未定稿

「首相秘書」AI 音声の真偽は情報幾何で弁別可能
・・・fake 規制に正しい科学的根拠と分別を！

伊東　乾
（東京大学大学院情報学環教授）

6 月 5 日（金） の参議院予算委員会）で、あろうことか首相答弁の一部として「通常の音声録音」と「機械学習システムによる合成音声」の「判断がつかず」

「これはどう考えても　確認のしようがない」

という発言があったと報じられています。

政治的な背景のある事案ですが、ここでもし、本当に高市氏が「確認しようがない」と思っているのであれば、かなりＡＩの基本リテラシーが不足していると、当該分野の国立大学教官として指摘せねばなりません。

末尾にも記しますが、まがりなりにも国会での首相答弁であれば、法に準ずる拘束力で我が国を縛る可能性（懸念）がありますので、明瞭に指摘しておきます。

自然な録音と合成音は適切な解析演算を実行すれば明瞭に弁別がつきます。

本稿は具体的な根拠の例とともに、これを広く公衆にお伝えし、未来を危ぶむことのないようリテラシー向上に資したいと考えるものです。

Fake の画像や音声が大きな社会問題となり、それに対する法制度整備も議論される国会で、初心者水準のリテラシーを欠く議論がまかり通ったりすれば、国難と言わねばなりません。

重ねて指摘するなら、それを追求するはずの野党も、また大手マスメディアから YouTube の独自動画まで、およそごく普通の背景となる科学も技術もすっ飛ばした談義に終始しているのは、極めて危険な状況と指摘するべきと思います。

今回は、当該分野に関する国立大学教授職として、技術的に明白な白黒がつくポイント

にはしっかり折り目をつけておくべきと考え、稿を準備しました。

＊　AI合成音声は初等的にも弁別可能

一連の経緯について、あまりとえいばあまりである、として郷原信郎弁護士が詳細に解説する動画　https://www.youtube.com/watch?v=2mIJfobZVQ4　を公開されています。

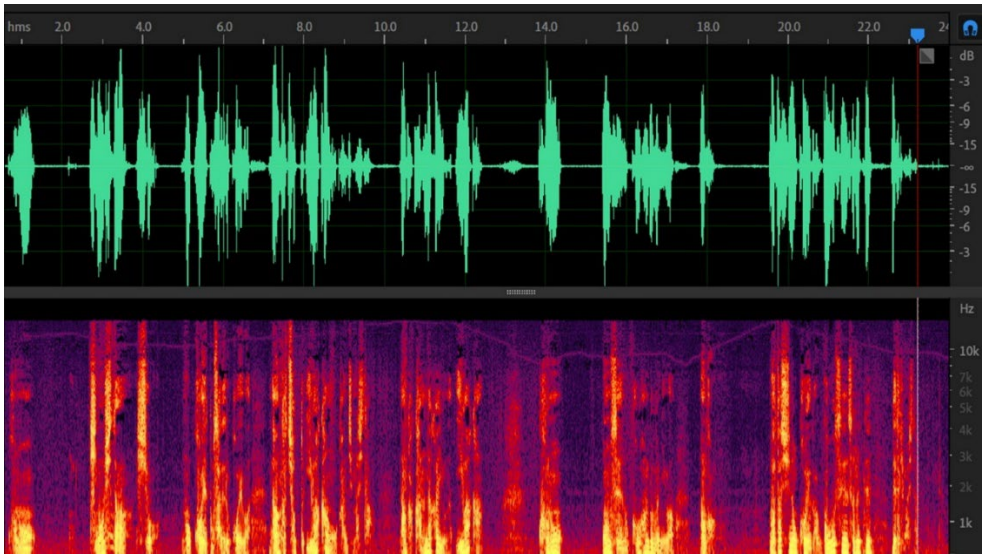
郷原さんは東京大学理学部の先輩にもあたり、親しくご指導いただきますので、今回は許可を頂いて、彼の声をサンプリングし、ＡＩに学習させたうえで「偽郷原弁護士」ヴォイスを試作してみました。

