دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

گزارش فاز دوم پروژه

نام و نام خانوادگی اعضای تیم

علیرضا فرشی ۸۱۰۱۰۴۲۲۳
ساراسادات نصر ۸۱۰۱۰۴۲۷۲
سعید رضائی نور ۸۱۰۱۰۴۱۴۵

۲۶ بهمن ۱۴۰۴

فهرست مطالب

۱ چکیده ۱

۲ مقدمه ۲

۱.۲ بیان مسئله و اهمیت موضوع ۲

۲.۲ مرور کلی بر پروژه ۲

۳.۲ پیش‌پردازش و استخراج ویژگی ۲

۴.۲ مدل‌های یادگیری ماشین ۳

۵.۲ ساختار گزارش ۳

۶.۲ اهداف و دستاوردهای مورد انتظار ۴

۳ پاک سازی داده و استخراج ویژگی ۴

۱.۳ پاک سازی داده ها ۵

۱.۱.۳ بارگذاری و یکسان‌سازی نرخ نمونه‌برداری ۵

۲.۱.۳ حذف سکوت ۵

۳.۱.۳ نرمال‌سازی دامنه ۵

۴.۱.۳ افزایش داده (Data Augmentation) ۶

۲.۳ ویژگی‌های استخراج‌شده و علت انتخاب آن‌ها ۷

۱.۲.۳ ویژگی‌های حوزه زمان ۷

۲.۲.۳ ویژگی‌های حوزه فرکانس ۷

۳.۲.۳ ویژگی‌های مبتنی بر طیف‌نگار مل (Mel-spectrogram) ۸

۴.۲.۳ ضرایب کپسترال مل (MFCC) ۸

۵.۲.۳ ویژگی‌های کروماتیک (Chroma Features) ۸

۶.۲.۳ ویژگی‌های تُنال (Tonnetz) ۹

۷.۲.۳ ویژگی‌های کنتراست طیفی (Spectral Contrast) ۹

۸.۲.۳ ویژگی‌های چندجمله‌ای (Poly Features) ۹
۹.۲.۳ ویژگی‌های ریتمیک (Tempogram) ۹
۱۰.۲.۳ جمع بندی و مشاهده ویژگی ها ۱۰

۴ طبقه بندی ۱۲
۱.۴ مدل های مورد استفاده در طبقه بندی ۱۳
۱.۱.۴ مدل‌های مبتنی بر فاصله و مشابهت ۱۳
۲.۱.۴ مدل‌های مبتنی بر مرز تصمیم‌گیری ۱۳
۳.۱.۴ مدل‌های مبتنی بر درخت و ensemble ۱۳
۴.۱.۴ مدل‌های شبکه عصبی ۱۴
۵.۱.۴ مدل‌های خطی و احتمالاتی ۱۴
۲.۴ معیارهای ارزیابی و تحلیل عملکرد مدل‌ها ۱۴
۱.۲.۴ تحلیل معیارهای ارزیابی ۱۵
۲.۲.۴ تحلیل ماتریس درهم‌ریختگی ۱۵

۵ خوشه بندی ۱۶
۱.۵ تعیین تعداد بهینه خوشه ها ۱۶
۲.۵ مدل های مورد استفاده در خوشه بندی ۱۷
۳.۵ معیار ارزیابی و کارایی مدل های مختلف ۱۸
۴.۵ تحلیل نتایج خوشه‌بندی ۱۹
۵.۵ نمایش بصری الگوریتم های خوشه بندی ۲۰

۶ ارزیابی ۲۴
۱.۶ ارزیابی مدل‌های طبقه‌بندی (Classification) ۲۴
۲.۶ ارزیابی مدل‌های خوشه‌بندی (Clustering) ۲۵
۳.۶ نتیجه‌گیری نهایی ۲۶

۷ مراجع ۲۷

فهرست تصاویر

۱ توزیع دادگان بر اساس معیارهای مختلف ۴
۲ فایل های صوتی پاک سازی شده به ازای هر زبان ۶
۳ ماتریس همبستگی ویژگی ها ۱۰
۴ توزیع ویژگی های مختلف بر اساس زبان ها ۱۱
۵ نمایش سه بعدی داده های صوتی بر اساس کلید ویژگی های انتخاب شده . ۱۱
۶ نمایش نمودار PCA ویژگی ها ۱۲
۷ نمایش LDA برای زبان های مختلف ۱۲
۸ مقایسه کلی مدل ها در معیارهای مختلف ۱۵
۹ ماتریس درهم‌ریختگی سه مدل برتر ۱۶
۱۰ نمودار معیارهای تعیین تعداد خوشه های بهینه ۱۷
۱۱ دسته بندی های انجام شده در الگوریتم های مختلف ۲۱
۱۲ توزیع داده ها در دسته های مختلف ۲۲
۱۳ مقایسه cluster purity بین الگوریتم های مختلف ۲۲
۱۴ نمایش داده ها با t-SNE به همراه برچسب های واقعی ۲۳

فهرست جداول

۱ معیارهای ارزیابی مدل‌های طبقه‌بندی ۱۴
۲ معیارهای ارزیابی مدل های مختلف خوشه بندی ۱۹
۳ معیارهای دقیق ارزیابی مدل های طبقه بندی ۲۴
۴ معیارهای دقیق ارزیابی مدل های خوشه بندی ۲۵

چکیده

هدف از این پروژه، طراحی و پیاده‌سازی یک سیستم تشخیص خودکار زبان‌های گفتاری (اسپانیایی، کره‌ای، ایتالیایی و آلمانی) بر پایه یادگیری ماشین است. پس از جمع‌آوری و مشاهده کلی دادگان و پیش‌پردازش داده‌های صوتی از جمله حذف نویز یا نرمال‌سازی صوت، ویژگی‌های زیادی از جمله ویژگی‌های مبتنی بر حوزه زمان و Mel-scaled spectrogram استخراج گردیده است. داده‌ها با روش نمونه‌برداری طبقه‌ای به مجموعه‌های آموزش (٪۸۰) و تست (٪۲۰) تقسیم شدند. در بخش طبقه‌بندی، مدل های k نزدیک‌ترین همسایه، SVM با کرنل rbf، random forest، MLP، Decision Tree، Logistic Regression و Naive Bayes با استفاده از کتابخانه Scikit-Learn پیاده سازی و ارزیابی شدند. نتایج ارزیابی نشان داد که k نزدیک ترین همسایه با دقت 98.7 درصد با مقدار k برابر با ۵، بالاترین کارایی را دارد. در بخش خوشه‌بندی بدون ناظر، الگوریتم DBSCAN با تعیین ۱۴ خوشه بهینه (مطابق تعداد زبان‌ها) و امتیاز خلوص ۱ بهترین انطباق با ساختار واقعی داده را نشان داد. همچنین کاهش ابعاد با t-SNE جداسازی نسبتاً متمایز خوشه‌ها را در فضای دو بعدی آشکار ساخت. این پژوهش گامی در جهت کاربرد پردازش سیگنال گفتار و یادگیری ماشین در شناسایی چندزبانه است.

مقدمه

۱.۲ بیان مسئله و اهمیت موضوع

در دنیای امروز، با توجه به گسترش ارتباطات بین‌المللی و افزایش محتوای چندزبانه در فضای دیجیتال، تشخیص خودکار زبان گفتاری به یکی از چالش‌های مهم در حوزه پردازش سیگنال و یادگیری ماشین تبدیل شده است. کاربردهای این فناوری از سیستم‌های ترجمه همزمان و دستیارهای صوتی هوشمند گرفته تا برچسب‌گذاری خودکار محتوای صوتی و پالایش اطلاعات در پلتفرم‌های چندزبانه گسترده است. با توجه به تنوع زبانی موجود در داده‌های صوتی کتابخانه‌های دیجیتال و پادکست‌ها، توسعه سیستمی که بتواند با دقت بالا زبان گفتاری را شناسایی کند، گامی اساسی در جهت سازماندهی و بازیابی مؤثر اطلاعات به شمار می‌رود.

۲.۲ مرور کلی بر پروژه

در این پروژه، مجموعه‌ای از داده‌های صوتی شامل قطعات کوتاه (حدود یک دقیقه) از کتاب‌های صوتی به چهار زبان اسپانیایی، کره‌ای، ایتالیایی و آلمانی مورد استفاده قرار گرفته است. هدف اصلی، طراحی و پیاده‌سازی یک سیستم کامل مبتنی بر یادگیری ماشین است که قادر به تشخیص زبان گفتاری در این قطعات صوتی باشد. برای این منظور، پروژه در چند گام اصلی شامل پاک‌سازی داده‌ها، استخراج ویژگی، طبقه‌بندی ،خوشه‌بندی و ارزیابی طراحی شده است.

۳.۲ پیش‌پردازش و استخراج ویژگی

صداهای خام به دلیل وجود نویز و تفاوت در بلندی صدا، مستقیماً برای مدل‌های یادگیری ماشین قابل استفاده نیستند. از این رو، در گام نخست، عملیات پیش‌پردازش شامل نویززدایی و نرمال‌سازی بر روی داده‌های صوتی اعمال می‌گردد تا کیفیت ورودی بهبود یابد. در ادامه، با استفاده از روش‌های استخراج ویژگی، سیگنال‌های صوتی به بردارهای عددی معنادار تبدیل می‌شوند. ویژگی‌های مرسومی همچون ضرایب کپسترال مل (MFCC)، طیف‌نگار، انرژی و zcr از جمله شاخص‌هایی هستند که می‌توانند تفاوت بین زبان‌های مختلف را به خوبی نمایش

دهند.

