話者ダイアライゼーション: 「誰が、いつ話したのか」の理解を可能にする高SNR MEMSマイクロフォン

Dr. Gunar Lorenz、インフィニオン テクノロジーズ

音声入力の品質を向上させることは、AIを日常的に利用する上で不可欠です。高SNR (信号対雑音比) のMEMSマイクロフォンで録音された音声であれば、AIは騒がしい環境でも会話中の異なる話者を容易に認識できます。音声入力の品質改善は、誤認識を減らし、翻訳の品質も向上させます。高SNRのMEMSマイクロフォンを用いたAIアプリケーションは、高度な環境認識、話者認識、そして音の区別を可能にし、デバイスを用いた世界との交流に革命をもたらします。

バイリンガルの子どもの親として、私は常にAIがプライベートと職場の両方において、言語の壁を打ち破る可能性に期待を寄せてきました。昨年、主要なスマホメーカーが取り上げたAIの話題を検証し、その後のアジア出張で最新のリアルタイム翻訳を実際に試せるのを待ち望んでいました。最新のフラッグシップ モデルのスマートフォンを持ち、画面に表示される韓国語の会話を追おうとしたところ、英語のテキストはわずかに遅れただけで、直ぐに表示されました。

翻 訳の速度はそれでも驚異的でしたが、元の文章の意味が完全に捉えられていない場合もありました。さらに、韓国語の会話に新たな参加者が加わった際、英語に翻訳された文章には、誰がいつ発言したか表示されませんでした。また、複数の発言者が同時に話す場合、発言と翻訳の間にわずかな遅れが生じることで、会話の流れを追うことが困難になりました。複数のライブ翻訳アプリを試しましたが、いずれも話者識別用のIDやラベルを追加していませんでした。AI用語では、誰がいつ話したかを理解する機能を「話者ダイアライゼーション」と呼びます。この機能がないと、ライブ翻訳はシェイクスピアの戯曲を、誰が誰なのか分からない状態で読むようなもので、意味を正しく理解できません。

ダイアライゼーションなしでAIが取得する情報:

聖人たちにも唇はあるし、巡礼者たちもそうではありませんか？ええ、巡礼者も祈るためにその唇を使います。それなら、聖なる方よ、唇にも手と同じことをさせてください。聖人の心は動きませんが、祈りのためなら許しを与えます。それでは、動かないでください、私の祈りが叶う瞬間をいただきますから。こうして、私の唇からあなたの唇へ罪が清められました。それでは今度は私の唇があなたから罪を受け取ったことになるのですね。あなたの唇から罪をもらったのですか？何と甘美な誘惑でしょう！それならもう一度私にその罪を戻してください。あなたのキスはお手本のようですね。

ダイアライゼーションを使ってAIが取得する情報:

話者 1: 聖人たちにも唇はあるし、巡礼者たちもそうではありませんか？
話者2: ええ、巡礼者さんも祈るためにその唇を使います。
話者 1: それなら、聖なる方よ、唇にも手と同じことをさせてください。
話者2: 聖人の心は動きませんが、祈りのためなら許しを与えます。
話者 1: それでは、動かないでください、私の祈りが叶う瞬間をいただきますから。こうして、私の唇からあなたの唇へ罪が清められました。
話者2: それでは今度は私の唇があなたから罪を受け取ったことになるのですね。
話者 1: あなたの唇から罪をもらったのですか？何と甘美な誘惑でしょう！それならもう一度私にその罪を戻してください。
話者2: あなたのキスはお手本のようですね。

