STT – TTS

Mohammad Mahdi Seyedmomeni

مدل : STT

- **اسم مدل : Speechmatics**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

Speechmatics **یک مجوز غیرانحصاری، غیرقابل انتقال به کاربران می‌دهد برای استفاده از نرم‌افزار و** API **در** «Subscription Term»**.**

- **هزینه های** API**:**

کاربران جدید می‌توانند با طرح Free **شروع کنند: تا** 480 **دقیقه رایگان در ماه برای سرویس** Speech-to-Text**.**

بر اساس سایت قیمت‌ها: برای سرویس ابری (cloud/API)**:**

طرح «Pro» **از حدود** US $0.24 **/ ساعت برای تراکنش شروع می‌شود.**

گزارش‌های دیگر: مثلاً حدود US$0.30**/ساعت برای** «batch transcription» **و** ≈ US$1.04**/ساعت برای** «real-time transcription» **ذکر شده‌اند.**

- **اجرای محلی :**

Speechmatics **امکاناتی برای اجرای محلی** (on-premises) **یا درون محیط سازمانی/سرور داخلی دارد: به عنوان مثال** «Virtual Appliance» **و** «Container» **برای** batch **یا** real-time**.**

- **هزینه محلی:**

قیمت پایه یا نرخ (مثلاً دلار/سال) برای مجوز on-premises **به‌صورت عمومی منتشر نشده است.**

هزینه اجرای محلی شامل چه مواردی می‌شود (مثلاً نرم‌افزار، پشتیبانی، به‌روزرسانی، تعداد نشست‌ها، تعداد کاربر، نوع سخت‌افزار) به‌صورت شفاف درج نشده است.

- **دقت فارسی:**

بسیار خوب (در آزمایش‌ها بهترین عملکرد بین مدل‌های تجاری داشت).

- **سرعت/تاخیر: کم؛ پردازش تقریباً هم‌زمان** (near real-time).
- **واژگان سفارشی: دارد ،می‌توان لیست کلمات خاص (نام برند، اصطلاح فنی و...) را اضافه کرد تا دقیق‌تر تشخیص دهد.**

مدل : STT

- **اسم مدل : ElevenLabs**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

کاملاً مالکیتی (Closed-source, Proprietary)

سورس‌کد و مدل‌ها در دسترس عموم نیستند.

استفاده فقط از طریق API **رسمی یا پلتفرم وب ممکن است**

- **هزینه های** API:

قیمت برای TTS/**ولید صوت: پلان‌های مختلف دارند: رایگان،** Starter، Creator، Pro، Scale، Business، Enterprise.

Free plan = 10,000 **اعتبار/ماه (برای** TTS **حدود ۱۰ دقیقه صوت)**

- **اجرای محلی :**

در منابع عمومی، اجرای محلی (On-Premises) **برای** ElevenLabs **به‌صورت رسمی و واضح نیافته‌ام یعنی اطلاعاتی درباره«نصب در سرور داخلی»، «مجوز جداگانه»، یا «هزینه ثابت اجرای محلی» به‌طور عمومی درج نشده است.**

- **هزینه محلی:**

قیمت دقیق محلی عمومی نیست؛ برای مشتریان سازمانی با استعلام تعیین می‌شود.

- **دقت فارسی: بسیار بالا؛ برای فارسی مدل «۷٫۸٪ WER» گزارش شده است.**
- **سرعت/تاخیر: مناسب برای فایل‌های ضبط‌شده ولی هنوز برای «گفتار زنده با تاخیر بسیار کم» بهینه نشده است.**

مدل : STT

- **اسم مدل : Gladia**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

Gladia یک سرویس مالکیتی (proprietary) است؛ سورس-کد مدل یا امکان دانلود آزاد مدل در دسترس عمومی نیست.

- **هزینه های API:**

طرح رایگان: تا ۱۰ ساعت/ماه برای توسعه‌دهندگان.

طرح «Pay-as-you-go / Scaling»: از حدود 0.50 $/ساعت (یا 0.55$/ساعت) برای سرویس‌های asynchronous/real-time شروع شده است.

طرح Enterprise: قیمت «با استعلام» و ترتیبات سفارشی دارد.