۴.۲ مدل‌های یادگیری ماشین

پس از آماده‌سازی داده‌ها، نوبت به بخش اصلی پروژه یعنی مدل‌سازی می‌رسد. در این پروژه، سه رویکرد اصلی دنبال می‌شود:

- **طبقه‌بندی نظارت‌شده:** در این بخش، داده‌ها با برچسب زبان به مدل‌های مختلف یادگیری ماشین داده می‌شوند تا الگوهای متمایزکننده زبان‌ها را بیاموزند. الگوریتم‌هایی مانند ماشین بردار پشتیبان (SVM)، جنگل تصادفی (Random Forest) و شبکه عصبی پرسپترون چندلایه (MLP) برای این منظور پیاده‌سازی و ارزیابی می‌شوند.
- **خوشه‌بندی بدون نظارت:** برای تحلیل ساختار ذاتی داده‌ها بدون استفاده از برچسب‌ها، الگوریتم‌های خوشه‌بندی مانند K-Means و DBSCAN به کار گرفته می‌شوند تا مشخص شود آیا داده‌های صوتی به طور طبیعی بر اساس زبان گروه‌بندی می‌شوند یا خیر.
- **ارزیابی و تحلیل:** در نهایت، عملکرد مدل‌ها با معیارهایی همچون دقت (Accuracy)، فراخوانی (Recall)، صحت (Precision) و امتیاز F۱ سنجیده شده و با رسم ماتریس درهم‌ریختگی، زبان‌هایی که بیشترین خطا در تشخیص آن‌ها رخ داده، شناسایی و تحلیل می‌شوند.

۵.۲ ساختار گزارش

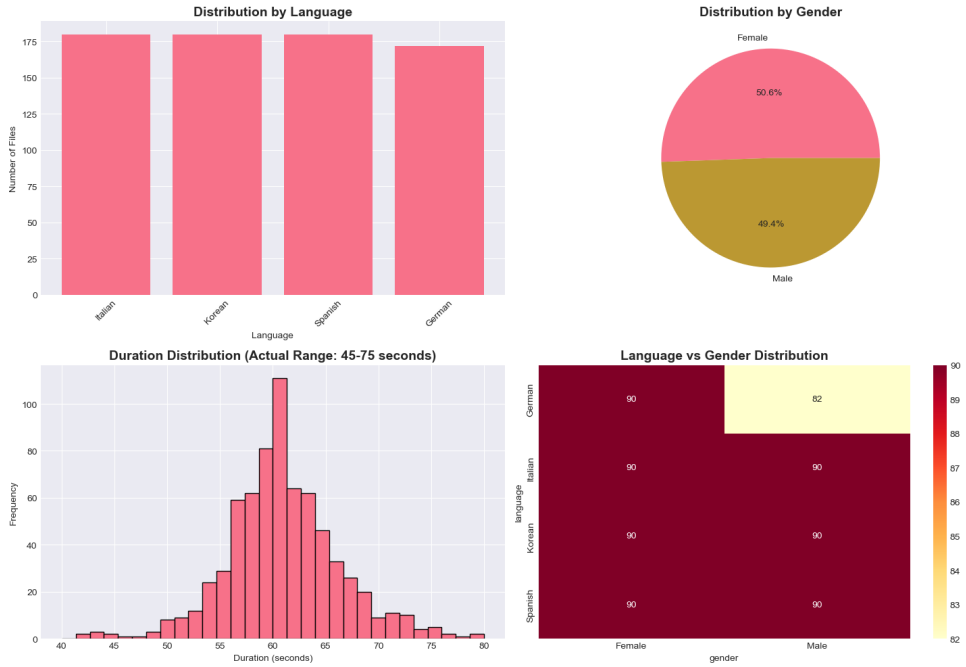
گزارش پیش رو مطابق با مراحل پروژه سازماندهی شده است. در بخش دوم، فرآیند پاک‌سازی داده‌ها و استخراج ویژگی تشریح می‌گردد. بخش سوم به طبقه‌بندی و مدل‌های نظارت‌شده اختصاص دارد. در بخش چهارم، خوشه‌بندی و تحلیل بدون نظارت ارائه می‌شود. بخش پنجم شامل ارزیابی جامع مدل‌ها و مقایسه نتایج است و در نهایت، بخش ششم به جمع‌بندی و نتیجه‌گیری اختصاص یافته است. همچنین کدهای پیاده‌سازی شده در قالب چندین notebook به صورت جداگانه ارائه گردیده است.

۶.۲ اهداف و دستاوردهای مورد انتظار

انتظار می‌رود که با اجرای این پروژه، ضمن آشنایی عملی با چالش‌های کار با داده‌های صوتی واقعی، بتوان سیستمی طراحی کرد که با دقت مناسبی زبان گفتاری را تشخیص دهد. همچنین تحلیل خوشه‌بندی می‌تواند نشان دهد که آیا ویژگی‌های استخراج‌شده به اندازه کافی قدرت تفکیک زبان‌ها را دارند یا نیاز به بازنگری در انتخاب ویژگی‌ها وجود دارد. مقایسه مدل‌های مختلف نیز دیدگاه روشنی نسبت به نقاط قوت و ضعف هر الگوریتم در مسئله تشخیص زبان ارائه خواهد داد.

پاک سازی داده و استخراج ویژگی

در این مرحله قبل از اینکه عملیات پاکسازی و استخراج ویژگی ها را انجام دهیم، دادگان را مشاهده می کنیم و سعی می کنیم اطلاعاتی کلی از آن استخراج کنیم. در شکل های زیر توزیع دادگان بر اساس زبان های مختلف، جنسیت و مدت زمان نمایش داده شده است.



شکل ۱: توزیع دادگان بر اساس معیارهای مختلف

نمودار بالا سمت چپ توزیع دادگان بر اساس زبان های مختلف، نمودار بالا راست توزیع

دادگان بر اساس جنسیت فایل های صوتی، نمودار پایین چپ نمودار توزیع مدت زمان فایل های صوتی که از یک توزیع نسبتا گاوسی می آید و نمودار پایین راست توزیع جنسیت در هریک از زبان ها را مشاهده می کنید.

۱.۳ پاک سازی داده ها

۱.۱.۳ بارگذاری و یکسان‌سازی نرخ نمونه‌برداری

ابتدا تمام فایل‌های صوتی با استفاده از کتابخانه Librosa بارگذاری شدند. نرخ نمونه‌برداری همه فایل‌ها به ۲۲۰۵۰ هرتز یکسان‌سازی گردید. این مقدار به عنوان نرخ استاندارد در بسیاری از پروژه‌های پردازش گفتار انتخاب می‌شود، زیرا تعادل مناسبی بین کیفیت سیگنال و حجم داده برقرار می‌کند. همچنین برای کاهش زمان پردازش و یکنواخت‌سازی طول نمونه‌ها، هر فایل صوتی تا ۵۰ ثانیه برش داده شد. این بازه زمانی برای استخراج ویژگی‌های معنادار کافی بوده و از ورود داده‌های اضافی و افزونه به مدل جلوگیری می‌کند.

۲.۱.۳ حذف سکوت

یکی از چالش‌های رایج در داده‌های صوتی، وجود بخش‌های سکوت یا نویزهای کم‌دامنه در ابتدا و انتهای فایل‌هاست. این بخش‌ها نه تنها اطلاعات مفیدی به مدل اضافه نمی‌کنند، بلکه می‌توانند باعث کاهش کارایی و افزایش خطا شوند. برای رفع این مشکل، از تابع librosa.effects.trim با آستانه ۲۰ دسی‌بل استفاده شد. این تابع با شناسایی بخش‌هایی که انرژی آن‌ها کمتر از آستانه تعیین‌شده است، آن‌ها را حذف کرده و تنها بخش‌های حاوی گفتار را نگه می‌دارد. انتخاب آستانه ۲۰ دسی‌بل بر اساس آزمایش‌های اولیه و مشاهده توزیع انرژی سیگنال‌ها صورت گرفته است.

۳.۱.۳ نرمال‌سازی دامنه

سیگنال‌های صوتی مختلف ممکن است دامنه‌های متفاوتی داشته باشند که این تفاوت می‌تواند ناشی از شرایط ضبط، فاصله میکروفون یا تنظیمات سخت‌افزاری باشد. برای رفع این ناهماهنگی، از نرمال‌سازی با استفاده از تابع librosa.util.normalize استفاده شد. این تابع دامنه سیگنال را به بازه [-1, 1] مقیاس‌دهی می‌کند و تضمین می‌کند که تمام نمونه‌ها از نظر بلندی صدا در یک سطح قرار گیرند. این امر باعث می‌شود مدل به جای تفاوت‌های دامنه‌ای، بر روی ویژگی‌های زبانی متمرکز شود.

۴.۱.۳ افزایش داده (Data Augmentation)

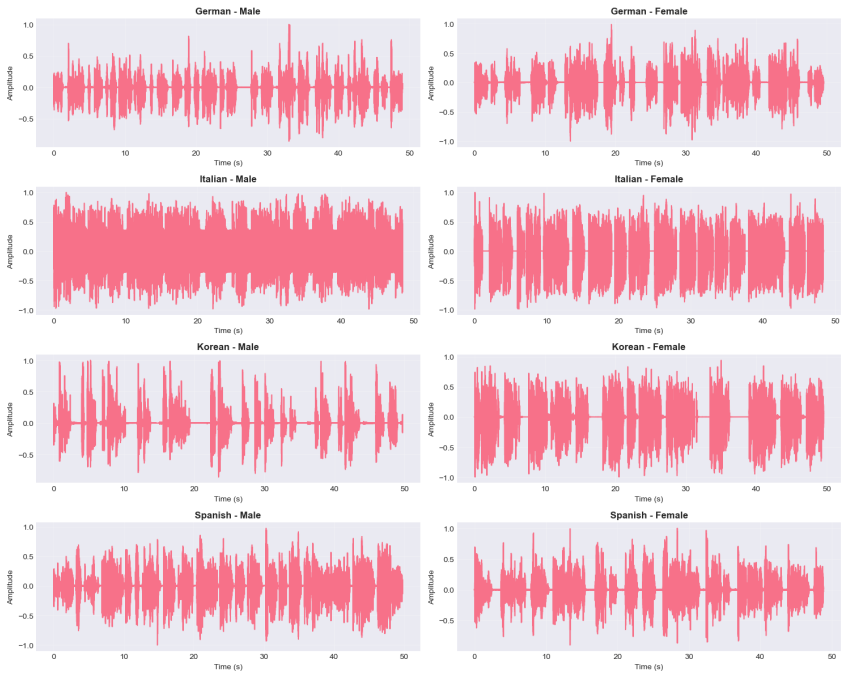
برای بهبود تعمیم‌پذیری مدل و افزایش تنوع داده‌های آموزشی، از تکنیک‌های افزایش داده استفاده شد. این کار به ویژه با توجه به محدودیت حجم داده‌های اولیه، می‌تواند عملکرد مدل را به طور چشمگیری بهبود بخشد. سه روش افزایش داده به کار گرفته شد:

- **افزودن نویز سفید:** نویز سفید با ضریب ۰/۰۰۵ به سیگنال اصلی اضافه شد. این کار باعث می‌شود مدل در برابر نویزهای محیطی مقاوم‌تر شده و توانایی تشخیص زبان در شرایط واقعی را بیاموزد. مقدار نویز انتخاب‌شده به اندازه‌ای است که تغییر محسوسی در سیگنال ایجاد کند اما گفتار را غیرقابل تشخیص نسازد.