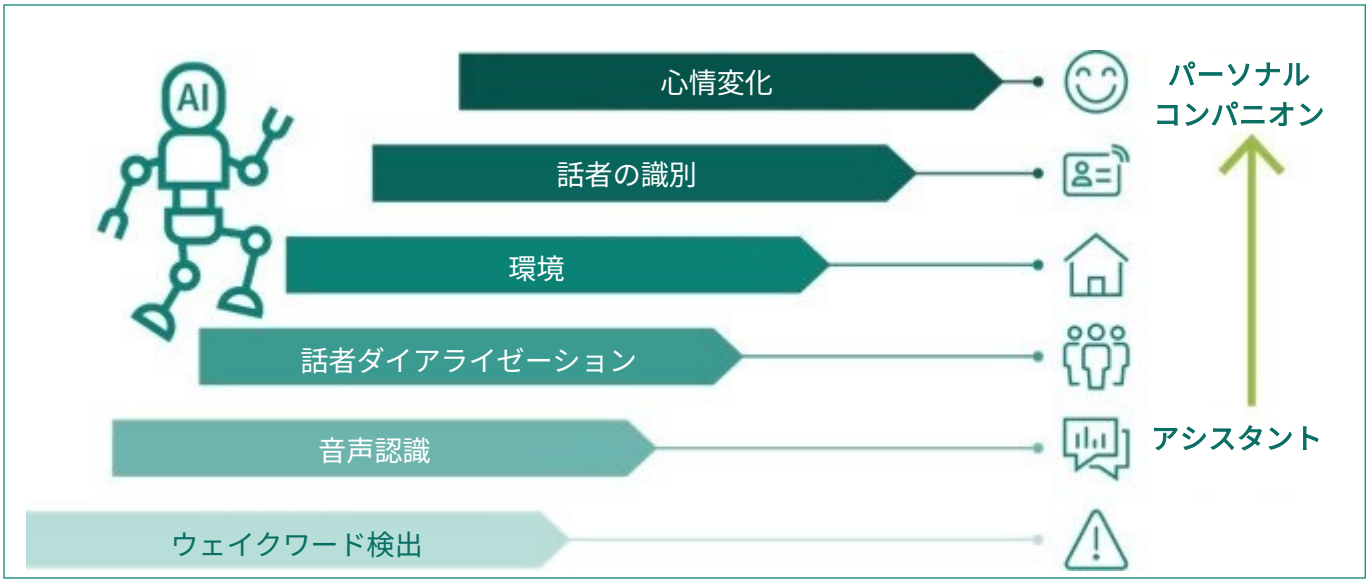
シェイクスピア本来の意:

ロミオ: 聖人たちにも唇はあるし、巡礼者たちもそうではありませんか？
ジュリエット: ええ、巡礼者さんも祈るためにその唇を使います。
ロミオ: それなら、聖なる方よ、唇にも手と同じことをさせてください。
ジュリエット: 聖人の心は動きませんが、祈りのためなら許しを与えます。
ロミオ: それでは、動かないでください、私の祈りが叶う瞬間をいただきますから。こうして、私の唇からあなたの唇へ罪が清められました。
ジュリエット: それでは今度は私の唇があなたから罪を受け取ったことになるのですね。
ロミオ: あなたの唇から罪をもらったのですか？何と甘美な誘惑でしょう！それならもう一度私にその罪を戻してください。
ジュリエット:あなたのキスはお手本のようですね。

スマートフォンで録音された、美しく演じられたシェイクスピアの劇からAIが得られる情報はその程度です。適切なダイアライゼーション (中央のテキスト) があれば、録音された音声に2人の人物の会話が含まれていることを、AIははっきりと認識できます。2人が会話している内容から、AIは会話のロマンチックな性質を推測するかもしれません。

しかし、適切なダイアライゼーションが提供されても、まだAIは、2人の性別と年齢を特定することはできません。

AIが音声解釈を行う能力は一連のスキルと見なされ、「ウェイクワード検出」は入力音声を解釈する最も基本的なAI能力です。AIの音声スキルを段階的に視覚化すると、「話者の識別」は「音声認識」の次の、レベル3に位置します。



録音された対話でどちらがロミオでどちらがジュリエットなのかを特定するには、話者の識別が必要です。話者の識別により、1ページ目のダイアログボックスの右側にシェイクスピアの本来の意図が表示されます。たとえ AI がロミオとジュリエットを正しく識別したとしても、俳優の声のトーンや演技のロマンチックさや、美しさはすべて失われてしまうことでしょう。少なくとも、学生時代の私が感じていたのがまさにそうです。ピーター マーティンス監督の映画『ロミオ + ジュリエット』で、レオナルド ディカプリオとクレア デインズの対話を見たとき、その対話の深さと美しさを初めて理解できたのです。

