博 士 論 文

光レーザマイクロホンを用いた音声強調システムに関する研究
(Speech Enhancement System with Optical Laser Microphone)

2023 年 9 月

立命館大学大学院情報理工学研究科
情報理工学専攻博士課程後期課程

CAI Chengkai

立命館大学審査博士論文

光レーザマイクロホンを用いた音声強調システムに関する研究
(Speech Enhancement System with Optical Laser Microphone)

2023 年 9 月

September 2023

立命館大学大学院情報理工学研究科

情報理工学専攻博士課程後期課程

Doctoral Program in Advanced Information Science and Engineering

Graduate School of Information Science and Engineering

Ritsumeikan University

CAI Chengkai

サイ　セイカイ

研究指導教員：西浦　敬信教授

Supervisor: Professor NISHIURA Takanobu

本論文は立命館大学大学院情報理工学研究科に
博士 (工学) 授与の要件として提出した博士論文である.

CAI Chengkai

審査委員：　主査　西浦 敬信 教授
副査　山下 洋一 教授
副査　谷口 忠大 教授

光レーザマイクロホンを用いた音声強調システムに関する研究 (Speech Enhancement System with Optical Laser Microphone)*

CAI Chengkai

内容梗概

音声収集は日常生活でよく使用される情報獲得の方法であり，録音するために様々なデバイスが開発されている．しかし，災害救助や防犯セキュリティなどでは，遠方から音声を取得することが必要となる．このような場合において，通常の音声収集技術，すなわち気伝導マイクロホンを使用すると，遠方からの音声を観測することが困難となるだけでなく，マイクロホン周辺の雑音も混入する．この問題を解決するために，光レーザマイクロホンを用いて，音源の近くにある物体にレーザ光を照射することで，音源により発生した振動を計測し，その振動を音に復元する技術が提案された．一般的なマイクロホンと比べて，レーザ光は高い直進性を持つため，遠方からの音を計測することが可能となり，マイクロホン周辺の雑音に影響を受けない特徴を持っている．

しかしながら，この技術には幾つの問題点がある．光レーザマイクロホンは音源付近の振動物体を介して音声を収録するため，観測した音声の音質は被照射物体の振動特性に依存する．例えば，振動物体の振幅が小さな周波数帯域において，音声成分は欠落する可能性が高い．また，表面の粗いの物体にレーザ光を照射すると，レーザ光の反射率が低下するため，反射光の強度は小さくなり，雑音が混入する．高音

*立命館大学大学院 情報理工学研究科 情報理工学専攻 博士論文.

質の音声を取得するために，光レーザマイクロホンの観測音声に対して，音声強調などの処理を適用することが必要となる．通常の音声強調手法は観測音声から雑音パワーを推定して，観測音声からその推定雑音成分を除去する．もしくは，機械学習を用いて，事前に観測音声とクリーン音声のパワースペクトルのマッピングを学習することで目的音声を強調する．両手法において，強調音声を生成する際に，観測音声の位相情報そのまま利用するため，高音質な強調音声を得ることは難しい．また，雑音パワー推定の手法では物体の振動により発生した周波数帯域成分の欠落を復元することは不可能である．機械学習を用いた手法は欠落した周波数帯域成分の復元は一部可能であるが，位相に起因した歪みや欠落には，対応が困難であった．加えて，両手法とも，未知の被照射物体に対しては，性能を補償できないという問題があった．

そこで，本論文では，深層学習を利用した光レーザマイクロホンのための 3 つ音声強調手法を提案する．既知の被照射物体において，スペクトルと時間波形それぞれに基づいて，二段階処理深層学習を利用した 2 手法を提案する．スペクトルに基づく手法は，パワースペクトルを復元した上で，位相スペクトルの復元も考慮し，2 つのネットワークを利用することで，パワースペクトルと位相スペクトルの各々を復元する．時間波形に基づいた手法は，観測音声の低域と高域における歪みの違いに着目する．まず，観測音声とクリーン音声の低域成分のマッピングを学習し，低域成分の復元を行う．その後，復元した低域成分を利用して，高域成分を復元する．未知の被照射物体においては，まず，パワースペクトルのフレーム方向の差分を利用して，パワースペクトルの包絡を推定する．そして，観測音声の基本周波数情報を抽出し，推定したパワースペクトルの包絡を用いて，強調音声のパワースペクトルを復元する．最後に，パワースペクトルから位相スペクトルを推定し，強調音声を生成する．提案した 3 つの手法において，客観的評価実験を行った結果，提案手法の有効性を確認した．

キーワード

光レーザマイクロホン，レーザドップラ振動計，音声強調，未知物体に対応，遠隔収録，深層学習

Speech Enhancement System with Optical Laser Microphone*

CAI Chengkai

Abstract

Speech measurement is a common method for information acquisition in daily life. For speech acquisition, various devices have been developed. However, in the case of disaster relief and crime prevention security, it is necessary to acquire speech from a long distance. In such a case, the ordinary speech acquisition method is to use an air-conduction microphone. It is not only difficult to acquire speech from a distance but also includes noise around the microphone. Here, an optical laser microphone was proposed to measure the vibration generated by the sound source by irradiating an object near the sound source using a laser Doppler vibrometer (LDV) and restoring the vibration to sound. Since the laser beam has strong straightness, it can measure the sound from very far away.

However, there are several problems with this technology. Since the sound quality of the observed speech depends on the vibration characteristics of the object, the speech component is lacking in the frequency bands with small amplitude responses. Also, when irradiating laser to an object with a rough surface, the intensity of the reflected light is reduced so that noise is included in the acquired speech. Moreover, the characteristics of the optical laser microphone (laser wavelength, sampling rate, etc.) can affect the sound quality. Therefore, in order to obtain high-quality speech, it is necessary to apply speech enhancement processing to the speech acquired by

*Doctoral Dissertation, Advanced Information Science and Engineering, Graduate School of Information Science and Engineering, Ritsumeikan University.

optical laser microphone. Conventional speech enhancement methods estimate noise components from observed speech or use machine learning to obtain the mapping between observed speech and clean speech. However, in both methods, the phase information of the observed speech is used when generating the enhanced speech. The noise power estimation method cannot restore the lacking frequency band components caused by the amplitude response. Methods using machine learning can restore lacking parts to some extent, but they cannot deal with phase-induced distortions. Moreover, existing models cannot be used when the irradiated object is unknown.

In this paper, three speech enhancement methods for optical laser microphone using deep learning were proposed. In the case that the irradiated object is known, considering the deterioration of observed speech, two speech-enhancement methods that are based on the frequency and time domains were proposed. Both methods use multiple deep neural networks to handle different types of deterioration. With our frequency-domain-processing method (hereafter, short-time Fourier transfer (STFT)-based method), the noise power in the power spectrum of the observed speech is first removed. The phase difference between the observed speech and clean speech is then calculated using the noise-suppressed power spectrum to obtain the phase spectrum of the enhanced speech. With our time-domain-processing method (hereafter, waveform-based method), the low-frequency waveform is first restored, and then the high-frequency waveform is estimated using the restored low-frequency waveform. In the case of the irradiated object is unknown, the envelope of the power spectrum is first estimated by using the difference in the frame direction of the power spectrum. Then, the fundamental frequency information of the observed speech is extracted, and the power spectrum of the enhanced speech is reconstructed using the estimated envelope of the power spectrum. Finally, the phase spectrum is estimated from the power spectrum and the enhanced speech is generated. Objective evaluations are carried out to evaluate the three proposed methods. As the results of the experiments, the effectiveness of the proposed method was confirmed.

Keywords:

Optical laser microphone, Laser Doppler vibrometer, Speech enhancement, Handling of unknown object, Speech acquisition in distance, Deep neural network

目 次

図 目 次

表 目 次

第1章　序論

1.1. 研究背景と目的

マイクロホンは日常生活で最も一般的な収音デバイスである．従来のマイクロホンは，空気を媒介して振動膜に到達した音波を電気信号に変換することで音を収録する．音は空気中を伝播する際に，距離の増加に伴いパワーの減衰が非常に大きいため，遠方にて発生した音を収録することは非常に困難である．遠距離の音を収録するために，パラボラマイクやショットガンマイク [1] が開発されている．これらのマイクロホンは，マイクロホンの形状と振動板の位置を制御することで，従来のマイクロホンと比べて，より遠方の音を観測することが可能となる．しかし，非常に遠方で発生した音を収録することは原理的に困難であり，収録する際に，マイクロホン周辺の雑音も同時に混入する．そのため，騒がしい環境での使用においては性能補償が難しい．この問題を解決するために，光レーザマイクロホンと呼ばれるレーザ光を用いた音響計測システムが提案された [2, 3]．光レーザマイクロホンは光のドップラ効果を利用して，参照光と反射光との位相差から被照射物体の振動を計測するデバイスでる．光レーザマイクロホンは音声収録だけでなく，建物や臓器などの振動も計測可能であるため，産業および医療領域でもよく使用されている [4, 5, 6]．図 1.1 に従来のマイクロホン，パラボラマイクロホン及び光レーザマイクロホンの各々を用いて遠距離から収録した場合のイメージ図を示す．光レーザマイクロホンは，従来のマイクロホンとは異なり，音源の近くにある物体のみにレーザ光を照射するため，マイクロホン周辺の雑音に影響を受けない．また，レーザ光は指向性が強く，距離が増加してもパワーの減衰が小さいため，非常に遠方の音も観測できる．以上の特徴を持つ光レーザマイクロホンは遠距離収録において有用となる一方で，その計測原理により発生した歪みは，収録音声の音質に強く影響を与える．図 1.2 に光

レーザマイクロホンにより収録した音声のスペクトログラムを示す．図 1.2 により，被照射物体表面の反射率の影響で混入した雑音が確認できる．また，被照射物体の振動特性により，振幅応答の小さな周波数帯域 (特に高周波数帯域) において，音声成分が欠落するという問題が存在する．これらの要因により，光レーザマイクロホンによる観測音声をそのまま音声認識システムにり認識することが困難となる．その結果，高音質に音声を取得するために，光レーザマイクロホンの観測音声に対して音声強調処理を行う必要がある．

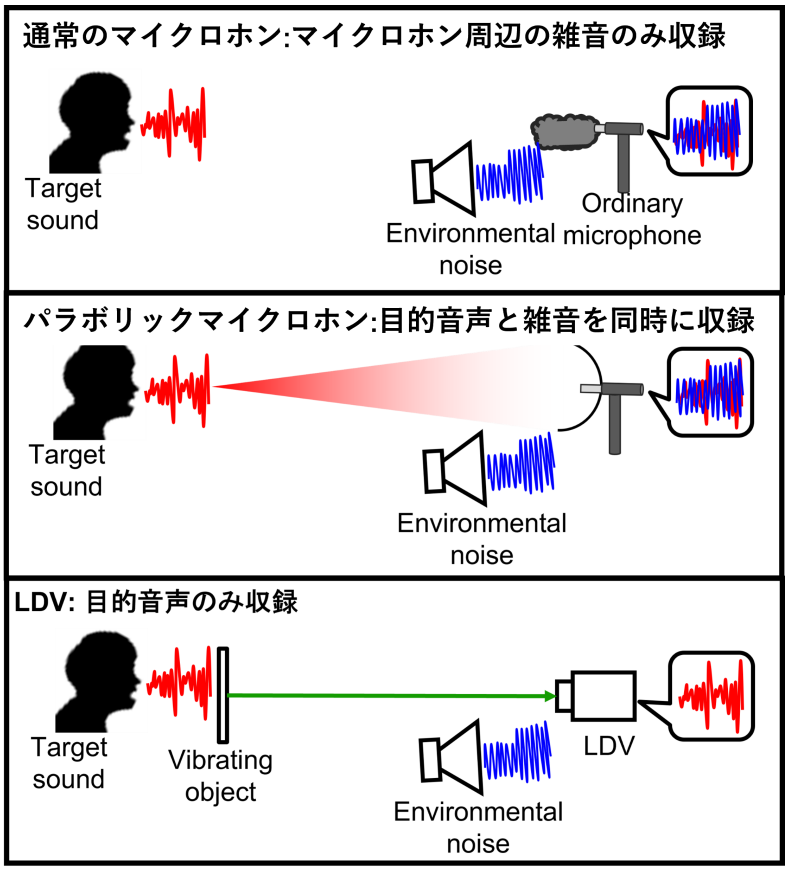
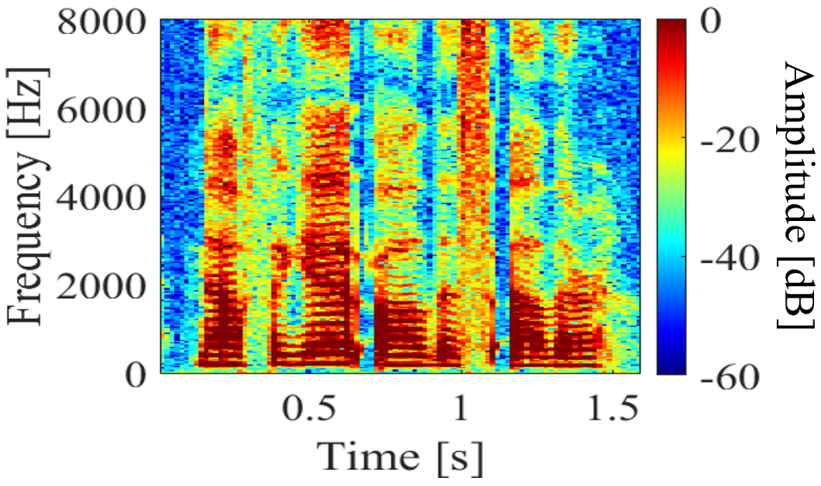
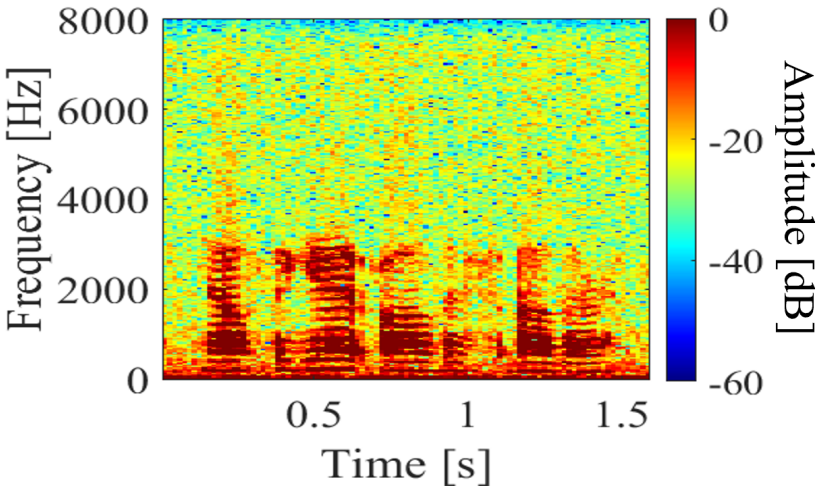


図 1.1 光レーザマイクロホン (LDV) の特徴



(a) クリーン音声



(b) 観測音声

図 1.2 光レーザマイクロホンの観測音声の特徴

これまでに，統計モデルやディープニューラルネットワーク (Deep neural network: DNN) を含む様々な音声強調手法が提案されてきた [7, 8, 9, 10]．しかし，従来の音声強調手法は光レーザマイクロホンの観測音声のような，多様な歪みを含む音声に対して，強調結果が劣化するだけではなく，その強調精度はトレーニングデータに依存するため，音声を観測する時の被照射物体がトレーニングデータを取得した時の被照射物体と異なる (被照射物体が未知) 場合，従来手法では，十分な性能の実現が難しい．一方，光レーザマイクロホンのための音声強調方法も提案されてきた [11, 12, 13, 14]．例えば，W. Li らの提案手法 [12] では，まず，バンドパスフィルタ (Bandpass filter: BPF) を利用し，観測音声のパワースペクトルにおいて，音声情報のない (振幅応答により完全に欠落した) 周波数帯域の成分を除去する．次にウィナーフィルターを用いて，観測音声の雑音を低減する方法を提案した．また，R. Peng らは [13]，2 つの LDV を用いて同時収録し，取得した 2 チャネル信号に coherent-to-diffuse ratio(CDR) と multi-channel linear prediction (MCLP) を適用することで雑音を低減する手法を提案した．しかし，これら 2 つの従来方法では，取得した音声に対して雑音抑圧処理を行うだけであり，欠落した音声成分は復元されず，未知の被照射物体の場合，十分な性能を達成できない．

そこで，本論文では，雑音除去と音声成分復元に対応可能な光レーザマイクロホンのための 3 つの音声強調手法を提案する．まず，被照射物体が既知の場合において，短時間フーリエ変換 (Short-time Fourier transform: STFT) と時間波形それぞれに基づく 2 つの音声強調手法を提案する．STFT に基づく手法は DNN を用いて，収録音声のスペクトルに着目し，振幅スペクトルと位相スペクトルそれぞれの劣化に対して，2 つのネットワークを使用し，二段階で処理することで，振幅と位相スペクトルぞれぞれを復元し，強調音声を得る．時間波形を用いた手法では，観測音声波形において，低域成分と高域成分の劣化の違いに着目し，2 つのネットワークで二段階処理により観測音声を強調する．まず，低域の雑音を除去する．その後，強調した低域成分を利用して高域成分を推定することで強調音声を推定する．さらに，未知の被照射物体における音声強調手法も提案する．提案手法は振動特性に影響されない特徴量を利用する．まず，観測音声から，ピッチとパワースペクトルの包絡情報を抽出する．そして，DNN を利用して振幅応答により劣化したスペクトル包絡

を復元する．次に，復元されたスペクトル包絡とピッチを用いて強調音声のパワースペクトルを再構成する．最後に，復元されたパワースペクトルを利用して位相スペクトルを生成し，強調音声を合成する．

1.2. 本論文の構成

本論文では，以下の 5 章から構成される．2 章では，光レーザマイクロホンを用いた音声収録の原理及び関連研究について述べる．そして 3 章では，被照射物体が既知の場合において，提案した 2 つの音声強調手法及び評価結果について述べる．4 章では，被照射物体が未知の場合の提案手法及び評価結果について述べる．最後に 5 章で結論と今後の課題について述べる．

第2章　光レーザマイクロホンを用いた音声強調システムの基礎

2.1. はじめに

光マイクロホンはレーザ光を利用し，音により発生した振動を計測することで，その振動から音響信号を得るシステムであり，大きく分けてカプセル型光マイクロホン [15] とレーザ型光マイクロホン [3, 16, 17] の 2 種類が開発されている．カプセル型は振動板とレーザ部を内包しており，レーザを用いて振動板に到達する音波を計測する．カプセル型光マイクロホンは一般的なマイクロホンと同様に，近距離で発生した音を計測するために使用される．特に磁気素材を使用していないため，高磁場環境でも使用可能である．レーザ型光マイクロホンは，振動板の代わりに，音波によって振動している音源付近の物体の表面にレーザ光を照射し，レーザドップラ振動計 (Laser Doppler vibrometer: LDV) を用いて，音波により振動している物体表面の振動を計測することで音響信号を取得する．そのため，マイクロホン周辺に存在する騒音源の影響を受けずに，目的音のみ計測することが可能である．本論文では，レーザ型光マイクロホンを研究対象として，レーザドップラ振動計を使用し，以下 LDV と表記する．

本章では，LDV のための音声強調システムの基礎について述べる．2.2 節では，LDV の収録原理について説明する．2.3 節では，LDV の観測音声の特徴と問題点について述べる．2.4 節では，本論文で使用した音声データの収録方法と収録条件について説明する．2.5 節では，従来のニューラルネットワークを用いた音声強調手法とそれらを LDV の収録音声に適用できない理由を説明する．最後に 2.6 節では本論文で使用した音質評価基準について説明する．