- **کشش زمانی:** با ضریب ۰/۹، سیگنال صوتی در زمان فشرده شد. این تکنیک سرعت گفتار را تغییر می‌دهد بدون اینکه زیروبمی صدا تحت تأثیر قرار گیرد. این کار به مدل کمک می‌کند تا الگوهای زبانی را مستقل از سرعت گفتار تشخیص دهد.

- **تغییر زیروبمی صدا:** با جابجایی ۲ نیم‌پرده، زیروبمی صدا تغییر داده شد. این تکنیک تفاوت‌های ناشی از جنسیت، سن و ویژگی‌های فیزیکی گویندگان را شبیه‌سازی می‌کند و باعث می‌شود مدل به این تغییرات حساس نباشد.

در نهایت خروجی بدست آمده از پاک سازی داده های صوتی در شکل زیر قابل مشاهده می باشد.



شکل ۲: فایل های صوتی پاک سازی شده به ازای هر زبان

۲.۳ ویژگی‌های استخراج‌شده و علت انتخاب آن‌ها

در این پروژه، برای تبدیل سیگنال‌های صوتی خام به بردارهای عددی قابل استفاده در مدل‌های یادگیری ماشین، مجموعاً ۱۰۲ ویژگی از حوزه‌های مختلف زمانی، فرکانسی و طیفی استخراج گردید. این ویژگی‌ها با هدف پوشش کامل جنبه‌های مختلف سیگنال گفتار شامل زیروبمی، تُن، ریتم، انرژی و ساختار فرکانسی انتخاب شده‌اند.

۱.۲.۳ ویژگی‌های حوزه زمان

ویژگی‌های حوزه زمان مستقیماً از سیگنال صوتی در بازه زمانی استخراج می‌شوند و اطلاعاتی درباره دامنه و نوسانات سیگنال ارائه می‌دهند:

- **نرخ عبور از صفر (Zero Crossing Rate):** این ویژگی تعداد دفعاتی را که سیگنال از صفر عبور می‌کند در واحد زمان اندازه‌گیری می‌کند. ZCR برای تشخیص نویز از گفتار و همچنین تفکیک صداهای گرفته‌دار (مانند همخوان‌ها) از صداهای کشیده (مصوت‌ها) مفید است. میانگین و انحراف معیار این ویژگی محاسبه شده است.

- **ریشه میانگین مربعات انرژی (RMS Energy):** این ویژگی میزان انرژی سیگنال را نشان می‌دهد و معیاری برای بلندی صدا محسوب می‌شود. RMS می‌تواند به تفکیک بخش‌های گفتاری از سکوت کمک کرده و تغییرات انرژی در طول زمان را نمایش دهد. میانگین و انحراف معیار این ویژگی نیز استخراج گردیده است.

۲.۲.۳ ویژگی‌های حوزه فرکانس

این ویژگی‌ها با تبدیل سیگنال به حوزه فرکانس (با استفاده از تبدیل فوریه) به دست می‌آیند و اطلاعاتی درباره توزیع انرژی در فرکانس‌های مختلف ارائه می‌دهند:

- **مرکزواره طیفی (Spectral Centroid):** این ویژگی مرکز ثقل طیف فرکانسی را نشان می‌دهد و معیاری برای زیروبمی صدا است. صداهای زیرتر centroid بالاتری دارند و صداهای بم‌تر مرکزواره پایین‌تری. این ویژگی در تفکیک زبان‌ها بر اساس ساختار آوایی آن‌ها مؤثر است.

- **پهنای باند طیفی (Spectral Bandwidth):** پهنای باند، گستردگی طیف فرکانسی حول مرکزواره را اندازه‌گیری می‌کند. این ویژگی می‌تواند تفاوت بین صداهای خالص و نویزی را نشان دهد.

- **نقطه برگشت طیفی (Spectral Rolloff):** این ویژگی فرکانسی را نشان می‌دهد که درصد مشخصی از انرژی طیفی (معمولاً ۸۵٪) زیر آن قرار دارد. این ویژگی برای تشخیص نویز و تمایز بین صداهای گرفته‌دار و کشیده مفید است.

۳.۲.۳ ویژگی‌های مبتنی بر طیف‌نگار مل (Mel-spectrogram)

طیف‌نگار مل نسخه‌ای از طیف‌نگار است که بر اساس مقیاس شنوایی انسان (مقیاس مل) تنظیم شده است. گوش انسان در فرکانس‌های پایین تفکیک‌پذیری بیشتری دارد و مقیاس مل این ویژگی را شبیه‌سازی می‌کند:

- **طیف‌نگار مل:** پس از محاسبه طیف‌نگار مل، آن را به مقیاس دسی‌بل تبدیل کرده و میانگین و انحراف معیار آن استخراج شده است. این ویژگی نمای کلی از توزیع انرژی در باندهای فرکانسی مختلف ارائه می‌دهد.

۴.۲.۳ ضرایب کپسترال مل (MFCC)

MFCC ها یکی از مهم‌ترین و پرکاربردترین ویژگی‌ها در پردازش گفتار و تشخیص زبان هستند. این ویژگی‌ها با اعمال تبدیل کسینوسی گسسته بر روی طیف‌نگار مل به دست می‌آیند و ساختار کانال صوتی را به صورت فشرده نمایش می‌دهند:

- **۲۰ ضریب MFCC:** برای هر نمونه صوتی، ۲۰ ضریب اول MFCC محاسبه و میانگین و انحراف معیار هر ضریب استخراج گردید. ضریب اول معمولاً نمایانگر انرژی کلی، ضرایب بعدی نمایانگر ساختار فرمت‌ها (رزونانس‌های کانال صوتی) و ضرایب بالاتر نمایانگر جزئیات ظریف‌تر هستند. MFCC ها به دلیل توانایی بالا در مدل‌سازی ویژگی‌های آوایی، هسته اصلی بردار ویژگی این پروژه را تشکیل می‌دهند.

۵.۲.۳ ویژگی‌های کروماتیک (Chroma Features)

این ویژگی‌ها توزیع انرژی در ۱۲ کلاس نت موسیقیایی را نشان می‌دهند و برای تحلیل زیروبمی و تُنالیتی صدا مفید هستند:

- **۱۲ ویژگی کروماتیک:** برای هر یک از ۱۲ کلاس کروماتیک، میانگین و انحراف معیار انرژی محاسبه شده است. این ویژگی‌ها می‌توانند در تشخیص زبان‌هایی که ساختار تُنال متفاوتی دارند (مانند زبان‌های آهنگین) مؤثر باشند.

۶.۲.۳ ویژگی‌های تُنال (Tonnetz)

ویژگی‌های Tonnetz (شبکه تُنال) نمایشی از مرکز ثقل تُنال در فضای ۶ بعدی هستند و روابط هارمونیک بین نت‌ها را مدل می‌کنند:

- **۶ ویژگی :Tonnetz** این ویژگی‌ها از روی ویژگی‌های کروماتیک محاسبه شده و تغییرات تُنال در طول زمان را نشان می‌دهند. میانگین و انحراف معیار هر ۶ بعد محاسبه گردیده است.

۷.۲.۳ ویژگی‌های کنتراست طیفی (Spectral Contrast)

این ویژگی‌ها تفاوت بین قله‌ها و دره‌های طیفی را در هر زیرباند فرکانسی اندازه‌گیری می‌کنند:

- **۷ زیرباند کنتراست طیفی:** برای هر زیرباند، میانگین و انحراف معیار کنتراست طیفی محاسبه شده است. این ویژگی‌ها می‌توانند ساختار هارمونیک و نویزی سیگنال را بهتر نشان دهند.

۸.۲.۳ ویژگی‌های چندجمله‌ای (Poly Features)

این ویژگی‌ها ضرایب یک چندجمله‌ای درجه ۱ هستند که بر روی پوش طیفی برازش داده می‌شود:

- **۲ ویژگی چندجمله‌ای:** ضرایب ۰ و ۱ (میانگین) استخراج گردیده که نمایانگر شیب و عرض از مبدأ پوش طیفی هستند.

