修士論文

特異スペクトル分析に基づいた音響ゼロ電子透かし

渡辺 瑠伊

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７年６月

Abstract

Digital content, encompassing audio, images, and video, has become an indispensable component across numerous domains in modern society, ranging from entertainment and communication to legal evidence and security systems. However, the inherent ease with which digital audio files, in particular, can be perfectly copied, seamlessly edited, and widely redistributed presents significant challenges. Consequently, issues related to copyright infringement, unauthorized usage, and critically, malicious tampering for disinformation or evidence manipulation, have escalated into serious societal and technological concerns. Traditional audio watermarking techniques attempt to address these by directly embedding identification or authentication data into the host audio signal itself. While conceptually straightforward, this approach often introduces perceptible distortions, degrading the listening experience, and crucially, the embedded watermarks frequently lack resilience against common signal-processing operations (attacks), whether incidental or intentional, such as compression, filtering, or noise addition.

To overcome these inherent limitations of conventional embedding, zero-watermarking schemes have emerged as a promising alternative. These methods fundamentally differ by not modifying the host signal. Instead, they generate a binary pattern, often termed a watermark key or robust hash, derived from perceptually significant, intrinsic features of the audio content itself, combined with an external secret key. This generated pattern is then securely stored alongside the original audio. For verification, the process is repeated on the potentially altered audio, and the newly derived pattern is compared to the stored original. This approach intrinsically preserves the original audio fidelity while providing a mechanism for robust authenticity and integrity verification.

This study focuses on the development and evaluation of a zero-watermarking framework specifically designed for audio signals, particularly speech. The primary objective is to achieve a delicate balance: the framework must exhibit strong robustness against common, benign processing operations like lossy compression and resampling, which are frequent in standard distribution channels, yet simultaneously possess deliberate fragility, meaning it should reliably detect malicious tampering attacks intended to alter the content's meaning or authenticity. The proposed method uniquely leverages the inherent sparsity characteristic of speech signals. It employs singular spectrum analysis (SSA), a powerful technique for time series analysis, to decompose the signal and extract temporally stable, low-rank component features that are less susceptible to noise and common distortions. These stable features form the basis for subsequent binary pattern generation, designed to capture the essential characteristics of the speech segment.

The performance of the proposed SSA based zero-watermarking system is evaluated against a notable baseline method developed by Ichikawa and Unoki, which utilizes auditory spectrum features. The primary evaluation metric employed is the bit error rate (BER) calculated between the original binary pattern and the pattern extracted from the processed or potentially tampered audio. Within this evaluation context, a BER value at or below 1% (BER $\leq$ 1%) is considered indicative of successful watermark detection, confirming the audio's authenticity despite potential benign processing. Conversely, significantly higher BER values are interpreted as evidence of potential tampering or unacceptable levels of distortion.

Experimental results demonstrate that the proposed method successfully achieves the target BER $\leq$ 1% under several common processing conditions, including MP3 and AAC compression, resampling operations, and requantization, thereby confirming its robustness to these specific modifications. However, the system exhibits significantly higher BER, indicating watermark failure, when subjected to low-bitrate speech codecs such as G.711 and G.719, suggesting that the features extracted by SSA are sensitive to the specific type of quantization and signal representation used by these codecs. While the proposed method shows performance improvements over the baseline for MP3 and AAC robustness, its overall robustness across a wider range of operations is found to be somewhat inferior to the established conventional method. Furthermore, a critical limitation emerged regarding tampering detection: the system demonstrated excessive robustness, meaning that even after malicious modifications were applied, the extracted features did not change sufficiently. This resulted in an insufficient escalation of the BER, leading to low detection accuracy for certain types of malicious tampering.

Future research directions will refining the zero-watermarking framework to address these limitations. A key focus will be the redesign of the key-generation process, specifically targeting the binarization stage. The current binarization approach appears to lack sufficient temporal localization, making it overly resilient to localized tampering. The goal is to introduce the necessary fragility to reliably detect intentional attacks by incorporating finer temporal details or adaptive thresholding mechanisms, while carefully maintaining the achieved robustness against common, benign signal processing operations essential for practical applications.

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

近年，情報通信技術の急速な発展とインターネットの普及に伴い，音声，画像，映像などのマルチメディアコンテンツが爆発的に増加している．これらのコンテンツは，私たちの社会において，コミュニケーション，エンターテイメント，教育，ビジネスなど，様々な分野で重要な役割を果たしている．特に，音響信号は，音楽配信サービス，ポッドキャスト，音声アシスタント，オンライン会議システムなど，日常生活でも多く利用されている．例えば，音楽ストリーミングサービスでは，無数の楽曲が配信され，ユーザーは聴きたい音楽を楽しむことができる．また，音声アシスタントは，スマートスピーカーやスマートフォンに搭載され，音声による情報検索やIoT端末の操作を可能にしている．オンライン会議システムは，離れた場所にいる人々がリアルタイムかつ遠隔でコミュニケーションを行うための重要なツールとなっている．このように，音響信号を用いたサービスは，現代社会において重要な役割を担っている．

しかしながら，デジタル情報は，容易に複製，編集，再配布が可能であるという特性を持つため，著作権侵害や不正利用，改ざんなどの問題が生じている．例えば，音楽ファイルがインターネット上で違法に共有されたり，音声記録が改ざんされて裁判の証拠としての信頼性が損なわれたりするなどの事例が発生している．したがって，デジタルコンテンツの真正性，完全性，および出所を保証するための技術が必要不可欠である．

音響信号の真正性保証は，特に重要な課題である．なぜなら，音響信号は，音楽作品などの著作物の保護だけでなく，音声記録による証拠としての利用，医療診断における音声分析，通信における音声伝送など，様々な分野で利用されているからである．例えば，裁判における音声記録の改ざんは，司法の公正性を大きく損なう可能性がある．また，医療分野における音声分析の誤りは，患者の診断や治療に重大な影響を与える可能性がある．さらに，音楽業界においては，デジタル音楽の違法コピーや改ざんが，アーティストの権利を侵害し，音楽市場の健全な発展を妨げる深刻な問題となっている．このように，音響信号の真正性を保証することは，社会の様々な側面において，公正性，信頼性，および安全性を確保するために不可欠な要素となっている．

音響信号の真正性を保証する技術の一つとして，電子透かし技術が広く研究されている．電子透かし技術は，著作権保護，改ざん検知，放送監視など，様々な

目的のために，音響信号に特定の情報を埋め込む技術である．従来の電子透かし技術では，音響信号に直接透かし情報を埋め込むため，埋め込みによる音質の劣化は避けられないことや，様々な音情報処理や攻撃に対する脆弱性などの課題が存在する．例えば，圧縮，ノイズ付加，フィルタ処理などの一般的な信号処理操作によって，埋め込まれた透かし情報が正しく検出できなくなる場合がある．また，悪意のあるユーザーによる電子透かしの除去や改ざんも懸念される．音質の劣化は，特に音楽コンテンツなどの高品質な音響信号を扱う場合には，大きな問題となる．また，脆弱性は，電子透かし技術の信頼性を損なうため，著作権保護などの真正性証明が不可能となる．

これらの問題を解決するために，近年，音響ゼロ電子透かし技術が提案されている．音響ゼロ電子透かし技術は，電子透かしを音響信号自体に埋め込むのではなく，電子透かし情報と音響信号の特徴量から生成されるバイナリパターンを用いて検出鍵を生成する．この検出鍵を用いて，音響信号から電子透かしを検出する．音響ゼロ電子透かし技術は，音響信号自体に電子透かしを埋め込まないため，従来の電子透かし技術と比較して，音質の劣化がしないという利点がある．また，信号の特徴量に基づいて検出鍵を生成するため，信号処理や攻撃に対する耐性を高めることができる．図 1.1 に音響ゼロ電子透かし技術の具体例を示す．この技術では，ユーザが保護したい音声データとそれを証明するための透かし情報を用いて検出鍵を作成する．そして，音声データがインターネット上に公開され，悪意を持った第三者が改ざん処理を加えた場合，持ち主である本人が生成した鍵を用いて照合することで，なりすましや改ざんの検知が可能となる．また，データが複製され二次的に流布された場合も，検出鍵を用いて同一のデータであることを証明することで，著作権の保護が可能となる．その他の応用例として，話者認証にも利用される．予め端末に音声データと検出鍵を登録しておくことで，端末の所有者の声のみに反応するような認証技術が実現可能となる．

音響ゼロ電子透かし技術においても，音響信号が様々な信号処理や改ざんを受けた場合に，電子透かしの検出精度が低下したり，誤検出が発生したりする可能性がある．特に，音響信号に対する悪意のある改ざんに対して，ロバスト性を確保することが重要な課題となる．例えば，特定の周波数成分の除去や追加，時間軸方向の歪みなど，様々な改ざん方法に対して，安定した電子透かし検出を可能にする必要がある．悪意のあるユーザーは，電子透かしを無効化したり，誤検出を発生させたりするために，様々な改ざんを試みる可能性がある．したがって，これらの改ざんに対してロバストな電子透かし技術を開発することが重要となる．

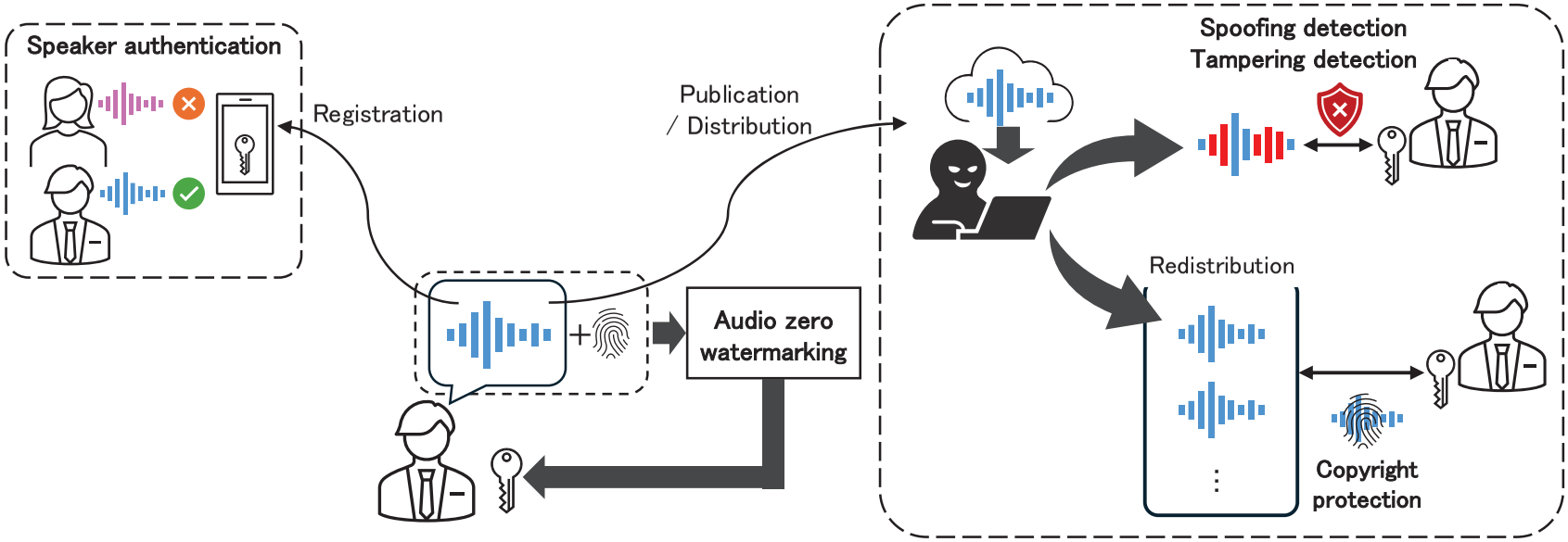

図 1.1: 音響ゼロ電子透かしの応用例

1.2 研究背景

電子透かし技術では，デジタルデータに対して特定の情報を埋め込むことで，その真正性や著作権の保護が可能となる．その中で，音楽や音声といった音響信号の真正性を担保するために，これまで様々な音響電子透かし法が提案されてきた．典型的な方法として，音声や音響信号をデジタル化したデータ（音声データ）のヘッダ部分に真正性を示す情報を埋め込む方法がある．音データは，音響信号そのものに加えて，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の属性情報をヘッダ領域に含んでいる．例えば，WAV 形式の音データにはファイルサイズ情報やチャンネル数，バイトレート，ブロックサイズ，量子化ビット数などの基本属性が含まれる．これらの情報は音データを適切に再生するために必要な基本的属性である．この方法では，埋め込む情報をヘッダ領域に記録することで音響電子透かしを実現している．この方法は実装が容易であり音質低下も避けられるが，ヘッダ領域にある情報だけを意図的に削除されると容易に不正が行えるという弱点がある．