その波形や「スペクトル（周波数分布）」を用いた各種演算、とくに私の研究室では甘利俊一先生が労作された情報幾何の手法を用いる解析を行っていますので、その結果や、材料工学で用いられるパーシステント・ホモロジーによるトポロジカル・マッピングなど進んだ解析による結果もご紹介しておきます。

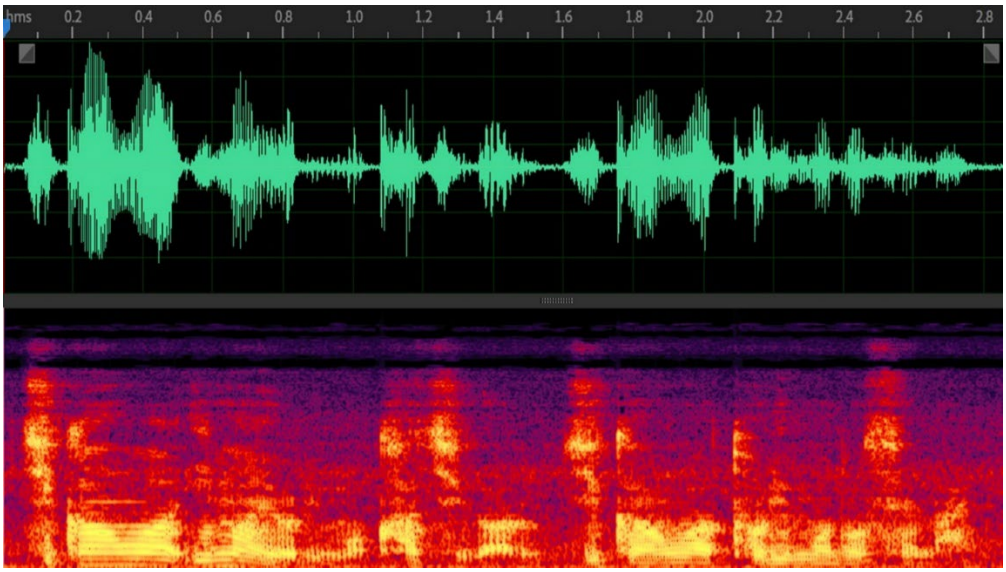
学習させるのは上のリンク　https://www.youtube.com/watch?v=2mIJfobZVQ4　の 7 分 12 秒～22 秒周辺で、以下のように郷原さんが語っている部分です。まず

「この・・・、木下秘書の声とされる声は、なんかちょっと違和感を感じました、なんか甲高いような違和感」

この部分の音声を取り出してシステムに学習させます。つぎに、そのデータをもとに新たにテキストを指定すれば「ＡＩゴウハラ弁護士」に好き勝手な内容を喋らせることが出来ます。おのおののデータを見てみましょう。まず生身の郷原さんによる自然音声ですが



