

セキュア音声コミュニケーションのための音声情報ハイディング法

Speech information hiding method for secure speech communication



鵜木 祐史 (Masashi UNOKI, Ph. D.)

北陸先端科学技術大学院大学 先端科学技術研究科
教授

(Professor, Japan Advanced Institute of Science and
Technology, Graduate school of Advanced Science and
Technology)

日本音響学会 アメリカ音響学会 電子情報通信学会 ISCA IEEE 他
受賞: IEICE Fellow (2022) 日本音響学会第40回佐藤論文賞 (2000年)
日本音響学会第50回佐藤論文賞 (2010年) 日本音響学会第54回佐藤論文
賞 (2014年) 電子情報通信学会論文賞 (2017年) 他
著書: 音声情報ハイディング, コロナ社出版 (2018年)
研究専門分野: 聴覚科学 音声情報処理 音声セキュリティ

あらまし

情報通信技術や AI 技術の急激な発展により、音声セキュリティの向上が喫緊の課題となっている。この課題を解決する方法として、音声情報ハイディング^{*1}に注目が集まっている。頑健な音声情報ハイディング法として直接スペクトル拡散 (DSS) 法が知られているが、情報埋め込みによる音質低下が大きな問題である。本研究では、音声合成法の一つである線形予測 (LP) 法に着目し、LP 残差を用いたスペクトル拡散型音声情報ハイディング法を提案する。この方法は、LP 残差が DSS 法の疑似乱数雑音 (PN) 信号と同様の統計的性質を持つことに着目して提案されたものであり、情報埋め込みによる音質低下を防ぎ、DSS 法と同様に高い頑健性を有する。さらに、セグメント間の LP 残差の統計的性質に着目することで、ブラインド検出やフレーム同期も同時に実現した技術である。

1. 研究の背景と狙い

近年、情報通信技術 (ICT) の急速な発達やインフラの整備により、インターネット上での音声情報の利用

が盛んである。特に、音声コンテンツの利用は、スマートフォンの普及やアプリ開発の増加、AI スピーカ (音声アシスタント) の登場とともに、この数年で急激な伸びを示している。このような急激な需要増加に対して、音声情報を安心・安全に利用するための技術革新や法整備は相当な遅れをとっており、コンテンツの違法配信や情報改ざんだけでなく、プライバシー侵害やディープフェイクによるなりすましといった社会問題も招いている (図1)。そのため、音声メディアを安心・安全に利用するための仕組みや、音声メディアやその中に含まれる音声プライバシー情報を様々な攻撃から守る革新的技術の確立が喫緊の課題である。

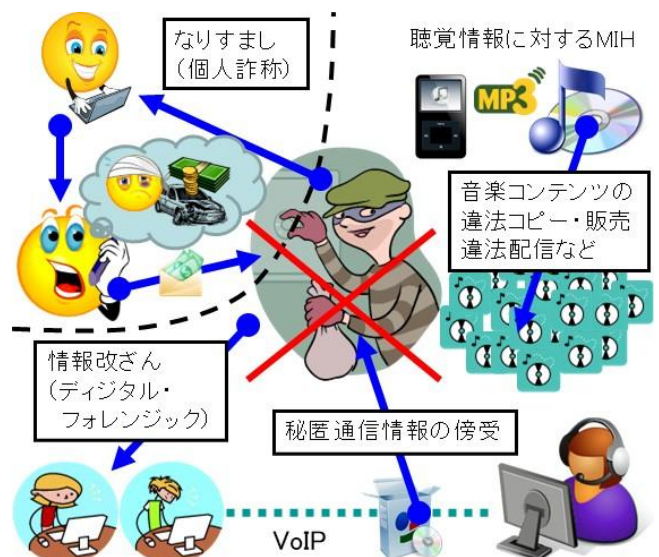


図1 ICTでの社会問題の事例

この課題を解決する一つの解として、マルチメディア情報ハイディング (MIH) 技術が注目を浴びている。MIH 技術は、これまでの暗号技術とはまったく異なるものであり、著作権保護のための電子透かし技術と、コンテンツ自体に別の情報を隠すためのステガノグラフィ技術に大別される。また、従来の暗号技術と併用が可能であるため、コンテンツ認証技術や情報秘匿技術にも応用可能である。現在まで、視覚情報に関するマルチメディア情報 (画像・映像) に対しては、成功例が多く、その技術基盤が整備されつつある。しかし、聴覚情報に関するマルチメディア情報 (音楽・音声) に対しては、視覚と聴覚のモダリティの違いから、音に何かを隠すことが視覚情報に比べて非常に難しく、

