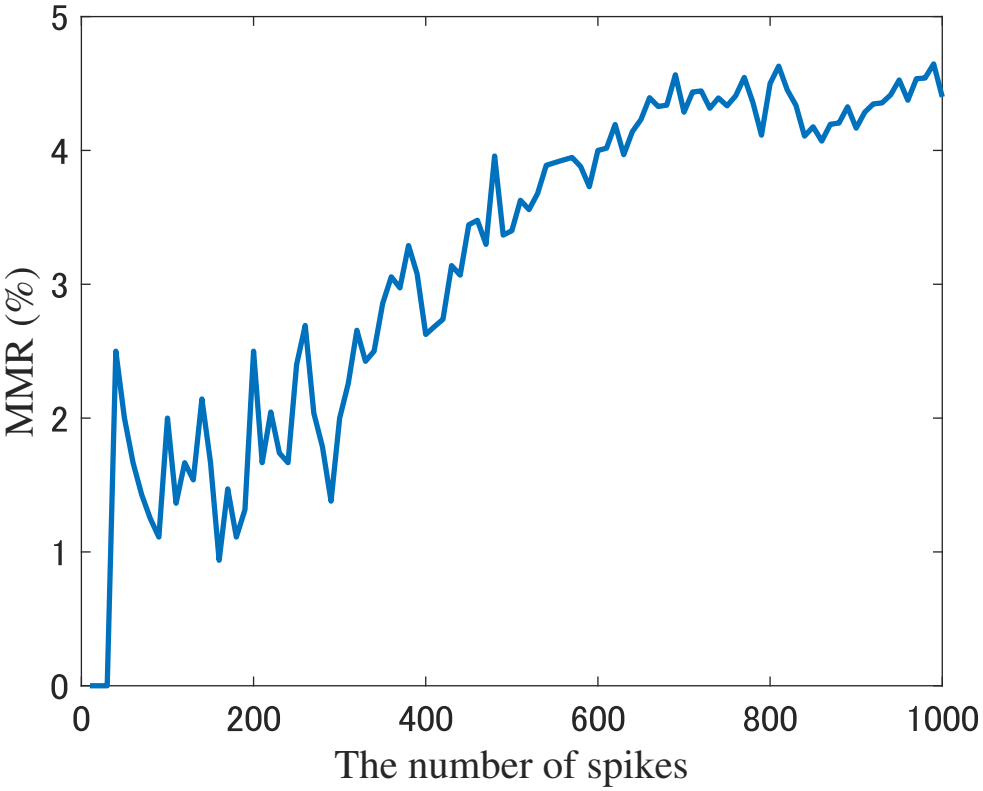

図 3.4: Spike 数と不一致率の関係.

3.5　音響ゼロ電子透かしへの利用

Spikegram の全ての Spike の中でも一部の Spike は，信号を構成する上で，特に重要な要素である．反復回数と Spike 強度の関係，Spike の数と Spike 位置の不一致率の関係から，Spike 強度が大きい Spike の位置（タイミングと場所）は，強度の小さい Spike の位置に比べ音情報処理に対して頑健であることがわかった．ここから，反復初期に得られる Spike は，特に音響信号にとって重要な情報であり，信号に欠かすことのできない本質的な要素であるといえる．したがって，この情報表現からバイナリパターンを生成するような方法を音響ゼロ電子透かしに採用すれば頑健性が向上すると考え，本研究では Spike の数を 100 個（反復回数の上限を 100 回）に絞り，これを信号の特徴として活用するような音響ゼロ電子透かし法を提案する．そして，Spikegram からバイナリパターンへの変換は，Spike の有無による二値変換処理で実現する．

Spikegram の計算に用いる基底は，Tran らの研究 [39] では Gammachirp フィルタ [35, 38] を用いてるが，事前検討の結果 Gammatone フィルタを利用した方が，音響ゼロ電子透かし法の頑健性は有意に高かった．そこで本研究では Gammatone フィルタを用いる．さらに，Tran らはマスキングの影響を考慮し，ヒトの知覚に特に重要と予想される Spike を導出するような方法を提案している．これを採用した音響ゼロ電子透かし法についても事前検討を行った．その結果，マスキングを考慮しない方法と比較したところ，頑健性に有意な差は見られなかった．今回の方法では Spike の導出を初期の 100 回に制限しているため，この範囲内ではマスキングの影響を受ける Spike の数が少ないことが一つの要因として考えられる．しかし，計算量に関しては，マスキングを考慮した Spikegram の計算を採用した音響ゼロ電子透かし法は，マスキングを考慮しない方法に比べはるかに大きかった．即時性の観点から，本研究ではマスキングを考慮しない Spikegram の計算を採用する．

第4章　聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法

4.1　提案法の全体構成

図 4.1 に聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法の構成を示す．図 4.1(a) は透かし情報の埋め込み処理，図 4.1(b) は透かし情報の検出処理である．ここでは，n $(1 \leq n \leq N)$ を信号のサンプル値，k $(1 \leq k \leq K)$ を周波数のチャンネル番号，m $(1 \leq m \leq M)$ を時間分割するセグメントの番号とする．

埋め込み処理では，はじめに，原信号 $x(n)$ から Spikegram $S(k,n)$ を導出する．次に，二値変換処理（Binalization）により Spikegram $S(k,n)$ からバイナリパターン $B(k,m)$ を生成する．最後に，埋め込み処理（XOR）を利用して，透かし情報 $W(k,m)$ とバイナリパターン $B(k,m)$ から鍵情報 $Y(k,m)$ を作成する．したがって，検出鍵のサイズはバイナリパターンのサイズ $k \times m$ と等しくなる．

一方，検出処理では，ターゲット信号の Spikegram $S(k,n)$ とバイナリパターン $B(k,m)$ を同様に導出し，図 4(a) で作成された鍵情報 $Y(k,m)$ を用いて，透かし情報 $W(k,m)$ を検出する．

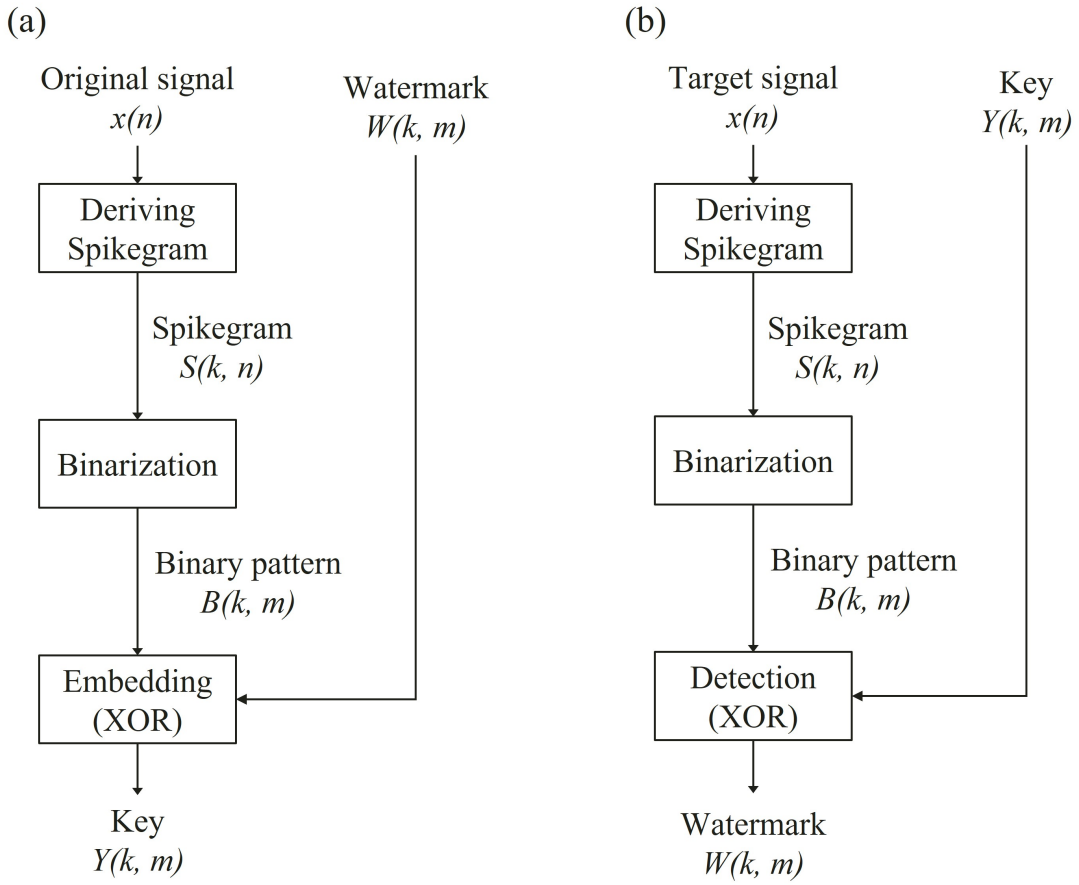

図 4.1: 提案法のブロックダイアグラム：透かし情報の (a) 埋め込み処理と (b) 検出処理.

