**テキスト読み上げの概要**

[アーティクル] • 2024/09/24

この概要では、Azure AI サービスに含まれる⾳声サービスのテキスト読み上げの利点と機能について説明します。

テキスト読み上げを使⽤すると、アプリケーション、ツール、またはデバイスで、テキストを⼈間に似た合成⾳声に変換できます。 テキスト読み上げ機能は⾳声合成とも呼ばれます。 ⼈間に似た事前構築済みのニューラル⾳声をそのまま使⽤するか、製品やブランドに固有のカスタム ニューラル⾳声を作成します。 サポートされている⾳声、

⾔語、ロケールの完全なリストについては、「⾳声サービスの⾔語と⾳声のサポート」を参照してください。

# コア機能

テキスト読み上げには、次の機能が含まれています。

| **機能** | **まとめ** | **デモ** |
| --- | --- | --- |
| 事前構築済みのニューラル音声 ([価格ページ](https://azure.microsoft.com/pricing/details/cognitive-services/speech-services/)では "ニューラル" と表記) | 非常に自然で、すぐに使用できる音声です。 Azure サブスクリプションと Speech リソースを作成した後、[Speech SDK](https://learn.microsoft.com/ja-jp/azure/ai-services/speech-service/get-started-text-to-speech) を使用するか、[Speech Studio ポータル](https://speech.microsoft.com/portal)にアクセスし、事前構築済みニューラル音声を選択して作業を開始します。 [価格の詳細](https://azure.microsoft.com/pricing/details/cognitive-services/speech-services/)に関するページを確認してください。 | [音声ギャラリー](https://speech.microsoft.com/portal/voicegallery)を確認して、ビジネス ニーズに合った適切な音声を決定してください。 |
| カスタム ニューラル音声 ([価格ページ](https://azure.microsoft.com/pricing/details/cognitive-services/speech-services/)では "カスタム ニューラル" と表記) | 自然な音声を作成するための使いやすいセルフ サービス。責任を持って使用していただくためアクセスが制限されています。 (S0 レベルで) Azure サブスクリプションと Speech リソースを作成し、カスタム ボイス機能の使用を[申し込み](https://aka.ms/customneural)ます。 アクセスを許可されたら、[Speech Studio ポータル](https://speech.microsoft.com/portal)にアクセスし、**[Custom Voice]** を選択して使用を開始します。 [価格の詳細](https://azure.microsoft.com/pricing/details/cognitive-services/speech-services/)に関するページを確認してください。 | [音声のサンプル](https://aka.ms/customvoice)を確認してください。 |

**ニューラル テキスト読み上げ機能についての詳細情報**

テキスト読み上げでは、ディープ ニューラル ネットワークを使⽤して、⼈間の録⾳とほとんど区別できないコンピューターの⾳声を⽣成します。 ニューラル テキスト読み上げでは、明瞭な発⾳により、ユーザーが AI システムと対話する際のリスニング疲労が⼤幅に軽減されます。話し⾔葉のアクセントとイントネーションのパターンを、"*韻律*" と呼びます。 従来のテキスト読み上げシステムでは、韻律を、独⽴したモデルによって管理される別々の⾔語分析と⾳響予測の⼿順に分解します。 その結果、こもったような⽿障りな⾳声合成になる可能性があります。

⾳声サービスのニューラル テキスト読み上げ機能と、従来のテキスト読み上げシステムの制限を克服する⽅法の詳細を次に⽰します。

**リアルタイム⾳声合成**: Speech SDK または REST API を使⽤して、あらかじめ構築されたニューラル⾳声またはカスタム ニューラル⾳声を使ってテキスト読み上げを⾏います。

**⻑時間の⾳声の⾮同期合成**: バッチ合成 API を使⽤すると、10 分を超えるテキスト読み上げファイル (たとえば、オーディオ ブック、講義など) を⾮同期に合成できます。 Speech SDK または Speech to Text REST API を使⽤して実⾏される合成とは異なり、応答はリアルタイムで返されません。 要求が⾮同期に送信され、応答がポーリングされること、および合成⾳声がサービスから利⽤可能になったときにダウンロードされることを想定しています。