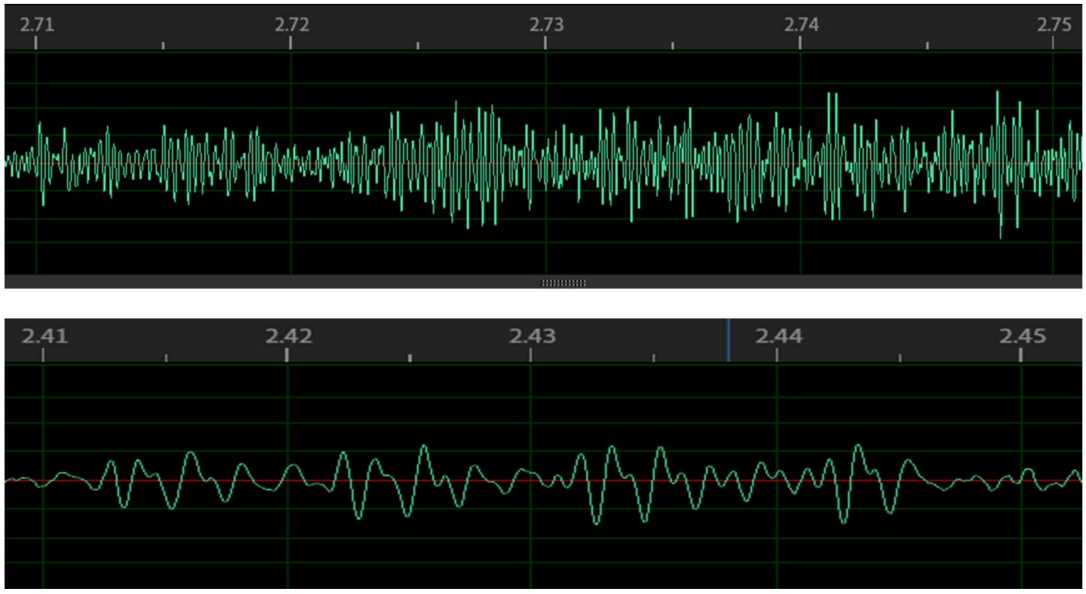
　マスコミ各位や国会議員が見ても判りやすいよう、平易な説明から始めてみます。
目視でも判るように、言葉と言葉の間の間隔がランダムだったり、母音がさまざまに変化してり、「一様でないムラ」が見られます・・・といっても判りにくいかもしれないので、合成音声のデータと比較してみましょう。



　一目見ただけでも、下半分の「スペクトル」の赤や黄色が横一線になってなっているのが分かるでしょう。このエリアは＜あ、え、い、お、う＞など「母音」に当たる周波数帯域ですが、「自然音声」の場合、生きた人間の声帯振動は決して一様にならないので、ムラが見えます。

これに対して合成音声は、事前学習データをもとに電気的に発振させた音ですからムラを創り出すのがむしろ難しい。結果的に単純な音にならざるを得ません。

これは「波形」を拡大することで、子供が見ても判りやすくなります。いっぱんに電子的に発生させた音声は、目的とする音しか録音されていません。これに対して自然な録音では、まず背景雑音があり、さらに、意図せざる物音が随時混ざってきますから「細かい」「汚い」波形になります。

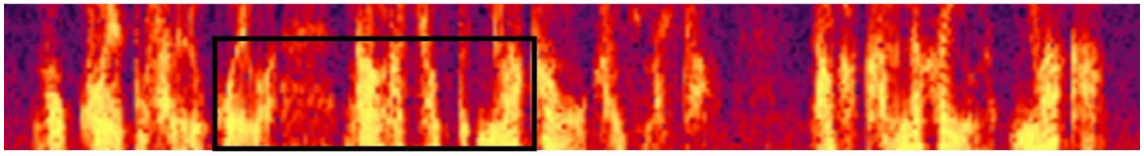


図の上に示したのが「自然な音声の録音」下に示したのが「合成音声のゴウハラ弁護士」のおなじ０．０５秒区間の波形です。まがいものは所詮シンプルにしか作れませんから、偽物であることは一目瞭然と思います。

じつはこうした事情は「ＡＩ画像」「ＡＩ動画」でも同様で、そこでは「空間周波数」という指標などを用いて、真偽を一発で見分けることが出来ます。

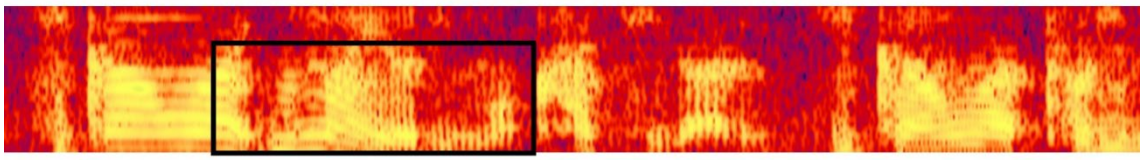
自然人格による生理的な音声

上の例：郷原信郎弁護士の自然な発話のスペクトル



機械学習による合成音声は一様、不自然なので一目で弁別可能

下の例：上と同じ郷原弁護士の声を学習して合成したもの、合成痕が明らか



＊　より厳密な解析：情報幾何、データ駆動科学などを用いた実証

以上は、新聞社のデスクなどマスコミ各位、また国会議員や議会関係者にも判りやすいよう、平易な事例で記しましたが、校閲係から「最先端のＡＩ音声は判別しにくいのでは？」うんぬんと質問がありましたので、追記した部分が以下に当たります。

