
بسم الله الرحمن الرحيم

دانشکده مهندسی برق و رباتیک

گروه الکترونیک

پایان نامه کارشناسی ارشد

طراحی و پیاده سازی سخت افزاری یک روش شناسایی برخط گوینده در بستر پردازشگر سیگنال TMS320C55xx

احمد معینی

استاد راهنما:

دکتر هادی گرایلو

مرداد ماه ۱۳۹۴

شماره : ۱۲۹۹/آ.ت.ب
تاریخ : ۹۴/۰۵/۰۶
ویرایش : ------

بسمه تعالی

مدیریت تحصیلات تکمیلی
فرم شماره (۶)

فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای :

احمد معینی رودبالی رشته : **بـرق - الکترونیک** گرایش : **سیستم**

تحت عنوان : **طراحی و پیاده سازی سخت افزاری یک روش شناسایی برخط گوینده در بستر پردازشگر سیگنال TMS۳۲۰C۵۵xx**

که در تاریخ **۹۴/۰۵/۰۶** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : عالی امتیاز ۱۹/۰۶) ☑ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹) ۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)

۳ـ خوب (۱۷/۹۹ ـ ۱۶) ۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)

۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضـــاء
۱ـ استـاد راهنمـا	هادی گرایلو	استادیار	
۲ـ استـاد مشـاور	ـــــ	ـــــ	ـــــ
۳ـ نماینده شورای تحصیلات تکمیلی	ساسان ناصح	استادیار	
۴ـ استـاد ممتحـن	حسن اورعی	استادیار	
۵ ـ استـاد ممتحـن	سید علی شهایی	دانشیار	

رئیس دانشکده :

تقدیم به پدر بزرگوار و مادر مهربانم...

تقدیر و تشکر:

در تمام مراحل انجام این پژوهش از راهنمایی و کمک های دکتر هادی گرایلو بهره مند بودم که از ایشان کمال تشکر را دارم. همچنین در زمینه مسایل پردازش گفتار از کمک های دکتر حسین مروی استفاده کردم و از ایشان سپاسگزاری می کنم. همچون سایر مراحل زندگی، از کمک و حمایت های خانواده ام بهره مند بودم و همچنین از کمک های دوستان خوبم استفاده کردم و از تمامی آن ها سپاسگزارم.

تعهد نامه

اینجانب **احمد معینی روبالی** دانشجوی دوره کارشناسی ارشد رشته **الکترونیک- سیستم** دانشکده **برق و رباتیک** دانشگاه شاهرود نویسنده پایان نامه **طراحی و پیاده سازی سخت افزاری یک روش شناسایی برخط گوینده در بستر پردازشگر سیگنال TMS320C55xx** تحت راهنمائی **دکتر هادی گرایلو** متعهد می شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه شاهرود می باشد و مقالات مستخرج با نام « دانشگاه شاهرود » و یا « Shahrood University» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.

چکیده

شناسایی گوینده یکی از شاخه‌های پردازش گفتار می‌باشد که کاربرد زیادی در سیستم‌های امنیتی مبتنی بر پردازش گفتار دارد. در طول چند دهه اخیر تلاش‌های زیادی برای بهبود کارایی و افزایش دقت سیستم‌های شناسایی گوینده انجام شده است. اگر چه در اغلب پژوهش‌ها تمرکز بر افزایش درصد شناسایی و بهبود کارایی سیستم می‌باشد، با این وجود کمتر بر اهمیت برخط بودن و پیاده سازی سخت-افزاری سیستم‌های شناسایی گوینده تاکید شده است. تمرکز این پایان نامه بر روی پیاده‌سازی برخط سیستم شناسایی گوینده بر روی یک سیستم سخت‌افزاری مبتنی بر پردازشگر سیگنال TMS320C5509A می‌باشد.

پس از بررسی‌های انجام شده، روش‌های طیفی زمان کوتاه[1] انتخاب شده و به صورت خاص از میان این دسته از روش‌های استخراج ویژگی، روش‌های MFCC و LPCC برای پیاده‌سازی سخت افزاری انتخاب شده‌اند. برای مدل کردن گوینده‌گان از مدل آمیخته گاوسی استفاده شده است.

نتایج آزمایش‌ها در محیط نرم افزار Matlab نشان دادند که روش استخراج ویژگی MFCC برای کاربردهای شناسایی گوینده نتایج بهتری نسبت به روش LPCC به دست می‌دهد. اگرچه روش استخراج ویژگی MFCC در محیط بدون نویز قابلیت شناسایی گویندگان را تا ۹۹ درصد دارد، اما در محیط‌های نویزی، کارایی سیستم به شدت کاهش پیدا می‌کند. علاوه بر این آموزش مدل‌های GMM با استفاده از پایگاه داده نویزی نیز باعث کاهش زیادی در درصد شناسایی سیستم می‌شود. به دلیل عدم دسترسی به امکانات لازم برای تهیه پایگاه داده بدون نویز و همچنین ضعف روش MFCC در مقابل نویز، از پایگاه داده TIMIT برای آزمایش سخت افزار استفاده شده است. در آزمایش‌های انجام شده، داده‌های آزمایش به صورت مستقیم بر روی پردازنده بارگذاری شده است. نتایج نشان دادند که الگوریتم انتخاب شده قابلیت استفاده به صورت بر خط را دارد. بیشترین درصد شناسایی بر روی سخت افزار طراحی شده برابر با ۷۸/۶ می‌باشد.

کلمات کلیدی: شناسایی گوینده، MFCC، LPCC، مدل آمیخته گاوسی، پردازنده سیگنال TMS320C5509A

[1] Short Time Spectral Features

مقالات استخراج شده از پایان نامه :

- معینی ا ، گرایلو ه، (۱۳۹۴) " مروری بر روش‌های شناسایی گوینده مستقل از گفتار با تاکید بر مدل آمیخته گاوسی"، کنفرانس ملی فناوری، انرژی و داده با رویکرد مهندسی برق و کامپیوتر، کرمانشاه.
- معینی ا ، گرایلو ه، (۱۳۹۴) " مقایسه روش‌های استخراج ویژگی MFCC و LPCC برای شناسایی گوینده مستقل از متن"، دومین همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، تهران.

فهرست مطالب

فصل اول: مقدمه ۱

۱-۱ پیش گفتار ۲

۱-۲ شناسایی هویت گوینده و تایید هویت گوینده ۲

۱-۳ شناسایی هویت گوینده ۳

۱-۴ تایید هویت گوینده ۳

۱-۵ اهمیت محتوای گفتار در سیستم شناسایی گوینده ۴

۱-۶ هدف پایان نامه ۵

۱-۷ ساختار پایان نامه ۶

فصل دوم: مروری بر مفاهیم و روش‌های شناسایی گوینده ۷

۲-۱ مقدمه ۸

۲-۲ معرفی بخش‌های سیستم شناسایی گوینده ۸

۲-۳ استخراج ویژگی ۹

۲-۴ روش استخراج ویژگی طیفی زمان کوتاه ۱۳

۲-۴-۱ پیش پردازش ۱۳

۲-۴-۲ دسته‌بندی روش‌های استخراج ویژگی طیفی زمان کوتاه ۱۷

۲-۴-۳ روش‌های استخراج ویژگی طیفی زمان کوتاه بر مبنای شنوایی انسان ۱۷

۲-۴-۳-۱ روش استخراج ویژگی ضرایب کپسترال در مقیاس مِل (MFCC) ۱۸
۲-۴-۳-۲ روش استخراج ویژگی ضرایب کپسترال در مقیاس خطی(LFCC) ۲۲
۲-۴-۳-۳ روش استخراج ویژگی ادراکی پیشگویی خطی(PLP)..... ۲۲
۲-۴-۳-۴ روش استخراج ویژگی ضرایب کپسترال در مقیاس گاماتن(GFCC) ۲۴
۲-۴-۳-۵ روش‌های بر مبنای تبدیل موجک ۲۵
۲-۴-۴ روش‌های استخراج ویژگی بر مبنای سیستم تولید گفتار انسان ۲۶
2-4-4-1 روش استخراج ویژگی کد کردن پیشبینی خطی(LPC) ۲۶
۲-۴-۴-۲ استخراج ویژگی به روش فرکانس طیف خطی(LSF)..... ۳۰
۲-۴-۵ پس پردازش..... ۳۰
۲-۴-۵-۱ نرمالیزه کردن میانگین کپسترال(CMN) ۳۱
۲-۴-۵-۲ مشتقات زمانی ۳۲
۲-۴-۵-۳ روش فیلتر کردن نسبی طیف(RASTA) ۳۳
۲-۴-۶ جمع بندی ۳۳
۲-۵ مدل کردن گوینده ۳۵
۲-۵-۱ مدل کردن گوینده با استفاده از VQ ۳۶
۲-۵-۲ مدل کردن گوینده با استفاده از مدل مخفی مارکوف ۴۰
۲-۵-۳ مدل کردن گوینده با استفاده از مدل آمیخته گاوسی..... ۴۲

2-5-3-1 مدل پس زمینه جامع(UBM) ۴۵

۲-۵-۳-۲ کلاسه بندی با استفاده از GMM ۴۸

۲-۵-۴ جمع بندی ۴۸

فصل سوم: معرفی سیستم سخت‌افزاری ۵۱

3-1 پردازشگرهای سیگنال ۵۲

۳-۱-۱ پردازنده‌های مهم شرکت TI ۵۲

۳-۱-۲ پردازنده TMS320C5509A ۵۴

۳-۲ مدار پردازشگر سیگنال ۵۶

۳-۲-۱ منبع تغذیه ۵۶

۳-۲-۲ مبدل داده‌ها ۵۸

۳-۲-۳ نمای کلی مدار ۶۱

فصل چهارم: بررسی نتیجه آزمایش‌ها..... ۶۳

۴-۱ معرفی پایگاه‌های داده استفاده شده در آزمایش‌ها ۶۴

۴-۱-۱ پایگاه داده TIMIT ۶۴

4-1-2 پایگاه داده vidTIMIT ۶۴

۴-۲ بررسی نتایج آزمایش‌ها ۶۵

4-2-1 مقایسه روش‌های استخراج ویژگی MFCC و LPCC ۶۵

۴-۲-۲ بررسی اثر پس پردازش بر کارکرد سیستم شناسایی گوینده ۷۶

۴-۳ انتخاب روش استخراج ویژگی برای پیاده‌سازی سخت افزاری ۸۴

۴-۴ بررسی پیاده سازی سخت‌افزاری الگوریتم ۸۷

فصل پنجم: جمع‌بندی و ارائه پیشنهادات برای کارهای آینده ۹۳

۵-۱ جمع‌بندی و نتیجه گیری ۹۴

۵-۲ پیشنهاد برای کارهای آینده ۹۶

پیوست: بررسی الگوریتم EM ۹۷

۶-۱ بیشینه شباهت ۹۸

۶-۲ بیشینه سازی امید ریاضی ۹۹

۶-۲-۱ محاسبه پارامترهای تابع شباهت برای ترکیب آمیخته ۱۰۰

مراجع ۱۰۴

فهرست شکل‌ها

شکل(۲-۱): دیاگرام بلوکی سیستم شناسایی گوینده ۹

شکل(۲-۲): دیاگرام بلوکی پیش پردازش سیگنال گفتار ۱۴

شکل(۲-۳): نمودار سیگنال گفتار قبل و بعد از اعمال فیلتر بالاگذر ۱۴

شکل(۲-۴): نمایش تبدیل فوریه گسسته سیگنال گفتار ۱۵

شکل(۲-۵): پنجره همینگ ۱۰۰ نقطه‌ای ۱۶

شکل(۲-۶): نمودار بلوکی استخراج ویژگی به روش MFCC ۱۹

شکل(۲-۷): نمودار مقیاس مِل بر مبنای مقیاس خطی ۲۰

شکل(۲-۸): یک نمونه فیلتر بانک در مقیاس مِل ۲۱

شکل(۲-۹): دیاگرام بلوکی استخراج ویژگی به روش LFCC ۲۲

شکل(۲-۱۰): دیاگرام بلوکی روش استخراج ویژگی PLP ۲۳

شکل(۲-۱۱): دیاگرام بلوکی استخراج ویژگی به روش GFCC ۲۵

شکل(۲-۱۲): دیاگرام بلوکی ساده شده سیستم تولید گفتار انسان ۲۷

شکل(۲-۱۳): ترکیب ضرایب CMN، دلتا و شتاب با یکدیگر ۳۳

شکل(۲-۱۴): روندنمای الگوریتم LBG ۳۸

شکل(۲-۱۵): نحوه عملکرد مدل آمیخته گاوسی ۴۴

شکل(۲-۱۶): دیاگرام بلوکی آموزش با استفاده از تطبیق MAP ۴۷

فصل سوم

شکل(۳-۱): JTAG جهت ارتباط کامپیوتر با پردازنده DSP..... ۵۵
شکل(۳-۲): مدار محافظ منبع تغذیه ۵۷
شکل(۳-۳): آی سی تغذیه و ادوات جانبی آی سی ۵۸
شکل(۳-۴): مدار کدک و ادوات جانبی همراه با آی سی ۶۰
شکل(۳-۵): مدار نهایی جهت شناسایی گوینده..... ۶۱

فصل چهارم

شکل(۴-۱) : مقایسه نتایج دو روش MFCC و LPCC با بردارهای ویژگی ۱۲ درایه‌ای ۷۱
شکل(۴-۲): مقایسه نتیجه‌های MFCC با بردارهای ۱۲ و ۳۶ درایه‌ای ۷۳
شکل(۴-۳): مقایسه نتیجه‌های LPCC با بردارهای ۱۲ و ۳۶ درایه‌ای ۷۵
شکل(۴-۴): مقایسه بین روشهای MFCC و LPCC ۷۵
شکل(۴-۵): مقایسه بین روشهای MFCC و MFCC + CMN ۷۷
شکل(۴-۶): مقایسه بین روشهای MFCC و MFCC + CMN و MFCC + CMN + Delta... ۷۹
شکل(۴-۷) : مقایسه روش‌های MFCC با ترکیبهای متفاوتی از پس پردازش‌ها..... ۸۱
شکل(۴-۸): مقایسه روش‌های MFCC با پس پردازش و بدون پس پردازش در محیط نویزی .. ۸۵
شکل(۴-۹): نمودار بلوکی الگوریتم پیاده‌سازی شده بر روی پردازنده DSP ۸۸

فهرست جدول‌ها

جدول(۴-۱): نرخ شناسایی گوینده به ازای بردار ویژگی ۱۲ درایه‌ای MFCC ۶۹

جدول(۴-۲): نرخ شناسایی گوینده به ازای بردار ویژگی ۱۲ درایه‌ای LPCC ۷۰

جدول(۴-۳): نرخ شناسایی گوینده به ازای بردار ویژگی ۳۶ درایه‌ای MFCC ۷۱

جدول(۴-۴): نرخ شناسایی گوینده به ازای بردار ویژگی ۳۶ درایه‌ای LPCC ۷۳

جدول(۴-۵): نرخ شناسایی با استفاده از MFCC جبران سازی شده توسط CMN ۷۷

جدول(۴-۶): نرخ شناسایی با استفاده از MFCC همراه با CMN و دلتا ۷۹

جدول(۴-۷): نرخ شناسایی به ازای استفاده از روش MFCC ، دلتا و دلتا – دلتا ۸۲

جدول(۴-۸): نرخ شناسایی به ازای روش MFCC با بردار ویژگی ۳۹ درایه‌ای ۸۳

جدول(۴-۹): بررسی تاثیر آموزش مدل‌های GMM با استفاده از داده‌های آموزشی نویزی ۸۶

جدول(۴-۱۰): زمان اجرای سخت‌افزاری بخش‌های مختلف الگوریتم ۹۰

جدول(۴-۱۱): زمان اجرای الگوریتم..... ۹۰

جدول(۴-۱۲): مقایسه درصد شناسایی در Matlab با پیاده‌سازی سخت افزاری ۹۱

فصل اول

مقدمه‌ای بر سیستم‌های تشخیص گوینده

۱-۱ پیش گفتار

ارتباط برقرار کردن با استفاده از گفتار یکی از ساده‌ترین روش‌های برقراری ارتباط در میان انسان‌ها می‌باشد. در طراحی سیستم‌های هوشمند، به کارگیری سیگنال گفتار برای ارتباط برقرار کردن با ماشین-های هوشمند موضوعی می‌باشد که در بحث هوشمند سازی ماشین‌ها مورد توجه بوده و با پیشرفت علم پردازش سیگنال و هوش مصنوعی و البته طراحی پردازنده‌های پرقدرت، این موضوع مورد بررسی‌های بیشتری نیز قرار گرفته است. به این دلیل که مشخصات سیستم تولید گفتار هر انسان با انسان دیگری متفاوت بوده و ویژگی‌های خاص خود را دارد، از سیگنال گفتار در سیستم‌های امنیتی و شناسایی افراد به صورت گسترده استفاده می‌شود. بحث شناسایی گوینده نیز در راستای تامین امنیت یک مجموعه کاربرد زیادی دارد. استفاده از سیگنال گفتار برای شناسایی هویت افراد راه مطمئنی می‌باشد که در عین حال استفاده از آن نیز برای کاربرهای یک مجموعه یا سازمان راحت می‌باشد. می‌توان با استفاده از روش‌های متنوع استخراج ویژگی و مدل کردن گوینده که در طول چند دهه گذشته ارائه شده است، یک سیستم مطمئن امنیتی بر مبنای صدای افراد تهیه کرد.

۱-۲ شناسایی هویت گوینده و تایید هویت گوینده

در بحث تشخیص گوینده[1]، دو موضوع اساسی مورد بحث قرار می‌گیرد. این دو موضوع عبارتند از شناسایی هویت گوینده[2] و تایید هویت گوینده[3]. فرض کنید یک سیگنال گفتار مربوط به یک گوینده ناشناس وارد سیستم می‌شود. وظیفه سیستم شناسایی گوینده این است که تعیین کند این سیگنال ورودی مربوط به کدام یک از گویندگان موجود در سیستم است، در صورتی که وظیفه سیستم تایید

[1] Speaker Recognition
[2] Speaker Identification
[3] Speaker Verification

گوینده این است که تعیین کند سیگنال گفتار ورودی به سیستم، به مجموعه گویندگان موجود در سیستم تعلق دارد و یا اینکه مربوط به یک گوینده خارج از این مجموعه باشد[1].

۱-۳ شناسایی هویت گوینده

سیستم شناسایی گوینده دارای دو بخش کلی می‌باشد. اولین بخش آموزش است. در بخش آموزش از صدای هر گوینده، که هویت او از قبل تایید شده است، استفاده می‌شود تا یک مدل برای آن گوینده ساخته شود. این عمل را برای تمام گویندگانی که قرار است عضوی از سیستم باشند انجام می‌دهیم. معمولا مرحله آموزش قبل از شروع به کار سیستم انجام می‌شود. دومین بخش آزمایش می‌باشد. در این بخش یک سیگنال گفتار وارد سیستم می‌شود و با هر کدام از مدل‌های موجود در سیستم به صورت جداگانه مقایسه شده و هر گوینده‌ای که بیشترین امتیاز را در این مقایسه‌ها کسب کند، سیگنال گفتار ورودی متعلق به او می‌باشد. در سیستم‌هایی که دارای یک مجموعه بسته می‌باشند، سیگنال‌های ناشناس ورودی حتما مربوط به یکی از گویندگان موجود در سیستم است و مساله این می‌شود که بررسی کنیم سیگنال گفتار ورودی به سیستم، به کدام یک از گویندگان مجموعه تعلق دارد. در اینگونه سیستم‌ها، معیار کیفیت کارایی سیستم، نرخ شناسایی گوینده است. سیستم‌های با مجموعه بسته در سازمان‌ها و مکان‌هایی استفاده می‌شود که گویندگان از قبل مشخص شده و هیچ گوینده خارج از سیستمی وجود ندارد[2].

۱-۴ تایید هویت گوینده

برای تایید هویت گوینده به این ترتیب عمل می‌کنیم که ابتدا فرض می‌کنیم سیگنال گفتار ورودی مربوط به یکی از گویندگان موجود در سیستم می‌باشد. در مرحله بعد سیگنال گفتار ورودی را هم با مدل

گوینده‌ای که فرض کرده‌ایم سیگنال گفتار متعلق به اوست و هم با یک مدل جامع که از تمام گویندگان موجود در سیستم ساخته می‌شود، مقایسه می‌کنیم. در ادامه نسبت این دو مقایسه را محاسبه و آن را با یک آستانه که از قبل تعریف کرده‌ایم مقایسه می‌کنیم. اگر نسبت این دو مقایسه از آستانه بیشتر باشد، فرض ما صحیح بوده و سیگنال گفتار ورودی مربوط به گوینده مورد نظر می‌باشد و اگر اندازه نسبت دو مقایسه کمتر از آستانه باشد، فرض ما اشتباه بوده است. برای بررسی سیستم تایید گوینده معمولا از دو معیار استفاده می‌شود. اولین معیار عبارت است از تعداد دفعاتی که گوینده صحیح به اشتباه رد شده است و دومین معیار، تعداد دفعاتی است که گوینده اشتباه به عنوان گوینده صحیح تایید شده است[1, 3].

در سیستم‌های با مجموعه باز، گویندگانی که از سیستم استفاده می‌کنند می‌توانند عضوی از مجموعه گویندگان موجود در سیستم باشند و یا اینکه خارج از این مجموعه باشند. در اینگونه سیستم‌ها وظیفه ابتدایی این است که بررسی کنیم سیگنال گفتار ورودی به سیستم، مربوط به گویندگان موجود در سیستم می‌باشد و یا اینکه خارج از مجموعه گویندگان است. در مرحله بعد اگر تایید کردیم که این سیگنال گفتار مربوط به گویندگان مجموعه است، آنگاه هویت او را مشخص می‌کنیم. وقتی که مرحله تایید گوینده را از چنین سیستمی حذف کنیم، هر سیگنال گفتاری که به سیستم وارد شود، فارغ از اینکه مربوط به مجموعه گویندگان باشد و یا اینکه خارج از این مجموعه باشد، به عنوان یکی از گویندگان موجود در سیستم شناسایی شده و به صورت تصادفی هویت یکی از اعضای مجموعه به آن تعلق می‌گیرد [1].

۱-۵ اهمیت محتوای گفتار در سیستم شناسایی گوینده

می‌توان برای بهبود کیفیت سیستم‌های شناسایی گوینده، اطلاعات مربوط به محتوای گفتار را نیز در نظر گرفت، به این ترتیب که سیگنال گفتار به کلمات خاص محدود شود(برای مثال یک رمز، یا اسم

گوینده و ....). به این روش شناسایی گوینده، شناسایی وابسته به متن می‌گویند. گرچه این سیستم کارایی خوبی دارد اما استفاده از آن محدویت‌هایی دارد. نوع دیگری از سیستم شناسایی گوینـده کـه کـار بـا آن راحت‌تر می‌باشد، سیستم شناسایی گوینده مستقل از محتوای گفتار است. در این سیستم بدون توجه بـه محتوای گفتار، عمل شناسایی گوینده انجام می‌شود[2].

۱-۶ هدف پایان نامه

در کارهایی که در زمینه شناسایی گوینده انجام شده اسـت، معمـولا موضـوع بهبـود کیفیـت ایـن سیستم و افزایش درصد شناسایی گویندگان مورد بررسی قرار گرفته است. این بهبود کیفیت می‌توانـد در زمینه بهبود ویژگی‌های استخراج شده، مدلی که برای مدل کردن گویندگان استفاده شده است و یـا نـوع کلاسه‌بندی که برای کلاسه‌بندی گوینده‌ها استفاده شده است، باشد. اما معمولا کمتر بـه موضـوع پیـاده-سازی سخت افزاری این سیستم پرداخته شده است.

هدف این پایان نامه پیاده سازی سخت افزاری یک سیستم شناسایی گوینده توسـط مـدار طراحـی شده بر مبنای پردازنده سیگنال TMS320C5509A می‌باشد. به همین دلیل ابتدا بـه بررسـی روش‌هـای مختلف شناسایی گوینده‌ای پرداخته شده است که توانایی استفاده به صورت برخط[1] را دارنـد و سـپس از میان این روش‌ها، روش‌هایی انتخاب شده‌اند که توانایی پیاده‌سازی سخت‌افزاری به صورت بهینه را داشته باشند. در ادامه، شبیه‌سازی‌هایی انجام شده و بهترین روش با توجه بـه امکانـاتی کـه پردازنـده سـیگنال موجود در اختیار کاربر قرار می‌دهد و کیفت شناسایی گوینده روش مورد بررسی، انتخاب شـده اسـت. در نهایت مداری بر مبنای پردازنده TMS320C5509A طراحی شده و الگوریتم شناسایی گوینده انتخابی بر روی آن پیاده سازی و بررسی شده است.

[1] Online

۱-۷ ساختار پایان نامه

در فصل دوم به بررسی مفاهیم سیستم‌های شناسایی گوینده پرداخته شده و روش‌های پرکاربرد استخراج ویژگی، مدل کردن و کلاسه‌بندی مورد بحث و بررسی قرار گرفته است. در ادامه در فصل سوم سیستم سخت‌افزاری پیشنهادی معرفی شده است. بعد از معرفی سخت افزار، در فصل چهارم، نتایج آزمایشات در محیط Matlab برای انتخاب مناسب‌ترین روش شناسایی گوینده و همینطور نتیجه آزمایش بر روی سخت افزار طراحی شده ارائه شده است. در فصل پنجم به جمع بندی و ارائه پیشنهادات برای کارهای آینده پرداخته شده است و در نهایت در قسمت پیوست در انتهای پایان نامه، در مورد الگوریتم EM و نحوه استفاده از آن به صورت مختصر صحبت شده است.

فصل دوم

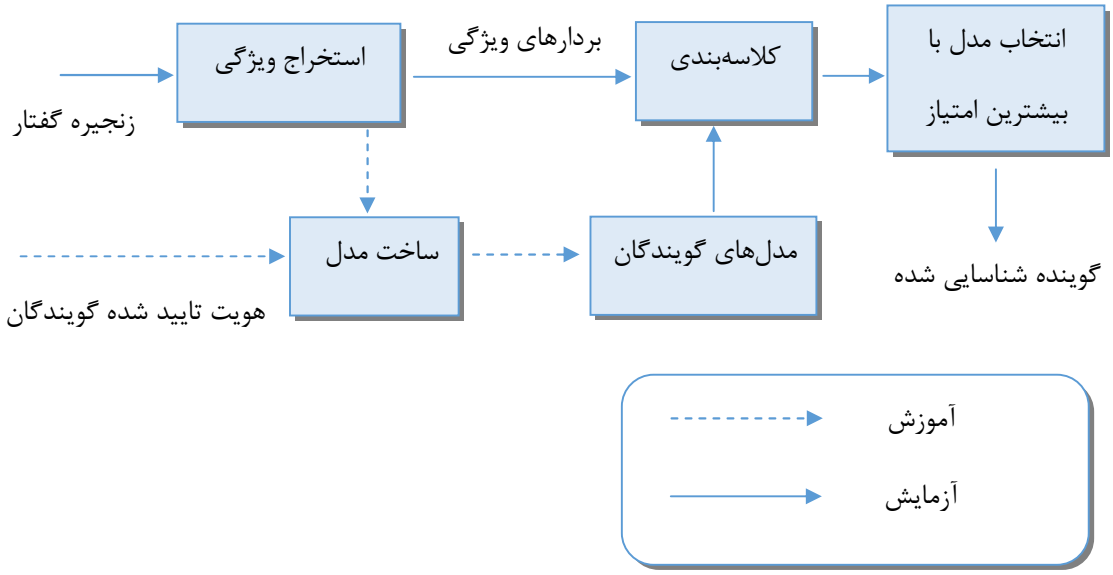
مروری بر مفاهیم و روش‌های شناسایی گوینده

۲-۱ مقدمه

در این فصل ابتدا به طور کلی به معرفی سیستم شناسایی گوینده پرداخته می‌شود و در ادامه قسمت‌های مختلف این سیستم معرفی می‌شود. روش‌های مختلف استخراج ویژگی، مدل کردن گوینده و کلاسه بندی که در مراجع گوناگون ارائه شده مورد بررسی قرار گرفته است.

۲-۲ معرفی بخش‌های سیستم شناسایی گوینده

در شکل (۲-۱) نمودار بلوکی یک سیستم شناسایی گوینده مشاهده می‌شود. در این شکل هر دو قسمت آموزش و آزمایش نمایش داده شده است. در قسمت آموزش ابتدا از سیگنال گفتار مربوط به هر گوینده استخراج ویژگی می‌شود و سپس با استفاده از روش‌های مختلف مدل کردن، آنها را مدل می‌کنند. در قسمت آزمایش، ابتدا ویژگی‌های سیگنال گفتار ورودی به سیستم را استخراج می‌کنند و در مرحله بعد با مدل‌هایی که در قسمت آموزش ساخته شده است تطبیق می‌دهند. هر کدام از مدل‌ها که امتیاز بیشتری را کسب کند، سیگنال گفتار مربوط به همان گوینده می‌باشد. در ادامه این فصل قسمت‌های مختلف سیستم شناسایی گوینده تشریح شده است[1].



شکل(۲-۱): دیاگرام بلوکی سیستم شناسایی گوینده[1]

۲-۳ استخراج ویژگی

اولین مرحله در یک سیستم شناسایی گوینده، استخراج ویژگی‌های مناسب از سیگنال گفتار می‌باشد. استخراج ویژگی یک مرحله حیاتی برای سیستم بوده و تاثیر بسزایی در کارایی سیستم دارد[4] . طبیعتا نمی‌توان از هر نوع روش استخراج ویژگی استفاده کرد و باید ویژگی‌هایی انتخاب شوند که به بهترین وجه بتوانند سیگنال گفتار را مدل کنند و البته برای هدف ما، یعنی شناسایی گوینده مستقل از گفتار، نیز مناسب باشند. در ضمن در کاربردهای مد نظر این پایان نامه، روش استخراج ویژگی که انتخاب می‌شود باید برای پیاده‌سازی سخت افزاری مناسب بوده و قابلیت پیاده‌سازی برخط را نیز داشته باشد. اما ویژگی‌هایی که برای استفاده در سیستم شناسایی گوینده استفاده می‌شوند باید چه خصوصیاتی داشته باشند؟ در ادامه به چند نمونه از این خصوصیات اشاره شده است. به ویژگی‌ای ویژگی مناسب می‌گویند که[2, 5, 6]:

- دارای تغییرات بین گوینده‌ای زیاد و تغییرات درون گوینده‌ای کم باشد

- در مقابل نویز و اعوجاج مقاوم باشد
- به دفعات زیاد و به صورت طبیعی در سیگنال گفتار اتفاق بیفتد
- اندازه‌گیری آن از سیگنال گفتار آسان باشد
- به راحتی قابل تقلید نباشد
- سلامتی(سیستم تولید گفتار) گوینده و یا سبک صحبت کردن شخص بر روی آن تاثیر نداشته باشد
- برای مدل کردن گوینده نیاز به تعداد بسیار زیادی بردار ویژگی نباشد

ابعاد بردار ویژگی و همینطور تعداد این بردارها به صورت نمایی بر میزان حجم محاسبات در هنگام آموزش و آزمایش سیستم تاثیر می‌گذارند و با افزایش تعداد و خصوصا ابعاد بردار ویژگی، بار محاسباتی بیشتر می‌شود[9-7]. اهمیت موضوع پیچیدگی محاسباتی در هنگامی که هدف سیستم شناسایی گوینده برخط باشد دوچندان می‌شود.