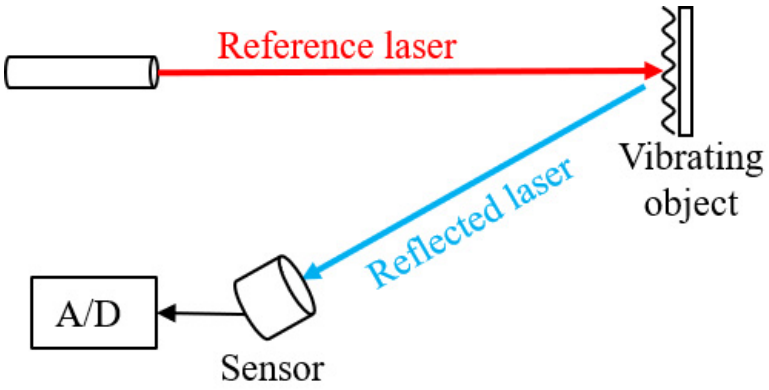


図 2.1　フォトダイオードを用いた音響計測

2.2. LDV を用いた音声収録の原理

話者が発話すると，その音声は空気中を伝搬し，空気の振動が音源付近の物体に伝わる．レーザ光を振動物体の表面に照射すると，その振動により，レーザ光の行路長が変化し，反射光の強度や位相も変化する．ここで，まず，フォトダイオードを利用した反射光の強度変化に着目した音響計測について述べる．フォトダイオードを利用した音響計測の概略図を図 2.1 に示す．この手法では，物体の振動によってレーザ光が振幅変調される．変調波 $S_r(t)$ を式 (2.1) に示す．これにより，物体振動から音響信号を取得することができる．

$$S_r(t) = A_c(1 + s(t))\cos(2\pi F_c t + \phi_c), \tag{2.1}$$

ここで，A_c，F_c，ϕ_c，t，$s(t)$ はそれぞれレーザ光の振幅，周波数，位相，時間指標と物体の振動を表す．

次に，反射光の位相に着目し，光のドップラ効果を利用した LDV について述べる．LDV による計測の概略図を図 2.2 に示す．ドップラ効果は，レーザ光が物体表面で反射する際に生じる．このとき，干渉計を利用することで入射光と反射光の周波数差 $f_{\mathrm{D}}(t)$ を検出できる．周波数差 $f_{\mathrm{D}}(t)$ と振動速度の関係は式 (2.2) により表される．

$$f_{\mathrm{D}}(t) = \frac{2v(t)}{\lambda_0} = \frac{2}{\lambda_0} \cdot \frac{\mathrm{d}L_1(t)}{\mathrm{d}t}, \tag{2.2}$$

ここで，λ_0 は参照光の波長，$v(t)$ は振動物体の速度，$L_1(t)$ は照射地点から LDV 内の検知器に到達するまでの光路長である．反射光と参照光を重ね合わせると，干渉縞が発生する．物体が振動している時，反射光の位相が連続的変化するため，干渉縞が移動する．検知器で観測されるレーザ光の強度 I_t は式 (2.3) により表される．

$$I(t) = I_1 + I_2 + 2\sqrt{I_1 I_2 \cos(\rho}(t)), \tag{2.3}$$

$$\rho(t) = 2\pi \frac{(L_2 - L_1(t))(\Delta f - f_D(t))}{c}, \tag{2.4}$$

ここで，I_1 は反射光の強度，I_2 は参照光の強度，L_2 は参照光の光路長，Δf は周波数シフタにより与えられるシフト量，c は光速を表す．以上 3 つの式により，物体の振動速度を計測できる．周波数シフタを利用する理由は，干渉縞の間隔のみを利用する場合，振動速度の符号を判別できない．よって，周波数シフタにより参照光の周波数を変更することで，振動方向も含めた振幅を計測することができる．LDV では，計測された振動速度が電気信号として出力される．この電気信号を A/D コンバータを用いてサンプリングすることで，デジタルの音響信号として取得できる．

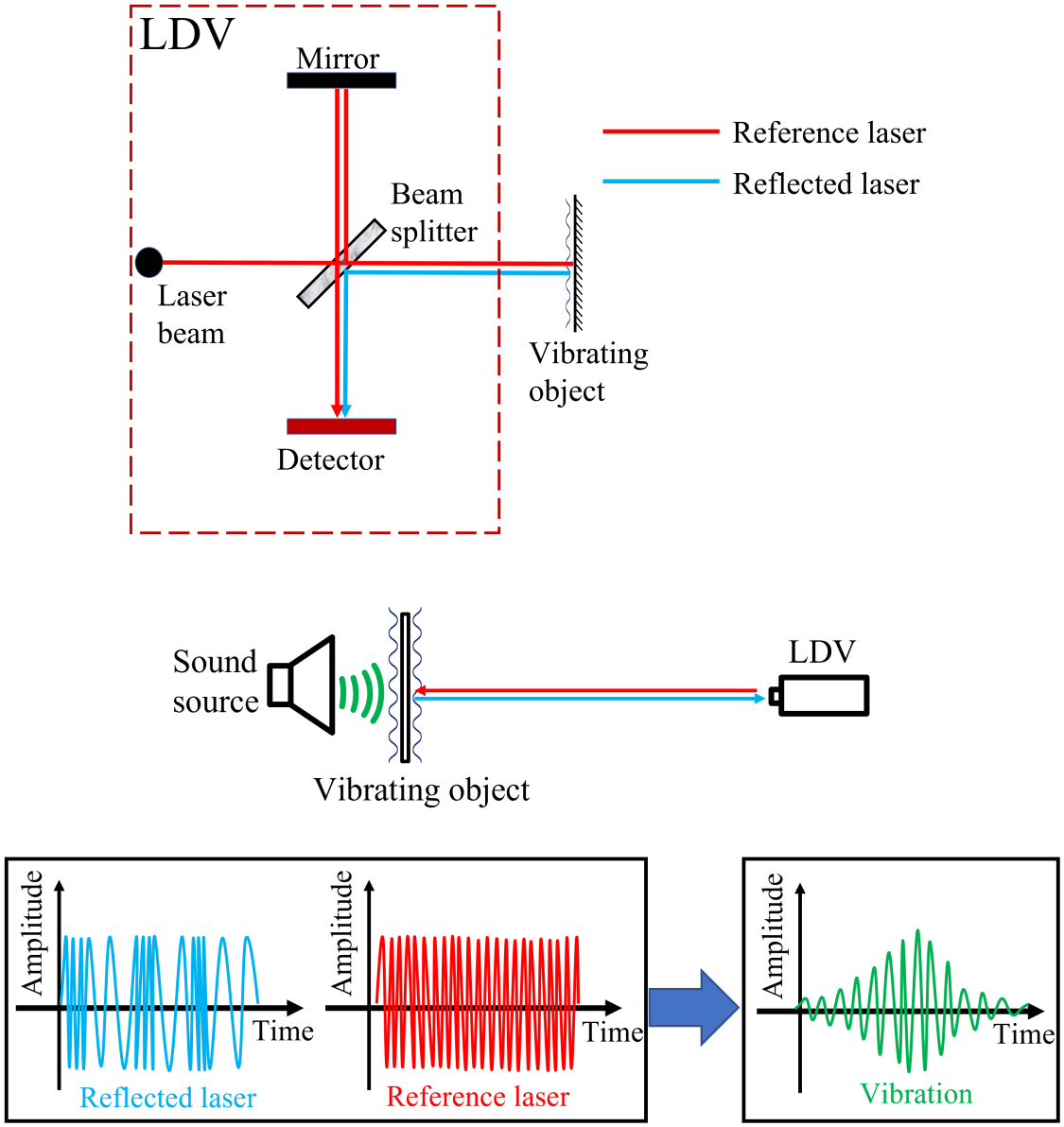
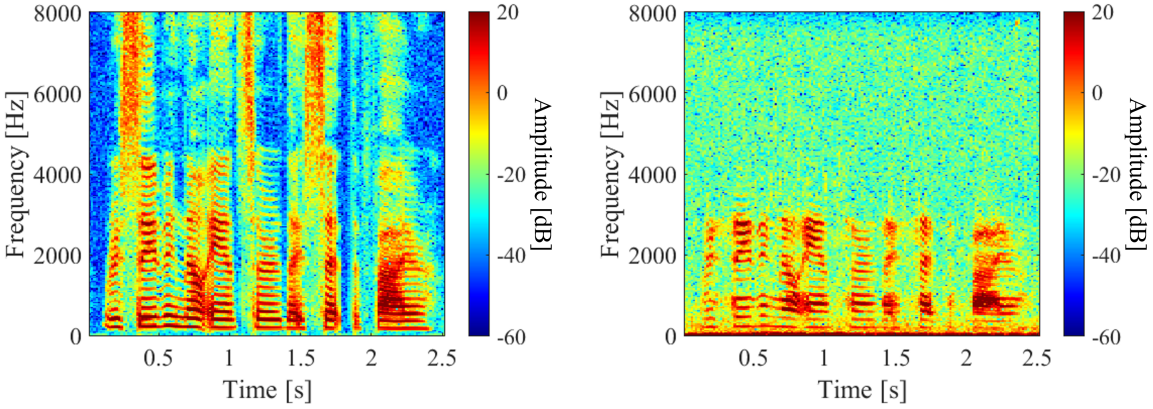


図 2.2　LDV を用いた音響計測

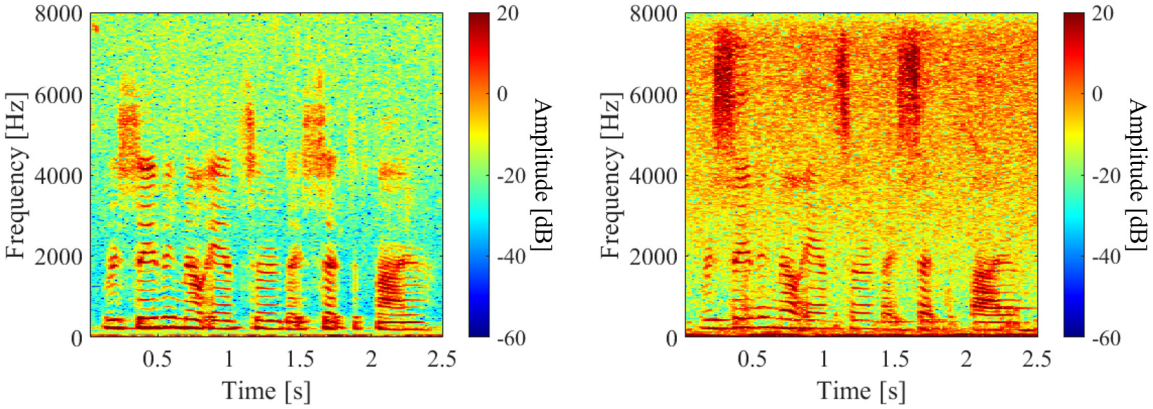
2.3. LDV による観測音声の特徴

LDV は，物体表面に照射した参照光と反射光の位相差を計測することで物体の振動速度を測定するデバイスである [2]，レーザ光は高い直進性を持つため，遠方で発生する振動を計測することが可能である．また，被照射物体周囲に発生した音のみ計測するため，LDV 周囲の雑音の影響を受けない．以上の特徴を持つため，遠隔発話を受音する場合において，LDV が有用であるといえる．しかし，振動物体を介して音響情報を取得するため，観測音声は被照射物体の形状と振動特性に依存する．例えば，表面の粗い物体が測定方向以外に振動したとき，$f_{\mathrm{D}}(t)$ は $v(t)$ だけでなく，物体表面の凹凸により変動する．また，物体から反射したレーザ光の光量が減少することにより，検知器において観測されるレーザ光の強度は減少し，定常雑音が混入する．

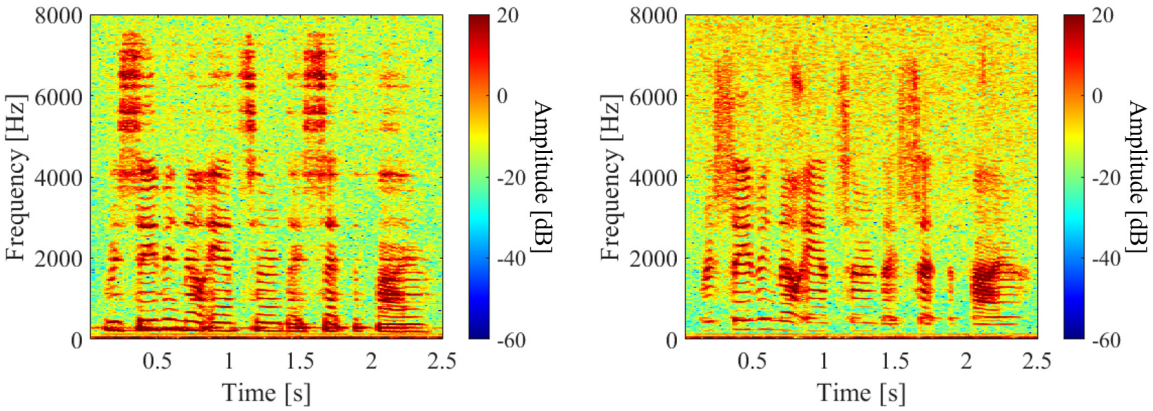
図 2.3 では，さまざまな照射対象物によって観測された音声のスペクトログラムを示している． 図 2.3 から，違うレベルの雑音と音声成分の欠落が確認できる．被照射物が未知の場合において，日常生活でよく見かける印刷用紙，アルミ板，プラスチック板，ダンボールを観察対象として選択した．各被照射物体を用いて音声を収録し，モデルを学習する．そして，その他の物体を未知物体として，学習したモデルを適用する．この 4 つの振動物体のうち，印刷用紙とアルミシートの表面は滑らかで，プラスチック板やダンボールの表面は粗い．



(a) クリーン音声　(b) ペットボトル



(c) 印刷用紙　(d) ダンボール



(e) アルミ板　(f) プラスチック板

図 2.3　様々な被照射物体による観測音声のスペクトログラム

2.4. LDV による音声データの収録

本論文で提案した 3 つの手法におけるネットワークのトレーニングで利用する音声データに対して，LDV を用いて収録した．収録条件を表 2.1，収録機材は表 2.2，機材配置図を図 2.4，収録風景を図 2.5 に示す．計測対象として，容積 0.5 リットルの空のペットボトル及び寸法は約 100 mm × 100 mm のアルミ板，印刷用紙，プラスチック板，ダンボール板を使用した．図 2.3(c),(d),(e),(f) より，異なる被照射物体にて収録を行う場合，音声成分の欠落帯域も異なる．また，物体表面の反射率も異なるため，雑音レベルの違いも確認できる．収録内容として，TIMIT Acoustic Phonetic[18] にある 4620 文に対して，各被照射物体毎に 2 回の収録を行い，それぞれ 9240 文を用いて実験データを作成した．その中の 9000 文 (約 4 時間) のデータをネットワークのトレーニングに使用し，240 文 (約 12 分) のデータを評価実験で使用した．

表 2.1 実験条件

Environment	Sound-proof room
Ambient noise level	20.8 dB
Sampling frequency	16 kHz
Quantization bit rate	16 bits
Temperature	25.9 °C
Humidity	19.4 %
Data	TIMIT Acoustic Phonetic Continuous Speech Corpus[18] 9,000 samples(4 hours) for training 240 samples(12 minutes) for validation

表 2.2 実験機材

LDV	Polytec NLV-2500-5
Loudspeaker	FOSTEX, FE83En
Loudspeaker amplifier	YAMAHA, P4050
Audio interface	ROLAND, UA-1010

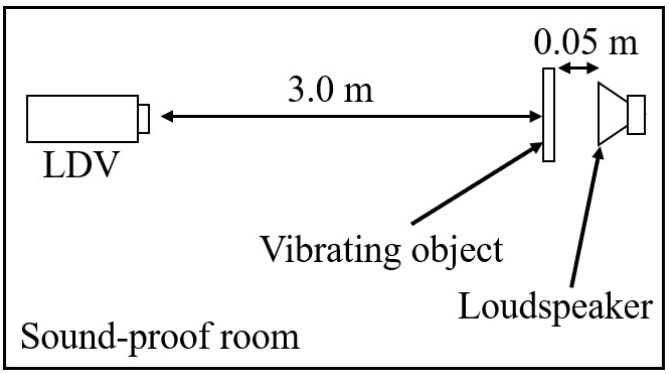


図 2.4　実験配置図

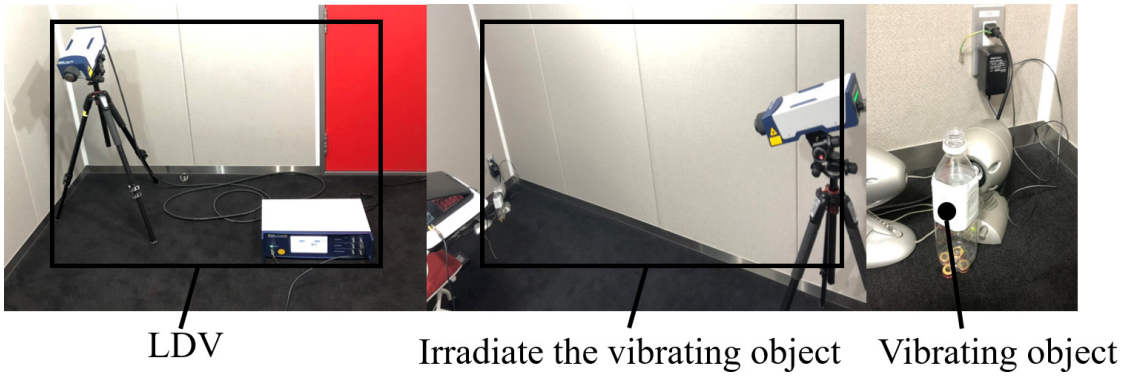


図 2.5　実験風景

2.5. ニューラルネットワークを用いた音声強調の従来手法

近年，DNN が新興分野となり，信号処理分野でも幅広く利用され，音声生成 [19]，声質変換 [20, 21]，音声強調 [22, 23, 24, 25, 26] などの分野で有効性が示されている．DNN を用いた音声強調は，STFT (Short-time Fourier transform) に基づく手法と時間波形に基づく手法に大別される．

STFT に基づいた手法は，収録音声およびクリーン音声の振幅スペクトル間の関係を学習する．そして，音声波形を再構築する際に，推定された振幅スペクトルと収録音声の位相スペクトルを用いて音声波形を復元する．しかし，周波数領域処理における音声強調の精度は，振幅スペクトルだけでなく位相スペクトルにも依存する．特に，LDV による収録音声において，各周波数における位相遅延が異なるため，音声波形を復元する際に収録音声の位相スペクトルを用いると，音声強調の精度が低下する傾向がある．そのため，位相復元手法も盛んに研究されており，その一つとして Griffin-Lim 法 [27, 28] が挙げられる．また，近年では，DNN を利用した位相推定手法 [29, 30] も提案されている．例えば，S. Takamichi らは，振幅スペクトルを用いて位相スペクトルを推定し，それを Griffin-Lim 法の初期値とする手法を提案している [29]．その他にも，複素スペクトルを用いて振幅と位相を同時に推定する手法 [30] も存在する．

一方，時間波形は振幅と位相の両方を含むため，時間波形を特徴量とすることで，位相に起因する歪みの発生を防ぐことができる．しかし，時間波形は複雑な構造を持つため，高精度な音声強調を実現するためには，多くの入力系列サンプルが必要となる．ネットワークでより多くのサンプルを処理させるため，A.V.D. Oord らは Dilated convolution 構造を有するネットワーク WaveNet[31] を提案した．WaveNet に基づいて，D. Rethage らと Y. Gu らは雑音除去手法 [32] と帯域拡大手法 [33] をそれぞれ提案した．また，長・短期記憶 (Long short-term memory: LSTM) を用いた再帰型ニューラルネットワーク (Recurrent neural network: RNN) に基づいて，Y.Gu らは音声の帯域拡大手法 [34, 35] を提案した．しかし，LDV による収録音声は，高域成分の欠落だけでなく，雑音や残響も混在するため，それらの手法を直接適用しても，高い復元精度を得ることは難しい．

2.5.1 STFT に基づくニューラルネットワーク

STFT に基づいたニューラルネットワークの入出力には，劣化音声とクリーン音声の対数パワースペクトル (Log-power spectrum: LPS) などの特徴量が使用されており，平均二乗誤差 (Mean square error: MSE) を最小化するように入出力間のマッピング関係を学習し，ネットワークの最適化が行われている．音声の波形を再構築する際に，ネットワークにより推定したパワースペクトルと入力位相を用いて逆離散フーリエ変換により音声波形を復元できるが，周波数領域における音声強調精度は，振幅だけでなく，位相にも依存する．特に，LDV による観測音声のような多種類の歪みが混在する信号において，音声波形を復元する際に劣化音声の位相スペクトルを用いると，復元後の音質が不十分となる．ゆえに LDV の観測音声に対して，位相の制御も必要となる．位相制御の代表的な方法として，Griffin-Lim 法 [27] が挙げられる．また近年では，DNN を利用した位相推定手法も提案された．複素スペクトルを用いて，振幅と位相を同時に強調する手法 [30]，または振幅スペクトルから位相を推定する [36] 手法もある．