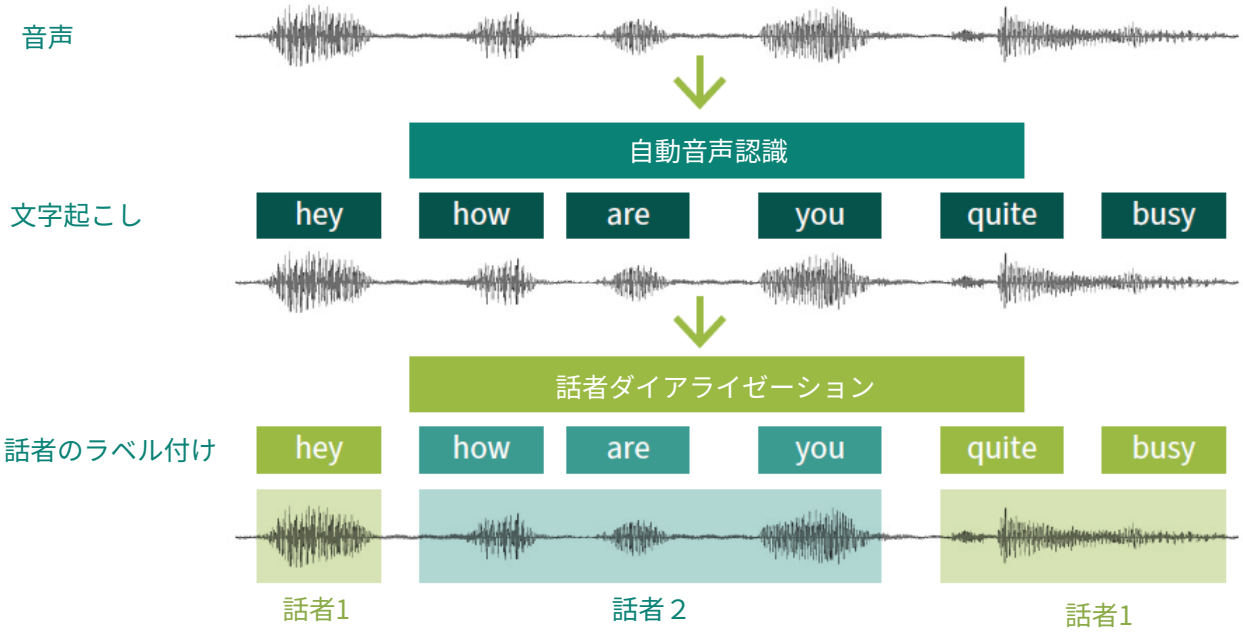
調査結果によると、コミュニケーションの大部分は、声のトーン (38%) やボディランゲージ (55%) など非言語的なもので占められています。私の最近の記事の一つでは、AIの聴解能力を向上させれば、機械がデジタルコンパニオンや同僚になり得ると述べました。機械は人間の会話を理解し、録音されたやり取りの要約を作成できるようになるでしょう。

インフィニオンは、高品質なオーディオが優れたユーザー体験を提供する鍵だと考えています。高SNR MEMSマイクロフォンの主要サプライヤーであるインフィニオンは、この考えがAIにも当てはまるのか検証しました。音声認識能力の向上が、AIが異なる話者を識別するのに、どのような影響を及ぼすのか検証したかったのです。Syntiant社の最近の発表は、高SNRマイクがウェイクワード検出と音声コマンド認識にもたらす利点について、インフィニオン社内で行った測定結果を裏付けるものでした。

最新AIの話者識別機能をよりよく理解するため、会話を録音した音声ファイルを使用して話者識別の誤差を測定できる一般公開されているAIツールを使用しました。これまでのテストと同様に、頻繁にリリース/更新される新たなAIツールに対応し続けるのは難しいですが、最終的に、NVIDIAのNeMo Frameworkを採用しました。NeMoは「より良く聴く」とはどういうことかを定量化するための実測データを取得する上で、特に役に立つことがわかりました。