روش‌های مختلفی برای طبقه‌بندی کردن ویژگی‌ها وجود دارد. برای نمونه به منظور دسته‌بندی ویژگی‌ها، می‌توان دسته‌های زیر را معرفی کرد[2]:

- ویژگی‌های طیفی زمان کوتاه[1].
- ویژگی‌های ریتمیک[2].
- ویژگی‌های سطح بالا

[1]Short Term Spectral Features
[2]Prosodic Features

سیگنال گفتار وقتی که در بازه‌های زمانی کوتاه(در حدود چند میلی ثانیه) بررسی شود، دارای خصوصیات شبه ایستا[1] می‌باشد. روش طیفی زمان کوتاه از این خاصیت بهره می‌برد و با بررسی سیگنال گفتار در بازه‌های زمانی کوتاه، سیگنال را در آن بازه مدل می‌کند و مشخصه‌های سیگنال را استخراج می‌کند. در این روش می‌توان ویژگی‌های سیگنال گفتار را در حوزه زمان و یا فرکانس استخراج کرد[10].

صدای یک گوینده شامل اطلاعات گوناگونی مانند استرس، لهجه، ریتم، نوع تلفظ، آهنگ صدا و ... می‌باشد. این مشخصات در هر گوینده با گوینده دیگر متفاوت است و در روش استخراج ویژگی ریتمیک سعی بر این است که با استفاده از این خصوصیات صدای گوینده، ویژگی‌هایی استخراج کنیم که بتوان از آن‌ها برای مدل کردن گوینده استفاده کرد[4].

گاهی برای مدل کردن یک سیگنال گفتار به خصوصیات پیشرفته‌تری در سطح محاوره مراجعه می‌کنند، برای مثال خصوصیاتی مانند معنای کلمات و یا نوع کلماتی که استفاده میکنیم. به ویژگی‌هایی که با استفاده از این خصوصیات استخراج می‌شوند، ویژگی‌های سطح بالا می‌گویند[2].

برای محققی که در زمینه پردازش گفتار کار می‌کند این سوال پیش می‌آید که کدام یک از این روش‌های استخراج ویژگی روش بهتری است؟ باید گفت که قبل از هر چیز انتخاب روش استخراج ویژگی به نوع کاربرد بستگی دارد. برای مثال وقتی که قرار است یک سیستم برخط طراحی شود، استفاده از ویژگی‌های طیفی زمان کوتاه انتخاب عاقلانه‌تری نسبت به ویژگی‌های سطح بالا می‌باشد. زیرا زمانی که صرف استخراج ویژگی‌های زمان کوتاه می‌شود زمان کمتری است. علاوه بر این انتخاب مدلی که قرار است با ویژگی‌های استخراج شده آموزش داده شود نیز در انتخاب روش تاثیر گذار است. مثلا برای آموزش دادن یک سیستم مبتنی بر مدل مخفی ماکوف (HMM)[2] به میزان داده‌های زیادی احتیاج است، در

[1]Quasi Stationary
[2]Hidden Markov Model

صورتی که برای سیستمی که از مدل آمیخته گاوسی[1](GMM) استفاده می‌کند، بـه داده‌هـای بـه نسـبت کمتری برای آموزش مدل‌ها نیاز است[1, 11]. روش‌های اسـتخراج ویژگـی کـه معرفـی شـد، بـه صـورت عمومی یک سری خصوصیات دارند که در هنگام انتخاب آنها باید به این خصوصیات دقت کـرد. در مـورد ویژگی‌های طیفی زمان کوتاه یکی از مهمترین خصوصیاتی که می‌توان به آن اشـاره کـرد، سـرعت بـالای استخراج ویژگی است. این سرعت بالای استخراج ویژگی‌ها مزیت مهمی در سیستم‌های بـرخط محسـوب می‌شود. در ضمن برای آموزش مدل به کمک ویژگی‌های طیفی زمان کوتاه، نیاز بـه مقـادیر زیـادی داده برای آموزش نیست و به راحتی نیز قابل استخراج هستند[12, 13].

بر خلاف روش‌های طیفی زمان کوتاه، روش‌های استخراج ویژگی سطح بالا به زمان بیشتری بـرای استخراج ویژگی نیاز دارند و این باعث می‌شود که برای کاربردهای برخط مناسب نباشـند. در ضـمن ایـن ویژگی‌ها به میزان زیادی داده برای استخراج ویژگی و مدل کردن گوینده نیاز دارند. شـاید بـه بزرگتـرین حسنی که برای ویژگی‌های سطح بالا در مقابل ویژگی‌های طیفی زمان کوتاه می‌توان اشاره کرد، مقاومـت آنها در مقابل نویز و اثرات کانال انتقال باشد. ویژگی‌های ریتمیک از نظر قدرت، در بین دو روش استخراج ویژگی طیفی زمان کوتاه و سطح بالا قرار می‌گیرند. برای مثال سرعت آنها از روش‌های سطح بـالا بیشـتر ولی از روشهای طیفی زمان کوتاه کمتر می‌باشد و یا در مقایسه با ویژگی‌های سطح بالا به میزان کمتـری داده برای استخراج ویژگی نیاز دارند[4, 14].

با توجه به مطالبی که ارائه شد، برای کاربردهـای شناسـایی گفتـار و خصوصـا شناسـایی گوینـده، معمولا از ویژگی‌های طیفی زمان کوتاه استفاده می‌شود[15-21]. دلیلی که در این پایان نامه، این دسته از روش‌ها برای استخراج ویژگی انتخاب شد این است که برای پیاده‌سـازی سـخت افـزاری مناسـب بـوده و همینطور می‌توان از آن‌ها برای کاربردهای برخط استفاده کرد. در ادامه روش‌های استخراج ویژگی طیفـی

[1]Gaussian Mixture Model

زمان کوتاه مورد بررسی قرار گرفته است.

۲-۴ روش استخراج ویژگی طیفی زمان کوتاه

۲-۴-۱ پیش پردازش

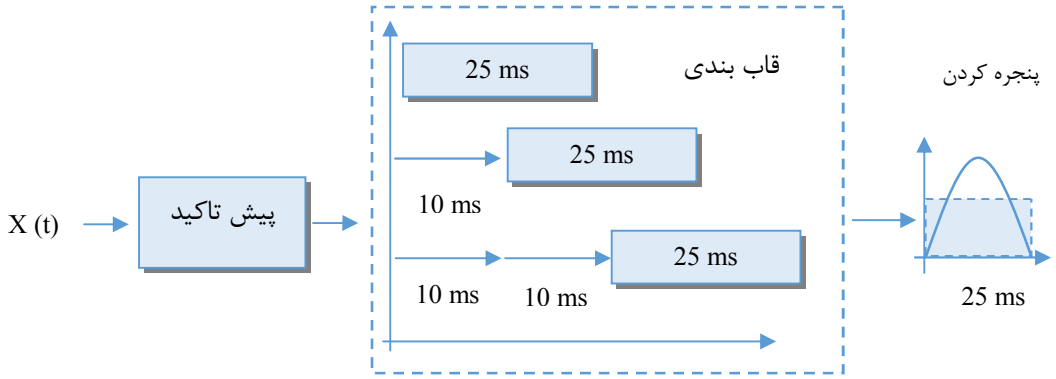
همانگونه که در بخش قبل گفته شد، در روش استخراج ویژگی طیفی زمان کوتاه سیگنال گفتار به قسمت‌های کوچک تقسیم می‌شود تا بتوان ویژگی‌های مورد نظر را از آنان استخراج کرد.

- ✓ **پیش تاکید**

طرح کلی این روش در شکل (۲-۲) نمایش داده شده است. مرحله اول عبارت است از یک پیش پردازش که عبارت است از اعمال یک فیلتر بالاگذر مرتبه اول به سیگنال. این فیلتر به صورت معمول به صورت معادله (۲-۱) می‌باشد[1].

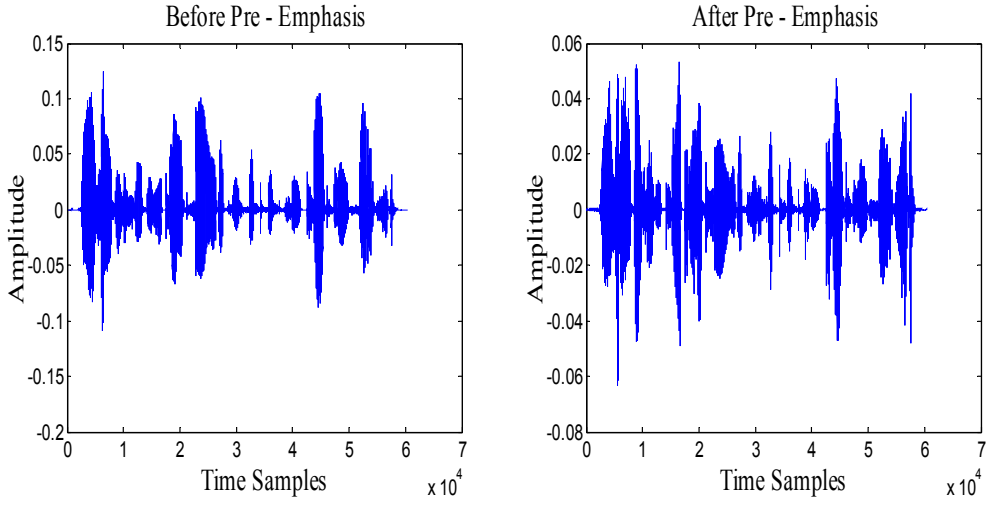
$$y(t) = x(t) - 0.97x(t-1) \qquad (۲-۱)$$

که در آن x(t) عبارت است از سیگنال فیلتر نشده و y(t) عبارت است از سیگنالی که فیلتر به آن اعمال شده است.



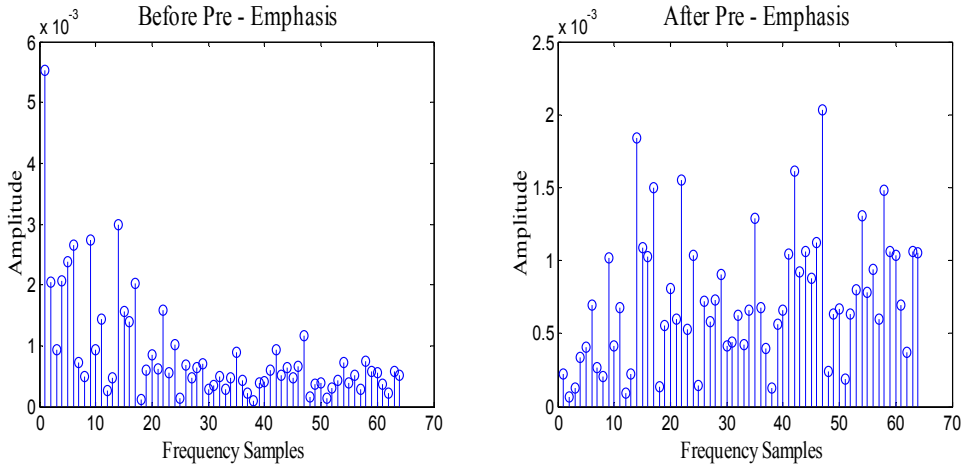
شکل(۲-۲): دیاگرام بلوکی پیش‌پردازش سیگنال گفتار[1]

سیستم تولید گفتار در انسان به طور طبیعی به صورتی می‌باشد که فرکانس‌های بالای سیگنال گفتار را تضعیف می‌کند. اعمال فیلتر معادله (۲-۱) به محتویات فرکانس بالای سیگنال گفتار اجازه می‌دهد که در مقابل محتویات مربوط به فرکانس‌های پایین‌تر خود را نشان دهند و اثر فرکانس‌های بالا در ویژگی‌های استخراج شده کم نشود. برای مثال در شکل (۲-۳) یک سیگنال گفتار قبل و بعد از پس پردازش نمایش داده شده است. شکل سمت چپ مربوط به قبل از اعمال فیلتر و شکل سمت راست مربوط به بعد از فیلتر کردن می‌باشد.



شکل(۲-۳): نمودار سیگنال گفتار قبل و بعد از اعمال فیلتر بالاگذر(شکل سمت چپ مربوط به قبل از اعمال فیلتر و شکل سمت راست مربوط به بعد از اعمال فیلتر می‌باشد)

در ادامه تبدیل فوریه هر دو را محاسبه و نتیجه در شکل (۲-۴) نمایش داده شده است. همانطور که مشاهده می‌شود، در شکل سمت چپ که مربوط به قبل از اعمال فیلتر بالا گذر می‌باشد، توزیع محتویات فرکانسی به صورت نامتقارن می‌باشد در صورتی که در شکل سمت راست که مربوط به بعد از اعمال فیلتر می‌باشد، محتویات فرکانسی توزیع متقارن‌تری دارند.



شکل(۲-۴): نمایش تبدیل فوریه گسسته سیگنال گفتار (شکل سمت چپ مربوط به قبل از اعمال فیلتر و شکل سمت راست مربوط به بعد از اعمال فیلتر می‌باشد)

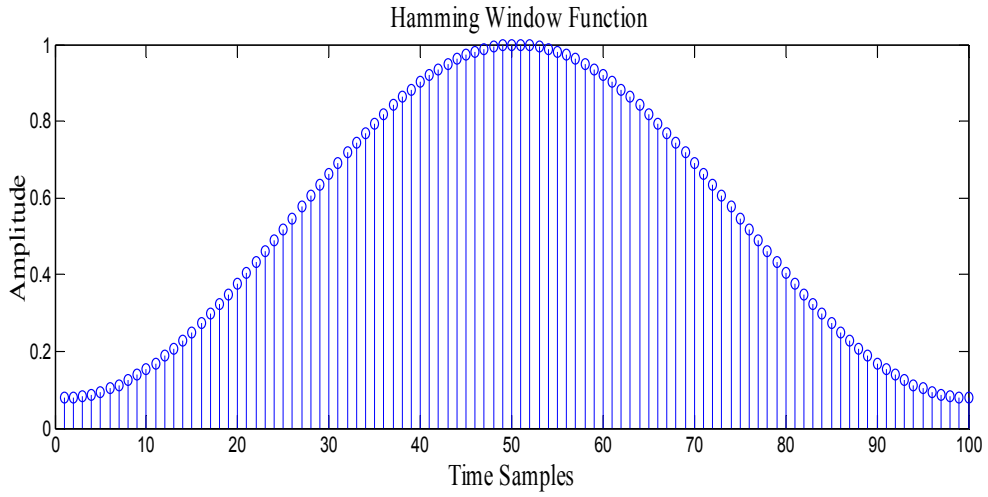
✓ قاب‌بندی

وقتی که فیلتر بالا گذر اعمال شد، در مرحله بعد باید سیگنال قاب[1] بندی شود. طول هر قاب را معمولا بین ۲۰ تا ۳۰ میلی ثانیه(به طور متوسط ۲۵ میلی ثانیه) در نظر می‌گیرند. معمولا برای اینکه انتقال از یک قاب به قاب بعدی به صورت آرام[2] انجام شود، قاب‌ها را به صورتی از سیگنال گفتار انتخاب می‌کنند که با هم همپوشانی داشته باشند. مثلا در شکل (۲-۲) هر دو قاب متوالی با یکدیگر ۱۵ میلی ثانیه همپوشانی دارند. پس با فرض اینکه قاب‌ها هر کدام ۲۵ میلی ثانیه طول دارند، هر ۱۰ میلی‌ثانیه یک بار قاب برداری انجام می‌شود[19] .

[1]Frame
[2]Smooth

✓ پنجره کردن

در مرحله پایانی پیش پردازش یک تابع پنجره به هر کدام از قاب‌های استخراج شده از سیگنال گفتار اعمال می‌کنند. این کار به این دلیل است که وقتی یک قاب را از سیگنال گفتار انتخاب می‌کنند عملا از یک تابع پنجره مستطیلی استفاده می‌شود، از طرفی بسیاری از روش‌های استخراج ویژگی طیفی زمان کوتاه بر اساس تبدیل فوریه عمل می‌کنند و استفاده از پنجره مستطیلی در حوزه زمان باعث اعوجاج در حوزه فرکانس می‌شود. برای غلبه بر این مشکل از توابع پنجره‌ای مانند همینگ[1] و یا هنینگ[2] استفاده می‌شود. این پنجره‌ها به آرامی دامنه دو سمت قاب را صفر می‌کنند و اثر یک‌باره صفر شدن تابع پنجره مستطیلی را تا حدودی رفع می‌کنند و باعث می‌شوند که درحوزه فرکانس اعوجاج کمتری ایجاد شود [22]. برای نمونه در شکل (۲-۵) پنجره همینگ ۱۰۰ نقطه‌ای نمایش داده شده است.



شکل(۲-۵): پنجره همینگ ۱۰۰ نقطه‌ای

با اعمال تابع پنجره به هر یک از قاب‌ها، مرحله پیش‌پردازش به اتمام می‌رسد و بسته به نوع روش استخراج ویژگی که انتخاب می‌شود، مراحل بعدی متفاوت است. در ادامه روش‌های مختلف استخراج

[1] Hamming
[2] Hanning

ویژگی طیفی زمان کوتاه معرفی شده است.

۲-۴-۲ دسته‌بندی روش‌های استخراج ویژگی طیفی زمان کوتاه

بعد از مرحله پیش پردازش نوبت استخراج ویژگی‌ها می‌رسد. روش‌های استخراج ویژگی طیفی زمان کوتاه را می‌توان به دو دسته کلی تقسیم کرد. دسته اول روش‌هایی هستند که بر اساس سیستم شنوایی انسان ویژگی‌های سیگنال گفتار را استخراج می‌کنند و دسته دوم روش‌هایی هستند که بر اساس مشخصات سیستم تولید صدای انسان، سیگنال گفتار را مدل می‌کنند[5]. در ادامه هر دو دسته و روش-های معروف و پر استفاده‌ای که بر اساس آن‌ها پیشنهاد شده، ارائه شده‌اند.

۲-۴-۳ روش‌های استخراج ویژگی طیفی زمان کوتاه بر مبنای شنوایی انسان

روش‌های استخراج ویژگی بر مبنای سیستم شنوایی انسان در چند قسمت کلی مشترک هستند. بعد از اینکه سیگنال از مرحله پیش‌پردازش عبور کرد، با استفاده از تبدیل فوریه گسسته[1](DFT) و یا تبدیل موجک گسسته[2] (DWT) سیگنال گفتار را به فضای دیگری انتقال داده و سپس از یک فیلتر بانک استفاده می‌شود[13]. فیلتر بانک‌ها شامل فیلترهای میان‌گذری هستند که بر اساس نوع روشی که استفاده می‌شود متفاوت می‌باشند و مقیاس‌بندی آن‌ها نیز بسته به اینکه از چه روشی استفاده می‌شود می‌تواند خطی باشد و یا اینکه از نوع مقیاسی که انتخاب شده است پیروی کند. وقتی که تبدیل فوریه و یا تبدیل موجک یک قاب محاسبه می‌شود، معمولا اطلاعات زیادی به دست می‌آید که همه این اطلاعات مورد نیاز نمی‌باشد و علاوه بر این حجم زیاد اطلاعات نیز باعث پیچیدگی محاسباتی زیادی می‌شود. برای مثال فرض کنید تبدیل فوریه یک قاب را محاسبه کرده و نتیجه این عمل ۲۵۶ عدد است. وقتی که از یک فیلتر بانک شامل ۳۰ فیلتر استفاده می‌شود، تمام اطلاعات این ۲۵۶ عدد در ۳۰ عدد ذخیره می‌شود.

[1]Discrete Fourier Transform
[2]DiscreteWavelet Transform

به این ترتیب اطلاعات زیادی از دست داده نمی‌شود و در عین حال پیچیدگی محاسباتی به میزان زیادی کاهش می‌یابد. روش‌ها را بر حسب اینکه از DFT و یا DWT استفاده کنند و اینکه چه نوع فیلتر بانکی را انتخاب کنند دسته‌بندی و نام‌گذاری می‌کنند. در ادامه چند روش پیشنهاد شده در مقالات مختلف بررسی شده است.

۲-۴-۳-۱ روش استخراج ویژگی ضرایب کپسترال در مقیاس مِل (MFCC)

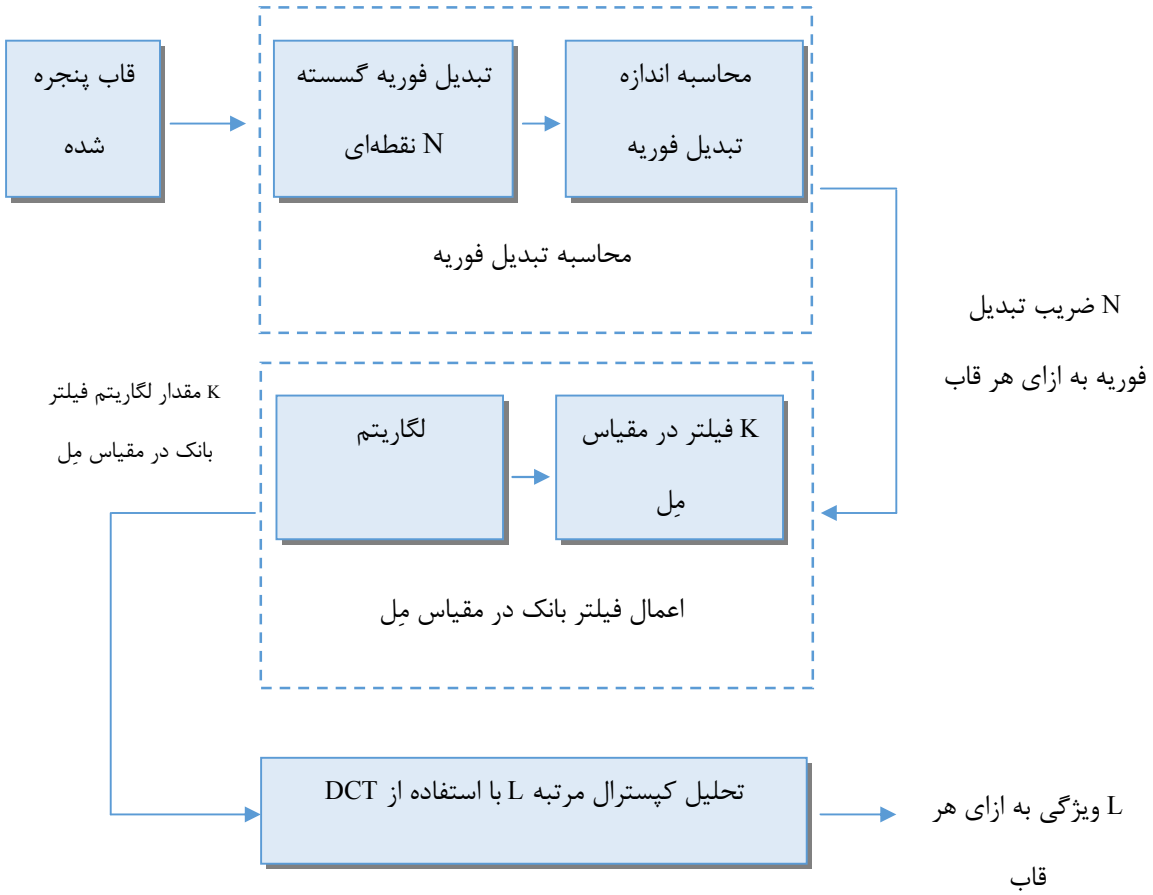
شناخته شده‌ترین روش این دسته که در بسیاری از مقالات به عنوان روش استاندارد استخراج ویژگی شناخته می‌شود روش ضرایب کپسترال در مقیاس مِل[1](MFCC) می‌باشد[6, 13, 19, 23, 24]. در روش MFCC از تبدیل فوریه گسسته استفاده می‌شود و فیلتر بانک مورد استفاده، فیلتر بانک مِل می‌باشد. اما روش به دست آوردن ویژگی‌های MFCC دقیقا چگونه است؟ در شکل (۲-۶) نمودار بلوکی این روش نمایش داده شده است[1].

هنگامی که در مرحله پیش پردازش سیگنال گفتار به قاب‌های پنجره شده تقسیم شدند، آنگاه تبدیل فوریه گسسته را برای هر قاب محاسبه می‌کنند. این کار با استفاده از الگوریتم‌های محاسبه سریع تبدیل فوریه گسسته[2](FFT) انجام می‌شود. برای مثال فرض شده که برای هر قاب، تبدیل فوریه گسسته N = 512 نقطه‌ای محاسبه شده است. در استخراج ویژگی برای شناسایی گوینده به اندازه تبدیل فوریه نیاز است و اطلاعات فاز در نظر گرفته نمی‌شود. پس اندازه هر ۵۱۲ عدد را محاسبه می‌کنیم. در ادامه به دلیل تقارن FFT نیمی از داده‌ها را نگه داشته و از نیمه دیگر صرفنظر می‌شود. پس اکنون به ازای یک قاب، ۲۵۶ عدد حقیقی به دست آمده است ولی هنوز هم این تعداد عدد زیاد است و نیازی به همه آن‌ها نیست. پس با استفاده از فیلتر بانک این تعداد داده را خلاصه‌تر کرده و در K = 30 عدد نمایش می‌دهند.

[1]Mel Frequency Cepstral Coefficient
[2]Fast Fourier Transform

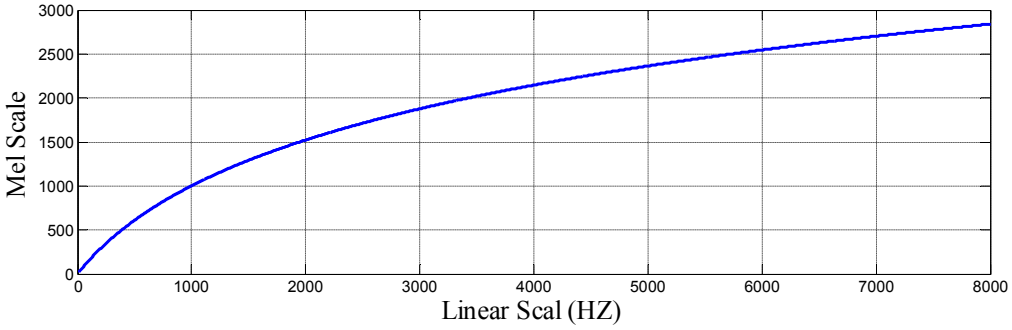
برای این کار از فیلتر بانک در مقیاس مِل استفاده می‌شود. برای تبدیل مقیاس‌بندی خطی به مقیاس مِـل از معادله (۲-۲) استفاده می‌شود[5].

$$F_{Mel} = 2595\, log_{10}\left(1 + \frac{F_{Linear}}{700}\right) \qquad (۲-۲)$$



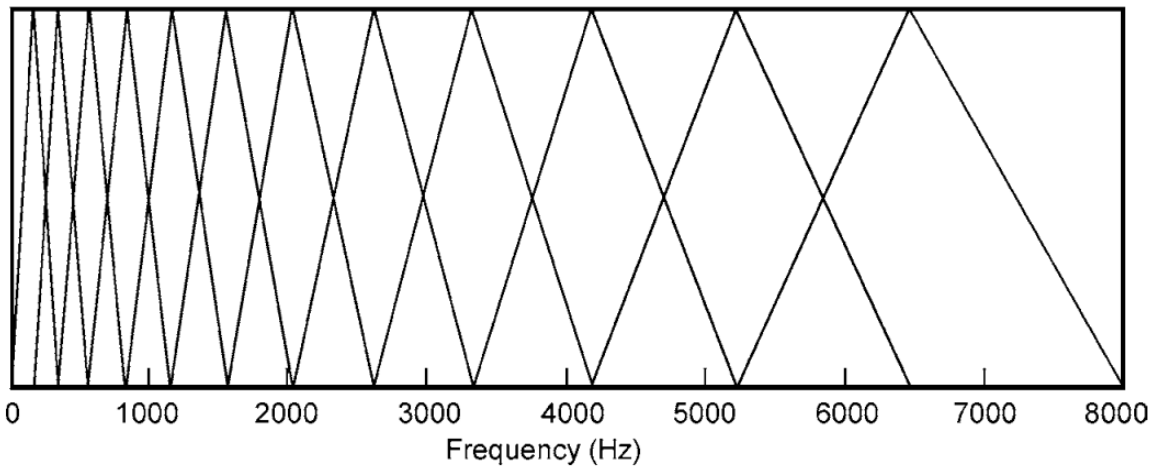
شکل(۲-۶): نمودار بلوکی استخراج ویژگی به روش MFCC [1]

همانطور که در شکل (۲-۷) مشاهده می‌شود، مقیاس مل بین ۰ تا ۱۰۰۰ هرتز، تقریبا به صورت خطی می‌باشد اما بعد از آن به صورت غیر خطی رفتار می‌کند.



شکل(۲-۷): نمودار مقیاس مِل بر مبنای مقیاس خطی

هر کدام از فیلترهای میان‌گذری که در فیلتر بانک استفاده می‌شود، یک فیلتر مثلثی می‌باشد و تمام فیلترها با همدیگر همپوشانی دارند. ابتدا باید مرکز هر فیلتر محاسبه شود. اگر فرض شود حداکثر فرکانس ۸۰۰۰ هرتز باشد(فرکانس نمونه‌برداری ۱۶۰۰۰ هرتز است)، با استفاده از معادله (۲-۲) اندازه حداکثر و حداقل فرکانس در حوزه مِل را بر اساس فرکانس خطی محاسبه می‌کنند. مثلا معادل فرکانس ۸۰۰۰ هرتز در حوزه مِل برابر با 2840 است . در ادامه محدوده بین کمترین و بیشترین فرکانسی در حوزه مل را به تعداد فیلتری که مورد نظر است تقسیم می‌کنند. مثلا در اینجا تعداد فیلترها۳۰ فرض شده است. با این کار مراکز فیلترها در حوزه مِل به دست می‌آید. سپس با استفاده از معکوس معادله (۲-۲) این مراکز را به حوزه فرکانس خطی هرتز انتقال می‌دهند. ابتدای هر فیلتر مثلثی از مرکز فیلتر قبل شروع شده و انتهای آن نیز در مرکز فیلتر بعد از خودش قرار می‌گیرد. یک نمونه از این فیلتر بانک در شکل (۲-۸) نمایش داده شده است.