4.2 透かし情報の埋込と検出

4.2.1 透かし情報の埋込

透かし情報の埋め込みは，以下の手順に沿って行われる．

Step. 1 $x(n)$ から $S(k,n)$ を導出する．このとき，Spikegram を導出する際の反復回数は 100 回とする．

Step. 2 $S(k,n)$ を時間方向に M 個のセグメントで分割する．セグメント内に一点でも Spike（非ゼロ要素）があれば，そのセグメントを 1 とし，なければ 0 とする．この二値変換処理によって，図 4.2 に示すような，Spikegram $S(k,n)$ からバイナリパターン $B(k,m)$ への変換を行う．

Step. 3 $B(k,m)$ と $W(k,m)$ の排他的論理和（XOR）を計算し，$Y(k,m)$ を得る．このとき，W もバイナリ行列である．

4.2.2 透かし情報の検出

透かし情報の検出は，以下の手順に沿って行われる．

Step. 1 透かし情報を検出したい対象となるターゲット信号を $x(n)$，埋め込み処理で作成された検出鍵を $Y(k,m)$ とする．

Step. 2 埋め込み処理における Step. 1 と同様に，$S(k,n)$ を導出する．

Step. 3 埋め込み処理における Step. 2 と同様に，$S(k,n)$ を $B(k,m)$ に変換する．

Step. 4 $B(k,m)$ と検出鍵 $Y(k,m)$ の XOR を計算し，透かし情報 $W(k,m)$ を検出する．

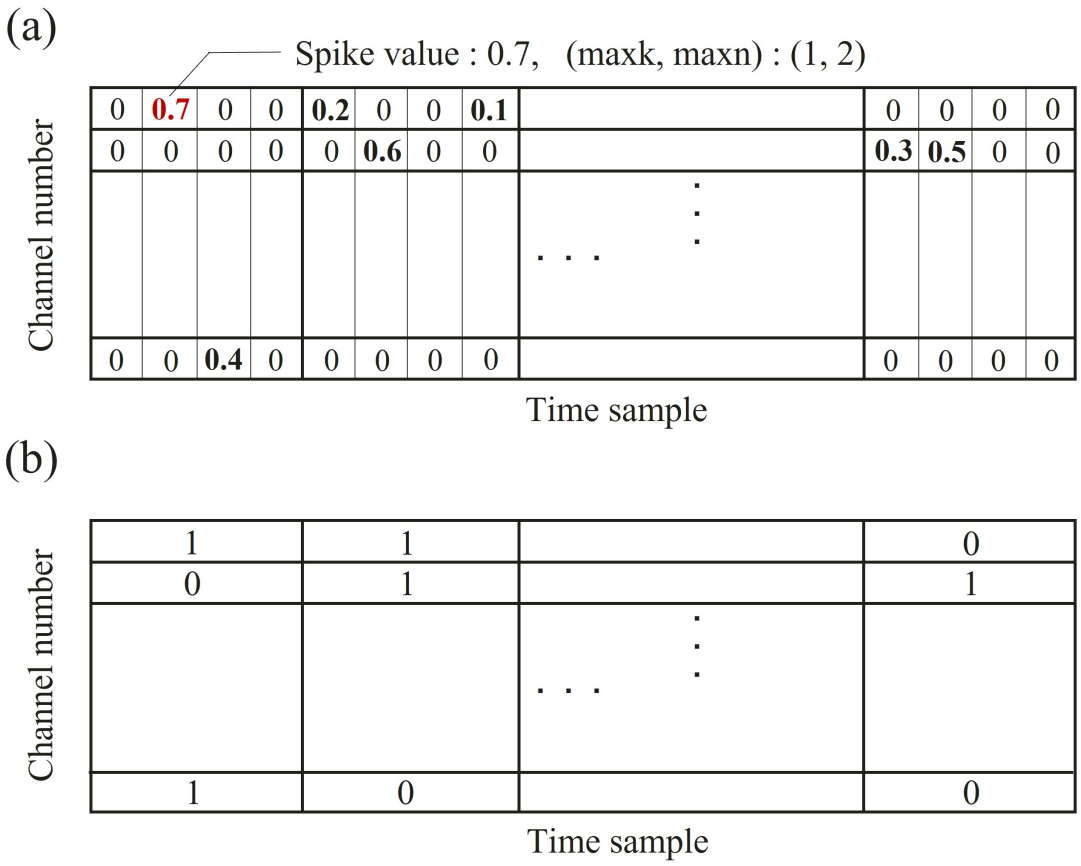

図 4.2: 二値変換処理：(a) Spikegram とそれから得られる (b) バイナリパターン.

4.3 動作の検証

提案法を用いて透かし情報の埋め込みと検出を行った例を，図 4.3 に示す．原信号にはサンプリング周波数が 16 kHz，信号長 2 秒の男性音声を用いた．男性音声の発話内容は「彼らにどんな未来が待っているのだろうか」であった．また，透かし情報には，図 4.3 のバイナリ画像 W（50×50）を用いた．透かし情報を埋め込むために，原信号から Spikegram を導出し，これを二値変換することで図 4.3 に示すバイナリパターン B が生成された．W と B の XOR を計算することで，検出鍵が作成された．検出処理に用いるターゲット信号を，原信号に MP3 圧縮符号化を施し作成した．ターゲット信号から Spikegram の導出，二値化変換処理を行い B' が生成された．埋め込み処理で作成した検出鍵と B' の XOR により，透かし情報 W' が検出された．埋め込んだ透かし情報 W と検出された透かし情報 W' から，MP3 圧縮符号化による信号変化に影響されずに，正しく透かし情報を検出できたことがわかる．

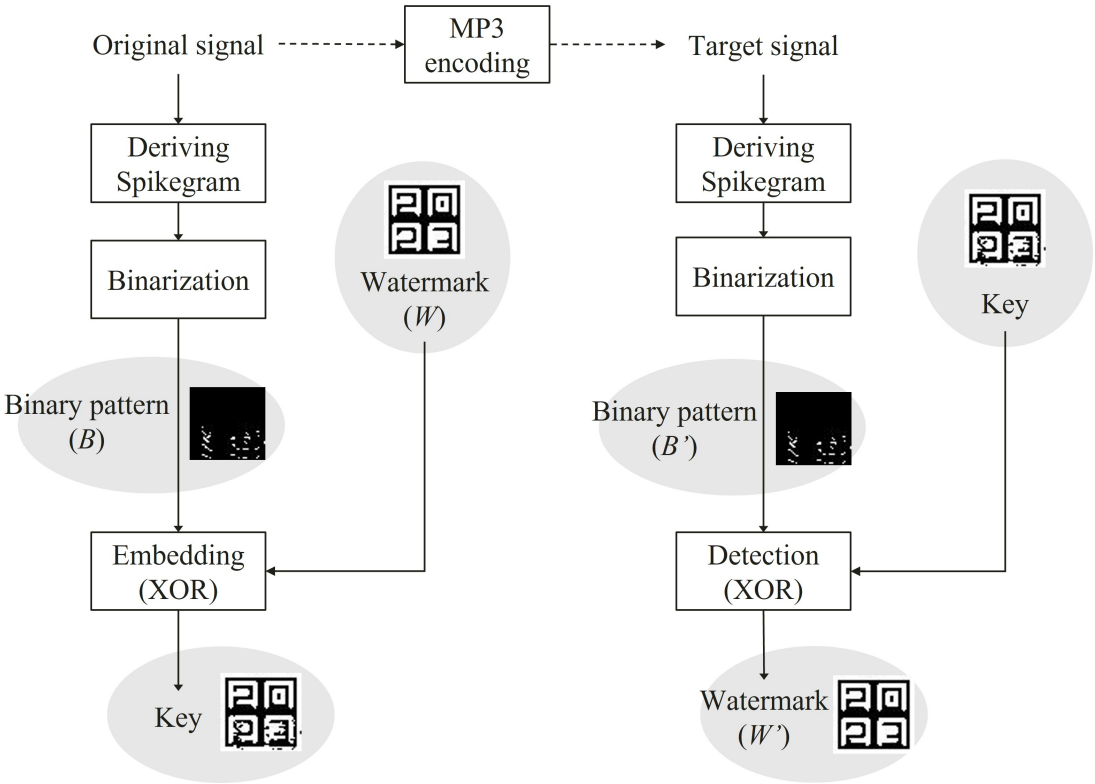

図 4.3: 提案法を用いた情報の埋め込みと検出の例.

第5章　聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法の評価

5.1　目的

本節では，提案法の有効性を確認するために，次に説明する方法で，頑健性および秘匿情報量の評価を行う．頑健性の評価では，音情報処理が施された信号から，透かし情報をどれだけ正しく検出できるかという観点で評価を行う．このとき，音情報処理の影響を受けず頑健に透かし情報を検出できることが望ましい．秘匿情報量の評価では，音響信号にどれだけの透かし情報を埋め込めるか，またそのときの頑健性を考慮し評価を行う．秘匿情報量は大きければ大きいほど，様々な情報を自由に埋め込むことができるため実用性に富む．しかし，秘匿情報量を大きくしたときに頑健性が低下するような方法は望ましくない．したがって，提案法の秘匿情報量の評価では，その大きさのみでなく頑健性との関係性も考慮する必要がある．

5.2　方法

図 4.3 のブロックダイアグラムに沿ってシミュレートを行い，提案法を評価する．埋め込み処理では，原信号となる音響信号から得られるバイナリパターンと，透かし情報であるバイナリ画像の XOR を計算し，検出鍵を作成する．検出処理では，表 5.1 の音情報処理が施されたターゲット信号からバイナリパターンを生成し，埋め込み処理で作成した検出鍵との XOR を計算し，バイナリ画像を検出する．誤って検出されたビットの数は，埋め込んだバイナリ画像 W と，検出されたバイナリ画像 W' の XOR 結果の 1 の総数である．この値を使って，どれだけ提案法が頑健に透かし情報を検出できたかを評価する．また，秘匿情報量は，1 秒間の音響信号に何ビットの情報を埋め込めるかを計算し，そのときの頑健性とともに評価する．

5.2.1 条件

頑健性の評価に用いる音情報処理の一覧を表5.1に示す．まず，音情報処理が施されず原信号がそのままターゲット信号となる条件で評価を行う．つぎに，代表的な音情報処理としてローパスフィルタ処理とリサンプリング，再量子化に対する評価を行う．ローパスフィルタ処理のカットオフ周波数は8 kHzとする．リサンプリングは，サンプリング周波数を16 kHzから8 kHzに下げ16 kHzに戻す処理と，44.1 kHzに上げてから16 kHzに戻す処理の二種類である．再量子化も，量子化ビット数を16 bitから8 bitに下げ16 bitに戻す処理と，24 bitに上げてから16 bitに戻す処理の二種類である．最後に，情報圧縮／音声符号化に対する評価を行う．情報圧縮処理には，ビットレートが32 kbps，128 kbpsのMP3とAAC [41]を用いる．音声符号化には，対数PCMを利用した符号化処理であるG.711 [42, 43]と，CELP方式に基づくG.729 [42, 43]を用いる．

秘匿情報量を評価するために，以下のサイズ（行数と列数）を持つランダムバイナリ行列（計12種類）を透かし情報として利用する．ランダムバイナリ行列は，0と1がランダムに配置された行列である．また，行数は提案法におけるチャンネル数K，列数はセグメント数Mに対応する．

- 行数（チャンネル数）：10, 20, 50
- 列数（セグメント数）：10, 20, 50, 80

提案法の有効性を確認するために，信号のパワー情報に着目した方法（FE法）[25]と，離散ウェーブレット変換と離散コサイン変換から得られる周波数情報に着目した方法（DWT-DCT法）[32]，ガンマチャープフィルタバンク [35]による聴覚的スペクトル表現に基づく方法（GCFB法，※付録参照）を比較対象として利用する．

表 5.1: 提案法の頑健性評価に利用する音情報処理の種類.