セキュア音声コミュニケーションのための音声情報ハイディング法

Speech information hiding method for secure speech communication

さらに符号化・圧縮技術や、悪意のある攻撃に脆弱であるため、革新的な技術の登場が待ち望まれている。

音声情報ハイディング技術 [1-3] の基盤整備には、聴覚科学の研究成果と音声知覚・音声合成の研究の成果が求められる。図2に示すような処理体系において、重要な要求項目は、知覚不可能性（情報の埋め込みに気づかないこと）、秘匿性（隠されていることを知られないこと）、頑健性（音声符号化・圧縮技術の他、悪意ある攻撃に耐性があること）を満たすことである。広範囲の応用を考えると、情報秘匿の埋込容量（ペイロード）も大きいことが望まれる。

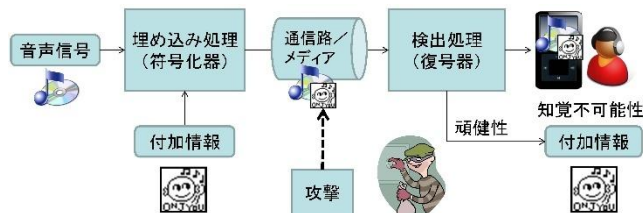


図2 音声情報ハイディングの処理体系

2. 研究の目的

本研究の目的は、デジタル表現された音声メディア情報を安心・安全に利用するために、音声信号への汎用性の高い情報ハイディング技術の基盤を確立することである。特に、音声符号化・圧縮技術に耐性を持たせることで、セキュア音声コミュニケーションを実現するための情報秘匿技術を整備する。

3. 研究の方法

3.1 アプローチ

本研究では、様々な音声符号化・圧縮技術に耐性のある音声情報ハイディング法の実現を目指している。そのため、図3に示すように、音声符号化技術の根本的な理念に原点回帰し、申請者のこれまでに実現した音声情報ハイディング法を俯瞰的に眺め、音声の情報表現の冗長性と聴知覚の頑健性の検討から、知覚不可能性／秘匿性／頑健性を満たすように情報を秘匿可能にする音響特徴を明らかにする。

現在までに、音声生成メカニズムに基づいた音声ボコーダにおいて、声道特性に係わるフォルマント情報と耳に到達するまでの伝達特性（エコー）、さらに聴知

覚特性として、位相変調（蝸牛遅延）の三つが有効であることを確認している。これらは相互に干渉することはないが、総合的な処理によりアーティファクトが発生し、それによって知覚的に検知される可能性がある。そのため、聴覚的顕著性に着目し、音の気づきに影響を与えない特徴を明らかにする。

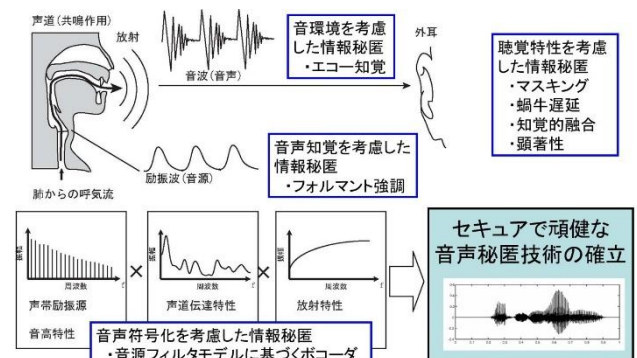


図3 音声情報ハイディングにおける音響特徴

3.2 線形予測残差を利用したスペクトル拡散法

音響情報ハイディングの中で最も頑健性の高いものは、スペクトル拡散（DSS: Direct Spread Spectrum）に基づく方法である。DSSは、デジタル変調で知られる代表的な変調方式であり、DSSに基づく音響情報ハイディングは、疑似乱数雑音（PN）にメッセージ情報を乗じて、対象信号に直接埋め込む方法である。秘匿情報がスペクトルに広範に埋め込まれるため、様々な処理に対して頑健性が高くなるが、スペクトル拡散による歪が音質劣化を招くという大きな問題がある。