2.5.2 時間波形に基づいたニューラルネットワーク

時間波形は振幅と位相情報の両方を含むため，時間波形を特徴量とすると，推定位相により発生したアーティファクトが発生しない．しかし，時間領域の特徴量は複雑な構造を持つため，高い推定精度を得るためには，多くの入力サンプル系列が必要となる．ここで，時間波形に基づくニューラルネットワークとして，WaveNet[31] について説明する．WaveNet は Dilated convolution 構造を用いた複数の畳み込み層のある残差ブロックから構成される．Dilated convolution の構造を図 2.6 に示す． 図 2.6 より，Dilated convolution では，フィルタと畳み込み処理する対象の間隔を空けるため，より大きな受容野 (receptive field： RF) を取得できる．ネットワーク層数の増加に伴い，ネットワーク RF は指数的に増大する．よって，畳み込みニューラルネットワーク (Convolutional neural network： CNN) を用いて時系列の予測も可能となる．

WaveNet を用いた音声強調は劣化音声を μ-law[37] により 8 ビットで量子化した結

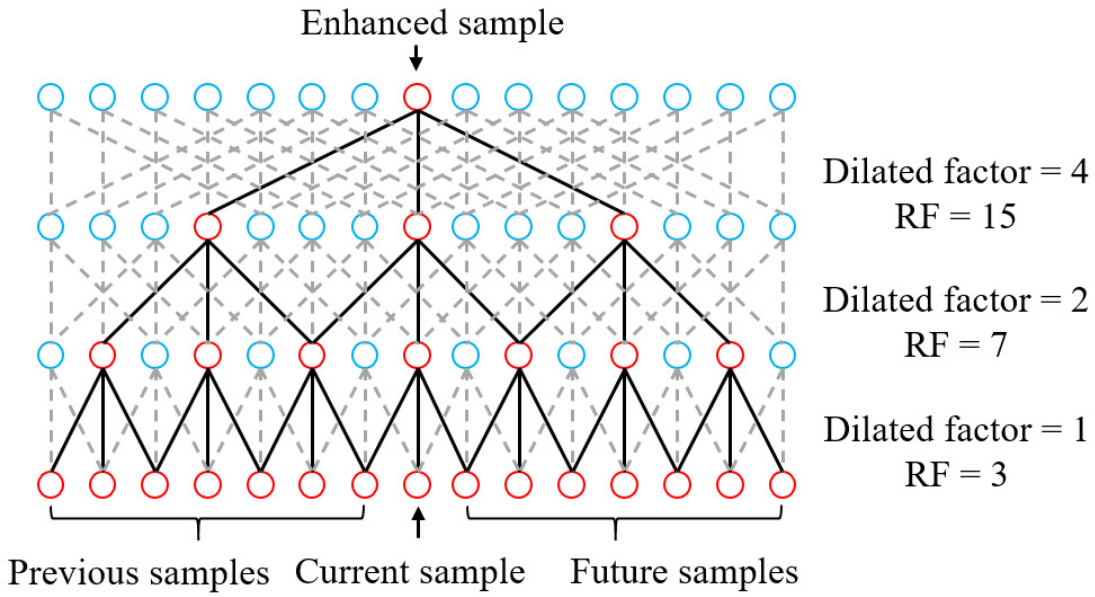


図 2.6 Dilated convolution の構造

果を入力する．出力する際に，活性化関数 softmax を用いて，式 (2.5) により，劣化音声 $\boldsymbol{x} = [x(0), x(1), ..., x(N-1)]$ の条件下の出力音声 $\hat{\boldsymbol{y}} = [\hat{y}(0), \hat{y}(1), ..., \hat{y}(N-1)]$ の離散確率分布を計算する．

$$p\left(\hat{\boldsymbol{y}}|\boldsymbol{x}\right) = \prod_{n=0}^{N-1} p\left(\hat{y}\left(n\right)|x\left(n-\frac{R}{2}\right), ..., x\left(n+\frac{R}{2}\right)\right), \tag{2.5}$$

ここで，$R+1$ は RF のサイズである．入出力間の交差エントロピー (Cross-entropy: CE) を最小になるようにネットワークをトレーニングする．しかし，LDV による収録音声は多種類の歪みが混入するため，推定精度が低下する．

2.6. 音声強調の評価基準

本論文では，提案手法の有効性を確認するために，LDV による収録音声を用いた音声強調実験を実施する．本実験では，評価指標として，Wideband perceptual

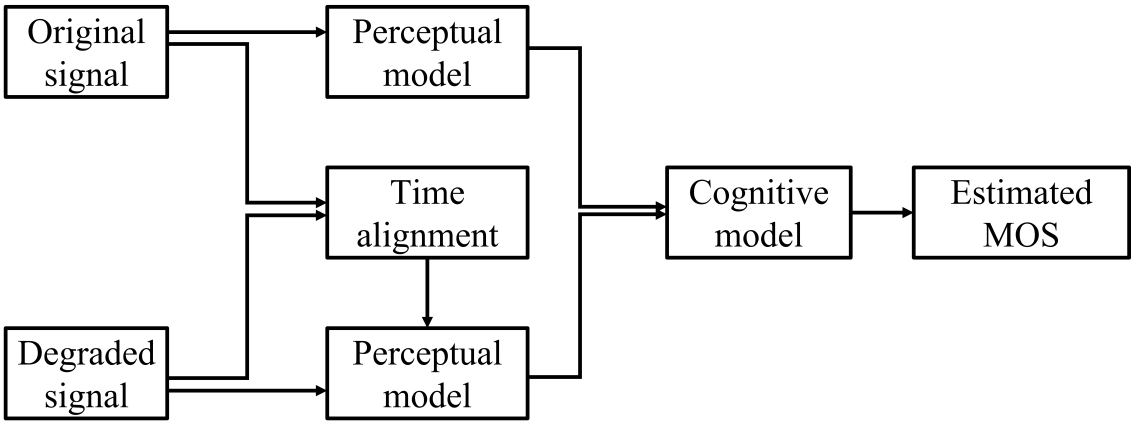


図 2.7 PESQ スコアの算出方法

evaluation of speech quality (PESQ)，Log-spectal distance (LSD)，Short-time objective intelligibility (STOI) の 3 指標を用いて，音質，劣化度，明瞭度を比較する．なお，PESQ は高いほど音質が高く，LSD は小さいほど音声の劣化が少ない．また，STOI は高いほど明瞭度が高いことを表す．次に，PESQ，LSD，STOI それぞれについて説明する．

- PESQ

PESQ[38] は，評価信号の主観的な品質を客観的に推定可能な評価指標である.PESQ スコアの計測アルゴリズムを図 2.7 に示す．はじめに，クリーン音声と劣化音声を知覚モデルを用いてセルと呼ばれる時間・バークスペクトル領域に射影する．そして，セル間の歪みから認知モデルを用いて主観 Mean opinion score (MOS) の推定値 (PESQ 値) を計測する．

- LSD

対数スペクトル距離 (LSD)[39] は，対数スペクトル歪みまたは二乗平均平方根対数スペクトル距離とも呼ばれ，2 つのスペクトル間の距離の尺度である．離散フーリエ変換を適用した信号がスペクトル領域に変換される場合，1 つのフレームに対

して，LSD は以下の式により定義される．

$$\mathrm{LSD} = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(10\log_{10}\frac{P(k)}{\hat{P}(k)}\right)^2}, \tag{2.6}$$

ここで，$\boldsymbol{P}$，$\hat{\boldsymbol{P}}$ はそれぞれ 2 つの信号のパワースペクトルであり，k は周波数インデクスである．

- STOI

短時間客観的明瞭度 (STOI)[40] は，音声明瞭度を測定するための指標である．音声信号中の単語は理解できる場合と理解できない場合の 2 通りしかないことから，明瞭度は 2 値であると考えられるため，STOI の値の範囲は 0 から 1 までの範囲で数値化される．値が 1 の場合は完全に理解できたことを意味する．STOI はクリーン音声と処理された音声を入力し，クリーン音声と処理と処理後の信号のスペクトルエネルギーの短期相関を計算し，低周波 (150 Hz〜 約 3800 Hz) のみを計算する．算出下平均相関係数を STOI の値とする．

2.7. まとめ

本章では，LDV のための音声強調システムの基礎について述べた．2.2 節では，LDV の収録原理について説明した．2.3 節では，LDV の観測音声の特徴と問題点について述べた．2.4 節では，本論文で使用した音声データの収録方法と収録条件について説明した．2.5 節では，従来のニューラルネットワークを用いた音声強調手法とそれらを LDV の収録音声に適用できない理由を説明した．最後に 2.6 節では本論文で使用した音質評価基準について説明した．

第3章　被照射物体が既知(特定)の場合における音声強調

3.1. はじめに

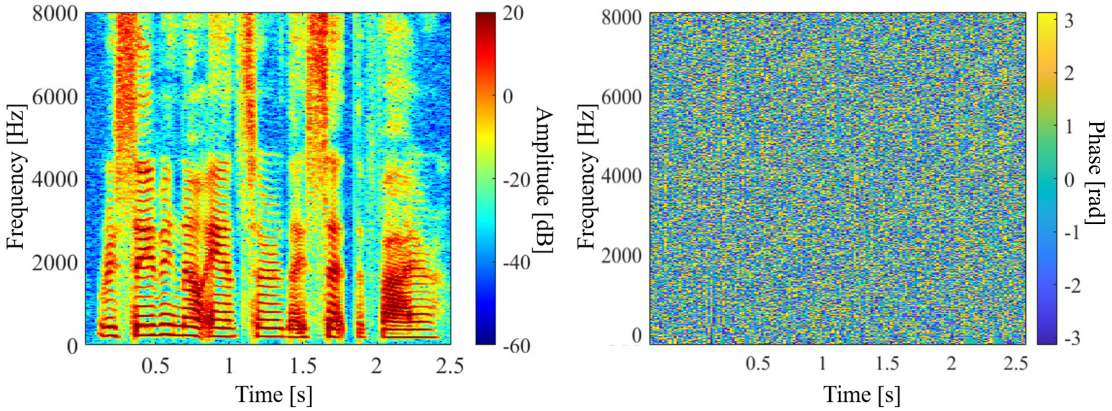
遠距離の音声収録は常に注目の研究テーマである．その中，LDV は非常に遠いところに発生したの音声を観測でき，マイクロホン周辺の雑音の影響を受けないという利点を持ち，高く評価されている．しかしながら，LDV により，観測された音声には雑音が混入し，被照射物体の振幅応答の小さな周波数帯域において，音声成分が欠落する．よって，LDV の観測音声に対して，音声強調処理が必要となる．LDV の観測音声には多様な劣化があるため，従来の音声強調方法は直接に適用することが困難である．したがって，この章では，被照射物体が既知 (特定) の場合において，複数の DNN を使用し，LDV の観測音声の多様な音声劣化に各々対応するという方法を提案する．提案手法は STFT と時間波形それぞれに基づき構成している．STFT に基づく方法では，まず従来方法を利用して，観測音声のパワースペクトル上の雑音を抑制し，欠落した音声成分を復元する．次に，処理後のパワースペクトルを利用して，観測音声とクリーン音声の間の位相差を計算する．時間波形に基づく方法では、まず低域部分の雑音を除去する．次に強調された低域信号を利用して高域信号を推定する．

本章では，3.2 節で，STFT に基づく音声強調手法，3.3 節で，時間波形に基づく音声強調手法それぞれについて説明する．そして，3.4 節で，2 つの手法で使用したネットワークの構造と学習条件を説明し，評価実験により 2 つの提案手法に対して，二段階処理の必要性を述べる．最後に，3.5 節で，2 つの提案手法を用いて観測音声を強調し，その結果を評価する．

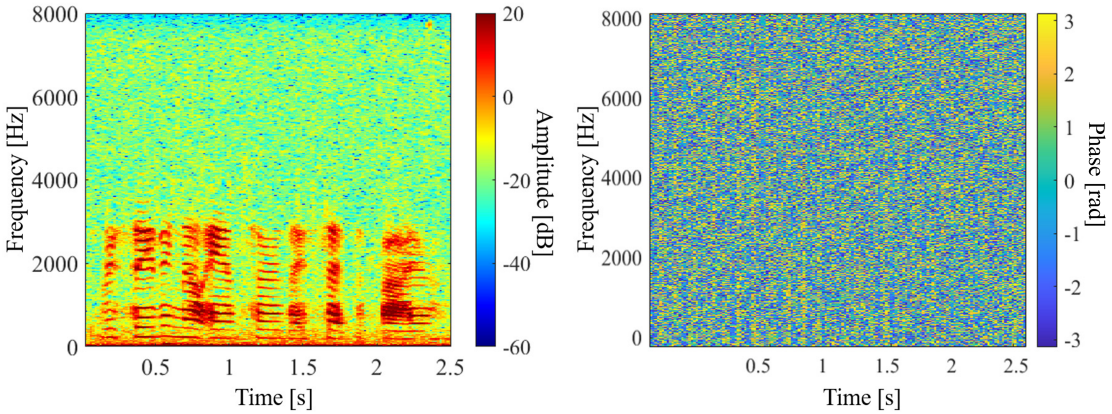
3.2. 提案法 1:STFT に基づく音声強調

前章で述べたように，従来手法では，音声波形を復元する際に収録音声の位相スペクトルを用いるため，音声強調の精度が低い傾向にある．ここでは，STFT に基づき，振幅と位相処理を有する二段階音声強調手法を提案する．音源の音波は空気中を伝播し，観測物体を振動させるため，観測物体の振動は音源により発生した減衰のある強制振動と考えられる．図 3.1 にペットボトルを観測物体とする場合の収録結果 (a)，(b)，観測物体の振幅特性 (c) 及び収録音声とクリーン音声の位相差 (d) を示す．(c) の振幅特性より，約 3kHz 以上の周波数帯域における振幅の応答は減少する傾向がある．(a) と (b) の左図より，音声成分の振幅は雑音より十分小さい場合，特に 4–8 kHz の帯域にあるランダム性の強い音声成分は観測不能となる．よって，振幅復元処理では，再帰型ニューラルネットワーク (Recurrent Neural Network: RNN) を利用して，収録音声とクリーン音声の振幅スペクトルの関係を学習し，振幅スペクトルを復元する．

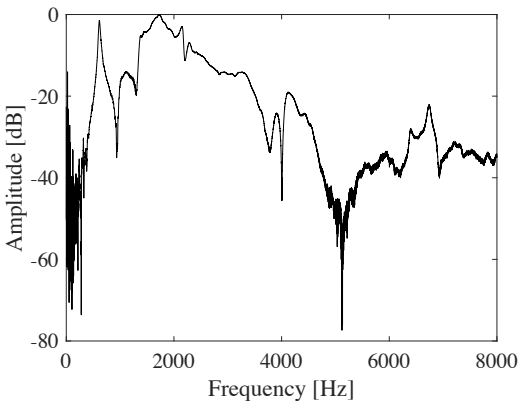
一方，図 3.1(a)，(b) の右図に示すように，位相スペクトルは複雑な構造を持ち，収録音声とクリーン音声の位相スペクトルの関係を直接学習することは困難である．ここで，図 3.1(d) より，収録音声とクリーン音声の位相差は収録音声の振幅に強く関連することが確認できる．収録音声の振幅スペクトルにおいて，高域成分 (4–8 kHz) は欠落しているが，その高域成分は概ね無声音成分であり，その構造はランダムが高いため，位相誤差が発生しても聴覚的には知覚できない．よって，位相復元処理では，畳み込みニューラル ネットワーク (Convolutional neural network: CNN) を利用して，収録音声の振幅スペクトルにより，収録音声とクリーン音声の低域成分 (0–4 kHz) の位相差のみを推定し，観測位相スペクトルと加算することで復元位相スペクトルを得る．



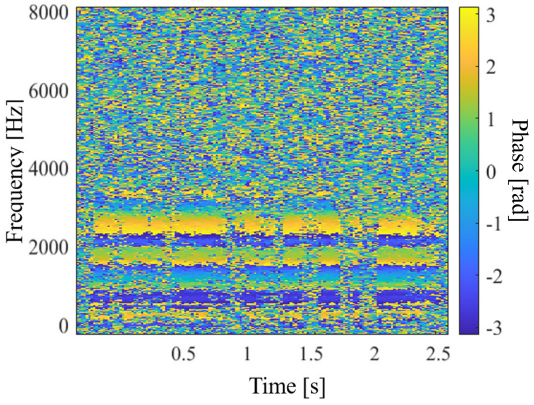
(a) クリーン音声のパワースペクトル (左) と位相スペクトル (右)



(b) 収録音声のパワースペクトル (左) と位相スペクトル (右)



(c) 観測物体の振幅特性



(d) 位相差

図 3.1 クリーン音声と収録音声の振幅スペクトルと位相スペクトル及び位相差

提案手法 1 の処理手順を図 3.2 に示す．学習段階では，まず，クリーン音声 $\boldsymbol{x}$ と収録音声 $\boldsymbol{y}$ をフーリエ変換により，それぞれの対数パワースペクトル $|\boldsymbol{X}^{\mathrm{LPS}}|$，$|\boldsymbol{Y}^{\mathrm{LPS}}|$ と位相スペクトル $\boldsymbol{\phi}_x$，$\boldsymbol{\phi}_y$ を抽出する．振幅スペクトル復元用 DNN において，ネットワークの入出力はそれぞれ $|\boldsymbol{Y}^{\mathrm{LPS}}|$ と $|\boldsymbol{X}^{\mathrm{LPS}}|$ であり，平均二乗誤差 (Mean Squared Error: MSE) を最小になるように最適化する．振幅復元用 DNN の構造は図 3.3 に示すように，2 層の LSTM 層と 3 層の全結合層により構成される．活性化関数は ReLU を利用する．位相スペクトル復元用 DNN の入力は $|\boldsymbol{Y}^{\mathrm{LPS}}|$ であり，出力は $\boldsymbol{\phi}_x$ と $\boldsymbol{\phi}_y$ から得られた位相差 $\boldsymbol{\phi}^{\mathrm{PD}}$ である．位相スペクトルは 2π の周期を持つため，学習の際に利用した損失関数は式 (3.1) に定義されるように，推定周波数帯域 0−4 kHz における，フレーム毎の目的位相差と推定位相差のコサイン距離の合計である．

$$\mathrm{Loss}_{\mathrm{PD}}(m) = \sum_{k=0}^{K}(1 - \cos(\boldsymbol{\phi}^{\mathrm{PD}}(k,m) - \hat{\boldsymbol{\phi}}^{\mathrm{PD}}(k,m))), \tag{3.1}$$

ここで，k，K，m，$\boldsymbol{\phi}^{\mathrm{PD}}$，$\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$ はそれぞれ，周波数インデクス，フーリエ点数，フレームインデクス，目的位相差，推定位相差である．図 3.4 に示すように，位相復元用 DNN の構造は 5 層の畳み込み層により構成される．活性化関数は Gated Linear Units (GLU)[41] を利用する．なお，勾配を計算する際に，両ネットワークともに Adaptive moment estimation(Adam)[42] を利用した．

また，音声強調の段階では，まず，収録音声の対数パワースペクトル $|\boldsymbol{X}^{\mathrm{LPS}}|$ を学習した振幅復元用 DNN と位相復元用 DNN に入力し，推定対数パワースペクトル $|\hat{\boldsymbol{X}}^{\mathrm{LPS}}|$ と推定位相差 $\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$ を算出する．復元位相は推定位相差 $\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$ と収録音声の位相 $\boldsymbol{\phi}^{\mathrm{PD}}$ を加算することで得られる．最後に，$|\hat{\boldsymbol{X}}^{\mathrm{LPS}}|$ と $\hat{\boldsymbol{\phi}}^{\mathrm{PD}}$ を用いて，逆フーリエ変換により強調音声を得る．