この問題を克服するために，ヘッダ領域ではなく，音響信号自体に情報を付加する音響電子透かし法が提案された．音響電子透かしの基本的な仕組みは次のとおりである．まず，対象となる音響信号に透かし情報（真正性情報）を埋め込む．透かし情報には，文字（バイナリデータ）や画像などが利用でき，著作者の情報や個人情報を証明するために使用される．次に，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．この時，正しく透かし情報が検出されれば，その音響信号と元の信号との同一性が証明でき，真正性が担保される．一方で，正しく透かし情報が検出されない場合，音響信号に何らかの攻撃が加えられているということが分かる．

従来の音響電子透かし技術は，音響信号に直接透かし情報を埋め込む方式であるため，音質の劣化が避けられないという課題がある．また，圧縮やノイズ付加などの信号処理や，悪意のある改ざんに対して脆弱であるという問題も指摘されている．音響電子透かし技術の研究は，初期段階では，主に画像信号を対象とした電子透かし技術を応用する形で発展してきた．その後，周波数領域に透かし情報を埋め込む方法や，人間の聴覚特性を利用して透かし情報を知覚されにくくする方法など，様々な技術が提案されてきた．近年では，機械学習は深層学習を用いた電子透かし技術も研究されている．これらの技術は，従来の電子透かし技術と比較して，高い秘匿性と頑健性を両立できる可能性があり，注目を集めている．

このような電子透かし技術に基づき，新たな発想を取り入れたゼロ電子透かし技術も提案されている．ゼロ電子透かしについても，画像処理分野や音響信号処理分野にわたって広く利用されている．音響信号に対して適用させた音響ゼロ電子透かしの基本的な仕組みは次のとおりである．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，音響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，音響信号に影響を与えな

い形式で透かし情報を埋め込むことができる．そして，この検出鍵を用いて，音響電子透かしと同様に真正性の担保から音響信号に破壊及び攻撃が加えられていることを検出できる．音響ゼロ電子透かしでは，音響信号のどういった特徴量を利用するか，つまりバイナリパターンの生成法を検討することで，秘匿情報量も制御できる．このような仕組みによって，音響電子透かしにおいて課題となっていた，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決することが期待できる．

音響ゼロ電子透かし技術は，音響信号に直接透かし情報を埋め込む従来の音響電子透かし技術とは異なり，音質の劣化を与えずに透かし情報を検出できる利点がある．しかしながら，音響ゼロ電子透かし技術においても，信号処理や改ざんに対する頑健性の向上が重要な課題となっていおり，現在も研究が行われている．

1.3 研究目的

本研究では，音声信号について，出所の保証や著作権保護，及びなりすまし検知に応用できる音響ゼロ電子透かし法の確立を目的とする．そのために，音声信号に対して圧縮符号化やリサンプリングといった処理が加えられたとしても，透かし情報が正しく検出できる頑健性を有することが必要とされる．一方で，改ざん攻撃や音声信号を破壊するような悪意を持った処理に対しては，抽出される透かし情報が崩れることで，それらを検知する．つまり，特定の攻撃に対しては脆弱である必要がある．これらの要求を満たすような音響ゼロ電子透かし法は幅広いデータに応用でき有用性が高いといえる．

このような頑健性及び脆弱性の特性は，音響ゼロ電子を行う際の音響信号の特徴量選択によって変化する．つまり，音声信号信号からどのような特徴量を抽出すれば，求められる頑健性と脆弱性のバランスを取ることができるのかを考えることが重要となる．そこで，音声信号の持つスパース性に着目し，スパースコーディングによる特徴量抽出を行うことで，音声信号の本質的な特徴を捉えることが可能となり，本研究の目的を達成できると考えられる．

1.4 論文構成

本論文は 6 章で構成される．図 1.2 に全体構成を示す．第 1 章では，電子透かし及び音響電子透かしの研究背景及び研究目的について述べる．第 2 章では，音響電子透かし及び音響ゼロ電子透かしに関する研究を示す．第 3 章では，音響ゼロ電子透かし技術の課題及び要求項目を説明し，それらを解決する方法論を示す．第 4 章では，特異スペクトル分析を用いた音響ゼロ電子透かし法について具体的な説明を行う．第 5 章では，本論文で提案した方法による実験内容を説明し，実験結

果から提案法の評価及び考察を行う．第 6 章では，本研究の結論と残された課題について述べる．

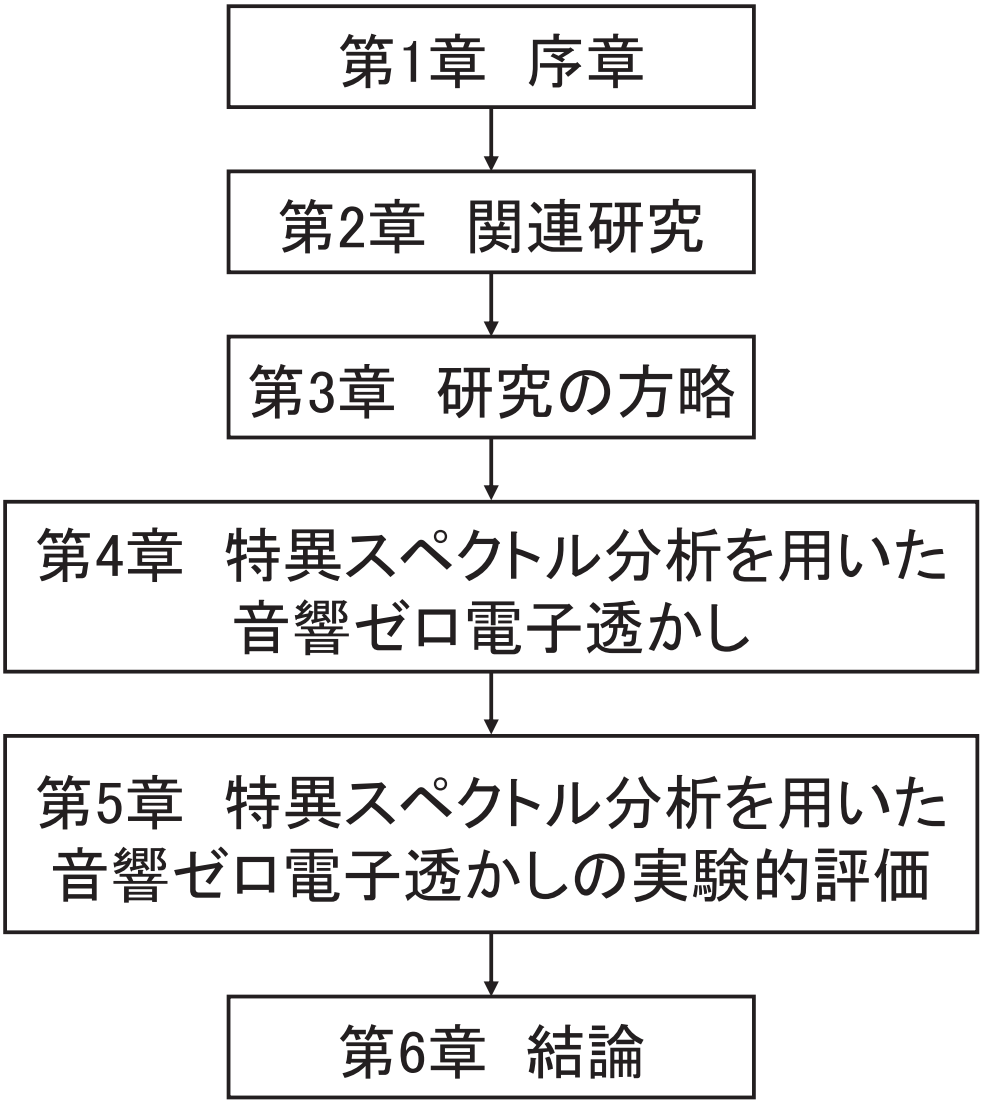
第1章　序章
第2章　関連研究
第3章　研究の方略
第4章　特異スペクトル分析を用いた
音響ゼロ電子透かし
第5章　特異スペクトル分析を用いた
音響ゼロ電子透かしの実験的評価
第6章　結論

図 1.2: 本論文の構成

第2章　関連研究

2.1　音響電子透かしの既存方法

音響電子透かしとは，音楽や音声といった音響信号に対し何らかの情報を埋め込み，抽出を行う技術である．音響電子透かしの方法は，大きく分けて時間領域における方法と周波数領域における方法に分類される．時間領域における方法は，音響信号の振幅や位相に直接変更を加えることで情報を埋め込む方法である．代表的な方法として，LSB (Least significant bit) 置換法 [1] がある．これは音響信号のサンプル値の LSB を透かし情報で置き換える方法であり，非常に単純な方法で処理負荷が低いのが特徴である．しかし，音質劣化が比較的大きく，耐攻撃性も低いという欠点がある．エコーハイディング法 [2, 3, 4] は，原音響信号に遅延と減衰を施したエコーを付加することで情報を埋め込む方法である．エコーの遅延時間や減衰率を調整することで，複数の情報を埋め込むことも可能である．人間の聴覚特性を利用することで，知覚されにくい形で情報を埋め込むことができるが，攻撃耐性は低い傾向にある．スペクトル拡散法 [5, 6, 7] は，透かし情報を拡散符号を用いて広帯域に拡散し，原音響信号に重畳することで情報を埋め込む方法である．埋め込む情報量が少ない一方で，耐攻撃性が高いことが特徴である．しかし，拡散処理により音質劣化が起こる可能性がある．振幅変調法 [8] は，原音響信号の振幅を透かし情報に応じて変調させることで情報を埋め込む方法である．比較的単純な方法であり，処理負荷も低いが，耐攻撃性はあまり高くない．また，人間の内耳に存在する蝸牛の周波数特性を利用した蝸牛遅延に基づく方法 [9, 10] も提案されている．この方法では，人間の聴覚特性に基づいているため，知覚的な歪みを抑えつつ，情報の埋め込みを行うことが可能となっている．

一方，周波数領域における方法は，音響信号をフーリエ変換やコサイン変換などの変換を施して周波数領域に変換し，その周波数成分に対して変更を加えることで情報を埋め込む方法である．周波数領域の方法として，周波数マスキング法 [11, 12] がある．これは，人間の聴覚特性である周波数マスキング効果を利用し，強い信号の近傍周波数に存在する微弱な信号が知覚されにくくなる現象を利用するものである．具体的には，マスキング効果が発生する範囲に透かし情報を埋め込むことで，知覚されにくい情報の埋め込みを実現する．この方法は，聴覚心理モデルに基づいているため，知覚的な透明性を高める上で効果的である．位相変調法 [13] は，周波数成分の位相を透かし情報に応じて変調させることで情報を埋め込む方法である．人間の聴覚は位相の変化に鈍感であるため，音質劣化が比較

的少ないという特徴がある．量子化インデクス変調法 [14, 15] は，周波数成分を量子化する際の量子化インデックスを透かし情報に応じて変更することで情報を埋め込む方法である．耐圧縮性に優れているという特徴がある．これらの方法はそれぞれ一長一短があり，用途や求められる性能 (音質，埋め込み容量，耐攻撃性など) に応じて適切な方法を選択する必要がある．近年では，これらの方法を組み合わせたり，機械学習を用いたりすることで，より高性能な音響電子透かし法の開発が進められている．さらに，人間の聴覚心理モデルを利用して，知覚的な透明性を高める工夫もなされている．