شکل(۲-۸): یک نمونه فیلتر بانک در مقیاس مِل

در ادامه مقادیر اندازه FFT را در اندازه فیلتر مثلثی مربوط به هـر فرکـانس ضـرب کـرده و نتـایج حاصلضرب هر مثلث را با هم جمع می‌کنند و به این ترتیب ۲۵۶ عدد به ۳۰ عدد تبدیل می‌شوند. معمولا به این دلیل که سیستم شنوایی انسان از یک منحنی لگاریتمی پیروی می‌کند، برای نشان دادن این اثر از خروجی هر کدام از فیلترها لگاریتم می‌گیرند.

با استفاده از تحلیل کپسترال می‌توان باز هم داده‌ها را خلاصه‌تر کرد و به ساده‌تر شدن محاسبات و سریع‌تر شدن روش استخراج ویژگی کمک کرد. در اینجا برای به دست آوردن ضرایب کپسترال از تبـدیل کسینوسی گسسته[1] (DCT) استفاده می‌شود. معادله‌ای که برای این تبدیل استفاده می‌شود عبارت اسـت از[25] :

$$C_n = \sum_{k=1}^{K} (S_k) \cos\left(n\,(k-0.5)\left(\frac{\pi}{K}\right)\right) \quad , \; n = 1,2,3\ldots,L \qquad (۳-۲)$$

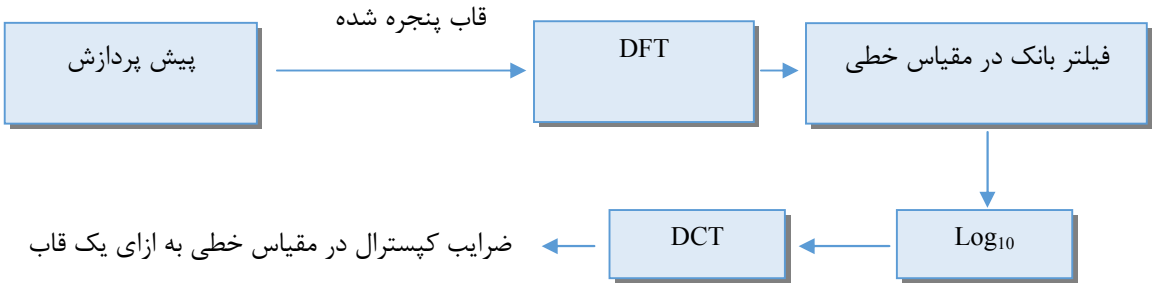
که در این فرمول S_k عبارت است از لگاریتم خروجی فیلتر k ام و C_n ضـریب کپسـترال n ام مـی-باشد. به طور معمول L = 12 فرض می‌شود و ضریب C_0 را بسته به کاربرد می‌توان استفاده کرد و یا از آن

[1]Discrete Cosine Transform

صرفنظر کرد.

۲-۳-۴-۲ روش استخراج ویژگی ضرایب کپسترال در مقیاس خطی[1](LFCC)

این روش همانند روش MFCC می‌باشد با این تفاوت که فیلتر بانکی که استفاده می‌شود دیگر بر مبنای مقیاس مِل نمی‌باشد بلکه در مقیاس خطی تقسیم‌بندی شده و طول تمام فیلترهای مثلثی با یکدیگر یکسان می‌باشد. دیاگرام بلوکی این روش در شکل (۲-۹) ارائه شده است[24].

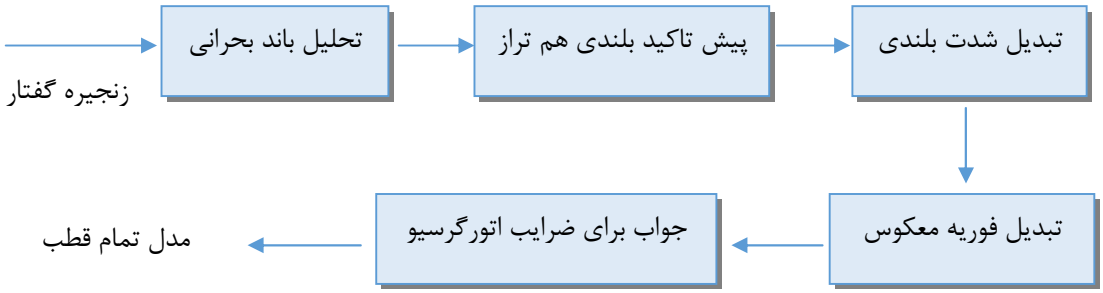


شکل(۲-۹): دیاگرام بلوکی استخراج ویژگی به روش LFCC

۲-۳-۴-۳ روش استخراج ویژگی ادراکی پیش‌گویی خطی[2](PLP)

روش PLP طیف شنیداری را به صورت یک مدل تمام قطب با مرتبه پایین مدل می‌کند. دیاگرام بلوکی این روش در شکل (۲-۱۰) نمایش داده شده است[15].

[1]Linear Frequency Cepstral Coefficients
[2]Perceptual Linear Prediction



شکل(۲-۱۰): دیاگرام بلوکی روش استخراج ویژگی PLP

در ادامه قسمت‌های مختلف این دیاگرام بلوکی شرح داده شده است.

✓ تحلیل باند بحرانی

در این قسمت ابتدا مرحله پیش‌پردازش، همانگونه که در قسمت (۲-۴-۱) شرح داده شده است، انجام می‌شود. سپس FFT قاب پنجره شده را محاسبه کرده و اندازه طیف را به دست می‌آورند. در مرحله بعد باید مقیاس خطی را به مقیاس بارک[1] انتقال داد. این کار را می‌توان با استفاده از معادله (۲-۴) انجام داد[5, 15].

$$\Omega(\omega) = 6\, ln\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right\} \qquad (۲-۴)$$

که در آن ω فرکانس زاویه‌ای با یکای رادیان بر ثانیه می‌باشد. در ادامه طیف توان انتقال یافته به مقیاس بارک را با تابع $\Psi(\Omega)$ کانولوشن می‌کنند. تابعی که برای کانولوشن استفاده می‌شود معمولا به صورت معادله (۲-۵) می‌باشد.

$$\Psi(\Omega) = \begin{cases} 0 & , \quad \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & , \quad -1.3 \le \Omega \le -0.5 \\ 1 & , \quad -0.5 < \Omega < 0.5 \\ 10^{-(\Omega-0.5)} & , \quad 0.5 < \Omega < 2.5 \\ 0 & , \quad \Omega > 2.5 \end{cases} \qquad (۲-۵)$$

[1]Bark Scale

- ✓ **پیش تاکید بلندی هم تراز**

در این مرحله نتیجه کانولوشن مرحله قبل را در تابع پیش تاکید$E(\omega)$ ضرب می‌کنند. این کار بـه منظور مدل کردن سیستم شنوایی انسان می‌باشد. تابع پیش تاکید با معادله (۲-۶) نمایش داده می‌شود.

$$E(\omega) = \frac{[(\omega^2 + 5.68 \times 10^6)\omega^4]}{[(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9) \times (\omega^6 + 9.58 \times 10^{26})]} \qquad (۲-۶)$$

- ✓ **تبدیل شدت بلندی و محاسبه ویژگی‌ها**

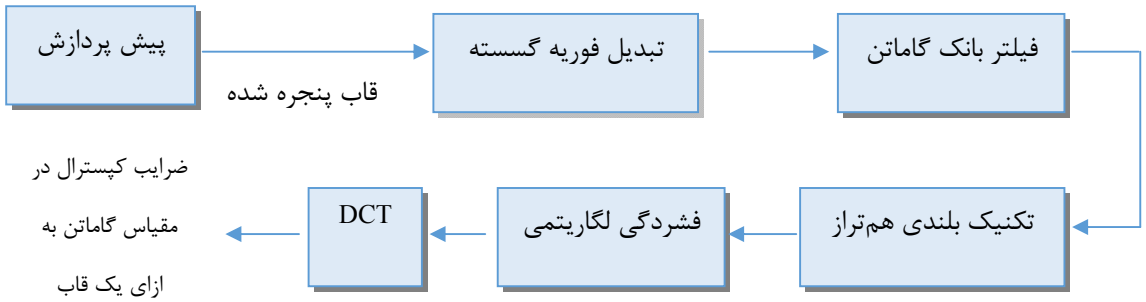
بعد از مرحله پیش تاکید، برای اینکه تبدیل شدت بلندی انجام شود، نتیجه ضرب مرحله قبل را به توان ۰/۳۳ می‌رسانند. این کار به شبیه سازی رابطه غیر خطی بین شدت صدا و بلندی صدای درک شده کمک می‌کند. در ادامه تبدیل فوریه معکوس گسسته برای مقادیر به دست آمده را محاسبه می‌کننـد و از M+1 ضریب اولیه خود همبستگی استفاده کرده و معادله Yule – Walker را بـرای ضـرایب اتورگرسـیو[1] یک مدل تمام قطب مرتبه M حل می‌کنند. ضرایب به دست آمده، ویژگی‌های به دست آمده به ازای یک قاب هستند[15].

۲-۴-۳-۴ روش استخراج ویژگی ضرایب کپسترال در مقیاس گاماتن[2](GFCC)

این روش ترکیبی از روش‌های MFCC و PLP می‌باشد. به این ترتیب که برای قسمتی که به جای فیلتر بانک در مقیاس مِل، از فیلتر بانک در مقیاس گاماتن استفاده می‌کنند. بعـد از اعمـال فیلتـر بانـک، مانند روش PLP از تکنیک بلندی هم‌تراز استفاده می‌کنند[27 ,26]. دیاگرام بلـوکی ایـن روش اسـتخراج

[1]Autoregressive
[2]Gammatone Frequency Cepstarl Coefficients

ویژگی در شکل (۲-۱۱) نمایش داده شده است.



شکل(۲-۱۱): دیاگرام بلوکی استخراج ویژگی به روش GFCC [26]

فیلتر بانک گاماتن عبارت است از یک سری فیلتر میان گذر که سیستم شنیداری انسان را مدل می‌کنند. تابع ضربه هر کدام از این فیلترها به صورت معادله (۲-۷) می‌باشد. در این معادله a یک ثابت است که معمولا برابر با ۱ در نظر گرفته می‌شود. φ عبارت است از جابه‌جایی فاز و n مرتبه فیلتر می‌باشد. در این معادله f_c و b به ترتیب فرکانس مرکزی و پهنای باند فیلتر در حوزه هرتز هستند[27].

$$g(t) = at^{n-1}e^{-2\pi t}\cos(2\pi f_c t + \varphi) \qquad (۲-۷)$$

۲-۴-۳-۵ روش‌های بر مبنای تبدیل موجک

در روش‌هایی که بر مبنای تبدیل موجک گسسته هستند، در طی فرایند استخراج ویژگی به جای بخشی که از تبدیل فوریه گسسته استفاده شده، از تبدیل موجک گسسته استفاده می‌شود. روش‌های این دسته نیز بر طبق اینکه از چه نوع فیلتری برای ساخت فیلتر بانک استفاده می‌شود، با همدیگر متفاوت هستند. برای این دسته می‌توان به روش‌هایی مانند روش Farooq & Datta و یا روش Sarikaya & Hansen اشاره کرد که هر دو به نام پیشنهاد دهندگانشان نامگذاری شده‌اند[24]. در مقالات، از روش‌های بر پایه تبدیل فوریه گسسته نتایج بهتری نسبت به روش‌های بر پایه تبدیل موجک گزارش شده است.

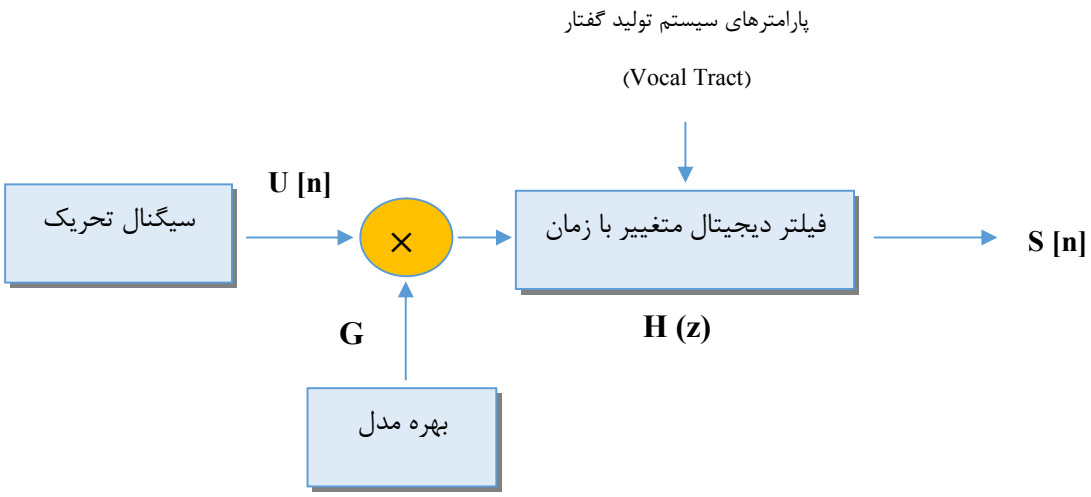
۲-۴-۴ روش‌های استخراج ویژگی بر مبنای سیستم تولید گفتار انسان

در روش‌های بر مبنای سیستم شنوایی انسان سعی می‌شود که سیستم شنوایی انسان را مدل کرده و پارامترهای آن را به دست آورد. در این بخش به بررسی دو روش کد کردن پیش‌بینی خطی[1](LPC) وفرکانس طیف خطی[2](LSF) پرداخته می‌شود.

۲-۴-۴-۱ روش استخراج ویژگی کد کردن پیش‌بینی خطی (LPC)

روش LPC بر این مبنا بنا شده است که می‌توان نمونه‌های سیگنال گفتار را با استفاده از یک ترکیب خطی از نمونه‌های قبل از آن‌ها تقریب زد. در این روش سیستم گفتار را به صورت یک مدل تمام قطب مدل می‌کنند و با استفاده از روش‌های مختلف، این پارامترها را تخمین می‌زنند[5]. فرض کنید مدلی که برای سیستم تولید گفتار انسان ارائه شده است به صورت شکل (۲-۱۲) باشد[25].

[1]Linear Predictive Coding
[2]Line Spectral Frequency



شکل(۲-۱۲): دیاگرام بلوکی ساده شده سیستم تولید گفتار انسان

برای این سیستم تابع تبدیل را به صورت معادله (۲-۸) تعریف می‌کنند.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \qquad (۸-۲)$$

نقطه قوت این مدل این است که می‌توان پارامترهای مدل، یعنی a_kها را با روش‌های سر راست و از نظر محاسباتی بهینه، از طریق تحلیل پیش‌بینی خطی به دست آورد. رابطه بین یک سیستم پیش‌بینی خطی و یک سیستم تولید گفتار در انسان در ادامه بررسی شده است.

برای سیستم شکل (۲-۱۲) می‌توان رابطه بین نمونه‌های تولید شده و سیگنال تحریک در حوزه زمان را به صورت معادله (۲-۹) نوشت[25].

$$s[n] = \sum_{k=1}^{P} a_k s[n-k] + Gu[n] \qquad (۹-۲)$$

یک سیستم پیش‌بینی خطی با ضرایب پیش‌بینی خطی α_k ، به صورت معادله (۲-۱۰) تعریف می-شود.

$$\tilde{s}[n] = \sum_{k=1}^{P} \alpha_k s[n-k] \qquad (۱۰-۲)$$

در ادامه برای یک سیستم پیش‌بینی خطی با مرتبه P ، خطای پیش‌بینی، یعنی e[n] به صورت معادله (۲-۱۱) تعریف می‌شود.

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{P} \alpha_k s[n-k] \qquad (۱۱-۲)$$

معادله (۲-۱۱) نشان می‌دهد که خطای پیش‌بینی را می‌توان به صورت خروجی یک سیستم با تابع تبدیل معادله (۲-۱۲) فرض کرد.

$$A(z) = 1 - \sum_{k=1}^{P} \alpha_k z^{-k} \qquad (۱۲-۲)$$

از مقایسه معادلات (۲-۸) و (۲-۱۲) می‌توان استنباط کرد که اگر سیگنال گفتار از مدل معادله (۲-۹) پیروی کند، و اگر فرض شود که $a_k = \alpha_k$، آنگاه می‌توان گفت $e[n] = Gu[n]$ است. بنابراین می‌توان گفت که تابع فیلتر خطای پیش‌بینی برابر است با عکس تابع معکوس سیستم تولید گفتار انسان. از همین خاصیت استفاده می‌شود و برای سیگنال گفتار با استفاده از روش پیش‌بینی خطی، پارامترهای مدل سیستم تولید گفتار را تخمین می‌زنند.

در سیستم پیش‌بینی خطی، هدف این است که یک سری α_k را یافت که با استفاده از آن‌ها بتوان خطای پیش‌بینی را به کمترین مقدر ممکن رساند[5]. برای انجام این کار می‌توان به سه روش اشاره کرد که این روش‌ها عبارتند از:

- روش حصیری[۱]
- روش کوواریانس[۱]

[1]Lattice Method

- روش خودهمبستگی[2]

از میان این روش‌ها، در کاربردهای پردازش گفتار به دلیل بهینه بودن محاسبات روش خودهمبستگی، معمولا از این روش استفاده می‌شود[5]. در این روش تابع خودهمبستگی را به صورت معادله (۲-۱۳) تعریف می‌کنند.

$$R_n(K) = \sum_{m=0}^{N-1-k} s_w(m)s_w(m+k) \qquad (۲-۱۳)$$

که در این معادله R تابع خودهمبستگی و s_w قاب پنجره شده است. برای به دست آوردن ضرایب α_k باید معادله (۲-۱۴) حل شود.

$$R_n(i) = \sum_{k=1}^{P} \alpha_k R_n(|i-k|) \qquad (۲-۱۴)$$

می‌توان با استفاده از روش Levinson – Durbin معادله (۲-۱۴) را حل کرده و پارامترهای سیستم را محاسبه کرد. پس از محاسبه پارمترهای مدل تولید گفتار، می‌توان با استفاده از معادلات (۲-۱۵) تا (۲-۱۷) ضرایب کپسترال را محاسبه کرد[28].

$$C_0 = \ln(G) \qquad (۲-۱۵)$$

$$C_m = \alpha_m + \sum_{k=1}^{m-1} \left(k/m\right) C_k a_{m-k} \quad , \quad 1 \le m \le P \qquad (۲-۱۶)$$

[1]Covariance Method
[2]Autocorrelation Method

$$C_m = \sum_{k=1}^{m-1} \left(k/m\right) C_k a_{m-k} \quad , \quad m > P \tag{۲-۱۷}$$

۲-۴-۴-۲ استخراج ویژگی به روش فرکانس طیف خطی (LSF)

روش استخراج ویژگی LSF یک روش فرعی است که از روش LPC مشتق شده است. فرض کنید که برای یک سیستم پیش‌بینی خطی، فیلتر خطای پیش‌بینی به صورت معادله (۲-۱۸) تعریف شده باشد.

$$A(z) = 1 - \sum_{k=1}^{P} a(k) A\left(z^{-k}\right) \tag{۲-۱۸}$$

که در این معادله $a(\mathrm{k})$ ها ضرایب پیش‌بینی خطی هستند. می‌توان دو معادله متقارن (۲-۱۹) و نامتقارن (۲-۲۰) را به صورت زیر تعریف کرد.

$$F_1(\mathrm{z}) = \mathrm{A(z)} + \mathrm{z}^{-(\mathrm{P}+1)} \mathrm{A}(\mathrm{z}^{-1}) \tag{۲-۱۹}$$

$$F_1(\mathrm{z}) = \mathrm{A(z)} - \mathrm{z}^{-(\mathrm{P}+1)} \mathrm{A}(\mathrm{z}^{-1}) \tag{۲-۲۰}$$

ریشه‌های چندجمله‌ای‌های معادلات (۲-۱۹) و (۲-۲۰) ویژگی‌های LSF هستند.

۲-۴-۵ پس پردازش[1]

بعد از اینکه ویژگی‌های مورد نیاز از سیگنال گفتار استخراج شد، معمولا یک سری عملیات بر روی آن‌ها به منظور بهبود کیفیت و یا کم کردن بار محاسباتی با فشرده کردن بردارهای ویژگی انجام می‌دهند. در ادامه به چند روش از روش‌های پرکاربرد اشاره شده است.

[1] Post Processing

۲-۴-۵-۱ نرمالیزه کردن میانگین کپسترال[1] (CMN)

وقتی که بردارهای ویژگی استخراج شدند و ضرایب کپسترال نیز برای آنها محاسبه شد، برای اینکه بتوان اثرات نویز موجود در پایگاه داده و یا میکروفن و البته اثرات اعوجاج مربوط به کانال انتقال را کمتر کرد، می‌توان از روش CMN استفاده کرد[1]. در این روش ابتدا ضرایب کپسترال را برای تمام قاب‌های سیگنال گفتار محاسبه و سپس میانگین تمام بردارهای ویژگی را بر روی هر یک از ابعاد بردار ویژگی محاسبه می‌کنند. این کار با استفاده از معادله (۲-۲۱) انجام می‌شود.

$$\vec{\mu}_T = \frac{1}{\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\overrightarrow{\mathrm{C_t}} \qquad (۲۱-۲)$$

در این معادله فرض شده است که سیگنال گفتار دارای T بردار ویژگی می‌باشد. بردارهای ویژگی به صورت زیر نمایش داده می‌شوند:

$$\overrightarrow{C_t} = [\,C_1, C_2, \ldots., C_L]^t \qquad (۲۲-۲)$$

در معادله (۲-۲۲) اندازه بردارهای ویژگی برابر با L می‌باشد.

حال که میانگین‌ها محاسبه شدند، بردارهای جبران‌سازی شده جدید را به صورت معادله(۲-۲۳) محاسبه می‌کنند.

$$\overrightarrow{C_t^{\,c}} = \overrightarrow{C_t} - \vec{\mu}_T \qquad (۲۳-۲)$$

این کار را برای تمام بردارهای ویژگی انجام می‌دهند و به این ترتیب ویژگی‌های جبران سازی شده CMN به دست می‌آیند. البته می‌توان به جای روش CMN از روش‌هایی مانند نرمالیزه کردن با استفاده از واریانس و میانه با هم نیز استفاده کرد. روش CMN محاسبات کمتری دارد و معمول‌تر می‌باشد[29].

[1] Cepstral Mean Normalization

۲-۴-۵-۲ مشتقات زمانی[1]

در پردازش گفتار یک نکته مهم این است که بتوان ارتباط بین بردارهای ویژگی و البته پویا بودن سیگنال گفتار را در ویژگی‌های به دست آمده اعمال کرد. برای این کار معمولا از مشتقات زمانی مرتبه اول، دوم و یا سوم استفاده می‌شود. مشتقات زمانی کمک می‌کنند که ارتباط بین بردارهای ویژگی و پویایی سیگنال گفتار در ویژگی‌های استخراج شده لحاظ شوند. گرچه می‌توان از مشتقات مرتبه سوم نیز استفاده کرد اما به دلیل ملاحظات محاسباتی و جلوگیری از بزرگ شدن بیش از اندازه بردار ویژگی معمولا از مشتقات مرتبه اول و دوم استفاده می‌شود. برای به دست آوردن مشتقات مرتبه اول از ضرایب جبران سازی شده CMN استفاده می‌شود تا بتوان علاوه بر نمایش پویایی سیگنال گفتار، مقاومت در مقابل نویز را نیز به ویژگی‌ها افزود. مشتقات مرتبه اول که با نام" ضرایب دلتا" نیز شناخته می‌شوند، از معادله (۲-۲۴) به دست می‌آیند[1, 23].

$$\overrightarrow{d_t} = \frac{\sum_{p=1}^{P} p\left(\overrightarrow{C^c}_{t+p} - \overrightarrow{C^c}_{t-p}\right)}{2\sum_{p=1}^{P} p^2} \qquad (۲۴-۲)$$

در این معادله معمولا p=2 فرض می‌شود. $\vec{C}^c{}_t$ عبارت است از ضرایب جبران سازی شده به روش CMN و $\vec{d}_t$ نیز ضرایب دلتا می‌باشند. برای محاسبه مشتق مرتبه دوم که با نام" ضرایب دلتا – دلتا" و یا "ضرایب شتاب[2]" شناخته می‌شوند، از معادله (۲-۲۵) استفاده می‌شود.

$$\overrightarrow{a_t} = \frac{\sum_{p=1}^{P} p\left(\overrightarrow{d^c}_{t+p} - \overrightarrow{d^c}_{t-p}\right)}{2\sum_{p=1}^{P} p^2} \qquad (۲۵-۲)$$

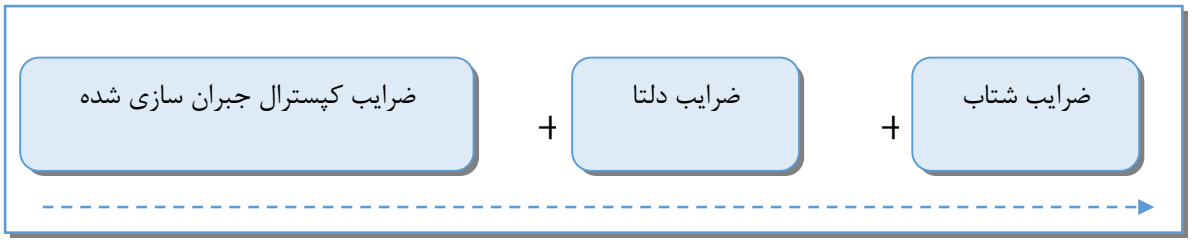
همانگونه که ملاحظه می‌شود، رابطه (۲-۲۵) همانند رابطه (۲-۲۴) می‌باشد که برای به دست آوردن ضرایب شتاب، یعنی $\overrightarrow{a_t}$ ، از ضرایب دلتا استفاده شده است. پارامتر P نیز همانند معادله (۲-۲۴) معمولا برابر ۲ در نظر گرفته می‌شود[1].

[1] Temporal Derivatives
[2] Acceleration

وقتی که ضرایب CMN و با استفاده از آن‌ها ضرایب دلتا و شتاب نیز محاسبه شدند، می‌توان با کنار هم قرار دادن این ویژگی‌ها در یک بردار، یک بردار ویژگی نهایی شامل هر سه دسته ضریب ایجاد کرد. شکل (۲-۱۳) این مطلب را نمایش می‌دهد.

۲-۴-۵-۳ روش فیلتر کردن نسبی طیف[1] (RASTA)

روش RASTA اولین بار توسط Hermansky برای بهبود ویژگی‌های استخراج شده از روش PLP پیشنهاد شد[4]. گرچه در ادامه برای سایر روش‌های استخراج ویژگی مانند MFCC و یا روش‌های بر مبنای تحلیل کپسترال نیز به کار رفت. روش RASTA درصد شناسایی گوینده را در حضور نویز حاصل از انتقال افزایش می‌دهد. این روش با اعمال فیلتر میان‌گذر بر سیگنال، طیف‌هایی از سیگنال که نرخ تغییرات آنها به نسبت نرخ تغییرات سیگنال گفتار کندتر و یا تندتر می‌باشد را حذف می‌کند[30].



شکل(۲-۱۳): ترکیب ضرایب CMN، دلتا و شتاب با یکدیگر

۲-۴-۶ جمع بندی

در این قسمت ابتدا روش‌های استخراج ویژگی دسته‌بندی شده و سپس با انتخاب روش استخراج ویژگی طیفی زمان کوتاه به عنوان یکی از پرکاربردترین روش‌های استخراج ویژگی، به تشریح این دسته از روش‌های استخراج ویژگی پرداخته شد. انواع روش‌های استخراج ویژگی طیفی زمان کوتاه شرح داده شدند و در ادامه روش‌های بهبود این ویژگی‌ها نیز بیان شد.

[1]Relative Spectra Filtering

حال سوال این است که برای استخراج ویژگی کدام یک از روش‌ها مناسب‌تر است؟ هدف این پایان نامه این است که یک سیستم برخط شناسایی گوینده بـر روی پردازنـده DSP پیـاده سـازی شـود. پـس طبیعتا روش استخراج ویژگی که انتخاب می‌شود باید سرعت زیادی داشته و در عین سریع بودن، درصـد شناسایی بالایی داشته باشد. در ضمن باید از الگـوریتمی پیـروی کنـد کـه قابلیـت پیـاده‌سـازی بـر روی پردازنده DSP را داشته باشد. با توجه به این معیارها و البته مطالب ذکر شده در مقالات مختلف و روش-هایی که انتخاب شده است[4, 5, 17, 18, 23, 27, 29, 31-33]، تصمیم گرفته شد که از دو روش پرکاربرد MFCC و LPCC استفاده شود. هر دو روش دارای ضرایب کپسترال بوده و می‌توان از آنها برای محاسـب ضرایب CMN نیز استفاده کرد. در ضمن برای قسمت پس پردازش از روش‌های CMN، پارامترهای دلتا و همینطور پارامترهای شتاب استفاده خواهد شد.

۲-۵ مدل کردن گوینده

بعد از به دست آوردن بردارهای ویژگی ناشی از استخراج ویژگی، با استفاده از ویژگی‌های به دست آمده برای هر گوینده یک مدل می‌سازند. در شناسایی گوینده وابسته به متن، مدلی که استفاده می‌شود وابسته به سیگنال گفتار و شامل وابستگی‌های زمانی بین بردارهای ویژگی می‌باشد ولی در شناسایی گوینده مستقل از متن، اغلب به جای وابستگی‌های زمانی، توزیع ویژگی‌ها را مدل می‌کنند. در شناسایی گوینده وابسته به متن می‌توان نمونه‌های آزمایش و آموزش را به صورت زمانی تطبیق داد، زیرا فرض بر این است که هر دو نمونه آزمایش و آموزش دارای دنباله کلمات یکسانی هستند، اما در روش مستقل از متن به این دلیل که ممکن است دنباله کلماتی که در مرحله آموزش به سیستم داده شده است با دنباله کلمات مرحله آزمایش متفاوت باشند، نمی‌توان از تطبیق زمانی استفاده کرد[2].

