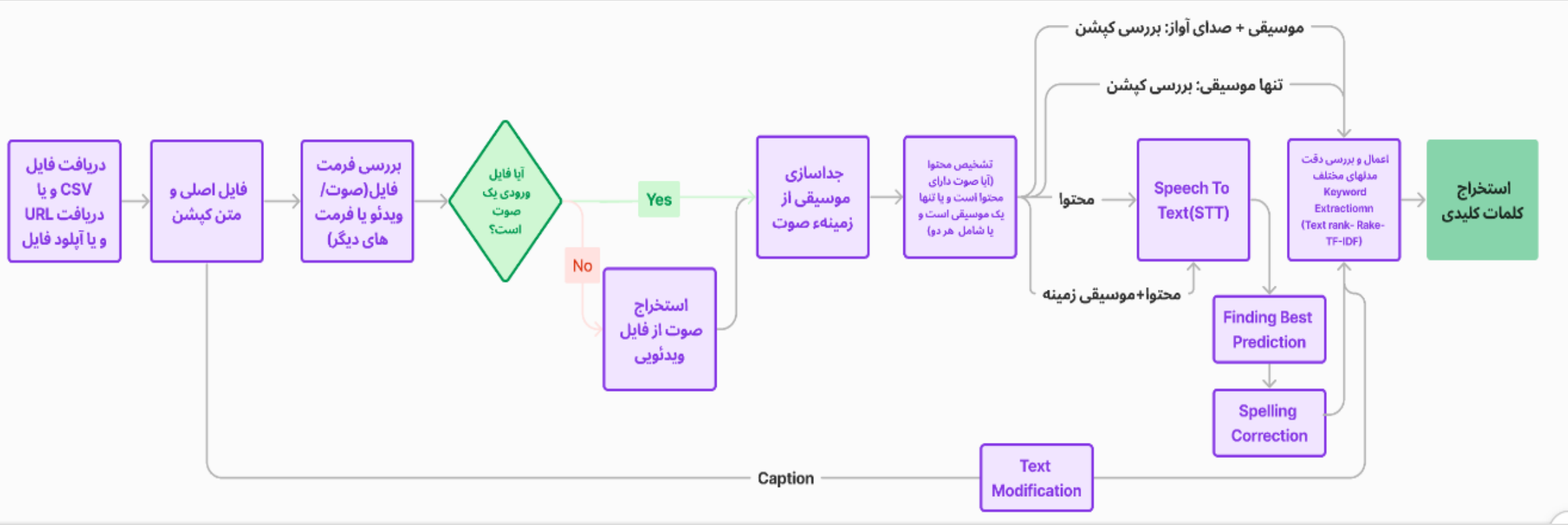
پروژه "استخراج کلمات کلیدی از منابع صوتی و ویدئویی"