- **اجرای محلی :**

Gladia در سایت خود اعلام کرده است که امکان «custom hosting / on-premises / air-gapped hosting» برای مشتریان سازمانی وجود دارد.

- **هزینه محلی:**

قیمت ثابت و عمومی برای این گزینه منتشر نشده است؛ یعنی برای هزینه اجرای محلی باید با تیم فروش تماس گرفته شود.

- **دقت فارسی: خوب، ولی نه در حد Speechmatics یا Whisper؛ برای فارسی به اندازه انگلیسی بهینه نشده.**
- **سرعت/تاخیر: بالا (پردازش لحظه‌ای ابری).**

- **واژگان سفارشی: دارد ،در مستندات API امکان افزودن glossary برای اصطلاحات خاص ذکر شده.**

مدل : STT

- **اسم مدل : Whisper**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

کد مدل و وزن‌های مدل Whisper تحت مجوز MIT License منتشر شده‌اند.

چون MIT یک مجوز آزاد (permissive open-source) است، به این معنی است که می‌توان:

مدل را دانلود و به‌صورت محلی اجرا کرد (در بسیاری از موارد) اصلاحش کرد، توزیع کرد (تحت شرایط حفظ حق کپی‌رایت و مجوز مطابق باشد)

برای مقاصد تجاری استفاده کرد، به شرطی که الزامات MIT رعایت شود.

بنابراین Whisper از منظر مجوز اوپن‌سورس است و برخلاف مدل‌های کاملاً مالکیتی، امکان استفاده محلی با آزادی بیشتری دارد.

- **هزینه های API:**

اگر از سرویس API.OpenAI برای Whisper استفاده شود، قیمت‌هایی گزارش شده‌اند: مثلاً حدود 0.006$ به ازای هر دقیقه صوت برای ترانویسی گفته شده است

- **اجرای محلی :**

از آنجا که مدل Whisper اوپن‌سورس است، می‌توان آن را به صورت محلی اجرا کرد (روی سرور خودتان) بدون نیاز به پرداخت هزینه API یا مجوز از OpenAI، اگر فقط از کد و وزن تحت MIT استفاده شود. مثال‌ها و راهنمایی‌هایی برای اجرای محلی وجود دارد.

- **هزینه محلی:-**
- **دقت فارسی:**

بسیار بالا (در نسخه‌های large و medium مخصوصاً). مدل فارسی را ذاتاً پشتیبانی می‌کند و با لهجه‌ها سازگار است.

- **سرعت/تاخیر: بسته به سخت‌افزار؛ روی GPU نسبتاً سریع، اما روی CPU سنگین است.**
- **واژگان سفارشی: به‌صورت مستقیم پشتیبانی نمی‌کند؛ فقط می‌توان با fine-tuning یا prompt-engineering کیفیت واژه‌های خاص را بهبود داد.**

مدل : STT

- **اسم مدل : Vosk**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

Vosk تحت مجوز Apache License 2.0 منتشر شده است.به این معنی که منبع باز (open-source) است، می‌توان مدل‌ها و کد را به صورت محلی نصب و اجرا کرد با شرایط این مجوز (مثلاً حفظ اعلامیه‌های مجوز، بدون الزام به پرداخت حق مجوز).مدل‌ها و کتابخانه‌ها و سرور Vosk-Server نیز با همین یا مجوز سازگار در دسترس‌اند.

- **هزینه های API:-**
- **اجرای محلی :**

یکی از نقاط قوت Vosk این است که قابل اجرا کاملاً محلی است روی سرور داخلی، یا حتی روی دستگاه‌های کم‌مصرف مانند Raspberry Pi.

- **هزینه محلی:**

چون Vosk اوپن‌سورس است، هزینه مجوز مستقیم صفر است (برای استفاده طبق مجوز Apache 2.0).

ولی هزینه واقعی شامل موارد زیر می‌شود:

هزینه سخت‌افزار (سرور، پردازشگر، حافظه، ذخیره‌سازی)

هزینه مدیریت، نگهداری، به‌روزرسانی مدل‌ها

هزینه نیروی انسانی (نصب، راه‌اندازی، تست و بهینه‌سازی)

دقت فارسی: متوسط تا خوب؛ بسته به مدل آکوستیک فارسی که استفاده می‌کنی.