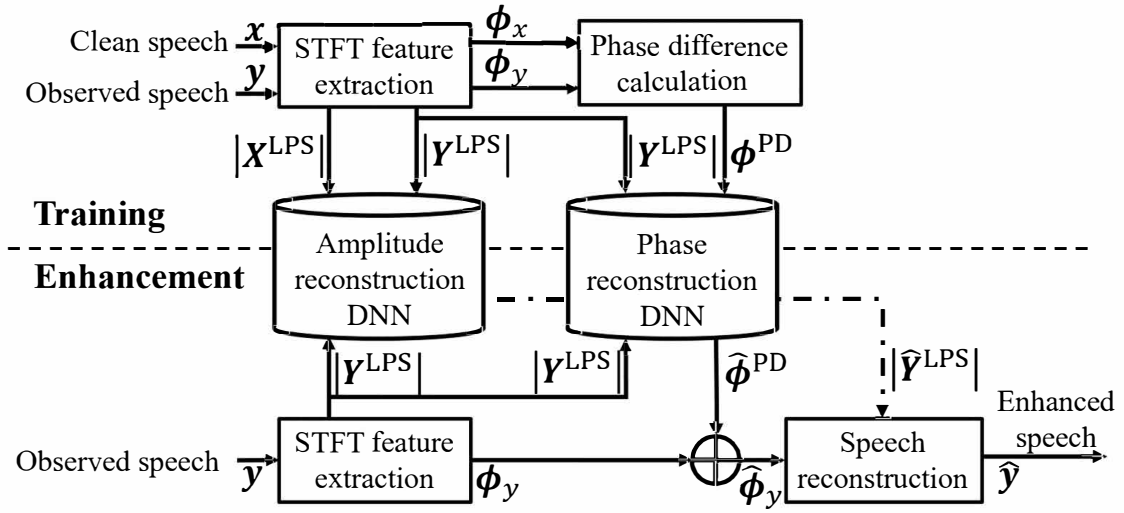


図 3.2 提案手法 1 のブロック図

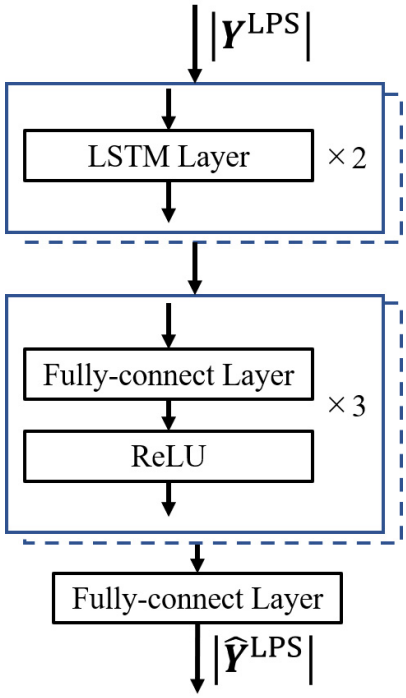


図 3.3　振幅復元用 DNN の構造

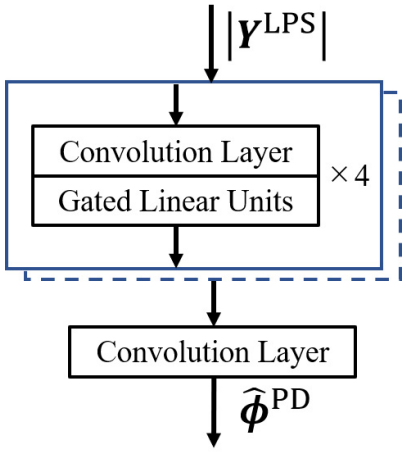


図 3.4　位相復元用 DNN の構造

3.3. 提案法 2: 時間波形に基づいた音声強調

LDV による収録音声には，主に雑音の混入と高域成分の欠落の 2 種類の劣化が存在する．一般的に，高域成分の復元手法は精確な低域成分が必要である．また，図 3.1(a) に示すように，音声の低域と高域は異なる構造を持つため，1 つのネットワークですべての劣化に対処することは困難である．そこで，低域の雑音抑圧および高域成分の復元の二段階処理による LDV の収録音声の音声強調法を提案する．

本手法の処理手順を図 3.5 に示す．学習段階において，収録音声 $\boldsymbol{x}$ の高周波成分には音声の特徴的な構造が存在しないため，ダウンサンプリングにより，$\boldsymbol{x}$ とクリーン音声 $\boldsymbol{y}$ それぞれの高域成分を除去する．その後，ダウンサンプリング後の音声 $\boldsymbol{x}^{\mathrm{NB}}$ と $\boldsymbol{y}^{\mathrm{NB}}$ を雑音抑圧用 DNN に入力し，ネットワークを学習させる．そして，クリーン音声の長さを合わせるように，低域強調音声 $\hat{\boldsymbol{x}}^{\mathrm{NB}}$ をアップサンプリングする．最後に，アップサンプリング後の音声 $\hat{\boldsymbol{x}}^{\mathrm{WB}}$ と $\boldsymbol{y}$ それぞれを μ-law[37] により 8 ビット量子化した結果を用いて，高域復元用 DNN を学習させる．

図 3.6 に示すように，雑音抑圧用 DNN は 8 層の Dilated convolution 構造を有する畳み込み層により構成される．損失関数と活性化関数はそれぞれ MSE と PReLU[43] を利用する．MSE を損失関数とする場合，その勾配は MSE の導関数 $\hat{\boldsymbol{x}}^{\mathrm{NB}} - \boldsymbol{y}^{\mathrm{NB}}$ により与えられる．ここで，音声信号のパワーが低域に集中するため，高周波数成分は低周波成分に対して十分小さいとみなせることに注目する．すなわち，$\hat{\boldsymbol{x}}^{\mathrm{NB}} - \boldsymbol{y}^{\mathrm{NB}}$ は低域の振幅差によりほぼ決まる．従って，MSE を用いた時間波形の学習は，低周波成分により行われ，ダウンサンプリングされた信号を用いることにより，観測信号の低周波数成分に含まれる雑音が除去され，低域の音声構造を強調することが可能である．

しかし，前節で述べたように，音声の高域にある無声音の構造はホワイトノイズのようなランダム性の高い雑音と類似する特徴を有する．そのため，畳み込みフィルタによる特徴量の抽出は困難である．さらに，ネットワークを MSE に基づき学習する場合，復元した高域成分の波形がクリーン音声よりも平滑化される可能性がある [44]．よって，高域成分の復元処理では，RNN を用いて，時刻 n までの情報により，時間 n におけるサンプルを予測する．ここで，入出力データである時間波形を 8 ビット量子化することで，時間波形の振幅が $0 - 255$ の整数となり，トレーニン

グを高速化できる．時間波形の 1 サンプルにおける全結合層の出力は，各真値に相当する確率分布であり，256 次元のベクトルとなる．ネットワーク構造を図 3.7 に示す．図 3.7 に示すように，高域復元用のネットワークは 2 層の LSTM と 2 層の全結合層から構成され，時間ステップ順入力波形を処理する．時刻 n における LSTM 層の処理を式 (3.2) に示す．

$$\boldsymbol{S}(n) = \mathcal{G}(\boldsymbol{S}(n-1), \hat{x}^{\mathrm{WB}}(n)), \tag{3.2}$$

ここで，$\boldsymbol{S}(n)$ は LSTM の出力，$\hat{x}^{\mathrm{NB}}(n)$ はネットワークの入力，$\mathcal{G}(\cdot)$ は LSTM の活性化関数である．活性化関数 softmax を用いて，全結合層の出力に対して，最も大きな確率を取る予測値を出力する．時刻 n における，出力 $\hat{y}^{\mathrm{WB}}(n)$ を式 (3.3) の条件付き確率に基づき算出する．

$$p(\hat{y}^{\mathrm{WB}}(n)|\hat{x}^{\mathrm{WB}}(1), \hat{x}^{\mathrm{WB}}(2), ..., \hat{x}^{\mathrm{WB}}(n)) = \mathrm{FC}(\boldsymbol{S}(n)), \tag{3.3}$$

ここで，FC(·) は全結合層の出力である．ネットワークのトレーニングでは，雑音抑圧した収録音声 $\hat{\boldsymbol{x}}^{\mathrm{WB}}$ とクリーン音声 $\boldsymbol{y}$ を入力する．高域成分の復元処理では，損失関数として Cross entropy (CE) を，最適化関数として Adam をそれぞれ用いた．

音声強調段階では，まず，ダウンサンプリングにより，収録音声 $\boldsymbol{x}$ から高域成分を除去する．その後，ダウンサンプリング後の音声 $\boldsymbol{x}^{\mathrm{NB}}$ に対して雑音抑圧用 DNN による雑音抑圧を行い，低域成分を強調する．そして，低域強調した結果 $\hat{\boldsymbol{x}}^{\mathrm{NB}}$ をアップサンプリングし，これを μ-law により 8 ビット量子化する．その結果を高域復元用 DNN に入力し，高域成分を再構築する．最後に，復元音声は高域復元用 DNN の出力 $\hat{y}^{\mathrm{WB}}$ を μ-law によりデコーディングしてから得られる．

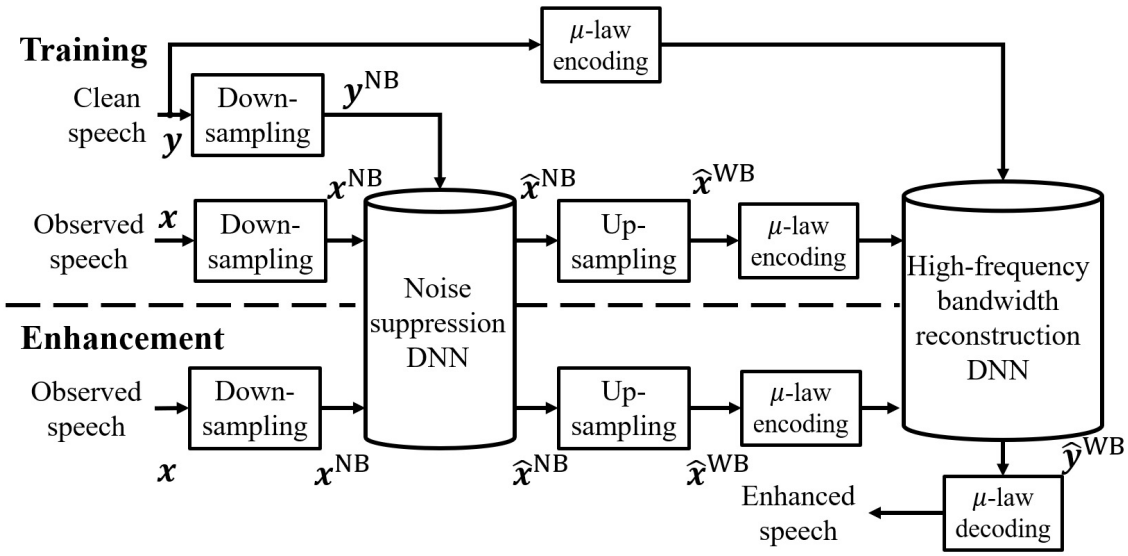


図 3.5 提案手法 2 のブロック図

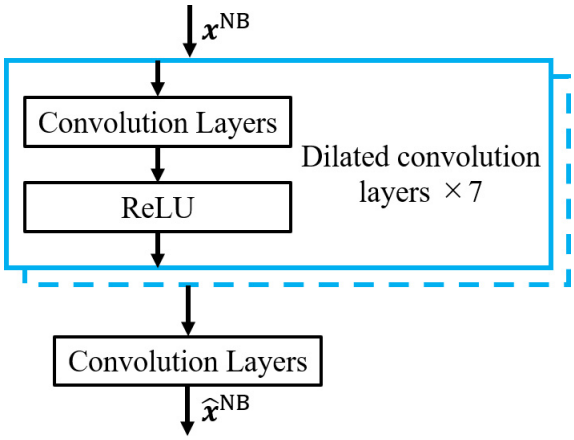


図 3.6 雑音抑圧のネットワークの構造

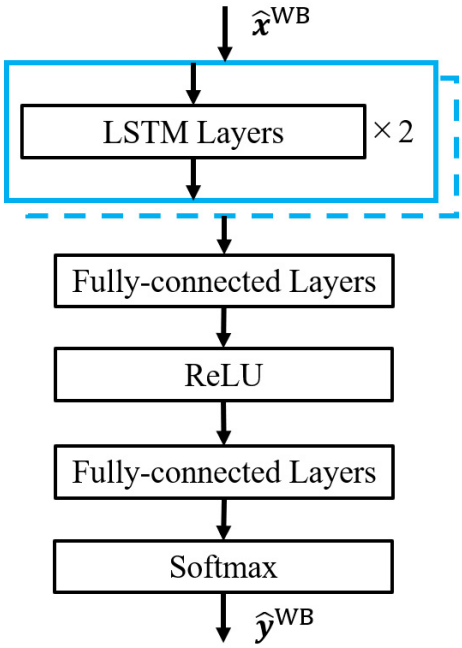


図 3.7 高域復元のネットワークの構造

3.4. 被照射物体が既知 (特定) の場合の音声強調実験

本節では，まず各提案手法における評価実験を実施することで 2 段階処理の必要性を説明する．その後，クリーン音声，収録音声，従来手法による処理結果，及び提案手法による処理結果を比較し，提案手法の有効性を確認する．なお，被照射物体が既知 (特定) の場合において，本実験ではペットボトルを用いた観測音声を使用した．ペットボトルによる観測音声は欠落程度の最も高い観測音声であり，内部の空気による残響も検討できる．

3.4.1 ネットワークの構造及び学習条件

提案手法 1 における DNN

STFT 特徴量を抽出する際に，フーリエ点数を 1024 とし，フレームシフト長は 256 点とした．なお，離散フーリエ変換の対称性を考慮し，各フレームに対して振幅復元用 DNN の入出力の次元を 513 に設定した．LSTM 層と中間の全結合層のユニット数は 1024 とし，学習率は 0.001 に設定した．

位相復元用 DNN において，スペクトルのフレーム方向の連続性を考慮した上，1 フレームの出力に対して前後 2 フレームを含めた，計 5 フレームを入力とした．畳み込み層において，1 層目の畳み込みフィルタサイズは 5×9 に設定し，それ以外は 1×9 に設定した．また，畳み込み層のフィルタ数において，出力層は 1 であり，それ以外は 128 とした．学習率は 0.00001 に設定した．

提案手法 2 における DNN

雑音抑圧用 DNN の入出力は 2048 サンプル数の時間波形とした．ネットワークは Dilated convolution 構造を有する畳み込み層 8 層から構成され，Dilated 係数は $2^n, n = 0, 1, ..., 7$ に設定した．畳み込みフィルタサイズは，9×1 である．出力層のフィルタ数は 1 であり，それ以外各層のフィルタ数は 128 でした．学習率は 0.0001 に設定した．

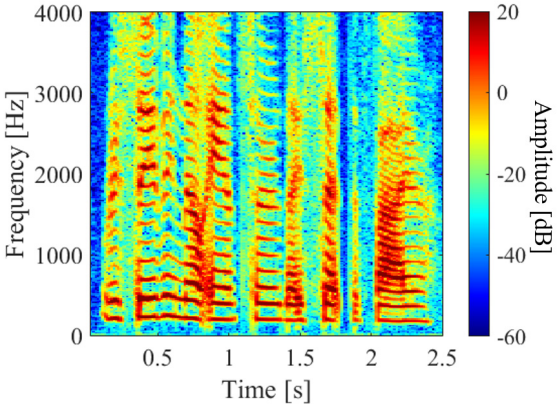
高域復元 DNN に使用される LSTM 層と全結合層のユニット数は全て 1024 に設定した．ネットワークを学習する際に，Backpropagation through time (BPTT)[45] を利用し，過去 480 点の時間ステップまで遡り，現在時刻の成分を予測した．学習率は雑音抑圧用 DNN と同様に 0.0001 に設定した．

3.4.2 提案手法 1 における評価実験

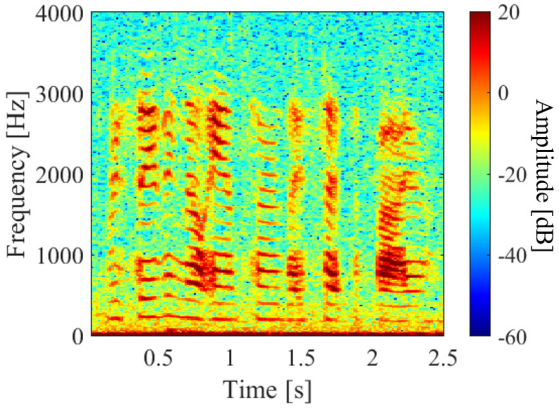
提案手法 1 の有効性を確認するために，収録音声，振幅スペクトルのみ復元した強調音声，振幅スペクトルと Griffin-Lim 法 (200 回反復) による位相スペクトル両方復元による強調音声及び提案手法 1 による強調音声を用いて評価実験を実施した．ま

ず，図 3.8 に各音声の振幅スペクトログラムを示す．図 3.8 より，提案手法 1 の結果 (e) は，振幅復元のみの結果 (c) と Griffin-Lim 法による位相復元を加えた結果 (d) に比べ雑音が抑圧されていることがわかる．特に，0−1 kHz 以下の帯域における雑音が抑圧され，音声特有の調波構造が明確になっていることも確認できる．

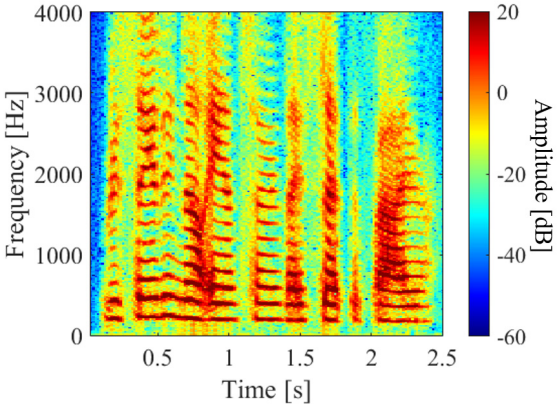
次に，提案した位相復元手法を確認するために，収録音声の位相，振幅強調後の振幅スペクトルにより，Griffin-Lim 法を用いて 200 回反復演算することで生成した位相及び提案手法により復元した位相それぞれとクリーン音声の位相のコサイン距離を用いて評価実験を実施した．復元位相とクリーン音声の位相のコサイン距離を図 3.9 に示す．本図において，コサイン距離が小さいほど位相復元の精度が高いことを表している．図 3.9 より，0−4 kHz の周波数帯域において，提案手法は Griffin-Lim 法に比べコサイン距離が 0.4 ほど小さくなっていることがわかる．また，0−8 kHz, すなわち復元処理を行っていない高域成分を含めても，提案手法は Griffin-Lim 法に比べコサイン距離が 0.2 ほど小さくなっていることがわかる．以上の結果によって，振幅復元および位相復元の両 DNN を用いる提案手法 1 は特定の被照射物体 (ペットボトル) を用いた観測音声の位相復元に有効であることを確認できた．



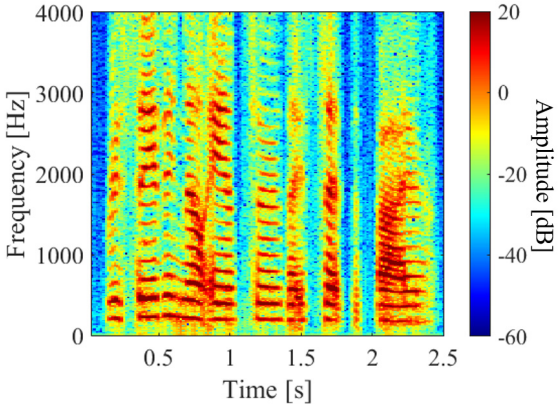
(a) クリーン音声



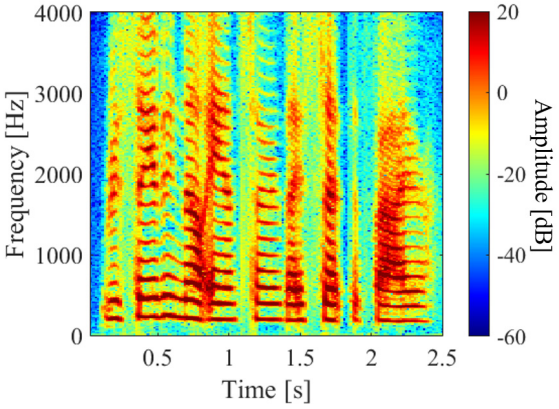
(b) 収録音声



(c) 振幅復元+観測位相



(d) 振幅復元+Griffin-Lim 法



(e) 振幅，位相の両方を復元

図 3.8　提案手法 1 における音声強調結果のスペクトログラム

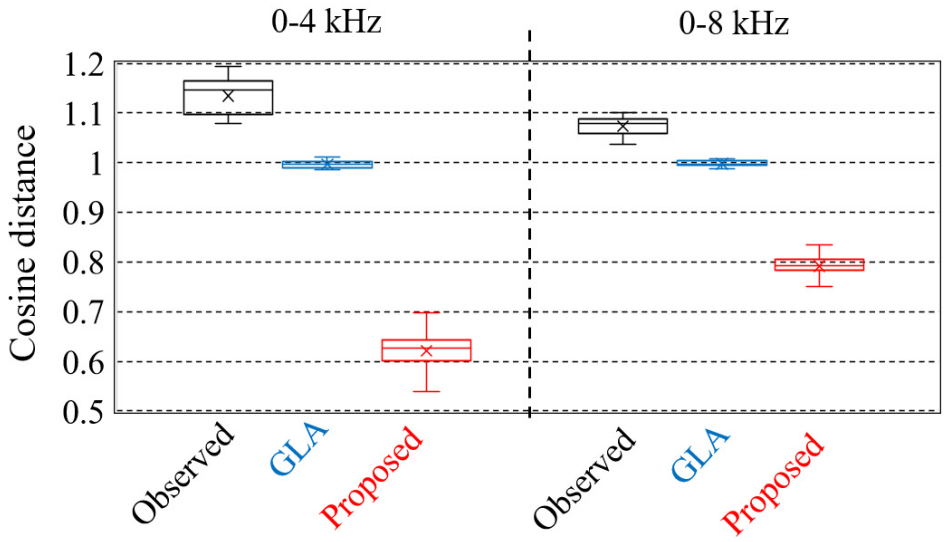
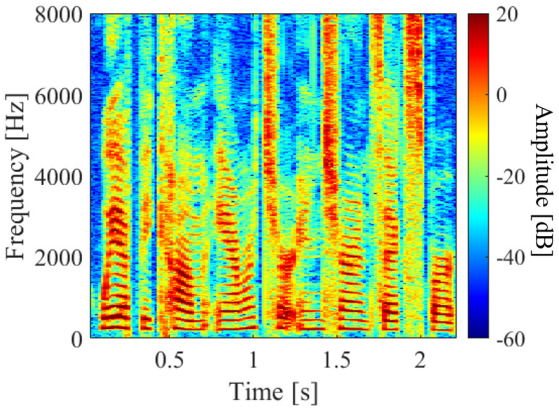