ارائهء دوم

فاطمه وحیدیونسی

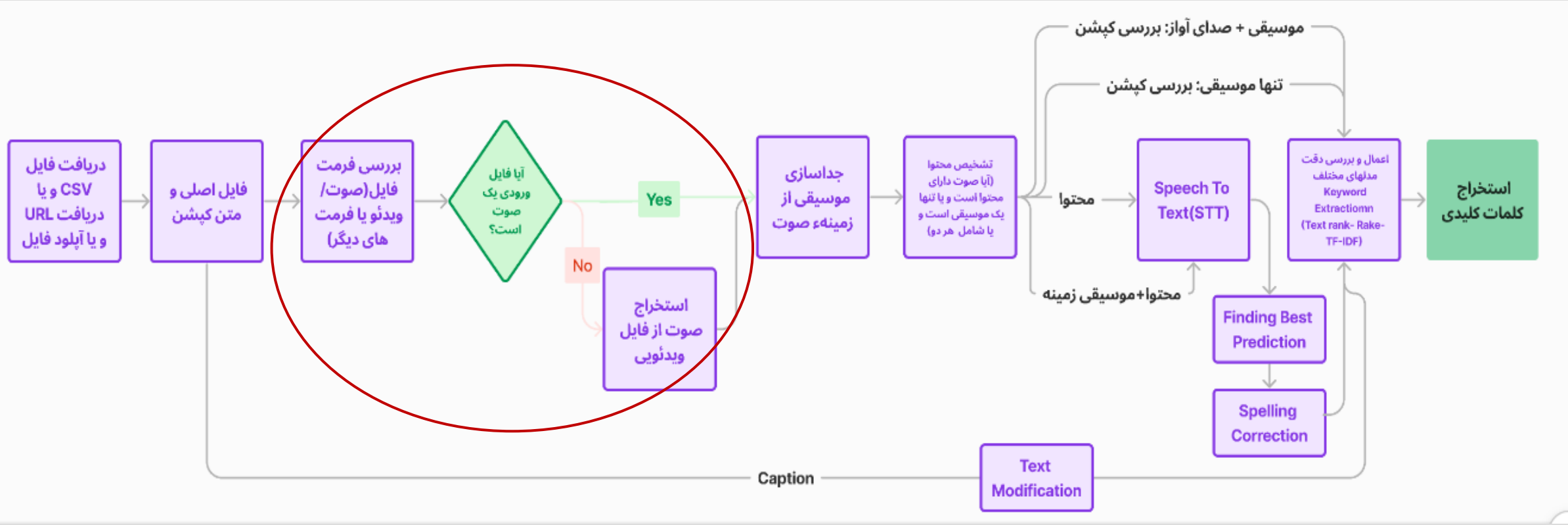
مرکز نوآوری علوم و فناوری های شناختی دانشگاه علم و صنعت

