様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ５ 月 １１ 日現在

機関番号：３４３０４
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００２３４

研究課題名（和文）音声の長時間位相スペクトルを利用した画像の音変換に関する研究

研究課題名（英文）Image to Sound Mapping Method Using Spectral Phase on Long-Term Fourier Transform

研究代表者
川村　新（KAWAMURA, Arata）

京都産業大学・情報理工学部・教授

研究者番号：６０３６２６４６

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：音声をフーリエ変換し，角周波数の振幅スペクトルを輝度として，時間方向に並べた画像をスペクトログラムと呼ぶ．本研究では，一般的な画像を音声のスペクトログラムに埋め込み，音声を合成する方法について検討した．画像をスペクトログラムに埋め込むと，音声の振幅スペクトルが失われる．しかし，音声の長時間位相スペクトルを利用すれば，明瞭度のある音声を合成することができる．そこで提案法では，画像を振幅スペクトル，音声を位相スペクトルに対応させることで，画像から明瞭度のある音声を合成した．提案法では，合成音声から埋め込み画像を復元する際に，一般の情報埋め込み技術とは異なり，振幅スペクトルそのものが画像を表現する．

研究成果の学術的意義や社会的意義
本研究では，画像を埋め込んだ合成音声をスピーカ等から放射し，受信側で音声から画像を復元する．この技術が完成すれば，音声から得られる言葉の情報とともに，画像情報も同時に伝達できる．また，WiFi環境が整備されていない場所でも受信が可能となり，受信可能範囲も，スピーカの音量調整により制御可能となる．応用例は多岐にわたり，防災用スピーカからの緊急放送に避難経路や災害現場の写真を埋め込む，ラジオの天気予報に天気図を埋め込む，絵本の読み聞かせに該当ページの絵を埋め込む，タイムセール放送に商品や売り場の地図を埋め込む，海外のバスや電車の音声アナウンスに翻訳情報を埋め込む，などが考えられる．

研究成果の概要（英文）：In this study, we proposed an image to sound mapping method. This technique treats an image as a spectrogram and maps it to a sound by taking inverse Fourier transform of the spectrogram. The embedded image destroyed speech spectral amplitude. We compensate the speech quality by using a speech spectral phase obtained by taking LTFT (Long-Term Fourier Transform). The speech spectral phase on LTFT contains speech intelligibility. The proposed method synthesis a speech signal with spectrogram consisting of an original image and speech spectral phase on LTFT. The synthesis speech signal is transmitted from a loudspeaker, and received at a microphone equipped on a mobile device. The received speech signal is transformed to a spectrogram which directly displays the transmitted image. The proposed method does not require any special transformation technique excepted of Fourier Transform.

研究分野：音声音響信号処理

キーワード：画像の音変換　スペクトログラム　位相スペクトル　長時間フーリエ変換

１．研究開始当初の背景

火災や地震など，災害時の避難放送では，音声による情報伝達が行われる．避難放送音声によって伝えられる緊急性は，受け取り側により個人差が生じる．緊急性の認識不足により，甚大な被害が生じる危険性がある．音声情報と共に，被害状況を視覚的に確認できる画像を伝送できれば，より正確に緊急性を伝えることが可能になると考えられる．

音声に画像を埋め込むことができれば，WiFi 環境が整備されていない場所でも音波による画像の送受信が可能となり，受信範囲も，スピーカの音量調整により制御できる．応用例は，前述のように避難放送に被害状況写真を埋め込むことや，ラジオの天気予報に天気図を埋め込む，絵本の読み聞かせに該当ページの絵を埋め込む，タイムセール放送に商品写真や売り場への経路を埋め込む，海外のバスや電車の音声アナウンスに翻訳情報を埋め込む，など多岐にわたる．

このような背景から，本研究では，音声に画像を効率的に埋め込む方法について検討する．一般的な音声への情報埋め込み技術では，音質の劣化は小さいものの，埋め込める情報量が少ないため，画像を送信する場合に伝送時間が非常に長くなる．一方，画像をスペクトログラムとみなして，逆フーリエ変換により，音を合成する方法がある．スペクトログラム自体が画像なので，音による画像の伝送時間は短縮できる．ただし，得られる合成音は，人間の音声とはかけ離れたものとなる．ここで，音声の長時間フーリエ変換（LTFT: Long-Time Fourier Transform）によって得られる「長時間位相スペクトル」を利用すると，振幅スペクトルが正確でなくとも，明瞭度のある音声が復元できることが知られている．そこで本研究では，画像で振幅スペクトルをつくり，音声で LTFT の位相スペクトルをつくる．そして両者を融合することで，画像から明瞭度のある音声を合成する．合成音から元の画像を復元する際には，単純に合成音のスペクトログラムを作成すればよい．一般の情報埋め込み技術と異なり，本手法では音声の振幅スペクトルそのものが画像を表現することに新規性がある．