ＡＩに限局せず、合成音声はほぼ永久に、個別にみれば作りものであることが直ちにわかってしまいます。大量の「ＡＩ合成音声」のなかから「ＡＩで自動的に自然な音声と弁別するシステム」などは、近年盛んに研究開発されている分野ですが、これは自動判定させようとすると、誤判定が出るというもので、ひとつひとつ取り出せば、まず100%作りものは判ります。

人間の音声言語には、例えば「母音の非一意性」といった特徴があります。どういうことか？

いま「ありがとー」と発音するとして、さいごの「おー」を幾分口を広めに開けた[O]

で発語したとしましょう。

で、この[O]の発音を、そのまま利用して、もう一度「ありがとー」と発音することが出来ます。試してみてください。教室でやってみせると多くの学生が興味を持ってくれるポ

イントです。

物理的には全く同じ音を発しているのに、前後の関係によって[O]が「お」にも聞こえ、また「あ」とも聞こえる。こうした不思議な知覚現象は「母音の非一意性」あるいは「音声知覚の多義性」「話者正規化による不変性」の問題などとして、関連専門人があまねく知る基礎の代表例のひとつです。

これは逆にいうと、生きた人間が話す言葉では、およそ様々な「音（音響＝物理的音波)」が同一の「あ」「い」「う」「え」「お」などの「音素（発音を示す字母)」と対応していることを意味する。

要するに、生きた人間は、多様なバラエティをもった音を、物理的な声帯や喉、舌や唇を震わせて、お喋りしている。

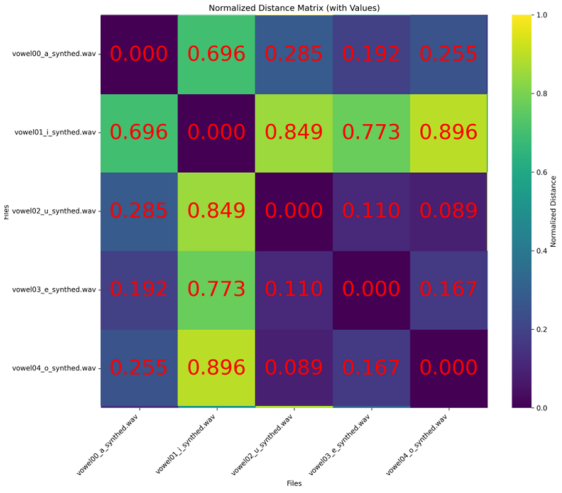
これに対してＡＩ合成音声は、限られた学習データから音素（字母）に対応する音響を合成して繋げるので、圧倒的に音のバラエティが少なく、一つの音素には類似した響きが割り当てられる形で合成音声が出力されます。

　いま、生身の郷原信郎弁護士が発音する「あ」「い」「う」「え」「お」の音と、それらを学習した音声合成ＡＩが出力した[a][i][u][e][o]の音のサンプルを作ってみます。以下判りにくい話が続きますが、編集部からの注文によりますので、ご容赦ください。

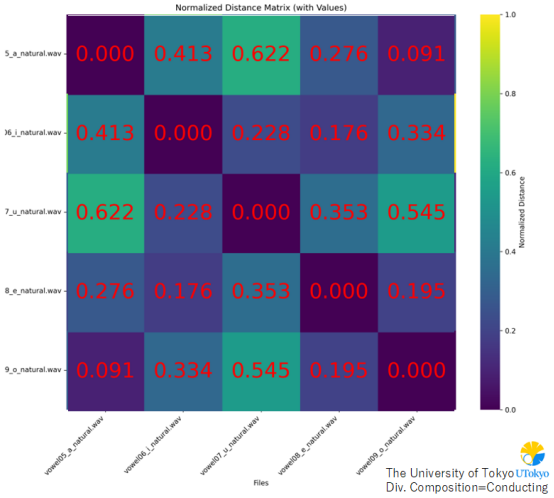
　これらを「周波数分解」して「スペクトル」になおし（たうえ、これを「確率密度関数」と見るために、スペクトルの値の合計が１になるように調整し）ます。