Label	Signal processing
NOP	No processing
LPF	Low-pass filtering (cutoff frequency : 8 kHz)
$\mathrm{RSM_D}$	Resampling (16 kHz $\rightarrow$ 8 kHz $\rightarrow$ 16 kHz)
$\mathrm{RSM_U}$	Resampling (16 kHz $\rightarrow$ 44.1 kHz $\rightarrow$ 16 kHz)
$\mathrm{RQZ_D}$	Requantization (16 bit $\rightarrow$ 8 bit $\rightarrow$ 16 bit)
$\mathrm{RQZ_U}$	Requantization (16 bit $\rightarrow$ 24 bit $\rightarrow$ 16 bit)
$\mathrm{MP3_{32}}$	MP3 (bitrate : 32 kbps) [44]
$\mathrm{MP3_{128}}$	MP3 (bitrate : 128 kbps)
$\mathrm{AAC_{32}}$	AAC (bitrate : 32 kbps) [41]
$\mathrm{AAC_{128}}$	AAC (bitrate : 128 kbps)
G711	G.711 (Fs : 8 kHz, log-PCM, μ-law) [42]
G729	G.729 (Fs : 8 kHz, CELP) [42]

5.2.2 評価指標

頑健性の評価指標には，誤検出率（Bit Error Rate, BER）を用いる．BER(%)は，検出された透かし情報の全てのビットにおいて，誤って検出されたビットの割合を示し，以下の式 5.1 で計算される．

$$\mathrm{BER} = \frac{\text{the number of error bits}}{\text{the number of channels} \times \text{the number of segments}} \times 100 \tag{5.1}$$

また，誤検出されたビットの数は，透かし情報 W と W' の XOR を計算し求めることができる．埋め込んだビットが 0 で検出されたビットが 1 のとき，または埋め込んだビットが 1 で検出されたビットが 0 のときに，XOR の計算結果は 1 となる．したがって，誤検出ビットは W と W' の XOR を計算し，これを E とすると，E の 1 のビットが誤検出ビットとなる．

$$E = \mathrm{XOR}(W, W') \tag{5.2}$$

また，透かし情報の埋め込み（検出鍵の作成）処理は以下で表せる．

$$K = \mathrm{XOR}(B, W) \tag{5.3}$$

このとき，K は検出鍵，B は原信号から生成されるバイナリパターンである．XOR の性質から (5.3) は以下の式 (5.4) と同値といえる．

$$W = \mathrm{XOR}(B, K) \tag{5.4}$$

さらに，透かし情報の検出処理は，以下で表せる．

$$W' = \mathrm{XOR}(B', K) \tag{5.5}$$

W' は検出された透かし情報，B' はターゲット信号から生成されるバイナリパターンである．(5.4) と (5.5) から，以下の式が導かれる．

$$\mathrm{XOR}(W, W') = \mathrm{XOR}(B, B') \tag{5.6}$$

(5.2)，(5.5) から以下が成り立つ．

$$E = \mathrm{XOR}(B, B') \tag{5.7}$$

したがって，E は透かし情報 W，W' を用いることなく，バイナリパターン B，B' からも同様に計算することができる．このことから，透かし情報の誤検出は，透かし情報の内容に依存することなく，信号から生成されるバイナリパターンによって決まることがわかる．

秘匿情報量（Payload）は，信号 1 秒あたりの埋め込み bit 数であるため，以下の式で計算する．

$$\mathrm{Payload} = \frac{\text{the number of total bits}}{\text{signal length}} \tag{5.8}$$

秘匿情報量の単位は bps（bit per second）であり，信号長（signal length）の単位は秒である．したがって，5 秒の信号に埋め込まれた透かし情報が 50×50 のサイズを持つ場合，秘匿情報量は，500 bps となる．

5.2.3 データセット

対象となる音声信号には，新聞記事読み上げ音声コーパス [45] から，男女各 50 名ずつ（計 100 名）の音声信号を利用する．音声信号は長さが 1〜10 秒，サンプリング周波数が 16 kHz，量子化ビット数が 16 bit のモノラル信号である．

5.3 結果

5.3.1 頑健性に関する評価の結果

50×50 のサイズを持つ透かし情報を用いた場合の結果を表 5.2 に示す．それぞれの BER の値は，100 個の音声信号に対する BER 結果の平均値である．結果から，処理無し条件において提案法の BER は 0%であり，誤差なく透かし情報を検出できることがわかった．また，提案法は従来法に比べ，全ての音情報処理に対し BER の平均が低かった．各音情報処理ごとに，提案法の結果と従来法の結果に有意な差があるかを調べるために，5%の有意水準を用いてマン・ホイットニーの U 検定を行った．検定の結果，提案法は全ての音情報処理に対して，FE 法および DWT-DCT 法に比べ有意に BER が低いことがわかった．また，提案法は GCFB 法に比べ，ローパスフィルタ処理，リサンプリング（16 kHz $\rightarrow$ 44.1 kHz $\rightarrow$ 16 kHz），AAC（128 kpbs, 256 bps）を除く全ての音情報処理に対して有意に BER が低かった．

つぎに，12 種類のサイズを持つ全ての透かし情報を用いたときの結果を，表 5.3 に示す．それぞれの BER の値は，100 個の音声信号に対する結果と，サイズが異なる 12 種類の透かし情報に対する結果を加味しているため，1200 個の BER の平均値である．処理無し条件において BER は 0%であり，誤差なく透かし情報が検出された．他の処理条件では，G.729 音声符号化を除くすべての処理に対して，BER の平均が 1%を下回ったが，特に再量子化（16 bit $\rightarrow$ 8 bit $\rightarrow$ 16 bit），低ビットレート条件の MP3・AAC（情報圧縮処理），G.711・G.729（音声符号化）条件の結果は，他の処理条件に比べ BER の平均が高かった．

従来法と比較して提案法は，情報圧縮処理を除く全ての音情報処理に対し BER の平均が下回っており，特に，従来法では顕著に BER が高かった再量子化（16 bit $\rightarrow$ 8 bit $\rightarrow$ 16 bit），G.711 音声符号化，G.729 音声符号化に対して，BER の大幅な低下が見られた．加えて，リサンプリング（16kHz $\rightarrow$ 8 kHz $\rightarrow$ 16 kHz）に関しても，BER の大きな低下が見られた．また，マン・ホイットニーの U 検定の結果，提案法はローパスフィルタ処理，リサンプリング，再量子化，音声符号化に対して，FE 法および DWT-DCT 法に比べ有意に BER が低いことがわかった．さらに，提案法は FE 法に比べ，MP3（32 kbps）に対して有意に BER が低いこともわかった．

総じて，全ての結果から次のことがわかった．処理無し条件の結果から，音情報処理が施されたいない場合，提案法は誤差なく透かし情報を検出できる．また，

ローパスフィルタ処理や，リサンプリング，再量子化の条件において，提案法の BER は従来法の BER を下回っており，提案法は従来法と同様これらの音情報処理に対し高い頑健性を有することがわかった．さらに，G.711 音声符号化，G.729 音声符号化に対する結果から，従来法が苦手とする音声符号化に対しても提案法は高い頑健性を有することがわかった．

表 5.2: 音情報処理を施した音声信号から検出された透かし情報の BER (%)．50×50 のサイズを持つ透かし情報を用いたときの結果.

Label	FE [25]	DWT-DCT [32]	GCFB	Proposed
NOP	0.000	0.000	0.000	0.000
LPF	0.012	0.011	0.003	0.002
$\mathrm{RSM_D}$	0.895	0.393	0.302	0.014
$\mathrm{RSM_U}$	0.024	0.009	0.002	0.002
$\mathrm{RQZ_D}$	2.402	2.443	1.093	0.213
$\mathrm{RQZ_U}$	0.000	0.000	0.000	0.000
$\mathrm{MP3_{32}}$	0.965	0.908	0.873	0.480
$\mathrm{MP3_{64}}$	0.187	0.210	0.221	0.144
$\mathrm{MP3_{128}}$	0.064	0.040	0.035	0.026
$\mathrm{MP3_{256}}$	0.065	0.039	0.034	0.028
$\mathrm{AAC_{32}}$	0.700	0.699	0.778	0.385
$\mathrm{AAC_{64}}$	0.174	0.164	0.158	0.125
$\mathrm{AAC_{128}}$	0.126	0.126	0.104	0.095
$\mathrm{AAC_{256}}$	0.126	0.126	0.104	0.095
G711	13.885	13.834	0.429	0.411
G729	22.690	22.414	4.942	1.074

表 5.3: 音情報処理を施した音声信号から検出された透かし情報の BER (%)．12 種類の異なるサイズを持つ透かし情報を用いたときの結果.