❑ فرآیند پروژه



نمودار۱- فلوچارت نهایی مراحل پروژه

❑ فرآیند پروژه



نمودار۱- فلوچارت نهایی مراحل پروژه

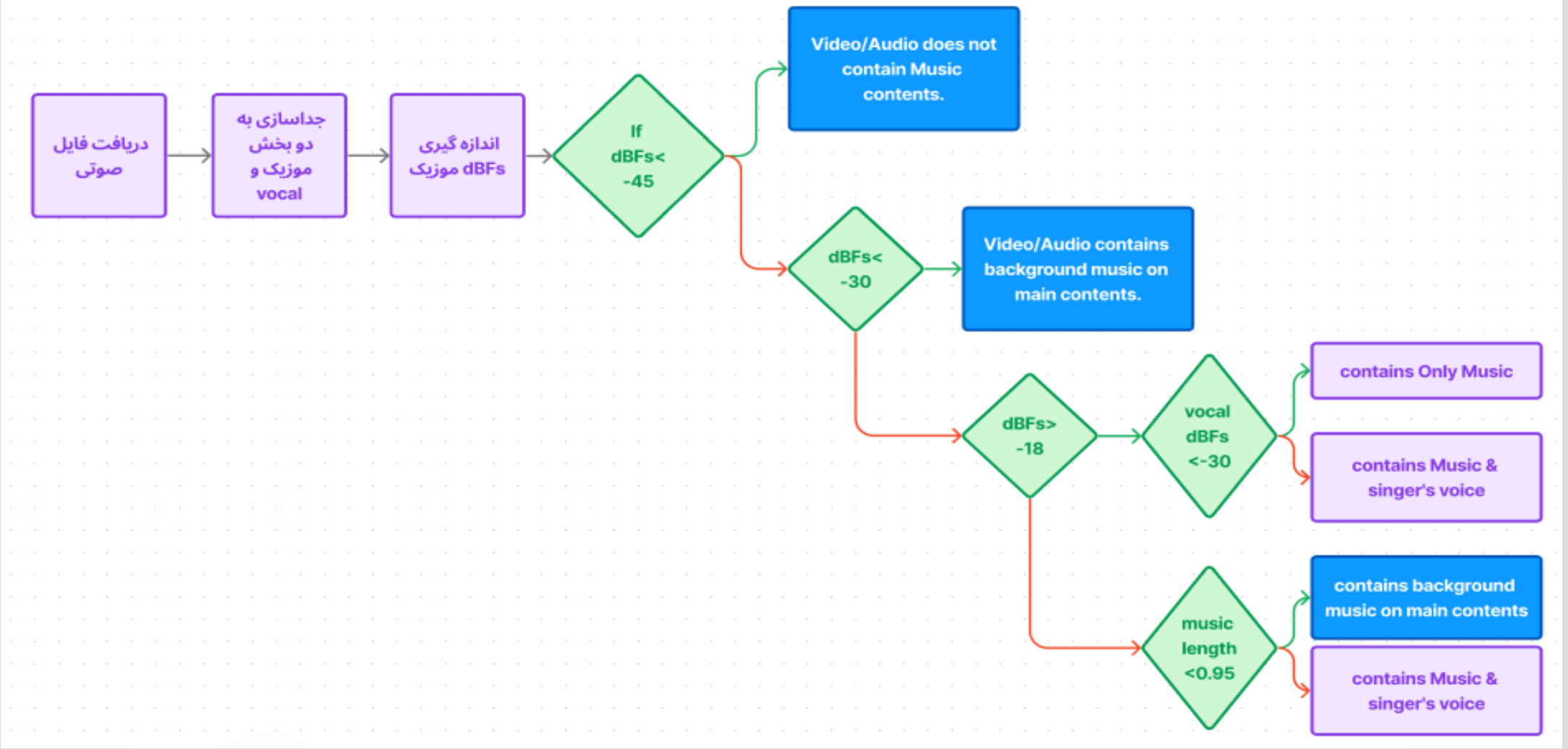
❑ بررسی جزییات هر مرحله از پروژه

بر اساس اجرا بر روی Tesla T4 GPU			
مرحله	متشکل از	حدود دقت	زمان تقریبی اجرا
1-تشخیص فرمت فایل ورودی	تابع is_video	100٪	کسری از ثانیه
2- تبدیل فایل ویدئویی به صوتی	تابع vid2audio	100٪	4٪ از کل زمان اجرا همزمان وابسته به حجم و مدت زمان ویدئو. بنابراین با کاهش حجم فایل و فشرده سازی زمان، می توان از زمان اجرای این مرحله کاست. (درصورتیکه تکنیک اجرا شده، زمان بر نباشد). در کل زمان اجرای این مرحله بیشتر وابسته به "طول زمانی" فایل است.

❑ بررسی جزییات هر مرحله از پروژه

بر اساس اجرا بر روی Tesla T4 GPU			
3- جداسازی موسیقی زمینه از صوت	استفاده از مدل HTdemucs(v4) جهت جداسازی محتوای صوت به 2 بخش Vocal و Music (بخش Music خود شامل 3 فایل مربوط به instrument های مختلف موسیقیایی است).	این مرحله دارای 2 معیار کیفی است: Mean opinion score (MOS)=4 Overall SDR[1]* =9.0	حدود 22% طول زمانی صوت 17٪ از کل زمان اجرا برای "محتواهای دارای موسیقی در متن اصلی"
4- تشخیص نوع صوت از نظر وجود محتوای مفید برای مرحله STT(صوت به متن)	- استفاده از dBFS جهت محاسبه تراز شدت صوتِ موزیک زمینه و vocal - استفاده از تابع music_length جهت یافتن طول زمان موسیقی در صوت - استفاده از تابع music_mode جهت تشخیص محتوای صوت(vocal- vocal و موزیک زمینه- موسیقی)	90٪ در فایل هایی که همزمان حاوی چندین نوع محتوا در کنار هم هستند خطا رخ می دهد. به عنوان مثال، در فایل هایی که بخشی از آن حاوی سخنرانی و بخش دیگر حاوی موسیقی باشد. راه حل: 1- برای مدیاهایی که دارای محتوای مشخص هستند(به عنوان مثال، صوت های سخنرانی در یک کانال تلگرامی) این مرحله قابل حذف است. 2- برای دیگر مدیاها(اینستاگرام) که معمولا دارای موسیقی هستند در حال بررسی تکنیک های دیگر اکوستیکی جهت افزایش دقت شناسایی محتوا و کاهش زمان اجرا هستیم.	4٪ از کل زمان اجرا برابر با 2.5٪ طول فایل برای محتواهای "فاقد موسیقی" در متن اصلی 39٪ از کل زمان اجرا برابر با 41٪ طول فایل برای محتواهای "دارای موسیقی" در متن اصلی