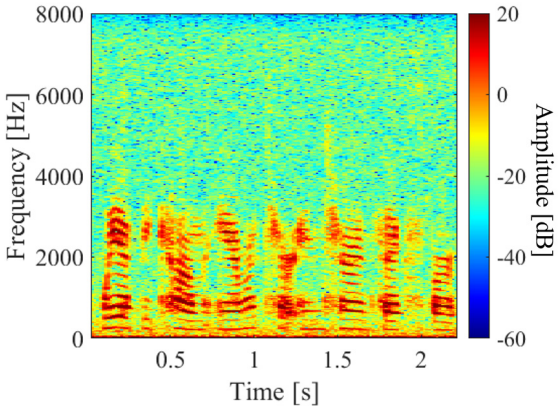
図 3.9 提案手法 1 における各手法により復元された位相とクリーン音声の位相のコサイン距離

3.4.3 提案手法 2 における評価実験

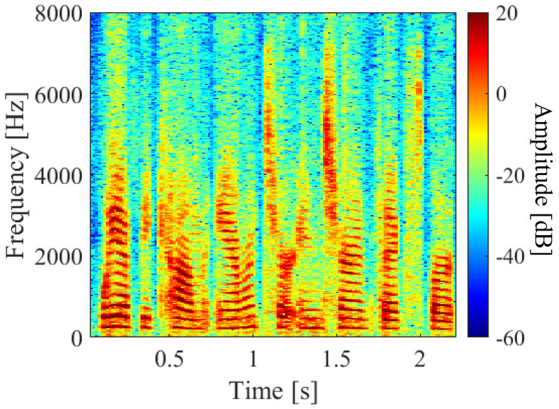
提案手法 2 の有効性を確認するために，収録音声，高域処理なしの結果及び提案手法 2 の結果，それぞれを用いて評価実験を実施した．まず，図 3.10 にクリーン音声，収録音声，雑音抑圧処理のみの結果及び提案手法 2 の結果の振幅スペクトルを示す．図 3.10 より，提案手法 2 により高域成分が復元されていることがわかる．また，図 3.10(c) より，提案手法 2 の雑音抑圧 DNN により，低周波数帯域における音声強調が十分に行われていることもわかる．次に，LSD 指標を用いて，各周波数帯域における，(b)，(c)，(d) を評価する．各周波数帯域における LSD を図 3.11 に示す．図 3.11 より，高域復元を行った 4–8 kHz において LSD が約 0.3 dB 小さくなっていることがわかる．また，0–4 kHz において，高域復元の有無による LSD の差は非常に小さいこともわかる．以上より，提案手法 2 を特定の被照射物体 (ペットボトル) を用いた観測音声に適用した結果，高域成分が強調されたことを確認できた．



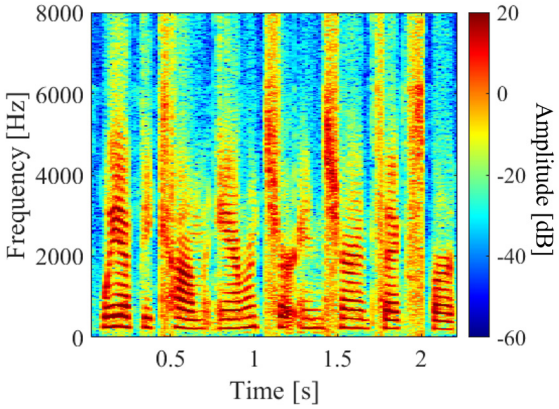
(a) クリーン音声



(b) 収録音声



(c) 雑音抑圧処理のみの音声



(d) 雑音抑圧と高域復元処理後の音声

図 3.10　提案手法 2 における音声強調結果のスペクトログラム

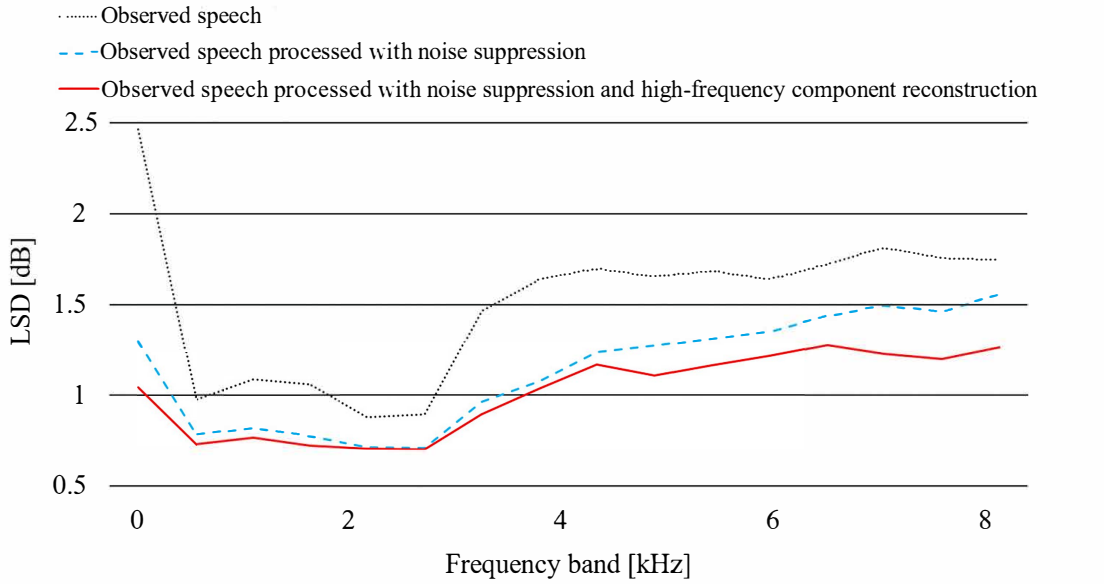


図 3.11　提案手法 2 における各周波数帯域における強調音声の LSD

3.5. 音声強調結果及び考察

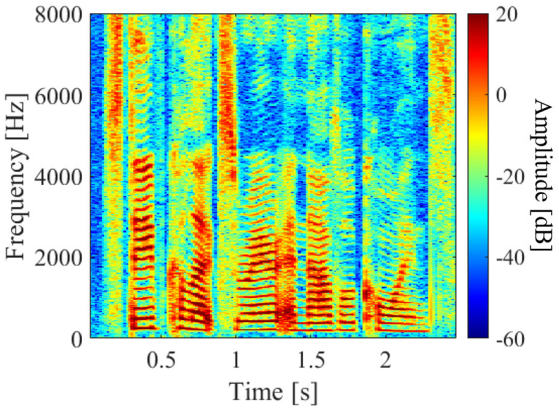
提案手法の有効性を確認するために，LDV による収録音声を用いた音声強調実験を実施した．本実験では，評価指標として，PESQ，LSD，STOI を用いて，音質，劣化度，明瞭度を比較した．比較対象はペットボトルを用いた収録音声，従来手法 [12] に基づく音声強調の結果，提案手法 1 の結果及び提案手法 2 の結果の計 4 音源であり，それぞれクリーン音声と比較した．各音声のスペクトログラムを図 3.12 に，各復元音声に対する評価指標を表 3.1 に示す．

図 3.12 より，提案手法 1，2 により，雑音が抑圧されていることが確認できる．また，振動物体の周波数応答により欠落した音声成分において，従来手法では，バンドパスフィルタを用いてその部分を完全に除去しているが，提案手法 1，2 により，その周波数帯域にある音声成分も復元されていることもわかる．次に，表 3.1 より，PESQ スコアにおいて，提案手法 1，2 は従来手法とほぼ同等の結果となっており，収録音声に比べ PESQ が 0.5 ほど高くなっていることが確認できる．また，LSD に

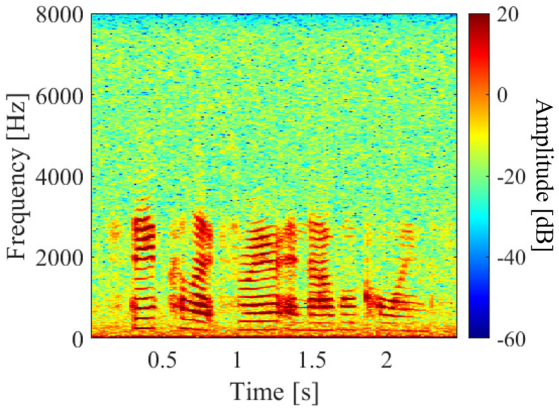
表 3.1 各手法に対する客観評価指標

	PESQ score	LSD[dB]	STOI score
観測音声	1.76±0.40	1.62±0.10	0.85±0.04
従来手法の結果	2.25±0.35	2.17±0.30	0.87±0.04
提案手法 1 の結果	2.25±0.30	1.09±0.08	0.93±0.02
提案手法 2 の結果	2.35±0.30	1.11±0.08	0.94±0.03

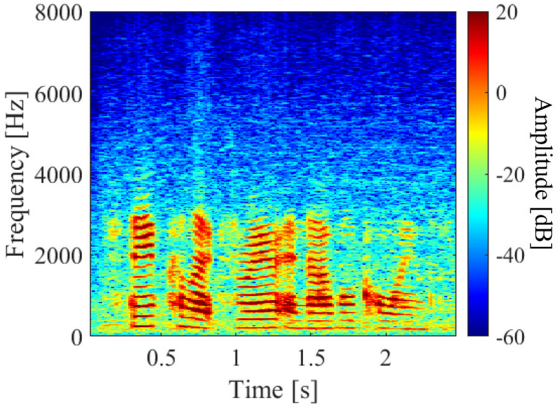
おいて，提案手法 1，2 が従来手法に比べ LSD が小さくなっており，音声の劣化が小さくなっていることがわかる．なお，従来手法では収録音声の 4−8 kHz の帯域，すなわち雑音が音声に比べ大きくなっている帯域での音声強調が困難であるため，提案手法 1，2 に比べ LSD が大きくなっている．最後に，STOI において，提案手法 1，2 は収録音声より 0.08 ほど向上し，従来手法よりも 0.06 ほど向上したことがわかる．以上の結果より，提案手法 1，2 が LDV による特定の被照射物体 (ペットボトル) を用いた観測音声の強調に有効であることが示された．



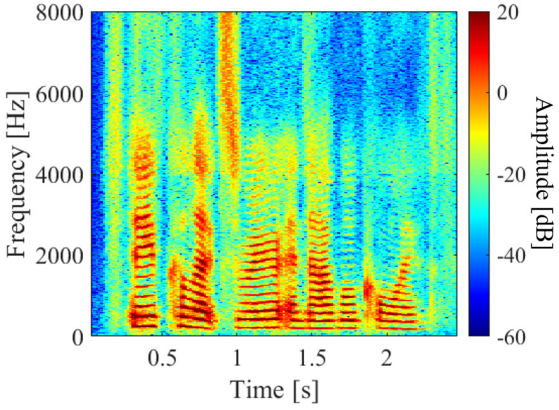
(a) クリーン音声



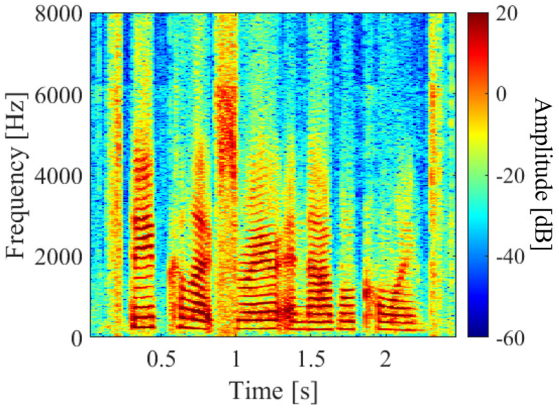
(b) 収録音声



(c) 従来手法の強調音声



(d) 提案手法 1 の強調音声



(e) 提案手法 2 の強調音声

図 3.12 クリーン音声，収録音声及び各手法による音声強調結果のスペクトログラム

3.6. まとめ

LDV はレーザ光を利用し，音波による振動を計測するため，計測物体の振動特性により，多種類の歪みが混入する．この多種類の歪みにより，従来の音声強調手法を用いても十分な音声強調精度が得られないという問題がある．そこで本章では，振幅復元，位相復元による二段階処理 DNN を用いた手法，雑音抑圧と高域情報復元の 2 つの DNN を利用した二段階処理による手法を提案した．3.2 節で，STFT に基づく音声強調手法，3.3 節で，時間波形に基づく音声強調手法それぞれについて説明した．そして，3.4 節で，2 つの手法で使用したネットワークの構造と学習条件を説明し，評価実験により 2 つの提案手法に対して，二段階処理の必要性を述べた．最後に，3.5 節で，2 つの提案手法を用いてペットボトルを用いた観測音声を強調し，その結果を評価した．評価実験の結果により，提案手法は特定の被照射物体 (ペットボトル) を用いた音声に対する有効性を示すことができた．

第4章　被照射物体が未知の場合における音声強調

4.1. はじめに

2.5 節で述べたように，DNN を用いた音声強調手法の強調精度は，トレーニングデータセットに大きく依存する． したがって，DNN を使用して音声を強調する従来手法では，被照射物体が未知の場合において，その音声強調結果は非常に低下している．観測音声からクリーン音声を直接推定する従来手法と異なり，本章で提案した手法は，まず，観測音声からロバストな特徴量を抽出し，パワースペクトルと位相スペクトル各々を復元する．最後に，各復元結果に基づき音声を合成する方法を提案する．これらの特徴量は被照射物体の振動特性による影響は小さいため，未知の被照射物体にも適用できる．4.2 節において，提案手法の各ステップの処理を詳しく説明する．そして，4.3 節では，各ステップに使用したネットワークの構造と学習条件を述べる．次に，4.4 節では，評価実験を行い，その結果により，提案手法の有効性を確認する．

4.2. 提案手法: 包絡補正を使用した音声再合成

図 4.1 に提案手法のブロック図を示す．前章にも述べたように，従来の音声強調方法は通常，観測音声のパワースペクトルのみを処理し，音声を再構築する際に観察音声の位相スペクトルをそのまま利用する．ただし，LDV によるの観測音声には多様な歪みが存在しており，特に振動特性により音声成分が欠落し，位相遅延も発生する．よって，観測音声の位相スペクトルを音声復元に直接適用すると，再構成された音声の音質が著しく低下する．したがって，LDV の観測音声において，振幅

と位相スペクトルの両方を推定する必要がある．振幅スペクトルについては，被照射物体の振幅応答が各周波数帯域の音声のエネルギー，いわゆるスペクトル包絡を決定するだけであり，音声のピッチ情報には影響しない．これは，図 2.3 より確認できる．したがって，まず，観測音声のパワースペクトルからピッチとスペクトル包絡情報を抽出し，スペクトル包絡のみを補正する．そして，ピッチと処理後のスペクトル包絡を用いて，音声パワースペクトルを再合成する．音声の位相スペクトルの構造は非常に複雑であるため，観測音声とクリーン音声の位相のマッピングを学習し，位相を直接復元することは困難である．したがって，位相スペクトルは，再構成された音声のパワースペクトルを基に推定する．提案手法の各ステップについては，次の項で説明する．

4.2.1 ピッチと包絡情報の抽出

2.3 節では，LDV によるの観測音声の歪みの原因を説明した．被照射物体表面の反射率は固有の特性であるため，振動しない場合でも観測音音声に雑音が存在する (図 1.1，観測音声の無音部分を参照)．そして，物体が振動すると，音源と相関した雑音が発生するが (例えば，物体の表面が粗いため，反射光の角度がわずかに変化する)，図 1.1 で示したように，雑音のレベルはほぼ安定しており，音源と相関する雑音はほとんど観察されない．さらに，4.3 節の実験結果により，音源と相関する雑音はほぼ実験結果に影響を与えない．したがって，雑音は定常の加法性雑音と仮定し，観測音声は式 (4.1) に表すことができる．

$$x(t) = y(t) * h(t) + n(t), \tag{4.1}$$

ここで，$x(t)$，$y(t)$，$n(t)$，$h(t)$，t は，それぞれ観測音声，クリーン音声，雑音，被照射物体のインパルス応答および時間インデックスであり，$*$ は畳み込み演算子である．

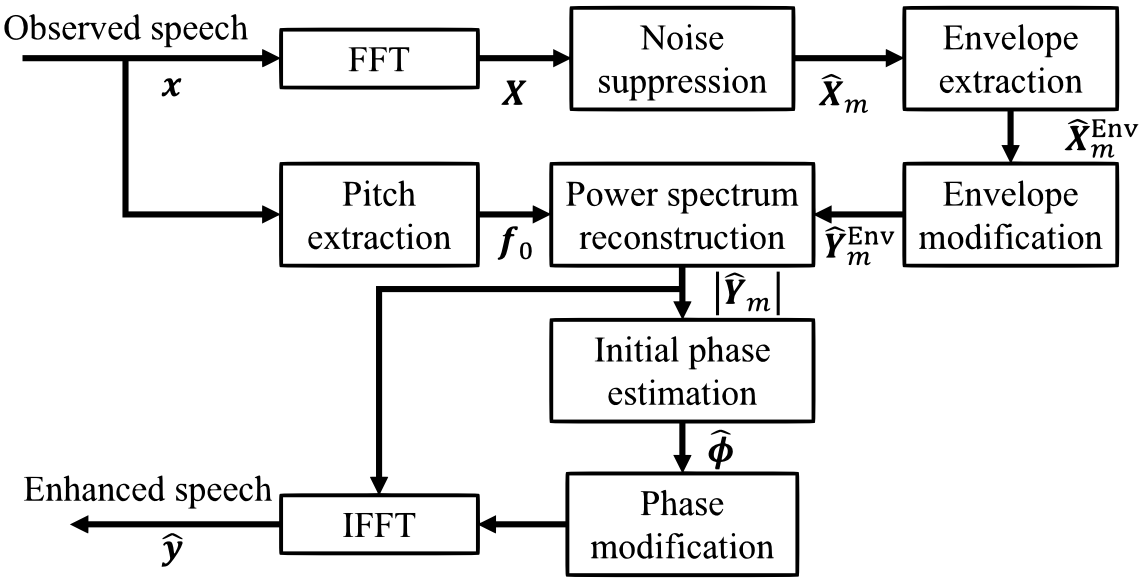


図 4.1 提案手法のブロック図

音声は調波構造を持つため，短時間音声信号は時間領域で明確な周期性を示す．したがって，その自己相関シーケンスを計算すると，ピッチは f_s/l_{Peak} として算出できる．ここで，f_s と l_{Peak} は，それぞれサンプリング周波数と自己相関シーケンスの隣接するピーク間の長さである．フレーム長とシフト長は 64 ms と 16 ms に設定する．図 1.1 から観測できるように，LDV の観測音声には 75 Hz 以下の低周波数帯域において，常に大きな雑音が含まれており，ピッチ算出の精度に大きく影響を与えるため，ピッチを計算する前に低周波雑音を除去する必要がある．よって，まず，カットオフ周波数 100 Hz のハイパスフィルターを適用する．さらに，現時点のフレームは無音声区間か否かを判別するために，自己相関係数の絶対値の平均としてパラメータ α を設定する．観測音声の高周波帯域には無声音声と同様の構造を持つ