2.2 音響ゼロ電子透かしの既存方法

音響ゼロ電子透かしは，音響信号自体に改変を加えることなく，対象信号から抽出された特徴量，例えばハッシュ値等を透かし情報として利用する技術である．図 2.1 に音響電子透かし及び音響ゼロ電子透かしの概略図を示す．音響電子透かしでは，対象となる信号に対して直接情報を埋め込まれる．音響信号への情報の埋め込みに際して，原理的に音質への影響を完全に排除することが困難であり，埋め込み情報量と音質劣化との間にトレードオフの関係が存在する．音響ゼロ電子透かしは，この課題を解消する有効な方法として位置づけられる．図 2.1（b）で示されるように，音響信号の特徴量と透かし情報により鍵を生成し，鍵を利用することによって透かし情報が検知される．音響ゼロ電子透かしの代表的なものとして，特徴量ベースの方法が挙げられる．これは音響信号から抽出した特徴量を基に，ハッシュ値を生成したり，バイナリパターンと呼ばれる二値（0 と 1）の特徴パターンを構築したりするものである．特徴量の選択や設計は，ロバスト性と検出能力のバランスを考慮して行う必要がある．離散ウェーブレット変換（discrete wavelet transform: DWT）と離散コサイン変換（discrete cosine transform: DCT）を利用して特徴量を抽出する方法 [16, 17] では，時間周波数領域における音響信号の特徴を捉えることができ，計算コストを抑えつつ頑健な音響ゼロ電子透かしを実現している．また，DWT，DCT に加えて特異値分解を組み合わせた方法 [18] では，DWT の近似係数から生成された行列に対して特異値分解を適用し，得られた特異値を特徴量として利用することでバイナリパターンを生成している．このように音響特徴量を行列分解しバイナリパターンを生成する方法として，非負値行列因子分解を用いた方法 [19] や K-特異値分解 [20] を用いた方法などがある．近年では，深層学習を用いた音響ゼロ電子透かしも提案されており，

これらの方法は，それぞれ異なる特性を持っており，利用する特徴量に応じて，適切モデル設計を行うことが重要である．また，複数の方法を組み合わせたり，パラメータを最適化したりすることで，より高性能なゼロ電子透かしシステムを構築することが可能である．以上で述べた音響電子透かし及び音響ゼロ電子透かしの特性，それらの代表的な方法についてまとめたものを表 2.1 に示す．音響電子透

かし及び音響ゼロ電子透かしを実用するうえでは，これらの特性を理解し適切な方法を選択する必要がある．

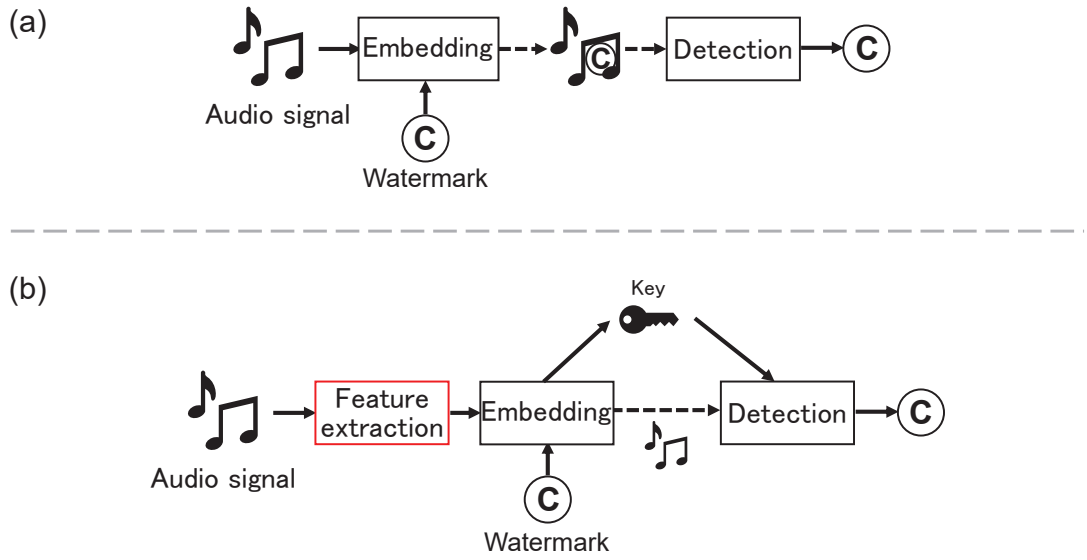

図 2.1: 音響電子透かしと音響ゼロ電子透かしの概略図：(a) 音響電子透かし, (b) 音響ゼロ電子透かし

表 2.1: 音響電子透かしと音響ゼロ電子透かしの特徴と比較

	Audio watermarking	**Audio zero-watermarking**
Principle	Direct embedding	Feature utilization
Signal change	Yes	No
Affects sound quality	Yes	No
Information capacity	High	Low
Main uses	Copyright protection, Tamper detection	Audio authentication, Identity verification
Robustness	Embedding dependent	Feature dependent
Methods	LSB, Echo hiding, Frequency masking, ...	DWT-DCT, NMF, K-SVD
Advantages	High information capacity, Versatile	Maintains audio quality
Disadvantages	Audio degradation	Limited information capacity

2.3 聴覚的スペクトル表現に基づいた音響ゼロ電子透かし

2.3.1 本方法の概要

市川・鵜木は，人間の聴覚特性を考慮した音響ゼロ電子透かし法 [21] を提案している．この方法は，従来の音響ゼロ電子透かし技術におけるロバスト性の課題に注目している．特に，情報の通信や保存過程で不可避な情報圧縮（MP3，AAC など）や音声符号化（G.711，G.729 など）といった処理に対して頑健性を向上させることを目指している．従来の音響ゼロ電子透かし技術では，音響信号の特徴量として，主に時間領域や周波数領域におけるエネルギーや振幅などの物理的な特徴量が用いられてきた．これらの特徴量は，情報圧縮などの信号処理によってデジタル信号そのものが大きく変化すると，特徴量も変動してしまい，電子透かしの検出精度が低下する可能性がある．

市川・鵜木の方法では，この課題を解決するため，人間の聴覚にとって本質的な信号要素は情報圧縮などの影響を受けにくいという考えに基づき，Spikegram と呼ばれる特徴量を採用している．Spikegram は，聴覚神経における神経細胞の発火パターンを模した聴覚的スペクトル表現である．Spikegram では，音響信号 $x(t)$ を，式 (2.1) のように，スパースコーディングによって複数の基底 $g_k(t)$ の組み合わせで表現する．

$$x(t) = \sum_{k=0}^{\infty} \langle R^k x(t), g_k(t) \rangle g_k(t) \tag{2.1}$$

この式において，$R^k x(t)$ は $k-1$ 回目までの反復処理で再構成された信号成分を元の信号から引いた残差信号であり，$\langle \cdot \rangle$ は残差信号と基底との内積（射影）を表す． この射影係数が Spike の強度となる． 基底 $g_k(t)$ には，ヒトの聴覚フィルタを近似した Gammatone フィルタが用いられ，以下の式で定義される．

$$g(t) = at^{n-1} \exp(-2\pi b \mathrm{ERB}_N(f_c)t) \cos(2\pi f_c t + \phi) \tag{2.2}$$

ここで a は振幅，n はフィルタ次数，b は帯域幅係数，f_c は中心周波数，ϕ は位相を表す． 帯域幅を決定する $\mathrm{ERB}_N(f_c)$（等価矩形帯域幅）は，人間の聴覚の周波数分解能をモデル化したものであり，中心周波数 f_c に依存して変化する．

Spikegram は反復更新処理により，発火点である Spike を順次導出する． 反復回数毎に得られる Spike の強度は，反復回数が増えると指数関数的に小さくなる特性がある．特に，反復処理の初期に得られる Spike は強度が大きく，信号にとって最も重要で本質的な要素であるため，音情報処理を施しても発火位置が変化しにくいことが実験的に確認されている．

本手法ではこの性質を利用し，反復初期に得られる強度の強い 100 個の Spike のみを用いてバイナリパターンを作成する．バイナリパターンの二値化処理では，

Spikegram を時間方向に一定区間のセグメントに分割し，セグメント内に Spike が一つでも存在している場合は ‘1‘，存在しない場合は ‘0‘という値を設定し，0/1 の二値を持つバイナリパターンを作成している．このバイナリパターンと透かし情報（二値画像）との排他的論理和（XOR）によって検出鍵を作成する．

2.3.2 評価と応用

シミュレーション評価の結果，本手法は量子化，リサンプリング，及び MP3・AAC 符号化など様々な音情報処理への高い頑健性を持つことが実験を通じて示されている．特に，従来の手法が苦手としていた再量子化や音声符号化（G.711, G.729）に対して，誤り率（BER）を大幅に低減できることが確認された．また，DWT-DCT 法や，振幅情報に着目した方法，及びガンマチャープフィルタバンクを用いた聴覚スペクトル表現に基づく方法といった従来法と比較して，多くの処理で有意に頑健性が向上していることも明らかとなっている．さらに，秘匿情報量（ペイロード）を増加させると頑健性はわずかに低下する傾向が見られたが，その影響は小さく，高い頑健性を維持できることが示された．この頑健性は，音声信号だけでなく音楽信号においても確認されている．

応用として，音声改ざん検出への適用も検討されている．この応用では，音情報処理に対する BER と，意図的な改ざん攻撃（雑音付加，無音化，ピッチシフトなど）に対する BER の差が大きいことを利用する．BER が特定の閾値（本研究では 1.15 評価の結果，ゼロ補間（無音化），ピッチシフト，サンプル置換といった攻撃に対して，一定の検出能力を持つことが確認された．

2.3.3 課題点

従来法では多くの音情報処理に対して頑健性が高いことと，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が，以下の課題点も存在する．

1. 全ての音情報処理で誤差なく透かし情報を検出できない．特に，G.729 符号化や音楽信号については，誤差が生じている．

2. Spike の発火位置のみを特徴として検出鍵を生成しており，Spike 強度を考慮したバイナリパターンの生成法については未検討である．

3. Spike 数が少ないと頑健性が増すが，改ざん攻撃に対する検出感度が低くなる．雑音付与及び切取り攻撃に対しての検出精度が低いことが実験により確認されている．

まず，1. について，従来法では高い頑健性を有しているが，MP3・AAC 符号化，及び G.729 符号化については，リサンプリングや再量子化と比較して，頑健性が

低い傾向が確認されている．従って，圧縮符号化に適した，従来法とは異なるスパースコーディング方法を用いることで性能向上が図れると考える．次に，2. 及び 3. について，バイナリパターンを生成する際に Spike 強度を考慮していない二値化処理を行っていることが原因であると考えられる．バイナリパターンを生成する際に，透かし情報（二値画像）との排他的論理和を取るため，音響信号の特徴量も二値にする必要がある．この時，なるべく元の特徴量を反映したまま二値化できるような変換を行うことができれば，頑健性を保ったまま改ざん検知能力も向上させられると考える．

第3章　研究の方略

3.1　音響ゼロ電子透かしにおける要求項目

音響ゼロ電子透かしでは，信号の同期ずれ，リサンプリング，及び圧縮・符号化といった一般的な音情報処理（非攻撃的処理）に頑健であること．また，音声改ざん処理のような攻撃的処理が行われた場合には，それを検出する必要がある．音響ゼロ電子透かし法は，信号に対しユニークな検出鍵を生成し，この鍵を使用して透かし情報の検出や照合を行う．つまり，非攻撃的処理を行った信号から得られる検出鍵は元の信号のものと一致する必要がある．一方で，攻撃的処理が行われた信号から得られる検出鍵が元の信号のものと異なれば，攻撃を検出することが可能となる．

このような要求を満たす音響ゼロ電子透かし法を確立することが本研究の目的となっている．そこで，本論では，音声信号のスパース性に着目し，スパースコーディングを利用した音響ゼロ電子透かし法を提案する．

3.2　スパースコーディングによる音声信号の表現

スパースとは,「疎」,「まばら」という意味であり，数理的には大部分がゼロであり，非ゼロ要素が少ない構造を指す．具体的には n 次元ベクトル $\boldsymbol{x} \in \mathbb{R}^n$ について，

$$\|\boldsymbol{x}\|_0 = |\{i|x_i \neq 0\}| \tag{3.1}$$

で定義される ℓ_0 ノルムが $\|\boldsymbol{x}\|_0 \ll n$ であれば $\|\boldsymbol{x}\|$ はスパースとなる．スパースコーディングとは，観測信号 $\boldsymbol{y}$ を，基底となる行列 $\boldsymbol{\Phi}$ と係数行列 $\boldsymbol{x}$ を用いて

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} \tag{3.2}$$

という線形モデルで表し，さらに，式 (3.1) のようなスパースな $\boldsymbol{x}$ を推定する方法である．一般的に以下のような ℓ_0 制約付き最適化として定式化される．

$$\min_x \|\boldsymbol{x}\|_0 \ \ \text{subject to} \ \ \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 \leq \epsilon \tag{3.3}$$

