#GLOBALAZURE 2026

#Milano

Dare un volto all'AI: creare avatar intelligenti con Azure Text-to-Speech

Alessio Iafrate

Microsoft MVP

#GLOBALAZURE 2026

#Milano

improove

TD SYNNEX

Grazie ai nostri sponsor 🙏

Azure AI ≠ Azure AI Services

Azure AI

Pre-Built AI

Conversational AI

Custom AI


AZURE AI SERVICES
VI RICORDERETE DI NOI COME I COGNITIVE SERVICES
imgflip.com

Frequently asked questions

Expand all | Collapse all

01/ **What are Azure AI services?**

02/ **Is Azure AI services the same as Azure Cognitive Services?**

Yes, Azure Cognitive Services was the original name for this product collection. We updated the name to Azure AI services to better match our compendium of products.

03/ **Does Azure AI services cost anything?**

04/ **What is the difference between Azure machine learning and Azure AI services?**

Fondamenti

Panoramica dei servizi Azure AI: speech, voci neurali e le idee alla base degli avatar parlanti.

AZURE AI SERVICES: LA PIATTAFORMA

una piattaforma completa per creare applicazioni intelligenti.

I servizi speech, vision, language e decision possono essere combinati in esperienze end-to-end.

Funzionalità pronte per l'enterprise aiutano i team a distribuire su larga scala con API e strumenti coerenti.

Insieme, questi servizi abilitano interazioni con avatar ricche e reattive su più canali.

AZURE AI SERVICES: LA PIATTAFORMA

Voce
Migliorare le esperienze dei clienti tramite funzionalità di riconoscimento vocale, sintesi vocale e traduzione vocale.

Visualizzare tutte le funzionalità del parlato

Lingua e traduttore
Analizza, riepiloga e traduci usando le funzionalità di elaborazione del linguaggio naturale basate su LLM.

Visualizzare tutte le funzionalità lingua e traduttore

Visione + documento
Scopri informazioni e insights da documenti, immagini e video grazie a OCR e all'intelligenza artificiale multimodale.

Visualizzare tutte le funzionalità di visione e documentazione

Sicurezza dei contenuti
Rileva i contenuti dannosi, offensivi o inappropriati generati dall'utente o dall'intelligenza artificiale nell'app, tra cui testo, immagini e API multimodali.

Visualizzare tutte le funzionalità di sicurezza dei contenuti

Novità

Traduzione di documenti
Consente di tradurre documenti dalla lingua di origine alla lingua di destinazione da tipi di file come .docx, .pptx, .xlsx, .txt, .html e altro.

Garantire la sicurezza dei contenuti per l'IA generativa
Rilevare contenuti dannosi, offensivi o inappropriati generati dall'intelligenza artificiale nell'applicazione.

Estrai informazioni personali
Identificare e redigere le entità sensibili associate a una singola persona.

Demo: portale AI services

https://ai.azure.com/explore/aiservices

Azure AI Speech Service

- **Speech-to-Text (STT)** — trascrizione in tempo reale e batch
- **Text-to-Speech (TTS)** — sintesi vocale con voci neurali
- **Speech Translation** — traduzione vocale in tempo reale
- **Speaker Recognition** — identificazione e verifica del parlante
- **Pronunciation Assessment** — valutazione della pronuncia
- **Talking Avatar** — sintesi vocale con avatar animato

Avatar Parlante: Il Concetto

Un avatar parlante combina la sintesi vocale con una rappresentazione visiva per creare un personaggio interattivo che può conversare con gli utenti in tempo reale.

Consente un’interfaccia più naturale presentando le risposte tramite una persona digitale, anziché interazioni solo testuali.