雑音が必ず存在するため，実験により α の閾値を 0.04 に設定した．$\alpha > 0.04$ の場合，現時点のフレームは音声ありフレームとして判断し，それ以外の場合は無音声フレームまたは無声音フレームとして定義される．人間の音声特性を考慮して，音声ありのフレームのピッチ範囲は 50–300 Hz に制限する．ピッチ検出はフレームごと行うため，現時点のフレームのピッチはこの範囲外の場合，またはフレームが無声音あるいは無音声として判断されている場合，ピッチを 0 Hz に設定する．

雑音は定常加法性雑音であると仮定されているため，スペクトル減算法 [46, 47] を利用するだけで雑音を十分に抑圧できると考えられる．観測音声のパワースペクトルの雑音成分をスペクトル減算法によって抑圧した後，パワースペクトルの包絡線を抽出する．スペクトルの包絡抽出 [48, 49] にはさまざまな方法があるが，観測された音声は劣化しているため，高精度で抽出する必要はない．よって，本論文では，対数振幅スペクトルに対して，逆フーリエ変換 (Inverse Fourier transform: IFT) を行い，結果の低周波数部分を包絡として抽出する．

4.2.2 包絡補正

スペクトル包絡の補正は，未知の照射物体に対処するための鍵となる．このステップでは，DNN を利用して被照射対象物の振動特性によって劣化したスペクトル包絡を補正する．被照射物体は未知であるため，異なる物体を照射することによって観測される音声信号の包絡線も異なる．入力と出力として観測音声とクリーンな音声のスペクトル包絡を直接トレーニングすることは不可能である．したがって，本論文では，未知の被照射物体に対応するために，振動特性の影響を受けない特徴を DNN の入力として検討する．式 (4.1) に基づき，観測音声とクリーンな音声を高速フーリエ変換 (Fast Fourier transform: FFT) によって得られる振幅スペクトルは次のように表せる．

$$|\boldsymbol{Y}_m| \times |\boldsymbol{H}| = |\boldsymbol{X}_m| - |\boldsymbol{N}_m|, \tag{4.2}$$

ここで，$|\boldsymbol{Y}_m|$，$|\boldsymbol{X}_m|$，$|\boldsymbol{N}_m|$ はそれぞれ，クリーン音声，観測音声及び雑音のパワースペクトルであり，m と $|\boldsymbol{H}|$ はフレームインデックスと振幅応答である．3.1 節で述べたように，$|\boldsymbol{N}_m|$ は先にスペクトル減算法によって推定される．ここで，$|\hat{\boldsymbol{X}}_m|$

を $|\boldsymbol{X}_m| - |\boldsymbol{N}_m|$ の結果とし，対数をとると，式 (4.2) は次のように書き換えられる．

$$\log_{10}|\boldsymbol{Y}_m| = \log_{10}|\hat{\boldsymbol{X}}_m| - \log_{10}|\boldsymbol{H}|. \tag{4.3}$$

$\log_{10}|\boldsymbol{X}|$ と $\log_{10}|\boldsymbol{Y}|$ をスペクトル包絡と微細構造の加算として表すと，$\log|\hat{\boldsymbol{X}}_m| = \hat{\boldsymbol{X}}_m^{\mathrm{Env}} + \boldsymbol{X}_m^{\mathrm{Det}}$，$\log|\boldsymbol{Y}_m| = \boldsymbol{Y}_m^{\mathrm{Env}} + \boldsymbol{Y}_m^{\mathrm{Det}}$ となる．よって，式 (4.3) は次のように書き換えることができる．

$$\boldsymbol{Y}_m^{\mathrm{Env}} + \boldsymbol{Y}_m^{\mathrm{Det}} = \hat{\boldsymbol{X}}_m^{\mathrm{Env}} + \hat{\boldsymbol{X}}_m^{\mathrm{Det}} - \log_{10}|\boldsymbol{H}|. \tag{4.4}$$

振動特性は主にスペクトル包絡に影響を与えるため、微細構造への影響は小さい．したがって，$\boldsymbol{Y}_m^{\mathrm{Det}} = \hat{\boldsymbol{X}}_m^{\mathrm{Det}}$ と近似的に考えることができる．よって，式 (4.4) は次のように表せる．

$$\boldsymbol{Y}_m^{\mathrm{Env}} = \hat{\boldsymbol{X}}_m^{\mathrm{Env}} - \log_{10}|\boldsymbol{H}|. \tag{4.5}$$

$\log_{10}|\boldsymbol{H}|$ は照射物体の固有の性質であるため，フレーム方向の差分をとると，振動特性の影響を軽減できると考える．よって，式 (4.5) は次のようになる．

$$\boldsymbol{Y}_m^{\mathrm{Env}} - \boldsymbol{Y}_{m-1}^{\mathrm{Env}} = \hat{\boldsymbol{X}}_m^{\mathrm{Env}} - \hat{\boldsymbol{X}}_{m-1}^{\mathrm{Env}}. \tag{4.6}$$

音声信号はフレーム方向で連続性を持つことを考慮し，再帰型ニューラルネットワーク (RNN) を利用して，図 4.2 に示すように，スペクトル包絡のフレーム方向の差分を利用してスペクトル包絡を推定することが可能と考えられる．

$$\boldsymbol{Y}_m^{\mathrm{Env}} = \mathbb{G}[\hat{\boldsymbol{X}}_1^{\mathrm{Env}} - \hat{\boldsymbol{X}}_0^{\mathrm{Env}}, ..., \hat{\boldsymbol{X}}_m^{\mathrm{Env}} - \hat{\boldsymbol{X}}_{m-1}^{\mathrm{Env}}], \tag{4.7}$$

ここで，$\mathbb{G}\cdot$ は RNN の計算を表す．つまり，提案手法は，観測音声とクリーン音声のスペクトル包絡間のマッピングの計算を，スペクトル包絡の差分とスペクトル包絡のマッピングに変換した．一般的に，初期ベクトルが既知であれば，スペクトル包絡を簡単に算出できることから，わざわざ DNN を使用する必要もないように考えられるが，前節で説明したように，観測音声にはさまざまな歪みが含まれるため，スペクトル減算法ではそれらすべてを除去できるわけではない．高精度にスペクトル包絡を推定するため，DNN の使用は依然として必要となる．本論文では，スペクトル包絡と微細構造が別々に処理されるため，前節で述べたように，低レベルの相

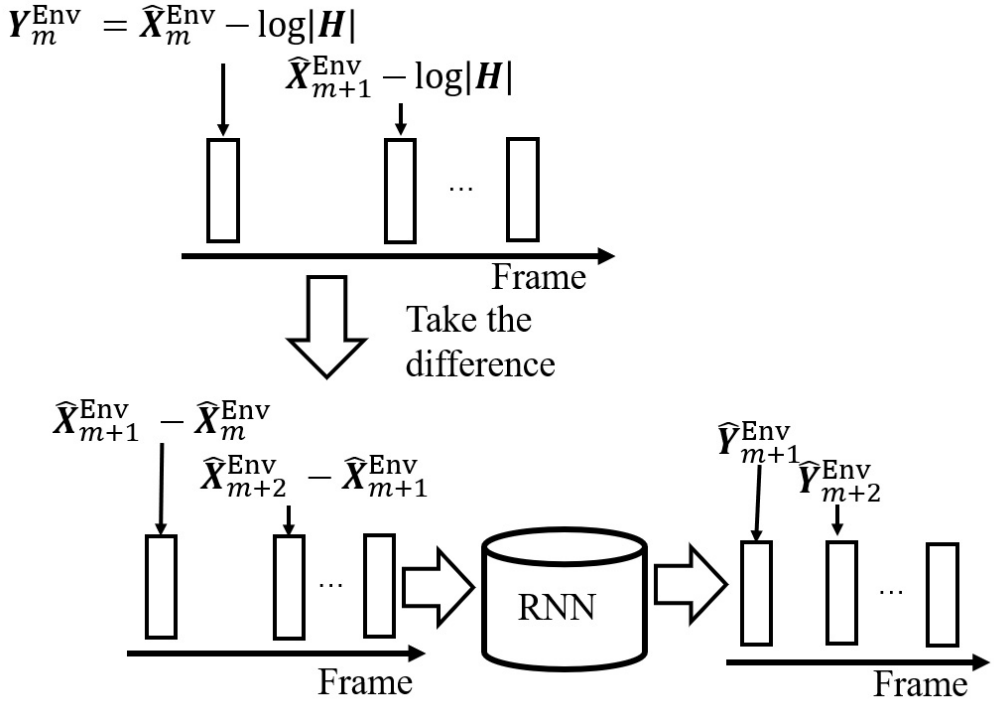


図 4.2 スペクトル包絡補正処理の手順

関雑音を無視しても，生成された音声の音質には影響しない．その理由は，物体が振動している場合，推定されたスペクトル包絡に含まれる微小な相関雑音成分のパワーは音声合成の際に，音声のエネルギーとみなされるからである．滑らかな音声包絡の形状を目指して，トレーニングする際に，スペクトル包絡推定の損失関数には MSE を利用する [44]．学習条件として，学習率を 0.0001 に設定し，FFT 長とシフト長はそれぞれ 64 ms と 16 ms とした．

4.2.3 パワースペクトル再構築

このステップの処理を図 4.3 に示す．音声の調波構造を復元するため，まず，観測音声から抽出されたピッチ情報を用いて，ピッチの整数倍の周波数を持つ正弦波

を生成し，重ね合わせる．生成された波形に FFT を適用し，4.2.1 項の処理と同様に，微細構造を抽出する．次に，抽出された微細構造と推定された包絡情報を利用して，初期のパワースペクトルを合成する．最後に，DNN を使用して，生成されたパワースペクトルとクリーン音声のパワースペクトルの関係をモデル化し，画像の超解像手法 [50] を利用し，パワースペクトルをさらに処理する．このステップの処理は，敵対的生成ネットワーク (Generative adversarial networks: GAN)[51] によって実行され，より自然な音声を生成するために，結果にさまざまなランダム微細構造を加える．生成器と判別器の損失関数はそれぞれ次のように定義する．

$$L_{\mathrm{G}} = \mathbb{E}_{\hat{\boldsymbol{S}}\sim p(\hat{\boldsymbol{S}})}\big[(\mathrm{D}(\mathrm{G}(\hat{\boldsymbol{S}}))-1)^2 + \lambda||\mathrm{G}(\hat{\boldsymbol{S}})-\boldsymbol{S}||_1\big], \tag{4.8}$$

$$L_{\mathrm{D}} = \mathbb{E}_{\boldsymbol{S}\sim p(\boldsymbol{S}),\hat{\boldsymbol{S}}\sim p(\hat{\boldsymbol{S}})}\big[\mathrm{D}(\boldsymbol{S}) + (\mathrm{D}(\mathrm{G}(\hat{\boldsymbol{S}}))-1)^2\big], \tag{4.9}$$

ここで，$\hat{\boldsymbol{S}}$，$\boldsymbol{S}$，p，および λ は，それぞれ生成されたパワースペクトル，クリーン音声のパワースペクトル，$\hat{\boldsymbol{S}}$，$\boldsymbol{S}$ の分布，および L1 ノルムの係数である．実験により，λ は 10 に設定した．$\mathrm{G}(\cdot)$ と $\mathrm{D}(\cdot)$ は，生成器と判別器の出力であり，学習率は両方とも 0.0001 に設定した．

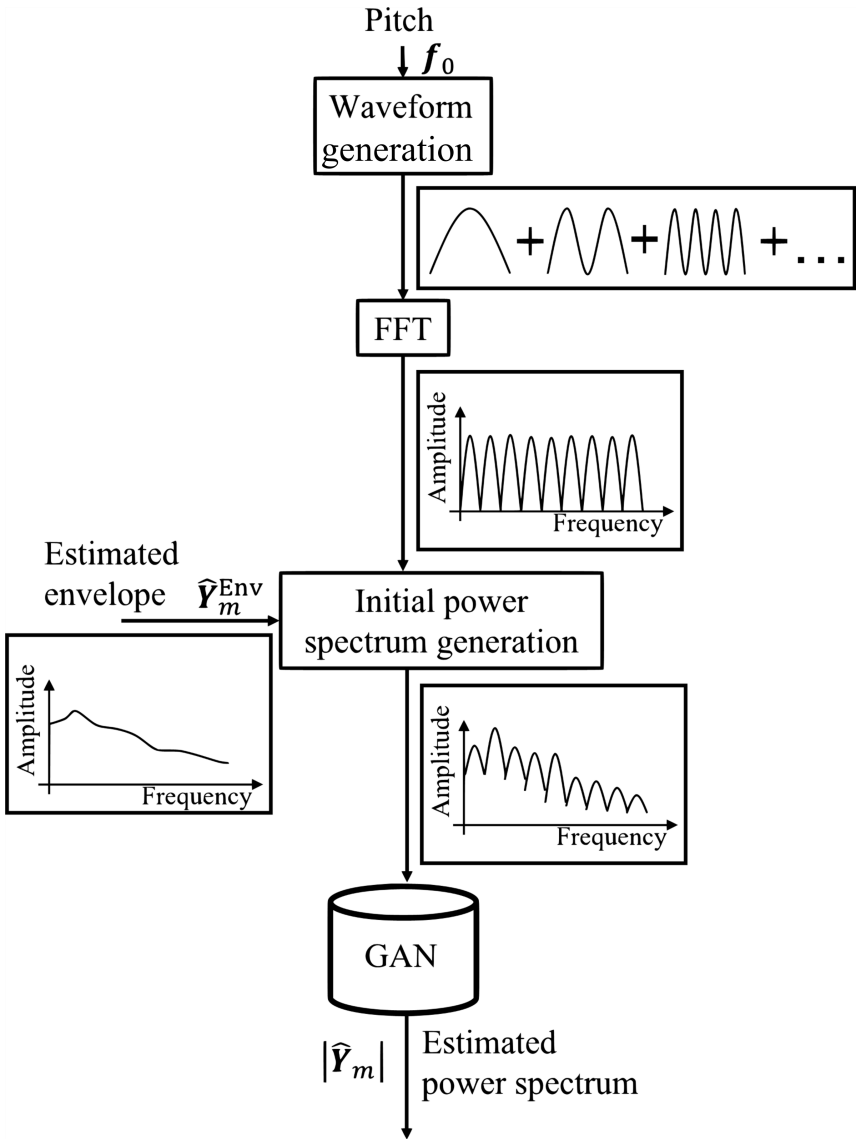


図 4.3 パワースペクトル推定の手順

4.2.4 位相スペクトル再構築

通常の位相強調の方法について，まず，初期位相スペクトルを推定し，それを Griffin-Lim アルゴリズム [27, 28] を適用して反復演算より調整する [36]．位相スペク

トルは複雑さとランダム性をもつため，DNN を直接使用して，観測音声とクリーンな音声をマッピングすることは困難である．よって，提案手法では，レーザ光が照射された物体が不明であるため，DNN を使用してパワースペクトル [29] から初期位相スペクトルを推定する．ネットワークの入力と出力は，それぞれ GAN によって生成されたパワースペクトルと推定された位相スペクトルである．パワースペクトルはフレーム方向で連続性を持つことを考慮し，推定精度を向上させるために，入力パワースペクトルは現フレームと前後 2 フレームを組み合わせたものとなる．損失関数 $L_{\text{Phase\_est}}$ は次のように定義する．

$$L_{\text{Phase\_est}} = L_{\text{Phase}} + \alpha L_{\text{GroupDelay}}, \tag{4.10}$$

$$L_{\text{Phase}} = \sum_{k=0}^{K-1} 1 - \cos(\boldsymbol{\phi}_k - \hat{\boldsymbol{\phi}}_k), \tag{4.11}$$

$$L_{\text{GroupDelay}} = \sum_{k=0}^{K-1} 1 - \cos(\Delta\boldsymbol{\phi}_k - \Delta\hat{\boldsymbol{\phi}}_k), \tag{4.12}$$

$$\Delta\phi_k = \boldsymbol{\phi}_{k+1} - \boldsymbol{\phi}_k, \Delta\hat{\boldsymbol{\phi}}_k = \hat{\boldsymbol{\phi}}_{k+1} - \hat{\boldsymbol{\phi}}_k, \tag{4.13}$$

ここで、k，$\boldsymbol{\phi}$，$\hat{\boldsymbol{\phi}}$ は周波数インデックス，クリーン音声の位相スペクトル，推定位相スペクトルであり，α は係数で，0.1 に設定する．音声の高周波成分の多くは無声音であり，その構造はホワイトノイズの構造に似ているため，DNN ではこれらの成分をモデル化することが困難となる．そのため，高域成分の推定精度は低周波成分の推定精度よりも低くなる．さらに，人間の聴覚は一般に，そのような無声音成分の位相に対して鈍感であると考えられる．したがって，提案手法では，位相スペクトルの推定を 0〜4000 Hz の低域のみに対して行い，高域の位相スペクトルは観測された音声の位相そのままを使用する．位相スペクトル再構成 DNN の学習率と Griffin-Lim アルゴリズムの反復回数は，それぞれ 0.0001 と 200 に設定した．

最後に，逆高速フーリエ変換 (Inverse fast Fourier transform: IFFT) によって，推定されたパワースペクトルと位相のスペクトルを用いて，強調音声を合成する．

4.3. 被照射物体が未知の場合の音声強調性能

この節では，まず，DNN の構造とトレーニング条件の設定について詳しく説明する．次に，提案手法の有効性を確認するために行った客観的評価実験を説明する．

4.3.1 ネットワークのトレーニング

ネットワークパラメータは表 4.1，4.2 に示す．包絡推定用 DNN は 6 つの全結合層と 1 つの LSTM 層から構成され，活性化関数としてパラメトリック整流線形ユニット (Parametric rectified linear unit: PReLu)[52] を選択した．パワースペクトル再構成ネットワークでは，図 4.4 に示すように，生成器として 4 つの畳み込み層と 4 つの転置畳み込み層を備えたオートエンコーダネットワークを使用する．エンコーダ層とデコーダ層が同じ次元で，スキップ接続構造を利用した．判別器の構造はジェネレータのエンコーダ部分と同様である．生成器と判別器は両方とも活性化関数として PReLU を使用する．位相スペクトル再構成ネットワークは 4 つの畳み込み層で構成され，ゲート線形ユニット (Gated linear units: GLU)[41] が活性化関数として使用される．

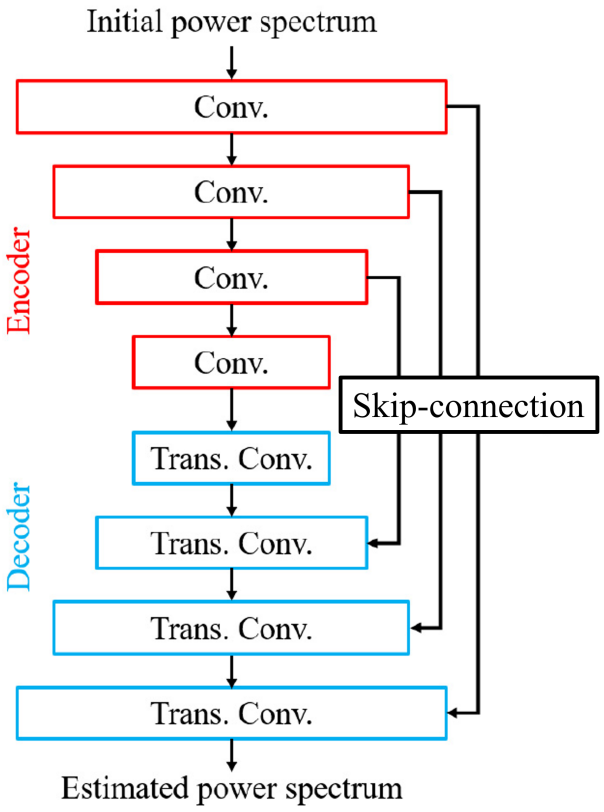

Initial power spectrum
Conv.
Conv.
Conv.
Conv.
Encoder
Skip-connection
Trans. Conv.
Trans. Conv.
Decoder
Trans. Conv.
Trans. Conv.
Estimated power spectrum

図 4.4 生成器の構造

表 4.1 ニューラルネットワークの構造 (1/2)