DSS法の埋め込み・検出のポイントはPN信号の統計的性質（平均が0、分散が1となる独立な確率変数）にある。本研究では、この点に着目し、図3における音声情報ハイディング法の方式を検討する。特に、線形予測（LP: Linear Prediction）に基づく音声符号化にて、LP残差がPN信号に似た統計的性質を有することに着目し、線形予測に基づくスペクトル拡散型音声情報ハイディング法（LP-DSS法）を検討する。

図4にLP-DSS法の情報埋め込みと検出の概要を示す[4]。この方法では、各フレームにて音声セグメントのLP分析を行い、得られたLP残差を利用してスペクトル拡散変調して、メッセージ信号をホスト信号に埋め込む（図4(a)）。メッセージ信号の検出では、埋

セキュア音声コミュニケーションのための音声情報ハイディング法

Speech information hiding method for secure speech communication

め込んだ信号に対して、同じ LP 残差を乗じ、その平均を求めることでメッセージが得られる (図 4 (b))。

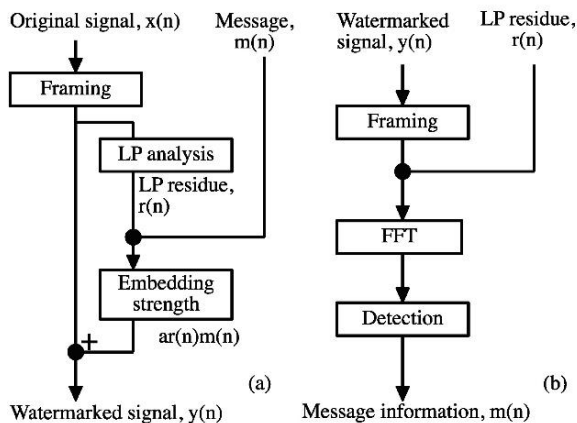


図 4 LP-DSS 法 : (a) 透かしの埋め込みと (b) 検出

この方法は、原則、埋め込みに利用した LP 残差が必要となり、ブラインド検出に制限がある。また、フレーム処理を行っているため、フレーム同期をどのように行うかも重要となる。

3.3 提案法 (ブラインド LP-DSS 法)

図 5 に提案法の処理体系を示す。この方法では上述した(1) ブラインド検出と(2) フレーム同期に対応している。図中の(a)はセキュリティ信号 (メッセージ信号) の埋め込み法を、(b)はその検出法を示す。

本処理体系では、対象音声 $x(n)$ を K 個のフレームに分割する。フレーム分割した信号を偶数次フレームと奇数次フレームに分け、奇数次フレームを用いて LP 残差を求め、セキュリティ信号をスペクトル拡散変調し、偶数次フレームに埋め込む。セキュリティ信号が埋め

込まれた偶数次フレームと奇数次フレームを用いて埋め込み信号を作成する。ここで、セキュリティ信号を頑健に埋め込むために、LP-DSS 法にて強度設定レベルによって埋め込み強度レベルを調整する。

セキュリティ信号の検出では、フレーム同期とフレーム分割を行い、奇数次フレームと偶数次フレームに分ける。奇数次フレームから LP 残差を求め、それを偶数次フレームに乗じることで、正確にセキュリティ信号を得ることができる。

図 6 にフレーム同期のブロックダイアグラムを示す。ここでは、情報を埋め込んだ信号を K 個のフレームに分割する。奇数次のフレームの埋め込み信号から LP 残差を求め、LP 残差を偶数次フレームの情報を埋め込んだ信号に乗じて、その期待値を求める。この処理をサンプルシフトしながら期待値が最大となる点を求めることでフレーム同期を行う。