ここで，ϵ は許容誤差を表す．音響信号に対しても，観測波形やスペクトログラムの各フレーム $\boldsymbol{y}_t$ を，辞書（基底行列）$\boldsymbol{\Phi} \in \mathbb{R}^{m \times K}$ と係数ベクトル $\boldsymbol{x}_t \in \mathbb{R}^K$ の線形

結合として以下のように適用できる．

$$\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{x}_t \tag{3.4}$$

ここで t は時間フレーム，m はフレーム次元数，K は基底数である．

音声信号では，時間領域で観測される波形が短時間フレーム毎に，ピッチ周期（周期的成分）や破裂音，アタック音といった局所成分を含むため，多くの場合フレーム長さに対し，高エネルギーとなるサンプル数は少ない．従って，音声信号が殆どの時間周波数で 0 に近く，スパースであることが明らかとなっている [22, 23]．従って，式 (3.4) のように表される時，適切な基底を選択すれば，係数ベクトル $\|\boldsymbol{x}_t\|_0$ はスパースな傾向となる．この性質を利用することで，冗長な情報を圧縮しつつ，音声信号を高精度に再構成できる．

3.3 特異スペクトル分析

特異スペクトル分析 (singular spectrum analysis: SSA)[24, 25, 26] は，時系列データの解析に使用されるスパースコーディング方法である．この方法は，時系列データから構造を抽出し，トレンド，周期性，及びノイズに分解する．SSA は，主に次の 4 つのステップから構成される．

1. 埋め込み

まず，長さ N の時系列データ $\{x_n\}_{n=1}^N$ を考え，遅延子 L $(2 \leq L \leq N/2)$ を選択し，時系列を L 次元ベクトルに埋め込む．具体的には，以下のように $K = N-L+1$ 個の遅延ベクトル $X_i \in \mathbb{R}^L$ を生成する．

$$X_i = (x_i, x_{i+1}, \ldots, x_{i+L-1})^T, \quad i = 1, 2, \ldots, K \tag{3.5}$$

次に，これらのベクトルを列ベクトルとして持つ $L \times K$ の軌跡行列 X を構成する．

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{bmatrix} = [X_1 : X_2 : \cdots : X_K] \tag{3.6}$$

軌跡行列は各反斜辺上の要素が全て等しいという特徴をもつ，ハンケル行列となる．

2. 特異値分解

軌跡行列 X に対して特異値分解（singular value decomposition: SVD）を適用する．X の SVD は次のように表される．

$$X = U\Sigma V^T = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T \tag{3.7}$$

ここで，$U = [U_1 : U_2 : \cdots : U_d]$ は $L \times d$ の左特異ベクトルからなる行列，$\Sigma = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d})$ は特異値 $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_d} > 0$ を対角成分に持つ $d \times d$ の対角行列（$d = \mathrm{rank}(X)$），$V = [V_1 : V_2 : \cdots : V_d]$ は $K \times d$ の右特異ベクトルからなる行列である．また，$\{U_i\}$ と $\{V_i\}$ はそれぞれ正規直交基底である．

3. グループ化

$I = \{1, 2, \ldots, d\}$ を m 個の互いに素な部分集合 $I_1, I_2, \ldots, I_m$ に分割する．各部分集合 I_p に対して，対応する行列を $X_{I_p} = \sum_{i \in I_p} \sqrt{\lambda_i} U_i V_i^T$ とする．これにより，元の行列 X を m 個の行列の和に分解できる．

$$X = X_{I_1} + X_{I_2} + \cdots + X_{I_m} \tag{3.8}$$

この操作がグループ化である．各グループは，時系列の異なる成分（トレンド，周期成分，ノイズなど）に対応すると考えられる．

4. 対角平均化

各グループの行列 X_{I_p} に対して，対角平均化を適用して，対応する成分を再構築する．$L \times K$ 行列 Y の要素を y_{ij} とし，$L^* = \min(L, K)$, $K^* = \max(L, K)$, $N = L + K - 1$ とする．また $y^*_{ij} = y_{ij}$ ($L < K$ の場合) $y^*_{ij} = y_{ji}$ ($L > K$ の場合) とする．対角平均化では，以下のようにして行列 Y から長さ N の時系列 $\tilde{y}_k$ を生成する．

$$\tilde{y}_k = \begin{cases} \frac{1}{k} \sum_{m=1}^{k} y^*_{m,k-m+1} & (1 \leq k < L^*) \\ \frac{1}{L^*} \sum_{m=1}^{L^*} y^*_{m,k-m+1} & (L^* \leq k \leq K^*) \\ \frac{1}{N-k+1} \sum_{m=k-K^*+1}^{N-K^*+1} y^*_{m,k-m+1} & (K^* < k \leq N) \end{cases} \tag{3.9}$$

この対角平均化により、各部分行列 X_{I_p} から長さ N の時系列を得る．これらを足し合わせることで，元の時系列の近似再構成が可能となる．

3.3.1 音響ゼロ電子透かしへの応用

音声信号に対してSSAを適用し，各成分に分解した例を図3.1に示す．図最上部にSSAの対象となる音声信号を示し，下部に$L = 5$としてSSAを用いて各成分に分解した結果を，特異値の高い順に図示している．特異値が最大となる成分（a）は信号全体の概形（トレンド成分）を表現している．成分（b）〜（d）は，高調波などの周期振動成分を持つ．そして，特異値が最小となる成分（e）は残差雑音や環境雑音といったノイズ成分を持つ．特異値の大きさは元信号を構成するうえでの寄与度を表しており，寄与度の高い成分を用いることで，元信号に近い品質で再合成が可能となる．従って，SSAに基づいて分解された上位成分には信号の本質的な特徴が集約されると考える．これら上位成分に透かし情報を埋め込むことで，リサンプリング，量子化，及び圧縮符号化に対して頑健性を獲得できると期待される．また，埋め込む際に，それぞれの成分に対して埋め込み処理を行うことで，音情報処理によって埋め込み情報が崩れた成分があったとしても，各成分を重畳することで誤りを補間でき，頑健となることが期待できる．

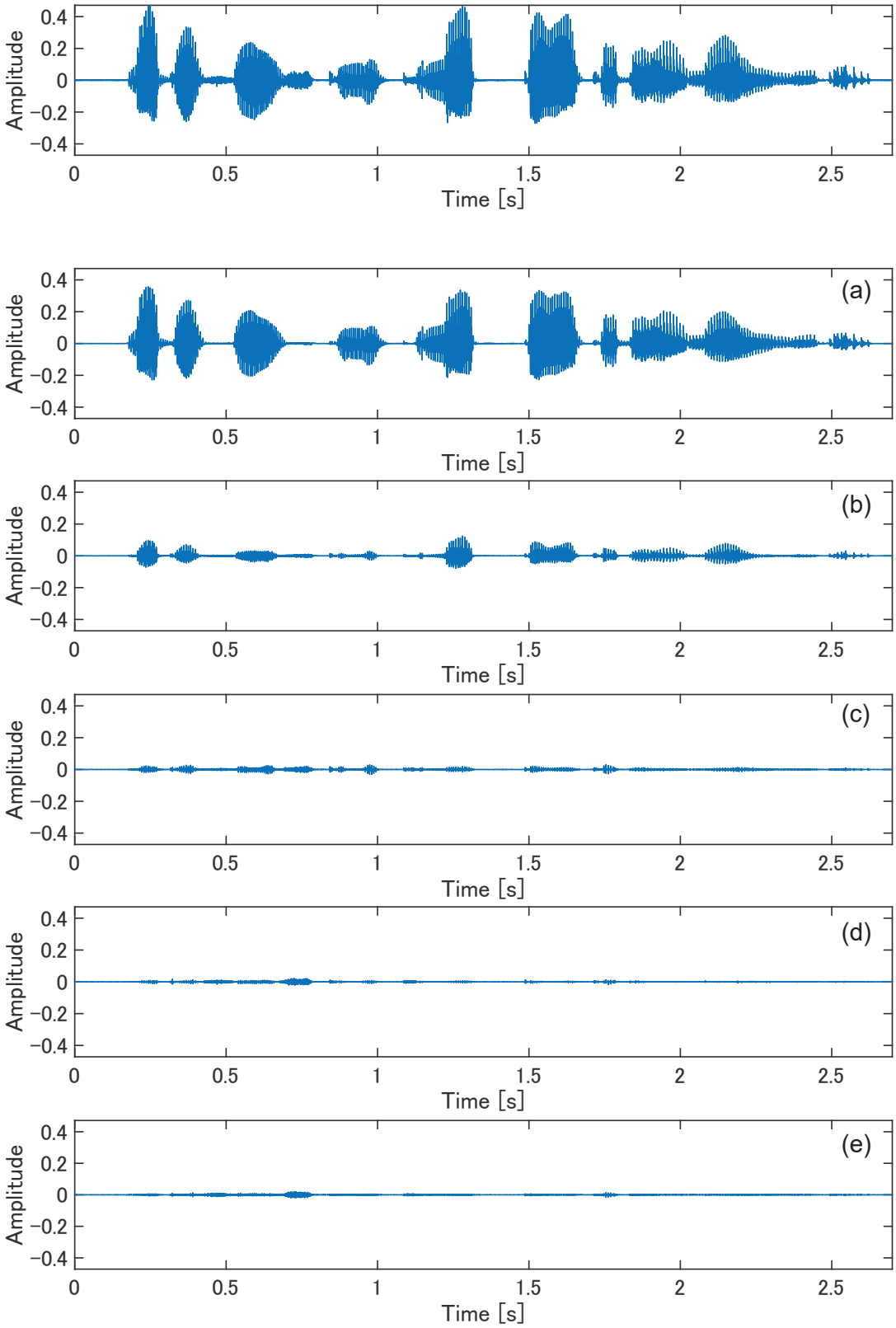

図 3.1: SSA を用いた信号分解

第4章　特異スペクトル分析を用いた音響ゼロ電子透かし

4.1　提案法の全体構成

音声信号のスパース性を考慮した上で，スパースコーディング法であるSSAを用いた音響ゼロ電子透かし法を提案する．図4.1に提案する方法のフレームワークを示す．図4.1(a)の埋め込み処理では，先ず音声信号に対してSSAを施し，各成分に分解する．次に，各成分の特異値の大きさを元信号を構成するうえでの寄与率として，寄与率の大きい上位成分を選択しバイナリパターンを生成する．バイナリパターン生成に際する二値化処理には，1-dimension local binary pattern（1DLBP）を用いる．そして，埋め込み処理として，排他的論理和（XOR）を用いてバイナリパターンと透かし情報（0/1の二値画像）から検出鍵を生成する．

図4.1(b)の検出処理では，照合の対象となる信号について，同様にSSAを施し，バイナリパターンを生成する．その後，図4.1(a)で得られた検出鍵を用いて透かし情報を抽出する．

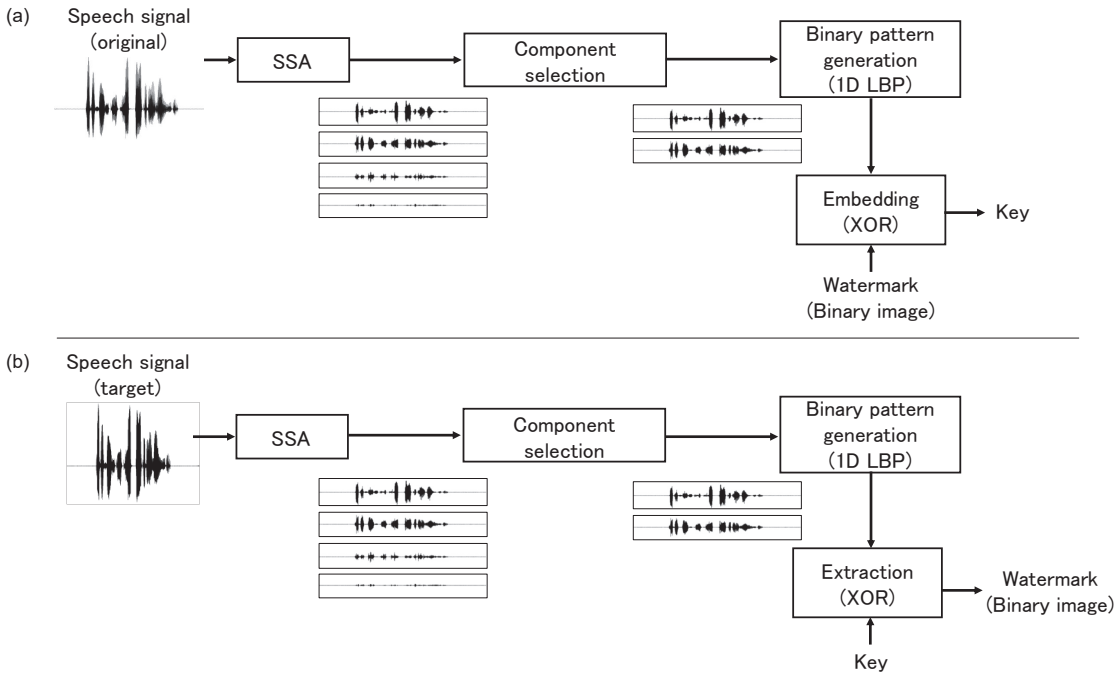

図 4.1: 提案法のフレームワーク：(a) 透かし情報の埋め込み処理，(b) 透かし情報の検出処理

4.2 検出鍵の生成法

図 4.1(a) の検出鍵の生成法では，以下の処理を行う．また，具体的な処理フローを図 4.2 に示す．

1. 音声信号を $x(t)$ とし，$x(t)$ に対し SSA を施し各成分 $Q(L)$ を得る．この時，SSA によって分解する成分数 L は任意に設定可能なパラメータとなる．