Workflow

analizzatore di testo

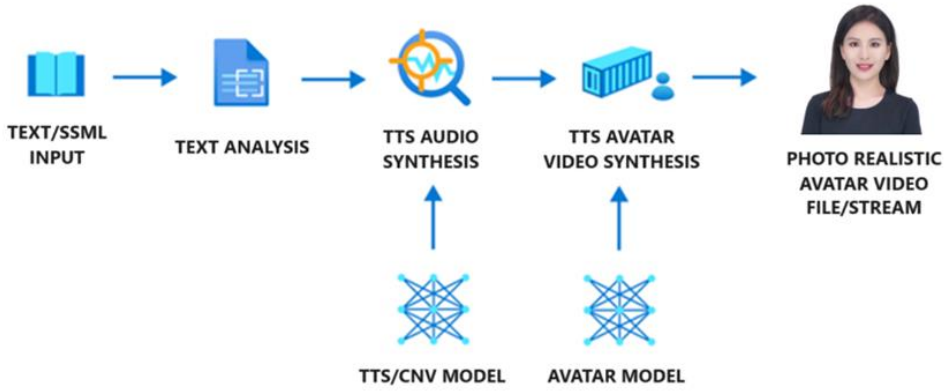
fornisce l'output sotto forma di sequenza di fonemi

il sintetizzatore audio TTS

predice le caratteristiche acustiche del testo di input e sintetizza la voce

sintetizzatore video avatar TTS

prevede l'immagine della sincronizzazione labiale con le caratteristiche acustiche, in modo da generare il video sintetico



Concetti principali

Architettura, gestione sicura dei token, flussi di connessione e demo: avatar di base e miglioramenti video.

WebRTC: streaming video in tempo reale

WebRTC (Web Real-Time Communication) è uno standard aperto che consente comunicazioni audio, video e dati **direttamente nel browser**, senza plugin.

Requisito	WebRTC
Bassa latenza	Streaming real-time (< 500ms)
Audio + Video	Canali multipli nativi
Nativo nel browser	Nessun plugin, nessuna installazione
Bidirezionale	Il server invia video, il client riceve

RTCPeerConnection: il cuore di WebRTC

RTCPeerConnection è l'oggetto JavaScript che gestisce l'intera connessione:

```
// Create WebRTC peer connection
let peerConnection = new RTCPeerConnection({
    iceServers: [{
        urls: [ iceServerUrl ],
        username: iceServerUsername,
        credential: iceServerCredential
    }],
    iceTransportPolicy: 'relay'
})

// Offer to receive 1 audio, and 1 video track
peerConnection.addTransceiver('video', { direction: 'sendrecv' })
peerConnection.addTransceiver('audio', { direction: 'sendrecv' })
```

ICE Token: attraversare firewall e NAT

Nella realtà, browser e server sono quasi sempre dietro **NAT** (Network Address Translation) o **firewall**. La connessione diretta peer-to-peer spesso **non è possibile**.

La soluzione: ICE framework

- **ICE** (Interactive Connectivity Establishment) è il protocollo che trova il **miglior percorso** di rete tra due peer:

Server	Ruolo
STUN	Scopre l'indirizzo IP pubblico del client (funziona se NAT è semplice)
TURN	Fa da **relay** — tutto il traffico passa attraverso il server TURN (funziona sempre)

ICE Token

```
{
    "Urls": ["turn:relay.communication.microsoft.com:443"],
    "Username": "credenziale-temporanea",
    "Password": "password-temporanea"
}
```

```
// Il browser usa il token per configurare RTCPeerConnection
fetch('/api/getIceToken').then(response => {
    response.json().then(data => {
        iceServerUrl = data.Urls[0]       // URL del server TURN
        iceServerUsername = data.Username  // Credenziali temporanee
        iceServerCredential = data.Password
    })
})
```

SDP : il "contratto" della connessione

SDP (Session Description Protocol) è il linguaggio con cui due peer WebRTC si dicono cosa sanno fare:

- Codec supportati (H264, VP8, Opus...)
- Indirizzi IP/porte per lo streaming
- Candidati ICE raccolti

SDP:la negoziazione (Offer/Answer)

Peer A crea un'offerta SDP Contiene:

1. codec supportati (VP8, H264, Opus…)
2. candidate ICE (IP/porte)
3. parametri RTP
4. modalità (sendrecv, recvonly…)

Peer A invia l'offerta tramite signaling (WebSocket, REST, SignalR, qualsiasi canale *fuori* da WebRTC)

Peer B riceve l'offerta e genera una risposta SDP (answer) Conferma:

1. quali codec accetta
2. quali candidate ICE usa
3. quali stream vuole inviare/ricevere

ICE + STUN/TURN trovano il percorso migliore Dopo la negoziazione SDP, parte la connessione P2P.