❑ الگوریتم تشخیص محتوای موسیقیایی:



❑ بررسی جزییات هر مرحله از پروژه

بر اساس اجرا بر روی Tesla T4 GPU			
STT+ Merging -5 Large & Medium Models	- Predict با استفاده از مدل های Medium و Large Whisper - استفاده از تابع Integrated_stt جهت ادغام نتایج دو مدل و یافتن بهترین پیش بینی حجم مدل Medium : 1.4Gb حجم مدل Large : 2.9 Gb	دقت برای مدل Large Whisper: به طور میانگین 36٪ "خطای" WER[2] و 12٪ "خطای" CER (88٪ دقت تشخیص حروف و 64٪ دقت تشخیص کلمات) **بررسی دقت در این بخش نیازمند تست بر روی داده های بیشتر بر روی سخت افزاری با سرعت بالاتر است. گزارش زیر، براساس بررسی های انجام شده "تاکنون" است: همانگونه که در جدول فوق نشان داده شده است، افزایش دقت در حالت ادغام نتایج دو مدل Medium , Large Whisper اندک بوده اما درصد زمان اجرای مرحله STT را تا 30٪ افزایش می دهد. در نتیجه، در صورت عدم محدودیت زمانی می توان از تابع Integrated_stt و ادغام دو مدل استفاده نمود. در غیر این صورت، درصورت وجود محدودیت زمانی در اجرا، مدل Large به تنهایی دقت مناسبی را ارائه می دهد.	41٪ از طول زمانی فایل، برابر با 56٪ از کل زمان اجرا برای محتواهای "فاقد موسیقی" در متن اصلی و 40٪ از کل زمان اجرا برای محتواهای "دارای موسیقی" در متن اصلی

	Large	Medium	Merging
Without spell checking	Wer 36٪ Cer 12.8٪	Wer 50٪ Cer 16٪	Wer 35% Cer 12%
With spell checking	Wer 29٪ Cer 12.3٪	Wer 39٪ Cer 13.9٪	Wer 26.9٪ Cer 11.7 ٪