2. 得られた $Q(L)$ のうち，特徴量として重要である成分を決め選択する．特異値の大きさをもとに，各成分の元信号に対する寄与率を算出し，寄与率が 90% となる成分までを選択する．

3. 選択した各成分に対して 1DLBP を用いてバイナリパターンを生成する．この時，1DLBP で設定する bit 幅によって，埋め込むことができる情報量が変化する．bit 幅を m とすると，バイナリパターン $B \in \{0,1\}^{2^{2m}}$ が生成される．

4. 各成分 $Q(l)$ の $B(l) \in \{0,1\}^{2^{2m}}$ に対して多数決処理を施し，透かし情報を埋め込むためのベクトルを生成する．例として，$B(l) \in \{0,1\}^{2^{2m}}$ の対応する次元のバイナリパターン列が $\{1,1,0,1\}$ であった場合，値として 1 を得る．

5. 透かし情報であるバイナリ画像を平坦化したベクトル $W \in \{0,1\}^{2^{2m}}$ と $B \in \{0,1\}^{2^{2m}}$ の XOR を算出することによって，検出鍵 $D \in \{0,1\}^{2^{2m}}$ を得る．

4.3 1-dimension local binary patternを用いたバイナリパターン生成法

Local binary pattern（LBP）[27] は本来二次元画像のテクスチャ解析として提案された方法である．画像の各ピクセルとその周辺のピクセルの輝度値を比較することで，局所的なテクスチャの特徴を捉えることが可能となる．時系列信号を対象とする場合には一次元化した 1-dimension local binary pattern（1DLBP）[28, 29] が利用される．1DLBP はノイズに対する頑健性を持ち，時間的な局所構造を捉えた有効な特徴量として扱われる．従って，バイナリパターン生成に利用することで，頑健性を獲得できると考えられる．

長さ N の信号

$$x = \{x_0, x_1, \ldots, x_{N-1}\}, \qquad x_i \in \mathbb{R} \tag{4.1}$$

を対象としたとき，解析中心点を i_c とおく．左右 R 点ずつ（計 $P = 2R$ 点）の近傍集合

$$\mathcal{N}(i_c) = \{\, i_c - R,\ i_c - R + 1,\ \ldots,\ i_c - 1,\ i_c + 1,\ \ldots,\ i_c + R\} \tag{4.2}$$

を用いる．そして，$p = 0, 1, \ldots, P-1$ に対し近傍インデックス $i_p = i_c - R + p + \mathbf{1}_{p \geq R}$ を導入し，以下の条件で 1 ビット符号 b_p を得る．

$$b_p = \begin{cases} 1 & (x_{i_p} \geq x_{i_c}) \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

ビット列 $\{b_p\}$ を下位ビット優先で重み付けすると，

$$\mathrm{LBP}_{P,R}(i_c) = \sum_{p=0}^{P-1} b_p \, 2^p \tag{4.4}$$

が 1DLBP 値となる．取り得る値は $0 \leq \mathrm{LBP}_{P,R} \leq 2^P - 1$ であり，P ビットの局所形状を 1 つの整数に圧縮している．

この 1DLBP 値に対して

$$H(k) = \sum_{i_c \in \mathcal{S}} \delta\big(\mathrm{LBP}_{P,R}(i_c),\, k\big), \qquad k = 0, \ldots, 2^P - 1 \tag{4.5}$$

の度数分布を求める．$H(k)$ の各ビンの出現回数の平均

$$\bar{H} = \frac{1}{K} \sum_{k=0}^{K-1} H(k) \tag{4.6}$$

を閾値として，最終的に二値のベクトル $B \in \{0, 1\}^{2^p}$ を得る．

提案法における 1DLBP の処理図を図 4.3 に示す．時間信号に対して，区間 R 毎に LBP の算出を行う．また，解析中心点の左右近傍の平均値を閾値とし，ビット符号 b_p を得る．

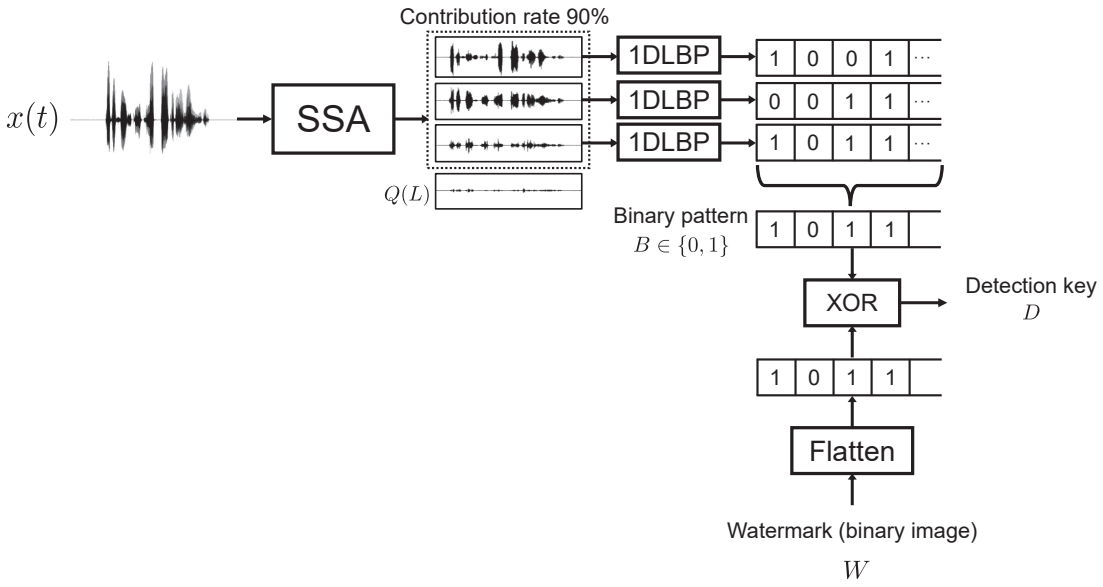

図 4.2: 検出鍵生成の処理フロー図

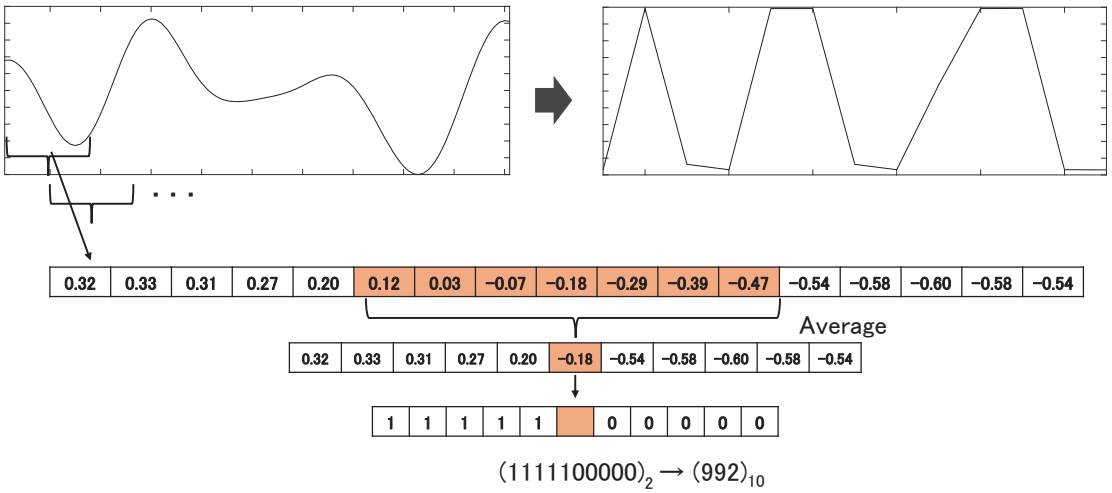

図 4.3: 提案法における 1DLBP の処理

4.4 透かし情報の検出法

図 4.1(b) の透かし情報の検出法では，以下の処理を行う．

1. 検出する対象の信号を $x'(t)$ とし，$x'(t)$ に対し SSA を施し各成分 $Q'(L)$ を得る．
2. 得られた $Q'(L)$ のうち，特徴量として重要である成分を決め選択する．
3. 選択した各成分に対して 1DLBP を用いてバイナリパターンを生成し，バイナリパターン $B' \in \{0,1\}^{2^{2m}}$ が得られる．
4. 各成分 $Q'(l)$ の $B'(l) \in \{0,1\}^{2^{2m}}$ に対して多数決処理を施し，透かし情報を埋め込むためのベクトルを生成する．
5. 検出鍵 $D \in \{0,1\}^{2^{2m}}$ と $B \in \{0,1\}^{2^{2m}}$ の XOR を算出することによって，透かし情報 $W' \in \{0,1\}^{2^{2m}}$ を検出する．

第5章　特異スペクトル分析を用いた音響ゼロ電子透かしの実験的評価

5.1　目的

提案した方法について，非攻撃的処理に対する頑健性と攻撃的処理に対する検出性能の評価を行う．非攻撃的処理とは，リサンプリングや再量子化，圧縮符号化といった音声信号の形式は変化するが，意味を持つ内容は損なわれないような一般的な音情報処理を指す．一方で，攻撃的処理は，音声改ざんやある区間を切取り，音声信号の内容を変化させるような処理を指す．また，ピッチを変化させることによって，話者情報である性別を変化させるような処理についても攻撃的な処理とみなす．

非攻撃的処理では，透かし情報を検出する際，なるべく誤差が少ないことが望ましい．逆に攻撃的処理では，検出された透かし情報の誤差が大きくなることが望ましい．また，透かし情報の情報量が大きいほど実用性が高いため，情報量を大きくした際に頑健性が低下しないことも要求される．

本論では，2.3 節で述べた従来法を比較対象とし，非攻撃的処理に対しての頑健性及び攻撃的処理に対しての検出精度についてそれぞれ評価する．

5.2　方法

提案法の性能を検証するため，図 4.1 に示すフレームワークを構築した．まず，図 4.1(a) の埋込処理では，入力音声信号と透かし情報 W（画像）から検出鍵を生成する．次に，図 4.1(b) の検出処理では，元信号に非攻撃的処理または攻撃的処理を施した信号と，前段で得られた検出鍵を用いて透かし情報 W' を抽出する．最後に W と W' の誤差を評価指標として提案法を定量的に評価する．頑健性が求められる音情報処理では誤差が小さいほど性能が高く，一方で脆弱性が求められる攻撃的処理に対しては誤差が大きいことが望ましい．

5.2.1 実験条件

実験に用いるデータセットには新聞読み上げコーパス [30] を用いる．このデータは男性及び女性の文章読み上げ音声が各 50 名ずつ計 100 名分収録されている．サンプリング周波数は 16 kHz，量子化ビット数 16 bit のモノラル信号を使用した．このデータセットに対して，オープンソースソフトウェアである FFmpeg [31] を用いて様々な音情報処理を施した．実験の評価に使用した音情報処理の一覧を表 5.1 に示す．非攻撃的処理では，ローパスフィルタ，リサンプリング処理，再量子化処理，及び音声符号化処理の条件で評価する．ローパスフィルタのカットオフ周波数は 8 kHz である．リサンプリング処理では，8 kHz にダウンサンプルした後，16 kHz に戻す処理と，44.1 kHz にアップサンプルした後，16 kHz に戻す処理の二種類を用意する．再量子化についても同様に，量子化ビット数を 8 bit に落としてから 16 bit に戻す処理，24 bit に上げてから 16 bit に戻す処理の二種類作成する．圧縮符号化処理については，MP3 符号化 [32, 33]，AAC 符号化 [34, 35]，G.711[36]，G.729[37, 38] の音声コーデックを用いた．

