氏　　　名	DUANGPUMMET, Suradej
学位の種類	博士（情報科学）
学位記番号	博情第 459 号
学位授与年月日	令和 3 年 12 月 24 日
論文題目	Blind Estimation of Room Acoustic Parameters and Speech Transmission Index Based on the Concept of the Modulation Transfer Function
論文審査委員	主査　鵜木　祐史　北陸先端科学技術大学院大学　教授
	赤木　正人　同　教授
	党　建武　同　教授
	吉高　淳夫　同　准教授
	KARNJANA Jessada　NECTEC　研究員
	KONGDRAWECHNON Waree　タマサート大学 SIIT　准教授

論文の内容の要旨

Assessment of the quality of auditory spaces is essential in room acoustics and speech signal processing. In room acoustics, the sound characteristics of auditoriums are related to the quality of life in different aspects. In emergency circumstances, e.g., earthquake or flood, emergency announcements and alarm sounds need to be easily audible and intelligible so that we can follow the safety procedures appropriately. In theaters or concert halls, excellent sound characteristics and superior acoustics are ideal environments for performances. The sounds of live performances should be clear and transparent so that attendees can enjoy the entertainment. Additionally, speech signal processing, such as speech dereverberation and noise suppression, would benefit in which the sound quality and speech intelligibility can be improved based on the room characteristics.

Intelligibility of speech and pleasure to music are subjective descriptions. It is difficult to convey such descriptions from listeners to architects who are responsible for designing auditoriums or diagnosing acoustic problems. Conventionally, speech intelligibility and sound quality can be determined by conducting listening experiments with a group of listeners. Unfortunately, the experiments are expensive, unreliable, and time-consuming. It is also impractical for real-time applications, such as hearing aids, automatic speech recognition, and speaker verification. Thus, the quality of a sound field and subjective aspects are defined through room acoustic parameters and objective indices related to the physical properties of a sound field. Hence, architects, acousticians, and signal processing algorithms, can justify acoustic conditions by measuring acoustical parameters.

Several useful room acoustic parameters and objective indices have been standardized. In IEC 60268-16 : 2020, the speech transmission index (STI), which is an objective index, is used to predict speech intelligibility from the quality of a speech transmission channel. The STI is calculated based on the concept of the modulation transfer function (MTF). The MTFs of seven-octave bands with their weighting values are converted to be a real number from 0 to 1. In addition, ISO 3382: 2009, specifies methods for measurement the reverberation times (T_{60} or T_{30}) and other room-acoustic parameters, including early decay time (EDT), clarity (early-to-late-arriving sound energy ratios: C_{80} or C_{50}), Deutlichkeit (early-to-total sound energy ratio: D_{50}), and center time (T_s). These parameters are derived from measuring the room impulse response (RIR).

In the time domain, an RIR completely describes the characteristics of a sound field. Similarly, a system transfer function in the frequency domain and the MTF in the modulation-frequency domain are the counterparts. In general, the RIR or MTF needs to be measured. However, it is difficult to measure RIR or MTF in daily-life places where people cannot be excluded, e.g., public stations, airports, and department stores. Moreover, by the nature of such public areas, room acoustics are prone to be a time-varying system. Sound absorption, reverberation, or other acoustical parameters are changed by varying occupants and object arrangements. Thus, acoustic parameters that were measured complying with the standards might be different from the current one. Hence, many methods have been proposed to estimate an acoustical parameter without measuring the RIR, known as *blind* estimation methods.

The blind estimation of an acoustical parameter is an ill-posed condition because both sound source and RIR are unknown. The ill-posed or blind inverse problem is challenging since it needs additional assumptions or complementary prior knowledge to formulate the estimation. Furthermore, the robustness of the estimator against various rooms (e.g., diffuse/non-diffuse field and connected chamber) and background noise need to be taken into account. To this end, this research presents blind estimation methods for estimating five-room acoustic parameters, STI, and SNR from a speech signal in noisy reverberant environments using a single-channel microphone and the concept of the MTF.

A speech signal can be decomposed into a fine structure and temporal structure. For temporal structure, a power envelope (PE) or temporal amplitude envelope (TAE) is used as a feature. On the basis of the MTF, PE or TAE represents the modulation distortion caused by reverberation and noise of the transmission channel (sound field). The TAE also plays an important role in speech intelligibility. In the proposed scheme, these features are extracted from an observed signal by using Hilbert transform and a low-pass filter. An observed signal in a given room is regarded as the output of the convolution between the RIR and speech signal. Hence, the modulation features (TAE/PE) and the convolution operation using one-dimensional convolutional neural networks (CNNs) were deployed. A more sophisticated deep neural network (DNN), such as a combined network between CNNs and long short-term memory (LSTM) networks, was also utilized. These DNNs were trained from the pairs of TAE/PE and the parameters of RIR models. In addition, data augmentation techniques were used for synthesising the dataset due to limited measured RIRs.