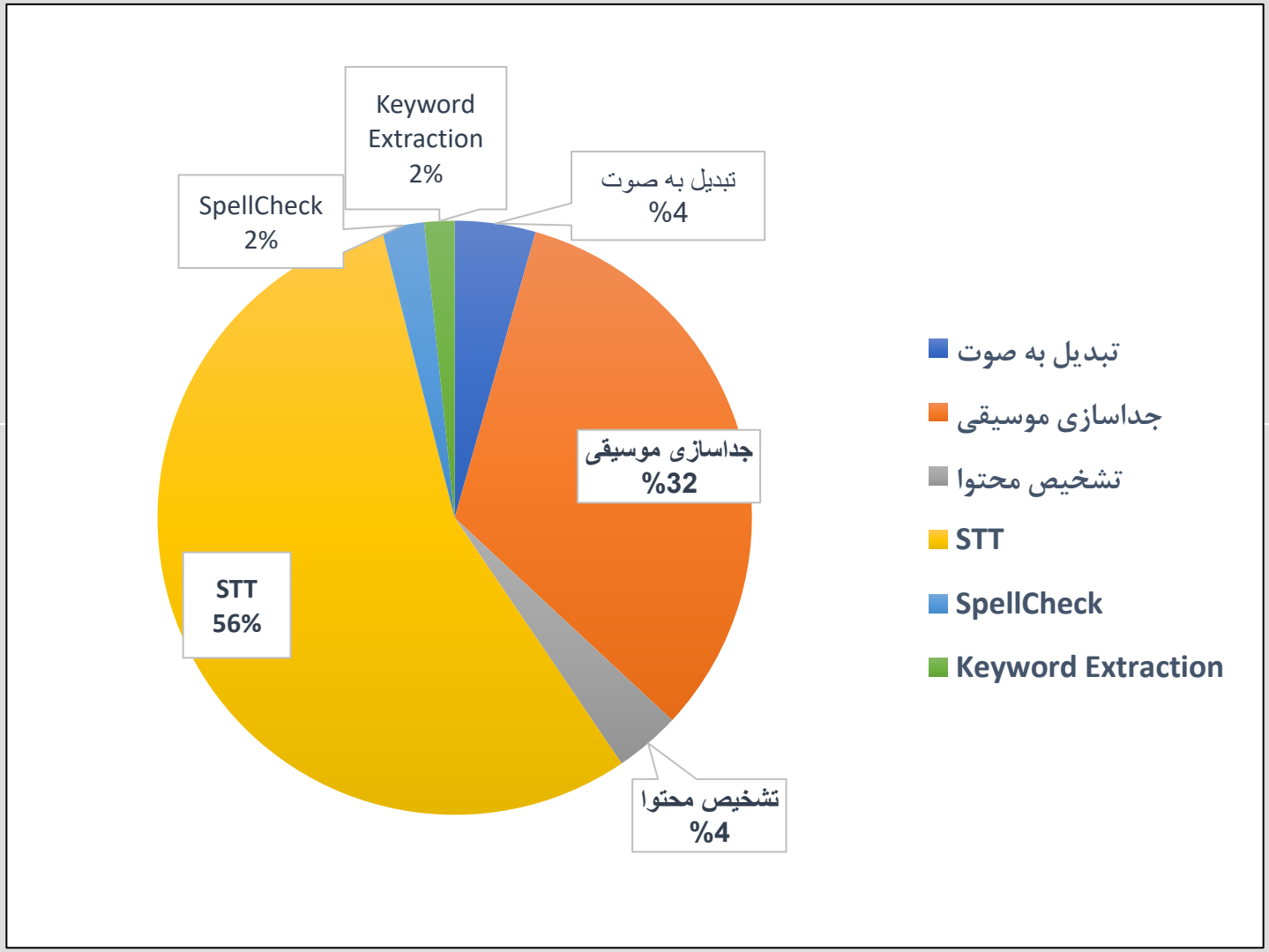
❑ بررسی جزییات هر مرحله از پروژه

بر اساس اجرا بر روی Tesla T4 GPU			
Spelling -6 Correction	اصلاح خطای املایی با استفاده از پکیج Parsivar حجم مدل های اصلاح خطاهای املایی Onegram و My_bigram به ترتیب 10 و157مگابایت	به طور میانگین دقت تشخیص کلمات را 7٪ افزایش می دهد. (خطای WER را 7٪ کاهش می دهد).	1.7٪ از طول زمانی فایل برابر با 2٪ از کل زمان اجرا برای محتواهای "فاقد موسیقی" در متن اصلی و 1٪ از کل زمان اجرا برای محتواهای "دارای موسیقی" در متن اصلی
(Caption)Text -7 Modification	برای محتوای دارای کپشن در کنار ویدیو: - حذف ایموجی از متن کپشن - یافتن هشتگ ها، ذخیره سازی و حذف underline - حذف آدرس های آیدی از کپشن	100٪	کسری از ثانیه

