التقرير التقني لتقييم نماذج تحويل النص لكلام (TTS) باللغة العربية على منصة Hugging face .

المقدمة :

يهدف هذا التقرير إلى تحديد أحد أفضل نماذج تحويل النص إلى كلام (Text-to-Speech - TTS) المتاحة للغة العربية على منصةHugging Face ، واقتراح إطار تقييم شامل لمقارنته مع نموذجين آخرين. الهدف النهائي هو تحديد النموذج الذي يحقق أفضل توازن بين دقة النطق، والطبيعية، ودعم اللهجات، والكفاءة التشغيلية، وقابلية الاستخدام العملي في تطبيقات توليد الكلام العربي.

افتراض :تعتمد مؤشرات الشعبية (عدد التنزيلات والإعجابات ومستوى التبني المجتمعي) على المعلومات العامة المتاحة في منظومة Hugging Faceخلال الفترة 2025–2026، وقد تتغير هذه الأرقام مع مرور الوقت

النماذج المختارة :

SILMA TTS (الأفضل) :

رابطه : https://huggingface.co/silma-ai/silma-tts

يُعد SILMA TTS من أقوى النماذج المخصصة للغة العربية المتوفرة حاليًا على منصة .Hugging Face يعتمد النموذج على معمارية F5-TTS الحديثة، وتم تطويره خصيصًا لإنتاج كلام عربي وإنجليزي عالي الجودة و نظرًا لتخصصه في اللغة العربية واعتماده على معمارية حديثة وازدياد تبنيه من المجتمع، تم اختياره كنموذج رئيسي للمقارنة.

XTTS-V2 (نموذج المقارنة الأول) :

رابطه : https://huggingface.co/coqui/XTTS-v2

يُعد XTTS-v2 أحد أكثر نماذج تحويل النص إلى كلام متعددة اللغات انتشارًا واستخدامًا , يوفر XTTS-v2 مقارنة مهمة لأنه يمثل نهجًا متعدد اللغات بدلًا من كونه نموذجًا متخصصًا بالعربية.

NAMAA Saudi TTS V2 (نموذج المقارنة الثاني) :

رابطه : https://huggingface.co/NAMAA-Space/NAMAA-Saudi-TTS-V2

يركز نموذج NAMAA Saudi TTS V2 على توليد الكلام العربي مع اهتمام خاص باللهجة السعودية والخليجية.

تنبع أهميته من قدرته على اختبار مدى دعم النموذج للهجات العربية المختلفة.

منهجية التقييم المقترحة :

لضمان عدالة المقارنة، يجب اختبار جميع النماذج باستخدام نفس مجموعة النصوص ونفس إعدادات التوليد ونفس بيئة التشغيل العتادية وسنقسم المنهجية لعدة رؤوس أقلام كالتالي :

تقييم دقة نطق الأصوات العربية الخاص :

وهي الأحرف المميزة التي يصعب إيجاد أصواتها في لغات أخرى مثل (ع , غ , ص , ض ,ء , الخ..) , أما بالنسبة لمجموعة الختبار فيجب أن تحتوي على عدة حالات مختلفة سواء كلمات مفردة او ازواج من الكلامات أو جمل قصير او طويلة بالإضافة للعبارات صعبة النطق.

ويتم تقييم ذلك بعدة طرق أولها وأكثرها قيمة ولكن الأصعب أيضا هو التقييم البشري اذ تططلب على الأقل 20 متحدثا باللغة العربية ويفضل من عدة لهجات . ثاني الطرق هي المحاذاة القسرية أو (Forced Alignment) اذ يتم مقارنة التسلسل الصوتي المتوقع بالتسلسل الصوتي الناتج من النموذج باستخدام أداة المحاذاة القسرية . وأخيرا هناك التقييم باستخدام الASR وهي عبارة عن الة تعرف كلامي تقوم بحساب معدل خطأ الكلمات (WER) ومعدل خطأ المحارف (ASR) وبشكل طبيعي كلما انخفضت هذه القيم كان ذلك أفضل.

تقييم الطبيعة والإيقاع (Prosody) :

تشير الطبيعية إلى مدى قرب الكلام الناتج من كلام الإنسان الحقيقي , أما الإيقاع فيشمل التنغيم والايقاع الزمني وسلاسة التدفق الكلامي والقوة التعبيرية .