- **سرعت/تاخیر: بسیار سریع روی CPU، مناسب اجرای آفلاین.**
- **واژگان سفارشی: دارد ،می‌توان لغات خاص را به دیکشنری مدل اضافه کرد.**

مدل : TTS

- **اسم مدل : ElevenLabs**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

ElevenLabs **سرویس** TTS **و صداسازی مبتنی بر مالکیتی** (proprietary) **است، نه سورس‌باز.**

- **هزینه های** API:

پلان‌های قیمت برای TTS **در** ElevenLabs **به‌صورت زیر هستند:**

Free: **رایگان، چند دقیقه/چند کاراکتر متن.**

Starter: **حدود** 5 $/**ماه با نمونه‌های محدودی از** TTS.

Creator: ~$22/**ماه برای حجم بیشتر.**

Pro: ~$99/**ماه برای حجم بزرگ‌تر.**

Scale / Business: **بالاتر از این برای حجم‌های بسیار زیاد (~**$330/**ماه و** $1320/**ماه) برای میلیون‌ها اعتبار.**

- **اجرای محلی :**

ElevenLabs **در صفحه** Enterprise **خود اعلام کرده است که برای سازمان‌ها «** Elevated support and custom deployments**» دارند. در بلاگشان نیز در مورد «**hybrid cloud**» نوشته‌اند که می‌توان ترکیبی از** Private Cloud / On-Premises + Public Cloud **داشت تا انعطاف بیشتری برای سازمانها فراهم شود.**

- **هزینه محلی:**

برای اجرای محلی (on-premises**) یا استقرار سفارشی در سرور سازمانی) قیمت عمومی و شفاف ندارد؛ فقط نوشته شده «**Contact Sales / Custom Pricing**» برای پلان** Enterprise.

- **طبیعی بودن: بسیار بالا**

جزو طبیعی‌ترین صداها در میان TTSها است (از نظر آهنگ، مکث‌ها، احساس و تأکید).

مدل‌های Flash و v3 prosody واقعی و near-human ارائه می‌دهند.

- **گوینده‌ها: بسیار متنوع**

بیش از 100 صدای از پیش آماده در زبان‌های مختلف.

قابلیت کلون صدای واقعی (voice cloning) در پلن‌های Creator و بالاتر.

برای فارسی نیز چند صدای طبیعی دارد، ولی تنوع فارسی کمتر از انگلیسی است.

مدل : TTS

- **اسم مدل : LOVO AI**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

LOVO AI یک سرویس مالکیتی (commercial, proprietary) است. سورس‌کد مدل به‌صورت عمومی آزاد نشده است.استفاده از پلتفرم و تولید صداها با مجوز تجاری امکان‌پذیر است.

- **هزینه های API:**

بر اساس پلن–های عمومی LOVO AI:

پلن Basic: حدود US $24/ماه (در سال اول) بُرای «۲ ساعت تولید صدا در ماه».

پلن Pro: قیمت بالاتر (~US $48/ماه بر اساس برخی منابع) برای حدود ۵ ساعت تولید صدای ماهانه.

پلن Pro+: ~US $75/ماه (یا تا ~US $149/ماه بسته به تخفیف) برای حدود ۲۰ ساعت تولید صدا در ماه..

- **اجرای محلی :**

امکان اجرای محلی (سرور داخل سازمان) برای مشتریان سازمانی در دسترس است اما قیمت عمومی برای این گزینه مشخص نیست، بلکه «Contact Sales» ذکر شده است.

- **هزینه محلی:**

ذکر نشده

- **طبیعی بودن: بالا**

از مدل‌های neural TTS استفاده می‌کند و احساس و زیر و بمی گفتار (prosody) را به‌خوبی منتقل می‌کند.

صداها نسبتاً گرم‌تر از ElevenLabs ولی کمی مصنوعی‌تر در جمله‌های بلند.

- **گوینده‌ها: زیاد**

بیش از 400 صدای آماده در 100+ زبان.

امکان انتخاب گوینده و تنظیم جنس، سن، و لحن.

در فارسی چند گوینده موجود است ولی کیفیتشان متفاوت است (برخی robotic-ترند).