3.4 評価

音声情報ハイディング法では、一般に、その性能評価において、埋め込みによる音質低下の回避と頑健な検出の実現が問われる。

前者では、知覚的音声品質評価指標 (PESQ) や対数スペクトル歪 (LSD) が利用される。PESQ は一種の MOS 値であり、 $-0.5 \sim 4.5$ (低～高) に付置される。品質維持として 3.0 が概ね基準となる。LSD は 1 dB 未満であると良好な品質であるといわれている。

後者では、検出誤差としてビット検出誤り率 (BER) が利用される。誤り訂正符号などを利用することも考えられるがここではそういった補償をせずに、BER 10% 未満であると良好であるといわれている。

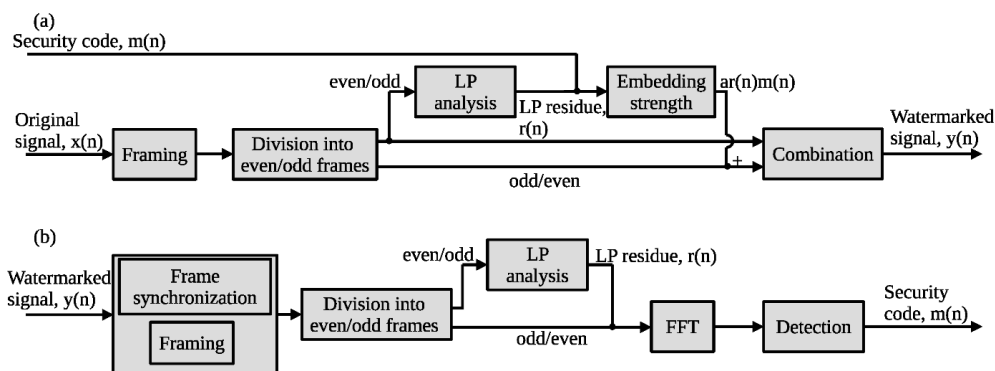


図 5 LP-DSS 音声情報ハイディングの処理体系

セキュア音声コミュニケーションのための音声情報ハイディング法

Speech information hiding method for secure speech communication

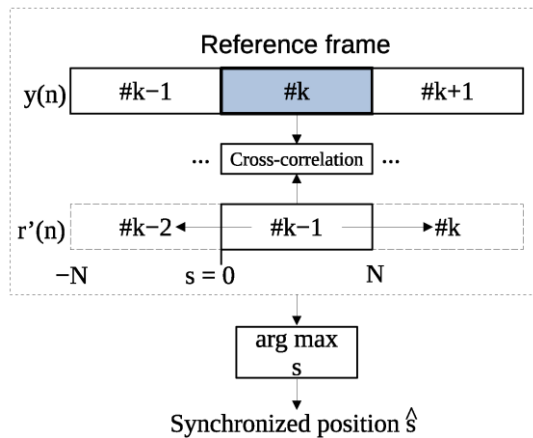


図6 フレーム同期

提案法の有効性を確認するために、次の三つの評価を行った。一つ目は埋め込み強度レベルに対応した音質評価と BER である。二つ目は、従来法と比較した音声の品質評価である。三つ目は、従来法と比較した提案法の頑健性の評価である。

これらの評価では、ATR 音声データベース (B セット) [5] にある 12 文章をホスト信号として利用した。セキュリティ情報(メッセージ情報)は“01010101...”, ペイロードを 4, 8, 16 bps とした。埋め込み強度レベルは -20 dB から 0 dB の 5 dB 刻みで設定した。

図7に一つ目の評価結果を示す。音声の品質評価では、PESQ 3.0 と LSD 1 dB を基準とすると埋め込み強度レベルが -15 dB が適切であることがわかる。一方で検出性能では、BER 10%を基準とすると、埋め込み強度レベルが -10 dB が適切であることがわかる。両者の間にはトレードオフの関係があり、埋め込み強度が -15 ~ -10 dB に適切な値があることがわかる。