２．研究の目的

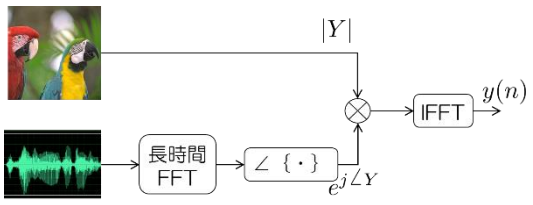
本研究では，画像を埋め込んだ合成音声をスピーカ等から放射し，マイクロホンで受信した音声から画像を復元することを目的としている．この際，受信者は，音声による言語情報と画像による視覚情報を同時に取得することができる．

最初に，音声のスペクトログラムの大部分を画像に置き換え，かつ合成音に明瞭度が得られる方法について検討する．音質劣化を最小限にするために，パワーの大きい音声の振幅スペクトルは保持し，それ以外のスペクトルに画像を埋め込む方法を確立する．

次に，別の手法として，超音波に近い高域のスペクトログラムにのみ，画像を埋め込む方法についても検討する．この方法では，音質劣化を小さくできるが，画像の伝送時間は長くなる．そこで，反復位相復元と呼ばれる位相スペクトルの生成法を導入し，伝送時間を半減する方法を確立する．

３．研究の方法

(1) 従来法では，図 1 に示すように，スペクトログラム（振幅スペクトル）を画像，位相スペクトルを音声として，両者を融合することで合成音声を得る．結果として，合成音は音声の明瞭度をある程度保持し，そのスペクトログラムは画像となる．



y(n)は音声として聞こえる
y(n)のスペクトログラムは画像になる

図１　画像と音声の融合

ここで，音声のスペクトログラムをすべて画像に変更すると，著しい音質劣化が生じるため，従来法では，音声のスペクトログラムの一部の領域に画像を埋め込んでいた．より具体的には，音声のスペクトログラムのうち，指定された矩形領域を画像に置き換え，合成音声を得ていた．当然ながら，指定領域を小さくすると音質は改善し，一定時間に伝送できる画像サイズは小さくなる．逆に，指定領域を大きくすると，画像の伝送時間は短縮するが，合成音声の音質劣化が著しくなるという，音質と伝送時間のトレードオフがあった．

そこで本研究課題では，音声の振幅スペクトルのうち，しきい値以上の振幅スペクトルを保持しておき，それ以下のスペクトルにのみ画像を埋め込む方法を検討した．図 2 に提案法の概要を示す．ここで，強いパワーをもつ振幅スペクトルはスペクトル包絡を保持する効果があり，音

声の明瞭度保持に貢献できると考えられる．提案法では，音質への寄与が小さい振幅スペクトルだけを選択して画像を埋め込むため，従来法と同じ容量の画像を埋め込んだ場合でも音質改善が可能となる．提案法の有効性を確認するため，シミュレーションを行い，提案法と従来法の音質を評価した[1]．

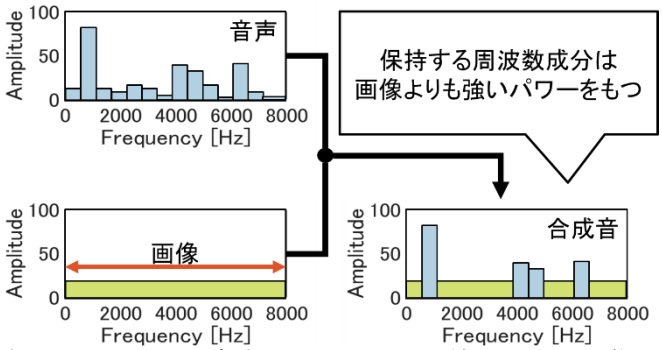


図 2　強いパワーをもつ音声スペクトルを保持したまま画像を埋め込む

(2) 方法(1)では，音声の振幅スペクトルのうち，小さいスペクトルだけを画像に置き換える手法により，音質劣化を抑制した．しかし，合成音の音質劣化は明らかに知覚できる．実用化のためには，さらなる音質改善が必要である．そこで，画像の埋め込み帯域を，超音波帯域に近い帯域に限定し，知覚できなくする方法を検討した．提案法のイメージを図 3 に示す．ここで，サンプリング周波数を 48kHz とし，19kHz 以上の帯域を埋め込みに利用する．