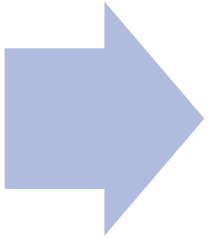
Architettura

Backend

ASP.NET Core, C#

Nuget: Microsoft.CognitiveServices.Speech

«orchestra la connessione, gestisce token e sintetizzatore vocale»



Frontend

HTML/CSS/JavaScript

Speech SDK per browser (https://aka.ms/csspeech/jsbrowserpackageraw)

«comunica via WebSocket con il servizio Azure TTS Avatar»

SDP : il codice

```javascript
// basic.js - Crea l'offerta e invia al backend
peerConnection.createOffer().then(sdp => {
    peerConnection.setLocalDescription(sdp)
})

// Quando l'ICE gathering è completato, invia al server
let localSdp = btoa(JSON.stringify(peerConnection.localDescription))
fetch('/api/connectAvatar', {
    method: 'POST',
    body: localSdp,
    headers: { 'AvatarCharacter': 'lisa', 'AvatarStyle': 'casual-sitting', ... }
}).then(response => {
    response.text().then(remoteSdp => {
        // Imposta la risposta del server → connessione stabilita
        peerConnection.setRemoteDescription(new RTCSessionDescription(JSON.parse(atob(remoteSdp))))
    })
})
```

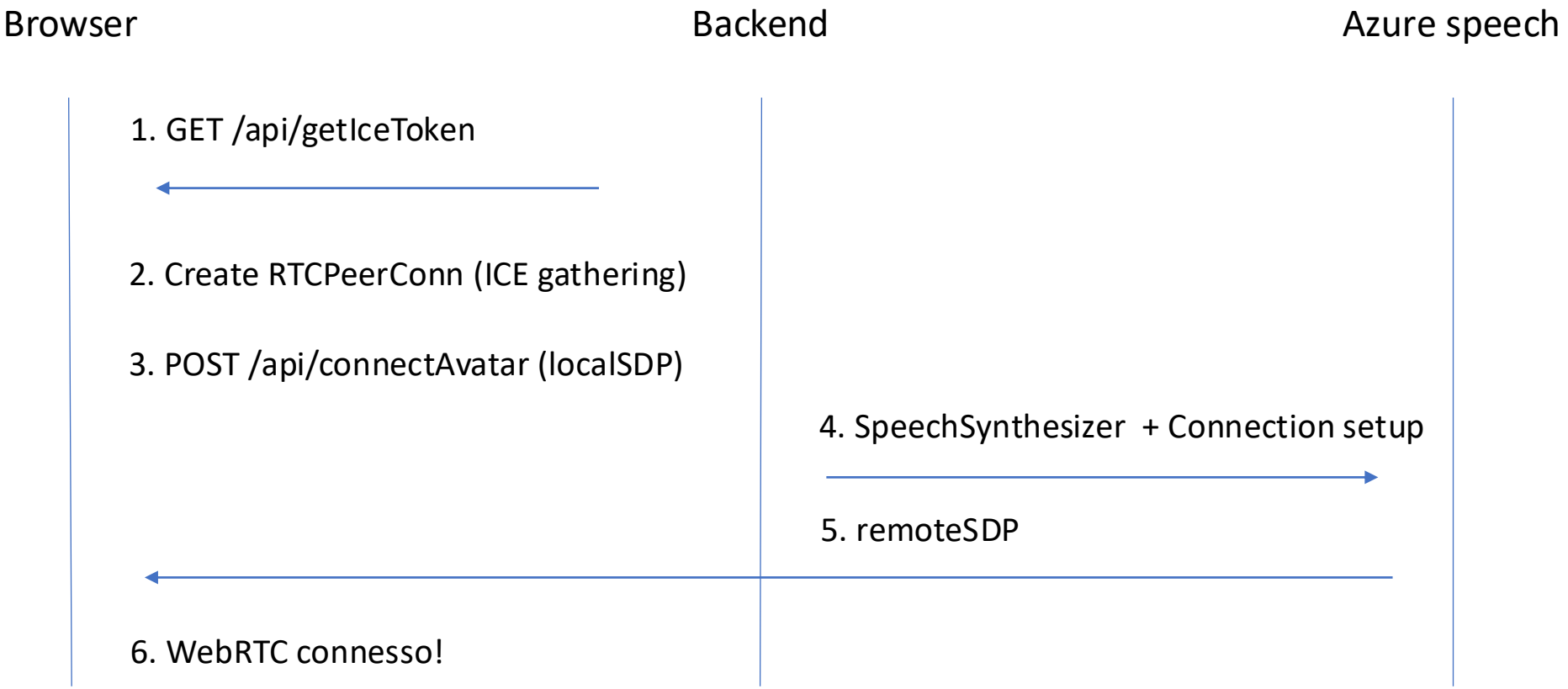
```csharp
// AvatarController.cs - Riceve localSDP, configura avatar, restituisce remoteSDP

// 1. Crea SpeechSynthesizer con configurazione WebRTC
var connection = Connection.FromSpeechSynthesizer(speechSynthesizer);
connection.SetMessageProperty("speech.config", "context",
    JsonConvert.SerializeObject(avatarConfig));  // Include ICE servers + avatar config

// 2. Avvia la connessione (invia empty text per trigger)
var result = speechSynthesizer.SpeakTextAsync("").Result;

// 3. Estrai il Remote SDP dalla risposta
var turnStartMessage = speechSynthesizer.Properties
    .GetProperty("SpeechSDKInternal-ExtraTurnStartMessage");
var remoteSdp = turnStartMessageJson?["webrtc"]?["connectionString"]?.ToString();
return Content(remoteSdp, "application/json");
```