مدل‌های گوینده را می‌توان به مدل‌های بر مبنای الگو[۱] و مدل‌های آماری[۲] تقسیم کرد. در روش بر مبنای الگو، بردارهای ویژگی داده‌های آموزش و آزمایش را مستقیما با هم مقایسه می‌کنند و فرض می‌کنند که یکی از آن دو، نمونه ناقصی از دیگری می‌باشد. میزان اعوجاج[۳] بین آن دو، تعیین کننده میزان شباهت آن دو می‌باشد. هر چقدر که این اعوجاج کمتر باشد، نمونه‌های آموزش و آزمایش بیشتر به یکدیگر شباهت دارند. برای دسته بر مبنای الگو می‌توان به روش‌های چندی‌سازی برداری[۴](VQ) و پیچش زمانی پویا[۵](DTW) اشاره کرد که به ترتیب برای شناسایی گوینده مستقل از متن و وابسته به متن مورد استفاده قرار می‌گیرند[34].

در مدل‌های آماری، هر گوینده را بر اساس یک توزیع احتمالاتی مدل می‌کنند. مرحله آموزش در

[1]Template Models
[2]Stochastic Models
[3]Distortion
[4]Vector Quantization
[5]Dynamic Time Warping

این مدل‌ها، تخمین زدن پارامترهای توزیع احتمالاتی از روی داده‌های آموزش برای گوینده مورد نظر است. مرحله آزمایش معمولا به صورت محاسبه شباهت [1] بین سیگنال گفتار ورودی به سیستم و مدل گوینده مورد نظر می‌باشد. مدل آمیخته گاوسی[2](GMM) و مدل مخفی مارکوف[3](HMM) از معمول-ترین روش‌های شناسایی گوینده آماری هستند که به ترتیب برای حالت‌های مستقل از متن و وابسته به متن مورد استفاده قرار می‌گیرند[34 ,2].

می‌توان از دید دیگری نیز مدل‌ها را دسته بندی کرد. در این دسته‌بندی مدل‌ها را به دو حالت تولید کننده[4] و متمایز کننده[5] تقسیم می‌کنند. مدل‌های تولید کننده مانند GMM یا VQ توزیع ویژگی مربوط به هر گوینده را تخمین می‌زند، اما در روش‌های متمایز کننده مانند شبکه‌های عصبی مصنوعی[6] (ANN) و ماشین بردار پشتیبان[7] (SVM) مرز بین گوینده‌ها را مدل می‌کنند[35, 36].

۲-۵-۱ مدل کردن گوینده با استفاده از VQ

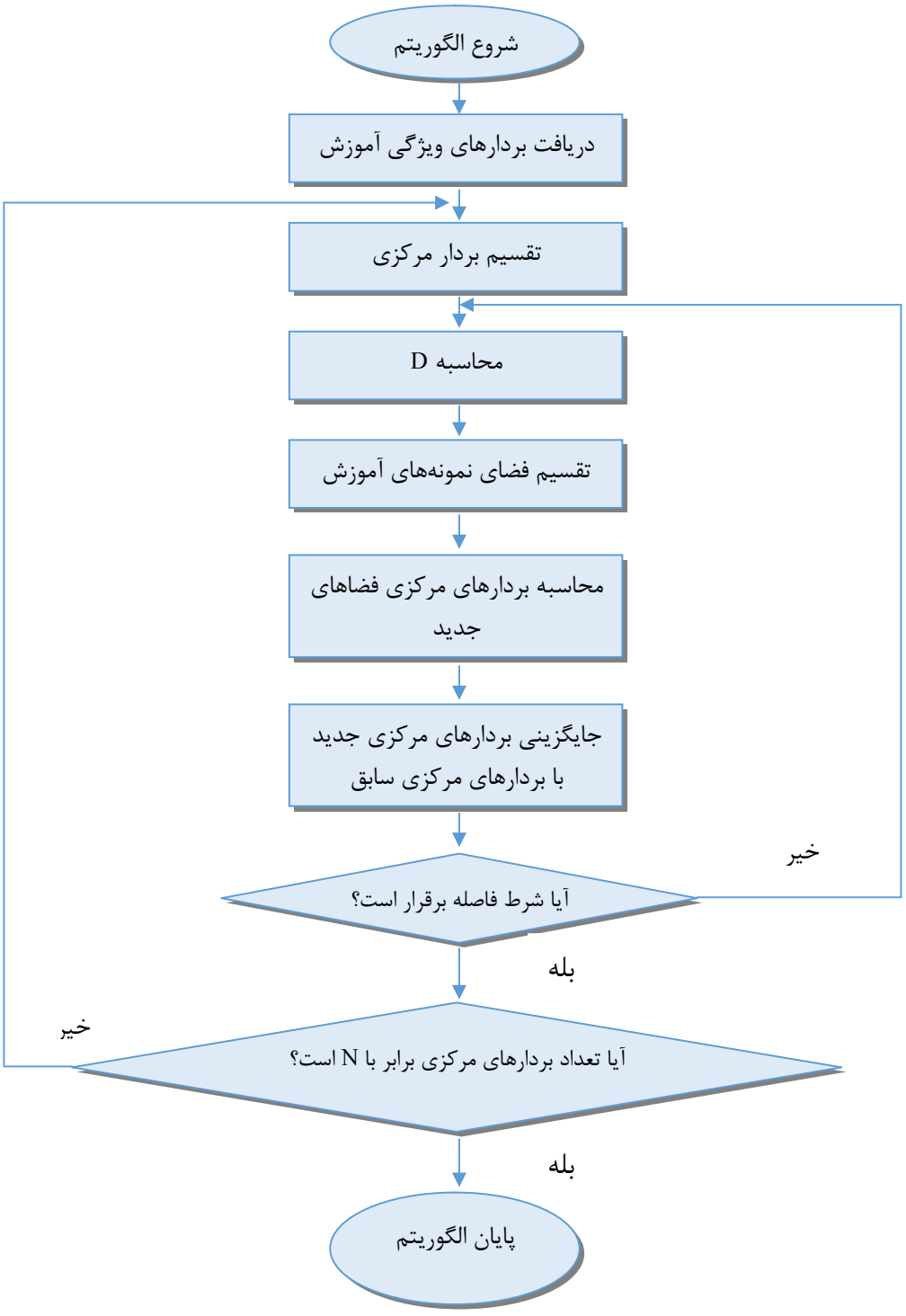
الگوریتم VQیک روش فشرده‌سازی با اتلاف می‌باشد که بر اساس کد کردن بلوکی عمل می‌کند. این روش یکی از روش‌های معمول برای نگاشت مقدار زیادی از بردارهای ویژگی به تعداد از پیش تعیین شده‌ای خوشه است که هر خوشه با یک بردار مرکزی[8] نمایش داده می‌شود. روش VQ کاربردهای گوناگونی مانند فشرده‌سازی صدا و تصویر، شناسایی گوینده و ... دارد. در این روش برای شناسایی گوینده، بردارهای ویژگی از سیگنال گفتار استخراج و به یک گروه کوچکتر از بردارهای اندازه‌گیری شده تبدیل

[1] Likelihood
[2] Gaussian Mixture Model
[3] Hidden Markov Model
[4] Generative
[5] Discriminative
[6] Artificial Neural Networks
[7] Support Vector Machine
[8] Centroid

می‌شوند که نمایانگر توزیع داده‌های آموزش هستند. هر کدام از این بردارها را با عنوان یک کلمه کد[1] و مجموع کلمات کد برای یک گوینده را با عنوان کتاب کد[2] می‌شناسند. برای آموزش یک سیستم شناسایی گوینده باید به ازای هر گوینده یک کتاب کد ساخته و در سیستم قرار داد. وقتی که یک سیگنال گفتار مربوط به یک گوینده ناشناس وارد سیستم می‌شود، با تمام کتاب کدهای موجود در سیستم مقایسه می‌شود و هر کدام از گویندگان که از این مقایسه امتیاز بیشتری کسب کند، سیگنال گفتار ورودی متعلق به آن گوینده می‌باشد. برای ساختن کتاب کد برای یک گوینده می‌توان از روش‌هایی مانند K-Means و یا الگوریتم LBG[3] استفاده کرد. در ادامه روش LBG به عنوان یک روش بهینه برای ساخت کتاب کد برای یک گوینده شرح داده شده است[9, 36-38].

روش کار الگوریتم LBG به این صورت است که یک بردار مرکزی را به عنوان یک کتاب کد اولیه تولید می‌کند. این مقدار اولیه معمولا میانگین بردارهای ویژگی مربوط به گوینده می‌باشد. در ادامه این مرکز را به دو مرکز تبدیل می‌کند و سپس دو مرکز را به چهار مرکز تبدیل می‌کند و این کار را تا جایی ادامه می‌دهد که یک کتاب کد با تعداد کلمه کد دلخواه را به دست آورد. روندنمای الگوریتم LBG برای محاسبه کتاب کد برای یک گوینده در شکل (۲-۱۴) نمایش داده شده است[9].

[1] Code Word
[2] Code Book
[3] Linde Buzo Gray



شکل(۲-۱۴): روندنمای الگوریتم LBG

در این روندنما:

- پارامتر D فاصله بردار مرکزی تا نمونه‌های آموزش است
- پارامتر D^1 فاصله نمونه‌های آموزش تا بردار مرکزی جدید می‌باشد
- فضای آموزش در هر مرحله به چند فضای جدید تقسیم می‌شود که تعداد فضاها برابر با تعداد بردارهای مرکزی است
- شرط فاصله عبارت است از : $\frac{D^1-\mathrm{D}}{\mathrm{D}} < TH$

وقتی که کتاب‌های کد ساخته شد، برای بررسی اینکه سیگنال گفتار ورودی به سیستم متعلق به کدام یک از گویندگان موجود در سیستم است، می‌توان از روش محاسبه فاصله اقلیدسی استفاده کرد. به این ترتیب که اگر فرض شود $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \overrightarrow{x_3}, \dots, \overrightarrow{x_T}\}$ بردارهای ویژگی سیگنال گفتار ورودی به سیستم باشند و $R = \{\overrightarrow{r_1}, \overrightarrow{r_2}, \overrightarrow{r_3}, \dots, \overrightarrow{r_K}\}$ کتاب کد برای گوینده مورد نظر باشد، آنگاه اعوجاج متوسط چندی‌سازی[1] را به صورت معادله (۲-۲۶) تعریف می‌کنند[2].

$$D_Q(X,R) = \frac{1}{T}\sum_{t=1}^{T} \min_{1 \le k \le K} d(x_t, r_k) \qquad (۲۶-۲)$$

که در آن $d(.,.)$ فاصله اقلیدسی بین x_t و r_k می‌باشد. این فاصله را به ازای سیگنال گفتار ورودی و تمام گویندگان موجود در سیستم محاسبه کرده و سیگنال گفتار متعلق به گوینده‌ای خواهد بود که کمترین مقدار D_Q را داشته باشد.

[1] Average Quantization Distortion

۲-۵-۲ مدل کردن گوینده با استفاده از مدل مخفی مارکوف

مدل مخفی مارکوف یک مدل احتمالاتی است که از آن می‌توان برای توصیف پدیده‌های طبیعی که ماهیت تصادفی دارند (مانند سیگنال گفتار) استفاده کرد. هر مدل HMM گسسته با استفاده از پارامترهای زیر تعریف می‌شود[39]:

- پارمتر N، عبارت است از تعداد حالت‌ها در مدل. گرچه حالت‌ها مخفی هستند اما معمولا در کاربردهای عملی ویژگی‌های فیزیکی‌ای وجود دارند که می‌توان از روی آنها حالت‌ها را شناسایی کرد. حالت‌های مجزا را با $S = \{s_1 , s_2 , \dots , s_N\}$ نمایش می‌دهند.
- پارامتر M که عبارت است از تعداد مشاهدات نمادهای مجزا به ازای هر حالت. نمادها، مشاهدات مربوط به خروجی سیستمی هستند که مدل می‌شود. نمادهای مجزا را با استفاده از $V = \{v_1, v_2, \dots, v_m\}$ نمایش می‌دهند.
- توزیع احتمال تغییر حالت‌ها، $A = \{a_{ij}\}$ ، که آن را به صورت معادله (۲-۲۷) نمایش می‌دهند.

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i) \quad , \quad 1 \le i, j \le N \tag{۲-۲۷}$$

در این معادله q معین کننده حالتی است که سیستم در زمان t در آن قرار دارد.

- توزیع احتمال نمادهای مشاهده شده در حالت j که به صورت $B = \{b_{ij}\}$ نمایش داده می‌شود و به صورت معادله (۲-۲۸) می‌باشد.

$$b_j(k) = P(V_k \ at \ t | q_t = s_i) \quad , \quad \begin{cases} 1 \le j \le N \\ 1 \le k \le M \end{cases} \tag{۲-۲۸}$$

- توزیع حالت اولیه، $\pi = \{\pi_i\}$ که به صورت معادله (۲-۲۹) تعریف می‌شود.

$$\pi_i = P(q_1 = s_1) \quad , \quad 1 \leq i \leq N \qquad (۲۹-۲)$$

به طورخلاصه یک HMM را به صورت $\lambda = (A, B, \pi)$ نمایش می‌دهند. وقتی که برای یک سیگنال گفتار مدل HMM می‌سازند، معمول است که توزیع احتمال مشاهده نمادها در هر حالت یعنی $b_j(k)$ را به صورت یک توزیع آمیخته گاوسی در نظر بگیرند[40]. با این فرض مدل به یک HMM پیوسته تبدیل می‌شود. پیوسته بودن مدل به این معنی است که احتمال مشاهده نمادها در حالت می‌تواند هر مقداری بین صفر تا یک داشته باشد، در صورتی که در حالت گسسته فقط می‌تواند چند مقدار خاص داشته باشد. استفاده از توزیع آمیخته گاوسی کمک زیادی به بالا بردن دقت مدل می‌کند اما از طرفی باعث افزایش بار محاسباتی و طولانی‌تر شدن زمان آموزش مدل می‌شود.

مدل HMM استفاده شده برای مدل کردن یک گوینده می‌تواند ارگودیک[1] و یا چپ به راست باشد. در مدل ارگودیک وقتی سیستم در یک حالت قرار دارد، می‌تواند در لحظه بعد به هر کدام از حالت‌های مدل انتقال پیدا کند. اما در مدل چپ به راست سیستم در هر انتقال حالت فقط می‌تواند به یک حالت بعد از خودش برود و یا اینکه در همان حالت باقی بماند. سیستم چپ به راست به دلیل ماهیت سیگنال گفتار، به صورت بهتری می‌تواند سیگنال گفتار را مدل کند[39].

برای استفاده از مدل HMM در سیستم شناسایی گوینده به این ترتیب عمل می‌شود که از بردارهای ویژگی استخراج شده از نمونه‌ها استفاده کرده تا پارامترهای مدل را تخمین بزنند. تخمین پارامترها را معمولا با استفاده از روش Baum – Welch انجام می‌دهند. البته روش‌های دیگری نیز مانند روش گرادیان برای تخمین پارامترها پیشنهاد شده است.

به ازای هر گوینده که قرار است در سیستم باشد باید یک مدل HMM آموزش داد. هنگامی که

[1] Ergodic

یک سیگنال گفتار متعلق به یک گوینده ناشناس وارد سیستم می‌شود، با استفاده از الگوریتم ویتربی[1]، سیگنال گفتار را با مدل‌های موجود در سیستم مقایسه کرده و هر مدلی که بیشترین امتیاز را کسب کند، سیگنال گفتار ورودی به سیستم متعلق به گوینده متناظر با آن مدل است[7, 41].

۲-۵-۳ مدل کردن گوینده با استفاده از مدل آمیخته گاوسی

مدل آمیخته گاوسی یک مدل احتمالاتی است که شامل ترکیب خطی تعداد محدودی توزیع گاوسی چند متغییره می‌باشد. مدل GMM به نوعی یک صورت پیشرفته‌تر از روش VQ می‌باشد که در این حالت، خوشه‌ها با هم هم‌پوشانی دارند و هر بردار ویژگی تنها به یکی از خوشه‌ها تعلق ندارد بلکه با درصدی احتمال به هر کدام از خوشه‌ها تعلق دارد[2, 42]. مدل آمیخته گاوسی فرض می‌کند که بردارهای ویژگی از یک توزیع گاوسی، که با یک میانه و انحراف معیار مشخص می‌شود، تبعیت می‌کنند. می‌توان با به دست آوردن میانگین‌ها و انحراف معیارهای ترکیب‌های گاوسی تشکیل دهنده GMM ، صدای یک گوینده را مدل کرد. در حقیقت روش مدل کردن گوینده با استفاده از GMM جزء یکی از اولین روش‌هایی بود که برای شناسایی گوینده پیشنهاد شد[43, 44]. این روش با گذر زمان کاملتر شده و با استفاده از روش‌های محاسباتی قدرتمندی که برای محاسبه پارامترهای یک مدل آمیخته گوسی وجود دارد، به مرور زمان به یکی از کارآمدترین روش‌های مدل کردن صدای گوینده تبدیل شده است[42, 45].

همانگونه که در شکل (۲-۱۵) نمایش داده شده است، مدل آمیخته گاوسی برای گوینده j، یعنی λ_j عبارت است از یک ترکیب خطی وزن‌دار از M توزیع گاوسی چند متغییره، با فرض اینکه GMM ما دارای M ترکیب می‌باشد. احتمال اینکه بردار ویژگی $\vec{x}_t$ متعلق به یک گوینده با مدل آمیخته گاوسی λ_j باشد برابر است با[46]:

[1] Viterbi

$$P(\vec{x}_t|\lambda_j) = \sum_{i=1}^{M} g_i N(\vec{x}_t|\overrightarrow{\mu_j}, \sum_j) \quad (۳۰-۲)$$

در این معادله g_i ها عبارتند از ضرایب ترکیب خطی که باید برای آن‌ها شرط (۲-۳۱) برقرار باشد. این شرط برای این است که معادله (۲-۳۰) به صورت یک تابع احتمالاتی باقی بماند.

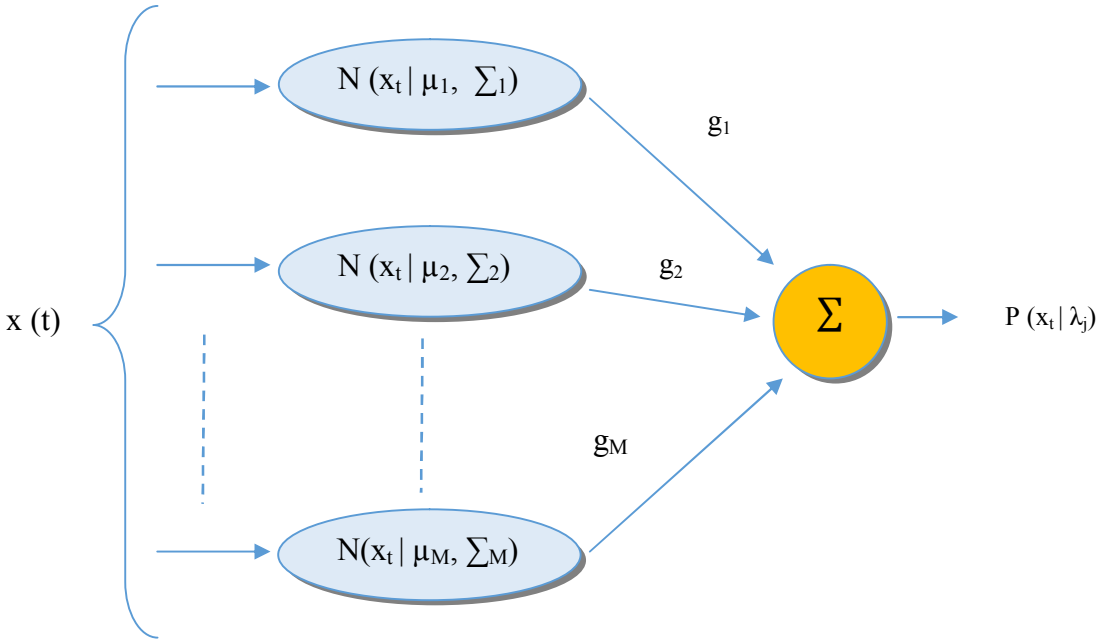
$$\sum_{i=1}^{M} g_i = 1 \quad (۳۱-۲)$$

در معادله (۲-۳۰)، $N(\vec{x}_t|\overrightarrow{\mu_j}, \sum_j)$ یک توزیع نرمال گاوسی D متغییره می‌باشد که به صورت (۲-۳۲) تعریف می‌شود.

$$N(\vec{x}_t|\overrightarrow{\mu_j}, \sum_j) = \frac{1}{(2\pi)^{D/2}|\sum_j|^{1/2}} e^{-0.5(\vec{x}_t - \overrightarrow{\mu_j})' \sum_j^{-1} (\vec{x}_t - \overrightarrow{\mu_j})} \quad (۳۲-۲)$$

در این معادله $\overrightarrow{\mu_j}$ عبارت است از میانگین ویژگی‌ها و $\sum_j$ ماتریس کواریانس می‌باشد. عناصر غیر از قطر اصلی این ماتریس کواریانس می‌توانند صفر و یا عددی غیر از صفر باشند. بسته به استفاده می‌توان از اطلاعات این اعضای ماتریس استفاده کرد، اما معمولا به دلایل محاسباتی از آن‌ها استفاده نمی‌کنند و ماتریس را به صورت قطری در نظر می‌گیرند. در برخی از مراجع اعلام شده است که استفاده از ماتریس کواریانس قطری، بر خلاف اینکه ممکن است تصور شود بر دقت مدل تاثیر منفی می‌گذارد، نتایج بهتری به دست داده است[8, 17]. در نهایت باید گفت که برای مدل کردن یک گوینده با استفاده از GMM نیاز است تا ماتریس کواریانس و میانگین به ازای هر کدام از M ترکیب تخمین زده شود و البته باید ضرایب ترکیب g_i را نیز برای GMM به دست آورد، یعنی برای گوینده j داریم:

$$\lambda_j = \{\overrightarrow{\mu_j}, \sum_j, g_i\}_j \quad (۳۳-۲)$$



شکل(۲-۱۵): نحوه عملکرد مدل آمیخته گاوسی

اکنون که روش مدل کردن یک گوینده به صورت کلی بیان شد، این سوال پیش می‌آید که چگونه باید پارامترهای صدای یک گوینده را برای یک مدل آمیخته گاوسی تخمین زد؟ برای تخمین زدن پارمترهای مدل GMM از الگوریتم بیشینه سازی امید ریاضی (EM)[1] استفاده می‌کنند[47, 48]. یکی از دلیل‌های محبوبیت روش GMM، الگوریتم EM می‌باشد زیرا الگوریتم EM با استفاده از مقدار نه چندان زیادی داده برای آموزش، پارامتراهای مدل میانگین، ماتریس کواریانس و ضرایب ترکیب را به خوبی و با دقت بالایی تخمین می‌زند.

برای استفاده از الگوریتم EM به این صورت عمل می‌شود که ابتدا یک مقدار اولیه به صورت تصادفی برای پارامترهای مدل آمیخته گاوسی، یعنی بردار میانگین، ماتریس کواریانس به ازای هر ترکیب، و ضرایب ترکیب خطی معین می‌کنند. البته این مقادیر اولیه شرطهایی را نیز باید برآورده کنند. مثلا برای ضرایب ترکیب باید شرط (۲-۳۱) برقرار باشد. برای میانگین‌ها نیز معمولا به صورت تصادفی این

[1]Expectation Maximization

مقداردهی اولیه انجام نمی‌شود بلکه با استفاده از الگوریتم‌های خوشه‌بند ابتدا بردارهای ویژگی را به تعداد ترکیب‌ها، در اینجا M، خوشه‌بندی می‌کنند و سپس مراکز هر خوشه را به عنوان میانگین یکی از ترکیب-ها به صورت مقدار اولیه به الگوریتم EM وارد می‌کنند. مرحله دوم بیشنه سازی شرطی می‌باشد. در این فاصله با استفاده از پارامترهای اولیه که به صورت تصادفی و خام مقداردهی شده‌اند، امید شرطی[1] را محاسبه کرده، در ادامه با استفاده از امید شرطی محاسبه شده، پارمترهای مدل GMM را تخمین می-زنند. این کار برای دوره تکرار اول می‌باشد. در دوره دوم با استفاده از پارامترهایی که در مرحله اول تخمین زده شده‌اند، دوباره امید شرطی را محاسبه می‌کنند و در ادامه با استفاده از امید شرطی محاسبه شده، دوباره پارامترهای مدل GMM را تخمین می‌زنند. این کار را به صورت تکراری انجام می‌دهند تا زمانی که پارمترهای مدل GMM با دقتی که از پیش معین شده است، تعیین شوند. الگوریتم EM تا حد زیادی به مقدار دهی اولیه وابسته می‌باشد و در صورت مقداردهی اولیه نامناسب، جواب‌های با دقت پایین به دست می‌دهد و یا اینکه واگرا می‌شود[49, 50]. برای اطلاعات بیشتر در مورد الگوریتم EM می‌توان به پیوست مراجعه کرد.

۲-۵-۳-۱ مدل پس زمینه جامع[2] (UBM)

هنگام طراحی یک سیستم تایید گوینده[3]، باید به ازای هر کدام از گویندگان موجود در سیستم یک مدل GMM ساخته شود. علاوه بر مدلی که برای هر گوینده ساخته می‌شود، یک مدل GMM هم با استفاده از داده‌های آموزشی تمام گویندگان به جز گوینده مورد نظر می‌سازند . به این مدل، مدل پس زمینه جامع یا UBM به ازای گوینده مورد نظر می‌گویند[45]. پس به ازای هر گوینده دو مدل به دست می‌آید. یک مدل، مدلی است که از داده‌های آموزشی خود گوینده ساخته شده است و مدل دوم مدلی

[1] Conditional Expectation
[2] Universal Background Model
[3] Speaker Verification

است که از داده‌های سایر گویندگان به جز گوینده مورد نظر ساخته شده است.

در مرحله بعد با استفاده از مدل‌های جداگانه مربوط به تمام گویندگان و مدل جامعی که ساخته شده اقدام به تایید هویت گویندگان می‌کنند. وقتی یک سیگنال وارد سیستم می‌شود، برای بررسی این موضوع که گوینده این سیگنال گفتار، عضوی از مجموعه گویندگان سیستم می‌باشد یا خیر، ابتدا ویژگی-های لازم از آن استخراج می‌شود و سپس این ویژگی‌ها را هم با مدل GMM گوینده و هم با مدل UBM که برای گوینده ساخته شده مقایسه می‌کنند و نبست این دو مقایسه را به دست می‌آورند. اگر اندازه این نسبت از آستانه‌ای که از قبل تعریف شده است بیشتر باشد، سیگنال ورودی مربوط به گوینده است در غیر این صورت سیگنال مربوط به گوینده نمی‌باشد[16, 51]. معمولا برای بهبود کیفیت سیستم‌های تایید گوینده، برای گویندگان مرد و زن به صورت جداگانه UBM آموزش می‌دهند[52].

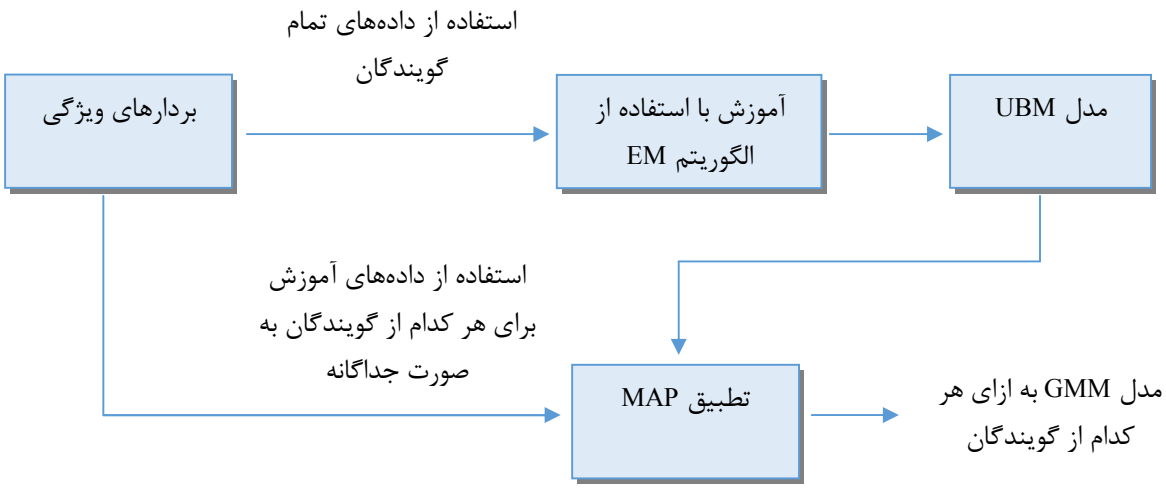
حال سوال این است که UBM در سیستم‌های شناسایی گوینده[1] چه کاربردی دارد؟ در سیستم-های با مجموعه گویندگان باز، یعنی سیستم‌هایی که با هرگوینده‌ای، خواه عضوی از پایگاه داده سیستم باشد و یا اینکه از خارج از مجموعه باشد، سروکار دارد، می‌توان از UBM برای بهبود کیفیت شناسایی سیستم استفاده کرد. به این ترتیب که وقتی یک نمونه سیگنال گفتار وارد سیستم شد، ابتدا هویت او را با استفاده از UBM تایید می‌کنند و سپس اگر هویت او تایید شد، آنگاه به شناسایی هویت او می‌پردازند. اما در سیستم‌های که نیازی به تایید گوینده نیست و فقط وظیفه سیستم، شناسایی گوینده می‌باشد، UBM چه استفاده‌ای می‌تواند داشته باشد؟

در کاربردهای شناسایی گوینده، تا وقتی که به ازای هر گوینده به اندازه کافی اطلاعات برای آموزش مدل‌ها داشته باشیم، نیازی به استفاده از UBM نمی‌باشد. اما وقتی که داده‌های آموزش برای گوینده‌ای به اندازه کافی وجود نداشته باشد و یا اینکه میزان داده‌های آموزش برای گویندگان در مجموعه

[1]Speaker Identification

با همدیگر توازن نداشته باشد و بعضی از گویندگان داده‌های آموزشی بیشتر و بعضی دیگر داده‌های آموزشی کمتر داشته باشند، می‌توان از UBM برای جبران این کمبودها استفاده کرد[55-53].

مدل‌های آماری مانند GMM را علاوه بر اینکه می‌توان با استفاده از الگوریتم‌هایی مانند EM آموزش داد و پارامترهای آنها را تخمین زد، می‌توان با میزان کمی داده آموزش، پارمترهای مدل GMM را با استفاده از روش‌هایی مانند بیشینه احتمال پسین[1](MAP) با داده‌های جدیدی تطبیق داد. پس یک روش برای آموزش دادن یک مدل GMM برای گوینده‌ای که دارای میزان کافی داده آموزشی نیست، عبارت است از اینکه یک مدل UBM به ازای تمام گویندگان موجود در مجموعه ساخته و سپس مدل UBM را با تمام مدل‌های GMM که برای هر کدام از گویندگان ساخته شده، تطبیق داد و مدل‌های جدیدی برای هر کدام از گویندگان ساخت[56]. نمودار تصویری این روش را می‌توان در شکل (۲-۱۶) مشاهده کرد.