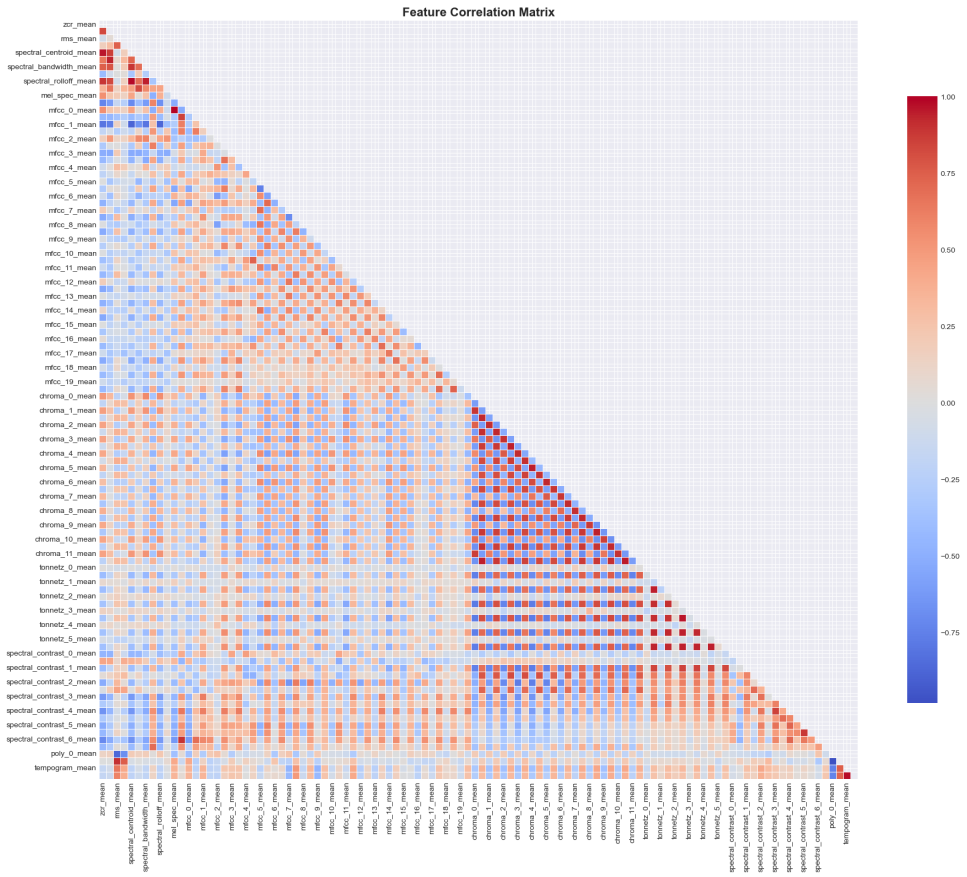
۹.۲.۳ ویژگی‌های ریتمیک (Tempogram)

تمپوگرام نمایشی از سرعت و ریتم سیگنال صوتی است و برای تحلیل ضرب‌آهنگ و ریتم گفتار مفید می‌باشد:

- **تمپوگرام:** با محاسبه پوشonset و سپس تمپوگرام، میانگین و انحراف معیار آن به عنوان ویژگی‌های ریتمیک استخراج شده‌اند. این ویژگی‌ها می‌توانند در تشخیص زبان‌هایی با ریتم متفاوت (مانند زبان‌های هجایی در مقابل فشارتکیه) مؤثر باشند.

۱۰.۲.۳ جمع بندی و مشاهده ویژگی ها

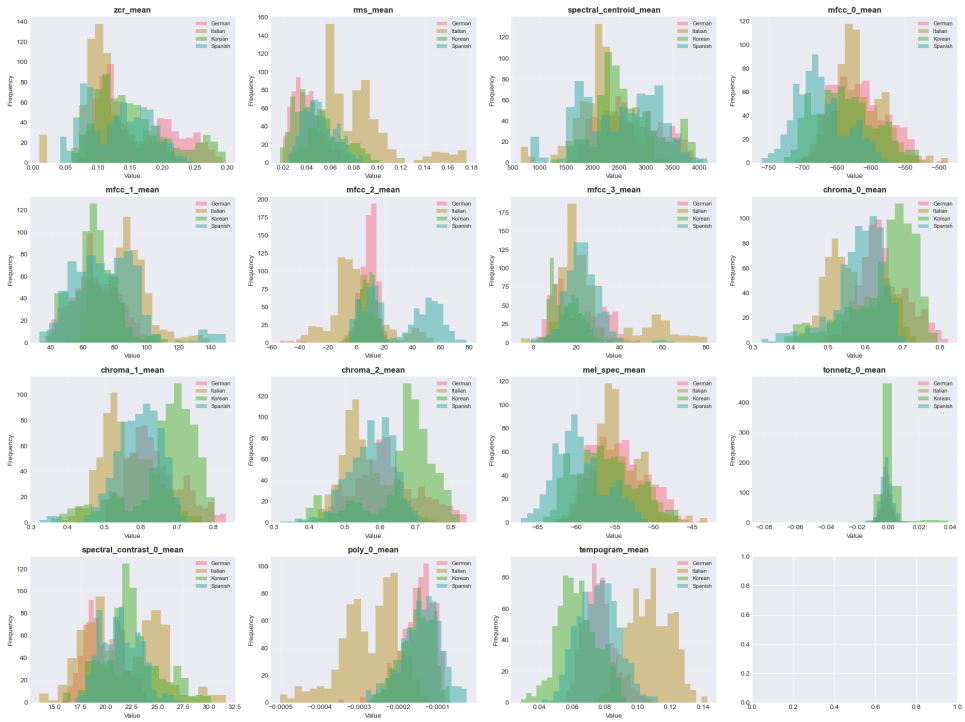
در شکل زیر ماتریس همبستگی ویژگی ها قابل مشاهده می باشد:



شکل ۳: ماتریس همبستگی ویژگی ها

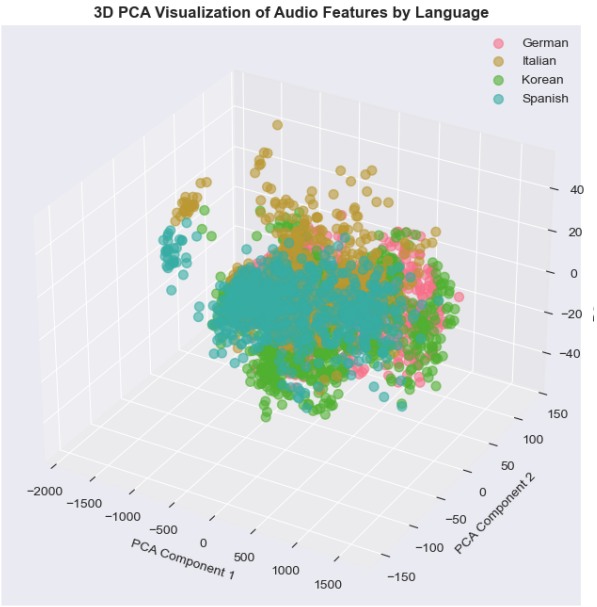
در شکل بالا همبستگی بین میانگین های ویژگی های مختلف محاسبه شده و با sns رسم شده است. رنگ های تیره تر نشان دهنده اندازه همبستگی بیشتر و رنگ های روشن تر نشان دهنده اندازه همبستگی کمتر می باشند.

در شکل زیر توزیع ویژگی های استخراج شده به ازای هر زبان مشاهده می شود:



شکل ۴: توزیع ویژگی های مختلف بر اساس زبان ها

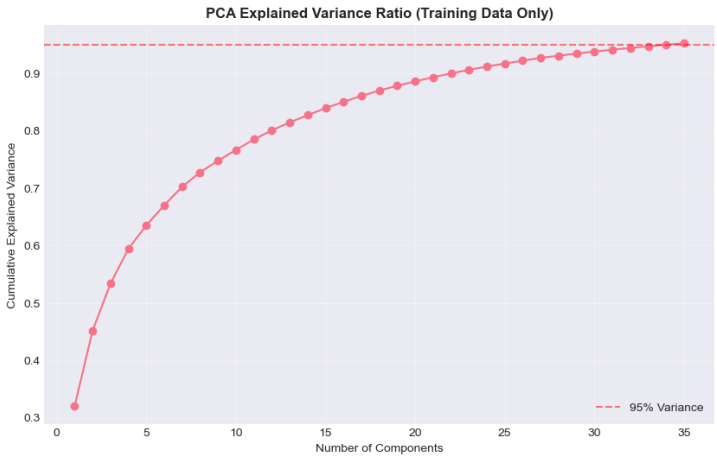
در نهایت نیز با انتخاب ویژگی های کلیدی استخراج شده و اعمال PCA روی آن ها به یک شکل ۳ بعدی به ازای زبان های مختلف رسیدیم که در زیر قابل مشاهده می باشد:



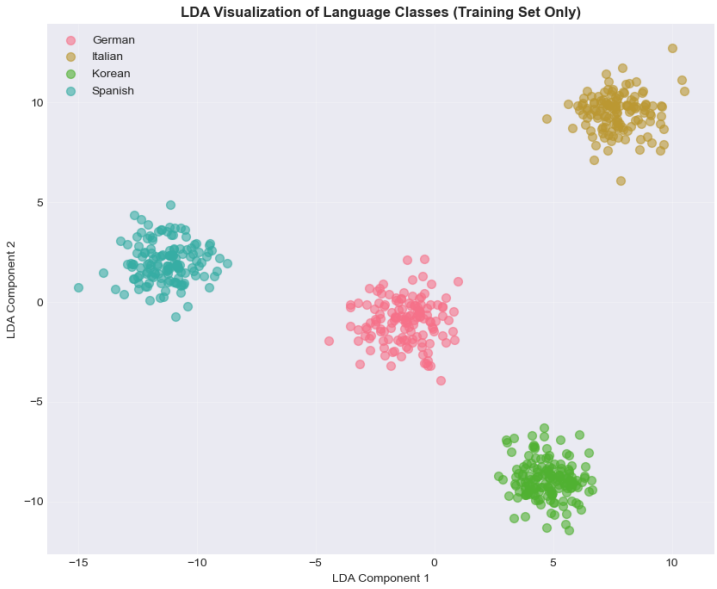
شکل ۵: نمایش سه بعدی داده های صوتی بر اساس کلید ویژگی های انتخاب شده

طبقه بندی

در ابتدا قبل از اینکه مدل هایی را برای آموزش در نظر بگیریم، روی داده ها دو پیش پردازش PCA و LDA را انجام می دهیم. با انجام PCA در نهایت امر به ۳۵ کامپوننت و با انجام LDA با در نظر گرفتن $n = 2$ برای تعداد کامپوننت ها، زبان های مختلف را از هم جدا کردیم. در دو شکل زیر PCA و LDA را مشاهده می کنید:



شکل ۶: نمایش نمودار PCA ویژگی ها



شکل ۷: نمایش LDA برای زبان های مختلف

۱.۴ مدل های مورد استفاده در طبقه بندی

در این پروژه، هفت الگوریتم مختلف طبقه‌بندی برای تشخیص زبان گفتاری انتخاب و پیاده‌سازی شدند. هدف از انتخاب این مجموعه متنوع، بررسی عملکرد الگوریتم‌های مختلف با رویکردهای گوناگون در مسئله تشخیص زبان و شناسایی بهترین مدل برای این کاربرد خاص بود. در ادامه به دلایل انتخاب هر دسته از مدل‌ها می‌پردازیم:

۱.۱.۴ مدل‌های مبتنی بر فاصله و مشابهت

- **K نزدیک‌ترین همسایه (KNN):** این الگوریتم به دلیل سادگی و عدم نیاز به مرحله آموزش صریح انتخاب شد. KNN با فرض اینکه نمونه‌های مشابه از نظر ویژگی‌ها، برچسب‌های مشابهی دارند، عمل می‌کند. با تنظیم پارامتر وزن‌دهی بر اساس فاصله (weights='distance')، به نقاط نزدیک‌تر تأثیر بیشتری داده شد. این مدل می‌تواند مرزهای تصمیم‌گیری غیرخطی و پیچیده را مدل کند.