図8に二つ目の結果を示す。従来法として、LSB 置換法(量子化における LSB のビット操作) [1, 2], DSS 法(従来の直接的スペクトル拡散法) [6], MC 法(McAdams 法を利用した LP 音声への秘匿法) [7] を利用した。図8(a)は PESQ の結果を、図8(b)は LSD の結果を示す。図中からもわかるように LSB 法が最も音声品質を高く(情報ハイディングによる歪が小さく)維持でき、DSS 法が最も性能が低いことがわかる。また、提案法は MC 法と同程度の音声品質をもち、評価基準を満たすものであることがわかる。

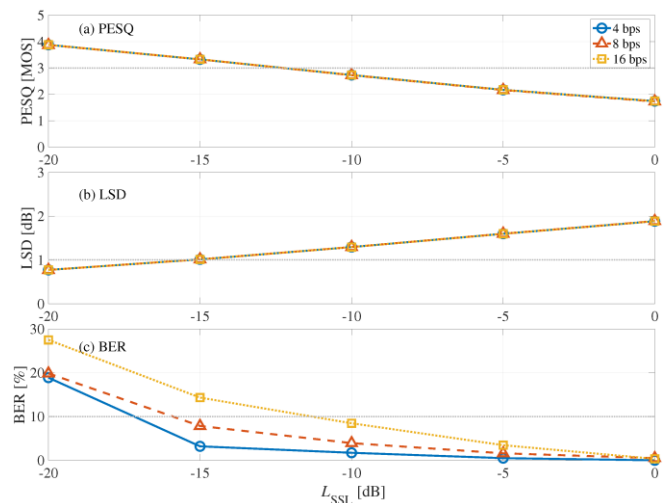


図7 埋め込み強度レベルに対する提案法の性能評価 : (a) PESQ, (b) LSD, (c) BER

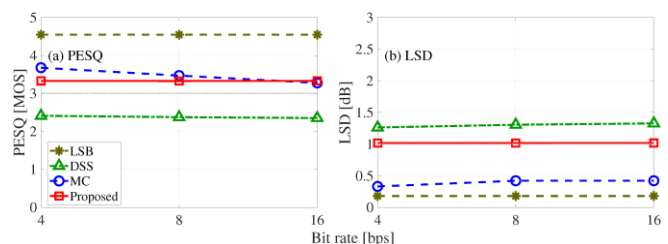


図8 情報ハイディングされた音声の品質評価

図9に三つ目の評価結果を示す。図9(a)は攻撃等がないときの結果であるが、いずれの方法とも正確に情報を検出できることがわかる。

次に、情報ハイディングされた音声に対する検出の耐性について検討する。ここでは、原音声のサンプリング周波数が 20 kHz、量子化ビットが 16 bits であるため、リサンプリングしたとき (b) 12 kHz ヘッドダウンサンプリング、(c) 24 kHz にアップサンプリング) と再量子化 ((d) 8 bits, (e) 24 bits) したときの頑健性について図示する。これらは一般に攻撃というよりは、情報表現の変更に伴う耐性試験である。次に、情報圧縮 ((f) OGG と(j) MP4) や音声符号化 ((g) G711, (h) G723.1, (i) G726) に対する頑健性について図示する。

これらの結果を俯瞰すると、LSB 法は音声の品質維持には非常に優れるが、情報変換や情報圧縮、音声符号化には脆弱であることがわかる。一方で、DSS 法は音声の品質維持において問題が見受けられるが、これらの耐性試験においては一番性能がよいことがわかる。