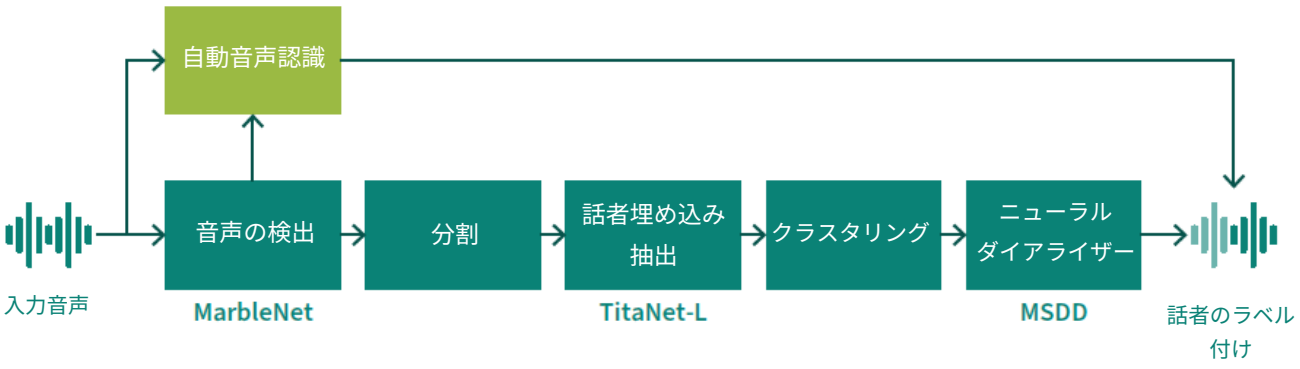
NVIDIAのウェブサイト上にはこう書かれています。「話者ダイアライゼーションは、音声記録に対して、話者にラベル付けするプロセスであり、『誰が、いつ話したか？』という疑問に答えるのを目的とします。話者ダイアライゼーションは、音声認識と比べたときに、明確に話者を区別されます」

下図で示すように、話者ダイアライゼーションの実行前には、「何が話されたか」は分かるものの、「誰が話したか」までは分かりません。つまり、話者ダイアライゼーションは、音声認識システムが書き起こした内容に話者のラベルを付け加えるのに不可欠な機能なのです。



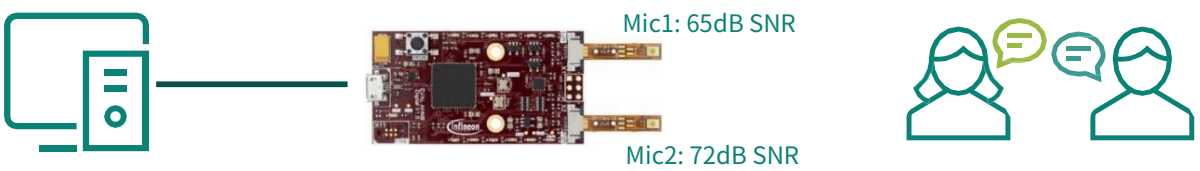
話者識別を行うライブ翻訳には、NVIDIAのウェブサイトにあるブロック図に示されるように、