۲.۱.۴ مدل‌های مبتنی بر مرز تصمیم‌گیری

- **ماشین بردار پشتیبان (SVM):** SVM با کرنل RBF به دلیل توانایی بالا در ایجاد مرزهای تصمیم‌گیری غیرخطی و عملکرد عالی در مسائل طبقه‌بندی با ابعاد بالا انتخاب شد. این الگوریتم با یافتن ابرصفحه‌ای که حاشیه جداسازی بین کلاس‌ها را بیشینه می‌کند، عمل می‌کند. تشخیص زبان به دلیل پیچیدگی ویژگی‌های صوتی، معمولاً نیازمند مرزهای غیرخطی است که SVM با کرنل RBF به خوبی از عهده آن برمی‌آید.

۳.۱.۴ مدل‌های مبتنی بر درخت و ensemble

- **جنگل تصادفی (Random Forest):** این الگوریتم به عنوان یک روش ensemble مبتنی بر درخت انتخاب شد که با ترکیب تعداد زیادی درخت تصمیم، دقت بالا و مقاومت در برابر بیش‌برازش را فراهم می‌کند. جنگل تصادفی با ۱۰۰ درخت و حداکثر عمق ۱۰، قادر به شناسایی تعاملات غیرخطی بین ویژگی‌ها و تعیین اهمیت آن‌ها است.

- **درخت تصمیم (Decision Tree):** این مدل به عنوان یک مدل پایه و تفسیرپذیر انتخاب شد تا بتوان ساختار تصمیم‌گیری را به صورت بصری تحلیل کرد. هرچند درخت‌های

تصمیم به تنهایی ممکن است دقت بالایی نداشته باشند، اما به عنوان بلوک ساختمانی مدل‌های ensemble اهمیت دارند.

۴.۱.۴ مدل‌های شبکه عصبی

- **پرسپترون چندلایه :(MLP)** شبکه عصبی پرسپترون چندلایه با دو لایه پنهان ۵۰ نورونی انتخاب شد تا توانایی یادگیری الگوهای بسیار پیچیده و غیرخطی را ارزیابی کنیم. با استفاده از regularization (alpha=۰.۰۱) و early stopping، از بیش‌برازش جلوگیری شد. MLP می‌تواند نمایش‌های سلسله‌مراتبی از ویژگی‌ها یاد بگیرد که برای داده‌های صوتی بسیار مفید است.

۵.۱.۴ مدل‌های خطی و احتمالاتی

- **رگرسیون لجستیک (Logistic Regression):** این مدل به عنوان یک طبقه‌بند خطی ساده و سریع انتخاب شد. با وجود محدودیت‌های خطی بودن، رگرسیون لجستیک می‌تواند به عنوان یک خط پایه (baseline) برای مقایسه با مدل‌های پیچیده‌تر عمل کند.

- **بیز ساده (Naive Bayes):** الگوریتم بیز ساده گاوسی بر اساس فرض استقلال ویژگی‌ها و توزیع نرمال آن‌ها انتخاب شد. این مدل با وجود فرض ساده‌کننده استقلال، در بسیاری از مسائل طبقه‌بندی متن و گفتار عملکرد خوبی دارد و بسیار سریع است.

۲.۴ معیارهای ارزیابی و تحلیل عملکرد مدل‌ها

برای ارزیابی جامع مدل‌ها، از معیارهای مختلفی استفاده شده است. نتایج به دست آمده در جدول یک خلاصه شده است.

مدل	دقت تست	F۱-Score
KNN	98.66	98.64
SVM	91.28	91.22
Random Forest	91.95	92.11
MLP	95.3	95.24
Decision Tree	82.55	81.76
Logistic Regression	92.62	92.60
Naive Bayes	83.22	82.26

جدول ۱: معیارهای ارزیابی مدل‌های طبقه‌بندی

۱.۲.۴ تحلیل معیارهای ارزیابی

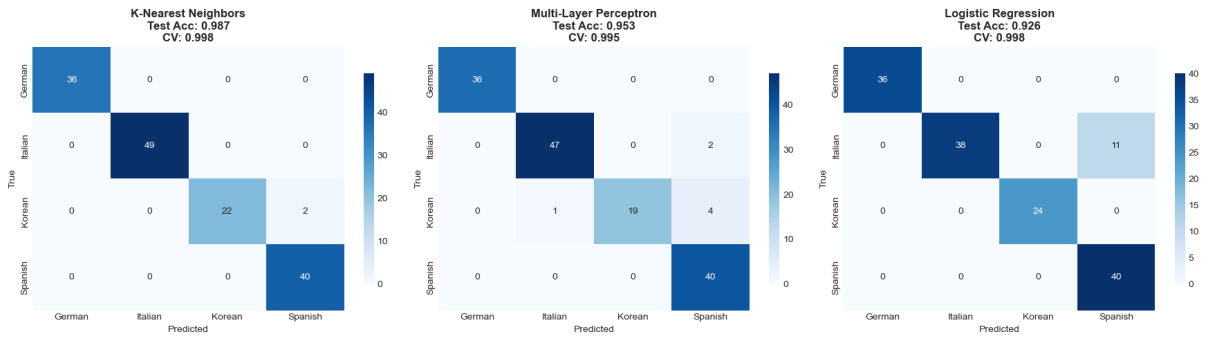
دقت (Accuracy): بالاترین دقت تست مربوط به مدل KNN با $k = 5$ با دقت 98.66 است. این نشان می‌دهد که KNN به خوبی توانسته مرزهای تصمیم‌گیری بین چهار زبان را یاد بگیرد. MLP و Logistic Regression نیز با دقت‌های 95.3 و 92.62 در رتبه‌های بعدی قرار دارند. پایین‌ترین دقت مربوط به Decision Tree با 82.55 است.

امتیاز F۱: امتیاز F۱ که میانگین وزنی دقت و فراخوانی است، نتایج مشابهی با دقت نشان می‌دهد. KNN با 98.64 بالاترین امتیاز را کسب کرده است. نزدیکی امتیاز F۱ به دقت نشان‌دهنده توزیع متوازن خطاها بین کلاس‌هاست. حال در شکل زیر نموداری از معیارهای مختلف دیگر یعنی stability و overfitting risk و cv means آمده است.

شکل ۸: مقایسه کلی مدل ها در معیارهای مختلف

۲.۲.۴ تحلیل ماتریس درهم‌ریختگی

در شکل زیر ماتریس درهم‌ریختگی سه مدل برتر نمایش داده شده است.



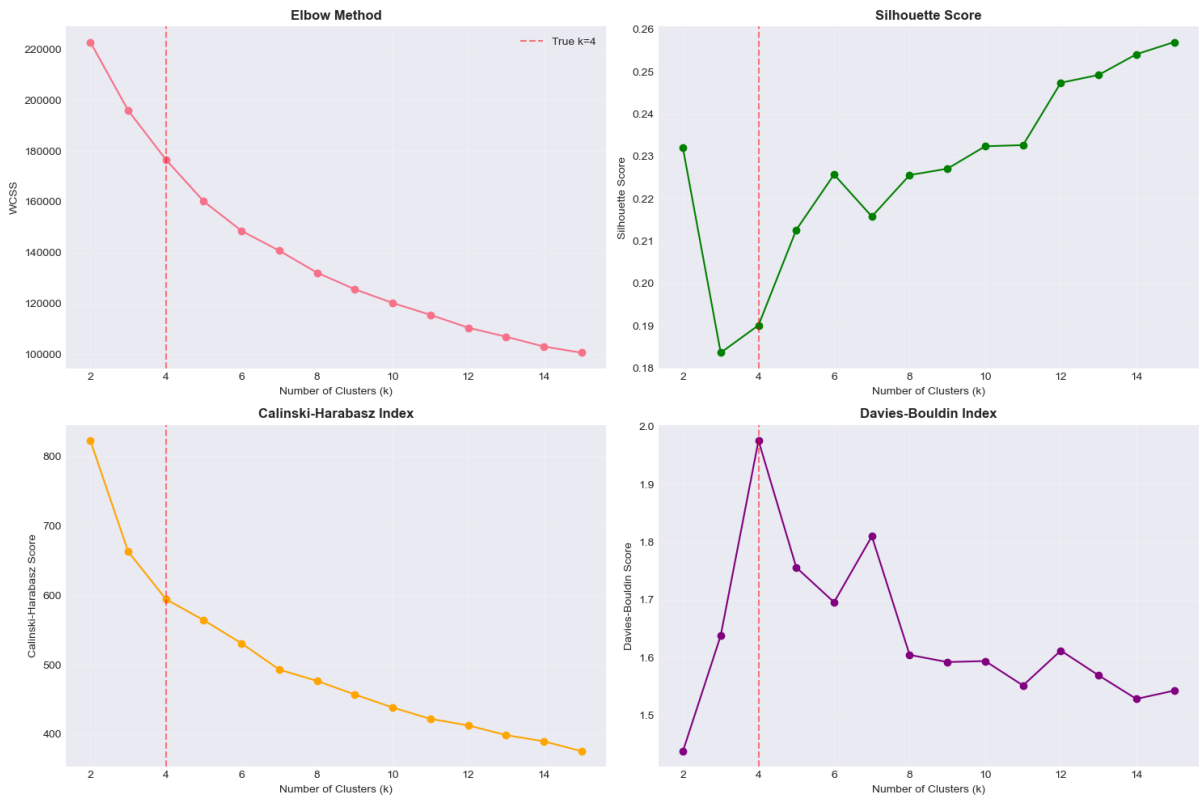
شکل ۹: ماتریس درهم‌ریختگی سه مدل برتر

خوشه بندی

در این بخش از پروژه، هدف اصلی گروه‌بندی نمونه‌های صوتی بر اساس ویژگی‌های استخراج‌شده از آن‌ها و مقایسه این گروه‌بندی با برچسب‌های واقعی زبان‌ها است. انتظار می‌رود که الگوریتم‌های خوشه‌بندی بتوانند چهار زبان مختلف (آلمانی، ایتالیایی، کره‌ای و اسپانیایی) را از یکدیگر تفکیک کنند.