**事前構築済みニューラル⾳声**: Azure AI ⾳声はディープ ニューラル ネットワークを使⽤して、話し⾔葉のアクセントとイントネーションに関する従来の⾳声合成の限界を克服します。 韻律予測と⾳声合成が同時に⾏われるため、より滑らかで

⾃然な⾳声出⼒が得られます。 事前構築済みの各ニューラル⾳声モデルは、24 kHz および⾼忠実度の 48 kHz で利⽤できます。 ニューラル⾳声を使⽤すると、次のことができます。

チャットボットや⾳声アシスタントとの対話をより⾃然で魅⼒的なものにします。

電⼦書籍などのデジタル テキストをオーディオ ブックに変換します。

⾞内ナビゲーション システムを強化します。

プラットフォーム ニューラル⾳声の完全な⼀覧については、「⾳声サービスの⾔語と⾳声のサポート」を参照してください。

**SSML を使⽤したテキスト読み上げ出⼒の改善**: ⾳声合成マークアップ⾔語 (SSML) は、テキスト読み上げ出⼒をカスタマイズするために使⽤される XML ベースのマークアップ⾔語です。 SSML を使⽤して、ピッチの調整、⼀時停⽌の追加、発⾳の改善、話す速度の変更、⾳量の調整、単⼀のドキュメントへの複数の⾳声の割り当てを⾏うことができます。

SSML を使⽤して、独⾃の辞書を定義したり、さまざまな話し⽅に切り替えたりできます。 多⾔語の⾳声では、SSML を介して話されている⾔語を調整することもできます。 ⾃分のシナリオでの⾳声出⼒を改善するには、「⾳声合成マークアップ⾔語を使⽤した合成の改善」および「Audio Content Creation ツールを使

⽤した⾳声合成」を参照してください。

**⼝形素**: ⼝形素は、特定の⾳素を⽣成するときの唇、顎、⾆の位置などの観察されたスピーチにおける主要な姿勢です。 ⼝形素は、⾳声と⾳素との強い相関関係を持っています。

Speech SDK で⼝形素イベントを使⽤すると、顔のアニメーション データを⽣成できます。 このデータは、読唇時のコミュニケーション、教育、エンターテイメント、カスタマー サービスでの顔をアニメーションで表現するために使⽤できます。 ⼝形素は現在、en-US 英語 (⽶国) のニューラル⾳声でのみサポートされています。

７ **注意**

2024 年に従来型/標準型⾳声とニューラル以外のカスタム⾳声は廃⽌される予定です。 その後、サポートは終了します。

アプリケーション、ツール、または製品が標準型⾳声とカスタム⾳声のいずれかを使⽤している場合は、ニューラル バージョンに移⾏する必要があります。 詳細については、**ニューラル⾳声への移⾏**に関するページを参照してください。

# 作業の開始

テキスト読み上げの使⽤を開始するには、「クイックスタート」を参照してください。

テキスト読み上げは、Speech SDK、REST API、および Speech CLI を介して利⽤できま

す。



**ヒント**

コードを使⽤しない⽅法でテキスト読み上げを⾳声に変換するには、

**S**

**peech**

**Studio**

の

**Audio Content Creation**

ツールを試してください。

# サンプル コード

テキスト読み上げのサンプル コードは、GitHub 上で⼊⼿できます。 これらのサンプルに含まれるテキスト読み上げ会話は、⼈気の⾼いプログラミング⾔語で作成されています。

テキスト読み上げのサンプル (SDK)

テキスト読み上げのサンプル (REST)

# カスタム ニューラル⾳声

事前構築済みニューラル⾳声に加え、製品やブランドに固有のカスタム ニューラル⾳声を作成することもできます。 作業を始めるにあたって必要なのは、少数のオーディオ ファイルと関連する⽂字起こしだけです。 詳細については、「カスタム ニューラル