Label	FE [25]	DWT-DCT [32]	GCFB	Proposed
NOP	0.000	0.000	0.000	0.000
LPF	0.005	0.006	0.002	0.002
$\mathrm{RSM_D}$	0.886	0.279	0.547	0.023
$\mathrm{RSM_U}$	0.022	0.003	0.002	0.001
$\mathrm{RQZ_D}$	1.208	1.228	1.336	0.281
$\mathrm{RQZ_U}$	0.000	0.000	0.000	0.000
$\mathrm{MP3_{32}}$	0.703	0.659	0.954	0.609
$\mathrm{MP3_{64}}$	0.145	0.129	0.232	0.189
$\mathrm{MP3_{128}}$	0.038	0.022	0.039	0.040
$\mathrm{MP3_{256}}$	0.039	0.022	0.038	0.040
$\mathrm{AAC_{32}}$	0.486	0.485	0.913	0.523
$\mathrm{AAC_{64}}$	0.127	0.124	0.178	0.154
$\mathrm{AAC_{128}}$	0.084	0.084	0.132	0.119
$\mathrm{AAC_{256}}$	0.084	0.084	0.132	0.119
G711	2.094	2.225	0.638	0.496
G729	10.584	10.281	5.309	1.640

5.3.2 秘匿情報量に関する評価の結果

図 5.1 は，透かし情報のサイズ（秘匿情報量）を変化させたときの BER の値をプロットした図である．それぞれの青いプロット点は，1 つの音声信号，1 つの透かし情報に対する，全ての音情報処理条件での BER の平均値である．提案法の結果は，秘匿情報量を増やすと BER が上昇する傾向を示し，秘匿情報量が最大のとき BER も最大の値となった．この結果から，提案法は秘匿情報量が大きくなるにつれ頑健性が低下することが示唆されたが，BER は最大でも 0.5%を下回っており，秘匿情報量が大きい場合でも十分な頑健性を保っていることがわかった．

従来法の結果と比べて提案法は，各点ごとの BER のばらつきが少なく，安定して BER が低いことがわかる．このとき，従来法の BER が大きく提案法と離れているように見える要因は，音声符号化（G.711, G.729）に対する結果の影響と考えられる．図 5.1 のプロット点は，全ての音情報処理の結果の平均値であるため，従来法において，他の音情報処理に比べ特に音声符号化に対する BER の結果が高かったことが起因している．

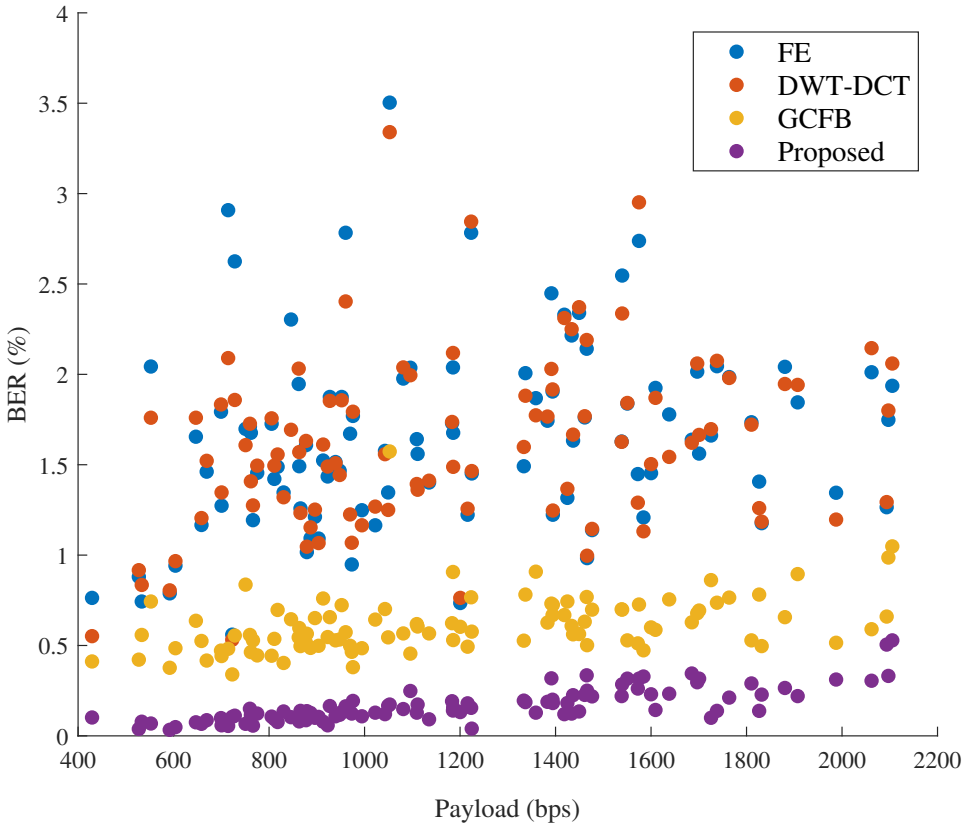

図 5.1: 秘匿情報量と BER の関係.

5.4　音楽信号への適用

提案法が音声信号だけでなく音楽信号にも利用できるかを調査するために，RWC 研究用音楽データベース [46, 47] の音楽信号を用いて，音声信号と同様の頑健性評価を行った．このとき，音楽信号はクラシック，ジャズ，日本語ポップス，英語ポップスの各ジャンルから 4 つ，計 16 の楽曲を選択し，曲の一部（長さ 10 秒）を利用した．音楽信号のサンプリング周波数は 44.1 kHz，量子化ビット数は 16 bit であり，チャネル数は元のステレオからモノラルに変換し利用した．透かし情報には，50×50 のサイズを持つランダムバイナリ行列を用いた．

評価の結果を表 5.4 に示す．結果から，提案法は G.729 音声符号化を除くすべての音情報処理に対して 1%を下回ることがわかった．さらに，50×50 のサイズを持つ透かし情報を用いたときの音声信号に対する結果と同様に，全ての音情報処理に対して，提案法の BER は従来法の BER を下回った．これらの結果から，提案法は音声信号だけでなく，音楽信号にも十分に適用することができることが示された．

表 5.4: 音情報処理を施した音楽信号から検出された透かし情報の BER (%).

Label	FE [25]	DWT-DCT [32]	GCFB	Proposed
NOP	0.000	0.000	0.000	0.000
LPF	0.408	0.225	0.013	0.000
RSM_D	1.320	1.065	0.413	0.000
RQZ_D	2.655	2.675	1.453	0.020
RQZ_U	0.000	0.000	0.000	0.000
$MP3_{32}$	2.503	2.483	1.523	0.303
$MP3_{64}$	1.320	1.293	0.863	0.200
$MP3_{128}$	0.268	0.238	0.293	0.043
$MP3_{256}$	0.040	0.035	0.045	0.005
AAC_{32}	3.598	3.525	2.615	0.515
AAC_{64}	1.658	1.693	1.365	0.223
AAC_{128}	0.520	0.513	0.480	0.080
AAC_{256}	0.320	0.290	0.288	0.035
G711	21.703	21.823	0.573	0.165
G729	23.783	23.920	7.470	1.110

5.5 考察

提案法の頑健性評価の結果から，三点考察を行う．まず，提案法が情報圧縮／音声符号化に対して高い頑健性を示した要因を考察する．従来法で大きな課題であったG.729音声符号化に対するBERの結果は，提案法において1.64%であり，予想を上回る良い結果だった．G.729音声符号化が採用しているCELP符号化は，A-b-S（Analysis-by-Synthesis）法により信号を再合成し表現する符号化であるため，符号化信号は原信号と大きく異なる．しかし，符号化信号と原信号はSpikegram上では近しい表現となり，これが提案法の高い頑健性に寄与していると考えられる．さらに，提案法が従来法に比べ，多くの音情報処理に対して高い頑健性を示した要因は，バイナリパターンの生成方法であると考えられる．従来法では，バイナリパターンを生成するために，部分信号の特徴を信号全体の特徴と比較し二値化する．このとき，音情報処理による影響で，それぞれの部分信号の特徴だけでなく，閾値となる信号全体の特徴も変化してしまうため，バイナリパターンに誤差が増え誤検出に繋がる．一方，提案法では，部分信号内のSpikeの有無で二値化するため，処理による影響を受けても，Spikeの位置さえ変化しなければ問題なく透かし情報を検出できる．これが提案法の高い頑健性に大きく寄与していると考えられる．

つぎに，提案法の改善について考察を行う．提案法はすべての音情報処理に対し高い頑健性を示した一方で，あらゆる音情報処理に対するBERを全て0%にはできていない．提案法の頑健性をさらに向上させるための手がかりは，Spikegramからバイナリパターンを生成する二値変換処理にあると考えられる．今回の提案法ではセグメント内にSpikeが一点でもあれば1，なければ0としたため，セグメント内のSpike数やSpike強度を考慮していない．しかし情報圧縮処理では，Spikeが全くない場所だけでなく，存在しても強度が小さいような場所であれば情報が間引かれる可能性が高い．したがって，Spikeの数（密集度合）や強度を考慮した二値変換処理を採用することで，Spikeが存在していても情報圧縮の対象となるような要素を排除した，より頑健な音響ゼロ電子透かし法の実現が期待できる．また，Spikegramを画像情報とみなすことで，画像としての特徴を利用しバイナリパターンを生成するような二値変換処理も検討する必要がある．具体的な方法としては，画像情報に対するゼロ電子透かし法 [48] を参考に，画像のテクスチャ分析などで用いられるローカルバイナリパターンを利用し，Spikegramからバイナリパターンを生成する二値変換処理などが考えられる．ローカルバイナリパターンでは，周囲のビットとの大小関係から，画像の局所的な特徴を表現できるため，先に述べたSpikeの密集度合や強度なども考慮することができる．

最後に，透かし情報について考察する．今回は透かし情報としてバイナリ画像を用いているが，XORの性質上，原信号から生成されるバイナリパターンとターゲット信号から生成されるバイナリパターンが等しければ，透かし情報は誤差なく検出できる．したがって，このときのバイナリ画像は，全て0あるいは1の画像