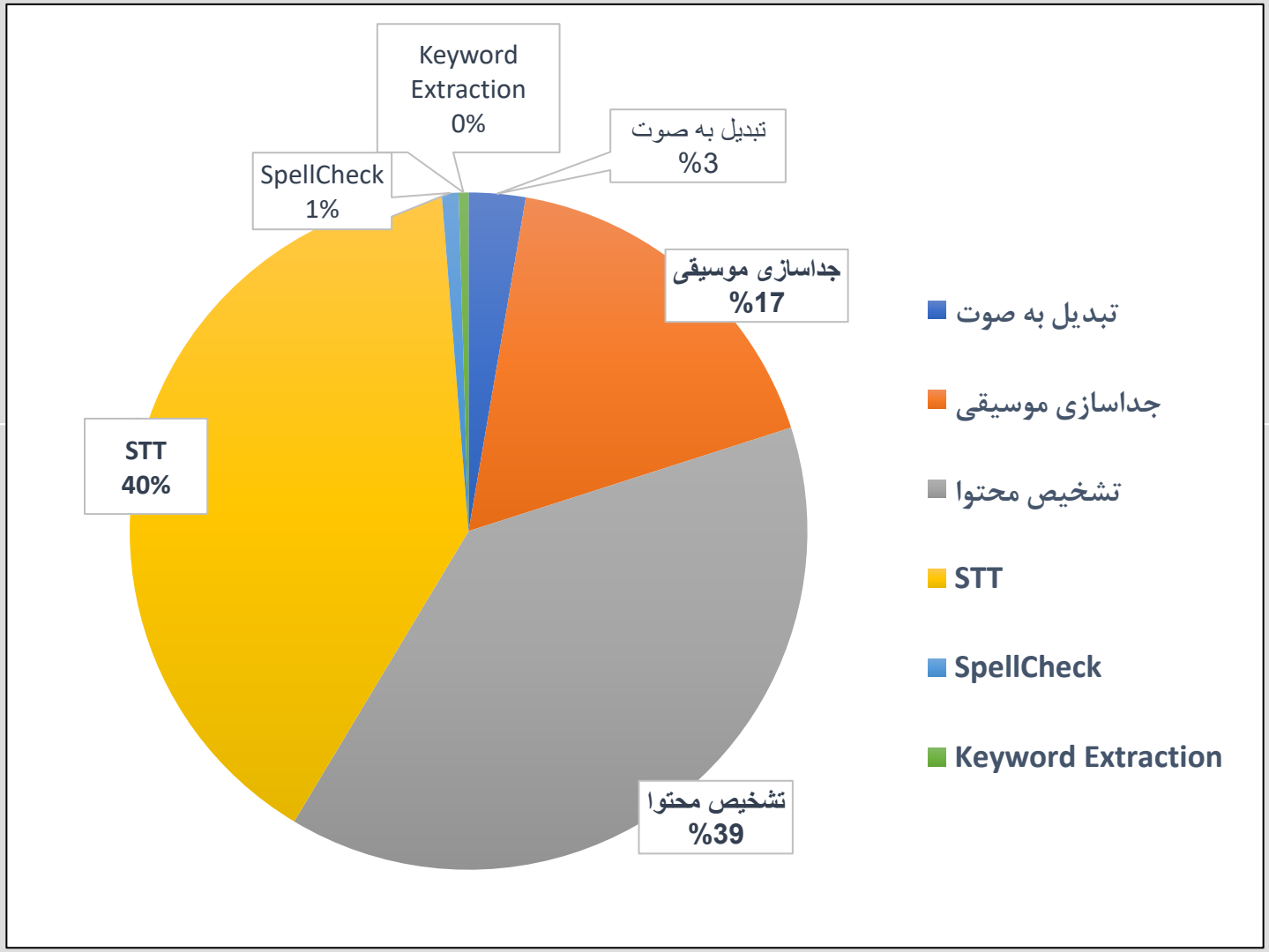
❑ بررسی جزییات هر مرحله از پروژه

بر اساس اجرا بر روی Tesla T4 GPU			
Keyword -8 Extraction	- ادغام متن حاصل از STT و کپشن(در صورت وجود) - استفاده از مدل TopicRank با استفاده از پکیج Perke جهت استخراج 20 کلمه کلیدی اصلی - رتبه بندی براساس کلمات موجود در هشتگ ها استفاده از مدل pos_tagger با حجم 19 مگابایت	ارزیابی دقت بر روی دیتاست Thesis Abstract شامل 450 مقاله از پایگاه Irandoc در زمینه علوم انسانی انجام گرفته است. دقت شود که با توجه به هدف پروژه، مرحله STT در اولویت بالاتر قرار داشته و تاکنون کمتر بر روی مرحله هشتم اقدام توسعه ای انجام گرفته، درصورت نیاز می توان دقت این مرحله را بالاتر برد: Partial F1 score at 10 extracted keywords (pF1@10)= 12.7% Partial Precision(pP@10)=10% Partial Recall(pR@10)=17.35%	1.2٪ از زمان فایل برابر با 1.6٪ از کل زمان اجرا برای محتواهای "فاقد موسیقی" در متن اصلی 0.5٪ از کل زمان اجرا برای محتواهای "دارای موسیقی" در متن اصلی