۱.۵ تعیین تعداد بهینه خوشه ها

برای یافتن بهترین تعداد خوشه (k) برای الگوریتم‌هایی که نیاز به تعیین این پارامتر دارند (مانند K-Means و GMM و Agglomerative)، از چندین معیار ارزیابی داخلی استفاده شد. نتایج این ارزیابی‌ها در نمودار شکل ۱۰ قابل مشاهده است.



شکل ۱۰: نمودار معیارهای تعیین تعداد خوشه های بهینه

همان طور که در نمودارها مشخص است، می توان هریک از معیارها را این گونه تحلیل کرد:

- **روش آرنج (Elbow Method)**: نمودار WCSS کاهش شدیدی تا $k = 4$ نشان می‌دهد و پس از آن کاهش شیب ملایم‌تر می‌شود. نقطه $k = 4$ به عنوان "آرنج" نمودار قابل شناسایی است.

- **امتیاز سیلهویت(Silhouette Score)**: این معیار که نشان‌دهنده انسجام درون خوشه‌ای و جدایی بین خوشه‌ای است، در $k = 15$ بیشترین مقدار را دارد که با مقدار واقعی خوشه‌ها ($k = 4$) مطابقت ندارد.

- **شاخص Calinski-Harabasz و Davies-Bouldin**: در دو شاخص مقدار $k = 2$ را به عنوان تعداد خوشه های بهینه گزارش می دهند.

۲.۵ مدل های مورد استفاده در خوشه بندی

پنج دسته اصلی از الگوریتم‌های خوشه‌بندی بر روی داده‌ها اعمال شدند تا عملکرد آن‌ها در جداسازی زبان‌ها مقایسه شود. دلیل انتخاب هر دسته به شرح زیر است:

- **الگوریتم‌های مبتنی بر مرکز (Centroid-based)**: از الگوریتم K-Means استفاده شده است. این الگوریتم به دلیل سادگی، سرعت بالا و کارایی مناسب برای خوشه‌های کروی و با اندازه‌های مشابه، یک انتخاب استاندارد و پایه‌ای برای هر مسئله خوشه‌بندی به شمار می‌رود. ما این الگوریتم را با تعداد خوشه‌های مختلف (k) اجرا کردیم تا بهترین تعداد خوشه را نیز بیابیم.

- **الگوریتم‌های سلسله‌مراتبی (Hierarchical-based)**: الگوریتم Agglomerative Clustering انتخاب شد. این روش یک درخت سلسله‌مراتبی از خوشه‌ها ایجاد می‌کند و برای داده‌هایی که ساختار سلسله‌مراتبی دارند یا زمانی که تعداد خوشه‌ها نامشخص است، بسیار مفید است.

- **الگوریتم‌های مبتنی بر توزیع (Distribution-based)**: مدل Gaussian Mixture Model (GMM) به کار گرفته شد. GMM فرض می‌کند که داده‌ها از ترکیبی از چند توزیع گاوسی تولید شده‌اند. این مدل نسبت به K-Means انعطاف‌پذیرتر است و می‌تواند خوشه‌هایی با اشکال و اندازه‌های مختلف را تشخیص دهد.

- **الگوریتم‌های مبتنی بر چگالی (Density-based)**: الگوریتم DBSCAN انتخاب شد. این الگوریتم می‌تواند خوشه‌هایی با اشکال دلخواه را پیدا کند و نسبت به داده‌های پرت (نویز) مقاوم است. هدف از استفاده از آن، بررسی وجود ساختارهای غیرکروی در داده‌ها و شناسایی نمونه‌های پرت بود.

- **الگوریتم طیفی (Spectral Clustering)** : ین الگوریتم با استفاده از تجزیه مقادیر ویژه یک ماتریس شباهت، داده‌ها را در یک فضای با ابعاد پایین‌تر بازنمایی کرده و سپس خوشه‌بندی را انجام می‌دهد. این روش می‌تواند ساختارهای غیرخطی و خوشه‌های پیچیده را بهتر از K-Means تشخیص دهد.

۳.۵ معیار ارزیابی و کارایی مدل های مختلف

برای ارزیابی عملکرد الگوریتم‌های مختلف خوشه‌بندی، از سه معیار ارزیابی داخلی (Internal Evaluation Metrics) استفاده شده است: امتیاز سیلهویت (Silhouette Score)، شاخص Calinski-Harabasz و شاخص Davies-Bouldin. نتایج این ارزیابی‌ها در جدول دو خلاصه شده است.

model	num of clusters	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means (k=4)	4	۰.190	594.271	1.976
K-Means (k=optimal)	15	0.257	375.534	1.543
Agglomerative (k=4)	4	0.195	574.167	1.597
Agglomerative (k=optimal)	15	0.264	383.334	1.544
Gaussian Mixture (k=4)	4	0.191	555.715	1.750
Gaussian Mixture (k=optimal)	15	0.246	371.494	1.525
DBSCAN (eps=2.5)	14	-0.219	41.429	1.358
DBSCAN (eps=3.0)	30	-0.072	57.097	1.319
DBSCAN (eps=3.5)	44	0.108	87.323	1.373
Spectral Clustering	4	0.002	82.969	1.509

جدول ۲: معیارهای ارزیابی مدل های مختلف خوشه بندی

۴.۵ تحلیل نتایج خوشه‌بندی

بر اساس نتایج ارائه‌شده در جدول دو، تحلیل‌های زیر قابل استخراج است:

- **عملکرد کلی مدل‌ها**: بر اساس امتیاز سیلهویت، اکثر الگوریتم‌ها عملکردی بهتر از حالت تصادفی (امتیاز نزدیک به صفر) داشته‌اند، اما هیچ‌کدام به امتیاز بالایی (نزدیک به ۱) دست نیافته‌اند. این نشان‌دهنده آن است که اگرچه خوشه‌ها تا حدودی از یکدیگر جدا هستند، اما همپوشانی قابل توجهی بین آن‌ها وجود دارد. این موضوع با نمودار t-SNE که مرزهای نسبتاً مشخص اما با نقاط همپوشان بین زبان‌ها را نشان می‌داد، همخوانی کامل

دارد.

- **مقایسه الگوریتم‌ها**: الگوریتم Agglomerative Clustering با ۱۵ خوشه، با امتیاز سیلهویت ۰.۲۶۴ و شاخص دیویس–بولدین ۱.۵۴۴ بهترین عملکرد را در بین تمامی مدل‌ها از نظر این معیارها داشته است. این بدان معناست که خوشه‌های ایجادشده توسط این الگوریتم، فشرده‌تر و از یکدیگر جداتر هستند.

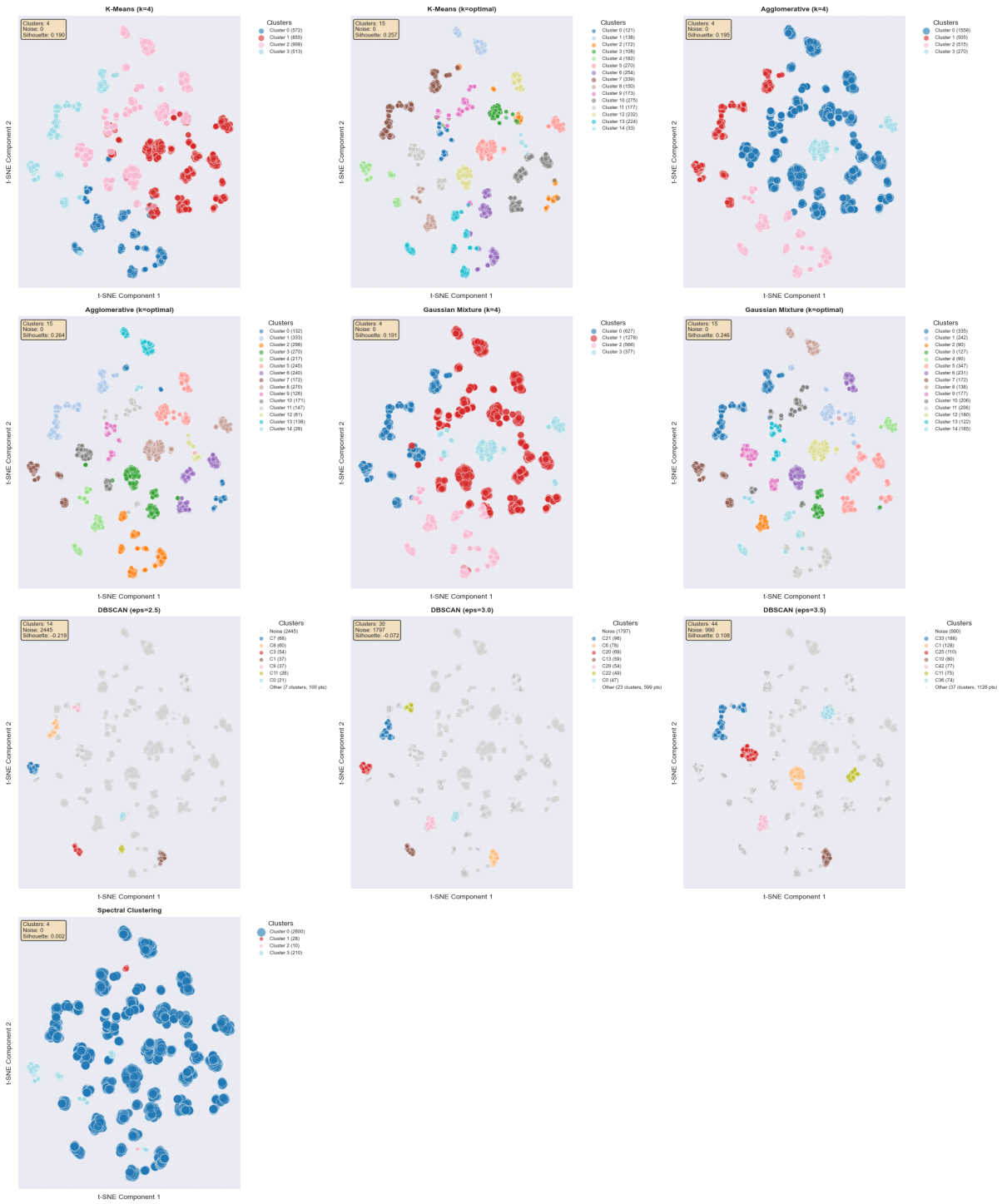
- **تأثیر تعداد خوشه‌ها**: در هر سه الگوریتم K-Means، Agglomerative و Gaussian Mixture، افزایش تعداد خوشه‌ها از ۴ به ۱۵ منجر به بهبود قابل توجه امتیاز سیلهویت و شاخص دیویس–بولدین شده است. این پدیده که در ارزیابی‌های داخلی خوشه‌بندی رایج است، نشان می‌دهد که با افزایش تعداد خوشه‌ها، داده‌ها به گروه‌های فشرده‌تری تقسیم می‌شوند. با این حال، این بهبود لزوماً به معنای تطابق بیشتر با برچسب‌های واقعی (۴ زبان) نیست و صرفاً نشان‌دهنده ساختار درونی بهتر از دید الگوریتم است.