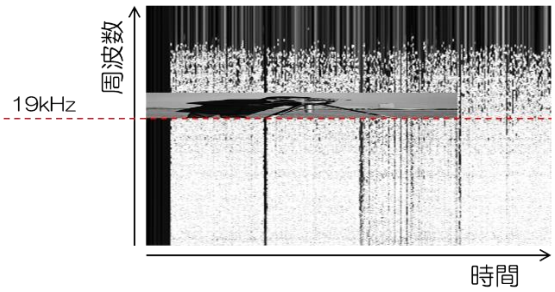


図 3　高域に画像を埋め込むイメージ

画像を埋め込む帯域を超音波に近い帯域に限定すると，全体の埋め込み量が縮小する．したがって，画像の伝送時間が長くなる．そこで，オーバーラップ加算を含むスペクトログラムを作成することで，画像の伝送時間を短縮する．しかし，図 4 に示すように，オーバーラップ加算により，画質が劣化するという新たな問題が生じた．

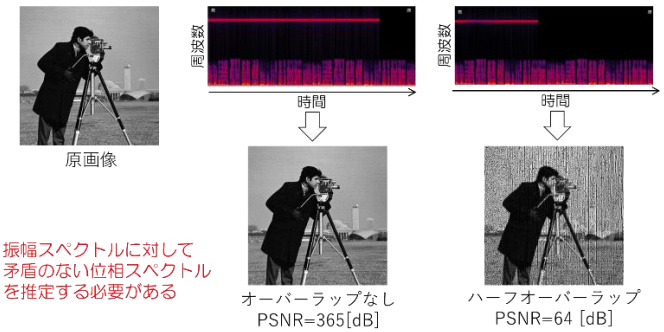


図 4　オーバーラップが含まれる場合の画質の劣化

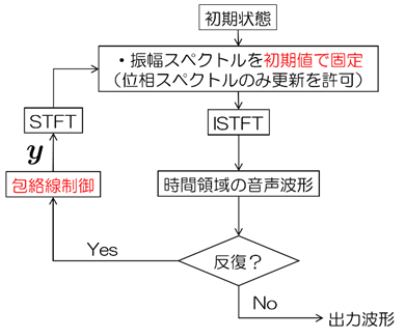


図 5　反復位相復元法

提案法では，この画質劣化を反復位相復元法の導入により改善する．提案する反復位相復元法の概略を図 5 に示す．ここで，反復回数が大きいほど，画質が改善される[2]．

４．研究成果

(1)　音声の振幅スペクトルのうちしきい値以上の振幅スペクトルを保持しておき，それ以下のスペクトルにのみ画像を埋め込む方法を検討した．提案法では，音質への寄与が小さい振幅スペクトルだけを選択して画像を埋め込むため，従来法よりも音質劣化を抑制できる．シミュレーションにより，提案法と従来法の合成音声の音質を比較した．ただし，両者の復元画像の画質は同

じである．音質の評価は，スペクトル包絡に対する MSE（最小二乗誤差）および時間波形に対する SNR（信号対雑音比）で行った．ここで，MSE は小さいほど性能が高く，SNR は大きいほど性能が高い．評価結果を図 6 に示す．結果から，20%の画像埋め込み率に対して，従来法の MSE は 13.4，提案法では 12.4 であり，約 1 ポイントの改善となった．また，SNR では従来法が 11.2dB，提案法 16.8dB で，5.6dB の音質改善を達成した．