شکل(۲-۱۶): دیاگرام بلوکی آموزش با استفاده از تطبیق MAP [56]

[1]Maximum A - Posteriori

۲-۵-۳-۲ کلاسه بندی با استفاده از GMM

در این روش برای کلاسه‌بندی گوینده‌ها از امتیازدهی به وسیله محاسبه بیشینه شباهت[1](ML) استفاده می‌شود. فرض کنید که یک سیگنال گفتار مربوط به یک گوینده ناشناس وارد سیستم می‌شود. برای تعیین هویت گوینده با استفاده از روش ML به این ترتیب عمل می‌شود که ابتدا ویژگی‌های سیگنال را استخراج می‌کنند. برای اینکه در فرایند شناسایی میزان خطا به کمترین مقدار برسد، به ازای هر یک از مدل‌های گوینده و بردارهای ویژگی استخراج شده از سیگنال آزمایش، احتمال پسین حساب می‌شود و گوینده‌ای برنده است که بیشترین احتمال پسین را به ازای بردارهای ویژگی سیگنال آزمایش داشته باشد. این موضوع در معادله (۲-۳۴) نمایش داده شده است.

$$\hat{j} = argmax\left[P(\lambda_j|X)\right] \quad , \quad 1 \le j \le N \tag{۲-۳۳}$$

در این معادله $X = (x_1, x_2, \ldots, x_T)$ مجموعه بردارهای ویژگی است و $\hat{j}$ گوینده شناسایی شده می‌باشد. می‌توان معادله (۲-۳۳) را به صورت معادله (۲-۳۴) ، یعنی امتیاز دهی ML، خلاصه کرد.

$$\hat{j} = argmax\left[\sum_{t=1}^{T} \log P(x_t|\,\lambda_j)\right] \quad , \quad 1 \le j \le N \tag{۲-۳۴}$$

مقدار $P(x_t|\,\lambda_j)$ را می‌توان با استفاده از رابطه (۲-۳۰) محاسبه کرد.

۲-۵-۴ جمع بندی

در این قسمت روش‌های معروف و پر استفاده مدل کردن گوینده را بررسی شد. هدف از این مرور و بررسی، پیدا کردن روشی بود که بتوان آن را به صورت برخط اجرا کرد و در عین حال قابلیت پیاده سازی بر روی پردازنده DSP را نیز داشته باشد. به همین دلیل روش‌های کلاسه‌بند مانند شبکه‌های عصبی مصنوعی به دلیل محاسبات زیاد در این فصل بررسی نشد و یا روش کلاسه بند SVM برای کلاسه‌بندی

[1]Maximum Likelihood

حالتی که از GMM برای مدل کردن گوینده‌ها استفاده می‌شود مورد بررسی قرار نگرفت.

از میان روش‌هایی که معرفی شد، روش VQ به نسبت دیگران محاسبات کمتری نیاز دارد ولی به نسبت روش GMM دقت کمتری در شناسایی دارد. روش HMM نیز بیشتر برای کاربردهای وابسته به گفتار مورد استفاده قرار می‌گیرد و البته پیچیدگی محاسباتی زیادی دارد و آموزش مدل‌ها نیز برای گوینده‌ها کار راحتی نمی‌باشد و نیاز به داده‌های زیادی برای آموزش مدل‌ها است. ولی برخلاف هر دو روش قبل، روش GMM هم دارای دقت بالایی می‌باشد و هم اینکه می‌توان به راحتی پارامترهای مدل GMM را برای گویندگان تخمین زد[8, 9, 31, 41, 51, 57].

برای انتخاب بین دو روش GMM و GMM-UBM باید به این موضوع اشاره کرد که پایگاه داده‌ای که در دسترس بود، پایگاه‌داده TMIT است که این پایگاه داده برای هر گوینده ۱۰ فایل صوتی ارائه داده است و مجموع این فایل‌ها زمان حدود یک دقیقه دارند. زمانی از GMM-UBM استفاده می‌شود که داده-های آموزش برای گویندگان مختلف با هم برابر نباشند و تعدادی از گویندگان به میزان کافی داده آموزشی در اختیار نداشته باشند. در پایگاه داده TMIT مدت زمان داده‌های آموزش برای گویندگان تقریبا با هم برابر است و از این بابت نیازی نیست که از GMM-UBM استفاده کرد. از طرفی برای استفاده از روش GMM – UBM معمولا به میزان زیادی داده آموزشی نیاز است. منظور از زمان زیاد در اینجا حدود چند ساعت می‌باشد. در آزمایشاتی که در مقالات مختلف ارائه شده است، هنگام آموزش با استفاده از GMM-UBM از مدل‌های آمیخته گاوسی با تعداد ترکیب بین ۶۴ تا ۲۵۶ استفاده شده است که برای تخمین زدن پارامترهای تمام این ترکیب‌ها باید میزان زیادی داده آموزشی در اختیار داشت [2, 57, 58]. برای روشن‌تر شدن موضوع فرض کنید که گوینده با استفاده از یک GMM با ۱۰ ترکیب مدل شده است. پس برای مدل کردن گوینده باید ۱۰ بردار میانگین و ۱۰ ماتریس کواریانس و ۱ بردار ضرایب ترکیب به دست آورد. ولی وقتی تعداد ترکیب‌ها برابر با ۲۵۶ می‌باشد باید ۲۵۶ بردار میانگین و ۲۵۶

ماتریس کواریانس را تخمین زد. بسته به اینکه ما از بردار ویژگی با چه ابعادی استفاده کنیم، تعداد پارامترها با افزایش تعداد ترکیب‌ها به صورت نمایی افزایش می‌یابد.

علاوه بر این هر چه تعداد ترکیب‌ها افزایش پیدا کند، میزان زمان لازم برای آزمایش نیز افزایش پیدا کرده و قابلیت برخط بودن سیستم کمتر می‌شود و البته حافظه بیشتری نیز نیاز می‌باشد. با توجه به مواردی که ذکر شد، استفاده از روش GMM به نسبت روش GMM – UBM منطقی‌تر می‌باشد و در این پایان نامه برای پیاده سازی سخت افزاری از روش مدل کردن گوینده با استفاده از GMM استفاده شده است.

فصل سوم

معرفی سیستم سخت‌افزاری

برای ساخت سخت افزار شناسایی گوینده، از پردازنده سیگنال TMS320C5509A استفاده شده که ساخت شرکت Texas Instrument می‌باشد. قبل از اینکه مدار سخت افزاری را معرفی کنیم، ابتدا خود پردازنده مورد استفاده و اطلاعات مربوط به آن را ارائه می‌دهیم.

۳-۱ پردازشگرهای سیگنال

در دهه ۷۰ میلادی همزمان با ساخت اولین پردازنده‌ها توسط شرکت‌های مختلف، شرکت TI نیز پردازنده‌های مخصوص پردازش سیگنال را وارد بازار کرد. این پردازنده‌ها که بیشتر با نام DSP[1] معروف هستند، همگی با نام TMS320 شروع می‌شوند. پردازنده‌های DSP در مدت زمان حدود ۴۰ سال که از حضورشان در بازار قطعات الکترونیک می‌گذرد بسیار تکامل یافته‌اند. اولین سری این پردازنده‌ها با نام TMS320C10 وارد بازار شد. پس از چند سال، حضور سری TMS320C25 باعث معروف شدن DSPها گردید. این پردازنده می‌توانست تبدیل فوریه را با سرعتی محاسبه کند که اولین سری‌های پردازنده پنتیوم ساخت شرکت اینتل ۲۰ سال بعد توانستند به آن سرعت برسند[۵۸].

۳-۱-۱ پردازنده‌های مهم شرکت TI

شرکت TI دارای چندین دسته مختلف از پردازنده‌ها می‌باشد که برای کاربردهای گوناگون استفاده می‌شوند و هر کدام از دسته‌های این پردازنده‌ها ویژگی‌های منحصر به فرد و مخصوص به خود را دارند. در ادامه به معرفی سری‌های مختلف پردازنده‌های شرکت TI می‌پردازیم.

الف– سری 5000 (یا 5XXXX): این سری دارای دو خانواده اصلی 55XX و 54XX می‌باشد. این پردازنده‌ها کم مصرف‌ترین پردازنده‌های شرکت TI می‌باشند که در بسیاری از تجهیزاتی که نیاز به

[1]Digital Signal Processor

پردازش با سرعت بالا و مصرف توان کم دارند استفاده می‌شوند. در سری 5000 سرعت پردازنده‌ها بین ۱۰۰ تا ۳۰۰ مگا هرتز می‌باشد و به صورت خاص در سری 55XX قدرت محاسبات ریاضی می‌تواند دو برابر فرکانس کاری پردازنده باشد. یعنی سری 55XX حداکثر توان انجام ۶۰۰ میلیون ضرب را در یک ثانیه دارد[۵۸].

کاربرد اصلی پردازشگرهای سری 5000 در پردازش صوت و الگوریتم‌هایی است که نیاز به سرعت بالا دارند. از بعضی سری‌ها که حجم حافظه بیشتر از ۱۲۸ کیلو بایت دارند می‌توان برای کاربردهای پردازش تصویر نیز استفاده کرد.

ب– سری 2000 (2XXX): این سری شامل دو خانواده اصلی 24XX و 28XX می‌باشد. سری 28XX یک خانواده با عملکرد نزدیک به میکروکنترلرها می‌باشد. ویژگی مهمی که این دسته را از سایر دسته‌ها مجزا می‌سازی، وجود حافظه فلش در این سری می‌باشد. وجود حافظه فلش داخلی سبب شده است که این سری از پردازنده‌ها را بتوان راحت‌تر از سری‌های دیگر DSP ها برنامه‌ریزی کرد. در این خانواده حجم حافظه داخلی از نوع SRAM کمتر از ۳۲ کیلوبایت بوده و کاربرد این سری از پردازنده‌ها بیشتر به عنوان یک میکروکنترلر پرسرعت می‌باشد. این میکروکنترلرهای به خوبی جای خود را در صنعت باز کرده‌اند.[۵۸].

ج – سری 6000 (6XXX): این سری شامل سه خانواده اصلی 62XX، 64XX و 67XX است. در این خانواده‌ها فرکانس کاری پردازنده بین ۱۵۰ مگاهرتز تا ۱/۲ گیگاهرتز می‌باشد اما سرعت کاری واقعی آنها ۸ برابر فرکانس کاری آنها است. در این پردازنده‌ها در هر کلاک تا حداکثر ۸ دستور به شکل همزمان قابل اجرا می‌باشد و به همین دلیل می‌توانند تا حدود ۱۰ گیگا دستورالعمل در ثانیه[1](GIPS) را اجرا

[1]Giga Instruction Per Second

نمایند. این خانواده برای تمامی انواع پردازش‌های پرسرعت مناسب می‌باشند، اما سری 64xx بـا قابلیـت-های خاص آن مناسب‌ترین سری برای پردازش‌های تصویر می‌باشد. [۵۸].

۳-۱-۲ پردازنده TMS320C5509A

از بین خانواده‌های معرفی شده، پردازنده 5509A برای طراحی سخت افزاری انتخاب شده است که از سری خانواده 5000 بوده و حالت بهینه شده پردازنـده 5509 مـی‌باشـد. مـی‌تـوان از خصوصـیات ایـن پردازنده به موارد زیر اشاره کرد[59] :

- قابلیت سه بار خواندن و دوبار نوشتن در هر ثانیه
- سیستم محاسباتی ممیز ثابت
- دو واحد MAC با قابلیت ضرب دو عدد ۱۷ بیتی
- فرکانس کاری قابل تنظیم تا حدود ۲۰۰ مگاهرتز
- واحد EMIF جهت دسترسی به حافظه
- معماری پیشرفته چندگذرگاهی شامل یک گذرگاه برنامه، سه گذرگاه داده و چهار گذرگاه آدرس
- قابلیت اجرای موازی چند دستور در یک سیکل
- توان مصرفی پایین
- اجرای دستورالعمل‌های پیچیده پردازش سیگنال مانند کانولوشن، تبدیل فوریه و ....
- قابلیت محاسبه ۴۰۰ میلیون محاسبه ریاضی در ثانیه در فرکانس ۲۰۰ مگاهرتز
- پشتیبانی از پروتکل‌های ارتباطی McBSP و I2C برای ارتباط با انواع مبدل‌ها

از این پردازنده می‌توان در کاربردهای مختلف پردازش سیگنال مانند کاربردهای زیر استفاده کرد.

- انواع کاربردهای پردازش گفتار شامل شناسایی گوینده و ....
- انواع الگوریتم‌های رفع نویز
- مدولاسیون و دمدولاسیون
- فشرده سازی صوت

برای برنامه نویسی و شبیه سازی و انتقال برنامه به DSP از نرم افزار Code Composer Studio استفاده می‌شود و برای اتصال کامپیوتر به پردازنده و همچنین به منظور انتقال برنامه از JTAG[1] استفاده می‌شود. مدل استفاده شده در این پایان نامه Spectrum Digital XDS510 USB می‌باشد که در شکل (۳-۱) نمایش داده شده است.

شکل(۳-۱): JTAG جهت ارتباط کامپیوتر با پردازنده DSP

[1]Joint Test Action Group

۳-۲ مدار پردازشگر سیگنال

مدار طراحی شده دارای قسمت‌های مختلفی می‌باشد که قسمت‌های اصلی آن عبارتند از:

- منبع تغذیه
- مبدل داده‌ها (کُدِک)
- پردازنده سیگنال

در ادامه هر کدام از این قسمت‌ها را بررسی کرده و در مورد ویژگی‌ها و کارکرد آن‌ها صحبت خواهیم کرد.

۳-۲-۱ منبع تغذیه

برای تامین انرژی مورد نیاز بخش‌های مختلف مدار از یک آی سی توان با نام TPS767D301 استفاده شده است. یکی از مهمترین ویژگی‌هایی که این آی سی دارد این است که توسط شرکت سازنده پردازنده سیگنال، یعنی Texas Instrument طراحی و ساخته شده است و دقیقا با مشخصات پردازنده سیگنال همخوانی داشته و نیازهای آن را برآورده می‌کند[60]. ویژگی‌هایی که می‌توان برای این آی سی نام برد عبارتند از:

- دارای دو ولتاژ خروجی است که برای استفاده‌های متفاوت در مدار مزیت مهمی می‌باشد.

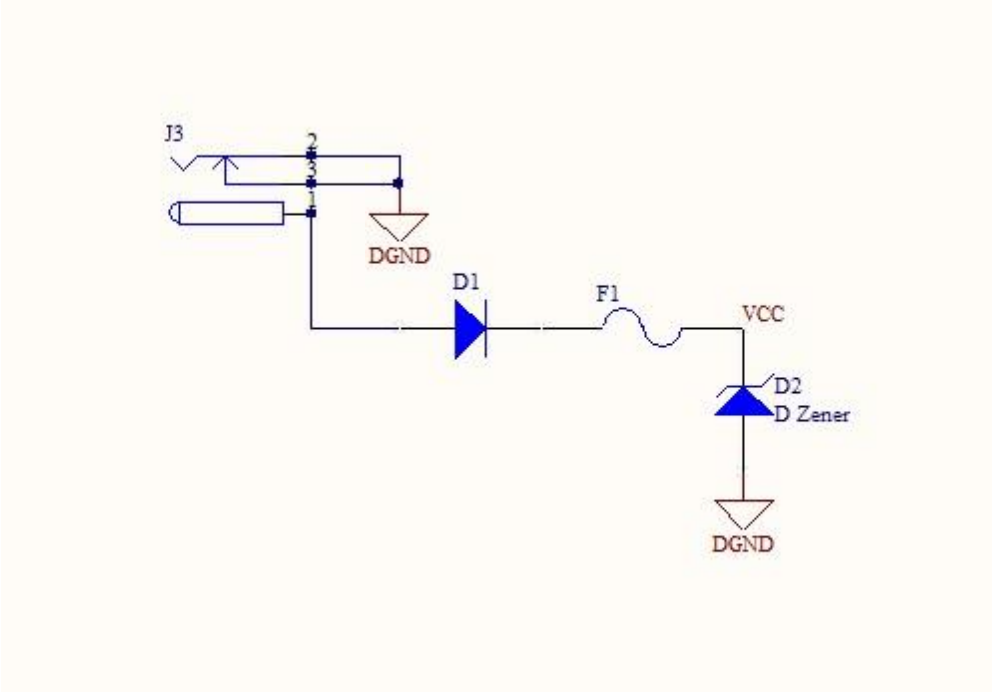
پردازنده DSP برای کار به دو سطح ولتاژ نیاز دارد که از ولتاژ ۳/۳ ولت برای استفاده در ادوات جانبی و پورت‌های ورودی و خروجی استفاده می‌شود و از ولتاژ قابل تنظیم که روی ۱/۶تنظیم شده است، برای کار پردازنده استفاده می‌شود.

- هر کدام از خروجی‌ها می‌توانند حداکثر ۱ آمپر جریان خروجی را تحمل کنند و به ازای هر

یک آمپر خروجی حداکثر ۳۵۰ میلی ولت افت جریان دارند.

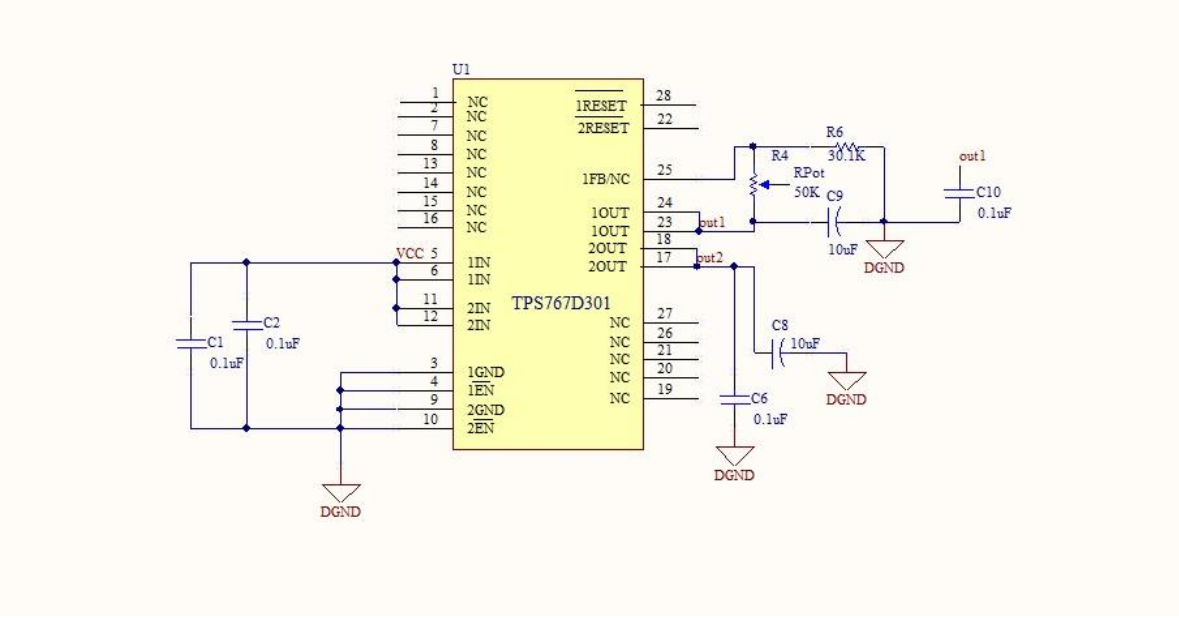
- زمان پایدار شدن رگولاتور ۲۰۰ میلی ثانیه می‌باشد.
- دارای پاسخ گذاری سریعی است.
- دارای محافظ حرارتی می‌باشد که اگر دما بیشتر از اندازه مجاز شود، به صورت خودکار ولتاژ خروجی را قطع می‌کند.

همانگونه که در شکل (۳-۲) نمایش داده شده است، در قسمت ورودی منبع تغذیه از یک مدار محافظ استفاده شده است که از یک فیوز و یک دیود زنر تشکیل شده است. این دو در کنار همدیگر وظیفه محافظت از مدار را به عهده دارند. مدار محافظ به گونه‌ای عمل می‌کند که فیوز از عبور جریان بیشتر از ۵۰۰ میلی آمپر جلوگیری می‌کند و دیود زنر نیز از افزایش ولتاژ بیشتر از ۵/۶ ولت جلوگیری می‌کند.



شکل(۳-۲): مدار محافظ منبع تغذیه

خروجی مدار شکل (۳-۱) وارد آی سی تغذیه می‌شود. در ادامه آی سی تغذیـه دو ولتـاژ خروجـی ۱/۶ ولت و ۳/۳ ولت را ایجاد می‌کند که از طریق کلیدها این دو ولتاژ به سرتاسر مدار انتقال پیدا می‌کنند. آی سی تغذیه و ادوات جانبی آن را می‌توان در شکل (۳-۳) مشاهده کرد.



شکل(۳-۳): آی سی تغذیه و ادوات جانبی آی سی

۳-۲-۲ مبدل داده‌ها

برای تبدیل داده‌های آنالوگ به دیجیتال از آی سی کُدِک[1] TLV320AIC23B استفاده شده است. برای استفاده از آی سی باید قبل از شروع به کار مدار، با استفاده از نـرم افـزار Code Composer Studio ثبات‌های درون کدک را به طور مناسبی برنامه‌ریزی کرد. کلاک[2] مورد نیاز برای آی سی نیز توسـط یـک کریستال نوسان‌ساز خارجی ۱۲ مگاهرتز تامین می‌شود. از ویژگی‌های این آی سی می‌توان به مـوارد زیـر

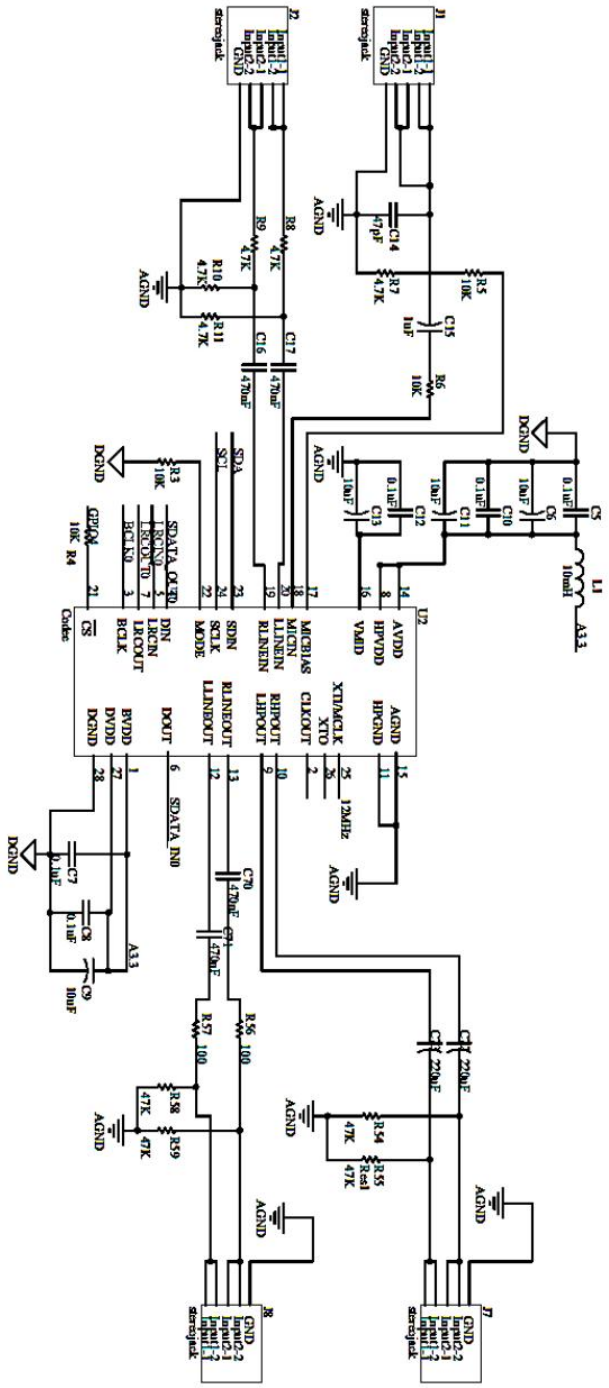
[1]Codec
[2]Clock

اشاره کرد[61]:

- توسط شرکت Texas Instrument ساخته شده است و با پردازنده‌های DSP ساخت این شرکت هماهنگی کامل دارد.
- کدک توانایی راه اندازی یک میکروفن به صورت مستقیم را داشته و ولتاژ بایاس میکروفن را نیز تامین می‌کند.
- شامل یک کانال خط ورودی[1] دوتایی، یک ورودی میکروفن به صورت جداگانه، یک کانال خروجی[2] دوتایی و یک خروجی هدفون می‌باشد.
- فرکانس نمونه‌برداری این آی سی قابل تنظیم بوده و بازه بین ۸ الی ۹۶ کیلوهرتز را پوشش می‌دهد.
- امکان ارسال و دریافت داده‌ها را به صورت همزمان و در اندازه‌های ۸، ۱۶، ۲۴ و ۳۲ بیت توسط پورت McBSP دارد.
- ولتاژ کاری کدک مشابه DSP و برابر با ۱/۶ است.
- تنظیم رجیسترهای کدک توسط روشهای مختلفی مانند SPI، I2C و McBSP قابل انجام است.
- شامل تقویت کننده داخلی جداگانه برای ورودی میکروفن و ورودی‌های دوتایی جهت تقویت سیگنال‌های ورودی می‌باشد.
- به منظور آزمایش عملکرد کدک، در داخل آن یک مسیر کنارگذر تعبیه شده است و می‌تواند هر سیگنالی که وارد آن می‌شود را به صورت مستقیم و بدون تغییر به خروجی بدهد.

[1] Line In
[2] Line Out

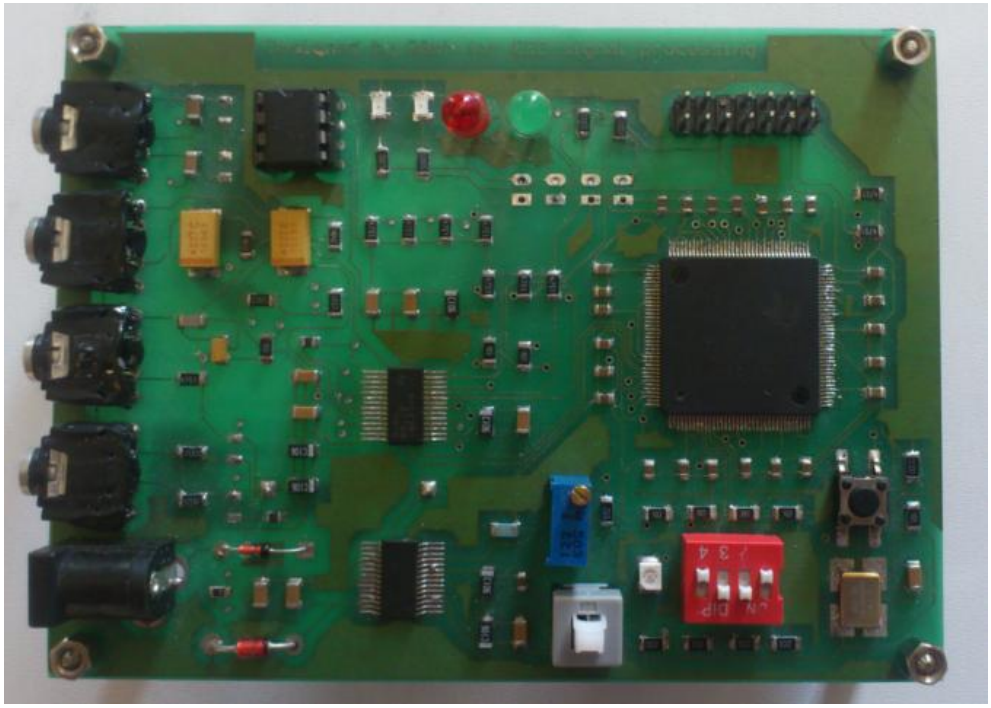
مدار کدک و ادوات جانبی آن را در شکل (۳-۴) مشاهده می‌کنید.



شکل(۳-۴): مدار کدک و ادوات جانبی همراه با آی سی

۳-۲-۳ نمای کلی مدار

در نهایت اصلی‌ترین قسمت مدار، پردازنده سیگنال می‌باشد که در بخش‌های قبل در مورد آن صحبت کرده‌ایم. نمای کلی مدار طراحی شده به صورت شکل (۳-۵) می‌باشد.



شکل(۳-۵): مدار نهایی جهت شناسایی گوینده

از دیگر ادوات نصب شده بر روی مدار می‌توان به موارد زیر اشاره کرد:

- تعداد ۳ عدد LED که بر روی پایه‌های GPIO 4, 6, 7 نصب شده‌اند و بـه منظـور نمـایش خروجی‌های مدار از آن‌ها استفاده می‌شود.
- یک عدد dip switch چهار کاناله که بر روی پایه‌های GPIO 1, 2, 3, 4 نصب شده و از آن برای تنظیم نوع Boot load استفاده می‌شـود. در مـدار طراحـی شـده، بـرای Boot load کردن پردازنده می‌توان از دو روش استفاده کرد. روش اول از طریق کامپیوتر(با استفاده از JTAG) می‌باشد و روش دوم از طریق حافظه EEPROM است.

فصل چهارم

بررسی نتیجه آزمایش‌ها

۴-۱ معرفی پایگاه‌های داده استفاده شده در آزمایش‌ها

۴-۱-۱ پایگاه داده TIMIT

پایگاه داده TIMIT یک پایگاه داده استاندارد می‌باشد که برای کاربردهای پردازش گفتار مانند شناسایی گوینده، جستجوی کلمه کلیدی و .... گردآوری شده است. این مجموعه شامل ۶۳۰ گوینده می-باشد که ۴۳۸ نفر مرد و ۱۹۲ نفر از این مجموعه زن هستند. هر کدام از گویندگان ۱۰ جمله را ادا کرده-اند که هر جمله به صورت جداگانه در یک فایل صوتی ذخیره شده است. در ضمن فایل‌های این پایگاه داده بدون نویز می‌باشند[62].

برای استفاده از این پایگاه داده به این ترتیب عمل شده که به صورت تصادفی ۵۰ گوینده از میان مجموعه انتخاب شده که ۲۵ نفر مرد و ۲۵ نفر زن هستند. به ازای هر گوینده، ۱۰ فایل وجود دارد که ۸ فایل برای آموزش و ۲ فایل باقی مانده برای آزمایش در نظر گرفته شده است.