攻撃的処理について，ゼロ補間処理では，発話内容の一部を無音にする処理である．ピッチシフトでは，音声の基本周波数 F_0 を変化させ，男性の声を女性の声になるよう，又はその逆となるような操作を行った．変換処理には，WORLD[39] を利用した．サンプル置換処理では，合成音声を用いて発話内容が変化するように一部を置き換える操作を行う．音声合成には VOICEVOX（© Hiroshiba Kazuyuki）を利用した．クロッピングでは，音声の一部を切り取り，切り取った箇所の前後を繋げることで，発話内容が変化するような操作を行った．

音声信号に SSA を施す際，分解する成分数 L は 30 とした．また，埋め込みを行う透かし画像のサイズは 16×16，32×32，64×64 の 3 種類を用いた．

表 5.1: 評価に使用する非攻撃的処理及び攻撃的処理の種類

Non-attack processing		
NOP	No processing	
LPF	Low-pass filter	Cut-off frequency: 8 kHz
$\mathbf{RSM_d}$	Resampling	16 kHz $\rightarrow$ 8 kHz $\rightarrow$ 16 kHz
$\mathbf{RSM_u}$	Resampling	16 kHz $\rightarrow$ 44.1 kHz $\rightarrow$ 16 kHz
$\mathbf{RQZ_d}$	Requantization	16 bit $\rightarrow$ 8 bit $\rightarrow$ 16 bit
$\mathbf{RQZ_u}$	Requantization	16 bit $\rightarrow$ 24 bit $\rightarrow$ 16 bit
$\mathbf{MP3_n}$	MP3 coding	Bit-rate n kbps, $n \in \{32, 64, 128, 256\}$
$\mathbf{AAC_n}$	AAC coding	Bit-rate n kbps, $n \in \{32, 64, 128, 256\}$
G.711	G.711 coding	$F_s = 8$ kHz, log-PCM, μ-law
G.729	G.729 coding	$F_s = 8$ kHz, CELP
Attack processing		
ZI	Zero interpolation	Zero padding / interpolation
$\mathbf{PSH}_{\pm 20\%}$	Pitch shift	Fundamental frequency F_0 shifted by $\pm 20\%$
RPM	Replacement	Sample replacement with synthetic speech
CRP	Cropping	Remove a specified segment

5.2.2 評価指標

評価指標に，Bit error rate（BER）を用いて，従来法と提案法の性能比較を行う．サイズ $m \times m$ の透かし情報 W と W' の誤差を測る際 BER（%）は以下のように算出される．

$$\mathrm{BER} = \frac{\sum_{i,j} \left(w_{i,j} \oplus w'_{i,j} \right)}{m \times m} \times 100 \tag{5.1}$$

$\oplus$ は排他的論理和を表し，以下のように定義される．

$$a \oplus b = (a + b) \bmod 2 \tag{5.2}$$

画像処理及び音響信号処理分野において，BER が 1%以下であれば，透かし情報の完全複合か可能であると報告されている [40]．従って，BER が 1%以下となれば頑健性を有すると評価する．

また，攻撃処理に対する検出精度の評価として F 値を用いる．F 値は，二値分類の性能を示す指標であり，真陽性を TP，偽陽性を FP，偽陰性を FN とするとき，適合率（Precision）P，再現率（Recall）R，および F 値は次式で定義される．

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5.3}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5.4}$$

$$\mathrm{F\ score} = \frac{2PR}{P + R} \tag{5.5}$$

F 値は 0 から 1 の範囲の値を持ち，1 に近いほど検出精度が優れていることを示す．本実験の場合，攻撃の見逃し（FN）をが少なく，誤検知（FP）が少ない場合，F 値が向上する．

5.3 結果

5.3.1 頑健性に関する評価

非攻撃的処理に対する頑健性評価の結果を表 5.2 及び表 5.3 に示す．100 名分のデータの平均 BER%を示している．3 種類のサイズの透かし画像を用いた結果を各行に示している．元信号と同一の信号を用いた場合，BER は 0%となっていることから，提案法について正常に実装できていることが確認できる．G.711 符号化及び G.729 符号化以外の非攻撃的処理について，提案法は頑健性の指標となる BER が 1%以下を達成しており，音響ゼロ電子透かしに有効であることが示唆される．また，透かし画像のサイズが増加する場合でも，BER は 1%を超えないた

め，情報量が増えた場合についても透かし情報を検出できていることが示された．従来法と比較した際，ローパルフィルタ処理，リサンプリング処理，及び再量子化処理では提案法の性能が従来法を上回らない結果となった．MP3 符号化に関して，低ビットレート（32 kbps，64 kbps）の場合，提案法が従来法よりも高い頑健性を持つことが分かる．一方で，高ビットレート（128 kbps，256 kbps）の場合は，提案法の方が頑健性を有する．AAC 符号化については全種類のビットレートにおいて提案法の方が高い頑健性を持つことが確認された．G.711 符号化については従来法と比較して提案法の BER が大幅に高く，BER が 1%を超える結果も確認された．G.729 符号化では，従来法に及ばない性能ではあるが，その差は僅かである．

表 5.2: 非攻撃的処理を施した音声信号から得られた透かし画像の平均 BER（%）及び従来法との比較結果（ローパスフィルタ，リサンプリング，及び再量子化）

	NOP		LPF		RSM_d		RSM_u		RQZ_d		RQZ_u	
	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA
16×16	0.00	0.00	0.00	0.03	0.02	0.55	0.01	0.01	0.17	0.31	0.00	0.00
32×32	0.00	0.00	0.00	0.03	0.03	0.72	0.00	0.03	0.24	0.35	0.00	0.00
64×64	0.00	0.00	0.00	0.05	0.03	0.81	0.00	0.08	0.29	0.40	0.00	0.00

表 5.3: 音声符号化処理を施した音声信号から得られた透かし画像の平均 BER（%）及び従来法との比較結果（MP3 符号化, AAC 符号化，G.711，G.729）

	MP3 (32 kbps)		MP3 (64 kbps)		MP3 (128 kbps)		MP3 (256 kbps)		AAC (32 kbps)		AAC (64 kbps)		AAC (128 kbps)		AAC (256 kbps)		G. 711		G. 729	
	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA
16×16	0.39	0.18	0.13	0.08	0.08	0.02	0.04	0.02	0.83	0.59	0.69	0.60	0.67	0.51	0.67	0.53	0.34	0.92	0.88	0.90
32×32	0.52	0.20	0.16	0.08	0.08	0.04	0.05	0.04	1.29	0.69	1.10	0.67	1.17	0.68	1.13	0.68	0.44	1.41	1.36	1.38
64×64	0.67	0.34	0.20	0.15	0.15	0.03	0.09	0.02	3.38	0.76	3.33	0.72	3.32	0.72	3.32	0.70	0.48	1.62	1.71	1.72

5.3.2　考察

前節の結果から頑健性に関する考察を行う．まず，ローパスフィルタ（LPF），リサンプリング（RSM），再量子化（RQZ）といった処理に対しては，従来法が全体的に低 BER を維持している一方，提案法は特にリサンプリングと量子化ノイズの影響を強く受け，最大で 0.81%（RSM_d）まで悪化した．提案法は SSA の構造上，音声信号の周期成分を主成分として鍵生成を行うため，時間軸の再サンプリングや量子化に伴う微細な周期構造の崩壊に大きく影響される．対照的に，従来法は時間位置情報（Spikegram）に依拠しており，振幅変動や高調波の変形に対して相対的に頑健である．

一方，MP3 符号化や AAC 符号化では，提案法が従来法に対して,BER 低減を達成している場合もあり，これは，MP3 符号化及び AAC 符号化が中低周波数帯の周期性を比較的保持するため，SSA が抽出する周期パターンが符号化後も整合しやすいことによると考えられる．しかし，狭帯域予測符号化型の G.711 及び G.729 では，8 kHz ローパスフィルタと線形予測残差により周期構造が大きく損なわれ，提案法の BER が 1.7 から 1.9% 程度まで上昇した．従来法も性能低下は免れないものの，相対的優位を保っている．

透かし画像のサイズの影響については，32×32 から 64×64 への拡大に伴い，従来法及び提案法の両方とも BER が上昇したが，BER は 1%以下を達成している．従って，情報量が増えたとしても透かし情報を正しく検出できるといえる．

また，周波数領域 SSA への拡張による時間軸操作への耐性強化や狭帯域保持成分への重み付け設計といった改良を行うことで，様々な音情報処理に対応可能となることが期待できる．

5.3.3　改ざん攻撃への脆弱性に関する評価

攻撃処理を施した後の音声信号から得られた透かし画像の平均 BER の結果を表 5.4 に示す．従来法では，全攻撃処理で BER が 1%よりも大幅に超えており，改ざんを高確率で検知することが可能である．一方，提案法では，BER が 0.8%から 1.7%にとどまり，BER の増加による改ざん検出を行う際に十分な値に達していないことが示される．

100 データの音声信号を改ざんあり，改ざん無しに分類した際の F 値の結果を表 5.5 に示す．提案法は，従来法と比較し検出精度（F 値）が低く，誤検知及び改ざんデータの見逃しが多いことが示される．従って，攻撃的処理に対してもある程度頑健となってしまい，透かし画像の破壊を通じて改ざん検出を行う場合実用性に欠けることが明らかとなった．

表 5.4: 攻撃的処理を施した音声信号から得られた透かし画像の平均 BER（%）及び従来法との比較結果

	CRP		PSH(H)		PSH(L)		RPM		Zi	
	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA
16×16	4.57	1.55	4.22	1.24	4.39	1.23	4.13	1.74	3.81	1.50
32×32	5.25	1.06	5.57	1.10	5.16	1.02	4.15	0.93	3.73	1.52
64×64	5.22	0.98	5.45	1.11	5.44	0.83	4.08	0.79	3.61	1.03

表 5.5: 攻撃的処理を施した音声の検出精度（F 値）

	CRP		PSH(H)		PSH(L)		RPM		Zi	
	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA	Spikegram	SSA
16×16	0.41	0.33	0.90	0.60	0.91	0.61	0.76	0.66	0.71	0.25
32×32	0.34	0.28	0.91	0.60	0.91	0.60	0.77	0.60	0.78	0.22
64×64	0.32	0.29	0.92	0.57	0.91	0.59	0.70	0.61	0.75	0.21

5.3.4 考察

攻撃的処理に対する実験結果について考察を行う．提案法では，SSA によって得られる成分に透かし情報を重畳している．また，1DLBP を用いて二値化を行い得られるバイナリパターンは時間情報を持たない．従って，音声信号の一部を改ざんするようなゼロ補間攻撃及び切取り攻撃では，SSA によって得られる分解成分の大域的な構造が変化せず，結果として透かし情報が崩れない結果となったと考えられる．また，各成分にそれぞれ 1DLBP を施し得られるバイナリパターンに対して多数決処理を行い最終的に透かし情報を埋め込むバイナリパターンを作成している．このような二値化処理が過度な頑健性を獲得しており，攻撃処理に対しても BER が大幅に増加しない結果になっていると考えられる．従って，攻撃的処理によって破壊されるバイナリパターンとなるような特徴量が反映されるような二値化処理法を改めて考える必要がある．

5.4 全体考察

従来法（Spikegram）と提案法（SSA+1DLBP）比較した．非攻撃的処理に対して，提案法はほぼすべての条件で BER $\leq 1\%$ を維持し，特に MP3 及び AAC の低ビットレート符号化では従来法より BER を低減した．これは，SSA が抽出する周期主成分が音声コーデック後も保存されやすく，1DLBP が振幅スケーリングや緩慢な位相ゆらぎに不変であるためであるといえる．一方，G.711，G.729，ローパスフィルタ，リサンプリング，再量子化に対しては従来法が優位であり，SSA の周期依存性が時間軸操作と量子化雑音に弱いことが考えられる．