مدل : TTS

- **اسم مدل : Tihu**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

Tihu یک پروژه متن‌باز (open-source) مخصوص زبان فارسی است: در گیت‌هاب منتشر شده تحت نام «tihu-nlp/tihu». استفاده از پلتفرم و تولید صداها با مجوز تجاری امکان‌پذیر است.

- **هزینه های API:**

این نرم‌افزار متن-باز است و مجوز آن اوپن‌سورس است، بنابراین هزینه فعلی برای مجوز نرم‌افزار وجود ندارد.

- **اجرای محلی :**

از آنجا که Tihu متن‌باز است، اساساً اجرای محلی به سادگی ممکن است: شما می‌توانید کد را در سرور خودتان کامپایل و اجرا کنید.

- **هزینه محلی:**

این نرم‌افزار متن-باز است و مجوز آن اوپن‌سورس است، بنابراین هزینه فعلی برای مجوز نرم‌افزار وجود ندارد.

- **طبیعی بودن: متوسط**

صدای تولیدی واضح ولی نسبتاً یکنواخت و بدون کنترل احساس یا لحن.

مناسب کاربردهای ساده (اعلان، راهنما) نه گفتار طبیعی انسانی.

- **گوینده‌ها: محدود**

یک یا چند صدای پایه (زن و مرد) دارد.

بدون امکان تنظیم لحن یا سبک گفتار.

مدل : TTS

- **اسم مدل : Coqui**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

بخش کد ابزار Coqui TTS **تحت مجوز** Mozilla Public License 2.0 (MPL-2.0) **منتشر شده است.**

در مورد مدل‌های پیش‌آماده (pre-trained models) **یا بخش‌های خاص مانند** XTTS**، مجوز متفاوتی به نام** Coqui Public Model License (CPML) **وجود دارد که ممکن است برای کاربرد تجاری محدودیت داشته باشد.**

- **هزینه های** API:

«قیمت‌های اشتراک» برای "Coqui AI**" را به‌عنوان سرویس تجاری بیان کرده است (ممکن است کاملاً دقیق شرکت رسمی نباشد): مثلاً** Free**،** ~$9.9 Starter**/ماه،** ~$19.9 Creator**/ماه،** ~$69.9 Pro**/ماه.**

- **اجرای محلی :**

یکی از نقاط قوت Coqui **این است که ابزار و مدل آن قابلیت اجرا محلی را دارند**

- **هزینه محلی:**

قیمت مشخص عمومی برای مجوز اجرای محلی یا پشتیبانی سازمانی Coqui TTS **یافت نشد.**

گرچه مدل ابزار آزاد است، اما برای مدل‌های پیش‌آماده تجاری یا تجاری سازی ممکن است هزینه مجوز داشته باشد.

- **طبیعی بودن: خوب تا خیلی خوب**

مدل‌های neural-based **با** prosody **طبیعی (به‌ویژه** XTTS-v2**).**

برای زبان‌هایی مثل انگلیسی و اسپانیایی عالی است، ولی فارسی پشتیبانی رسمی ندارد.

- **گوینده‌ها: زیاد**

پشتیبانی از multi-speaker **و** cloning**.**

می‌توان گوینده جدید آموزش داد (speaker embedding)**.**

برای فارسی نیاز به آموزش سفارشی دارد.

مدل : TTS

- **اسم مدل : eSpeak NG**
- **پشنیبانی به زبان فارسی : بله**
- **وضعیت مجوز :**

eSpeak NG **تحت مجوز** «GNU General Public License v3 (GPLv3) **منتشر شده است.**

این یعنی نرم‌افزار آزاد (open-source) **است، ولی یکی از ویژگی‌های** GPLv3 **این است که اگر شما نرم‌افزار را با بخش‌های دیگری لینک یا ترکیب کنید که تجاری هستند، ممکن است مجبور باشید کل محصول را تحت** GPL **منتشر کنید یا مجوز متفاوت بگیرید.**

- **هزینه های API:**

چون این نرم‌افزار آزاد است و می‌توان آن را محلی نصب کرد، بیشتر کاربران از آن به صورت محلی یا در پروژه‌های کوچک استفاده می‌کنند و هزینهٔ مجوز نرم‌افزاری ندارد.

- **اجرای محلی :**

از آنجا که eSpeak NG آزاد است، امکان اجرای محلی آن تقریباً بی‌قید است: شما می‌توانید آن را دانلود، کامپایل و روی سرور یا دستگاه خودتان نصبش کنید.