۴-۱-۲ پایگاه داده vidTIMIT

این پایگاه داده، به صورت ترکیبی از صدا و ویدیو ساخته شده است که برای کاربردهایی نظیر لب خوانی خودکار، تشخیص چهره، و انواع کاربردهای پردازش گفتار مانند شناسایی گوینده، تایید گوینده و .... قابل استفاده می‌باشد. این مجموعه شامل ۴۳ گوینده (۲۴ مرد و ۱۹ زن) است که به ازای هر گوینده ۱۰ فایل صوتی وجود دارد. این پایگاه داده به صورت نویزی تهیه شده است. نویز موجود عموما حاصل از وسایل موجود در اتاق ضبط صدا مانند صدای پنکه[1] کامپیوتر می‌باشد[63, 64].

برای استفاده از این مجموعه از تمام ۴۳ گوینده موجود استفاده شده و به ازای هر گوینده ۸ فایل

[1] Fan

برای آموزش و ۲ فایل برای آزمایش در نظر گرفته شده است.

۴-۲ بررسی نتایج آزمایش‌ها

در بررسی‌هایی که در فصل دوم داشتیم، دو روش را برای استخراج ویژگی انتخاب کردیم. این دو روش عبارتند از روش MFCC و روش LPCC که هر دو می‌توانند بدون استفاده از هیچ پس پردازشی استفاده بشوند و یا اینکه از پس پردازش نیز برای بهبود کیفیت این ویژگی‌ها و افزایش مقاومت آن‌ها در مقابل نویز و اثرات کانال انتقال استفاده کرد. برای به دست آوردن بهترین نتیجه، در آزمایشات از هر دو روش MFCC و LPCC استفاده کرده‌ایم و برای پس پردازش از روش‌های محاسبه CMN، ضرایب دلتا و ضرایب شتاب استفاده کرده‌ایم.

از میان روش‌های مختلف مدل کردن گوینده، روش مدل کردن با استفاده از GMM را انتخاب کرده‌ایم و نتایج شناسایی با استفاده از این مدل و هر دو روش MFCC و LPCC را ارائه داده‌ایم. در ادامه به بررسی نتایج آزمایشات می‌پردازیم.

۴-۲-۱ مقایسه روش‌های استخراج ویژگی MFCC و LPCC

اولین موضوعی که بررسی شده این است که از بین روش‌های استخراج ویژگی MFCC و LPCC کدام یک برای کاربردهای شناسایی گوینده مناسب‌تر بوده و درصد شناسایی بیشتری را ارائه می‌دهد. برای این کار ابتدا از پایگاه داده TIMIT به صورت تصادفی ۲۵ مرد و ۲۵ زن انتخاب شده است. برای هر گوینده ۸ فایل برای آموزش و ۲ فایل برای قسمت آزمایش در نظر گرفته شد. پس ۱۰۰ نمونه آزمایش وجود دارد. به این ترتیب طول سیگنال آموزش برای هر گوینده به طور متوسط حدود ۲۵ ثانیه خواهد بود و طول سیگنال آزمایش را حداکثر حدود ۱/۵ ثانیه در نظر گرفته شد.

برای استخراج ویژگی طول قاب‌ها را ۲۵ میلی ثانیه در نظر گرفته شده و میـزان هـم‌پوشـانی بـین قاب‌ها نیز ۱۵ میلی ثانیه انتخاب شده است. پس هر ۱۰ میلی ثانیه یک بار قاب برداری انجـام مـی‌شـود. برای بررسی اثر طول نمونه آزمایش بر درصد شناسایی سیستم، آزمایش‌ها برای هر نمونه به ازای ۳ تعداد قاب مختلف انجام شده است. آزمـایش اول بـه ازای تعـداد ۵۰ قـاب، آزمـایش دوم بـه ازای ۱۰۰ قـاب و آزمایش سوم را به ازای ۱۵۰ قاب انجام شد.

در روش MFCC به طور معمول اندازه بردار ویژگی ۱۳ فرض می‌شود[5,8]. این عـدد بـا احتسـاب ضریب انرژی، (C_0)، است. در آزمایش‌های ابتدایی از ضرایب انرژی صرفنظر شده و در انتهای فصل اثر این ضرایب بر روی نتیجه بررسی شده است. در ادامه برای این بـردار ویژگـی بـه عنـوان اولـین مرحلـه پـس پردازش، عملیات CMN را انجام می‌دهند. پس تا اینجا یک بـردار ویژگـی ۱۲ درایـه‌ای داریـم کـه روش CMN نیز بر روی آن اعمال شده است. اکنون در مرحله دوم از پس پردازش، برای بهبود ویژگی‌ها از آن‌ها ضرایب دلتا را استخراج می‌کنیم. و در مرحله سوم پس پردازش، از ضرایب دلتا، ضرایب شتاب را استخراج می‌کنیم. وقتی که پس پردازش تمام شد و ضرایب جبران سازی شده به وسـیله CMN، ضـرایب دلتـا و ضرایب شتاب محاسبه شدند، آن‌ها را در یک بردار قرار داده و یک بـردار ویژگـی نهـایی ۳۶ درایـه‌ای بـه دست می‌آید. ترتیب قرار دادن این ضرایب در بردار نهایی هم معمولا بـه ایـن صـورت اسـت کـه ضـرایب جبران سازی شده با استفاده از CMN را به عنوان درایه‌های ۱ تا ۱۲، ضرایب دلتا را به عنوان درایـه‌هـای ۱۳ تا ۲۴ و ضرایب شتاب را به عنوان درایه‌های ۲۵ تا ۳۶ در نظر می‌گیرند. اگر ضرایب انرژی نیز در نظر گرفته شوند، این ضرایب در محاسبات ضرایب دلتا و شتاب نیـز اسـتفاده مـی‌شـوند و در بـردار نهـایی در مکان‌های اول، ۱۴ و ۲۷ قرار می‌گیرند.

برای روشLPCC نیز به این منظور که بـا روش MFCC هماهنـگ باشـد، بردارهـای ویژگـی ۱۳ درایه‌ای در نظر گرفته شده‌اند. ضریب (C_0) نیز برای روش LPCC در نظر گرفته نشده است. همانند روش

MFCC، برای ضرایب کپسترال LPC، جبران سازی به روش CMN انجام شده و از ایـن ضـرایب جبـران سازی شده، ضرایب دلتا و شتاب محاسبه شده و یک بردار ویژگی ۳۶ درایه‌ای شده است.

در این آزمایش برای ویژگی‌ها دو حالت را در نظر گرفته شده است. در حالت اول از هیچ نوع پـس پردازشی استفاده نشده و فقط از ضرایب MFCC و LPCC استفاده شـده اسـت. یعنـی اینکـه بردارهـای ویژگی دارای ۱۲ درایه می‌باشند. در حالت دوم، هر ۳ روش پس پردازش اعمال شده و بردارهـای ویژگـی ۳۶ درایه‌ای ساخته شده است. برای بررسی طول زمان نمونه آزمایش بر روی درصد شناسایی نیز ۳ حالت مختلف در نظر گرفته شد.

اکنون که ویژگی‌ها استخراج شده‌اند، باید با استفاده از این ویژگی‌ها بـرای هـر گوینـده یـک مـدل GMM ساخته شود. این کار با استفاده از الگوریتم EM انجام می‌شـود. بـرای هـر گوینـده بـا اسـتفاده از الگوریتم EM مدل‌های GMMای با تعداد ترکیب‌های مختلف آموزش داده شده است. برای کاهش حجـم محاسباتی، از ماتریس کواریانس قطری استفاده شده است. مدل‌ها به ازای تعداد ترکیب‌های ۵، ۱۰، ۱۵، ۲۰ و ۲۵ آموزش داده شدند. این اعداد به صورت تجربی به دست آمده‌اند. برای اینکه دلیـل انتخـاب ایـن اعداد را شرح بدهیم، باید بگوییم که برای ساختن مدل GMM برای یک گوینده، متناسـب بـا هـر تعـداد ترکیبی که انتخاب می‌کنیم باید به همان میزان نیز داده‌های آموزشی در اختیار الگوریتم EM قرار بدهیم. وقتی که تعداد ترکیب‌ها را افزایش می‌دهیم، تعداد پارامترهایی که باید تخمین بزنیم نیز افزایش پیدا می-کنند و برای اینکه این پارامترها را تخمین بزنیم، نیاز داریم که میزان داده بیشتری به الگـوریتم EM وارد کنیم. فرض کنید که این کار را انجام ندهیم و بدون توجه به این موضوع، داده‌های آموزشی را ثابـت نگـه داشته و تعداد ترکیب‌ها را افزایش بدهیم، در این صورت پارامترهای مدلی که به دست می‌آوریـم، مقـدار صحیحی نخواهند داشت و این باعث می‌شود که درصد شناسایی گوینده کاهش پیدا کند. در آزمایش‌هایی که انجام داده‌ایم، با توجه به میزان داده‌های آموزشی که در دسترس داشتیم، وقتی که تعداد ترکیب‌هـا را

از ۲۵ بیشتر می‌کنیم، درصد شناسایی شروع به کمتر شدن می‌کند. به همین دلیل از آزمایش به ازای ترکیب‌های بیشتر از ۲۵ خودداری کردیم. در ضمن دلیل اینکه ما فقط به ازای این ۵ عدد آزمایشات را انجام داده و مقادیر بین این اعداد را پوشش ندادیم این است که آموزش با استفاده از الگوریتم EM کار زمان‌بری می‌باشد و با توجه به میزان آزمایشات، به این تعداد ترکیب اکتفا کردیم تا روند کلی تاثیر افزایش تعداد ترکیب‌ها را بر روی درصد شناسایی سیستم بررسی کنیم.

هنگامی که با استفاده از داده‌های آموزشی مدل‌ها ساخته شدند، باید درصد شناسایی برای نمونه-های آزمایش محاسبه شود. روش آزمایش به این ترتیب است که هر کدام از نمونه‌ها را با تمام مدل‌های ۵۰ گوینده مقایسه کرده و هر کدام که امتیاز بیشتری بیاورد، نمونه آزمایش متعلق به آن گوینده می‌باشد. برای این کار از معادله (۲-۳۴) استفاده شده است.

در ادامه نتایج آزمایش ارائه شده است. نتیجه آزمایش به ازای بردارهای ویژگی ۱۲ درایه‌ای و روش استخراج ویژگی MFCC در جدول (۴-۱) قابل مشاهده می‌باشد.

جدول(۴-۱): نرخ شناسایی گوینده به ازای بردار ویژگی ۱۲ درایه‌ای MFCC بر حسب درصد

شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۶۳	۸۸	۹۱
آزمایش دوم	۱۰	۷۴	۹۱	۹۶
آزمایش سوم	۱۵	۷۷	۹۴	۹۸
آزمایش چهارم	۲۰	۷۴	۹۱	۹۵
آزمایش پنجم	۲۵	۷۸	۸۹	۹۱

در جدول (۴-۱) نتایج آزمایش را به ازای نمونه‌های آزمایش با تعداد قاب متفاوت و مدل‌های GMM با تعداد ترکیب‌های متفاوت مشاهده می‌کنید. همانگونه که انتظار می‌رود، با افزایش تعداد ترکیب-ها، درصد شناسایی نیز افزایش پیدا کرد. این موضوع در حالت‌های با تعداد ترکیب‌های ۵، ۱۰ و ۱۵ قابل مشاهده است. اما در ادامه با افزایش تعداد ترکیب‌ها، درصد شناسایی کاهش پیدا کرده است. این کاهش نیز در راستای تایید مطالب گفته شده در پاراگراف‌های قبل می‌باشد. وقتی تعداد ترکیب‌ها از ۱۵ بیشتر می‌شود، به دلیل عدم تعادل بین تعداد ترکیب‌ها و میزان داده‌های آموزش، الگوریتم EM توانایی تخمین صحیح پارامترهای مدل GMM را ندارد و این موضوع باعث می‌شود که درصد شناسایی کاهش پیدا کند. از طرفی اثر تعداد قاب‌های نمونه‌های آزمایش بر درصد شناسایی نیز در این جدول قابل مشاهده می‌باشد. با افزایش طول زمان نمونه آزمایش، جواب‌ها نیز به میزان قابل توجهی بهبود پیدا کرده‌اند.

در ادامه در جدول (۴-۲)، نتایج آزمایش را به ازای روش استخراج ویژگی LPCC که در آن

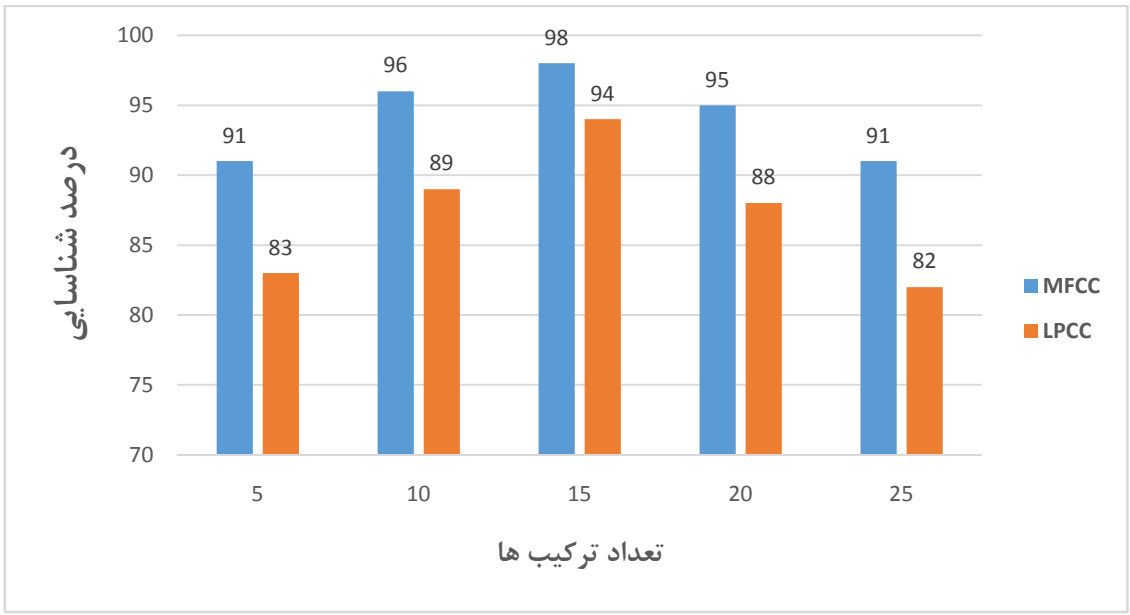
بردارهای ویژگی ۱۲ درایه‌ای می‌باشند، مشاهده می‌کنید.

جدول(۴-۲): نرخ شناسایی گوینده به ازای بردار ویژگی ۱۲ درایه‌ای LPCC بر حسب درصد

شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۵۰	۷۸	۸۳
آزمایش دوم	۱۰	۶۲	۸۱	۸۹
آزمایش سوم	۱۵	۷۰	۸۵	۹۴
آزمایش چهارم	۲۰	۶۳	۸۰	۸۸
آزمایش پنجم	۲۵	۵۵	۷۹	۸۲

نتیجه مشاهده شده در جدول (۴-۲) بیانگر این موضوع می‌باشد که روش MFCC در شرایط یکسان نرخ شناسایی بیشتری نسبت به روش استخراج ویژگی LPCC دارد. در جدول (۴-۲)، حداکثر درصد شناسایی به ازای ترکیب ۱۵ و تعداد قاب ۱۵۰ برابر با ۹۴ درصد می‌باشد که ۴ درصد از روش MFCC با همین متغییرها کمتر است. نتیجه مقایسه این دو آزمایش را در حالتی که نمونه‌های آزمایش دارای ۱۵۰ قاب می‌باشند، در شکل (۴-۱) مشاهده می‌کنید.

در ادامه همین مقایسه برای حالتی انجام شده است که در هر دو روش MFCC و LPCC از ۳ روش پس پردازش محاسبه ضرایب CMN، دلتا و شتاب استفاده شده است. نتایج آزمایش برای روش MFCC وقتی که بردارهای ویژگی دارای ۳۶ درایه می‌باشند در جدول (۴-۳) نمایش داده شده است.



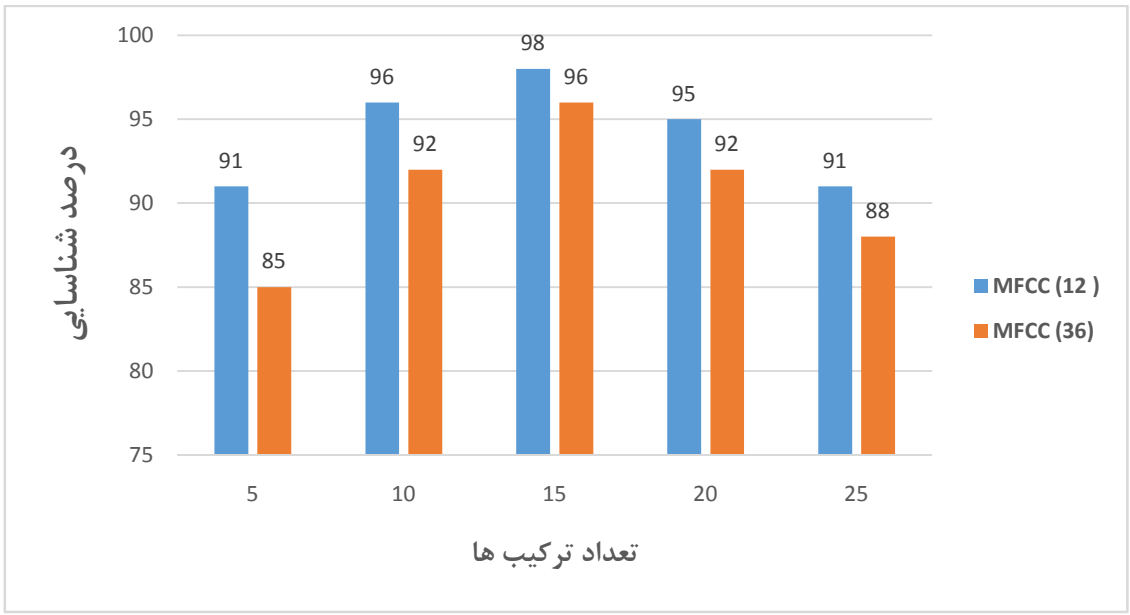
شکل(۴-۱) : مقایسه نتایج حاصل از شبیه سازی با استفاده از دو روش MFCC و LPCC هنگامی که بردارهای ویژگی برای هر دو روش دارای ۱۲ درایه می‌باشند

جدول(۴-۳): نرخ شناسایی گوینده به ازای بردار ویژگی ۳۶ درایه‌ای MFCC بر حسب درصد

شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۵۷	۷۸	۸۵
آزمایش دوم	۱۰	۶۸	۸۳	۹۲
آزمایش سوم	۱۵	۸۰	۸۸	۹۶
آزمایش چهارم	۲۰	۷۶	۸۶	۹۲
آزمایش پنجم	۲۵	۷۸	۸۲	۸۸

در جدول (۴-۳) بهترین نتیجه شناسایی مربوط به حالتی است که تعداد ترکیب‌ها ۱۵ و طول نمونه تست برابر ۱۵۰ قاب باشد که نتیجه برابر با ۹۶ درصد می‌باشد. در این آزمایش نسبت به آزمایشی که از بردار ویژگی MFCC با ۱۲ درایه استفاده شده، نتایج ضعیف‌تری کسب شده است در صورتی که ویژگی‌ها در این آزمایش بهبود داده شده‌اند. دلیل این کاهش درصد شناسایی سیستم این است که با افزایش اندازه بردار ویژگی، تعداد پارامترهایی که باید تخمین زده شوند نیز افزایش پیدا می‌کند. در حالت کلی، در مدل GMM برای یک گوینده با بردار ویژگی N درایه‌ای، ابعاد بردارهای میانگین، برابر $N \times 1$ و اندازه ماتریس کواریانس $N \times N$ می‌باشد. بنابراین وقتی که ابعاد بردار ویژگی افزایش پیدا می‌کند، میزان پارامترهایی که باید تخمین زده شوند نیز افزایش پیدا خواهد کرد. وقتی ابعاد بردار ویژگی از ۱۲ به ۳۶ تغییر کرده است، تعداد پارامترهایی که باید تخمین زده شود، زیادتر شده است اما میزان اطلاعاتی که بردارهای ویژگی ۳۶ درایه‌ای از داده‌های آموزشی استخراج می‌کنند به اندازه‌ای نیست که بتواند پارامترهای اضافه شده را تخمین بزند. به همین دلیل مشاهده می‌شود درصد شناسایی نسبت به حالتی که از بردار ویژگی ۱۲ درایه‌ای استفاده شده است، ۲ درصد کمتر شده است. مقایسه بین روش‌های استخراج ویژگی MFCC با بردار ویژگی ۱۲ درایه‌ای و ۳۶ درایه‌ای در حالتی که نمونه‌های آزمایش ۱۵۰ قاب می‌باشند در شکل (۴-۲) نمایش داده شده است.

در ادامه روش استخراج ویژگی LPCC را برای زمانی که از پس پردازش‌های CMN، محاسبه ضرایب دلتا و ضرایب شتاب نیز استفاده شده، بررسی شده است. در این حالت بردار ویژگی شامل ۳۶ درایه می‌باشد. نتایج آزمایش با این روش استخراج ویژگی در جدول (۴-۴) ارائه شده است.



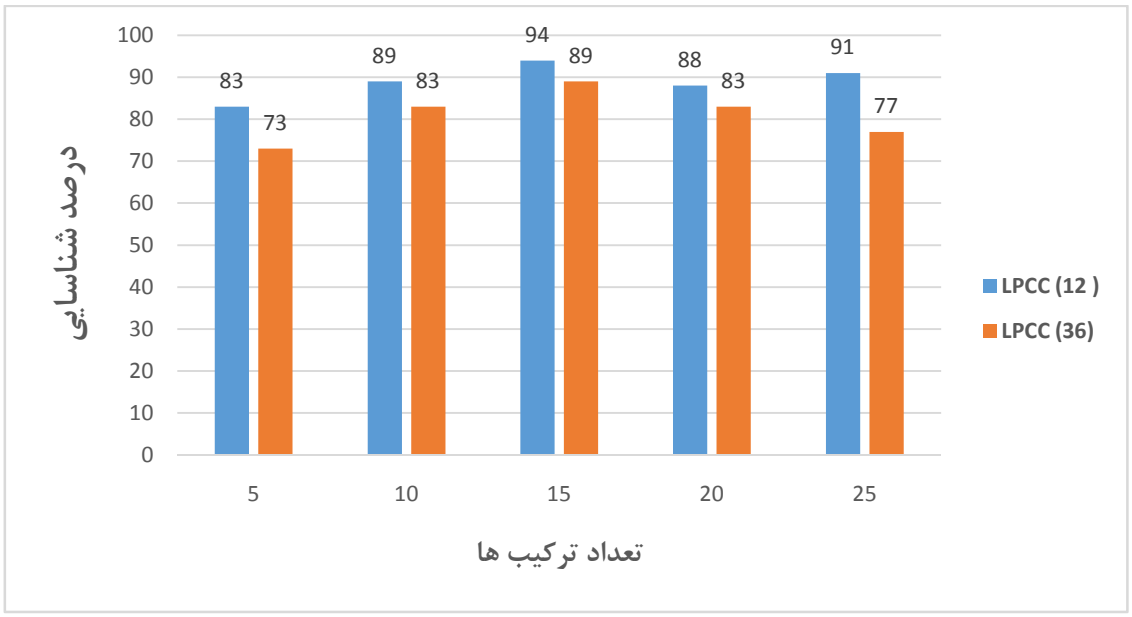
شکل(۴-۲): مقایسه نتیجه‌های حاصل از روش استخراج ویژگی با بردارهای ویژگی ۱۲ و ۳۶ درایه ای در روش استخراج ویژگی MFCC

جدول(۴-۴): نرخ شناسایی گوینده به ازای بردار ویژگی ۳۶ درایه‌ای LPCC بر حسب درصد

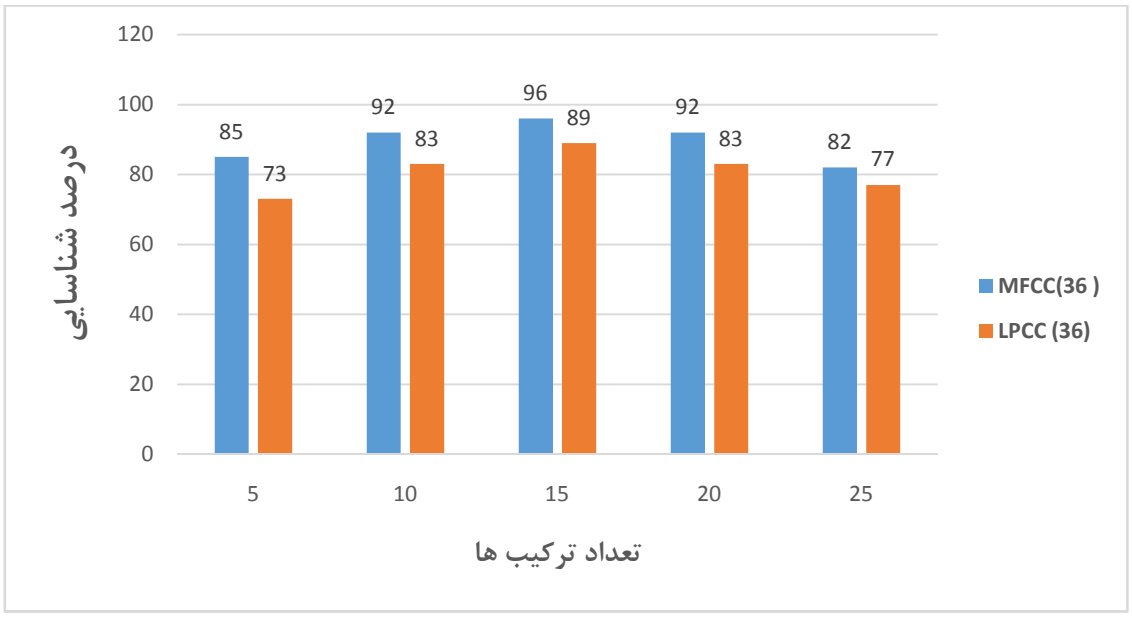
شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۴۸	۷۰	۷۳
آزمایش دوم	۱۰	۵۷	۷۵	۸۳
آزمایش سوم	۱۵	۶۹	۸۰	۸۹
آزمایش چهارم	۲۰	۶۴	۷۹	۸۳
آزمایش پنجم	۲۵	۵۵	۷۴	۷۷

همانگونه که در جدول (۴-۴) مشاهده می‌شود، روند کاهش درصد شناسایی نسبت به حالتی که از بـردار ویژگی ۱۲ درایه‌ای استفاده شده باشد، در آزمایش مربوط به روش LPCC نیز تکرار شده اسـت. حـداکثر درصد شناسایی در جدول (۴-۴) برابر ۸۹ درصد می‌باشد که نسبت به حالـت مشـابه وقتـی کـه از بـردار ویژگی ۱۲ درایه‌ای استفاده شده باشد، ۵ درصد کاهش پیدا کرده است. در ادامه با مقایسه جدول(۴-۴) با جدول (۴-۳) این نتیجه به دست می‌آید که استفاده از روش MFCC نسبت بـه روش LPCC در شـرایط یکسان، درصد شناسایی بالاتری را به دست می‌دهد. در شکل (۴-۴) مقایسه بـین روش LPCC بـا بـردار ویژگی‌های ۱۲ و ۳۶ درایه‌ای به ازای نمونه‌های آزمایش با ۱۵۰قاب نمایش داده شده است. نتیجه مقایسه بین روش MFCC و LPCC که از بردارهای ویژگی ۳۶ درایه‌ای و نمونه‌های آزمایش ۱۵۰ قـابی اسـتفاده کرده‌اند نیز در شکل (۴-۵) نمایش داده شده است.

برای جمع‌بندی می‌توان گفت که با توجه بـه نتیجـه آزمایشـاتی کـه انجـام داده شـد، اسـتفاده از روش MFCC وقتی که از هیچ نوع پس پردازشی استفاده نشده باشد و یا وقتی که از پس پردازش استفاده شده باشد، به نسبت روش LPCC در سیستم شناسایی گوینده درصد شناسایی بیشتری به دست می‌دهد.



شکل(۴-۳): مقایسه نتیجه‌های روش استخراج ویژگی LPCC در دو حالت بردار ویژگی ۱۲ و ۳۶ درایه‌ای



شکل(۴-۴): مقایسه بین روش‌های MFCC و LPCC وقتی که بردارهای ویژگی ۳۶ درایه‌ای باشند

۴-۲-۲ بررسی اثر پس پردازش بر کارکرد سیستم شناسایی گوینده

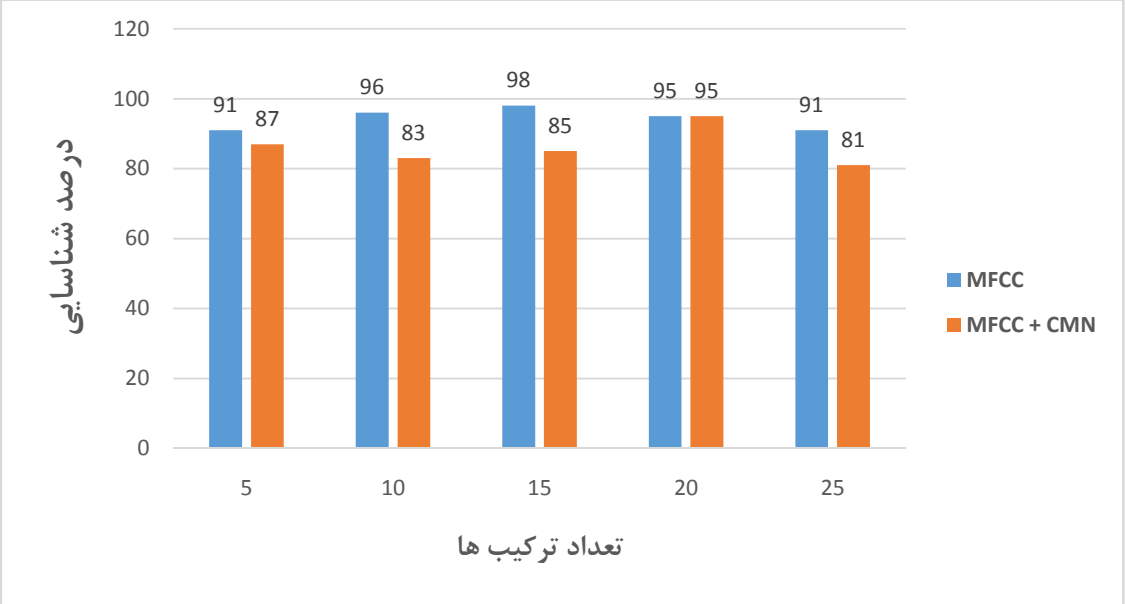
تا این لحظه آزمایش‌ها مشخص کرده‌اند که استفاده از روش استخراج ویژگی MFCC به نسبت روش LPCC در کاربردهای شناسایی گوینده نتایج بهتری به دست می‌دهد. در ادامه به بررسی تاثیر روش‌های پس پردازش در کیفیت سیستم شناسایی گوینده پرداخته می‌شود.