⾳声の概要」を参照してください。

# 価格に関する注意

**課⾦対象⽂字**

テキスト読み上げ機能を使⽤している場合、句読点を含めて、⽂字が⾳声に変換されるごとに課⾦されます。 SSML ドキュメント⾃体は課⾦対象外ですが、テキストが⾳声に変換される⽅法を調整するために使⽤される省略可能な要素 (⾳素やピッチなど) は、課⾦対象としてカウントされます。 課⾦対象の⼀覧を次に⽰します。要求の SSML 本⽂でテキスト読み上げ機能に渡されたテキスト

<speak> と <voice> タグを除く、SSML 形式の要求本⽂のテキスト フィールド内のすべてのアークアップ

⽂字、句読点、スペース、タブ、マークアップ、すべての空⽩⽂字

Unicode で定義されているすべてのコード ポイント

詳細については、「Speech Services の価格 」を参照してください。

） **重要**

⽇本語で使⽤される漢字、韓国語で使⽤される漢字、その他の⾔語で使⽤される繁体字を含め、中国語の各⽂字は、課⾦では 2 ⽂字としてカウントされます。

**カスタム ニューラル⾳声でのモデルのトレーニングとホスティングの時間**

カスタム ニューラル⾳声のトレーニングとホスティングは、どちらも時間単位で計算され、1 秒単位で課⾦されます。 課⾦単価については、「Speech Services の価格」を参照してください。

カスタム ニューラル⾳声 (CNV) のトレーニング時間は、"コンピューティング時間" (マシンの実⾏時間を測定する単位) によって測定されます。 通常、⾳声モデルをトレーニングする場合、2 つのコンピューティング タスクが並列で実⾏されます。 そのため、計算されたコンピューティング時間は、実際のトレーニング時間よりも⻑くなります。

平均では、CNV Lite ⾳声をトレーニングするのにかかるコンピューティング時間は 1 時間未満です。CNV Pro の場合、通常、単⼀スタイルの⾳声をトレーニングするには

20 から 40 時間、マルチスタイルの⾳声をトレーニングするには約 90 時間かかります。 CNV のトレーニング時間は、上限 96 コンピューティング時間で課⾦されます。そのため、⾳声モデルのトレーニングにかかったコンピューティング時間が 98 時間の場合、96 コンピューティング時間に対してのみ課⾦されます。

カスタム ニューラル⾳声 (CNV) エンドポイントのホスティングは、実際の時間 (時間単位) で測定されます。 各エンドポイントのホスティング時間 (時間) は、毎⽇ 00:00 UTC に、過去 24 時間について計算されます。 たとえば、エンドポイントが 1 ⽇⽬に 24 時間アクティブになっていた場合、2 ⽇⽬の 00:00 UTC に 24 時間分課⾦されます。 エンドポイントが新しく作成された場合、またはその⽇中に中断された場合、2 ⽇⽬の

00:00 UTC までの累積実⾏時間に対して課⾦されます。 エンドポイントが現在ホストさ

れていない場合は、課⾦されません。 毎⽇ 00:00 UTC の⽇次の計算に加えて、エンドポイントが削除または⼀時停⽌されるた場合も、課⾦がすぐにトリガーされます。 たとえば、12 ⽉ 1 ⽇の 08:00 UTC に作成されたエンドポイントの場合、ホスティング時間は 12 ⽉ 2 ⽇の 00:00 UTC で 16 時間、12 ⽉ 3 ⽇の 00:00 UTC で 24 時間として計算

されます。 ユーザーが 12 ⽉ 3 ⽇の 16:30 UTC にエンドポイントのホスティングを⼀時停⽌した場合、12 ⽉ 3 ⽇の 00:00 から 16:30 UTC までの期間 (16.5 時間) が課⾦対象として計算されます。