攻撃的処理ではクロッピング，ゼロ補間，ピッチシフト，サンプル置換など局所改ざんを施しても，提案法の BER は 0.8%から 1.7%にとどまり，改ざん検出には十分な上昇が得られなかった．SSA が大域周期構造を鍵とするため局所破壊後も特徴量が保持されやすく，さらに 1DLBP が隣接サンプルの大小関係のみを符号化するため，短時間の置換や位相ずれではパターンが変化しにくいことが主因と考えられる．結果として F 値も 0.21 から 0.66 と低く，従来法に大きく劣った．また，透かし画像サイズの拡大に伴い BER が低下する傾向も確認され，情報量増加が局所誤りを平均化して攻撃感度をさらに弱めることが示された．以上より，提案法では著作権保護などの透かしの可逆復元を優先する用途に適している．しかし，透かし破壊による改ざん検出を目的では感度が不足しており，時間位置情報に依存する Spikegram に劣る．

第6章　結論

6.1　本研究で明らかにしたこと

本研究では，音声信号のスパース性を考慮した特徴量を用いる音響ゼロ電子透かしとして，スパースコーディング法であるSSAを用いた方法を提案した．評価実験の結果，以下の事項が明らかとなった．

- 提案法は，MP3，AACなどの符号化やローパスフィルタ，リサンプリング，及び再量子化に対してBER $\leq 1\%$ を維持し，透かし情報を検出可能であったが，G.711符号化及びG.729符号化ではBER $> 1\%$ となり，透かし情報の検出に誤差が生じた．
- 前段のBER $\leq 1\%$ となった音情報処理に対して頑健性を有していることから，1DLBPを利用することで，音響信号から得られた特徴量を反映させつつ二値化処理（バイナリパターン生成）を行えることが示唆された．
- 透かし情報の情報量を増やすと，BERが増加するが，依然BERは1%以下であり，情報量増加に対しても頑健であった．
- MP3符号化やAAC符号化では，一部のビットレート条件において従来法を上回る性能となった．その他の処理については従来法に及ばなかったが，頑健性と判断する条件BER $\leq 1\%$ を達成している．
- 攻撃的処理では，提案法のBERが1%前後にとどまり，十分な改ざん検出精度（F値）が得られなかった．

以上の結果から，提案法は，SSAで抽出した周期成分に透かし情報を付与し，1DLBPを用いてバイナリパターンを生成することで，MP3・AAC符号化やローパスフィルタなど多くの非攻撃的処理に対しBER $\leq 1\%$ を維持し得ることが確認された．したがって，透かし情報を真正性担保に利用する音声コンテンツ配信や著作権保護に有効であると評価できる．一方，クロッピングやゼロ補間など局所改ざん攻撃ではBERの上昇が限定的で，改ざん検出精度（F値）が十分に向上しなかった．結論として，従来法よりも頑健性を向上させることは達成できなかったが，BER$\leq 1\%$ の基準を満たすため非攻撃的処理に対する頑健性は有していることが分かった．しかし，改ざん検知能力の向上は未達成であった．原因として，

SSA が信号の大域周期構造を保持しやすいことと，1DLBP が局所的不変性を持つことが考えられる．従って，改ざん攻撃に対応するための改良を検討する必要がある．

6.2 残された課題

本研究の結果から，提案法には解決すべき課題が残されている．まず，クロッピングやゼロ補間に代表される局所的な改ざん攻撃に対しては，透かし復元が過度に頑健であるため BER が十分に増加せず，改ざん検出感度（F 値）が低いという問題がある．次に，G.711 や G.729 のような狭帯域予測コーデックや粗い再量子化処理を通すと，音声の周期構造が崩れ，提案法の BER が 1%を超える弱点が顕在化する．最後に，SSA が大域的周期成分に，1DLBP が局所的不変性に依存するという特性が改ざん検出を困難にしている．これらを克服するためには，得られた特徴量の重要な成分を反映させられる二値化処理法を依然検討する必要がある．

付 録A　各種スパースコーディング法と再合成品質に関する評価

A.1　経験的モード分解

経験的モード分解（emperical mode decomposition: EMD）[41] は，非線形かつ非定常な時系列 $x(t)$ を固有モード関数（intrinsic mode function: IMF）$c_i(t)$ と残差 $r_N(t)$ に逐次的に分解し，

$$x(t) = \sum_{i=1}^{N} c_i(t) + r_N(t) \tag{A.1}$$

という自己適応的な表現を得る方法である．IMF を得るためには二つの条件が必要となる．第一に任意の時刻における零交差点と極値の個数の差が高々一つであること，第二に上包絡線 $u(t)$ と下包絡線 $l(t)$ の平均

$$m(t) = \frac{u(t) + l(t)}{2} \tag{A.2}$$

が時間軸上で常にゼロとなることである．これらの条件を満たすまでシフティングと呼ばれる操作を反復し，反復回数 k に対する中間信号

$$h^{(k)}(t) = h^{(k-1)}(t) - m^{(k-1)}(t), \qquad h^{(0)}(t) = x(t)$$

を更新する．収束判定には

$$SD = \sum_{t} \frac{|h^{(k-1)}(t) - h^{(k)}(t)|^2}{h^{(k-1)}(t)^2} \tag{A.3}$$

が所定の閾値 ε（10^{-3} 程度）を下回るかどうかを用いるのが一般的である．

EMD で得られる IMF はヒルベルト変換

$$\hat{c}_i(t) = \mathcal{H}[c_i](t) = \frac{1}{\pi}\,\mathrm{p.v.}\int_{-\infty}^{\infty} \frac{c_i(\tau)}{t-\tau}\,d\tau \tag{A.4}$$

によって解析信号 $z_i(t) = c_i(t) + j\,\hat{c}_i(t) = a_i(t)e^{j\theta_i(t)}$ を構成でき，瞬時周波数

$$f_i(t) = \frac{1}{2\pi}\frac{d\theta_i(t)}{dt} \tag{A.5}$$

によってヒルベルト–フアン変換も導出できる．フーリエ基底のように事前に形状を仮定せず，信号の局所統計量のみに基づいて基底関数を生成する点で，EMD はデータ駆動型かつ非線形適応的な分解法である．

スパースコーディングの観点から見ると，観測信号 $\mathbf{y}$ を基底行列 $\mathbf{\Phi}$ と係数ベクトル $\mathbf{x}$ による線形モデル

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x}, \qquad \|\mathbf{x}\|_0 \ll \dim(\mathbf{x}) \tag{A.6}$$

として仮定した際，ℓ_0 ノルム $\|\mathbf{x}\|_0$ を最小化することがスパースコーディングの目的となる．EMD が生成する IMF 群 $\{c_i(t)\}$ は線形独立ではないものの，それぞれが狭帯域で単一振動様の局所特徴を担うため，$\mathbf{\Phi}$ を $\begin{bmatrix} c_1(t) & \dots & c_N(t) & r_N(t) \end{bmatrix}$ とみなすと，式 (A.1) は係数ベクトル $\mathbf{x} = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^{\mathsf{T}}$ によって表される．このとき IMF の数 N が少なければ $\|\mathbf{x}\|_0 = N$ が小さくなり，EMD が暗黙的にスパース近似を実現していることが理解できる．さらに各 IMF において瞬時振幅 $a_i(t)$ が時間–周波数平面上でまばらに分布する場合には，式 (A.5) に由来する分布に対しても高次のスパース性が顕在化する．

A.2　主成分分析

主成分分析（principal component analysis: PCA）[42] は，多変量観測行列 $\mathbf{X} \in \mathbb{R}^{d\times M}$（$d$ 次元の観測が M 個）から直交基底を抽出し，分散が最大となる座標系へ変換することで次元圧縮と特徴抽出を同時に行う方法である．まず各観測 $\mathbf{x}_m$ を平均ベクトル $\bar{\mathbf{x}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_m$ で中心化し，中心化行列を $\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_M - \bar{\mathbf{x}} \end{bmatrix}$ とおく．分散共分散行列は

$$\mathbf{\Sigma} = \frac{1}{M}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}} \tag{A.7}$$

で定義され，固有値分解

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0 \tag{A.8}$$

によって得られる固有ベクトル $\mathbf{u}_i$ が主成分軸，固有値 λ_i が対応する分散量となる．固有ベクトルを列に並べた直交行列を $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1, \dots, \mathbf{u}_d \end{bmatrix}$ とすると，観測の主成分スコアは

$$\mathbf{z}_m = \mathbf{U}^{\mathsf{T}}\left(\mathbf{x}_m - \bar{\mathbf{x}}\right) \tag{A.9}$$

で与えられる．$K < d$ 個の主要成分を選択した近似では $\mathbf{U}_K = \left[\mathbf{u}_1, \ldots, \mathbf{u}_K\right]$ とし，

$$\mathbf{x}_m \approx \bar{\mathbf{x}} + \mathbf{U}_K \mathbf{z}_m^{(K)}, \qquad \mathbf{z}_m^{(K)} = \mathbf{U}_K^{\mathsf{T}} \left(\mathbf{x}_m - \bar{\mathbf{x}}\right) \tag{A.10}$$

となる．式 (A.10) は再構成誤差 $\sum_{m=1}^{M} \|\mathbf{x}_m - \bar{\mathbf{x}} - \mathbf{U}_K \mathbf{z}_m^{(K)}\|_2^2$ を最小化する直交射影であり，$\lambda_1 + \cdots + \lambda_K$ が保存される分散総量を示す．

数値実装では特異値分解 $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}_s\mathbf{V}^{\mathsf{T}}$ を用いて主成分軸 $\mathbf{U}$ と固有値 $\lambda_i = \sigma_i^2/M$ を同時に得る方法が一般的である．ここで $\mathbf{\Sigma}_s = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ は特異値行列であり，σ_i が信号エネルギーの大きさを規定する．

スパースコーディングでは，$\mathbf{U}_K$ を辞書行列，$\mathbf{z}_m^{(K)}$ を係数ベクトルとみなせば

$$\mathbf{y}_m = \mathbf{U}_K \mathbf{x}_m^{\mathrm{sp}}, \qquad \mathbf{x}_m^{\mathrm{sp}} = \mathbf{z}_m^{(K)} \tag{A.11}$$

という線形モデルを形成する．係数次元が K に制限されるため $\|\mathbf{x}_m^{\mathrm{sp}}\|_0 = K$ であり，$K \ll d$ の場合には明示的な正則化項を導入せずともスパース近似が達成される．ただし PCA は基底ベクトルがデータ分散を最大化するように連続値で学習されるため，EMD に比べると係数ベクトルは一般に疎ではないため，非負制約や局所性制約を課したスパース PCA の拡張が併用されることも多い．

A.3　非負値行列因子分解

非負値行列因子分解（nonnegative Matrix Factorization: NMF）は，観測行列 $\mathbf{X} \in \mathbb{R}_{\geq 0}^{d \times M}$ を二つの非負行列 $\mathbf{W} \in \mathbb{R}_{\geq 0}^{d \times K}$ と $\mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times M}$ の積で近似する方法である．ここで $\mathbf{W}$ は基底行列，$\mathbf{H}$ は係数行列に相当し，

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} \quad \text{with} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0 \tag{A.12}$$

が基本的なモデルである．非負制約は要素の相加的な構成を保証するため，画像や音響スペクトルのように定義域が非負であるデータを部品ごとに分解する「足し算のみ」の表現が得られる．

最も広く用いられる目的関数は Frobenius ノルムの二乗

$$J_F = \frac{1}{2}\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \tag{A.13}$$

であり，これを最小化する問題は Lee と Seung が提案した乗法更新規則

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^{\mathsf{T}}\mathbf{X}}{\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{H}}, \tag{A.14}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\mathbf{X}\mathbf{H}^{\mathsf{T}}}{\mathbf{W}\mathbf{H}\mathbf{H}^{\mathsf{T}}} \tag{A.15}$$

によって非負性を保ったまま単調に収束する．ここで $\odot$ は要素積，分数は要素ごとの商を表す．ポアソン分布を仮定した場合には Kullback–Leibler（KL）ダイバージェンス

$$J_{\mathrm{KL}} = \sum_{i,m} \left[X_{im} \log \frac{X_{im}}{(WH)_{im}} - X_{im} + (WH)_{im} \right] \tag{A.16}$$

が対数尤度に対応し，これに対しても乗法更新規則が導出される．音響信号処理ではスケール不変性を備えた Itakura–Saito（IS）ダイバージェンスを用いる場合が多く，領域ごとに適切な統計モデルを選ぶことで応用範囲が拡張される．