でも頑健性には影響しないと考えられる．また，透かし情報の次元数はバイナリパターンの次元数と一致していれば問題なく埋め込み／検出が可能であることから，用いる透かし情報は画像以外にも，文字コードを利用し文字列を符号化した一次元のビット列などでも問題はない．つまり，提案法においては，ビット系列の形で表現できる情報であれば，あらゆる情報を埋め込むことができる．

提案法における秘匿情報量と頑健性の関係性を評価した結果，提案法は秘匿情報量を増やすとBERが上昇する傾向があることがわかった．しかし，提案法のBER結果は，安定して（ばらつきが小さく）従来法よりも低かった．従来法のBER結果はばらつきが大きく，これは音響信号そのものの特性（信号長や波形の特徴）に影響されやすいことを示唆している．一方でばらつきの小さい提案法は，音響信号の特性に影響を受けにくいことがわかる．

第6章　音声改ざん検出への応用

提案した音響ゼロ電子透かし法の応用として，音声改ざん検出を試みる．前章において提案法は，情報圧縮／音声符号化のような音情報処理を含む，様々な音情報処理に対して高い頑健性を有することが確認された．提案法において音響信号が改ざん攻撃を受けた場合，透かし情報の検出誤りが生じるが，改ざん攻撃ではない一般的な信号処理（音情報処理）に対する頑健性が担保されているため，この検出誤り情報は音情報処理ではなく改ざん攻撃に起因するものが多くを占めていると考えることができる．つまり，提案法による検出誤り情報が，改ざん攻撃の有無や改ざん位置などを検出するための重要なヒントとなる．

6.1　聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法の改ざん攻撃に対する頑健性調査

提案法の音情報処理に対する頑健性を前章で評価した．ここでは，提案法の改ざん攻撃に対する頑健性（脆弱性）を調査する．提案法を用いて，音情報処理を施した音響信号から検出される透かし情報の BER と，改ざん攻撃を施した音響信号から検出される透かし情報の BER に明確な差がある場合，ある BER の値を閾値とする改ざん判定が可能である．改ざん攻撃を施したターゲット信号から計算される Spikegram と，原信号から計算される Spikegram は大きく異なるため，生成されるバイナリパターンにも大きな差が生じる．その結果，改ざん攻撃を施したときの BER は，音情報処理を施したときの BER に比べ大きくなることが予想される．そこで，改ざん攻撃を施した音声信号（ターゲット信号）から検出される透かし情報から BER を求め頑健性（脆弱性）評価を行った．このとき，音声信号には，音情報処理に対する頑健性評価に用いた新聞記事読み上げ音声を用いた．また，以下の改ざん攻撃を対象とした．

- 雑音付加（SN 比が 15 dB，0 dB，−15 dB）
- ゼロ補間
- ピッチシフト（−20 %，−20 %）

雑音付加では，SN 比が 15 dB，0 dB，−15 dB となるように白色ガウス雑音を原信号に付加する．ゼロ補間では，原信号の振幅を全て 0 にする．ピッチシフトで

は，WORLD [49] (D4C edition [50]) を利用し，原信号の f0 を下降（0.8 倍），上昇（1.2 倍）させる．

結果を表 6.1 に示す．結果から，SN 比が 15 dB と 0 dB の雑音付加条件を除き，提案法の改ざん攻撃に対する BER は音情報処理に対する BER を大きく上回った．このことは，提案法が音情報処理に対しては頑健であり，改ざん攻撃に対しては脆弱であることを示唆している．特にピッチシフトに対しては，Kasorn ら [23] の結果には届かないものの，3.5%を超える結果となった．一方で，SN 比が 15 dB の雑音付加に対する提案法の BER の結果は 0.5%を下回り，Kasorn ら [23] の結果に比べ大きく下回った．ここから，付加される雑音が信号に対して大きくない場合，雑音を加えても提案法では頑健に透かし情報を検出できてしまうことがわかった．しかし，雑音を音声信号の改ざんに用いる場合，ある特定の発話内容の位置に雑音を付加し，発話内容を誤解（あるいは認識不能に）させるような攻撃が想定される．このとき，一般的に，付加する雑音は元の信号よりも十分大きいものにすると考えられる．表 6.1 の結果から，SN 比を下げていく（雑音を大きくする）と BER は上昇し，特に-15 dB の条件では BER が 3.5%を超える結果となっているため，元の信号をかき消すような雑音を用いた改ざん攻撃に対して，提案法は十分な脆弱性を有することがわかる．

表 6.1: 改ざん攻撃を施した音声信号から検出された透かし情報の BER(%).

Tampering attack	BER
Additive white Gaussian noise (SNR:15 dB)	0.34
Additive white Gaussian noise (SNR:0 dB)	0.83
Additive white Gaussian noise (SNR:−15 dB)	3.82
Zero interpolation	2.39
Pitch shift (−20%)	3.77
Pitch shift (+20%)	4.01

6.2　改ざんの判定に用いる閾値の調査

改ざん攻撃に対する頑健性を調査した結果，改ざん攻撃に対し提案法は脆弱であることがわかった．提案法を用いた改ざん検出には，検出された透かし情報のBERの値を用いる．あるBERの値を閾値とし，計算されたBERが閾値を上回っていれば改ざん箇所であると判定し，下回っていれば非改ざん箇所と判定する．このときの適切な閾値を調査する．はじめに，BERを変化させたときのROC曲線から適切な閾値を求めた．閾値となるBERを，0%から5%まで0.1%刻みで変化させたときのROC曲線が図6.1である．このとき，曲線上の点と座標（0, 1）の距離が最短になるときのBERを求め，これを最適な閾値とする．その結果，ROC曲線から求まる最適な閾値（BER）は，1.2%となった．

つぎに，BERを変化させたときの改ざん見過ごし率（FRR:False Rejected Rate）と，改ざん誤検出率（FAR:False Acceptance Rate）から最適な閾値を求めた．改ざん見過ごし率は，実際には改ざん箇所である部分を，非改ざん箇所と判定した割合であり，改ざん誤検出率は，実際には非改ざん箇所である部分を，改ざん箇所であると判定した割合である．閾値となるBERを，0%から5%まで0.1%刻みで変化させたときの結果が図6.2である．このとき，各x座標の値（BER）におけるFRRとFARの差を求め，これが最小となるBERを最適な閾値とする．結果，最適な閾値（BER）は1.1%となった．

二つの方法で最適な閾値を求めた結果，BERは1.2%と1.1%が妥当であることが示された．ここから，改ざん判定に用いる閾値のBERは，1.15%とする．

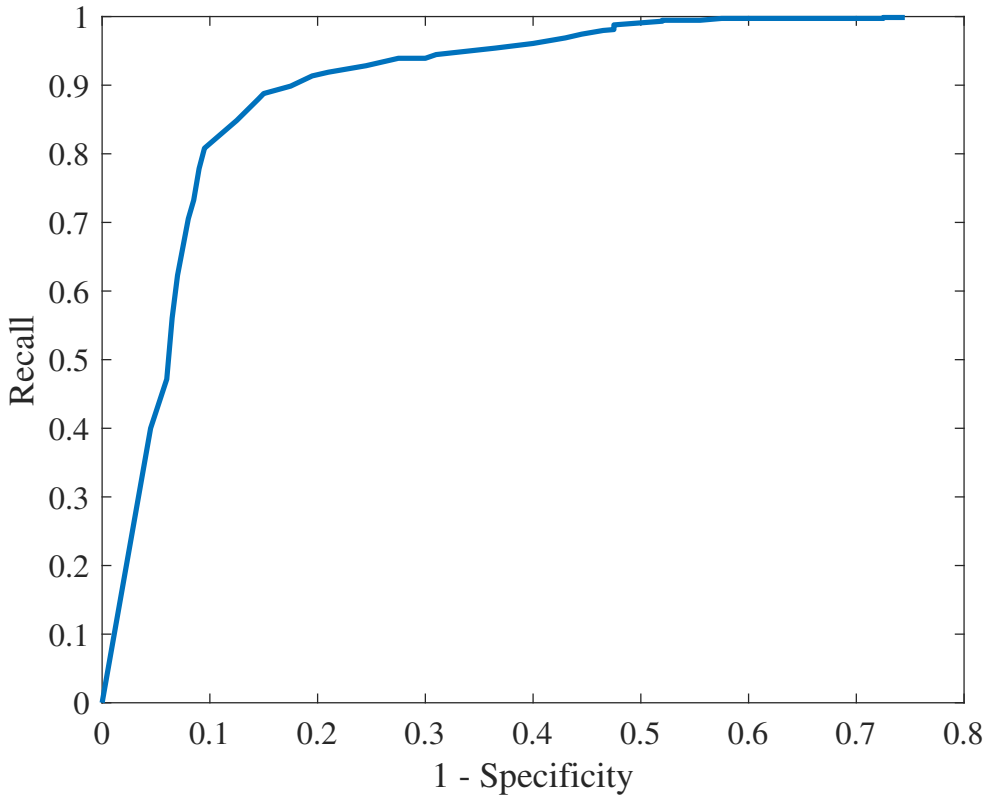

図 6.1: 閾値（BER）を変化させたときの再現率と特異度から求めた ROC 曲線.