次に、このようなスペクトル（確率密度）同士が、どれくらい「違っているか」を測る指標として「Wasserstein 計量」
https://ja.wikipedia.org/wiki/%E3%83%AF%E3%83%83%E3%82%B5%E3%83%BC%E3%82%B9%E3%82%BF%E3%82%A4%E3%83%B3%E8%A8%88%E9%87%8F　という一種の「距離」がありますので、これで「天然郷原弁護士」の「あいうえお」と「ＡＩゴウハラ氏」合成音声の「aiueo」の「自己距離行列」というものを計算すると、以下のようになります。

自然音声の
正規化Wasserstein距離行列例
行列要素の平均　～0.431

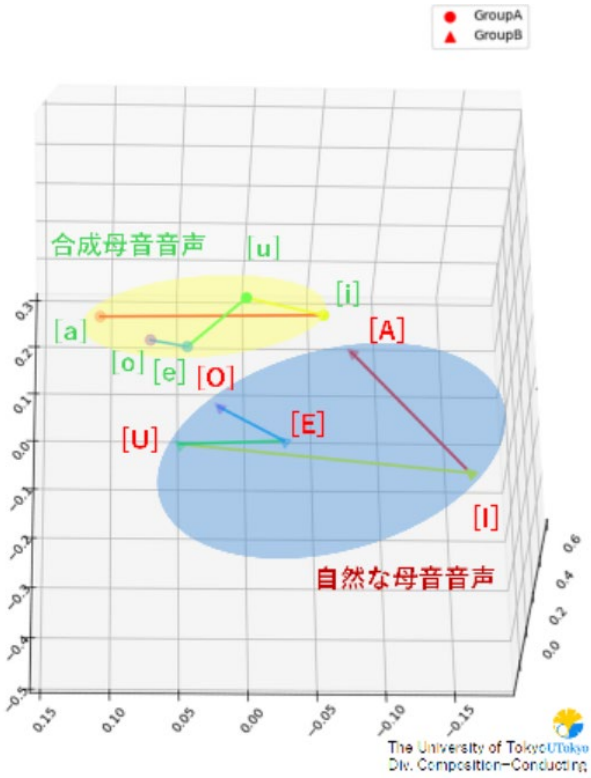


合成音声の
正規化Wasserstein距離行列例
行列要素の平均　～0.323



この図表、正確には「行列」が何を意味するかというと、「AI ゴウハラ氏」の合成音声は似たようなスペクトルで音声らしき音響を合成しているのに対して、「天然郷原弁護士」の発音は、よりバラエティに富んだ音で構成されていることを示す一例になっている。何分編集部から注文がついてから、私自身ならびに小ラボの修士２年生、田村優成君とで計算した例で、これだけでは十全とはいいがたいです。ですが、「最適輸送問題」の情報幾何を用いて合成音声の特徴を端的に示す一例にはなっていると思います。

同じ「天然郷原弁護士」と「ＡＩゴウハラ氏」の母音スペクトルが Wasserstein 距離で測ってどの程度かけ離れているかをパーシステント・ホモロジー https://en.wikipedia.org/wiki/Persistent_homology　の手法を用いてトポロジカル・マッピング　https://en.wikipedia.org/wiki/Topological_data_analysis　した結果を示します。