روش‌های پس پردازشی که مورد استفاده قرار گرفته‌اند، روش جبران سازی CMN، روش محاسبه ضرایب دلتا و روش محاسبه ضرایب شتاب هستند. اثر هر کدام از این سه روش به صورت جداگانه و یا ترکیبی بر روی کارایی سیستم شناسایی گوینده که از روش استخراج ویژگی MFCC استفاده می‌کند، بررسی شده است تا در نهایت بهترین روش که بیشترین درصد شناسایی را به دست می‌دهد انتخاب شده و بر روی پردازنده DSP پیاده‌سازی شود.

اولین روش پس پردازش، روش جبران سازی CMN می‌باشد. هنگامی که ویژگی‌های MFCC استخراج شوند، که در اینجا فرض شده است بردار ویژگی MFCC دارای ۱۲ درایه است، با استفاده از روش جبران سازی CMN، اولین مرحله پس پردازش انجام می‌شود. در روش CMN ابعاد بردار ویژگی تغییر نمی‌کند و تنها با استفاده از میانگین تمام بردارهای ویژگی، هر کدام از بردارهای ویژگی را نرمالیزه می‌کنند. پس در این حالت بردار ویژگی دارای ۱۲ درایه می‌باشد. با استفاده از بردار ویژگی MFCC که به روش CMN جبران سازی شده است، آزمایشی انجام شده و تاثیر این روش پس پردازش بر روی کارکرد سیستم شناسایی گوینده بررسی شده است. نتایج حاصل از این آزمایش در جدول (۴-۵) نمایش داده شده است. در ادامه یک مقایسه بین این روش با وقتی که تنها از روش استخراج ویژگی MFCC بدون جبران سازی استفاده شده است، انجام شده و در شکل (۴-۵) قابل مشاهده می‌باشد. در این مقایسه تعداد قاب‌های نمونه‌های آزمایش ۱۵۰ انتخاب شده است.

جدول(۴-۵): نرخ شناسایی گوینده با استفاده از MFCC جبران سازی شده توسط CMN بر حسب درصد

شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۶۸	۸۱	۸۷
آزمایش دوم	۱۰	۶۷	۷۷	۸۳
آزمایش سوم	۱۵	۷۶	۸۲	۸۵
آزمایش چهارم	۲۰	۸۵	۹۲	۹۵
آزمایش پنجم	۲۵	۷۳	۷۹	۸۱



شکل(۴-۵): مقایسه بین روش‌های MFCC و MFCC + CMN

داده‌های جدول (۴-۵) نشان می‌دهند که روش CMN به تنهایی تاثیر مثبتی در بهبود کیفیت ویژگی‌های استخراج شده ندارد و باعث کاهش درصد شناسایی سیستم می‌شود. در شکل (۴-۵) مقایسه-ای بین روشMFCC بدون استفاده از CMN و روش MFCC که از CMN استفاده کرده است را مشاهده می‌کنید. همانگونه که از این مقایسه آشکار است، روش استخراج ویژگی MFCC بدون استفاده از CMN درصد شناسایی بیشتری را به دست می‌دهد. حداکثر درصد شناسایی روش MFCC که از CMN نیز استفاده کرده است برابر ۹۵ درصد می‌باشد که نسبت به روش MFCC بدون استفاده از CMN کاهش ۳ درصدی دارد.

در مرحله بعد اثر پارامترهای دلتا بر روی کیفیت ویژگی‌های استخراج شده بررسی شده است. روش کار به این ترتیب است که ابتدا بردارهای ویژگی MFCC استخراج و توسط روش CMN جبران سازی می‌شوند، از ویژگی‌های جبران سازی شده پارامترهای دلتا استخراج می‌شوند و در نهایت با پشت سر هم قرار دادن ویژگی‌های جبران سازی شده و پارامترهای دلتا استخراج شده، یک بردار ویژگی ۲۴ درایه‌ای ساخته می‌شود. ۱۲ درایه اول ضرایب جبران سازی شده به روش CMN و ۱۲ درایه دوم نیز پارامترهای دلتا هستند.

نتایج آزمایش به ازای روش استخراج ویژگی MFCC که از روش‌های پس پردازش CMN و محاسبه پارامترهای دلتا استفاده کرده است، در جدول (۴-۶) نمایش داده شده است. مقایسه نتایج استخراج ویژگی MFCC، MFCC همراه با CMN و MFCC همراه با CMN و پارامترهای دلتا به ازای نمونه‌های آزمایش ۱۵۰ قاب در شکل (۴-۶) قابل مشاهده است.

جدول(۴-۶) :نرخ شناسایی گوینده با استفاده از MFCC جبران سازی شده همراه با CMN و پارامترهای دلتا بر حسب درصد

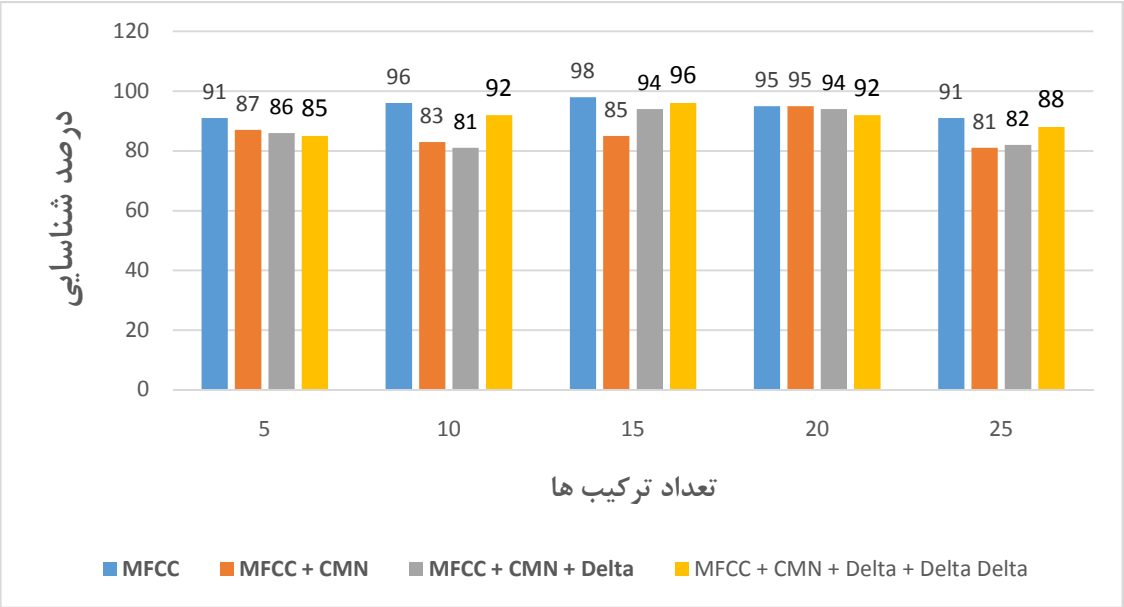
شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۶۸	۷۹	۸۶
آزمایش دوم	۱۰	۶۸	۷۶	۸۱
آزمایش سوم	۱۵	۷۶	۹۲	۹۴
آزمایش چهارم	۲۰	۷۹	۹۱	۹۴
آزمایش پنجم	۲۵	۷۰	۷۹	۸۲

شکل(۴-۶): مقایسه بین روش‌های MFCC و MFCC + CMN و MFCC + CMN + Delta

نتایج آزمایش در جدول (۴-۶) نشان می‌دهند که استفاده از ضرایب دلتا نه تنها باعث افزایش درصد شناسایی نشده است، بلکه باعث کاهش ۱ درصدی آن نسبت به حالتی که از ضرایب دلتا استفاده نکرده‌ایم شده است. گرچه کارایی سیستم یک درصد کاهش پیدا کرده است اما در شکل (۴-۶) مشاهده می‌شود که تغییرات نسبت به روشی که فقط از روش جبران سازی CMN استفاده کرده‌ایم، زیاد نیست.

در ادامه از پارامترهای شتاب (یا دلتا - دلتا) نیز برای پس پردازش استفاده شده است. همانند مرحله قبل، در این روش ابتدا پارامترهای دلتا استخراج می‌شود و سپس از پارامترهای دلتا، پارامترهای دلتا - دلتا استخراج شده است.در نهایت از کنار هم قرار دادن ویژگی‌های جبران سازی شده به روش CMN، پارامترهای دلتا و پارامترهای دلتا – دلتا یک بردار ویژگی ۳۶ درایه‌ای حاصل شده است. نتایج آزمایش با استفاده از این روش در جدول (۴-۳) قابل مشاهده می‌باشد و مقایسه بین این روش با روش-های قبل، وقتی که نمونه‌های آموزش دارای ۱۵۰ قاب هستند در شکل (۴-۷) ارائه شده است.

آزمایشات با استفاده از این روش استخراج ویژگی نشان می‌دهند که در این حالت کیفیت کارکرد سیستم بهبود پیدا کرده و حد اکثر به ۹۶ درصد رسیده است. علاوه بر این وقتی که تنها از روش CMN و یا از روش CMN همراه با پارامترهای دلتا استفاده شده است، در نتایج نوعی بی نظمی مشاهده می‌شود. منظور از بی نظمی این است که در حالتی که مدل GMM دارای ۵ ترکیب است، و سپس تعداد ترکیب را به ۱۰ افزایش می‌یابد، برخلاف انتظار درصد شناسایی کاهش پیدا می‌کند و سپس وقتی که تعداد ترکیب‌های GMM به از ۱۰ به ۱۵ تغییر داده می‌شود، درصد شناسایی افزایش پیدا می‌کند. اما این بی-نظمی وقتی که از پارامترهای دلتا – دلتا استفاده شده است، از بین رفته و درصد شناسایی نسبت به تعداد ترکیب‌ها به صورت منطقی تغییر می‌کند.



شکل(۴-۷) : مقایسه روش‌های MFCC با ترکیب‌های متفاوتی از پس پردازش‌ها

گرچه با استفاده از ضرایب دلتا - دلتا نرخ شناسایی گوینده یک درصد افزایش پیدا کرد و به ۹۶ درصد رسید، اما هنوز هم در مقایسه با روش MFCC که بدون استفاده از هیچ نوع پس پردازشی دارای نرخ شناسایی ۹۸ درصد می‌باشد، نتایج ضعیف‌تری می‌دهد. اگر نتایج آزمایش‌هایی که تا به حال انجام شده را بررسی کنیم، می‌توان از این ایده که ویژگی‌های MFCC به تنهایی نرخ شناسایی مناسبی به دست می‌دهند و همینطور این ایده که ترکیب روش‌های پس پردازش دلتا و دلتا – دلتا باعث بهبود ویژگی‌ها می‌شوند استفاده کرد و روش MFCC را به تنهایی با روش‌های دلتا و دلتا-دلتا استفاده کرده و مرحله پس پردازش به روش CMN را حذف کرد. به این ترتیب، بعد از اینکه ویژگی‌های MFCC استخراج شد، باید از این ویژگی‌ها مستقیما ضرایب دلتا و سپس دلتا – دلتا را محاسبه کرد. در نهایت یک بردار ویژگی ۳۶ درایه‌ای ایجاد می‌شود. نتایج آزمایش با استفاده از این روش استخراج ویژگی در جدول (۴-۷) نمایش داده شده است.

جدول(۴-۷): نرخ شناسایی به ازای استفاده از روش MFCC ، دلتا و دلتا – دلتا بر حسب درصد

شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۵۰	۷۶	۸۴
آزمایش دوم	۱۰	۶۴	۸۴	۹۴
آزمایش سوم	۱۵	۷۳	۹۱	۹۷
آزمایش چهارم	۲۰	۷۲	۸۹	۹۷
آزمایش پنجم	۲۵	۷۳	۸۶	۹۰

همانگونه که در جدول(۴-۷) مشاهده می‌شود، حذف روش جبران سازی CMN باعث شده که درصد شناسایی یک درصد افزایش پیدا کند و به ۹۷ درصد برسد. می‌توان برای بهبود بردارهای ویژگی، ضریب انرژی در بردار ویژگی MFCC را نیز به ویژگی‌ها اضافه کرد. پس از چند آزمایش و انتخاب ترکیب‌های مختلف از روش‌های پس پردازش توانستیم درصد شناسایی را به ۹۹ درصد افزایش دهیم، یعنی یک درصد بیشتر از حالتی که از روش MFCC بدون پس پردازش استفاده شده است. ویژگی‌های که استخراج کرده‌ایم ترکیبی از تمام روش‌های پس پردازش و ضریب انرژی هستند. روش استخراج ویژگی به این ترتیب می‌باشد که ابتدا ویژگی‌های MFCC را استخراج کردیم، این ویژگی‌ها شامل ضریب انرژی نیز می‌باشند. سپس روش جبران سازی CMN را بر روی بردار ویژگی اجرا کردیم و مانند آزمایشات قبل از این ضرایب جبران‌سازی شده، پارامترهای دلتا و از پارامترهای دلتا، پارامترهای دلتا – دلتا را محاسبه کردیم. در نهایت به جای اینکه از ویژگی‌های جبران سازی شده در بردار ویژگی نهایی استفاده کنیم، از ویژگی‌های MFCC استفاده کردیم. یعنی بردار نهایی ما شامل ویژگی‌های MFCC به علاوه

پارامترهای دلتا و دلتا – دلتایی می‌باشد که از بردار MFCC جبران سازی شده به روش CMN محاسبه شده‌اند. نتایج حاصل از این آزمایش را در جدول (۴-۸) مشاهده می‌کنید.

جدول(۴-۸): نرخ شناسایی به ازای روش استخراج ویژگی MFCC با بردار ویژگی ۳۹ درایه‌ای

شماره آزمایش	تعداد ترکیب GMM	تعداد قاب‌های نمونه‌های آزمایش		
		۵۰	۱۰۰	۱۵۰
آزمایش اول	۵	۶۶	۸۱	۸۶
آزمایش دوم	۱۰	۷۳	۸۹	۹۱
آزمایش سوم	۱۵	۷۸	۹۱	۹۷
آزمایش چهارم	۲۰	۸۸	۹۶	۹۹
آزمایش پنجم	۲۵	۸۴	۹۵	۹۸

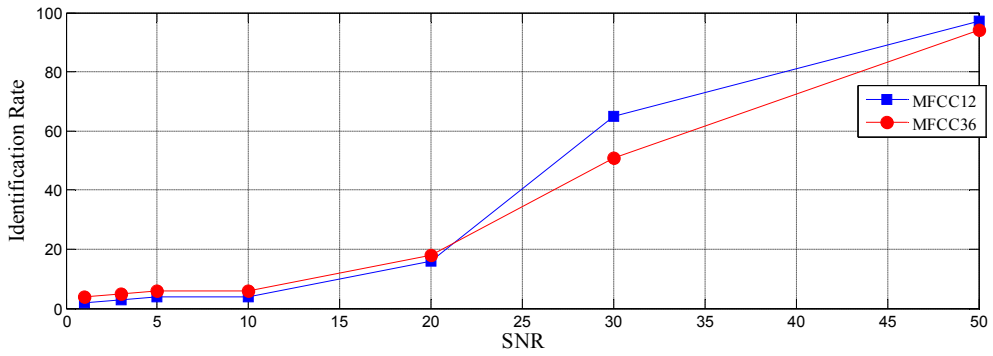
همانگونه که مشاهده می‌شود در این روش درصد شناسایی در بیشترین مقدار خود به ۹۹ درصد رسیده است که به نسبت روش MFCC بدون استفاده از پس پردازش، یک درصد افزایش داشته است.

۴-۳ انتخاب روش استخراج ویژگی برای پیاده‌سازی سخت افزاری

در بخش قبل تاثیر روش‌های استخراج ویژگی، پس پردازش، تعداد ترکیب‌های GMM و تعداد قاب‌های نمونه‌های آزمایش را بررسی کرده و نتایج شبیه سازی مربوط به هر کدام از روش‌ها ارائه شد. بیشترین درصد شناسایی مربوط به هنگامی بود که از ضریب انرژی، و ترکیبی از روش‌های پس پردازش استفاده شد. در نتایجی که ارائه شده، مشاهده می‌شود که اثر استفاده از روش‌های پس پردازش گاها به صورتی است که درصد شناسایی کاهش پیدا می‌کند، در صورتی که هدف استفاده از این روش‌ها بهبود ویژگی‌های استخراج شده می‌باشد. باید گفت که این روش‌ها برای بهبود کیفیت ویژگی‌ها در مقابل نویز و اعوجاج حاصل از انتقال استفاده می‌شوند و تاثیر خود را هنگامی نشان می‌دهند که سیگنال آزمایش نویزی شده باشد. برای بررسی اثرات نویز، با استفاده از پایگاه داده TIMIT مدل‌های GMM آموزش داده شدند و سپس داده‌های آزمایش را با استفاده از نرم افزار Matlab نویزی کرده و آزمایشات قبل دوباره برای این مدل‌ها و داده‌های جدید انجام شد. برای نمونه، در شکل (۴-۸) مقایسه‌ای بین دو روش استخراج ویژگی MFCC بدون پس پردازش و روش استخراج ویژگی MFCC همراه با هر سه روش پس پردازش CMN، دلتا و دلتا – دلتا وقتی که داده‌های آزمایش با نرخ‌های SNR مختلف نویزی شده‌اند، ارائه شده است.

همانگونه که در شکل نمایش داده شده است وقتی که اندازه SNR بزرگ باشد، یعنی در حدود ۵۰، نتایج شناسایی تقریبا با نتایج آزمایش در محیط بدون نویز برابر هستند. اما هر چقدر که اندازه SNR کوچک‌تر می‌شود، درصد شناسایی کاهش پیدا می‌کند. البته نکته قابل توجه این است که درصد شناسایی در روش MFCC همراه با پس پردازش، وقتی که اندازه SNR از ۲۲ کوچکتر می‌شود، نسبت به روش

MFCC بدون پس پردازش بیشتر می‌شود اما به هر حال وقتی که اندازه SNR کوچکتر می‌شود، نتایج به ازای هر دو روش استخراج ویژگی بسیار کاهش می‌یابد.



شکل(۴-۸): مقایسه روش‌های MFCC همراه با پس پردازش و بدون پس پردازش در محیط نویزی

برای آزمایش برخط مدار طراحی شده، باید یک پایگاه داده ساخته می‌شد. منظور از آزمایش برخط این است که با استفاده از میکروفن متصل شده به کدک سیگنال گفتار آزمایش را دریافت کرده، و به پردازنده سیگنال انتقال داد. آن‌گاه درون پردازنده ویژگی‌ها استخراج شده، و با مدل‌هایی که در حافظه قرار دارند مقایسه شده و مشخص شود که سیگنال ورودی به سیستم متعلق به کدام یک از گویندگان می‌باشد.

برای اینکه کارکرد مدار آزمایش شود، از روش استخراج ویژگی MFCC همراه با ضریب انرژی، پارامترهای دلتا، دلتا – دلتا و روش جبران سازی CMN استفاده شد. هنگامی که برنامه بر روی پردازنده DSP بارگذاری شد، جوابی از مدار حاصل نشد. در بررسی‌های که انجام شد، این نتیجه به دست آمد که به دلیل استفاده از پایگاه داده نویزی، که خودمان ساخته بودیم، و البته به دلیل نویز محیط بر روی نمونه‌های آزمایش، مدار با استفاده از این روش استخراج ویژگی و بدون استفاده از الگوریتم‌های رفع نویز قادر به کارکرد صحیح نمی‌باشد. برای بررسی اثر آموزش دادن مدل‌ها با استفاده از یک پایگاه داده نویزی، از پایگاه داده vidTIMIT استفاده شد. این پایگاه داده به تبعیت از پایگاه داده NTIMIT ساخته

شده است و می‌توان تاثیر نویز محیط را بر روی آموزش مدل‌ها با استفاده از آن بررسی کرد. نتایج آزمایش مدل‌ها با استفاده از این روش برای وقتی که از روش استخراج ویژگی MFCC همراه بـا ضـریب انـرژی و جبران سازی CMN و پارامترهای دلتا و دلتا – دلتا استفاده شده باشد، در جدول (۴-۹) ارائـه داده شـده است.

در این آزمایش تعداد گویندگان برای آزمایش ۱۰ نفر در نظر گرفته شده است و نمونه‌های آموزش دارای ۱۵۰ قاب می‌باشند. نمونه‌های آزمایش مستقیما از خود پایگاه داده استفاده شده و هیچ نـویزی بـه آن‌ها اضافه نگردیده است. مشاهده می‌شود که در بهترین حالت درصد شناسایی برابر با ۶۰ درصد می‌باشد که در مقایسه با زمانی که از پایگاه داده بدون نویز استفاده شده باشد، ۳۹ درصد کاهش دارد. در کنار این موضوع در هنگام کار با مدار، تاثیرنویز محیط بر روی نمونه‌های آزمایش نیز باعث شد که جـوابی از مـدار دریافت نشود.

جدول(۴-۹): بررسی تاثیر آموزش مدل‌های GMM با استفاده از داده‌های آموزشی نویزی

شماره آزمایش	تعداد ترکیب GMM	درصد شناسایی
آزمایش اول	۵	۲۵
آزمایش دوم	۱۰	۴۰
آزمایش سوم	۱۵	۵۰
آزمایش چهارم	۲۰	۶۰
آزمایش پنجم	۲۵	۴۵

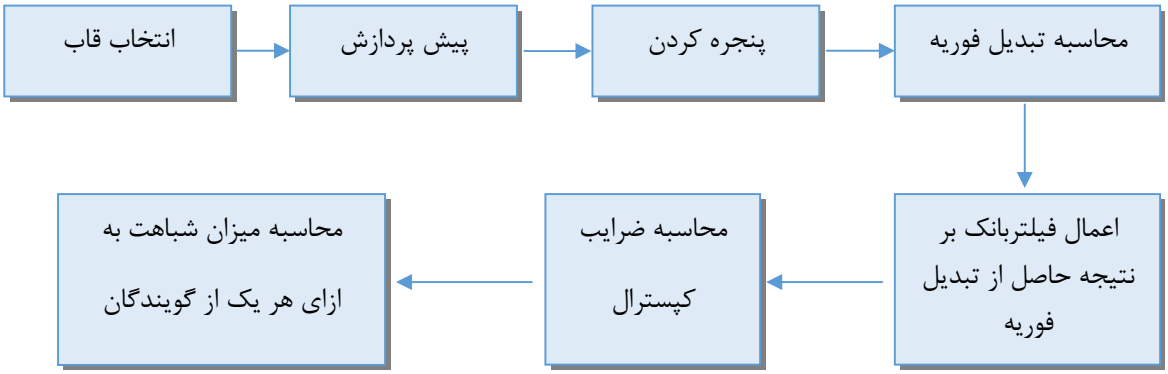
روش کار پردازنده DSP بر مبنای اعداد ممیز ثابت می‌باشد. به همین دلیـل بایـد اعـدادی کـه بـه

سیستم داده می‌شود ابتدا به اعداد صحیح نرمالیزه شده و سپس در DSP استفاده شوند. به این دلیل که پردازنده حجم حافظه محدودی در اختیار کاربر قرار می‌دهد، وقتی که اختلاف بین بزرگترین و کوچکترین عدد موجود در مجموعه ویژگی‌ها و یا مدل‌ها زیاد باشد، هنگام نرمالیزه کردن، مقدار زیادی از اطلاعات از دست می‌رود. در ویژگی‌های MFCC استخراج شده وقتی که بردارهای ویژگی دارای ۳۹ درایه می‌باشند و از ضریب انرژی در آنها استفاده شده است، به دلیل بزرگ بودن بیش از حد ضریب انرژی به نسبت اعداد دیگر، در هنگام نرمالیزه کردن ویژگی‌ها و انتقال آن‌ها به DSP اطلاعات زیادی از دست می-رود.

به دلیل نداشتن امکانات لازم برای تهیه پایگاده داده بدون نویز و البته توانایی پایین روش استخراج ویژگی MFCC در مقابل نویز، تصمیم بر این شد که از کدک استفاده نشود. برای مدل‌کردن گویندگان از پایگاه داده TIMIT استفاده شد و نمونه‌های آزمایش نیز بر روی پردازنده به صورت مستقیم بارگذاری شدند تا کمبود در اختیار نداشتن پایگاه داده بدون نویز جبران شود. برای استخراج ویژگی نیز از روش MFCC دارای بردارهای ویژگی ۱۲ درایه‌ای بدون استفاده از ضریب انرژی و هیچ نوع پس پردازشی استفاده شد که در محیط بدون نویز با بار محاسباتی و اشغال حجم حافظه کمتر، بیشترین درصد شناسایی را نسبت به سایر روش‌ها به دست می‌دهد.

۴-۴ بررسی پیاده سازی سخت‌افزاری الگوریتم

همانگونه که در قسمت قبل ذکر شد، در آزمایشی که بر روی مدار انجام شده، از کدک استفاده نشده است و سیگنال آزمایش به همراه مدل‌های گویندگان و کد نوشته شده برای الگوریتم بر روی پردازنده بارگذاری شده است. نحوه کارکرد الگوریتم در شکل (۴-۸) نمایش داده شده است.



شکل(۴-۹): نمودار بلوکی الگوریتم پیاده‌سازی شده بر روی پردازنده DSP

برای بررسی اینکه این مدار با استفاده از الگوریتمی که انتخاب شده است توانایی اجرای برخط شناسایی گوینده را دارد، برنامه الگوریتم به صورتی نوشته شد که بتواند فرایند دریافت و پردازش برخط را شبیه سازی کند. همانگونه که اشاره شد، سیگنال آزمایش همراه با برنامه بر روی پردازنده بارگذاری شده است. روند اجرای برنامه بر طبق شکل (۴-۸) می‌باشد. به این ترتیب که ابتدا از سیگنال آزمایش یک قاب انتخاب می‌شود و سپس این قاب به مرحله پیش‌پردازش رفته و پیش پردازش بر روی آن انجام می‌شود. در ادامه تبدیل فوریه این قاب محاسبه شده و نتیجه به قسمت فیلتر بانک انتقال پیدا می‌کند. در قسمت فیلتر بانک با استفاده از فیلتر بانک مِل، ضرایب خلاصه‌تر شده و تعداد آنها کمتر می‌شود. در ادامه، نتیجه خروجی فیلتر بانک به قسمت محاسبه ضرایب کپسترال انتقال می‌یابد. در این قسمت با استفاده از تبدیل کسینوسی گسسته، ضرایب کپسترال محاسبه می‌شود. وقتی ضرایب کپسترال محاسبه شود، فرایند استخراج ویژگی از قاب تمام می‌شود. در نهایت در قسمت آزمایش، با استفاده از معادله (۲-۳۰)، میزان شباهت[1] برای قاب مورد نظر به صورت جداگانه با هر یک از گویندگان موجود در پایگاه داده محاسبه

[1] Likelihood

می‌شود. بعد از اتمام کل فرایند استخراج ویژگی و محاسبه شباهت برای قاب اول، سیستم قاب بعدی را از سیگنال آزمایش انتخاب کرده و همین مراحل را بـر روی آن اجـرا مـی‌کنـد و در نهایـت نتـایج شـباهت محاسبه شده برای هر گوینده را با نتایج شباهت محاسبه شده به ازای قاب قبلی جمع می‌کند. این کار به همین صورت ادامه پیدا می‌کند تا زمانی که تعداد قاب‌ها به عددی برسد کـه از قبـل معـین شـده اسـت. وقتی که تعداد قاب‌ها به اندازه از پیش تعیین شده رسید، سیستم ایـن فراینـد را بـه اتمـام مـی‌رسـاند و مجموع شباهت‌های محاسبه شده به ازای هر گوینده را در نظر می‌گیرد. هر کدام از گویندگان کـه امتیـاز بیشتری کسب کرده باشد سیگنال ورودی به سیستم متعلق به همان گوینده می‌باشد. در اینجا منظـور از امتیاز، میزان حاصل جمع شباهت‌های هر گوینده به ازای تمام قاب‌های ورودی به سیستم می‌باشد.

در آزمایشی که بر روی سیستم انجام شد، از ۷ گوینده به عنوان پایگاه داده استفاده شـده اسـت و مدل GMM مربوط به هر گوینده دارای ۱۵ ترکیب مـی‌باشـد. بـرای اسـتخراج ویژگـی از روش MFCC بدون هیچ نوع پس پردازش استفاده شده و بردارهای ویژگی دارای ۱۲ درایه می‌باشـند. تعـداد قـاب‌هـای سیگنال آزمایش برابر ۱۵۰ در نظر گرفته است. نتیجه تعداد دوره‌های ساعت[1] برای هر کدام از قسمت‌های نمایش داده شده در شکل(۴-۸) به ازای آزمایش انجام داده شده، در جدول (۴-۱۰) ارائه شده است.

[1] Clock cycles

جدول(۴-۱۰): زمان اجرای سخت‌افزاری بخش‌های مختلف الگوریتم

مرحله الگوریتم	تعداد دوره‌های ساعت	زمان به میلی‌ثانیه
پیش پردازش	۴۳۴۱	۰/۰۳
پنجره کرده	۲۷۲۸۴	۰/۱۹
تبدیل فوریه	۱۳۵۴۸۸	۰/۹۴
فیلتر بانک	۶۷۷۶	۰/۰۴۷
ضرایب کپسترال	۴۹۳۲	۰/۰۳۴
محاسبه شباهت	۴۱۴۸۷۳	۲/۸۸

مقادیری که در جدول (۴-۱۰) وارد شده است، زمان به ازای یک قاب می‌باشد و برای محاسبه زمان کل باید مجموع این مقادیر را در ۱۵۰، یعنی تعداد قاب‌ها، ضرب کرد. این نتیجه در جدول (۲-۱۱) نمایش داده شده است.

جدول(۴-۱۱): زمان اجرای الگوریتم

نوع آزمایش	تعداد دوره‌های ساعت	زمان به میلی‌ثانیه
آزمایش به ازای یک قاب	۵۹۳۶۹۴	۴/۱۲۱
آزمایش به ازای ۱۵۰ قاب	۸۹۰۵۴۱۰۰	۶۱۸/۴۳۱

همانگونه که در آزمایش مشاهده می‌شود، زمان کل آزمایش حدود ۶۱۸ میلی ثانیه می‌باشد. در این

آزمایش از ۱۵۰ قاب استفاده شده است و طول سیگنال آزمایش حدود ۱/۵ ثانیه می‌باشد. با توجه به این نتایج می‌توان گفت که مدار طراحی شده با این پردازنده و الگوریتم استفاده شده، توانایی استفاده به صورت برخط برای شناسایی گوینده را دارد.