「音声の検出」から始まる音声処理過程に複数の要素が必要です。

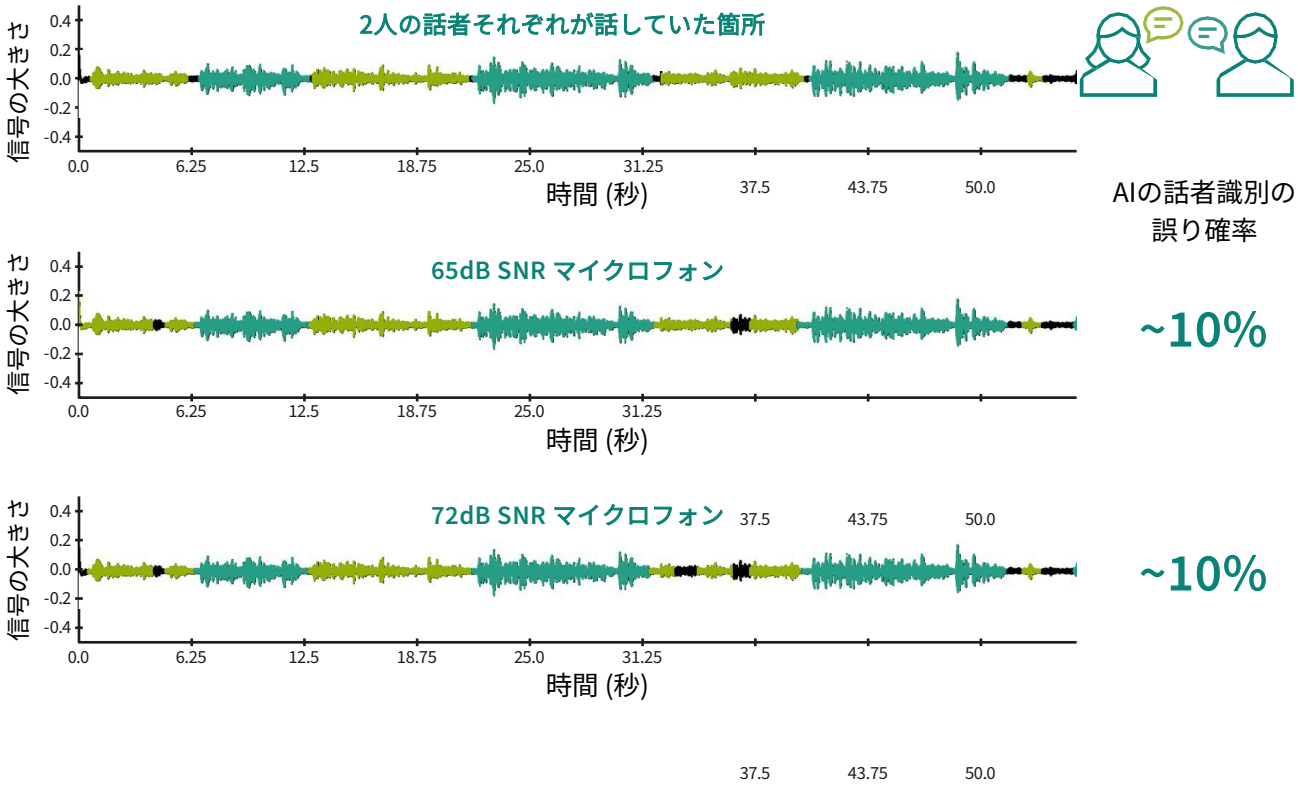


最初のテスト環境では、インフィニオンのマイク評価ボードAudiohub Nano を使用しました。

このボードは、2つの異なるMEMSマイクで同時に音声を録音できます。



最初のマイクロフォンは、多くのハイエンドのスマートフォンに搭載されているものとほぼ同じで、SNRは65 dB (A) です。マイクロフォンの信号対雑音比 (SNR) は、聴収したい音 (例: 音声) の強度とマイクロフォンの自己雑音を比較し、周りの雑音と音を識別する能力を示します。SNRが高いほど、録音された信号の音質は良くなります。インフィニオンで実施するテストにおける高性能なものの基準として、最新のデジタルXENSIV™マイクロフォンIM72D128 V [8] を使用し、SNRは72 dB（A）です。2つのマイクロフォンで録音された音声ストリームは、NVIDIAの話者ダイアライゼーションシステムNeMoを用いて解析されました。図には記録された会話の音声信号が示されています。



2つのグラフは2人の話者を表しており、1人目の話者の声は緑色で、2人目の話者の音声は青色で表しています。この参照データは、手動でラベル付けしたもので、NeMoへ提供されました。各話者が発言するたびに、ボタンが押されました。

中央のグラフは、標準のスマートフォンに搭載されているマイクを使用したNeMoの自動で話者にラベル付した結果であり、下部のグラフはインフィニオンの高SNRマイクを使用した結果を示しています。