天然と人工、双方の音声スペクトル同士の距離（「Wasserstein 距離」）を保ったまま、空間上にマッピングしてみると、合成音声のグループと自然な音声のグループが明瞭に分離されるのが分かるかと思います。なにぶん、先ほど言われて短時間で計算した例ですので最善とは言えないかもしれませんが、合成音声は自然音声と異なるスペクトル構造を持ち、その特徴は情報幾何の手法や、トポロジカル・マッピングなどの手法を適切にもちいれば、明瞭に示すことが可能である一例をお目にかけました。
＊　本来、閣僚の国会答弁は法に準ずる重さを持つ

こんな具合で、中身を知っている人間には、AI 合成とそうでないものの弁別は明瞭なものです。他方、不特定多数向けに配信された Fake 情報を視聴する、圧倒的多数のユーザには、こうした識別能力を期待することはできません。

そこでしっかりしてほしいのが、マスコミであり報道陣なのですが・・・

今回、高市氏が

「秘書の声かどうか、判断は難しゅうございます」
「これはどう考えても　確認のしようがございません」

などと述べた次の瞬間

「AI 合成と自然音声の区別など、瞬時に出来るにきまっとるだろうが」

と突っ込みを入れたメディアは、私が見た範囲（10 件以上検索してみましたが）では皆無。局として責任を取りたくないので、読んでくる「政治ジャーナリスト」その他のコメンテータも、およそ「群盲象をなでる」ＡＩや情報のリテラシー欠如が著しく、頭を抱えざるを得ません。

「確認のしようがございません」などということが、初歩の１の１で、ありえません。

＊立法府＝国会は fake 規制に正しい科学的根拠と分別を！

かつて、刑法の團藤重光先生と「反骨のコツ」
https://www.amazon.co.jp/%E5%8F%8D%E9%AA%A8%E3%81%AE%E3%82%B3%E3%83%84-%E6%9C%9D%E6%97%A5%E6%96%B0%E6%9B%B8-69-%E5%9C%98%E8%97%A4-%E9%87%8D%E5%85%89/dp/4022731699　などの書籍やプロジェクトを通じて、さまざまな法制度検討の議論を伺ったのを思い出します。

その中で、閣僚の国会答弁は、それ自体に国民や外部の裁判所を直接縛るような「法的拘束力」はないけれど。行政内部や、法解釈・政策の方向性を決定づける場においては実質的に法に準ずる非常に強い統治上の拘束力・責任を持つ、という論点がありました。

例えば 1970（昭和 45）年 3 月 18 日、第 3 次佐藤栄作内閣の中曽根康弘防衛庁長官（当時）は衆議院予算委員会、日本社会党の楢崎弥之助代議士（当時）の質問に対し、我が国の防衛の基本概念である「専守防衛」について「戦略上も戦術上も防御であり、相手国を直接攻撃するような兵器は持たない」と明言します。

この答弁により日本の安全保障は「相手から攻撃を受けて初めて防衛力を行使する」という受動的な国家方針に 100%拘束されました。今日に至るまで実質的な安全保障基本法の役割を果たし続けています。

その程度に閣僚の国会答弁は、一つ一つがその程度の重みを本来は持つべきものであることを、團藤先生は、ご自身が判事を務められた最高裁判所での判例（法と同様の拘束力

を以後に残す）と対照して強調しておられました。

増していわんや、内閣総理大臣の国会答弁は、その場の勢いで為されるようなものではありません。一個人あるいは一議員事務所にＡＩのリテラシーが欠如していたとしても、このような非科学的なやり取りが議事録に残るなら、大学教官として大いに危惧の念を抱かざるを得ません。

録音音声に関して、それがＡＩ合成か、そうでないか「判断は難しい」などということはあり得ません。まず自然人格の発声であるか否かは、適切にデータ解析すれば明瞭に判ります。

次いで、警察が犯罪捜査で用いるような分析システムを用いれば、刑事法定で有罪を言い渡す根拠となる程度の確かさで、ある人物であるか、否かを弁別することも可能でしょう。

我が国のＡＩコンテンツリテラシーの未来のため、また適切な制度整備が立法府＝国会で為されることも期待して、今回は専門の観点から根拠とともに記しました。