## Personal Voice

パーソナル ⾳声機能を使⽤すると、プロファイルのストレージと合成の両⽅に対して課⾦されます。

**プロファイル ストレージ**: パーソナル⾳声プロファイルが作成されると、システムから削除されるまで課⾦されます。 課⾦単位は 1 ⽇あたりの⾳声ごとです。

⾳声ストレージが続く期間が 24 時間未満の場合は、1 ⽇分として課⾦されます。

**合成**: ⽂字ごとに課⾦されます。 課⾦対象⽂字の詳細については、上記の課⾦対象⽂字を参照してください。

**テキスト読み上げアバター**

⾳声合成のアバター機能を利⽤すると、ビデオ出⼒の⻑さに基づいて料⾦が発⽣し、秒単位で課⾦されます。 ただし、リアルタイム アバターの場合、料⾦は、アバターがアクティブな時間に基づき、話しているか、サイレント状態であるかに関係なく、1 秒ごとに課⾦されます。 リアルタイムのアバターの使⽤コストを最適化するには、サンプル コードで提供されているヒントを参照してください ("Use Local Video for Idle" を検索します)。 アバター ホスティングは、エンドポイントごとに 1 秒単位で課⾦されます。 エンドポイントを⼀時停⽌してコストを節約できます。 エンドポイントを⼀時停⽌する場合は、直接削除できます。 もう⼀度使⽤する場合は、エンドポイントを再デプロイするだけです。

# Azure のテキスト読み上げメトリックを監視する

テキスト読み上げサービスに関連付けられている主要なメトリックを監視することは、

リソースの使⽤状況を管理しコストを制御するために重要です。 このセクションで

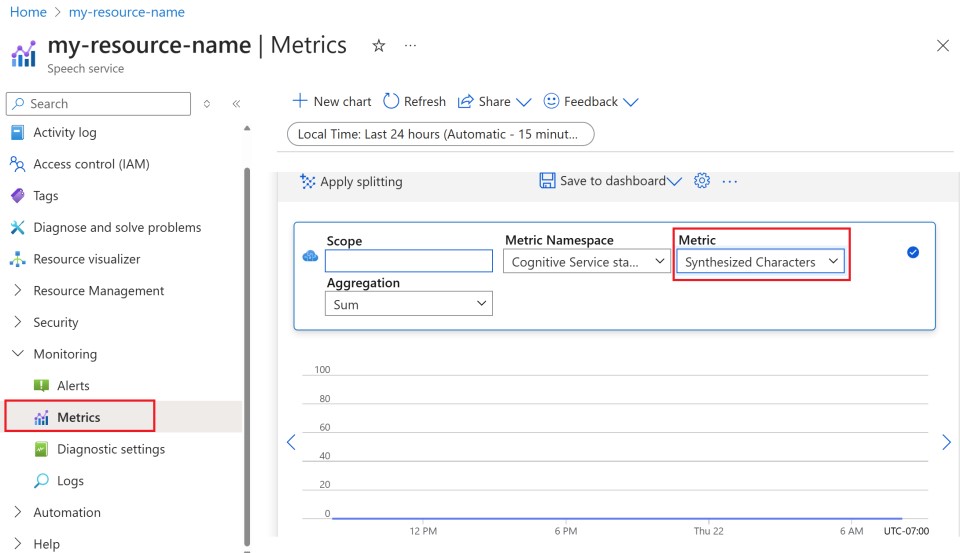
は、Azure portal で使⽤状況に関する情報を⾒つけ、主要なメトリックの詳細な定義を指定する⽅法について説明します。 Azure Monitor メトリックの詳細については、

「Azure Monitor メトリックの概要」を参照してください。

**Azure portal で使⽤状況に関する情報を⾒つける⽅法**

Azure リソースを効率的に管理するには、使⽤状況に関する情報に定期的にアクセスして確認することが不可⽋です。 使⽤状況に関する情報を⾒つける⽅法を以下に⽰します。