SDP : il "contratto" della connessione

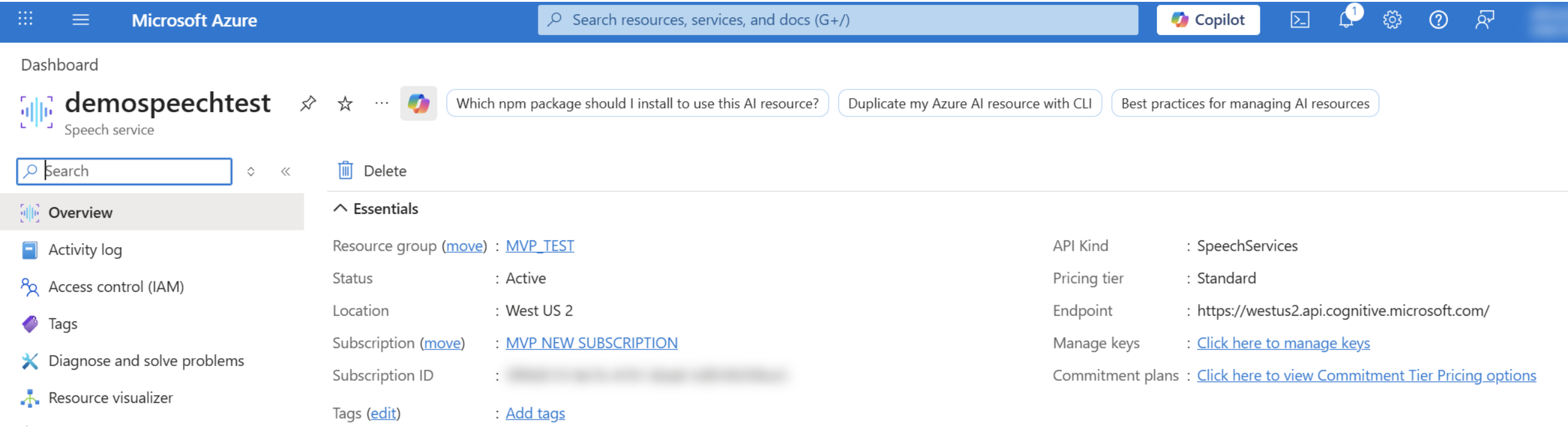
```
// Configurazione avatar inviata nella connessione

var talkingAvatar = new
{
    character = "lisa",              // Personaggio avatar
    style = "casual-sitting",        // Stile animazione
    background = new
    {
        color = "#FFFFFFFF",         // Colore sfondo
        image = new { url = "backgroundImageUrl" }
    }
};
```

Una volta connesso, il browser riceve **video** (avatar animato) e **audio** (voce TTS) come media stream, più un **data channel** per eventi

Architettura della Soluzione



Last step: keys



Demo: Avatar di base

Avatar conversazionale

Azure OpenAI, flussi di chat, RAG con Azure AI Search e Speech-to-Text.