セキュア音声コミュニケーションのための音声情報ハイディング法

Speech information hiding method for secure speech communication

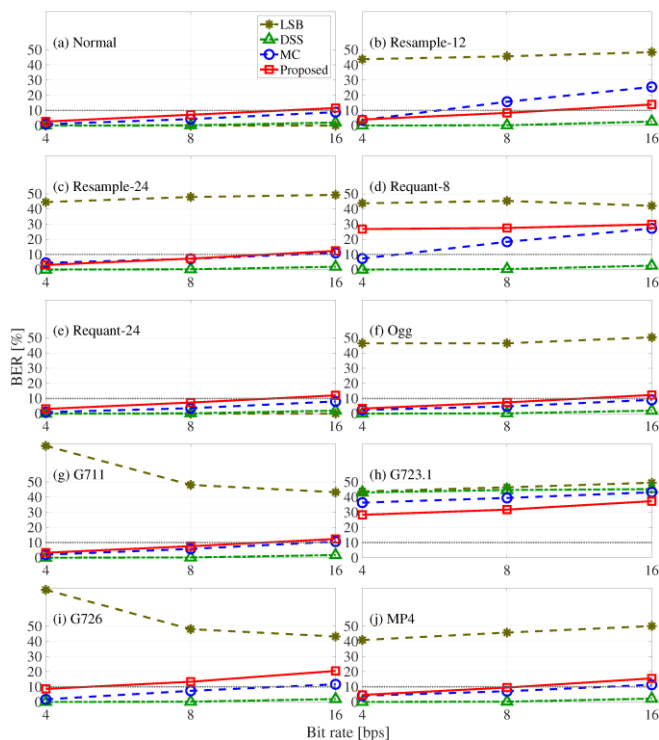


図9 音声情報ハイディングの頑健性の評価

提案法とMC法では、音質評価において同程度の性能を持つが、耐性試験においては提案法がMC法よりも優れていることがわかる。

以上の点から、総合評価の観点から、音声情報ハイディングとして提案法がすぐれた性能を有することがわかる。

3.5 改善点と残された課題

提案法(LP-DSS)は、原信号の利用を許したノンブラインド法であれば、音声の品質は現状のものよりPESQで0.5、LSDで0.5 dB低減することがわかっている。合わせて埋め込み強度レベルもブラインド法より10 dBほど弱く埋め込みしてもよいことがわかっている。しかし、音声セキュリティの点から利用範囲を拡充するためにはブラインド法であることが望まれる。特に、フレーム同期も同時に行うことから、埋め込み強度レベルを強めにする必要もあり、結果的に音声の品質評価においてノンブラインド法より性能低下してしまうことになる。

この点に関しては、LP-DSS法でホスト信号にセキュリティ情報を埋め込む際、マスキング特性を加味し

て緩和する方法がある[8]。この方法によりブラインド法での音声の品質評価を上げることが可能である。

最後に本研究で残された課題は、提案法のペイロードの向上(埋め込み情報量の増加)と音声合成技術への組み込みである。これらについては継続して取り組む予定である。

4. 将来展望

将来的な発展可能性として、可逆性を満たす音情報ハイディングを実現できれば、ロスレス情報ハイディングや埋め込み情報の再編集が可能である。さらに埋め込みの大容量化を実現できれば、音声そのものに、様々な付加価値情報や、その量にもよるが別の音声をこっそり隠すこともできる。また、音声圧縮や極低ビットレート音声合成の要素技術としても応用可能である。これらの実現にはまだ多くの高いハードルがある。

一方で、音声情報ハイディング技術を利用した具体的な応用課題[9, 10]がある。昨今話題になっている音声改ざん検出や音声なりすましといった、重要な音声メディアのセキュリティ問題を解決するものである。

図10に示すように、音声データの真正性を担保するために、音声符号化には耐性があるが、音声改ざんといった攻撃には脆弱な音声情報ハイディング法を利用する。音声改ざん検出法では、音響情報ハイディングの埋め込み部、検出部、識別部の三つの処理ブロックからなる。もしこのような方法を実現できれば、音声データの真正性担保を実現でき、コンテンツの法科学的な利用にも役立つかもしれない。また、検出された透かし情報の欠落箇所からどこが改ざんされたか(改ざんの場所)、どのような改ざんを受けたのか(改ざんの様態)まで検出可能になれば、その利用価値もさらに高まるものと考えられる。

最後に、図11に示すように、現在、ニュース等でも話題になっているディープフェイク音声といったなりすまし音声²の対策にも活用できる。流通データに事前に重要な秘匿情報を音声に埋め込んでおき、それらが流通されたあとに何らかの音声処理(例えば深層学習ベースの音声合成)によって改変された場合、それがフェイクなのか本物なのかを判定する(メディアクローン検出)必要がある。この課題で重要なことは、