- **عملکرد الگوریتم DBSCAN**: نتایج الگوریتم DBSCAN با سه مقدار متفاوت برای پارامتر ε (eps) گزارش شده است. در مقادیر پایین‌تر ε (۲.۵ و ۳.۰)، امتیاز سیلهویت منفی است که نشان‌دهنده تخصیص نادرست نقاط به خوشه‌ها و شکل‌گیری خوشه‌های ناپخته است. با افزایش ε به ۳.۵، امتیاز سیلهویت به ۰.۱۰۸ مثبت می‌رسد و شاخص Calinski-Harabasz نیز بهبود می‌یابد. با این حال، تعداد خوشه‌های یافت‌شده (۴۴) بسیار بیشتر از تعداد واقعی زبان‌ها است. این امر نشان می‌دهد که تنظیم پارامترهای DBSCAN برای این مجموعه داده دشوار بوده و ساختار چگالی داده‌ها به گونه‌ای نیست که این الگوریتم بتواند ۴ خوشهٔ منطبق با زبان‌ها را تشکیل دهد.

- **عملکرد الگوریتم Spectral Clustering**: این الگوریتم با امتیاز سیلهویت ۰.۰۰۲، ضعیف‌ترین عملکرد را در بین مدل‌های با ۴ خوشه داشته است. این نتیجه نشان می‌دهد که فرضیات این الگوریتم با ساختار داده‌ها سازگار نبوده و یا پارامتر affinity به درستی انتخاب نشده است.

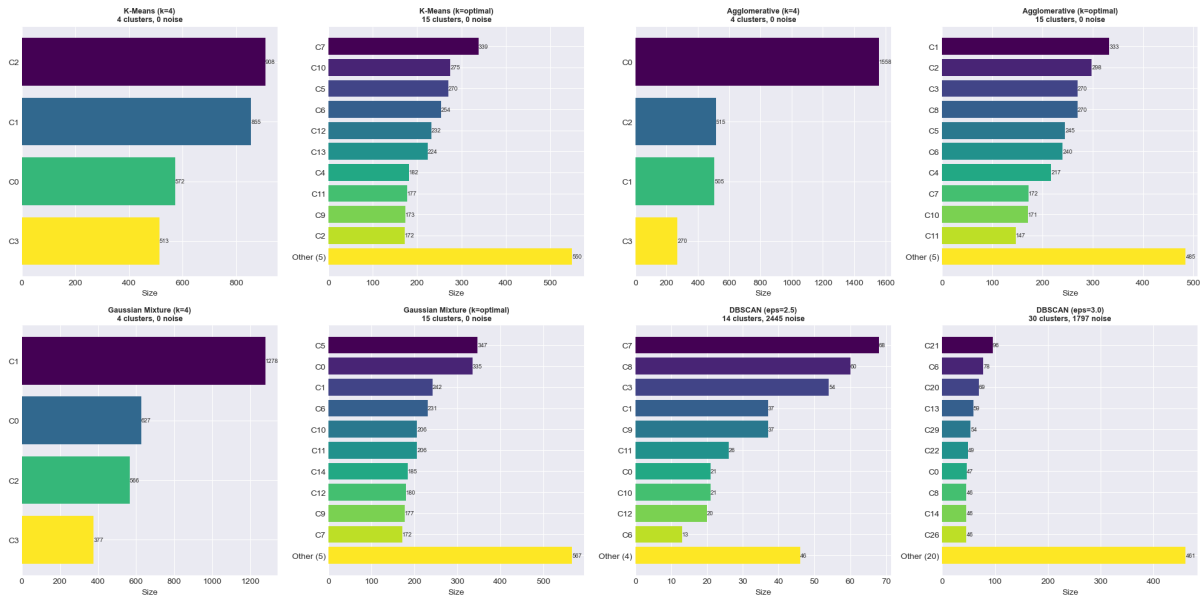
۵.۵ نمایش بصری الگوریتم های خوشه بندی

با اجرای این الگوریتم های خوشه بندی، دسته های زیر بدست آمد که می توان در شکل زیر آن ها را مشاهده کرد:



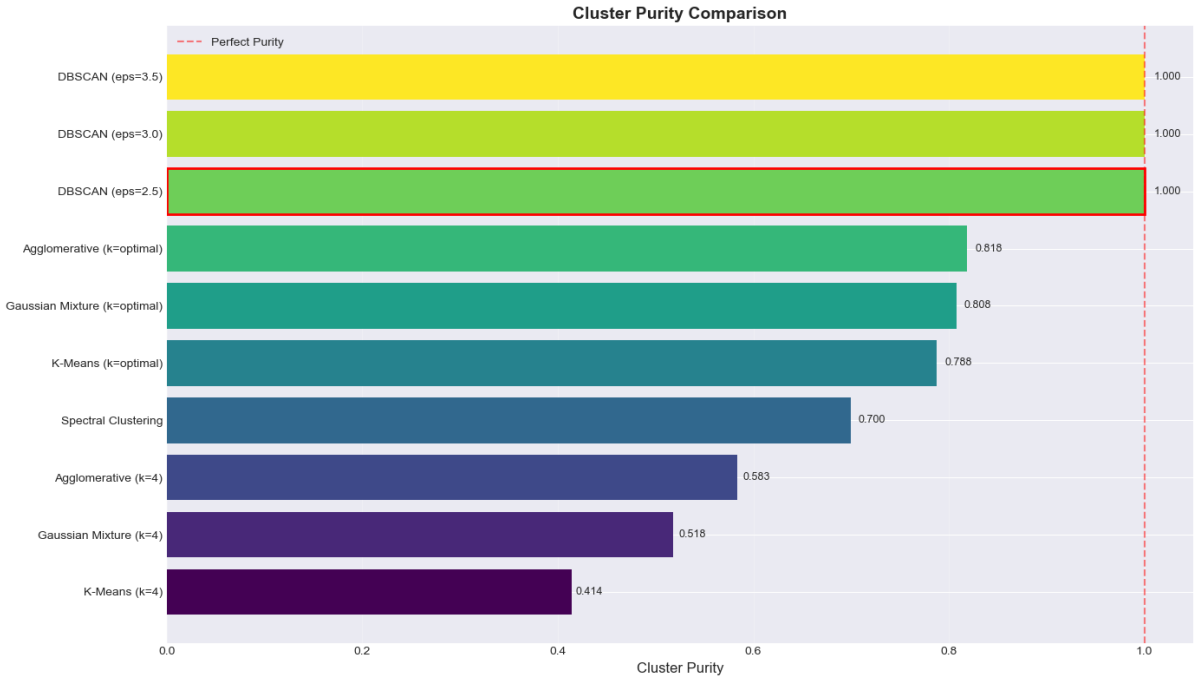
شکل ۱۱: دسته بندی های انجام شده در الگوریتم های مختلف

در هر نمودار تعداد خوشه ها و داده های مربوط به هر خوشه با رنگ مشخص شده است.
حال تفکیک داده های موجود در هریک از یک خوشه ها نیز در شکل زیر آمده است:



شکل ۱۲: توزیع داده ها در دسته های مختلف

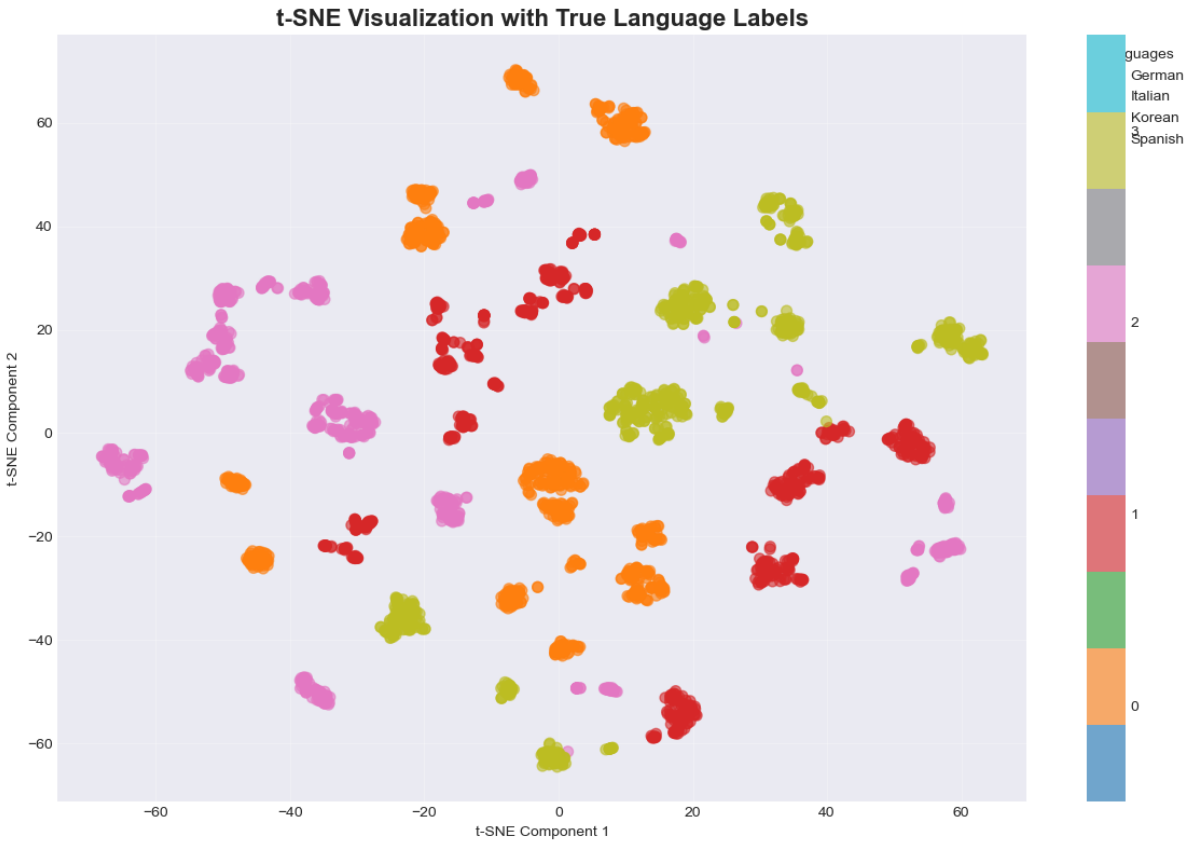
حال در ادامه معیار cluster purity برای تمام الگوریتم ها محاسبه شد و به صورت بصری در شکل زیر قابل مشاهده می باشد:



شکل ۱۳: مقایسه cluster purity بین الگوریتم های مختلف

طبق شکل بالا، اگر بخواهیم بهترین الگوریتم را با معیار purity بیان کنیم، الگوریتم DB-SCAN با eps برابر با 2.5 جواب مساله می باشد.

برای درک بصری از نحوه عملکرد الگوریتم‌های خوشه‌بندی، داده‌ها با استفاده از روش t-SNE به فضای دو بعدی تصویر شدند. تصویر زیر توزیع واقعی داده‌ها را بر اساس زبان نشان می‌دهد.



شکل ۱۴: نمایش داده ها با t-SNE به همراه برچسب های واقعی

همان‌طور که در تصویر t-SNE مشخص است، چهار زبان از یکدیگر تفکیک‌پذیر هستند، اما مرز بین آن‌ها کاملاً واضح و جدا نیست. این امر خوشه‌بندی کامل و بدون خطا را برای الگوریتم‌ها دشوار می‌سازد. بنابراین همان طور که در بخش های قبل دیدیم، هرکدام از الگوریتم های خوشه بندی تعداد خوشه های متفاوتی را تخمین زدند.

ارزیابی

در این قسمت به جمع بندی گزارش و ارزیابی کلی مدل های مورد استفاده در این پروژه می پردازیم.

۱.۶ ارزیابی مدل‌های طبقه‌بندی (Classification)

در این بخش، ۷ الگوریتم مختلف یادگیری ماشین بر روی داده‌های آزمایشی ارزیابی شده‌اند. معیارهای اصلی سنجش شامل دقت (Accuracy)، صحت (Precision)، یادآوری (Recall) و معیار F۱ (F1-Score) می‌باشند.

model	(Accuracy)	(Precision)	(Recall)	F1-Score
k-nearest_neighbors	**0.987**	**0.987**	**0.987**	**0.986**
multi-layer_perceptron	0.953	0.958	0.953	0.952
logistic_regression	0.926	0.942	0.926	0.926
random_forest	0.919	0.938	0.919	0.921
support_vector_machine	0.913	0.934	0.913	0.912
naive_bayes	0.832	0.882	0.832	0.823
decision_tree	0.826	0.865	0.826	0.818

جدول ۳: معیارهای دقیق ارزیابی مدل های طبقه بندی

- **عملکرد برتر:** مدل **k-nearest_neighbors** با کسب بالاترین امتیاز در تمامی معیارها (حدود ۹۹%)، بهترین عملکرد را در تشخیص دقیق زبان‌ها داشته است. این نشان‌دهنده آن است که داده‌ها در فضای ویژگی به خوبی تفکیک شده‌اند و نمونه‌های مشابه هر کلاس در کنار یکدیگر قرار گرفته‌اند.
- **عملکرد قوی:** مدل‌های **multi-layer_perceptron** (شبکه عصبی) و **logistic_regression** نیز با دقتی بالای ۹۲%، عملکرد بسیار خوبی از خود نشان داده‌اند.
- **عملکرد متوسط:** مدل‌های مبتنی بر درخت (**random_forest** و **decision_tree**) و ماشین بردار پشتیبان (**support_vector_machine**) دقت قابل قبولی در حدود ۹۱% و بالاتر دارند.

- **عملکرد ضعیف‌تر:** مدل **naive_bayes** و به ویژه **decision_tree** پایین‌ترین دقت را کسب کرده‌اند. این موضوع می‌تواند به دلیل فرضیات ساده‌کننده مدل naive_bayes و overfitting در مدل decision_tree باشد.

۲.۶ ارزیابی مدل‌های خوشه‌بندی (Clustering)

در این بخش، عملکرد ۱۰ روش و پیکربندی مختلف خوشه‌بندی بر روی داده‌ها ارزیابی شده است. ارزیابی خوشه‌بندی به دلیل ماهیت بدون نظارت آن، با معیارهای متفاوتی انجام می‌شود که کیفیت خوشه‌های ایجاد شده را می‌سنجند.

model	(num clusters)	(Silhouette)	(Davies-Bouldin)	Cluster Purity
Agglomerative (k=optimal)	15	0.264	1.544	**۰.818**
K-Means (k=optimal)	15	0.257	1.543	0.788
Gaussian Mixture (k=optimal)	15	0.246	1.525	0.808
DBSCAN (eps=2.5,3.0,3.5)	14 to 44	-0.108 to 0.108	1.319 to 1.373	**1.000**
Agglomerative (k=4)	4	0.195	1.597	0.583
K-Means (k=4)	4	0.190	1.976	0.414
Gaussian Mixture (k=4)	4	0.191	1.75	0.518
Spectral Clustering	4	0.002	1.509	0.7

جدول ۴: معیارهای دقیق ارزیابی مدل های خوشه بندی

- **تعیین تعداد بهینه خوشه :(k=۱۵)** الگوریتم‌هایی که با تعداد خوشه بهینه (k=optimal) اجرا شده‌اند (که توسط notebook به عدد ۱۵ تعیین شده)، عملکرد بسیار بهتری نسبت به حالت‌های با ۴ خوشه دارند. امتیاز سیلهوت بالاتر و شاخص دیویس_بولدین پایین‌تر در این موارد، نشان‌دهنده خوشه‌های فشرده‌تر و جدا از هم است. این نتیجه قویاً نشان می‌دهد که داده‌ها به طور طبیعی به ۱۵ گروه تقسیم می‌شوند (که احتمالاً منطبق بر ۱۵ speaker در دیتاست است).

- **بررسی خلوص خوشه (Purity):**

  - **DBSCAN:** این الگوریتم با خلوص ۱.۰ عملکرد خیره‌کننده‌ای دارد، به این معنی که تمام داده‌های درون خوشه‌های اصلی آن (غیر نویز) متعلق به یک کلاس واحد هستند. با این حال، امتیاز سیلهوت پایین (و حتی منفی) و تعداد خوشه‌های

بسیار زیاد (تا ۴۴ خوشه) نشان می‌دهد که DBSCAN داده‌ها را بسیار ریز و جزئی خوشه‌بندی کرده و احتمالاً داده‌های پرت (نویز) زیادی را نیز شناسایی کرده است.

- **Agglomerative (k=optimal):** این الگوریتم با بالاترین خلوص (۰.۸۱۸) و امتیاز سیلهوت (۰.۲۶۴) در میان روش‌های با k بهینه، بهترین عملکرد کلی را در ایجاد خوشه‌های معنادار و همگن دارد.

- **تعداد نادرست خوشه‌ها (k=۴):** همان‌طور که انتظار می‌رفت، تنظیم تعداد خوشه‌ها روی ۴ (که بسیار کمتر از مقدار بهینه است) منجر به افت شدید کیفیت خوشه‌بندی در تمام معیارها شده است. این امر بر اهمیت انتخاب صحیح تعداد خوشه‌ها تأکید می‌کند.

۳.۶ نتیجه‌گیری نهایی

- **برای طبقه‌بندی:** مدل **k-nearest_neighbors** به عنوان بهترین و قابل‌اعتمادترین مدل برای تشخیص زبان‌ها شناخته می‌شود. دقت بسیار بالای آن، آن را به گزینه‌ای ایده‌آل برای این کاربرد تبدیل می‌کند. مدل‌های **multi-layer_perceptron** و **logistic_regression** نیز جایگزین‌های بسیار مناسبی هستند.

- **برای خوشه‌بندی:** نتایج به شدت تأیید می‌کنند که تعداد بهینه خوشه‌ها **۱۵** است. در این میان، الگوریتم Agglomerative با ۱۵ خوشه با کسب بالاترین خلوص و امتیاز سیلهوت، بهترین عملکرد را در گروه‌بندی داده‌ها به خوشه‌های همگن و معنادار داشته است. هرچند DBSCAN در خلوص بی‌نقص عمل می‌کند، اما این به قیمت ایجاد خوشه‌های بسیار ریز تمام می‌شود.

مراجع