AGGIUNGERE INTELLIGENZA: AZURE OPENAI

Cosa fa: I servizi Azure OpenAI abilitano comprensione e generazione avanzate del linguaggio naturale.

Perché è importante: Questa intelligence alimenta gli avatar con capacità di conversational AI, per interazioni più naturali e reattive.

Aggiungere intelligenza: Azure OpenAI

Demo: Chat Avatar

Customizzazione

Custom TTS

L'avatar di sintesi vocale personalizzato ti consente di creare un avatar parlante sintetico personalizzato e unico nel suo genere per la tua applicazione. Con l'avatar personalizzato di sintesi vocale, puoi creare un avatar unico e dall'aspetto naturale per il tuo prodotto o marchio fornendo i dati di registrazione video degli attori selezionati. Se crei anche una voce neurale personalizzata per lo stesso attore e la usi come voce dell'avatar, l'avatar sarà ancora più realistico.

La creazione di un avatar di sintesi vocale personalizzato richiede almeno 10 minuti di registrazione video dell’attorecome dati di addestramento ed è necessario prima ottenere il consenso dell’ attore.

https://learn.microsoft.com/en-us/azure/ai-services/speech-service/text-to-speech-avatar/what-is-custom-text-to-speech-avatar

Custom TTS steps

Ottieni video di consenso

- ottieni una registrazione video della dichiarazione di consenso. La dichiarazione di consenso è una registrazione video del talento avatar che legge una dichiarazione, dando il consenso all'utilizzo della propria immagine e dei dati vocali per addestrare un modello di avatar vocale personalizzato.

Preparare i dati di allenamento

- assicurarsi che la registrazione video sia nel formato corretto. È una buona idea girare la registrazione video in uno studio di ripresa video di qualità professionale per ottenere un'immagine di sfondo pulita. La qualità dell'avatar risultante dipende fortemente dal video registrato utilizzato per l'allenamento. Fattori come la velocità di parola, la postura del corpo, l'espressione facciale, i gesti delle mani, la coerenza nella posizione dell'attore e l'illuminazione della registrazione video sono essenziali per creare un accattivante avatar di sintesi vocale personalizzato.

Addestrare il modello avatar

- inizieremo ad addestrare il modello di sintesi vocale personalizzato dopo aver verificato la dichiarazione di consenso del talento avatar. Nella fase di anteprima di questo servizio, questo passaggio verrà eseguito manualmente da Microsoft. Riceverai una notifica una volta che il modello sarà stato addestrato con successo.

Distribuisci e utilizza il tuo modello di avatar nelle tue APP

https://learn.microsoft.com/en-us/azure/ai-services/speech-service/text-to-speech-avatar/custom-avatar-record-video-samples

Batch synthesis

L'API di sintesi batch per avatar di sintesi vocale consente la sintesi asincrona del testo in un avatar parlante come file video. Gli editori e le piattaforme di contenuti video possono utilizzare questa API per creare contenuti video avatar in batch. Questo approccio può essere adatto a vari casi d'uso come materiali di formazione, presentazioni o pubblicità.

Il video avatar sintetico verrà generato in modo asincrono dopo che il sistema avrà ricevuto l'input di testo. L'output video generato può essere scaricato in modalità di sintesi batch. Invii il testo per la sintesi, esegui il polling sullo stato della sintesi e scarichi l'output video quando lo stato indica il successo. I formati di input del testo devono essere testo normale o testo SSML (Speech Synthesis Markup Language).

Batch synthesis - API

Operation	Method	REST API call
Create batch synthesis	PUT	avatar/batchsyntheses/{SynthesisId}?api-version=2024-04-15-preview
Get batch synthesis	GET	avatar/batchsyntheses/{SynthesisId}?api-version=2024-04-15-preview
List batch synthesis	GET	avatar/batchsyntheses/?api-version=2024-04-15-preview
Delete batch synthesis	DELETE	avatar/batchsyntheses/{SynthesisId}?api-version=2024-04-15-preview

Operazioni e adozione

Operazioni multi-client, ottimizzazione della latenza, costi, scenari e risorse per l’adozione.