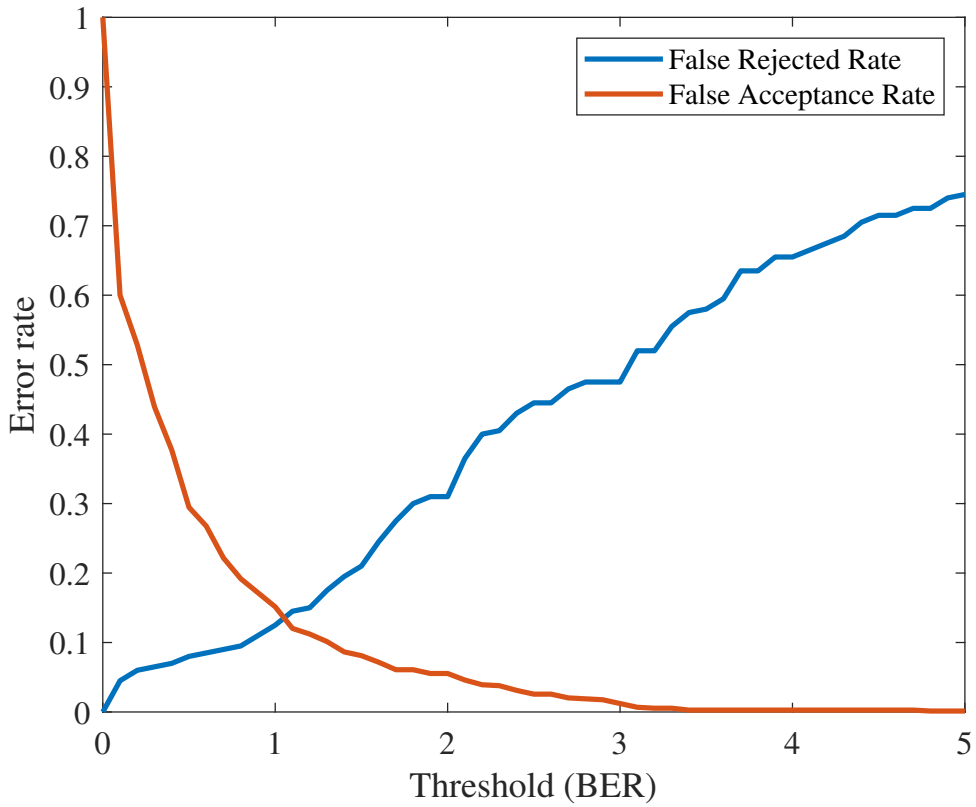

図 6.2: 閾値（BER）を変化させたときの改ざん見過ごし率（FRR）と改ざん誤検出率（FAR）.

6.3　音声改ざん検出法の構成

提案法を利用した音声改ざん検出は，以下の手順で行われる．また，改ざん判定は音声信号 1 秒ごとに行う．このとき，ターゲット信号は改ざんされた可能性がある音声信号であり，透かし情報 W と検出鍵は既知とする．

Step. 1 検出鍵を用いて透かし情報 W' を検出する．

Step. 2 W と W' の XOR を計算する．

Step. 3 Step. 2 の結果から，1 秒あたりの BER を計算する．

Step. 4 BER が 1.15%を超えた部分を改ざん箇所と判定する．

6.4　評価

6.4.1　目的

音声改ざん検出では，改ざん箇所を漏れなく改ざん箇所と判定できることが望ましい．一方で，改ざん箇所と判定したが，その判定が誤りであっては意味がない．そこで，提案法を利用した音声改ざん検出法の検出能力を，感度および正確性の観点から評価する．

6.4.2　方法

4 秒以上の音声信号（原信号）のランダムに選択された 1 秒間を対象に以下の改ざん攻撃を施し，これをターゲット信号とする．原信号と透かし情報 W から作成した検出鍵を用いて，ターゲット信号から透かし情報 W' を検出する．埋め込んだ透かし情報 W と，検出された W' の XOR を計算する．このとき，XOR の結果が 1 のビットが検出誤りビットである．W と W' の XOR 結果から，1 秒ごとの BER を計算し，閾値 1.15%により改ざん箇所と非改ざん箇所のニクラスに分類する．

改ざん攻撃には，以下を用いる．

- 雑音付加
- ゼロ補間
- ピッチシフト
- サンプル置換
- 切り取り

雑音付加による攻撃として，SN 比が −15 dB となるように白色ガウス雑音を改ざん箇所に付加する．ゼロ補間では，改ざん箇所の振幅を全て 0 にする．ピッチシフトでは，WORLD [49] (D4C edition [50]) を利用し，改ざん箇所の f0 を下降（0.8 倍），上昇（1.2 倍）させる．サンプル置換では，音声合成ソフト VOICEVOX（© 2021 Hiroshiba Kazuyuki）を利用し，合成音声で改ざん箇所をサンプル置換する．このとき，文中の固有名詞を自然な形で別の固有名詞に変更したり，文末の表現を変えることで肯定文を否定文にするような，発話内容が変化するサンプル置換を行う．切り取りでは，改ざん箇所を切り取り，以降のサンプルを切り取られた分だけ前にずらす．

評価指標には，検出率（Recall）と適合率（Precision），F 値（F-measure）を用いる．検出率は，改ざん箇所を正しく改ざん箇所と予測した割合であるため，改ざん検出法を感度の面から評価できる．検出率は以下で計算される．

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6.1}$$

改ざんを陽性とすると，TP（True positive）は実際に改ざん箇所であり改ざん箇所と判定された数，FP（False positive）は実際には非改ざん箇所であるにも関わらず改ざん箇所と判定された数，FN（False negative）は実際には改ざん箇所であるにも関わらず非改ざん箇所と判定された数を表す．また，適合率は，改ざん箇所と判定した部分が実際に改ざん箇所である割合であるため，どれだけ正しく改ざん判定ができるか（精度）を表す指標である．適合率は以下で計算される．

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6.2}$$

F 値は検出率と適合率から導かれる値であり，トレードオフの関係である検出率と適合率の両者を同時に評価することができる．

$$\text{F-measure} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{6.3}$$

透かし情報には，50×50 のサイズを持つランダムバイナリ行列を用いる．また，音声信号には，提案法の評価に用いた新聞読み上げ音声の中から，4 秒以上（男女各 20 名ずつ）の音声を用いる．さらに，音情報処理が施された場合の改ざん検出能力も調査するために，改ざん攻撃に加え，MP3 圧縮符号化（32 kbps）および G.729 音声符号化を施した音声信号から，改ざん検出を行う．

6.5 結果

音情報処理を加えず，改ざん攻撃のみを施した 40 の音声信号から，改ざん検出を行ったときの検出率を表 6.2，適合率を表 6.3，F 値を表 6.4 に示す．検出率は，雑音付加，ピッチシフト，切り取りに対して 0.9 以上となった．また，サンプル置

換に対する結果が最も低く 0.75 であった．適合率は，雑音付加と切り取りに対して 0.4 を下回る結果となった．また，ゼロ補間とサンプル置換は 0.7 以上の結果となり，ピッチシフトのみ 0.9 以上であった．F 値は，ピッチシフトに対して上昇／下降を問わず 0.9 を超える優れた結果となった．つぎにゼロ補間およびサンプル置換に対する結果が高く，0.7 を超えた．雑音付加と切り取りの結果は低く，特に切り取りに対する F 値は 0.4 を下回った．

つぎに，32 kpbs のビットレートの MP3 圧縮符号化と，改ざん攻撃を施した 40 の音声信号から，改ざん検出を行ったときの検出率を表 6.5，適合率を表 6.6，F 値を表 6.7 に示す．検出率は，改ざん攻撃のみの場合の結果（表 6.2）とほぼ変わらず，ピッチシフトに対する結果のみわずかに上昇した．適合率は，ピッチシフトに対してのみ 0.7 を上回っており，そのほかに 0.5 を超えた攻撃はゼロ補間とサンプル置換だった．また，改ざん攻撃のみの場合の適合率（6.3）と比較し，切り取りを除く全ての攻撃に対して適合率が低下した．F 値は，ゼロ補間，ピッチシフト，サンプル置換に対して 0.5 を超える結果となり，特にピッチシフトに対しては 0.8 を上回った．切り取りに対する F 値は，音情報処理が無い場合（表 6.4）に比べわずかに上昇したが，その他の攻撃に対しては全て低下した．特に，低下の度合が大きかったのはゼロ補間だった．

最後に，G.729 音声符号化と改ざん攻撃を施した 40 の音声信号から，改ざん検出を行ったときの検出率を表 6.8，適合率を表 6.9，F 値を表 6.10 に示す．検出率は，改ざん攻撃のみの場合（表 6.2）に比べ，雑音付加に対する結果のみ低下し，その他の攻撃に対する結果は変化なし，または上昇した．適合率は，改ざん攻撃のみの場合（表 6.3）に比べ，切り取りに対する結果はわずかに上昇したものの，その他の攻撃に対しては MP3 を施した場合の結果よりもさらに低下する結果となった．F 値は，適合率の低下が影響し，切り取りを除く全ての改ざん攻撃に対して，改ざん攻撃のみの結果（表 6.4）を大きく下回った．

表 6.2: 改ざん攻撃のみを施したときの検出率.

Tampering attack	Recall
Additive white Gaussian noise	0.900
Zero interpolation	0.800
Pitch shift (−20%)	0.900
Pitch shift (+20%)	0.900
Sample replacement	0.750
Cropping	1.000

表 6.3: 改ざん攻撃のみを施したときの適合率.

Tampering attack	Precision
Additive white Gaussian noise	0.371
Zero interpolation	0.762
Pitch shift (−20%)	0.923
Pitch shift (+20%)	0.923
Sample replacement	0.790
Cropping	0.226

表 6.4: 改ざん攻撃のみを施したときの F 値.

Tampering attack	F-score
Additive white Gaussian noise	0.526
Zero interpolation	0.781
Pitch shift (−20%)	0.911
Pitch shift (+20%)	0.911
Sample replacement	0.769
Cropping	0.369

表 6.5: MP3 圧縮符号化と改ざん攻撃を施したときの検出率.

Tampering attack	Recall
Additive white Gaussian noise	0.900
Zero interpolation	0.800
Pitch shift (−20%)	0.925
Pitch shift (+20%)	0.925
Sample replacement	0.750
Cropping	1.000

表 6.6: MP3 圧縮符号化と改ざん攻撃を施したときの適合率.