Envelope estimation:
Input: The first order difference of observed speech spectrum envelope after noise suppression process (513)
LSTM (1024) × 1
Fully connected (512) × 2, PReLU
Fully connected (256) × 2, PReLU
Fully connected (512) × 2, PReLU
Fully connected (513) × 1
Output: Estimated speech spectrum envelope (513)
Power spectrum reconstruction:
Generator input: Initial power spectrum (1 × 32 × 513)
Conv. 1D (512), kernel width: 9, stride: 2, PReLU
Conv. 1D (512), kernel width: 5, stride: 2, PReLU
Conv. 1D (1024), kernel width: 3, stride: 2, PReLU
Conv. 1D (1024), kernel width: 4, stride: 2, PReLU
Trans. Conv. 1D (1024), kernel width: 4, stride: 1, PReLU
Trans. Conv. 1D (1024), kernel width: 4, stride: 2, PReLU
Trans. Conv. 1D (512), kernel width: 4, stride: 2, PReLU
Trans. Conv. 1D (513), kernel width: 4, stride: 2
Generator output: Estimated power spectrum (1 × 32 × 513)
Discriminator input: Generated power spectrum (1 × 32 × 513)
Conv. 1D (128), kernel width: 3, stride: 2, ReLU
Conv. 1D (256), kernel width: 3, stride: 2, ReLU
Conv. 1D (512), kernel width: 3, stride: 2, ReLU
Conv. 1D (1), kernel width: 4, stride: 1
Discriminator output: Real/Fake score

表 4.2 ニューラルネットワークの構造 (2/2)

Initial phase estimation:
Input: Generated power spectrum (256 × 5)
Conv. 2D (128), kernel width: 9 × 5, stride: 1, padding: [4,4,0,0], GLU
Conv. 1D (128), kernel width: 9, stride: 1, padding: same size, GLU
Conv. 1D (128), kernel width: 9, stride: 1, padding: same size, GLU
Conv. 1D (128), kernel width: 9, stride: 1, padding: same size
Output: Estimated initial phase (256)

4.3.2 包絡推定結果評価

この実験では，既知の照射対象物として印刷用紙，プラスチック板，アルミ板を使用して，それぞれにより観察された音声を用いて 3 つのモデルをトレーニングした．1 つの既知の被照射物体については，他の物体は未知の被照射物体と扱い，トレーニングされたモデルの有効性を評価する．例えば，印刷用紙の観測音声でトレーニングしたモデルで，プラスチック板とアルミ板の観察音声を入力して評価する場合，印刷用紙は既知の物体であり，プラスチック板とアルミ板は未知の物体である．ただし，本研究は材質の振動特性による歪みを低減する方法を検討することが目的であるため，被照射物体は単一の材質にて構成され，50 mm × 50 mm 板形状とし，形状に起因した残響などの影響を防ぐ対策を講じた．

評価基準は以下の式により定義し，値が小さいほど推定したスペクトル包絡はクリーン音声の包絡に近いことを示す．

$$L = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(10\log\frac{Y^{\mathrm{Env}}(k)}{\hat{Y}^{\mathrm{Env}}(k)}\right)^2}, \tag{4.14}$$

ここで，Y^{Env} と $\hat{Y}^{\mathrm{Env}}$ はそれぞれクリーン音声のパワースペクトル包絡と推定したパワースペクトルの包絡である．評価結果を表 4.3 に示す．表 4.3 により，観測音声の L の値は約 8–10 dB であることがわる．テスト用音声とトレーニング用音声が同じ対象の場合，5–7 dB 程度 L の値が改善し，最良の結果となった．テスト用の音

表 4.3 スペクトル包絡推定の結果

Training data \ Test data	Printing paper			Aluminum sheet		
	obs.	con.	pro.	obs.	con.	pro.
Printing paper	9.00	**1.08**	**1.63**	7.69	6.68	3.09
Aluminum sheet	9.00	8.91	3.34	7.69	**1.06**	**1.55**
Plastic plate	9.00	16.81	3.38	7.69	15.81	3.43

Training data \ Test data	Plastic plate			Cardboard box		
	obs.	con.	pro.	obs.	con.	pro.
Printing paper	8.93	7.20	3.27	9.87	7.15	4.32
Aluminum sheet	8.93	7.20	3.39	9.87	6.80	3.90
Plastic plate	8.93	**1.43**	**1.85**	9.87	10.06	4.22

obs.: 観測音声, con.: 従来手法の結果, pro.: 提案手法の結果

声とトレーニング用の音声が異なる場合，つまり被照射物体が未知の場合でも，L の値が約 4–5 dB 向上することを確認できた.

4.4. 音声強調結果及び考察

この実験では，LSD，PESQ，STOI を評価基準として使用した．従来の LDV のための音声強調方法は，被照射物体が既知 (特定) であることを前提としてパラメータを決めていた．よって，提案手法の有効性を徹底的に調査するために，従来の DNN に基づく音声強調手法も比較対象とする [23]．結果を図 4.5，4.6，4.7 に示す．LSD 評価はパワースペクトルにのみ適用されるため，パワースペクトル補正の結果も図 4.6 の LSD 結果を参照できると考えられる．

図 4.6 により，LSD 評価結果において，照射対象が既知である場合でも，提案手法の結果が従来手法よりも向上している．これは，最初に雑音除去による前処理が要因の 1 つで，さらにスペクトル包絡情報のみをトレーニングするため，モデルが観測音声とクリーン音声の関係をマッピングし易いこともさらなる要因の 1 つと考えられる．図 4.5 PESQ の評価結果において，照射対象物が既知 (特定) の場合には提案手法と従来手法はほとんど差がない一方で，照射対象物が未知の場合には，提案手法が従来手法よりも大幅に向上した．プラスチック板による観測音声を使用してモデルをトレーニングした場合において，従来手法の結果では，未知の被照射物体の観測音声よりもさらに低下していた．この場合でも，提案手法の結果は，観測音声の結果よりも約 0.5 ほど PESQ の値が向上した．図 4.7 STOI の評価結果において，トレーニングデータとテストデータが同一の物体であるの場合には，提案手法の結果が従来手法に比べて若干劣る結果となった．しかし，被照射物体が未知の場合には，従来手法の結果は観測音声より低下していた一方で，提案手法の結果では観測音声に比べて STOI の値が向上したことが確認できる．

例として，図 4.8 と 4.9 は，アルミ板とダンボール板による観測音声を用いて，トレーニングした 3 つのモデルを用いた結果のパワースペクトルを示す．図 4.8 と 4.9 からわかるように，物体の振幅応答とその表面の反射率の違いにより，従来の方法をすべての未知の照射物体に適用することはできない．ここで，提案法の結果は，雑音抑圧と欠落した音声成分の復元の両方において高い性能を示した．これらの結果より，提案手法の有効性は上記の実験結果によって確認できた．

4.5. まとめ

本章では，被照射物体が未知の場合でも適用可能な LDV の音声強調手法を提案した. 従来手法はバンドパスフィルタなどを利用して雑音パワーを推定することで観測音声の音質を向上させることができたが，未知の観測物体に適用すると音質は十分ではなく，また観測音声より低下することもあった．そこで提案手法では，被照射物体の振幅応答の影響を受けない特徴量を利用した．具体的には，未知の被照射物体にも対応するために，4.2 節では，提案手法の各ステップの処理を詳しく説明した．そして，4.3 節では，各ステップに使用したネットワークの構造と学習条件を述べた．次に，4.4 節では，評価実験を行い，その結果により，提案手法の有効性を確認した.