NMF は辞書学習の一種とみなすことができ，$\mathbf{W}$ を辞書，$\mathbf{H}$ を係数ベクトル集合と解釈すると，式 (A.12) はスパースコーディングの線形モデルに対応する．非負制約だけでは係数の疎性は保証されないが，L1 正則化

$$J_{\mathrm{sp}} = J_F + \beta \sum_{k,m} H_{km} \tag{A.17}$$

を付加すると $\mathbf{H}$ が自動的にまばらになり，選択的に少数の基底のみが活性化される．β を調整することでスパース度合いを制御できる．

A.4 K-特異値分解

K-特異値分解（k-singular value decomposition: K-SVD）は，観測行列 $\mathbf{Y} \in \mathbb{R}^{d \times M}$ を過完備辞書 $\mathbf{D} \in \mathbb{R}^{d \times K}$ とスパース係数行列 $\mathbf{X} \in \mathbb{R}^{K \times M}$ に分解する辞書学習アルゴリズムであり，

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \quad \text{s.t.} \quad \forall m,\ \|\mathbf{x}_m\|_0 \le s, \quad \forall k,\ \|\mathbf{d}_k\|_2 = 1 \tag{A.18}$$

を交互最適化によって近似的に解く．ここで s は許容する非零係数の最大数，$\mathbf{d}_k$ は辞書の k 列目（アトム），$\mathbf{x}_m$ は信号 $\mathbf{y}_m$ に対応する係数ベクトルである．

最適化は二段階の反復で構成される．スパースコーディング段階では辞書 $\mathbf{D}$ を固定し，各 $\mathbf{x}_m$ を

$$\min_{\mathbf{x}_m} \|\mathbf{y}_m - \mathbf{D}\mathbf{x}_m\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}_m\|_0 \le s \tag{A.19}$$

で推定する．実際には直交マッチング追跡が広く用いられ，計算量を抑えつつ疎解を得る．

辞書更新段階では係数行列 $\mathbf{X}$ を固定し，アトム $\mathbf{d}_k$ と対応する係数行 $\mathbf{x}_k^{\mathsf{T}}$ を同時に更新する．非零要素を持つ信号集合を $\Omega_k = \{\, m \mid X_{km} \neq 0 \,\}$ と定め，残差行列を

$$\mathbf{E}_k = \mathbf{Y} - \sum_{j \neq k} \mathbf{d}_j \mathbf{x}_j^{\mathsf{T}} \tag{A.20}$$

とする．$\mathbf{E}_k$ の列を Ω_k で抽出した行列 $\mathbf{E}_k^{\Omega}$ に対し，

$$\mathbf{E}_k^{\Omega} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{A.21}$$

という特異値分解を行い，第一左特異ベクトル $\mathbf{u}_1$ を新しいアトム $\mathbf{d}_k$ とし，第一特異値 σ_1 と第一右特異ベクトル $\mathbf{v}_1$ の積を

$$\mathbf{x}_k^{\Omega} = \sigma_1 \mathbf{v}_1^{\mathsf{T}} \tag{A.22}$$

として係数を更新する．これにより目的関数 (A.18) の減少が保証され，非零パターンを保持したまま最小二乗近似が達成される．

アルゴリズムは残差エネルギーの十分な減少，または最大反復回数に達するまで続けられる．複数の初期辞書を試行することで局所最適解を回避し，収束後には $\|\mathbf{x}_m\|_0$ が均質に抑えられた高スパース表現が得られる．式 (A.18) はスパースコーディングの一般形であり，K-SVD は辞書学習と係数更新を同時に行うことで再構成誤差とスパース度合いを両立させる．

A.5　信号の再合成品質に関する評価

A.1 から A.4 で述べたスパースコーディング方法及び 3.3 節で述べた SSA について，信号の再合成品質の観点から，音響ゼロ電子透かしに有効な方法であるかを検討，評価する．スパースコーディングによって，信号を分解し得られた成分を用いて再び元信号を構成するように信号を再合成する．この時，元信号を完全に再現できていない場合，対象の信号に重要な特徴量を正しく抽出し成分分解できていないといえる．つまり，音響ゼロ電子透かしに利用するにあたり，音声信号の本質的な特徴量を捉えることができないため，符号化処理などが施された場合，頑健性を獲得できない．このような観点から各方法について信号の再合成品質の評価を行った．品質の評価指標には，時間領域での信号間の誤差を表す信号対誤差比（signal to error ratio: SER）[43]，スペクトル領域での歪みを表す対数スペクトル歪み（log spectral distance: LSD）[44]，及び音声信号の客観評価指標である perceptual evaluation of speech quality（PESQ）を用いた．SER は 20 dB 以上，LSD は 0.1 dB，PESQ は 4.5 以上であれば高品質な信号の再合成であるとされている．

100 名分の音声データ [30] を各方法に適用した際の再合成品質を表 A.1 に示す．各方法のうち，SSA，PCA，及び EMD では，3 つの評価指標の基準を満たしており，高品質な信号の再合成が可能であることが確認できる．

表 A.1: 各スパースコーディング法における音声信号の再合成品質

Method	Average SER [dB]	Average LSD [dB]	Average PESQ
SSA	303.41	0.00	4.63
EMD	316.24	0.00	4.64
PCA	22.33	0.00	4.63
NMF	10.21	0.23	2.29
K–SVD	29.74	3.84	3.36

謝辞

本研究を遂行するにあたって，主指導教員の鵜木祐史教授にはご指導とご助言を数多く賜りました．深く感謝いたします．また，研究室会議では，ご助言やご指摘を賜りました，木谷俊介講師，上江洲安史特任助教に心から感謝いたします．そして，鵜木研究室の皆様には，公私共に大変お世話になりましたこと，大変感謝いたします．最後に，学生生活を支えていただいた家族の皆様に感謝いたします．

参考文献

[1] Cvejic N. *et al.*, *Digital audio watermarking techniques and technologies: applications and benchmarks: applications and benchmarks*. IGI Global, 2007.

[2] Gruhl D., Lu A., and Bender W., "Echo hiding," in *Information Hiding: First International Workshop Cambridge, UK, May 30–June 1, 1996 Proceedings 1*. Springer, 1996, pp. 295–315.

[3] 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.

[4] Wang S., Yuan W., and Unoki M., "Multi-subspace echo hiding based on time-frequency similarities of audio signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2349–2363, 2020.

[5] 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. IEEE, 2004, pp. 533–537.

[6] Kirovski D. and Malvar H. S., "Spread-spectrum watermarking of audio signals," *IEEE transactions on signal processing*, vol. 51, no. 4, pp. 1020–1033, 2003.

[7] Xiang Y., Natgunanathan I., Rong Y., and Guo S., "Spread spectrum-based high embedding capacity watermarking method for audio signals," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 12, pp. 2228–2237, 2015.

[8] Nishimura A., "Audio watermarking based on subband amplitude modulation," *Acoustical science and technology*, vol. 31, no. 5, pp. 328–336, 2010.

[9] Unoki M. and Miyauchi R., "Method of digital-audio watermarking based on cochlear delay characteristics," in *Multimedia Information Hiding Technologies and Methodologies for Controlling Data*. IGI Global, 2013, pp. 42–70.

[10] Unoki M. and Miyauchi R., "Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay characteristics," *IEICE TRANSACTIONS on Information and Systems*, vol. 98, no. 1, pp. 38–48, 2015.

[11] Boney L., Tewfik A. H., and Hamdy K. N., "Digital watermarks for audio signals," in *Proceedings of the third IEEE international conference on multimedia computing and systems*. IEEE, 1996, pp. 473–480.

[12] 中山彰，陸金林，中村哲，鹿野清宏，"心理音響モデルに基づいたオーディオ信号の電子透かし"，電子情報通信学会論文誌 D，vol. 83，no. 11，pp. 2255–2263，2000.

[13] 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.

[14] Kanhe A. and Gnanasekaran A., "A qim-based energy modulation scheme for audio watermarking robust to synchronization attack," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 3415–3423, 2019.

[15] Budiman G., Novamizanti L., Alief R. N., Ansori M. R. R. *et al.*, "Qim-based audio watermarking using polar-based singular value in dct domain," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 2019, pp. 216–221.

[16] 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, 2016, pp. 352–357.

[17] Yu Y., Lei M., Liu X., Qu Z., and Wang C., "Novel zero-watermarking scheme based on dwt-dct," *China Communications*, vol. 13, no. 7, pp. 122–126, 2016.

[18] Lei M., Yang Y., Liu X., Cheng M., and Wang R., "Audio zero-watermark scheme based on discrete cosine transform-discrete wavelet transform-singular value decomposition," *China Communications*, vol. 13, no. 7, pp. 117–121, 2016.

[19] Guo X., Huang D., and Xu L., "A robust zero-watermarking scheme based on non-negative matrix factorization for audio protection," *Plos one*, vol. 17, no. 7, p. e0270579, 2022.

[20] Xu L., Huang D., Guo X., Rao W., Ji Y., Li R., and Lu X., "A novel robust zero-watermarking algorithm for audio based on sparse representation," *China Communications*, vol. 18, no. 8, pp. 237–248, 2021.

[21] 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"，信学技報，vol. 122，no. 368，pp. 20–25，2023.

[22] Makino S., Lee T.-W., and Sawada H., *Blind speech separation*. Springer, 2007, vol. 615.

[23] Yilmaz O. and Rickard S., "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[24] Broomhead D. S. and King G. P., "Extracting qualitative dynamics from experimental data," *Physica D: Nonlinear Phenomena*, vol. 20, no. 2-3, pp. 217–236, 1986.

[25] Golyandina N., "Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 12, no. 4, p. e1487, 2020.

[26] Hassani H., "Singular spectrum analysis: Methodology and comparison," *Journal of Data Science*, vol. 5, pp. 239–257, 2007.

[27] Ojala T., Pietikainen M., and Maenpaa T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[28] Kaya Y., Uyar M., Tekin R., and Yıldırım S., "1d-local binary pattern based feature extraction for classification of epileptic eeg signals," *Applied Mathematics and Computation*, vol. 243, pp. 209–219, 2014.

[29] McCool P., Chatlani N., Petropoulakis L., Soraghan J. J., Menon R., and Lakany H., "1-d local binary patterns for onset detection of myoelectric signals," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 499–503.

[30] 日本音響学会 音声データベース調査委員会, (新聞記事データ 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.

[31] Tomar S., "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[32] Brandenburg K. and Stoll G., "Iso/mpeg-1 audio: A generic standard for coding of high-quality digital audio," in *Proceedings of the 92nd Convention of the Audio Engineering Society*, Vienna, Austria, Mar. 1992.

[33] *Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio*, ISO/IEC Std. 11 172-3, 1993.

[34] Bosi M., Brandenburg K., Segers S., Herre J., Johnston J. *et al.*, "Iso/iec mpeg-2 advanced audio coding," *Journal of the Audio Engineering Society*, vol. 45, no. 10, pp. 789–814, Oct. 1997.

[35] *Information technology—Generic coding of moving pictures and associated audio – Part 7: Advanced Audio Coding (AAC)*, ISO/IEC Std. 13 818-7, 1997.

[36] *Pulse Code Modulation (PCM) of Voice Frequencies*, ITU-T Std. Recommendation G.711, 1972.

[37] Bessette B., Salami R., Lefebvre J., Lamblin C. *et al.*, "The itu-t 8 kb/s cs-acelp speech coding standard," *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 29–53, May 1998.

[38] *Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, ITU-T Std. Recommendation G.729, 1996.

[39] 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. 99, no. 7, pp. 1877–1884, 2016.

[40] Hernández J. R., Pérez-González F., Rodrıguez J. M., and Nieto G., "Coding and synchronization: A boost and a bottleneck for the development of image watermarking," in *Proc. of the COST# 254 workshop on Intelligent Communications*. Citeseer, 1998, pp. 77–82.

[41] Norden E. H., Zheng S., Steven R. L., Manli C. W., Hsing H. S., Quanan Z., Nai-Chyuan Y., Chi C. T., and Henry H. L., "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[42] Jolliffe I. T. and Cadima J., “Principal component analysis: a review and recent developments,” *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[43] Vincent E., Sawada H., Bofill P., Makino S., and Rosca J. P., “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.

[44] 福井勝宏，島内末廣，日岡裕輔，中川朗，羽田陽一，大室仲，片岡章俊，“雑音抑圧のための雑音対信号比率に基づく雑音パワー推定”，Journal of Signal Processing，vol. 18，no. 1，pp. 17–28，2014.

研究業績

国内発表

1. 渡辺瑠伊，鵜木祐史，“非負値行列因子分解に基づく音響ゼロ電子透かし,” 電気・情報関係学会北陸支部連合大会，H2-4，2024.

受賞

1. 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，2024.