A-031

音声判定法によるプライバシー保護と発話区間認識の高精度化
Enhance Accuracy for Privacy-Sensitive Communication Detection Technology Using Locally Time-Reversed Speech

河合 夏美† 田﨑 誠人‡ 三木 良雄†
Natsumi Kawai Masato Tasaki Yoshio Miki

1. はじめに

著者らは、西新宿地区の公開空地を対象とし、資産価値向上の取り組みを行っている。資産価値、すなわち生産性のある場所として機能するためには、コミュニケーションの存在が不可欠となる。そして、公開空地における賑わいを定量化するためには会話の測定が不可欠となる。

一方で、コミュニケーションの実態を把握する際には、個人情報およびプライバシー保護の観点を慎重に考慮する必要がある。一般に個人情報とは、氏名や住所、生年月日など、特定の個人を識別できる情報を指す。日常的な会話の多くはこの定義に直接当てはまらないが、内容に個人を特定できる情報が含まれる場合、それは個人情報として扱うべきである。さらに、個人情報に該当しない場合でも、個人の私的情報が含まれる可能性もあり、無断で収集・分析を行うことはプライバシー侵害のリスクを伴う。その点に関して、音を短い区間に分け、局部時間反転音声の技術を使用して音声の秘匿化が提案されているが，音声区間検出精度に限界があった。

本研究では、提案されている手法に対して、セキュリティと音声区間検出精度とのトレードオフを考慮し、音声区間検出精度向上の方法を提案する。

2. 本研究のアプローチ

音声データの加工と認識容易性については、電話等の通信回線品質分野で永らく研究されている[1]。通信回線の場合には伝送中に音声データがどこまで悪化しても人間が内容を認識可能かという観点の研究であるが、その成果として、音声の一部分を時間的に反転したとしても、その部分が短時間であると人間の認識能力により、内容を正しく認識できることが知られている。反転させる区間は局部時間と呼ばれているがその時間はおよそ 100ms であるといわれている。つまり、100ms 以上の局部時間を音声反転させると、人間には内容が認識できないということになる。

この観点から、従来手法[2]では 100ms から 199ms の局部時間をランダムに決定し、その区間の音声を時間反転していた。この手法では隣接している区間の接している音声が、元のデータでは無音であったとしても、それは反転しているからであって、実は発話している部分である可能性がある。特に反転区間が長くなると発話としては別のものが反転によって近づく可能性が高くなり発話区間の誤認識につながる。

以上述べた従来研究の結果から、本研究では反転区間を短くしつつも、人間にとって内容が認識できない方法を確立する。具体的には時間反転区間は短時間とし、その区間の順番を改めて変更することを考える。

3. 提案技術

本研究による提案手法では次に示す手順により音声秘匿化を実現する。

(1) マイクユニットからの音声を A/D 変換によりデジタル化する
(2) デジタル化した音声をバッファに格納する。このバッファに格納する音声の時間を区間時間 1 とする
(3) 次にバッファ内のデータを区間時間 2 で区切り、この区間のデータを時間反転する
(4) #3 の手順で作成した複数の時間反転されたデータを区間時間 3 の範囲でグループを作成する
(5) #4 で作成したグループ内に含まれる反転音声をランダムに入れ替える

ここで、区間時間 1 ＞区間時間 3 ＞区間時間 2 という関係を持たせており、特に区間時間 2 は 200ms よりも短い時間とする。区間時間 3 を先行研究[2]の音声反転区間よりも短くすることで、先行研究の誤判定可能性を低減する。

続いて、システム全体としては、内容秘匿化されたデータを Demucs（Meta Research）や Adobe Audition（Adobe）といった音源分離技術を用い、人の声（vocal）の抽出を行う。抽出された区間を会話区間とみなすことで、原音を直接扱うことなく、会話の有無や活動区間の推定が可能となる。図 1 に、本提案手法による音声秘匿化の概要図を示す。

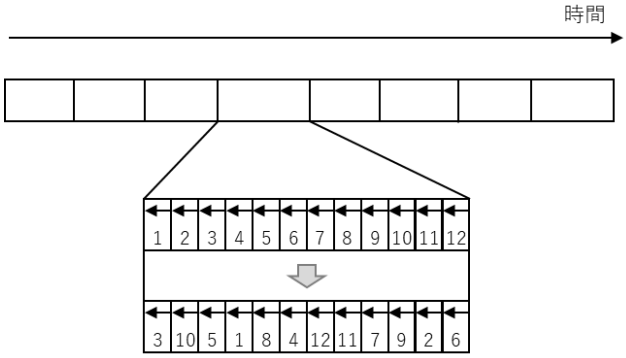

図 1：提案手法の概要図

4. 単語了解度の評価

提案手法が人間にとって認識困難な音声であることを確

† 工学院大学 情報学部 システム数理学科 Kogakuin University, Department of Information Science, Schools of Information.
‡ 工学院大学大学院 工学研究科 情報学専攻 Informatics Program, Kogakuin Graduate School of Engineering.