در ادامه در مورد دقت سیستم شناسایی گوینده طراحی شده باید گفت که به دلیل استفاده از متغییرهای ممیز ثابت در پردازنده DSP، مقداری از اطلاعات در حین هر عمل ریاضی از بین می‌رود و چون میزان محاسبات در این الگوریتم بسیار بالا می‌باشد، متغییرهای ممیز ثابت تاثیر زیادی بر کاهش درصد شناسایی گوینده خواهند گذاشت. در این آزمایش به ازای استفاده از مدل‌های GMM با تعداد ترکیب ۱۵ و روش استخراج ویژگی MFCC با ۱۲ درایه و بدون پس پردازش، بهترین نتیجه‌ای که حاصل شد برابر با ۷۸/۶ درصد بود که به نسبت نتایج شبیه‌سازی که در نرم‌افزار Matlab انجام شد، حدود ۱۹ درصد کاهش داشت.

جدول(۴-۱۲): مقایسه درصد شناسایی در Matlab با پیاده‌سازی سخت افزاری

نوع آزمایش	نرم‌افزار Matlab	نتیجه سخت‌افزار
درصد شناسایی	۹۸	۷۸/۶

برای بهبود درصد شناسایی می‌توان از یک حافظه جانبی استفاده کرد. استفاده از حافظه جانبی، با افزایش میزان حافظه در دسترس، این امکان را به کاربر می‌دهد که متغییرها را با دقت بیشتری به حالت ممیز ثابت انتقال دهد و با این کار میزان اطلاعاتی که در حین محاسبات از بین می‌رود، کاهش می‌یابد.

فصل پنجم

جمع‌بندی و ارائه پیشنهادات برای کارهای آینده

۵-۱ جمع‌بندی و نتیجه گیری

در این پایان نامه ابتدا مروری بر انواع روش‌ها شناسایی گوینده شامل روش‌های استخراج ویژگی، مدل کردن و کلاسه‌بندی انجام شد. از میان روش‌های استخراج ویژگی، روش‌های طیفی زمان کوتاه به نسبت دیگر روش‌ها برای پیاده سازی سخت افزاری مناسب‌تر بوده و قابلیت پیاده‌سازی برخط را نیز دارند. برای مدل کردن گوینده از روش مدل کردن با استفاده از مدل آمیخته گاوسی استفاده شد.

در ادامه برای بررسی روش‌های انتخابی، آزمایش‌هایی انجام شد که بر اساس آن‌ها می‌توان به نتایج زیر اشاره کرد:

- روش استخراج ویژگی MFCC به نسبت روش استخراج ویژگی LPCC برای کاربردهای شناسایی گوینده مناسب‌تر بوده و درصد شناسایی بالاتری ارائه می‌دهد.
- روش MFCC به تنهایی و بدون استفاده از هیچ نوع پس پردازشی، در محیط‌های بدون نویز تا ۹۸ درصد گوینده‌ها را به صورت صحیح شناسایی می‌کند.
- از میان روش‌های پس پردازش، استفاده از ضرایب دلتا و دلتا - دلتا با همدیگر، نتیجه بهتری به نسبت وقتی که از ضرایب دلتا به تنهایی استفاده شود، می‌دهد.
- می‌توان با ترکیب مناسب روش‌های پس پردازش در محیط بدون نویز، تا ۹۹ درصد کارایی سیستم شناسایی گوینده بر پایه GMM را افزایش داد.
- استفاده از روش‌های پس پردازش، خصوصا روش CMN، برای محیط‌های نویزی توصیه می‌شود. هنگامی که محیط بدون نویز باشد، استفاده از روش‌های پس پردازش تاثیر چندانی در بهبود کارایی سیستم شناسایی گوینده ندارد و علاوه بر این باعث افزایش بار محاسباتی نیز می‌شوند.

- روش استخراج ویژگی MFCC در مقابل نویز بسیار ضعیف بوده و باید برای کاربردهای در محیط‌های نویزی از سایر روش‌های استخراج ویژگی استفاده کرد. در صورت استفاده از روش استخراج ویژگی MFCC، باید از الگوریتم‌های رفع نویز نیز همراه با آن در سیستم استفاده شود.
- در استفاده از روش GMM برای مدل کردن گوینده، الزاما نیازی به GMM با ترکیب بالا نمی‌باشد و بسته به این که چه میزان داده آموزشی در اختیار باشد، باید تعداد ترکیب متناسب با آن داده‌ها مشخص شود.
- روش‌های بر پایه GMM و روش استخراج ویژگی طیفی زمان کوتاه برای کاربردهای برخط مناسب می‌باشند.

بعد از اتمام آزمایش‌ها در محیط نرم افزار Matlab، مداری بر پایه پردازنده سیگنال DSP طراحی شد و نتایج شبیه‌سازی‌ها بر روی آن برسی شد. همانگونه که ذکر شد، نتایج نشان دادند که روش انتخابی برای پیاده‌سازی سخت افزاری مناسب بوده و به صورت برخط قابل اجرا می‌باشد. به دلیل در اختیار نداشتن امکانات لازم برای تهیه پایگاه داده بدون نویز، از کدک این مدار استفاده نشد و داده‌های آزمایش به صورت مستقیم بر روی پردازنده بارگذاری شدند. بهترین درصد شناسایی به دست آمده از مدار به ازای بردارهای ویژگی MFCC بدون پس پردازش با ۱۲ درایه و مدل‌های GMM با تعداد ۱۵ترکیب، برابر ۷۸/۶ درصد بود.

۵-۲ پیشنهاد برای کارهای آینده

در این پژوهش مروری بر روش‌هاش شناسایی گوینده به منظور پیاده‌سازی سخت افـزاری سیسـتم شناسایی گوینده انجام شد و مداری نیز طراحی شده و نتایج بـر روی آن بررسـی شـد. بـرای ادامـه ایـن پژوهش در آینده می‌توان به موارد زیر اشاره کرد:

- یافتن یک روش استخراج ویژگی مناسب برای بهبود کارایی سیستم در مقابل نویز
- تهیه امکانات لازم برای ساخت یک پایگاه داده که مناسب استفاده در کابردهای شناسـایی گوینده باشد. پایگاه داده باید بدون نویز باشد و به میزان لازم برای آزمایش انواع روش‌های شناسایی گوینده، داده‌های آموزشی داشته باشد.
- تهیه قطعات سخت افزاری با کیفیت بالا، مانند میکروفن‌های مناسب برای استفاده در مدار پردازش سیگنال
- استفاده از حافظه جانبی برای کمک به بهبود درصد شناسایی
- بهینه کردن برنامه الگوریتم نوشته شده برای کاهش محاسـبات غیـر ضـرروی بـه منظـور کاهش زمان اجرای برنامه و البته افزایش میزان درصد شناسایی گوینده

پیوست

مروری بر الگوریتم EM

در این پیوست به صورت خلاصه در مورد الگوریتم EM و نحوه کارکرد آن صحبت شده و روش استفاده از این الگوریتم برای تخمین پارامترهای یک مدل آمیخته گاوسی شرح داده شده است.

۶-۱ بیشینه شباهت[1]

برای بررسی مفهوم بیشینه شباهت فرض کنید که یک تابع چگالی احتمال به صورت $p(X|\Theta)$ داریم (برای مثال p می‌تواند ترکیب تعدادی توزیع گاوسی باشد که به صورت چگالی آمیخته گاوسی یا GMM شناخته می‌شود و پارامتر Θ می‌تواند مجموعه‌ای از بردارهای میانگین و ماتریس‌های کواریانس باشد). اگر یک مجموعه داده به صورت $X = \{x_1, x_2, \ldots, x_N\}$ داشته باشیم و فرض کنیم که داده‌ها نسبت به همدیگر مستقل باشند، آنگاه توزیع احتمال برای داده‌های X به صورت معادله (۶-۱) نوشته می‌شود[47]:

$$p(X|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta) = \mathcal{L}(\Theta|X) \qquad (۱-۶)$$

در این معادله، تابع $\mathcal{L}(\Theta|X)$ را شباهت[2] پارامترهای Θ و داده‌های X می‌گویند. در محاسبه تابع شباهت، داده‌های X ثابت و پارامترهای Θ متغیر می‌باشند. در محاسبه بیشینه شباهت، هدف این است که پارامترهای Θ به گونه‌ای تعیین شوند که $\mathcal{L}$ ماکزیمم شود. معمولا برای سادگی محاسباتی، ماکزیمم $\log(\mathcal{L}(\Theta|X))$ را به جای $\mathcal{L}(\Theta|X)$ محاسبه می‌کنند. بستگی به نوع توزیع احتمالاتی $p(X|\Theta)$، این فرایند می‌تواند ساده و یا پیچیده باشد. برای مثال وقتی که تابع $p(X|\Theta)$ یک توزیع گاوسی ساده باشد، تنها نیاز به محاسبه یک واریانس و میانگین می‌باشد، یعنی باید فقط $\Theta = (\mu, \delta^2)$ می‌باشد. اما وقتی که توزیع $p(X|\Theta)$ یک توزیع آمیخته گاوسی باشد، آنگاه باید به تعداد ترکیب‌های توزیع آمیخته گاوسی،

[1] Maximum Likelihood
[2] Likelihood

میانگین و ماتریس کواریانس محاسبه کرد که به نسبت حالت قبل بسیار پیچیده‌تر می‌باشد.

۶-۲ بیشینه سازی امید ریاضی[1]

الگوریتم EM یک روش عمومی برای پیدا کردن بیشینه شباهت پارامترهای یک توزیع بر مبنای یک مجموعه داده می‌باشد، وقتی که تعدادی از داده‌ها ناکامل و یا ناقص باشند. از الگوریتم EM می‌توان به دو صورت استفاده کرد. حالت اول زمانی است که واقعا به دلیل محدویت‌های مشاهدات، تعدادی از داده‌ها ناقص باشند. حالت دوم زمانی است که به دست آوردن پارامترهای تابع شباهت به صورت تحلیلی مشکل بوده ولی می‌توان با فرض وجود داده‌های مخفی و ساده سازی تابع شباهت، تابع شباهت را تخمین زد. در کاربردهای شناسایی الگو، حالت دوم بیشتر کاربرد دارد[47].

مانند قبل، فرض می‌کنیم که X داده‌های مشاهده شده بوده و توسط یک توزیع احتمال تولید شده است. می‌توان X را داده‌های ناقص نامگذاری کرد. در ادامه فرض می‌شود که یک مجموعه داده کامل $Z = (X, Y)$ وجود داشته و تابع احتمال متصل (۶-۲) برقرار است.

$$p(Z|\Theta) = p(X, Y \mid \Theta) = p(Y|X, \Theta)\, p(X|\Theta) \qquad (۶-۲)$$

اکنون با استفاده از این تابع احتمال جدید می‌توان یک تابع شباهت جدید تعریف کرد که این تابع را تابع شباهت کامل می‌نامیم و به صورت (۶-۳) می‌باشد.

$$\mathcal{L}(\Theta|Z) = \mathcal{L}(\Theta|X, Y) = p(X, Y| \Theta) \qquad (۶-۳)$$

الگوریتم EM ابتدا لگاریتم امید تابع شباهت اطلاعات کامل، یعنی $\log p(X, y|\Theta)$ را با توجه به داده‌های مخفی Y، داده‌های مشاهده شده X و پارامترهای حاضر محاسبه می‌کند. این امید به صورت (۶-۴) تعریف می‌شود[47].

[1] Expectation Maximization

$$Q(\Theta, \Theta^{i-1}) = E[\log\, p(X, Y|\, \Theta)\, |X, \Theta^{i-1}] \qquad (۴-۶)$$

در این معادله Θ^{i-1} عبارت است از پارامترهای حاضر که با استفاده از آن‌ها مقدار امید را محاسبه می‌کنند و Θ پارامترهایی هستند که قرار است با بهینه کردن آنها، اندازه Q افزایش پیدا کند. علاوه بر این پارامترهای Θ^{i-1} و X ثابت بوده و Y یک متغییر تصادفی می‌باشد.

الگوریتم EM دو مرحله دارد. اولین مرحله مرحله محاسبه امید، یا $E - Step$، می‌باشد که باید معادله (۶-۴) حل شود. مرحله دوم به صورت ماکزیمم کردن امیدی می‌باشد که در مرحله اول محاسبه شده است. مرحله دوم به صورت (۶-۵) می‌باشد[47].

$$\Theta^{i} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta, \Theta^{i-1}) \qquad (۵-۶)$$

این دو مرحله تا زمانی پارامترها با دقتی مناسبی تعیین شوند، به صورت مداوم و پشت سر هم اجرا می‌شوند. هر دوره تکرار تضمین می‌کند که اندازه لگاریتم شباهت افزایش پیدا کند و ساختار الگوریتم تضمین می‌کند که تابع شباهت به یک ماکزیمم محلی همگرا می‌شود[47, 50].

۶-۲-۱ محاسبه پارامترهای تابع شباهت برای یک ترکیب آمیخته با استفاده از الگوریتم EM

محاسبه پارامترهای یک چگالی آمیخته احتمالا یکی از پرکاربردترین استفاده‌ها از الگوریتم EM می‌باشد. در این حالت چگالی آمیخته به صورت معادله (۶-۶) تعریف می‌شود[47].

$$p(X|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(X|\theta_i) \qquad (۶-۶)$$

در این معادله $\Theta = \{\alpha_1, \ldots., \alpha_M, \theta_1, \ldots\ldots, \theta_M\}$ است و شرط $\sum_{i=1}^{M} \alpha_i = 1$ باید برقرار باشد. علاوه بر این هر p_i یک چگالی احتمال است که پارامتر مربوط به آن θ_i می‌باشد. به عبارت دیگر فرض

شده است که تابع چگالی آمیخته از ترکیب M مولفه تشکیل شده که با ضرایب α_i با هم ترکیب شده‌اند.

لگاریتم تابع شباهت مربوط به داده‌های ناکامل به صورت (۶-۷) می‌باشد که البته برای محاسبه پیچیده است[47].

$$\log \mathcal{L}(\Theta|X) = \log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{M} \alpha_j p_j(X_i|\theta_j) \right) \quad (۶-۷)$$

به دلیل پیچیده بودن محاسبات معادله (۶-۷)، از لگاریتم تابع شباهت داده‌های کامل با فرض وجود داده‌های مشاهده نشده $Y = \{y_i\}^N_{i=1}$ ، استفاده می‌شود. لگاریتم تابع شباهت داده‌های کامل در معادله (۶-۸) نمایش داده شده است[47].

$$\log \mathcal{L}(\Theta|X,Y) = \log\big(p(X,Y|\Theta)\big) \quad (۶-۸)$$
$$= \sum_{i=1}^{N} \log\big(P(x_i|y_i)\, P(y)\big) = \sum_{i=1}^{N} \log\left(\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})\right)$$

برای حل معادله (۶-۸) ابتدا باید توزیع احتمال داده‌های مخفی را معین کرد. در اولین مرحله باید برای پارامترهای مدل یک مقدار اولیه تعیین شود. این مقادیر اولیه به صورت $\Theta^g = \{\alpha^g_1, \ldots, \alpha^g_M, \theta^g_1, \ldots, \theta^g_M\}$ نمایش داده می‌شود. وقتی که مقادیر Θ^g مشخص باشد، می‌توان به راحتی $p_j(X_i|\theta^g{}_j)$ را به ازای تمام i و j ها محاسبه کرد. علاوه بر این می‌توان فرض کرد که پارامترهای ترکیب، یعنی α_jها، به صورت احتمال‌های پیشین هر ترکیب می‌باشند، یعنی $\alpha_j = p\left(\text{مولفه } j\right)$ به عنوان احتمال پیشین فرض شده است. بنابراین با استفاده از قانون بِیز داریم:

$$p(y_i|x_i,\Theta^g) = \frac{\alpha^g_{y_i}\, p_{y_i}(x_i|\theta^g_{y_i})}{p(x_i|\Theta^g)} = \frac{\alpha^g_{y_i}\, p_{y_i}(x_i|\theta^g_{y_i})}{\sum_{k=1}^{M} \alpha^g_k p_k(x_i|\theta^g_k)} \quad (۶-۹)$$

و برای تمام مجموعه داده‌های Y داریم:

$$p(Y|X,\Theta^g) = \prod_{i=1}^{N} p(y_i|x_i,\Theta^g) \qquad (۱۰-۶)$$

در معادله (۶-۱۰) $Y = \{y_1, \dots, y_N\}$ فرض شده است. حال کـه توزیـع احتمـال داده‌هـای مخفـی تعریف شد، می‌توان معادله (۶-۴) را به صورت زیر بازنویسی کرد[47]:

$$Q(\Theta,\Theta^g) = \sum_{y\,\epsilon\,\Psi} \log \mathcal{L}(\Theta|X,Y)\ p(Y|X,\Theta^g) \qquad (۱۱-۶)$$

در معادله (۶-۱۱) نماد Ψ نشان دهنده مجموعه‌ای می‌باشد که y می‌تواند عضو آن باشد.

برای کاربرد مد نظر ما در این پایان نامه، یعنی مدل کردن گوینده با استفاده از یک مـدل آمیختـه گاوسی، نیاز است تا هر کدام از توزیع‌های تشکیل دهنده مدل آمیختـه، یـک توزیـع احتمـال گاوسـی d متغییره باشد که با استفاده از یک میانگین و ماتریس کواریانس مشخص می‌شـود. تـابع احتمـالاتی یـک توزیع گاوسی d متغیره با بردار میـانگین μ و مـاتریس کواریـانس Σ زمـانی کـه فـرض شـود ایـن تـابع احتمالاتی، ترکیب l ام از یک مدل آمیخته گاوسی باشد، به وسیله معادله (۶-۱۲) نمایش داده می‌شود.

$$p_{\,l}(x|\mu_l,\sum_l) = \frac{1}{(2\pi)^{d/2}|\sum_l|^{1/2}}\ e^{-0.5\,(x-\,\mu_l)^T\Sigma_l^{-1}(x-\,\mu_l)} \qquad (۱۲-۶)$$

بنابراین برای هر مدل آمیخته گاوسی با M ترکیب، به تعداد M بردار میانگین d درایه‌ای و تعـداد M ماتریس کواریانس با ابعاد $M \times M$ نیاز داریم. می‌توان نشان داد که با حل معادله (۶-۱۱)، مقادیر ایـن پارامترها را می‌توان به دست آورد. نحوه محاسبه این پارامترها در معادلات (۶-۱۳) تا (۶-۱۵) نمایش داده شده است. این معادلات هر دو مرحله الگوریتم EM یعنی مرحله تخمین امید ریاضـی و بیشـینه سـازی امید را به صورت همزمان انجام می‌دهند. برای اطلاع در مورد نحوه حل و ساده‌سازی معادله (۶-۱۱) می‌توان به مرجع [47] مراجعه کرد.

$$\alpha_l = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g) \qquad (۱۳-۶)$$

$$\mu_l = \frac{\sum_{i=1}^{N} x_i \, p(l|x_i, \Theta^g)}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)} \qquad (۱۴-۶)$$

$$\sum_l = \frac{\sum_{i=1}^{N} p(l|x_i, \Theta^g)\,(x_i - \mu_l)(x_i - \mu_l)^T}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)} \qquad (۱۵-۶)$$

به منظور تخمین زدن پارامترهای یک مدل آمیخته گاوسی برای یک گوینده، ابتدا ویژگی‌های داده‌های آموزش را استخراج می‌کنیم. در ادامه به صورت تصادفی پارامترهای مدل آمیخته گاوسی را مقداردهی کرده و با استفاده از این پارامترها و بردارهای ویژگی و معادلات (۶-۱۳) تا (۶-۱۵)، مقادیر جدید را برای پارامترهای مدل تخمین می‌زنیم. در دوره تکرار دوم، از پارامترهایی که در مرحله اول محاسبه شده است استفاده کرده، و دوباره پارامترهای مدل را با استفاده از معادلات (۶-۱۳) تا (۶-۱۵) تخمین می‌زنیم. در دوره‌های تکرار بعدی نیز همین مراحل را انجام داده تا زمانی که پارامترها با دقتی که لازم داریم تخمین زده شوند[50].

مراجع

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE,* vol. 11, pp. 23-61, 2011.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication,* vol. 52, pp. 12-40, 2010.

[3] M.-W. Mak and W. Rao, "Utterance partitioning with acoustic vector resampling for GMM–SVM speaker verification," *Speech Communication,* vol. 53, pp. 119-130, 2011.

[4] I. Trabelsi and D. Ben Ayed, "On the use of different feature extraction methods for linear and non linear kernels," in *Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on*, 2012, pp. 797-802.

[5] M. P. Kesarkar, "Feature extraction for speech recognition," in *M. Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay*, 2003, pp. 1-11.

[6] V. S. Baidwan and S. Gujral, "Comparative Analysis of Prosodic Features and Linear Predictive Coefficients for Speaker Recognition Using Machine Learning Technique," in *Devices, Circuits and Communications (ICDCCom),* 2014, pp. 1-8.

[7] M. Nilsson and M. Ejnarsson, "Speech recognition using hidden markov model," 2002.

[8] P. Qi and L. Wang, "Experiments of GMM based speaker identification," in *Ubiquitous Robots and Ambient Intelligence (URAI), 2011 8th International Conference on*, 2011, pp. 26-31.

[9] S. N. A., "Text-Independent Speaker Identification using Vector Quantization," *International Journal of Engineering Research & Technology (IJERT),* vol. 2, pp. 1787 - 1792, 2013.

[10] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *arXiv preprint arXiv:1305.1145,* 2013.

[11] F.-L. Huang, "A Novel Method for Speaker Independent Recognition Based on Hidden Markov Model," 2011.

[12] B. Singh, R. Kaur, N. Devgun, and R. Kaur, "The Process Of Feature Extraction In Automatic Speech Recognition System For Computer Machine Interaction With Humans: A Review," *International Journal Of Advanced Research In Computer Science And Software Engineering, ISSN,* vol. 2277, 2012.

[13] S. Shanthi Therese and C. Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition," *International Journal of Scientific Engineering and Technology,* vol. 2, pp. 479-484, 2013.

[14] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative Modeling of Voice Fundamental Frequency Contours," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* vol. 23, pp. 1042-1053, 2015.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

[16] A. S. Naini, M. M. Homayounpour, and A. Samani, "A real-time trained system for

robust speaker verification using relative space of anchor models," *Computer Speech & Language,* vol. 24, pp. 545-561, 2010.
[17] T. Kinnunen, I. Sidoroff, M. Tuononen, and P. Fränti, "Comparison of clustering methods: A case study of text-independent speaker modeling," *Pattern Recognition Letters,* vol. 32, pp. 1604-1617, 2011.
[18] W.-T. Hong, "HCRF-UBM approach for text-independent speaker identification," *Computers & Mathematics with Applications,* vol. 64, pp. 1120-1127, 2012.
[19] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *International Journal for Advance Research in Engineering and Technology,* vol. 1, pp. 1-4, 2013.
[20] A. F. Pour, M. Asgari, and M. R. Hasanabadi, "Gammatonegram based speaker identification," in *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*, 2014, pp. 52-55.
[21] S. O. Sadjadi and J. H. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Communication,* 2015.
[22] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, "Discrete-Time Signal Processing," 1999.
[23] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Computer speech & language,* vol. 17, pp. 381-402, 2003.
[24] I. Mporas, T. Ganchev, M. Siafarikas, and N. Fakotakis, "Comparison of speech features on the speech recognition task," *Journal of Computer Science,* vol. 3, pp. 608-616, 2007.
[25] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*: Prentice Hall, 2010.
[26] E. B. Tazi, A. Benabbou, and M. Harti, "Efficient text independent speaker identification based on GFCC and CMN methods," in *Multimedia Computing and Systems (ICMCS), 2012 International Conference on*, 2012, pp. 90-95.
[27] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7204-7208.
[28] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*: Prentice Hall PTR, 2001.
[29] J. Hai and E. M. Joo, "Improved linear predictive coding method for speech recognition," in *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, 2003, pp. 1614-1618.
[30] S. S. Nidhyananthan and R. S. S. Kumari, "Text independent voice based students attendance system under noisy environment using RASTA-MFCC feature," in *Communication and Network Technologies (ICCNT), 2014 International Conference on*, 2014, pp. 182-187.
[31] K. Y. Lee, "Local fuzzy PCA based GMM with dimension reduction on speaker identification," *Pattern recognition letters,* vol. 25, pp. 1811-1817, 2004.
[32] S. Chakroborty, A. Roy, and G. Saha, "Fusion of a complementary feature set with

MFCC for improved closed set text-independent speaker identification," in *Industrial Technology, 2006. ICIT 2006. IEEE International Conference on*, 2006, pp. 387-390.

[33] M. Sinith, A. Salim, K. Gowri Sankar, K. Sandeep Narayanan, and V. Soman, "A novel method for Text-Independent speaker identification using MFCC and GMM," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, 2010, pp. 292-296.

[34] N. P. Jawarkar, R. S. Holambe, and T. K. Basu, "On the use of classifiers for text-independent speaker identification," in *Automation, Control, Energy and Systems (ACES), 2014 First International Conference on*, 2014, pp. 1-6.

[35] K. S. Ahmad, A. S. Thosar, J. H. Nirmal, and V. S. Pande, "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network," in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*, 2015, pp. 1-6.

[36] B. A. Sonkamble and D. Doye, "Speech Recognition Using Vector Quantization through Modified K-meansLBG Algorithm," *Computer Engineering and Intelligent Systems,* vol. 3, pp. 137-144, 2012.

[37] A. M. Ahmad and L. M. Yee, "Vector quantization decision function for Gaussian Mixture Model based speaker identification," in *Intelligent Signal Processing and Communications Systems, 2008. ISPACS 2008. International Symposium on*, 2009, pp. 1-4.

[38] M. K. Gill, R. Kaur, and J. Kaur, "Vector quantization based speaker identification," *International Journal of Computer Applications,* vol. 4, pp. 0975-8887, 2010.

[39] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE,* vol. 77, pp. 257-286, 1989.

[40] I. Shahin, "Speaker identification in shouted talking environments based on novel Third-Order Hidden Markov Models," in *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*, 2014, pp. 352-357.

[41] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and trends in signal processing,* vol. 1, pp. 195-304, 2008.

[42] H. S. Mohammadi and R. Saeidi, "Speaker identification performance enhancement using Gaussian mixture model with GMM classification post-processor," in *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*, 2007, pp. 504-507.

[43] D. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on,* vol. 3, pp. 72-83, 1995.

[44] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech communication,* vol. 17, pp. 91-108, 1995.

[45] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing,* vol. 10, pp. 19-41, 2000.

[46] C. Zeng and Z. Li, "Application of GMM in the Speaker Identification System," in *2011 7th International Conference on Wireless Communications, Networking and Mobile Computing*, 2011.

[47] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," 1998.
[48] M. A. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 381-396, 2002.
[49] R. Duda, E. Peter, and D. Stork, "Pattern Classification," 2001.
[50] M. A. Figueiredo, "Lecture Notes on the EM Algorithm," 2004.
[51] T. Kinnunen, J. Saastamoinen, V. Hautamäki, M. Vinni, and P. Fränti, "Comparative evaluation of maximum a posteriori vector quantization and Gaussian mixture models in speaker verification," *Pattern Recognition Letters,* vol. 30, pp. 341-347, 2009.
[52] T. Pekhovsky and A. Sizov, "Comparison between supervised and unsupervised learning of probabilistic linear discriminant analysis mixture models for speaker verification," *Pattern Recognition Letters,* vol. 34, pp. 1307-1313, 2013.
[53] R. Zheng, S. Zhang, and B. Xu, "Text-independent speaker identification using GMM-UBM and frame level likelihood normalization," in *Chinese Spoken Language Processing, 2004 International Symposium on*, 2004, pp. 289-292.
[54] W. Chen, Q. Hong, and X. Li, "GMM-UBM for text-dependent speaker recognition," in *Audio, Language and Image Processing (ICALIP), 2012 International Conference on*, 2012, pp. 432-435.
[55] K. Matsumoto, N. Hayasaka, and Y. Iiguni, "Noise robust speaker identification by dividing MFCC," in *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*, 2014, pp. 652-655.
[56] W.-Q. Zhang and J. Liu, *Discriminative Universal Background Model Training for Speaker Recognition*.
[57] H. Tolba, "A high-perfor mance text-ind ependent speaker identification of arabic speakers using a CHHM-based approch," *Alexandria Engineering Journal,* pp. 43 - 47, 2011.
[۵۸] ک. شفاعی, مرجع کامل پردازنده های سری ۲۰۰۰ و ۵۰۰۰ و ۶۰۰۰. تهران: انتشارات آستان قدس, ۱۳۸۹.
[59] T. Instruments, "TMS320C5509A Data Manual," ed, 2008.
[60] T. Instruments, "TPS767D301-EP Data Sheet," ed, 2010.
[61] T. Instruments, "TLV320AIC23B Data Manual," ed, 2004.
[62] "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT)," ed, updated 10/12/90.
[63] "VidTIMIT Dataset Documentation," ed, 2009.
[64] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Advances in Biometrics*, ed: Springer, 2009, pp. 199-208.

Abstract

Speaker identification is one of most important branches of speech recognition which has many applications in speech-based secure systems. In last decades, many efforts have been done in order to improve performance of such systems. Much of these efforts have focus on improving recognition rate and have not paid much attention on other important parameter such as hardware implementation and being online. This thesis concerns to online implementation of a speaker identification algorithm on TMS320C5509A DSP processor platform.

After intensive investigation, we choose spectral short time methos, especially MFCC and LPCC for feature extraction and Gaussian Mixture Model (GMM) for speaker modeling.

Simulation results in MATLAB show that MFCC method has higher recognition rate than that of LPCC. Although MFCC has 99% recognition rate in noiseless conditions, but in presence of noise it degrades dramatically. In addition, learning GMM models in noisy conditions decreases the recognition rate noticeably.

Because of sensitivity of MFCC against noise, we used TIMIT noiseless database. In addition, we had not capability to generate a noisy standard. In hardware experiments, we directly loaded the program on DSP processor using JTAG cable. Results show that the algorithm can be performed online. Maximum recognition rate was obtained as 78.6%.

Keywords:

Speaker Identification, MFCC Feature Extraction, LPCC Feature Extraction, Gaussian Mixture Model, TMS320C5509A Digital Signal Processor

University of Shahrood

Faculty of Electrical and Robotic Engineering

Design and Hardware Implementation of an Online Speaker Identification on TMS320C55xx Platform

Ahmad Moeini

Supervisor:

Dr. Hadi Grailu

July 2015