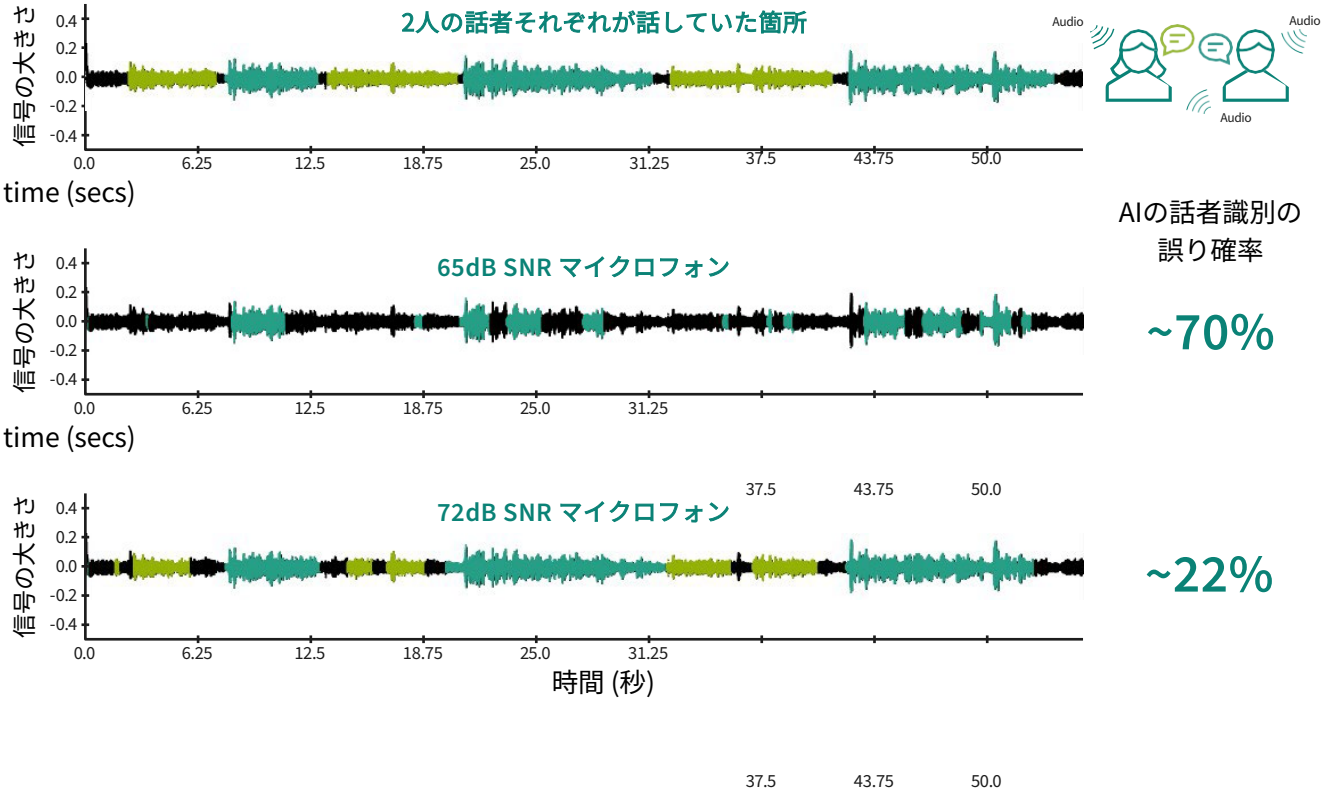
両グラフは、NeMoが話者1と2をほぼ正しく識別したことを示しています。ただし、中央に何か所か見られる黒の部分は、いずれの話者としても認識されなかった箇所です。両方のマイクにおいて、NeMoは比較的低い話者識別の誤り確率 (約10%) が算出されました。そのため、使用したマイクに関わらず、NeMoは90%の確率で2人の話者を正しく識別できました。

なお、留意願いたいのは、2人の話者は、話し始める前にわずかな間を置くことにより、特にNeMoが識別しにくいようなことはしませんでした。また、両者が同時に話すことは一度もありませんでした。

2回目のテストでは、非常に騒がしいインフィニオンのカフェテリアで同じ会話の実験を行い、バックグラウンドノイズを追加しました。



下のグラフでは再び3つの音声シーケンスを示しています。この上段には再び２人の話者が話していた箇所が示され、話者シーケンスでは１人目の話者が緑色、２人目の話者を青色で表しています。



AIの話者識別の誤り確率

~70%

~22%

バックグラウンドノイズ下では、中程度のSNRマイクで録音した場合、NeMoは2人目の話者を検出することさえ困難でした。さらに、話者識別の誤り確率は最大約70%にも達しました。

