
بسم الله الرحمن الرحيم

دانشکده فنی و مهندسی

پایان‌نامه دوره کارشناسی ارشد مهندسی پزشکی–بیوالکتریک

بازشناسی گفتار نویزی با اصلاح روش خوشه بندی در دادگان مفقود

صادق مسجودی

استاد راهنما:

دکتر منصور ولی

تابستان ۱۳۹۰

دانشگاه شاهد

دانشکده فنی و مهندسی

صورت جلسه هیئت داوران رساله کارشناسی ارشد

جلسه دفاعیه پروژه کارشناسی ارشد مربوط به آقای/خانم **صادق مسجودی** به شماره دانشجویی**۸۷۷۵۲۰۵۰۵**در رشته **مهندسی پزشکی** با عنوان "**بازشناسی گفتار نویزی با اصلاح روش خوشه بندی در دادگان مفقود**" به ارزش **۶ واحد** در روز **۹۰/۶/۱۶** در دانشکده فنی و مهندسی با حضور افراد ذیل تشکیل شد، نتیجه به قرار زیر است :

☑ پروژه نامبرده با نمره ۱۹٫۵ قابل قبول می‌باشد.

☐ پروژه نامبرده مردود می‌باشد.

☐ پروژه نامبرده به شرط انجام اصلاحات جزیی قابل قبول می‌باشد. نمرهٔ دانشجو متعاقباً اعلام می‌شود.

☑ نام استاد راهنمای اول دکتر منصور ولی	دانشگاه : شاهد	امضاء	سهم استاد (به درصد):
☐ نام استاد راهنمای دوم	دانشگاه :	امضاء	سهم استاد (به درصد):
☐ نام استاد مشاور اول	دانشگاه :	امضاء	سهم استاد (به درصد):
☐ نام استاد مشاور دوم	دانشگاه :	امضاء	سهم استاد (به درصد):
☐ نام داور اول دکتر جهانشاه کبودیان	دانشگاه : رازی کرمانشاه	امضاء	
☑ نام داور دوم دکتر علی مطیع نصرآبادی	دانشگاه : شاهد	امضاء	
☐ نام داور سوم	دانشگاه :	امضاء	
☐ نام داور چهارم	دانشگاه :	امضاء	
☑ نام نماینده معاونت پژوهشی محمد یزدیان		امضاء	

تذکر:تعیین سهم اساتید در صورت وجود بیش از یک استاد راهنما و مشاور ضروری است.

تقدیم به:

مادرم

و پدرم

آن دو که هر آنچه دارم بواسطه وجود آنهاست.

حمد و سپاس خداوند بلند مرتبه را که رحمت واسعه اش همواره شامل حال عالمیان بوده است. بر خود لازم میدانم از زحمات و راهنمایی های دلسوزانه استاد راهنمای خود جناب آقای دکتر ولی تشکر نمایم که در طول مدت آشنایی با ایشان در عرصه زندگی نیز از راهنمایی های سودمند ایشان بهره گرفتم. ضمنا از اساتید محترمی که زحمت داوری این پایان نامه را به عهده گرفتند تشکر می نمایم.

چکیده

بر خلاف سیستم شنوایی انسان‌ها، سیستم‌های خودکار بازشناسی گفتار نسبت به نویز زمینه بسیار حساس هستند. این اثر ناشی از تفاوت مابین آمارگان مدل‌های گفتار تعلیمی است با آنچه که در شرایط واقعی از آنها استفاده می‌شود. جبران نکردن چنین عدم انطباقی، دقت سیستم‌های بازشناسی را به شدت کاهش می‌دهد.