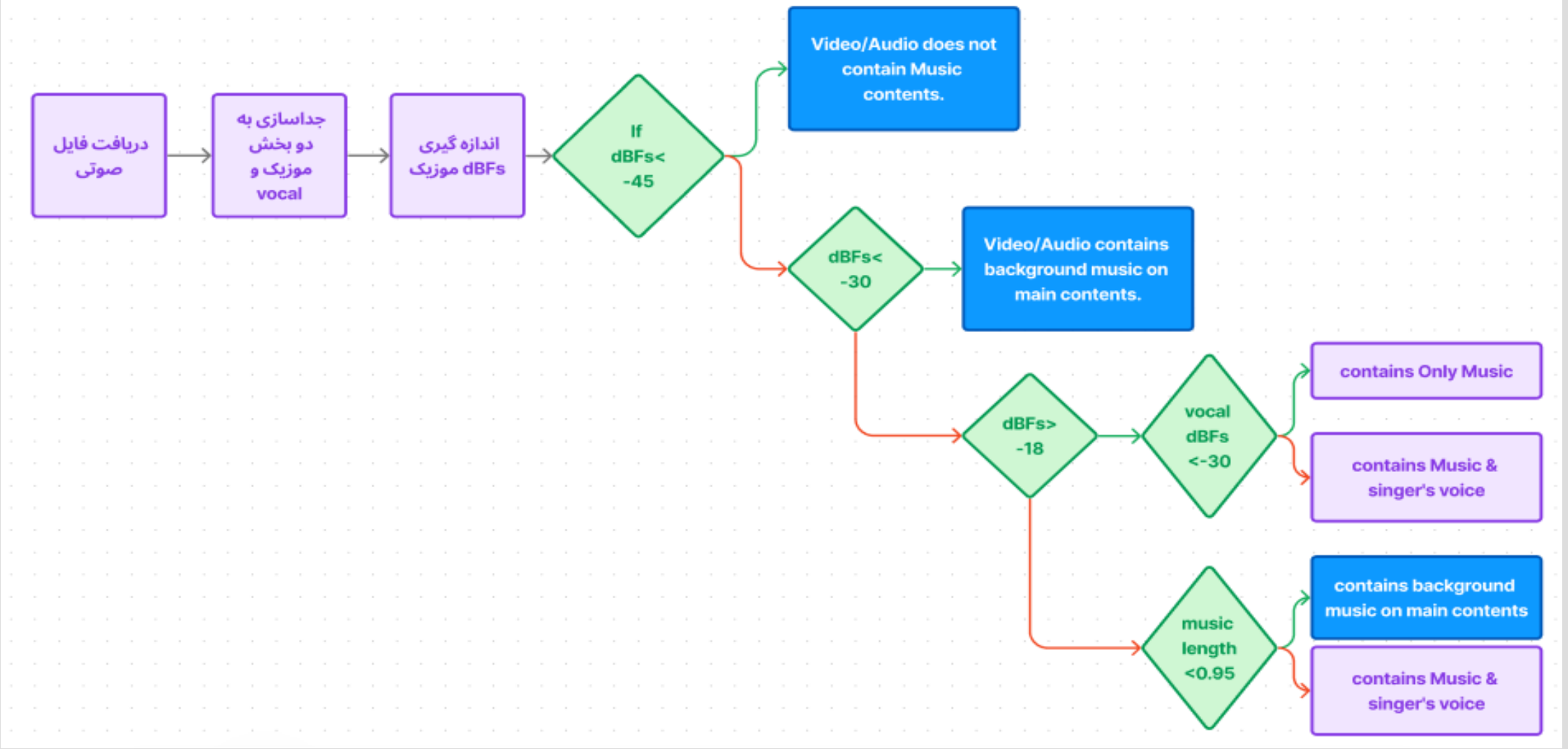
سهم هر مرحله از زمان کل اجرا (برای محتواهای فاقد موسیقی در متن اصلی)