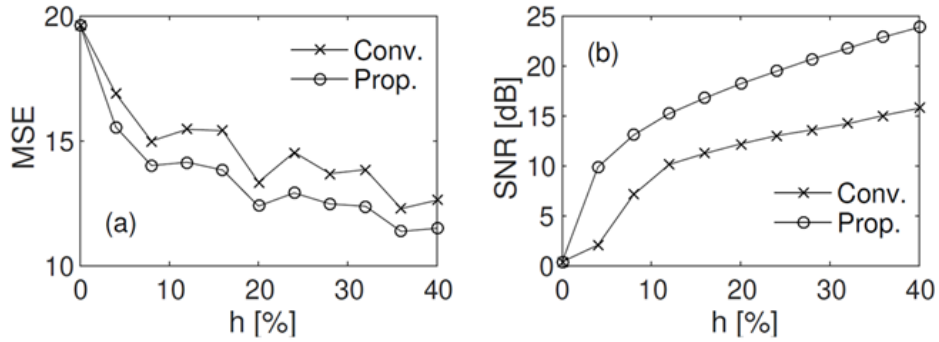


図 6　(a) MSE による画質の評価結果．(b) SNR による音質の評価結果．h はスペクトログラム全体に対する画像の埋め込み率を表す．

(2) 次に，音質劣化をさらに抑制するため，音声のスペクトログラムのうち，超音波に近い帯域にのみ画像を埋め込む方法を検討した．画像の埋め込み帯域が縮小されるため，オーバーラップを含むスペクトログラムに埋め込むことで，埋め込み容量を増やした．一方で，オーバーラップに起因する画質劣化が生じる．そこで，反復位相復元法により，伝送時間の短縮と画質の改善を行った．結果を図 7 に示す．同図では，オーバーラップの深さと反復位相復元の反復回数による，PSNR（画質評価）の変化を示している．PSNR が大きいほど画質が良いことを表している．

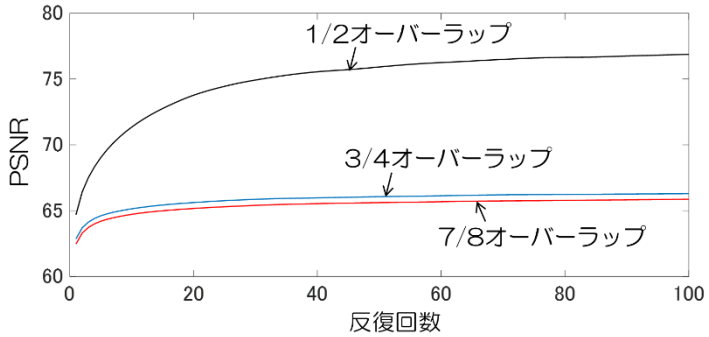


図 7　反復回数と PSNR の結果

結果から，反復回数が大きいほど画質は改善し，オーバーラップが深いほど，画質の改善は小さくなることがわかった．結果の一例を図 8 に示す．

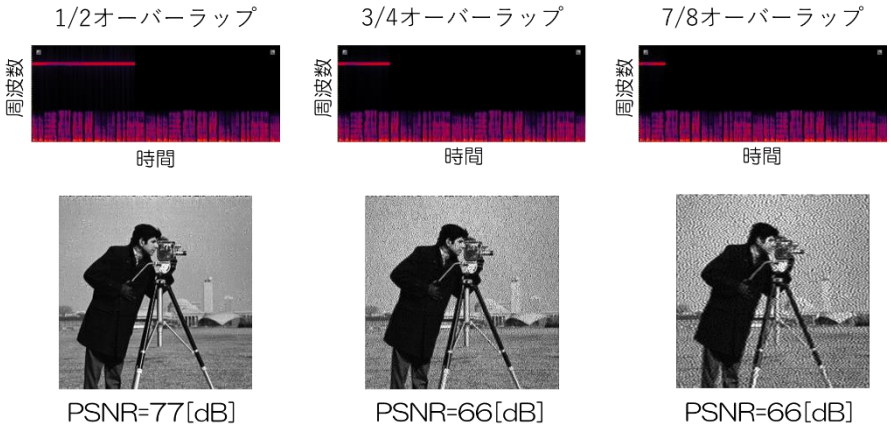


図 8　反復位相復元による埋め込み画像の確認（反復 100 回）

さらに，合成音の受信機を制作し，屋外実験を行った[3]．受信機および屋外実験の様子を図 9 に示す．ここで，受信機は Raspberry Pi 2 を用いて制作した．屋外実験ではスピーカから合成