Scenari reali

Scenario	Descrizione
Info Point / Kiosk	Totem informativi in aeroporti, musei, fiere
Assistenti virtuali	Customer care con volto umano
E-learning	Tutor virtuali che spiegano contenuti
HR & Onboarding	Assistente per nuovi dipendenti
Healthcare	Supporto informativo per pazienti
Accessibilità	Comunicazione per persone con disabilità uditive (lip reading)
Marketing	Testimonial virtuali personalizzati

Disponibilità

Region	Real-time avatar	Batch avatar	Custom avatar	Custom avatar training[1]
eastus2	✅	✅	✅	
northeurope	✅	✅	✅	
southcentralus	✅	✅	✅	
southeastasia	✅	✅	✅	✅
swedencentral	✅	✅	✅	
westeurope	✅	✅	✅	✅
westus2	✅	✅	✅	✅

https://learn.microsoft.com/en-us/azure/ai-services/speech-service/regions?tabs=ttsavatar

Costi e considerazioni

Comprendere i modelli di prezzo e il consumo di risorse aiuta nel budgeting e nella pianificazione per distribuire avatar intelligenti su larga scala.

Considera l'inferenza del modello e l'uso dei token; lo streaming multimediale in tempo reale e la larghezza di banda; storage, logging e analytics; orchestrazione, hosting e autoscaling.

Pianifica controlli di sicurezza e conformità; monitoraggio e supporto; licenze e impegni con i fornitori; guardrail dei costi (quote, caching e avvisi di utilizzo).

Prezzi (WE)

Area: Europa settentrionale

Valuta: Zona euro - Euro (€) EUR

1 USD = 0.868 EUR

Text to Speech[7]	Standard Voice	Neural (real-time and batch): **€13,021** per 1M characters Neural HD (real-time and batch)[4]: N/D
	Custom Voice	Professional Voice: Synthesis (real-time and batch): **€20,833** per 1M characters Synthesis (neural HD real-time and batch): **€41,665** per 1M characters Voice model training: **€45,137** per ora di calcolo, up to **€812,465** per training Endpoint hosting: **€3,50** per modello all'ora
		Personal Voice[6]: Synthesis (real-time and batch): N/D Voice creation: Free Voice profile storage: N/D
	Enhanced Add-on feature: Avatar	Standard: Interactive avatar (real-time): **€0,435** per minute Interactive 4K avatar (real-time): **€0,608** per minute Avatar video (batch): **€0,869** per minute 4K avatar video (batch): **€1,172** per minute
		Custom: Avatar model training: N/D Interactive avatar (real-time): **€0,521** per minute Interactive 4K avatar (real-time): **€0,695** per minute Avatar video (batch): **€1,737** per minute 4K avatar video (batch): **€2,344** per minute Endpoint hosting: **€0,521** per model per hour

Prezzi (WE) Old

Text to Speech[8]	Standard Voice	Neural: **€13.871** per 1M characters Neural HD[4]: N/A per 1M characters
	Custom Voice	Professional Voice: Synthesis: **€22.194** per 1M characters Voice model training: **€48.086** per compute hour, up to **€4,616.239** per training Endpoint hosting: **€3.73** per model per hour
		Personal Voice[6]: Synthesis: **€22.194** per 1M characters Voice creation: Free Voice profile storage: **€554.837** per 1,000 voice profiles per month
	Enhanced Add-on feature: Avatar[Preview]	Standard: **€0.925** per minute
		Custom: Real-time synthesis: **€0.925** per minute Batch synthesis: **€1.850** per minute Endpoint hosting: **€0.6** per model per hour

Conclusione

Punto chiave: Creare avatar intelligenti con Azure text-to-speech dà vita alle interazioni con l’AI, migliorando il coinvolgimento degli utenti e aprendo nuove possibilità per la comunicazione e l’erogazione dei servizi.

Domande?

Contatti

Alessio Iafrate

X: alessioiafrate

Github: a-iafrate

Linkedin: alessio-iafrate

Sessionize: alessio-iafrate

Email: alessioiafrate@hotmail.com

#GLOBALAZURE
2026
#Milano

Slide e video:
https://www.globalazuremilano.it