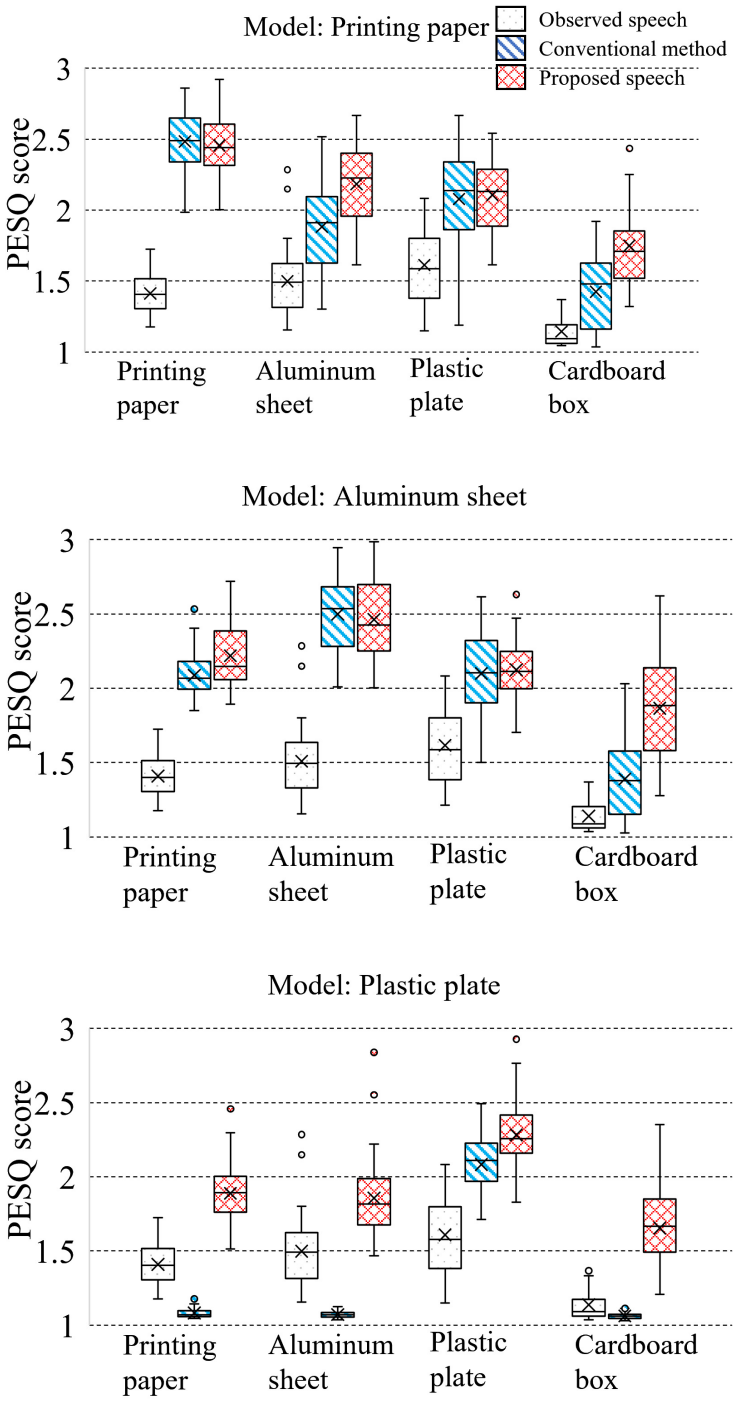


図 4.5 PESQ 評価における提案手法の結果

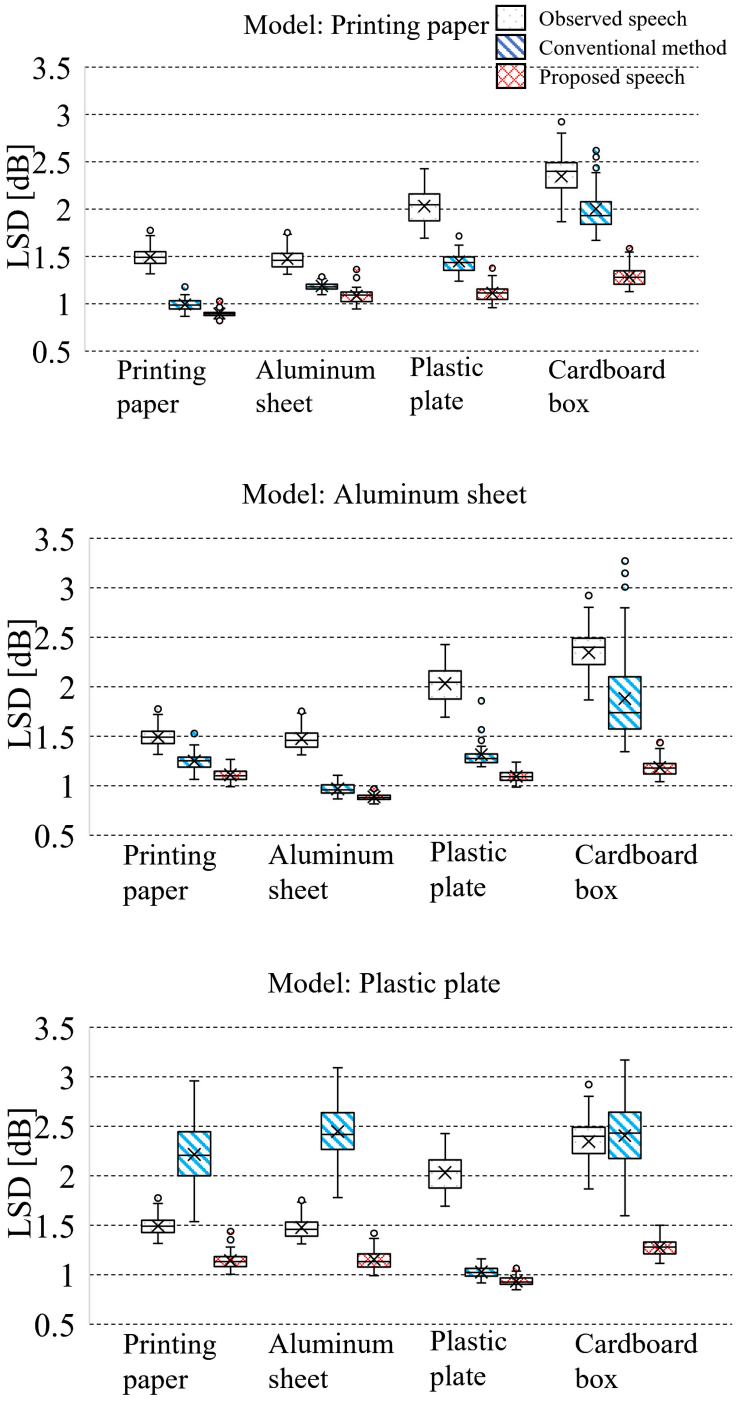


図 4.6 LSD 評価における提案手法の結果

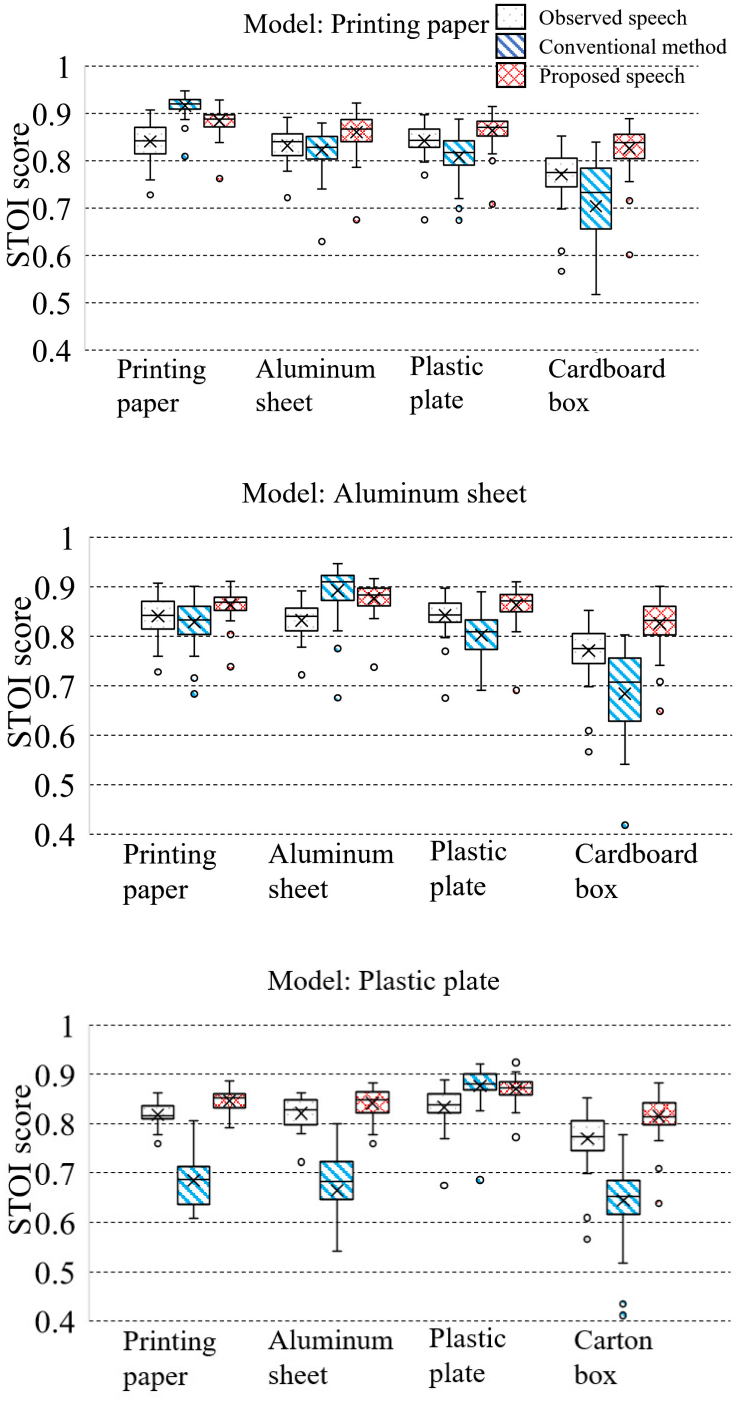


図 4.7 STOI 評価における提案手法の結果

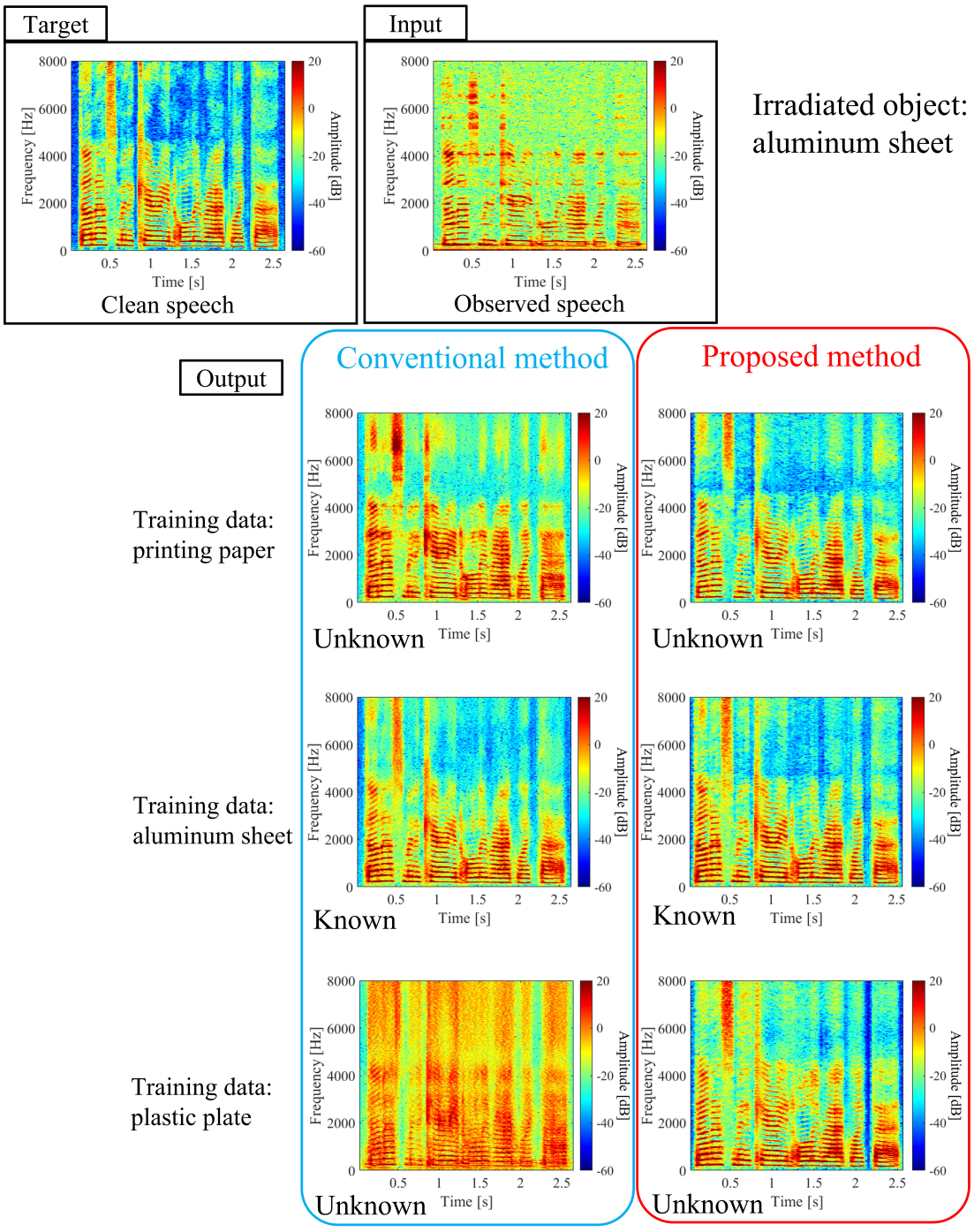


図 4.8 各モデルを用いたアルミ板による強調音声のスペクトログラム

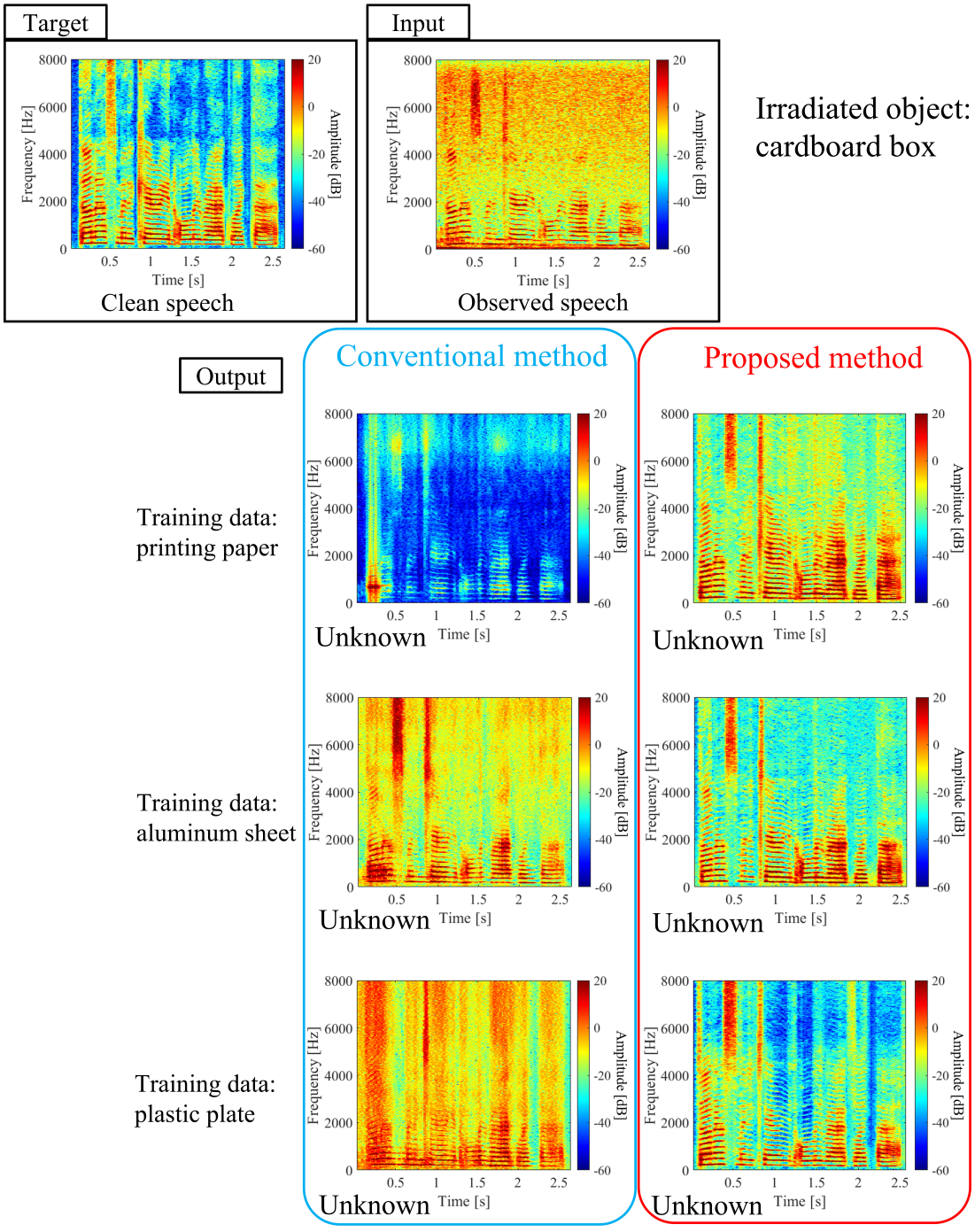


図 4.9 各モデルを用いたダンボール板による強調音声のスペクトログラム

第5章　結論

5.1. 本論文のまとめ

マイクは日常生活で最も一般的な音声取得デバイスである．従来のマイクは，空気を介して振動板に伝播した音波を電気信号に変換することによって音を収音する．音が空気中を伝播するにつれて，距離の増加とともにパワーの減衰が大きいため，遠方から収音することが困難である．そこで，遠方からの音を観測するため，レーザ光を用いたLDVにより収音する手法が提案された．レーザ光は強い直進性を持つため，遠方の音も感知できるが，その計測原理により歪みが発生し，収音した音声の劣化は大きい．高音質な音声をLDVにより取得するには，観測音声に対し，音声強調処理が必要となる．

近年では，音声強調手法として，DNNを利用する方法が数多く提案されている．しかし，これらの手法では多様な歪みを含むLDVの観測音声にそのまま適用することは困難であり，被照射物体が未知の場合，さらに対応が難しい．この問題を解決するために．2章では，まずLDVの原理と問題点について説明した．次に，従来の音声強調手法と従来手法の問題点を述べた．そして，本論文で用いられた音質評価基準も説明した．

3章では，被照射物体が既知 (特定) の場合において，2つの手法を提案した．提案手法1では，まず，DNNを用いて，収録音声のパワースペクトルを復元した．その後，復元されたパワースペクトルを用いて，収録音声とクリーン音声との位相差を推定することで音声を強調した．提案手法2では，まず，DNNを用いて収録音声の低域成分における雑音抑圧処理を行った．その後，得られた低域成分を用いて高域成分を推定することで音声を強調した．そして，ペットボトルを用いた観測音声を用いて，実験を実施することより，特定の被照射物体という限定的条件において，

提案手法の有効性を確認した．

4 章では，被照射物体が未知の場合においての音声強調手法を提案した．提案手法はロバスト性の高い特徴量を用いることで，いかなる被照射物体でも共通の特徴量を利用して音声強調を試みた．具体的には，まず，観測音声のピッチとスペクトル包絡情報を抽出する．そして，DNN を利用して振幅応答により劣化したスペクトル包絡を復元する．次に，復元されたスペクトル包絡とピッチを用いて強調音声のパワースペクトルを生成する．最後に，復元されたパワースペクトルを基に位相スペクトルを生成し，強調音声を合成する．評価実験の結果により，提案手法の有効性を示した．

5.2. 今後の課題

今後，被照射物体が既知の場合において，雑音抑圧，高域復元による二段階処理手法における損失関数による強調精度の違いについて検討し，高域成分の復元精度を向上させる予定である．また，本論文においては，ペットボトルを用いた観測音声のみの評価であったが，さらに計測対象の材質変化，収録音と LDV 間の距離の増加による収録音声の劣化を考慮した音声強調法を検討する予定である．また，被照射物体が未知の場合において，今後，物体の振幅応答と表面粗さを定量化し，ネットワークの設計時に振幅応答と表面形状を考慮し，より高精度化を追求する．

今後も引き続き研究を行い，さらに良い音質で LDV による観測音声を獲得することで，実用化につなげ，実社会に貢献できれば幸いである．

謝辞

本博士論文は，立命館大学大学院情報理工学研究科博士後期課程において筆者が行った研究の成果をまとめたものです．本研究を遂行するにあたり，学内，学外を問わず多くの方にお世話になりました．ここに深厚なる感謝の意を表します．

立命館大学情報理工学部西浦敬信教授には，筆者の本学在学中における研究活動を通じて多大なご指導を頂きました．西浦敬信教授には，博士課程後期課程入学から修了までの 5.5 年間，研究テーマの決定から最終論文の完成まで指導していただきました．学術的な研究だけでなく，職場や社会のルールからもいろいろ教えられました．また，今後の職場環境にスムーズに行けるように，企業とのプロジェクトなど各種の機会を与えられ，ここに心から感謝の意を表します．

同学部山下洋一教授には，筆者が本学在学の間，終始懇切丁寧なご指導を頂きました．山下先生は音，音声分野のエキスパートとして，筆者の研究過程において、特にテーマの選定において多くのアドバイスをいただき，本論文をより良い方向へ進歩させることができました．ここに深甚なる感謝の意を表します．

同学部谷口忠大教授には，本論文の副審査委員として本論文の執筆におけるご指導を頂きました．その上，筆者に R-GIRO プロジェクトに入る機会を頂き，心より深く御礼申し上げます．

同学部福森隆寛講師には，本研究の開始時に多くの関連知識を教えていただき，その後の研究をスムーズに進めることができました．その上，正課外のプロジェクト活動でも，ご指導をいただきました．ここに厚く御礼申し上げます．

同学部岩居健太講師には，毎週の研究進捗ミーティングや論文の執筆に至るまで，常日頃から懇切なる御指導，御助言を頂きました. 会議，ジャーナル，博士論文の執筆にあたり，有益なご助言を頂きました．心より感謝申し上げます．

個々には御名前を申し上げられませんが，筆者の研究上の議論に付き合っていた

だき，また筆者の至らない点を御援助頂きました立命館大学情報理工学部音情報処理研究室の多くの先輩，同期，後輩，秘書の皆様，そして多くの励ましを頂いた学内外の友人に心より御礼申し上げます.

最後になりましたが，深い愛情と広い心で今日まで筆者を支えて頂いた家族と友人に心から感謝いたします.

参考文献

[1] M. A. Clark, “An acoustic lens as a directional microphone,” Transactions of the IRE Professional Group on Audio, vol. AU-2, no. 1, pp. 5–7, January-February 1954.

[2] J. H. Shang, Y. He, D. Liu, H. G. Zang and W. B. Chen, “Laser Doppler vibrometer for real-time speech-signal acquirement,” Chinese Optics Letters, vol.7, no.8, pp.732–733, 2009.

[3] Q. Leclere and B. Laulagnet, “Nearfield acoustic holography using a laser vibrometer and a light membrane,” Journal of the Acoustical Society of America, vol.126, no.3, pp.1245–1249, 2009.

[4] A. Malekjafarian, D. Martinez and E. J. OBrien, “The feasibility of using laser Doppler vibrometer measurements from a passing vehicle for bridge damage detection,” Shock and Vibration, vol.2018, no.PT.5, pp.1–10, 2018.

[5] D. M. Chen, Y. F. Xu and W.D. Zhu, “Identification of damage in plates using full-field measurement with a continuously scanning laser Doppler vibrometer system,” Journal of Sound and Vibration, vol.422, pp.542–567, 2018.

[6] H. Aygün and A. Apolskis, “The quality and reliability of the mechanical stethoscopes and Laser Doppler Vibrometer (LDV) to record tracheal sounds,” Applied Acoustics, vol.161, pp.1–9, 2020.

[7] K. H. Li and C. H. Lee, “A deep neural network approach to speech bandwidth expansion,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, Australia, pp.4395–4399, Apr. 2015.

[8] T. Lotter and P. Vary, "Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-Gaussian speech modelling," 12th European Signal Processing Conference, Vienna, Austria, pp.1457–1460, Sept. 2004.

[9] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.22, no.12, pp.1931–1940, 2014.

[10] D. Rethage, J. Pons, and X. Serra. "A wavenet for speech denoising," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, pp.5069–5073, 2018.

[11] Z. Xie, J. Du, I. McLoughlin, Y. Xu, F. Ma and H. Wang, "Deep neural network for robust speech recognition with auxiliary features from laser-Doppler vibrometer sensor," 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, pp.1–5, Oct. 2016.

[12] W. H. Li, M. Liu, Z. G. Zhu and T. S. Huang, "LDV remote voice acquisition and enhancement," 18th International Conference on Pattern Recognition, Hong Kong, China, pp.262–265, Aug. 2006.

[13] R. Peng, B. Xu, G. Li, C. Zheng, and X. Li, "Long-range speech acquirement and enhancement with dual-point laser Doppler vibrometers," 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, pp.1–5, Nov. 2018.

[14] T. Lü, J. Guo, H. Y. Zhang, C. H. Yan and C. J. Wang, "Acquirement and enhancement of remote speech signals," Optoelectronics Letters, vol.13, no. 4, pp. 275–278, 2017.

[15] J.A. Bucaro and N. Lagakos, "Lightweight fiber optic microphones and accelerometers," Review of Scientific Instruments, vol.72, no.6, pp.2816–2821, 2001.

[16] J. Shang, Y. He, D. Liu, H.G. Zhang and W.B. Chen, “Laser Doppler vibrometer for real-time speech-signal acquirement,” Chinese Optics Letters, vol.7, pp.732–733, 2009.

[17] Y. Avargel and I. Cohen, “Speech measurement using a laser Doppler vibrometer sensor: Application to speech enhancement,” 2011 Joint Workshop Hands-free Speech Communication and Microphone Arrays, Edinburgh, UK, pp. 109–114, May 2011.

[18] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett and N. L. Dahlgren, “Acoustic-phonetic continuous speech corpus CD-ROM NIST speech disc 1-1.1,” NASA STI/Recon Technical Report N, LDC93S1, vol. 93, 1993.

[19] H. Ze, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, pp.7962–7966, Oct. 2013.

[20] L.H. Chen, Z.H. Ling, L.J. Liu, and L.R. Dai, “Voice conversionusing deep neural networks with layer-wise generative training,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.22, no.12, pp.1859–1872, Dec. 2014.

[21] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Voice conversion in high-order eigen space using deep belief nets,” Interspeech, Lyon, France, pp.369–372, Aug. 2013.

[22] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” Interspeech, Lyon, France, pp.436–440, Aug. 2013.

[23] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.1, pp.7–19, 2015.

[24] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, pp.7092–7096, May, 2013.

[25] F. Bolner, T. Goehring, J. Monaghan, B. van Dijk, J. Wouters and S. Bleeck, “Speech enhancement based on neural networks applied to cochlear implant coding strategies,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp.6520–6524, Mar. 2016.

[26] Y. Xu, J. Du, L.R. Dai and C.H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Processing Letters, vol.21, no.1, pp.65-68, 2014.

[27] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.32, no.2, pp.236–243, 1984.

[28] N. Perraudin, P. Balazs and P. L. Sondergaard, “A fast Griffin-Lim algorithm,” 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA, pp. 1–4, Oct. 2013.

[29] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura and H. Saruwatari, “Phase reconstruction from amplitude spectrograms based on von-mises-distribution deep neural Network,” 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, pp.286-290, Sept. 2018.

[30] D.S. Williamson, Y. Wang and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.24, no.3, pp.483-492, 2016.

[31] A.V.D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv:1609.03499, 2016.

[32] D. Rethage, J. Pons and X. Serra, "A wavenet for speech denoising," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, pp. 5069-5073, Apr. 2018.

[33] Y. Gu and Z.H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension," Interspeech, pp.1123-1127, Stockholm, Sweden, Aug. 2017.

[34] Y.Gu, Z.H. Ling, and L.R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," Interspeech, pp.297-301, San Francisco, USA, Sept. 2016.

[35] Z.H. Ling, Y. Ai, Y. Gu, and L.R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.26, no.5, pp.883-894, 2018.

[36] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on directional-statistics deep neural networks," Signal Processing, vol.169, pp.107368, 2020.

[37] I. Recommendation, "G. 711: Pulse code modulation (PCM) of voice frequencies," International Telecommunication Union, 1988.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, vol.2, pp.749–752, May 2001.

[39] P. Wang, Y. Wang, H. Liu, Y. Sheng, X. Wang and Z. Wei, "Speech enhancement based on auditory masking properties and log-spectral distance," 3rd International Conference on Computer Science and Network Technology, Dalian, China, pp. 1060–1064, Oct. 2013.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, USA, pp.4214–4217, Mar. 2010.

[41] Y.N. Dauphin, A. Fan, M. Audi and D. Grangier, "Language modeling with gated convolutional networks," International Conference on Machine Learning, pp.933–941, Sydney, Australia, Aug. 2017.

[42] D.P. Kingma, J. Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations, San Diego, USA, May 2015.

[43] K. He, X.Y. Zhang, S.Q. Ren and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," IEEE International Conference on Computer Vision, pp.1024-1034, Santiago, Chile, Dec. 2015.

[44] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," IEEE Conference on Computer Vision and Pattern Recognition, pp.105–114, Honolulu, Jul. 2017.

[45] P.J. Werbos, "Backpropagation through time: what it does and how to do it," IEEE, vol.78, no.10, pp.1550–1560, Oct. 1990.

[46] S. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, USA, vol.4, pp.200–203, Apr. 1979.

[47] R. Martin, "Spectral subtraction based on minimum statistics," Proceedings of EUSIPCO-94, Edinburgh, vol.6, no.8, pp.1182–1185, 1994.

[48] F. Villavicencio, A. Robel and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," 2006 IEEE Inter-

national Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, vol.1, pp. 869–872, May 2006.

[49] H. Y. Gu and S. F. Tsai, "A discrete-cepstrum based spectrum-envelope estimation scheme and its example application of voice transformation," International Journal of Computational Linguistics, vol.14, no.4, pp.363–382, 2009.

[50] Y. Shi, L. Han, L. Han, S. Chang, T. Hu and D. Dancey, "A Latent Encoder Coupled Generative Adversarial Network (LE-GAN) for Efficient Hyperspectral Image Super-Resolution," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–19, 2022.

[51] J. Gui, Z. Sun, Y. Wen, D. Tao and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3313–3332, 2023.

[52] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, " Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, " IEEE International Conference on Computer Vision, Santiago, Chile, pp.1026–1034, Dec. 2015.

研究業績

学術論文

1. **Chengkai Cai**, Kenta Iwai, and Takanobu Nishiura, “Speech enhancement based on two-stage processing with deep neural network for Laser Doppler Vibrometer.” Applied Sciences, vol.13, no.3:1958, pp.1–15, 2023, https://doi.org/10.3390/app13031958

2. **Chengkai Cai**, Kenta Iwai, and Takanobu Nishiura, “Speech enhancement for laser Doppler vibrometer dealing with unknown irradiated objects.” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E106.A, no.4, pp.647–656, 2022, https://doi.org/10.1587/transfun.2022EAP1036.

3. Takahiro Fukumori, **Chengkai Cai**, Yutao Zhang, Lotfi El Hafi, Yoshinobu Hagiwara, Takanobu Nishiura and Tadahiro Taniguchi, “Optical laser microphone for human-robot interaction: speech recognition in extremely noisy service environments,” Advanced Robotics, vol.36, no.5-6, pp.304–317, 2022, doi: 10.1080/01691864.2021.2023629.

国際会議

1. **Chengkai Cai**, Kenta Iwai, Takanobu Nishiura, Yoichi Yamashita, “Speech enhancement for optical laser microphone with deep neural network,” 2020 Asia-Pacific Signal and Information Processing Association Annual Summit

and Conference (APSIPA ASC), Auckland, New Zealand, pp. 449-454, Dec., 2020.

2. **Chengkai Cai**, Hiroki Shinndo, Koichi Terano, Shoji Ueda, Shunsuke Yamada, Kenta Iwai, Takahiro Fukumori, Takanobu Nishiura, and Yoichi Yamashita, “Pin spot sound capture with optical laser microphone,” 16th International Workshop on Acoustic Signal Enhancement, D9, Tokyo, Japan, Sep., 2018.

3. **Chengkai Cai**, Makoto Nakashizuka, Yosuke Sugiura, Takanobu Nishiura, Yoichi Yamashita, “Wind noise reduction with morphological noise power estimation,”Western Pacific Conference on Acoustics, New Delhi, India, Program ID: EN0/02, Nov., 2018.

大会発表

1. **Cai Chengkai**, 服部　新栄, 中静　真, 杉浦　陽介, “モフォロジカルフィルタによる雑音スペクトル推定に基づく風雑音除去,”電子情報通信学会ソサイエティ大会, A-8-12, Sep., 2017.

2. **Cai Chengkai**, 福森　隆寛, 西浦　敬信, 山下　洋一, “光レーザマイクロホンのための Residual Network を用いた CNN による雑音除去,”日本音響学会 2020 年春季研究発表会, pp. 343-344, Mar., 2020.

3. **Cai Chengkai**, 岩居　健太, 西浦　敬信, 山下　洋一, “光レーザマイクロホンのための深層学習による多段階音声強調の検討,”日本音響学会 2020 年秋季研究発表会, pp. 339-340, Aug., 2020.

4. **Cai Chengkai**, 岩居　健太, 西浦　敬信, 山下　洋一, “パワースペクトルを用いた GAN による光レーザマイクロホンのための音声強調,”日本音響学会 2021 年春季研究発表会, pp. 421-422, Feb., 2021.