氏名	Elhard James Kumalija
学位の種類	博士（応用情報科学）
学位記番号	博情第 73 号
学位授与年月日	令和 5 年　9 月　27 日
学位授与の要件	学位規則第 4 条第 1 項該当（課程博士）
論文題目	A Study of the Effect of Noise and Network Distortions in VoIP Speech Signals
論文審査委員	（主査）准教授　大島裕明 （副査）教授　原口亮 （副査）准教授　栗原淳 （副査）名誉教授　中本幸一

学位論文の要旨

現在、Voice over IP（VoIP）は、多くのサービスで広く利用されている技術である。VoIP アプリケーションでは、インタラクティブ音声応答や VoIP 通話における会話などにおいて、環境ノイズだけでなく伝送ネットワークのエラーやエンコード・デコードアルゴリズムによる歪みにさらされ、音声信号の劣化が起こってしまう。

環境ノイズとネットワークの歪みが VoIP アプリケーションに与える影響は、そのアプリケーションが持つ性質に依存する。たとえば、公共放送のようなアプリケーションでは、ユーザが受信する音声の品質が問題となるだろう。その場合、送信された音声の連続的な自動品質評価を行うことが必要となる。音声品質のモニタリングを行うことによって、VoIP アプリケーションは環境ノイズレベルの変化やネットワーク状況の変更に適応するということが可能になるだろう。他方で、VoIP 通話やの対話型音声応答のようなアプリケーションでは、音声認識システムの性能が問題となるだろう。また、受信環境が高性能なコンピュータであったり、モバイルフォンのようなメモリや計算能力が低い環境であったりするなど、VoIP の利用環境は多様なものであるといえる。

本研究では、環境ノイズとネットワークの歪みの双方を考慮して、VoIP アプリケーションにおいて、自動音声認識と音声品質評価に与える影響を検討することを目指した。特に、高い頑健性を持つ自動音声認識手法と音声品質予測手法について、それぞれ、深層学習を利用したモデルの開発を行った。

はじめに、機械学習モデルを訓練するためのデータについての分析と新しいデータの作成についての提案を行った。クリーンな音声によるデータセットと、環境ノイズとネット

ワークの歪みの双方を含んだ音声によるデータセットを用意した。それぞれで訓練された音声認識システムについて、比較の分析を行った。環境ノイズとネットワークの歪みの双方を含んだ音声によるデータセットで音声認識モデルを訓練することで、様々な種類のノイズにきちんと対応して、音声認識を行う能力が向上することが明らかとなった。クリーンな音声によるデータセットで訓練された音声認識モデルの性能は、ノイズの種類に依存して、うまく対応できることもあれば、うまく対応できないこともあるということが明らかとなった。

環境ノイズとネットワークの歪みの双方を含んだ音声によるデータセットで訓練されたモデルは、クリーンな音声で訓練されたモデルと比較して、単語誤り率（WER)、単語一致率（MER)、単語情報損失（WIL）において改善が見られた。また、ある程度のジッタやパケットロスにおいて、認識性能の低下が押さえられるという頑健性を持つことが示された。ジッタやパケットロスが増えてきて、性能が低下する場合においても、クリーンな音声でトレーニングされたモデルと比較して、性能の低下が押さえられることが明らかとなった。

次に、MiniatureVQNet という単一エンドの音声品質評価手法について提案した。提案手法は、軽量な深層ニューラルネットワークモデルに基づいており、リソースの限られた環境においても実行可能である。この音声品質評価手法については、環境ノイズとネットワークの歪みの双方を含んだ音声によるデータセットによって訓練されたモデルと、環境ノイズのみを含んだ音声によるデータセットによって訓練されたモデルを構築し、それらの比較を行った。総じて、様々な環境ノイズやネットワーク状況において、従来手法よりもよい性能が得られた。訓練に用いたデータによる違いについては、環境ノイズとネットワークの歪みの双方を含んだ音声によるデータセットを用いた場合の方が、様々な VoIP 環境において性能が向上するということが示された。

論文審査の結果の要旨

本論文で取り組まれている問題は、Voice over IP（VoIP）における音質劣化の要因として、環境ノイズとともに、ネットワークを要因とする音の歪みがあるということである。VoIP の様々なアプリケーションにおいて、その 2 つの音質劣化の要因を考慮した環境適応が必要となってきている。深層学習などを用いた新しい音声認識技術が発展する中、本論文では、それら双方の要因に対応する音声認識問題に取り組まなくてはいけないということを明らかにした。そして、その具体的な研究課題として、2 つの課題に取り組んでいる。1 つ目の研究課題は、双方の要因を考慮した音声データを作成できる仕組み

を構築し、それを用いて得られたデータによって自動音声認識システムを構築するということである。深層学習技術を用いた音声認識技術においては、データセットの構築が非常に重要になる。提案するデータセット構築手法によって、先述した音質劣化の双方の要因に対してより頑健に認識を行うことができる手法を実現することができるということを明らかにした。2つ目の研究課題は、VoIPアプリケーションにおける音声品質のリアルタイム認識を行う手法である。これまでに、すでに、深層学習を用いて音声品質を認識する手法が提案されてきているが、リアルタイム性を追求するために、これまでの手法よりも軽量に音声品質を認識することができる手法を提案した。

本博士論文では、まず、1 章において、本研究の位置づけを行い、続いて、2 章において関連する研究についての紹介を行っています。3 章と 4 章では、深層学習モデルの学習を行うためのデータセットについて述べている。既存のデータセットの紹介と、それらを用いることの問題点を明らかにするとともに、環境ノイズとネットワークを要因とする音の歪みの双方を考慮した音声認識のためのデータセットを構築する手法について説明している。5章と6章では、VoIPアプリケーションにおける音声品質のリアルタイム認識を行う手法の提案と、その手法についての評価について述べている。7 章では、先述したデータセットを用いた自動音声認識手法の構築とその評価について述べている。8 章で、研究のまとめを行っている。

VoIP アプリケーションは昨今のオンラインミーティングの活用の広がりなどを考慮しても、今後、ますます重要性が増すと考えられる。そのような社会情勢の中で、本研究は、問題の所在を明らかにするとともに、それらの問題に対して一定の効果を持つ手法を提案しており、社会や産業における進展に貢献すると認められる。以上を総合して、本審査委員会は、本論文が博士（応用情報学）の学位授与に値するものと全員一致で判定した。