氏　　名　Li Haoyu

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2368 号

学位授与の日付　2022 年 9 月 28 日

学位授与の要件　複合科学研究科　情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Improving Neural-Network-Based Speech Enhancement for Noise Reduction and Intelligibility Boosting

論 文 審 査 委 員　主　査　山岸　順一
情報学専攻　教授
越前　功
情報学専攻　教授
Yu Yi
情報学専攻　助教
戸田　智基
名古屋大学 情報基盤センター　教授
亀岡　弘和
日本電信電話株式会社
NTT コミュニケーション科学基礎研究所
上席特別研究員

(Form 3)

Summary of Doctoral Thesis

Name in full : Li Haoyu

Title : Improving Neural-Network-Based Speech Enhancement for Noise Reduction and Intelligibility Boosting

Speech is the main media used by humans to communicate in daily life. However, in real-life world, environmental factors such as noise, reverberation, and poor microphone characteristics, inevitably degrade the speech signals. In application scenarios such as telephony and teleconference, speaker and listener often locate in different physical places with different surrounding ambient noises, which further degrades the quality of voice communication experience. Speech enhancement techniques, which aim to improve the quality and intelligibility of speech, therefore play an indispensable role in real-world speech communication scenarios.

Depending on the usage scenario, this thesis further divides speech enhancement into two sub-tasks: (1) noise reduction and (2) intelligibility boosting. More specifically, noise reduction is supposed to work in the speaker side, where the microphone receives the mixture signal of speaker's voice and environmental background noise (such as additive background noise or multiplicative reverberant noise). The goal of noise reduction is thus to suppress noise and recover the clean speech signal from the noisy mixture input. On the other hand, intelligibility boosting is supposed to work in the listener side, where noise reduction techniques cannot work since noise sources are physically present near the listener. Therefore, instead of suppressing noise, the goal of intelligibility boosting is to only modify the speaker's speech signal to improve its intelligibility when exposed to the background noise and reverberation.

This thesis focuses on the neural network-based speech enhancement. First, this thesis lists three limitations of conventional noise reduction systems: (1) As a data-driven model, the generalization ability to the mismatched noise conditions is not satisfactory; (2) Many noise reduction systems mainly operate on spectrogram magnitude, while directly incorporating noisy phase results in speech quality degradation; and (3) Many existing works focus on additive or reverberant noise, but lack consideration on general device degradation, such as bad frequency characteristics of the common consumer-grade recording devices. This thesis tackles these limitations in three aspects. For limitation (1), this thesis proposes the noise token module, which is composed of a set of trainable neural noise templates, to dynamically encode the noise information and thus enrich the model's generalization. For limitation (2), this thesis uses neural vocoder-based waveform generation module to directly generate speech signals from the predicted mel-spectrogram, which avoids the use of noisy phase. For limitation (3),

this thesis directly models the joint degradation effect of common device-degraded speech, which includes not only additive noise but also reverberation and the bad frequency response of microphone. This thesis proposes an encoder-decoder neural network to enhance device-degraded speech, in which the encoder first filters out the channel characteristics of input speech and then the decoder predicts the target high-quality mel-spectrogram, and the final speech waveform is synthesized by using a neural vocoder. For above-mentioned limitations, experimental results showed that the proposed methods well suppress noise and produces high-quality speech waveform.

Next, this thesis proposes a novel neural network-based system for intelligibility boosting, in which a neural surrogate model is introduced to jointly optimize multiple intelligibility and quality metrics. Compared to the traditional signal processing-based approaches, this new system can be automatically built from speech data without relying on any expert knowledge. Specifically, a generative adversarial network framework is introduced to simultaneously optimize multiple advanced speech metrics, including both intelligibility- and quality-related ones, which results in notable improvements in performance and robustness. This system can not only work in non-real-time mode for offline audio playback but also support practical real-time speech applications. Experimental results using both objective measurements and subjective listening tests indicated that the proposed system can lead to significant intelligibility gains and perform much better than the state-of-the-art signal processing-based baseline method.

Finally, this thesis combines noise reduction and intelligibility boosting techniques to address the full-end speech enhancement task where both speaker and listener environments are noisy. This thesis investigates a novel joint framework, in which noise reduction module first suppresses noise and the intelligibility boosting module then modifies the denoised speech, i.e., the output of the noise reduction module, to further improve speech intelligibility. Such a model can fully benefit from the powerful modeling capabilities of neural networks and get rid of unnecessary assumptions. Experiments showed that this joint model significantly improves speech quality and intelligibility and clearly outperforms the disjoint pipeline methods.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　Li Haoyu

Title
論文題目　　Improving Neural-Network-Based Speech Enhancement for Noise Reduction and Intelligibility Boosting

本学位論文は、「Improving Neural-Network-Based Speech Enhancement for Noise Reduction and Intelligibility Boosting」と題し、全７章で構成されている。本論文は、話し手と聞き手が別の環境に存在するコミニュケーションチャネルにおいて、話し手側の雑音を除去（音声強調）し、同時に聞き手側の環境に適した音声へ変換（音声明瞭性強調）する枠組みに関して研究成果をまとめたものである。第１章では、本論文で扱う問題の重要性、位置付けおよび貢献について説明がなされ、第２章では音声強調および音声明瞭性強調の従来手法の説明がなされた。第３章では音声強調に着目し、未知の雑音に対する汎化性能を改良させる方法、および、雑音により劣化した位相スペクトルを利用せずにニューラルボコーダにより音声波形を合成する方法が提案された。第４章では加法性雑音に加え、反響や収録機器機の影響による劣化も考慮したモデルが提案された。第５章では音声明瞭性強調のために、ニューラルネットワークにより近似された音声明瞭性指標を最大化する様に音声変換を行うモデル構造が提案され、その有効性が実験から示された。第６章では、話し手側の雑音除去モデルと聞き手側の音声明瞭性強調モデルとを連結した手法が提案され、これらを同時最適化する事でより良い結果が得られる事を実験的に示した。７章ではまとめと今後の課題を述べている。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会および口述試験では、審査員からの質疑に対して適切に回答がなされた。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では出願者の博士研究に学位論文として十分なレベルの新規性や有効性があると評価した。

以上を要するに、本学位論文は、音声強調および音声明瞭性強調分野を学術的に発展させる内容であると同時に、音声情報処理を利用したサービス等にも直結する内容であり、その科学的貢献は大きいと言える。また本学位論文の成果は学術雑誌論文１編、国際会議論文４編として発表され、社会的な評価も得ている。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。