يتم تقييم ذلك بشكل مشابه للتجربة السابقة بداية باستخدام مستمعين بشر بعدد مشابه بين 20 و 30 مستمع ويمكن وضع معيار تقييم بين ال 1 وال 5 حيث سيقيمون سهولة فهمهم ومدا طلاقة الكلام وسهولة الاستماع . ويمكن أيضا استخدام تقييم موضوعي باستخدام معايير مثل Mel Cepstral Distpration(MCD) والذي يقوم بقياس المسافة بين الصوت الناتج والصوت المرجعي . كما ويمكن تحليل الايقاع ومقارنتها بالتسجيلات البشرية المتوفرة لملاحظة أي فروقات بمنحنيات طبقة الصوت وتوزيع الطاقة الصوتية أو مدا امتداد وطول كل مقطع صوتي.

تقييم التشكيل :

وهذا من أصعب التحديات اذ أنه من المعتدات كتابة الكلمات بدوت تشكيل لذلك تمرين النماذج على توقع التشكيل الصحيح يعد تحديا حقيقيا ولكن أرى أن أفضل طريقة هي وضع كل كلمتين بتشكيل مشابه بشكل زوج لكي يتدرب النموذج على الفرق بينهما بشكل قطعي وهذا سوف يعطينا معيارا سهلا للتقيم لاحقا اذ نستطيع الحصول على نسبة مؤية لدقة الأزواج المنطوقة

تقييم اللهجات العامية مقابل الفصحى :

وهذه ميزة أخرى من ميزات اللغة العربية والتي تزيد صعوبة تدريب النماذج بشكل كبير فمن السهل نسبيا تدريب نموذج على الفصحى ولكن من الصعب جعله قادرا على تمييز جميع اللهجات ونطقها بشكل صحيح وبرأيي هذا يتطلب عددا أكبر بكثير من المراجعات البشرية ربما أكثر من 100 شخص على الأقل وفي النهاية يمكن لنا أن نستخدم المعيار MOS (متوسط تقييم المستمعين)والذي نخصصه للهجات ويتم احتساب متوسط تقييم كل لهجة بشكل مستقل .

معامل الزمن الحقيقي(RTF) :

كما وأنه من الضروري أن يكون النموذج يولد ويتكلم بسرعة منطقية ومطابقة قدر الامكان للبشر فهذا المعيار هو عبارة عن زمن التوليد على طول الصوت المولد أو RTF = Synthesis time / Audio duration .وكلما كان الرقك أقرب ل 1 كلما كان ذلك أفضل .

اختبار المزج اللغوي :

كثير من الأحيان يدمج العرب بكلامهم كلمات من اللغات الأجنبية في حديثهم اليومي فانه من الجيد أن يكون هنالك معيار أيضا لتقييم النموذج من هذه الناحية وهذا سيعني قدرة النموذج على تمييز عدة لغات ويجب أن يتم تقييم سلاسة الانتقال بين اللغتين خلال النطق وهذا مهم جدا في النماذج ثنائية اللغة ك Silma الذي اخترته .

تقييم استنتساخ الصوت :

أقترح أيضا أن يتم تقييم قدرة النموذج على استنتساخ الصوت بمدا تشابه صوت المتحدث معه والحفاظ على هويته وقد يفيد ذلك في انشاء بيانات بمستوا دقة مقبولة تساعد في تدريب التماذج ولكن يتطلب ذلك مراجعة دقيقة للغاية لهذه الملفات .

تقييم المشاعر وأنماط الكلام :

فالكلام الطبيعي يتغير حسب الحالة الشعورية (الحزن , السعادة , الحماسة , التعب , الخ..)

اختبار معدلات أخذ العينات المختلفة :

فقد يتم نشر النموذج في عدة بيئات مختلفة وبدرجات صوتة مختلفة (16khz. 22.05khz,25khz الخ..)

فيجب تقييمها جميعا باستخدام نفس المعايير MOS و RTF ويضاف لذلك أيضا موضوع استهلاك الذاكرة للتأكد من أن النموذج ما زال عمليا في كل الحالات .

الخلاصة :

استنادًا إلى مستوى التبني المجتمعي، وجودة المعمارية، والتركيز على اللغة العربية، والنشاط الحديث للمشروع، يُعتبر Silma TTS

المرشح الأقوى حاليًا لتقييم أنظمة تحويل النص إلى كلام باللغة العربية.

وتمثل هذه النماذج ثلاثة اتجاهات مختلفة في بناء أنظمة تحويل النص إلى كلام: نموذج ثنائي اللغة متخصص بالعربية، ونموذج متعدد اللغات واسع الانتشار، ونموذج متخصص باللهجات الخليجية. وتوفر منهجية التقييم المقترحة إطارًا علميًا متكاملًا لتحديد أفضل نموذج من حيث جودة النطق، والطبيعية، ودعم اللهجات، والكفاءة التشغيلية، والجاهزية للاستخدام في التطبيقات الواقعية.