- **هزینه محلی:**

هزینه نرم‌افزار: صفر (کاملاً رایگان و متن‌باز تحت GPLv3)

هزینه اجرای محلی: فقط هزینه سخت‌افزار (CPU معمولی کافی است، نیاز به GPU ندارد)

- **طبیعی بودن: پایین**

سنتز مبتنی بر فرمات (formant synthesis)، آهنگ گفتار یکنواخت و ماشینی.

- **بدون prosody واقعی.**
- **گوینده‌ها: بسیار محدود**

چند صدای ماشینی ساده (بدون احساس یا تنوع).

تغییر صدا صرفاً با پارامترهای pitch/speed ممکن است، نه گوینده انسانی.

بررسی نتیجه نهایی مدل های STT

Whisper large-v3 و large-v3-turbo

جملات به‌هم‌ریخته‌اند، مثلاً در جمله‌ها حروف فارسی و اشتباهات ترکیبی مثل «یقهوه»، «یهست ییه دفعه گفت» و «یماش یویرفتم» زیاد دیده می‌شود.

کلمات به هم چسبیده یا ناقص‌اند و برخی جملات عملاً غیرقابل خواندن‌اند.

روانی و انسجام معنایی پایین است.

مشکل اصلی: پردازش اشتباه فاصله و segmentation **در زبان فارسی (چون مدل** Whisper **فارسی را به‌صورت اسکریپت لاتین یا تلفظی یاد گرفته است).**

Speechmatics (standard & enhanced)

متن خروجی بسیار منسجم‌تر است. جملات به‌صورت پیوسته آمده‌اند.

ساختار جمله درست‌تر است و بسیاری از جملات کامل هستند، مثلاً:

«یه لحظه به خودم گفتم کاش با مترو می‌رفتم. رادیو یه لحظه به خودم گفتم کاش با مترو میک وحشتناک بود. بیرون طبق معمول ترافیک...»

این‌ها خوانا هستند، حتی اگر علائم نگارشی کم باشد.

نسخه‌ی enhanced **نسبت به** standard **کمی روان‌تر و طبیعی‌تر است.**

خطاهای تایپی وجود دارد ولی معنی گفتار تقریباً حفظ شده است.

در فارسی غیررسمی، خروجی واقعاً نزدیک به گفتار است.

Vosk

دقت واژگانی بالاست (کلمات درست‌اند)،

ولی جمله‌سازی ساده است، مثلاً:

«اولش مامان زنگ زد گفت یادت نره امروز بری قبض برق پرداخت کنیم بعدش رئیس شرکت پیام داد که جلسه صبح زود شروع میشه...»

از نظر معنا درست است ولی گاهی توقف‌گاه‌ها، ضمایر و تن صدا را حذف می‌کند.

جمله‌ها خیلی ماشینی‌اند، روانی طبیعی گفتار را کمتر دارند.

برای کاربرد ماشینی (تحلیل داده) عالی است، اما برای متن طبیعی یا زیرنویس، کمی خشک است.

مدل	دقت واژگان	روان بودن	انسجام معنایی	مناسب ترین کاربرد
Speechmatics enhanced	90% – 85%	بالا	بالا	متن طبیعی و گفتار محاوره‌ای
Vosk	95%	متوسط	کمی خشک	استفاده‌ی تحلیلی یا آفلاین
Speechmatics standard	80%	خوب	خوب	محلی و سبک‌تر از enhanced
Whisper large-v3-turbo	70%	پایین	پایین	فقط برای تست سرعت
Whisper large-v3	65%	پایین	پایین	فقط برای آزمایش اولیه

جمع بندی :

Speechmatics **(ویژه نسخه** enhanced**) از نظر روانی و نزدیک بودن به گفتار واقعی عملکرد بهتری نسبت به** Whisper **داشته است،**

و فقط کمی از نظر دقت واژگان از Vosk **پایین‌تر است.**

بنابراین اگر هدف تولید خروجی خوانا و شبیه گفتار انسانی باشد ، Speechmatics enhanced **برتر است.**

اما اگر هدف دقت عددی بالا برای تحلیل یا جستجو باشد ، Vosk **دقیق‌تر است.**