音を放射して，受信機で音声を受信し，画像を復元した．また，スピーカから放射する合成音のスペクトログラムを図 10 に示す．今回は 3 種類の合成を作成した．

(a) 受信機　　(b) 屋外実験．スピーカから合成音を放射．

図 9　制作した受信機と屋外実験の様子

- 12kHz~15kHz，15kHz~18kHz，17kHz~20kHzに画像を埋め込む。
- 音声はGLの反復回数を75回、オーバーラップ数は2で合成する。

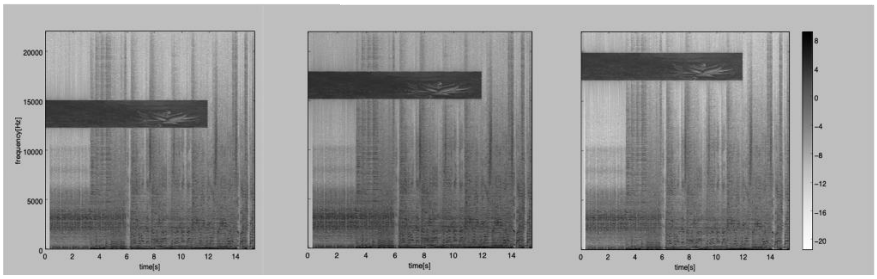


図 10 スピーカから放射する合成音のスペクトログラム．埋め込み上限を 15kHz，18kHz，20kHz として 3 種類の合成音を作成．反復回数は 75 回，オーバーラップは 1/2.

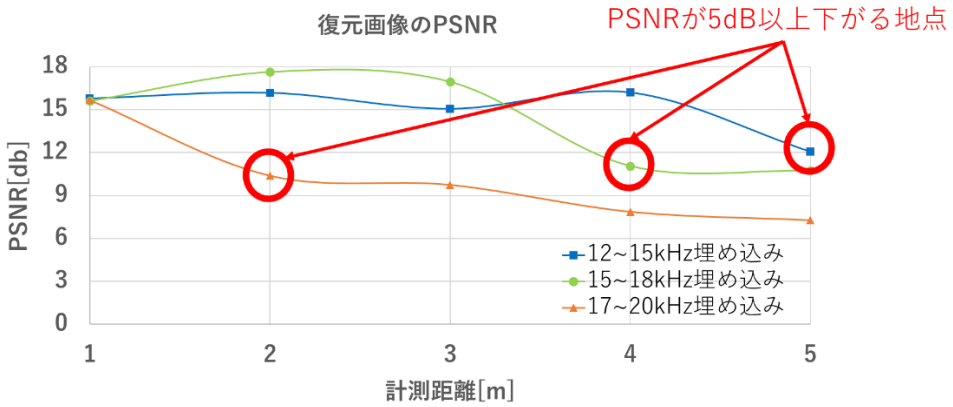


図 11　距離に対する受信画像の PSNR

図 11 に受信音声から復元した画像の評価結果を示す．ここで，距離に対する PSNR を示している．結果から，超音波帯域に近い帯域に埋め込むほど，伝送距離が短くなることがわかった．PSNR が 5dB 低下する地点を考えると，埋め込み上限 20kHz であれば 2m，15kHz であれば，5m 程度であった．本研究課題は実機の製作までを通し，概ね，予定通り進めることができた．しかし，提案法の実用化に向けては，さらに到達距離を延長する必要がある．

＜引用文献＞

[1] Y. Hosoda, A. Kawamura, and Y. Iiguni, "An efficient image to sound mapping method preserving speech spectral envelope", IEICE Trans. Fundamentals, vol.E103-A, no.3, pp.629-630, March 2020.

[2] 川村新，"反復位相復元を利用した音声スペクトログラムへの画像埋め込み"，電子情報通信学会，信学技報，pp.163-168，March 2020.

[3] 小野幸大，画像埋め込み音声の送受信システム，京都産業大学コンピュータ理工学部特別研究報告，March 2021.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 HOSODA Yuya、KAWAMURA Arata、IIGUNI Youji	４．巻 E103.A
２．論文標題 An Efficient Image to Sound Mapping Method Preserving Speech Spectral Envelope	５．発行年 2020年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 629～630
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.2019EAL2139	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Arata Kawamura	４．巻 117
２．論文標題 On Sound Signal Processing in Image to Sound Mapping Technique	５．発行年 2017年
３．雑誌名 Elsevier Applied Acoustics	６．最初と最後の頁 1-11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.apacoust.2016.10.014	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 KAWAMURA Arata、IGARASHI Hiro、IIGUNI Youji	４．巻 E100.A
２．論文標題 An Efficient Image to Sound Mapping Method Using Speech Spectral Phase and Multi-Column Image	５．発行年 2017年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 893～895
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E100.A.893	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 川村新
２．発表標題 反復位相復元を利用した音声スペクトログラムへの画像埋め込み
３．学会等名 電子情報通信学会　信号処理研究会　技術報告会資料
４．発表年 2020年

１．発表者名 Yuya Hosoda, Arata Kawamura, and Youji Iiguni
２．発表標題 Image-to-sound transformation using inpainting technique
３．学会等名 The 2018 International Symposium on Nonlinear Theory and its Applications（国際学会）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関