認するため、単語了解度を測定実験を実施した。

2 者による 45 秒の対話音声を、提案手法により局部時間反転処理を行った。なお、区間時間 3 を 300ms、区間時間 2 の中の時間反転した区間数(n)を 5 とした。

単語了解度の測定には 18-24 歳の男女 23 名を被験者とした。事前に、日本語の音声を正確に聞き取る能力があることを確認した上で、評価を実施した。評価対象となる音声データ（2 者による 45 秒の対話）から「の」や「を」といった助詞を除外し、意味のある単語のみを抽出した結果、対象単語数は 57 語となった。単語了解度は式 1 により算出した。

$$\text{単語了解度} = \frac{\text{回答者が正確に答えた単語の総和}}{\text{回答者数} * 57} \cdots \text{式1}$$

式 1 に基づき提案手法における単語了解度を測定した結果、その値は 0.008 であった。すなわち、認識された単語は全体の 1%未満であり、提案手法を用いて局所的に時間反転された音声を聴取した場合、会話全体の理解は極めて困難であることが示唆される。

さらに、正確に単語を認識できた被験者は 23 名中 3 名のみであり、これら 3 名はいずれも同一の約 4 単語を認識していたことが確認された。第 3 章の提案手法(5)で述べたとおり、反転音声は単語をランダムに入れ替える方式を採用しているため、単語の認識率はランダム化された順序によって影響を受ける可能性がある。しかし、単語了解度が 0.008 という極めて低い値であったことからも明らかなように、認識可能な被験者は限定的であり、仮に一部の単語が認識されたとしても、会話内容全体を把握することは極めて困難であると考えられる。

5. 結果

反転処理を施していない音声データを正解データとし、反転処理後の音声から検出された会話区間の精度を、Recall(再現率)，Precision(適合率)，F1-Score，False Positive Rate(FPR,誤検出率)，および Accuracy(正解率)の 5 つの指標で評価した。

音声データとして、15 分程度のオンライン会議音声に、環境音（歌詞のない音楽）を合成したものを用いた。提案手法による音声秘匿化の前後で、Demucs v4 を用いて人の声(Vocal)と環境音に分離した。

音声信号の振幅が一定の閾値以下の区間を無音と判定し、これを超える区間を会話区間として抽出した。会話区間の最小長を 300ms と定義し、短い無音の区間を発話の一部として扱うことで、検出の連続性を向上させた。また、反転処理による誤差を考慮し、反転処理有音声の無音区間および会話区間の判定において±0.2 秒の誤差を許容した。これにより、時間反転処理が検出精度に与える影響を適切に評価しつつ、検出結果の一貫性を確保することが可能となる。

本提案手法において会話区間検出を実施した結果、Recall が 1.0、Precision が 0.96，F1-Score が 0.98，FPR が 0.09，Accuracy が 0.97 という高い精度を得た。

これらの結果を従来手法の結果と比較したものを表 1 に示す。また、提案手法による混同行列を表 2 に示す。

表 1 ：従来手法と提案手法における検出性能の比較

指標	従来手法	提案手法
Recall	1.00	1.00
Precision	0.93	0.96
F1-Score	0.96	0.98
FPR	0.19	0.09
Accuracy	0.94	0.97

表 2：提案手法の混同行列

(単位：秒)		予測値	
		Positive	Negative
真値	Positive	688.0	0
	Negative	27.6	266.0

表 1 より、Recall が従来手法・提案手法ともに 1.00 であり、いずれも非常に高い精度で会話区間の検出漏れがなかったことがわかる。Precision についても、従来手法の 0.93 から提案手法の 0.96 と向上したことから、反転処理有音声から検出された会話区間のうち誤検出の割合が低減されたことが確認された。また、FPR も 0.19 から 0.09 に減少し、無音区間に対する誤検出率が 19%から 9%に低減された。これらの結果から、提案技術は会話区間の検出漏れを防ぎつつ、誤検出を抑制し、全体の検出精度を向上させたことが示された。

6. 結論と課題

本研究では、ランダムな区間長で局部時間反転処理を施し、さらに時間反転区間を無作為に並び替えることで、従来手法と比較して会話区間の検出精度が向上することを明らかにした。

精度向上の要因としては、従来手法では反転区間が 100-199ms と長いため、音声と雑音が同一区間内に混在しやすく、局部時間反転によって両者が混ざった不自然な音声となりやすかった。この結果、音源分離処理において会話音声と雑音の判別が困難となり、誤検出が増加し精度低下の一因となっていたと考えられる。一方で、本提案手法では、より短い区間で局部時間反転し、さらにそれらを一定区間内でランダムに並び替えることで、音声と雑音が 1 つの区間に混在しにくくなった。このことにより、音源分離の際に会話と環境音がより正確に分離され、全体として精度が向上したと考えられる。

今後の課題として、反転区間長や分割数などのパラメータについて最適化を図る必要がある。また、実際に公開空地で収音実験を実施し、多様な環境音が存在する状況下で本手法の有効性と検出精度を検証することが今後の重要な課題である。

参考文献

[1] Ueda, K., et.al, ” Intelligibility of locally time-reversed speech: A multilingual comparison”, Scientific Reports, Vol.7, no.1, pp.1-8, 2017.
[2] 田﨑誠人, 三木良雄, “局部時間反転音声によるプライバシーに考慮したコミュニケーション検出技術” , 情報処理学会第 87 回全国大会, pp.391-392, 2025.