同じ状況で高SNRマイクを使用すると、識別成功率は劇的に向上しました。静かな環境ではうまく機能しているように見えても、騒がしい環境では難しくなります。

この結果から、軽度の難聴を抱える高齢者が、騒がしい環境で集団の会話についていく難しさを考えさせられました。

私たちも年を重ねれば、聴力が徐々に低下し、高音域の音を聞き取ったり、騒がしい環境での会話を理解したりするのが難しくなることが増えます。低SNRのMEMSマイクロフォンを使用する場合、AIにも、同様のことが起こるようです。

AIと人間の聴覚とを比較すると、現在使用されているほとんどのMEMSマイクロフォンは、音声の微細な詳細を検出する能力において、60歳男性の聴力レベルとほぼ同等の性能を示します。詳細はこちら。

まとめと展望

AI 技術が急速に進歩しているにもかかわらず、AI が人間の会話を真に「理解」するのには、さらに時間が必要なようです。音声認識技術の習得は最初のステップに過ぎず、人間の会話を、人間がするように理解するためには、話者識別技術の習得が不可欠です。
当社の実験から、AIを日常的に利用できるようにするには、今以上に優れた音声入力の必要性がわかりました。こうした進歩において、高SNR (信号対雑音比) の音声入力は特に重要です。SNRが高いほど音声が明瞭になるため、AIがより正確に音声を処理できるようになるからです。このことは、騒がしい環境において特に重要となります。AIが異なる話者を容易に区別できるようになり、誤認識が減り、翻訳の品質を向上させられるでしょう。そして、将来的には、高SNRのMEMSマイクロフォンを使用したAIアプリケーションにより、高度な環境認識や音の区別が可能になることで、デバイスを用いた世界中の人々との交流に大きな変革をもたらすでしょう。

MEMSマイクロフォン技術の革新は、自然に聞き取り、理解し、共感できるAIを実現し、AIは日常生活に欠かせない存在になっていくはずです。

購入に関して

インフィニオンの販売拠点または販売パートナーまでお問い合わせください:
www.infineon.com/WhereToBuy

サービス ホットライン

インフィニオンはフリーダイヤルの0800/4001サービス ホットラインを代表番号として提供しており
英語、中国語、ドイツ語で24時間利用可能です。

- ドイツ 0800-951-951-951 (ドイツ語/英語)
- 中国本土 4001-200-951 (北京語/英語)
- インド 000-800-4402-951 (英語)
- 米国 1-866-951-9519 (英語/ドイツ語)
- その他の国 00*-800-951-951951 (英語/ドイツ語)
- ダイレクトアクセス +49-89-234-0 (相互接続料金、ドイツ語／英語)

*ご注意: 一部の国から上記国際電話番号にアクセスする際「00」以外のコードをダイヤルする必要があります。
詳細は https://www.infineon.com/service/をご覧ください。

モバイル製品カタログ
iOS/Android用モバイル アプリ

発行元:
インフィニオン テクノロジーズ ジャパン株式会社

© 2025 Infineon Technologies AG.
All rights reserved.

Public

Date: 03 / 2025

Stay connected!

Scan QR code and explore offering
www.infineon.com

Please note!
This Document is for information purposes only and any information given herein shall in no event be regarded as a warranty, guarantee or description of any functionality, conditions and/or quality of our products or any suitability for a particular purpose. With regard to the technical specifications of our products, we kindly ask you to refer to the relevant product data sheets provided by us. Our customers and their technical departments are required to evaluate the suitability of our products for the intended application.

We reserve the right to change this document and/or the information given herein at any time.

Additional information
For further information on technologies, our products, the application of our products, delivery terms and conditions and/or prices, please contact your nearest Infineon Technologies office (www.infineon.com).

Warnings
Due to technical requirements, our products may contain dangerous substances. For information on the types in question, please contact your nearest Infineon Technologies office.

Except as otherwise explicitly approved by us in a written document signed by authorized representatives of Infineon Technologies, our products may not be used in any life-endangering applications, including but not limited to medical, nuclear, military, life-critical or any other applications where a failure of the product or any consequences of the use thereof can result in personal injury.