Here, an unknown RIR is modeled by using a stochastic RIR model. Two RIR models were investigated: Schroeder's RIR model and the extended RIR model. The reverberation time is an only parameter in Schroeder's RIR as a simple exponential decay (T_R). The extended RIR model is an extended version of Schroeder's RIR model. It consists of three parameters, including rising parameter (T_h), peak position (T_0), and exponential decay parameter (T_t). Thus, the extended RIR model is much more accurate and flexible. Here, the parameter T_R in Schroeder's RIR and the three parameters of the extended RIR model are blindly estimated. Sub-band analysis is used as the same as the algorithm for calculating the STI. The distortion in seven-octave bands is estimated through the parameters of the RIR model. The approximated RIR for each sub-band can be reconstructed from their envelope modulated with band-limited noise. The wide-band RIR is also approximated from the summation of the sub-band signals based on the superposition principle. Therefore, the estimated acoustical parameters and STI for both sub-band and wide-band can be derived.

The effectiveness and performance of the proposed methods were evaluated. Simulations were performed by estimating the parameters from reverberant and noisy reverberant speech signals. The accuracy of the estimated acoustical parameters was compared with baselines calculated from measured RIRs and existing works. The robustness against various background noise was also evaluated by adding four types of noise with different SNR levels into the reverberant speech signals. The experimental results suggest that the proposed method can correctly, blindly, and simultaneously estimate five-room acoustic parameters, STI, and SNR from a speech signal in reverberant and noisy reverberant environments. The accuracy in terms of standard derivation of the error of the estimator for each parameter, i.e., T_{60}, EDT, C_{80}, D_{50}, T_s, and STI, was 9.4%, 10.5%, 2.7 dB, 14%, 45 ms, and 0.05, respectively. These results of the estimated parameters were close to the standard measurement derived from the RIR.

Keywords: room impulse response, speech transmission index, blind parameter estimation, modulation transfer function, convolutional neural networks.

論文審査の結果の要旨

音バリアフリーやユニバーサルデザインの志向から，利用者に係わらず円滑な音声コミュニケーションを実現するための要素・統合技術の確立が急務となっている．そのため，音環境設計における質的な検討（音声の明瞭性の向上，誘導音の検知力向上など）は，最重要課題である．室内音響の設計を行なう際，古典的にみると最適残響時間の利用が主であったが，現在は室内インパルス応答から求められる音響指標を利用することが主となっている．ISO 3382-1 では，残響時間（T60）や音の透明性(C 値)，音の明確性（D 値）があり，IEC 60268-16 では，音声伝送指標（STI）がある．これらは，音声了解度や聴き取り難さといった室内音響の「質」を主観的に評価するのに重要なものであり，音バリアフリーを目指した室内音響設計では非常に重要な検討項目である．しかし，室内インパルス応答の実測を前提とするため，聴力保護等の目的から人を排除して，人がいない室の特性が比較的安定しているうちに測定するか，あるいは複数回測定した後でその平均値を実測値として利用しなければならない．そのため，人を排除できないような音環境や時々刻々変化するような音環境（駅構内といった公共環境）には，既存の手法を適用することができない．

本研究では，これらの問題を解決するために，室内インパルス応答を測定せずに，観測された信号そのものから室内の変調伝達関数（MTF）を逆推定することで，室内音響パラメータ（T60，STI，C 値，D 値）や STI を逆推定する方略を明らかにした．特に，観測された信号の時間包絡線情報に着目し，(1) 室内の MTF の逆推定に必要な統計的室内インパルス応答のパラメータ推定法と(2) 背景雑音といった外乱に頑健な室内音響パラメータ・STI の推定法を確立した．これらの方法は，MTF の概念に基づくものであり，推定法の各要素技術を畳み込みニューラルネットワークといった機械学習ベースの方法で実現した．推定法のリアルタイム処理を実現するための枠組みが提案され，準リアルタイム処理としてプロトタイプも実装された．

以上，本論文は，音環境設計における質的な検討の一つとして，室内音響パラメータならびに STI をブラインドで正確に推定する革新的技術を提供した．本技術は，公共空間における音声アナウンスや非常時災害情報の効果的な提供のための音環境の性能予測，仮想音響空間における音声伝送の評価といった重要課題に対して利用可能であり，応用範囲が広く，学術的に貢献するところも大きい．よって博士

（情報科学）の学位論文として十分価値あるものと認めた．