Tampering attack	Precision
Additive white Gaussian noise	0.343
Zero interpolation	0.508
Pitch shift (−20%)	0.787
Pitch shift (+20%)	0.787
Sample replacement	0.625
Cropping	0.229

表 6.7: MP3 圧縮符号化と改ざん攻撃を施したときの F 値.

Tampering attack	F-score
Additive white Gaussian noise	0.497
Zero interpolation	0.621
Pitch shift (−20%)	0.851
Pitch shift (+20%)	0.851
Sample replacement	0.682
Cropping	0.372

表 6.8: G.729 音声符号化と改ざん攻撃を施したときの検出率.

Tampering attack	Recall
Additive white Gaussian noise	0.800
Zero interpolation	0.800
Pitch shift (−20%)	0.950
Pitch shift (+20%)	0.925
Sample replacement	0.775
Cropping	1.000

表 6.9: G.729 音声符号化と改ざん攻撃を施したときの適合率.

Tampering attack	Precision
Additive white Gaussian noise	0.471
Zero interpolation	0.368
Pitch shift (−20%)	0.535
Pitch shift (+20%)	0.536
Sample replacement	0.443
Cropping	0.229

表 6.10: G.729 音声符号化と改ざん攻撃を施したときの F 値.

Tampering attack	F-score
Additive white Gaussian noise	0.593
Zero interpolation	0.504
Pitch shift (−20%)	0.685
Pitch shift (+20%)	0.679
Sample replacement	0.564
Cropping	0.372

6.6 考察

音情報処理が無い条件で改ざん攻撃を施した場合と，音情報処理と改ざん攻撃どちらも施した場合の結果から，音情報処理（今回はMP3とG.729）による影響は適合率に与える影響が大きいことがわかった．改ざん検出の感度を表す検出率は，音情報処理の有無で差は生じなかった．検出の精度を表す適合率は大きく低下し，その結果F値も低下する傾向が見られた．適合率は式6.2で計算されるため，分母と分子に両方存在するTP（改ざん箇所を改ざん箇所と判定）の変化よりも，FP（非改ざん箇所を改ざん箇所と判定）の増加が適合率の低下の大きな要因となる．提案した音響ゼロ電子透かし法において，改ざん攻撃を施していなくても，音情報処理を施すだけで検出誤りが生じること，またどの程度生じるかは前章で明らかにした．つまり，改ざん攻撃に加え音情報処理を施した場合は，改ざん攻撃に起因する検出誤りだけでなく，音情報処理による検出誤りも生じる．また，提案法を用いた改ざん検出処理では，BERの値を基準に1.15%を超えれば改ざんありと判定している．そのため，音情報処理の影響でBERが増加することで，改ざんありと判定する箇所が増え（減ることがなく），感度を表す検出率は変化せず，さらに誤って改ざん判定した数であるFPは増加するため，適合率が低下したと考えられる．

つぎに，雑音付加に対するF値が他の改ざん攻撃に比べ低い要因について考察を行う．再現率に関しては，他の改ざん攻撃の結果に比べ，小さいわけではない．一方で，適合率は他の改ざん攻撃の結果に比べ大きく下回っている．この理由は，付加する雑音が信号に比べかなり大きいことが要因と考えられる．今回の改ざん検出の評価では，SN比が-15 dBとなる雑音を信号の一部に付加している．これにより，雑音を付加した部分（改ざん箇所）に対応する部分のSpike数が大幅に増え，代わりに雑音を付加していない部分（非改ざん箇所）のSpike数が減少する．結果的に，ターゲット信号の改ざん箇所に対応するバイナリパターンは，Spike数の増加により1の数が増加しBERが上昇する．また同時に，ターゲット信号の非改ざん箇所に対応するバイナリパターンは，Spike数の減少により1の数が減少し，BERが低下する．したがって，改ざん箇所だけでなく，非改ざん箇所のBERも上昇してしまうため，適合率が低下し，これがF値の低下に繋がったと考えられる．

改ざん攻撃の中でも唯一音声信号のサンプル数が変化する切り取り攻撃は，検出率は全て1.0であったが適合率が低く，結果としてF値も低くなった．検出率が1.0となったのは，改ざん攻撃におけるBERの増加が要因と考えられる．切り取り攻撃においては，切り取った箇所以降の信号を前に詰めるため，改ざん箇所以降の信号全てが元の信号と異なる信号になってしまう．さらに，信号長が変わることでバイナリパターンを構成するセグメントの長さも変化するため，結果的に信号全体のBERが増加する．これにより，信号のあらゆる箇所においてBERが閾値を超え改ざん判定されたためTPが増大し，検出率1.0を達成したと考えられる．しかし，同時にFPも増大するため，適合率は音情報処理の有無を問わず0.3

を下回っており，結果としてF値も低い値となっている．

第7章　結論

7.1　まとめ

本研究では，情報圧縮／音声符号化を含む，あらゆる音情報処理に頑健な音響ゼロ電子透かし法の確立を目的とし，ヒトの聴神経発火の情報表現である Spikegram を用いた，聴覚的スペクトル表現に基づく音響ゼロ電子透かし法を提案した．提案法の評価の結果，以下のことが明らかになった．

- 提案法は，音情報処理が施されない場合，誤差なく透かし情報を検出できる．
- 提案法は，音情報処理が施された場合，わずかな誤差が生じるものの頑健に透かし情報を検出できる．
- 提案法は，音響信号（音声信号，音楽信号）に対し利用可能である．
- 提案法は，秘匿情報量を増やすと頑健性が低下するが，その影響は小さい．

ここから，提案法は，透かし情報に真正性情報を利用することで，音響信号の真正性担保に活用できることが示された．さらに，提案法を利用した音声改ざん検出法の評価の結果は，部分的な無音化や声の高さを変えるピッチシフト，別音声による置換などに対し一定の検出能力を示し，提案法が音声改ざん検出技術へ応用可能であることが示唆された．

7.2　残された課題

提案法は，あらゆる音情報処理に対して頑健性が高いことがわかったものの，全ての処理に対して完全に誤差なく透かし情報を検出できるわけではない．さらに改善するためには，バイナリパターンの生成に利用する信号特徴を，より強固な特徴に絞ることが妥当であると考えられるが，特徴を限定しすぎた場合，信号が改ざんされても検出誤りが極端に減りすぎてしまうことが予想される．提案法を用いた改ざん検出では，透かし情報の検出誤り情報をヒントとして用いるため，ある種の脆弱性を持つことが望ましい．この音情報処理に対する頑健性と，改ざん攻撃に対する脆弱性のバランスを保った上で，聴覚的スペクトル表現上の特徴をどのように用いるかは，引き続き検討する必要がある．

今回の提案法を利用した音声改ざん検出法の評価の結果から，信号のサンプル数が変化する切り取り攻撃や，元の音声をかき消すような大きい雑音を付加する攻撃に対して，検出能力が低いことが明らかになった．想定される要因を前章で述べたが，これに対する策として，利用するSpike数を増やす，あるいは信号の特徴（大きさや長さなど）に応じて可変的に閾値を決定する方法が有効であると考えている．本研究では，Spike数を反復初期の100個に制限しているため，切り取りや雑音付加によって本来そこにあったSpikeが拡散・集中することで，他の信号部分にまで検出誤差が伝搬した．元々のSpike数が多ければ，このSpikeの拡散・集中による検出誤差の伝搬の影響を抑えられる．しかし，Spike数を増やすことは，頑健でない（脆弱な）Spikeを増やすことに繋がりかねないため，このトレードオフの関係を考慮し，検討を行う必要がある．

参考文献

[1] 一般社団法人電子情報技術産業協会，“CPS とは．”[Online]. Available: https://www.jeita.or.jp/cps/about

[2] 暦本純一ほか，オーグメンテッド・ヒューマン：AI と人体科学の融合による人機一体、究極の IF が創る未来　エヌ・ティー・エス，2018.

[3] 荒井隆行，森大毅，“小特集「音声は何を伝えているか」にあたって，” 日本音響学会誌，vol. 71，no. 9，pp. 459–460，2015.

[4] 森大毅，前川喜久雄，粕谷英樹，音声は何を伝えているか　コロナ社，2014.

[5] 森勢将雅，音声分析合成　コロナ社，2018.

[6] Sisman, B., Yamagishi, J., King, S., and Li, H., “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[7] Wijethunga, R., Matheesha, D., Noman, A. A., De Silva, K., Tissera, M., and Rupasinghe, L., “Deepfake audio detection: A deep learning based solution for group conversations,” in *2020 2nd International Conference on Advancements in Computing (ICAC)*, vol. 1, 2020, pp. 192–197.

[8] Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., Wang, T., Tian, Z., Bai, Y., Fan, C., Liang, S., Wang, S., Zhang, S., Yan, X., Xu, L., Wen, Z., and Li, H., “Add 2022: the first audio deep synthesis detection challenge,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9216–9220.

[9] 鵜木祐史，西村竜一，西村明，近藤和弘，薗田光太郎，音響情報ハイディング技術　コロナ社，2018.

[10] Wen, Q., SUN, T.-f., and Wang, S.-x., “Concept and application of zero-watermark,” *ACTA ELECTONICA SINICA*, vol. 31, no. 2, p. 214, 2003.

[11] Chen, N., and Zhu, J., “A robust zero-watermarking algorithm for audio,” *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–7, 2007.

[12] 川西隆仁，“音響指紋技術とデジタルメディアコンテンツ流通への適用事例，” 情報の科学と技術，vol. 69，no. 5，pp. 189–193，2019.

[13] Cvejic, N., and Seppanen, T., “Increasing robustness of lsb audio steganography using a novel embedding method,” in *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, vol. 2, 2004, pp. 533–537 Vol.2.

[14] Boney, L., Tewfik, A., and Hamdy, K., “Digital watermarks for audio signals,” in *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, 1996, pp. 473–480.

[15] Unoki, M., “Audio information hiding techniques based on human auditory characteristics,” *IEICE ESS Fundamentals Review*, vol. 13, no. 4, pp. 284–293, Apr. 2020.