سهم هر مرحله از زمان کل اجرا (برای محتواهای دارای موسیقی در متن اصلی)



❑ الگوریتم تشخیص محتوای موسیقیایی:



❑ پیشبرد و توسعه آتی در پروژه:

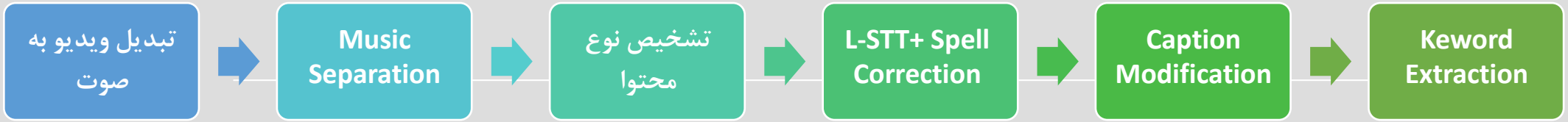
جهت افزایش دقت، بررسی تغییر تنظیمات مدل "کاهش نویز" بر روی دقت خروجی مدل STT جزو پیشبردهای آتی قرار دارد.

جهت افزایش سرعت اجرا، روش های زیر توصیه می شود و در دست بررسی است:

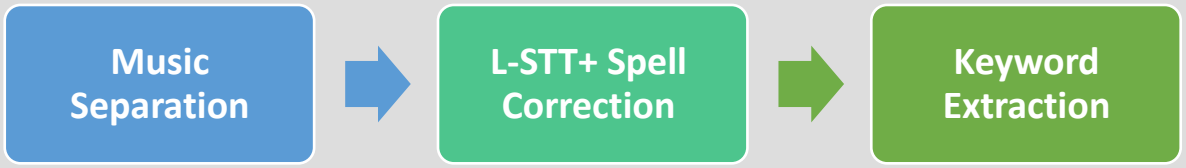
- همانگونه که گفته شد زمان بندی های فوق براساس اجرا بر روی پلن رایگان GPU گوگل کولب است. برای افزایش سرعت اجرا، می بایستی سخت افزار(GPU) قوی تری را به کار برد. (متعاقب آن، امکان تست بر روی مجموعه دادگان بیشتر را نیز ممکن می سازد).
- بررسی تاثیر کاهش حجم فایل و یا فشرده سازی محتوا(افزایش سرعت صوت) بر روی زمان اجرای هر یک از مراحل. درمورد مرحله تبدیل صوت به متن(STT) باید بررسی شود که این عملیات باعث کاهش کیفیت و دقت این مرحله نگردد.
- بررسی تکنیک های اکوستیکی جهت افزایش سرعت مرحلهء "تشخیص محتوا"

❑ پیشبرد و توسعه آتی در پروژه:

- حذف فرآیندهای کم تاثیر روی دقت(مانند ادغام مدل های STT). همچنین درمورد مدیاهایی با محتوای مشخص (مانند یک کانال سخنرانی در تلگرام) مرحله جداسازی و تشخیص محتوای موسیقیایی لازم نبوده و قابل حذف است. به مثال زیر توجه کنید، با توجه به تفاوت محتوا در شبکه های اجتماعی و مدیاهای مختلف، معماری زیر پیشنهاد می شود:
  - حالت اول(بررسی محتوای اینستاگرام/آپارات: محتوای دارای caption):



  - حالت دوم(بررسی محتوای یک کانال سخنرانی در تلگرام، محتوای صوتی و فاقد caption):



همانگونه که در نمودار بالا نمایش داده شده است، بسته به نوع محتوا و سورس آن، می توان برخی مراحل اجرا را حذف نمود و سرعت اجرا را بالاتر برد.

„Artificial intelligence helps you live better!