セキュア音声コミュニケーションのための音声情報ハイディング法

Speech information hiding method for secure speech communication

あらゆる音声符号化処理に対し頑健であり、かつなりすまし処理に関しては脆弱であるような音声情報ハイディング法を確立する必要がある。

これらはまだ研究段階であるが、継続的な課題として取り組む予定である。

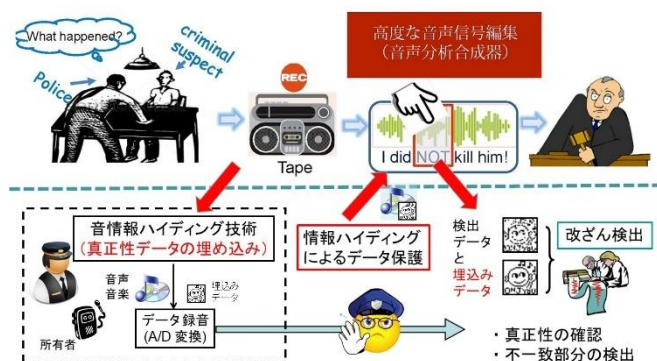


図 10 音声改ざん検出の例

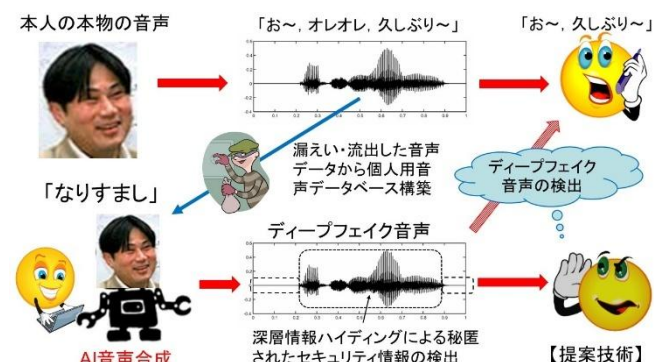


図 11 ディープフェイク音声の対策の例

用語解説

*1 音声情報ハイディング

デジタル音声信号にセキュリティに関わる情報を聴こえないように隠し、必要ときにその情報を頑健で正確に検出できる技術の総称

*2 なりすまし音声（ディープフェイク音声）

AI 音声合成を利用して本物そっくりに作り上げた偽物の音声のこと。サイバーセキュリティのリスクが非常に高いもの。

参考文献

- [1] Cvejic, N. and Seppanen, T. L., Digital Audio Watermarking Techniques and Technologies, IGI Global, Hershey, PA, USA, 2007.
- [2] Hua, G., Huang, J., Shi, Y. Q., Goh, J., and Thing, V. L. L., “Twenty years of digital audio watermarking - a comprehensive review,” Signal Processing, 128, 222–242, 2016.
- [3] 鵜木祐史, 他 5 名, 音響情報ハイディング技術, コロナ社, 2018.
- [4] Namikawa, R. and Unoki, M., “Non-Blind Speech Watermarking Method Based on Spread Spectrum Using Linear Prediction Residue,” IEICE Trans. Inf. Syst., E103.D, 63–66, 202.
- [5] Takeda, K. et al., “Speech Database User’s Manual, ATR Technical Report TR-I-0028, 2010.
- [6] Boney, L., Tewfik, A. H., Hamdy, K. N., “Digital Watermarks for Audio Signals,” Proc. of 3rd IEEE International Conference on Multimedia Computing and Systems, 473–480, 1996.
- [7] Mawalim, C. O., and Unoki, M., “Speech Watermarking Method Using McAdams Coefficient Based on Random Forest Learning,” Entropy, 23, 1246, 2021.
- [8] Adila, A., Mawalim, C.O., Isoyama, T., and Unoki, M., “Study on Inaudible Speech Watermarking Method Based on Spread Spectrum Using Linear Prediction Residue,” J. Signal Processing, 28(6), 309-313, 2024.

関連文献

- [9] 鵜木祐史, “音声メディアへの情報ハイディングとその応用,” 日本音響学会誌, 77, 6, 382-389, 2021.
- [10] 鵜木祐史, “聴覚特性に基づいた音響情報ハイディング技術,” IEICE Fundamental Review, 13(4), 284-293, 2020.

この研究は、令和3年度SCAT研究助成の対象として採用され、令和4～6年度に実施されたものです。