[16] Gruhl, D., Lu, A., and Bender, W., “Echo hiding,” in *Proceedings of the First International Workshop on Information Hiding*. Berlin, Heidelberg: Springer-Verlag, 1996, p. 293–315.

[17] Ko, B.-S., Nishimura, R., and Suzuki, Y., “Time-spread echo method for digital audio watermarking,” *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 212–221, 2005.

[18] Unoki, M., and Hamada, D., “Method of digital-audio watermarking based on cochlear delay characteristics,” *International Journal of Innovative Computing, Information and Control*, vol. 6, 03 2010.

[19] 鵜木祐史，宮内良太，“蝸牛遅延特性に基づいた音響電子透かし，” 日本音響学会誌，vol. 71，no. 1，pp. 15–22，2014.

[20] 西村竜一，鈴木陽一，“周期的位相変調に基づく音響電子透かし，” 日本音響学会誌，vol. 60，no. 5，pp. 268–272，2004.

[21] 鵜木祐史，ワンシュンベイ，宮内良太，“蝸牛遅延に基づいた電子音響透かし法を利用した音声信号の改ざん検出の検討，” 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報，vol. 112，no. 420，pp. 65–70，01 2013.

[22] Wang, S., Yuan, W., Wang, J., and Unoki, M., "Detection of speech tampering using sparse representations and spectral manipulations based information hiding," *Speech Communication*, vol. 112, pp. 1–14, Sep. 2019.

[23] Galajit, K., Karnjana, J., Unoki, M., and Aimmanee, P., "Semi-fragile speech watermarking based on singular-spectrum analysis with CNN-based parameter estimation for tampering detection," *APSIPA Transactions on Signal and Information Processing*, vol. 8, no. 1, 2019.

[24] Ali, Z., Imran, M., Alsulaiman, M., Zia, T., and Shoaib, M., "A zero-watermarking algorithm for privacy protection in biomedical signals," *Future Generation Computer Systems*, vol. 82, pp. 290–303, May 2018.

[25] Tsai, S.-M., "A robust zero-watermarking scheme for digital audio," *International Journal of Information and Electronics Engineering*, vol. 5, no. 2, pp. 117–121, 2015.

[26] Gao, L., Zhao, W., Wen, X., and Wang, L., "An audio zero-watermarking algorithm based on FFT," in *2010 International Conference on Networking and Digital Society*. IEEE, May 2010.

[27] Electa Alice Jayarani, A., Bhatt, M. R., and Geetha, D., "Zero watermarking on audio based on stft," in *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, 2018, pp. 253–256.

[28] Xi, Z., Xianghong, T., and Hengli, Y., "Zero-watermark scheme based on audio's statistical character," in *2007 International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications*, 2007, pp. 1227–1230.

[29] Yang, Y., Lei, M., Cheng, M., Liu, B., Lin, G., and Xiao, D., "An audio zero-watermark scheme based on energy comparing," *China Communications*, vol. 11, no. 7, pp. 110–116, July 2014.

[30] Yigiun, X., and Rangding, W., "An audio zero-watermark algorithm combined dct with zernike moments," in *2008 International Conference on Cyberworlds*, 2008, pp. 11–15.

[31] Wang, K., Li, C., and Tian, L., "Audio zero watermarking for mp3 based on low frequency energy," in *2017 6th International Conference on Informatics, Electronics and Vision 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, 2017, pp. 1–5.

[32] Panda, J., Choudhary, S., Nath, K., and Kumar, S., “Audio zero watermarking scheme based on sub band mean energy comparison using DWT-DCT,” in *2016 International Conference on Signal Processing and Communication (ICSC)*. IEEE, Dec. 2016.

[33] Painter, T., and Spanias, A., “Perceptual coding of digital audio,” *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[34] Erfani, Y., Pichevar, R., and Rouat, J., “Audio watermarking using spikegram and a two-dictionary approach,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 840–852, Apr. 2017.

[35] 森周司，香田徹，日比野浩，任書晃，倉智嘉久，入野俊夫，鵜木祐史，鈴木陽一，牧勝弘，津崎実，聴覚モデル　コロナ社，2011.

[36] 古川茂人，堀川順生，入野俊夫，鈴木陽一，飯田一博，津崎実，柏野牧夫，小澤賢司，森周司，北川智利，日高聡太，坂田俊文，白石君男，聴覚　コロナ社，2021.

[37] 大串健吾，音響聴覚心理学　誠心書房，2019.

[38] 入野俊夫，“はじめての聴覚フィルタ (やさしい解説)，” 日本音響学会誌，vol. 66，no. 10，pp. 506–512，2010.

[39] Tran, D. K., and Unoki, M., “Matching pursuit and sparse coding for auditory representation,” *IEEE Access*, vol. 9, pp. 167084–167095, 2021.

[40] Huber, R., and Kollmeier, B., “Pemo-q-a new method for objective audio quality assessment using a model of auditory perception,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[41] ISO/IEC 13818-7:2006, “Information technology - generic coding of moving pictures and associated audio information - part 7: Advanced audio coding (aac),” 2006. [Online]. Available: https://www.iso.org/standard/43345.html

[42] 守谷健弘，音声符号化　コロナ社，1998.

[43] 尾知博，川村新，黒崎正行，ディジタル音声&画像の圧縮/伸張/加工技術　CQ出版社，2013.

[44] ISO/IEC 11172-3:1993, “Information technology - coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s - part 3: Audio,” 1993. [Online]. Available: https://www.iso.org/standard/22412.html

[45] 日本音響学会，毎日新聞社，“日本音響学会 新聞記事読み上げ音声コーパス（JNAS），”2006．[Online]. Available: https://doi.org/10.32130/src.JNAS

[46] 後藤真孝，橋口博樹，西村拓一，岡隆一，“RWC 研究用音楽データベース: 音楽ジャンルデータベースと楽器音データベース，” 情報処理学会研究報告音楽情報科学（MUS），vol. 2002，no. 40（2002-MUS-045），pp. 19–26，2002.

[47] 後藤真孝，橋口博樹，西村拓一，岡隆一，“RWC 研究用音楽データベース: クラシック音楽データベースとジャズ音楽データベース，” 情報処理学会研究報告音楽情報科学（MUS），vol. 2002，no. 14（2001-MUS-044），pp. 25–32，2002.

[48] Vaidya, S. P., “Fingerprint-based robust medical image watermarking in hybrid transform,” *The Visual computer*, pp. 1–16, Jan. 2022.

[49] Morise, M., Yokomori, F., and Ozawa, K., “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[50] Morise, M., “D4c, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, Nov. 2016.

謝辞

本研究の遂行にあたり，常に的確なご指導とご助言を賜りました指導教官の鵜木祐史教授に深く感謝いたします．また，定期的な研究室会議の場に限らず，日々の研究活動において様々なご助言を賜りました赤木正人先生，木谷俊介講師，上江洲安史特任助教，大田恭士修士，磯山拓都修士に心から感謝いたします．そして公私にわたりお世話になりました鵜木研究室の皆様に改めてお礼申し上げます．最後に，長い学生生活を精神面・金銭面で支えてくれた家族に心から感謝を申し上げます．

研究業績

発表

1. 市川敦暉，鵜木祐史，“聴覚的スペクトル表現に基づく音響ゼロ電子透かし法の検討，” 2022 年度 電気・情報関係学会北陸支部連合大会，G-21，2022.
2. 市川敦暉，鵜木祐史，“聴覚的スペクトル表現に基づく音響ゼロ電子透かし法，” 電子情報通信学会技術研究報告書，EMM2022-65，pp. 31–36，2023.

受賞

1. 市川敦暉，2022 年度 電気・情報関係学会北陸支部連合大会 音響部門優秀発表賞，2022 年 9 月.

付録　Gammachirpフィルタバンクを用いた聴覚的スペクトル表現に基づいた音響ゼロ電子透かし法

Gammachirp フィルタは，ヒトの聴覚の非線形性を考慮した聴覚フィルタであり，以下の式で計算される．

$$g(t) = at^{l-1}\exp(-2\pi b\mathrm{ERB_N}(f_c)t)\cos(2\pi f_c t + c\ln t + \phi) \tag{7.1}$$

このとき，t $(t > 0)$ は時間，a は振幅である．また，l はフィルタ次数，b は帯域幅係数，c は周波数変化の係数，ϕ は位相である．ガンマトーンのインパルス応答との違いは，$c\ln t$ のみで，係数 $c = 0$ のとき等しくなる．

Gammachirp フィルタバンクを用いた音響ゼロ電子透かし法（GCFB）の透かし情報の埋め込みは，以下の手順で行う．

Step. 1 $x(n)$ から $G(k,n)$ を導出する．このとき G は，k 個の異なる中心周波数および帯域幅 $\mathrm{ERB_N}$（Hz）を持つ Gammachirp フィルタバンクの出力である．

Step. 2 $G(k,n)$ を時間方向に M 個のセグメントで分割する．各セグメントのパワー（振幅二乗値）の平均が $x(n)$ 全体のパワー平均を上回れば 1，下回れば 0 とする．この二値化処理によって，フィルタバンク出力 $G(k,n)$ からバイナリパターン $B(k,m)$ への変換を行う．

Step. 3 $B(k,m)$ と $W(k,m)$ の排他的論理和（XOR）を計算し，$Y(k,m)$ を得る．このとき，W もバイナリ行列である．

GCFB 法における透かし情報の検出は，以下の手順で行う．

Step. 1 透かし情報を検出したい対象となるターゲット信号を $x(n)$，埋め込み処理で作成された検出鍵を $Y(k,m)$ とする．

Step. 2 埋め込み処理の Step. 1 と同様に，$G(k,n)$ を導出する．

Step. 3 埋め込み処理の Step. 2 と同様に，$G(k,n)$ を $B(k,m)$ に変換する．

Step. 4 $B(k,m)$ と検出鍵 $Y(k,m)$ の XOR を計算し，透かし情報 $W(k,m)$ を検出する．