1. Azure portal に移動し、Azure アカウントでサインインします。
2. **[リソース]** に移動し、監視したいリソースを選択します。
3. 左側のメニューの **[監視]** の下の **[メトリック]** を選択します。



1. メトリック ビューをカスタマイズします。リソースの種類、メトリックの種類、時間の範囲、およびその他のパラメーターでデータをフィルター処理して、監視のニーズに合ったカスタム ビューを作成できます。 さらに、頻繁に使⽤されるメトリックに簡単にアクセスできるように、 **[ダッシュボードに保存]** を選択することで、メトリック ビューをダッシュボードに保存できます。
2. アラートの設定に関するページを参照してください。

使⽤状況をより効率的に管理するには、左側のメニューから **[監視]** の下の **[アラート]** タブに移動してアラートを設定します。 アラートでは、使⽤量が特定のしきい値に達したときに通知を⾏うことで、予期しないコストを防ぐことができます。

**メトリックの定義**次の表は、Azure テキスト読み上げサービスの主要なメトリックをまとめたものです。

ﾉ **テーブルを展開する**

|  |  |
| --- | --- |
| **メトリック名** | **説明** |
| **合成された⽂字数** | 事前構築済みのニューラル⾳声やカスタム ニューラル⾳声を含め、⾳声に変換された⽂字数を追跡します。 課⾦対象の⽂字の詳細については、「課⾦対象の⽂字」を参照してください。 |
| **合成されたビデオの秒数** | バッチ アバター合成、リアルタイム アバター合成、カスタム アバター合成を含め、合成されたビデオの合計時間を測定します。 |
| **アバター モデル ホスティング秒数** | カスタム アバター モデルがホストされた合計時間を秒単位で追跡します。 |
| **⾳声モデル ホスティング時間** | カスタム ニューラル⾳声モデルがホストされた合計時間を時間単位で追跡します。 |
| **⾳声モデル トレーニング分数** | カスタム ニューラル⾳声モデルのトレーニングの合計時間を分単位で測定します。 |

# リファレンス ドキュメント

Speech SDK

REST API: テキスト読み上げ

# 責任ある AI

AI システムには、テクノロジだけでなく、それを使用する人、それによって影響を受ける人、それがデプロイされる環境も含まれます。 「透過性のためのメモ」を読み、システムでの責任ある AI の使用とデプロイについて確認してください。

* [カスタム ニューラル音声の透過性のためのメモとユース ケース](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/transparency-note-custom-neural-voice?context=/azure/ai-services/speech-service/context/context)
* [カスタム ニューラル音声を使用する場合の特性と制限](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/characteristics-and-limitations-custom-neural-voice?context=/azure/ai-services/speech-service/context/context)
* [カスタム ニューラル音声への制限付きアクセス](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/limited-access-custom-neural-voice?context=/azure/ai-services/speech-service/context/context)
* [合成音声テクノロジの責任あるデプロイのためのガイドライン](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/concepts-guidelines-responsible-deployment-synthetic?context=/azure/ai-services/speech-service/context/context)
* [ボイス タレントに関する開示](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/disclosure-voice-talent?context=/azure/ai-services/speech-service/context/context)
* [開示設計のガイドライン](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/concepts-disclosure-guidelines?context=/azure/ai-services/speech-service/context/context)
* [設計パターンの開示](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/concepts-disclosure-patterns?context=/azure/ai-services/speech-service/context/context)
* [テキスト読み上げ統合の倫理規定](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/tts-code-of-conduct?context=/azure/ai-services/speech-service/context/context)
* [カスタム ニューラル音声のデータ、プライバシー、セキュリティ](https://learn.microsoft.com/ja-jp/legal/cognitive-services/speech-service/custom-neural-voice/data-privacy-security-custom-neural-voice?context=/azure/ai-services/speech-service/context/context)

**次のステップ**

テキスト読み上げのクイックスタート

Speech SDK を取得する