در این گزارش، جهت جبران عدم انطباق بین دادگان تعلیم و تست از یکی از روش‌های ویژگی مفقود تحت عنوان بازسازی مبتنی بر خوشه یابی استفاده شده است. در روش‌های معمول بازسازی مبتنی بر خوشه یابی، بردارهای بازنمایی لگاریتم انرژی فیلتر بانک‌های ورودی مربوط به دادگان تمیز تعلیمی خوشه بندی می‌شوند. سپس برای بازسازی بردارهای بازنمایی نویزی دو مرحله پردازش روی آنها صورت می‌پذیرد. در مرحله نخست مؤلفه‌های هر بردار بازنمایی نویزی ورودی به دو بخش قابل اطمینان و غیر قابل اطمینان تفکیک می‌شوند. در مرحله دوم مؤلفه‌هایی که برچسب غیرقابل اطمینان(ماسک‌ها) خورده‌اند بر اساس آمارگان نزدیکترین خوشه به آن بردار که بر مبنای مؤلفه‌های قابل اطمینان بردار شناسایی شده است، با استفاده از یکی از روش‌های ماکزیمم احتمال پسین[1] بازیابی می‌شوند. در این پایان نامه ایده‌ای جدید مبنی بر استفاده مفیدتر از اطلاعات واقع در مؤلفه‌های غیر قابل اطمینان مطرح شده است. در این روش با استفاده از یک تابع عضویت فازی با فازی نمودن نسبت سیگنال به نویز(SNR) تخمینی مؤلفه‌های مختلف بردارهای بازنمایی ضریبی تحت عنوان ضریب تصحیح به بردار بازسازی شده و متمم فازی آن به بردار بازنمایی نویزی تخصیص داده شده و بردار حاصل جمع این دو بردار به عنوان بردار بازسازی شده نهایی تلقی می‌شود. با فرض اینکه گام اول بازسازی مبتنی بر خوشه یابی یعنی تفکیک مؤلفه‌های هر بردار به دو کلاس قابل اطمینان و غیرقابل اطمینان به نحو مناسبی انجام شده باشد، ایده مطرح شده نویدبخش بهبود در نتایج بازشناسی بود. به منظور بازشناسی بردارهای بازسازی شده، ابتدا از آنها تبدیل گسسته کسینوسی گرفته و ویژگی‌های MFCC را از آنها استخراج نمودیم. جهت ارزیابی و مقایسه روش‌ها از ویژگی‌های MFCC روی دو مدل بازشناسی، یکی شبکه عصبی و دیگری مدل مخفی مارکف استفاده گردید. مطابق انتظار ارزیابی‌ها نشان دادند بازسازی‌ها هنگامی موثر واقع می‌شوند که گام نخست یعنی شناسایی مؤلفه‌های مفقود به نحو مناسبی صورت پذیرفته باشد. اگر مؤلفه‌های مفقود به طریق مناسب معین شده باشند، ایده فازی مطرح شده دقت بازشناسی شبکه عصبی را در نسبت‌های پایین سیگنال به نویز تا ۵٪ و دقت بازشناسی حاصل از مدل مخفی مارکف را بین یک تا دو درصد نسبت به روش غیر فازی بهبود می‌بخشد. در ادامه پیاده سازی‌ها از روش سری تیلور برداری برای تخمین میانگین نویز استفاده شد و به کمک آن مقادیری را به عنوان نسبت سیگنال

[1] MAP(Maximum A posteriori Probability)

به نویز(SNR) به مؤلفه‌های مختلف بردارهای بازنمایی تخصیص دادیم. سپس با قرار دادن آستانه‌ای به عنوان حد آستانه قابل اطمینان بودن روی مقادیر به دست آمده، مؤلفه‌ها را به دو کلاس قابل اطمینان و غیرقابل اطمینان تفکیک نمودیم. با تکرار مراحل قبل جهت بازسازی و بازشناسی بردارها بهبودی در نتایج حاصل نشد. این امر دلالت بر عدم کارایی روش به کارگرفته شده جهت تعیین ماسک‌ها دارد.

فهرست جداول

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل ۱- بازشناسی خودکار گفتار

۱-۱- مقدمه

انسان‌ها توانایی شگفت‌آوری در برقراری ارتباط گفتاری با دیگران دارند. گفتار به عنوان طبیعی‌ترین شکل ارتباط بین انسان‌ها طی سالیان دراز مورد توجه محققین بوده است. در دهه‌های اخیر دانشمندان و مهندسین گفتار در جهت تولید سیستم‌هایی که ارتباط بین انسان‌ها و ماشین‌ها تسهیل نمایند کارهای فراوانی انجام داده‌اند. تلاش‌های قابل ملاحظه‌ای نیز توسط زبان‌شناسان، فیزیولوژیست‌ها و روان‌شناسان برای بررسی ساختار گفتار و مکانیسم‌های تولید و دریافت آن صورت گرفته است. ترکیب این یافته‌ها با پیشرفت‌های صورت پذیرفته در زمینه‌های محاسبات دیجیتال و پردازش سیگنال منجر به شکل‌گیری حوزه *تکنولوژی گفتار* شده است. تکنولوژی گفتار با سرعت زیادی در حال پیشرفت بوده و کاربرد آن در زندگی روزمره فراگیر شده است.

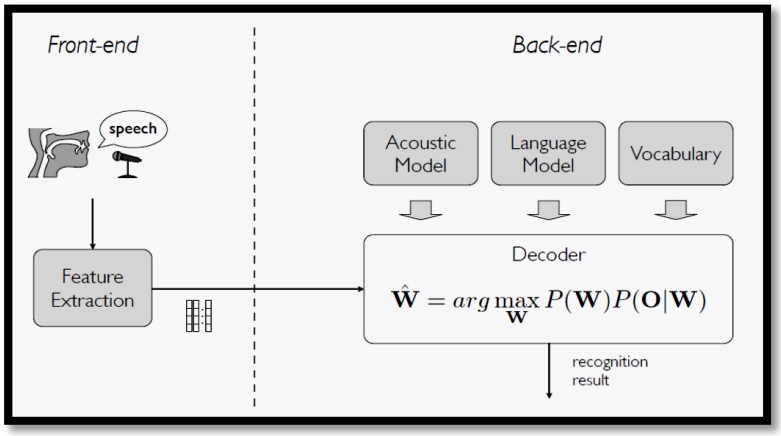
سه شاخه اصلی در زمینه گفتار تحت عناوین *سنتز گفتار*، *کدینگ گفتار* و *بازشناسی خودکار گفتار* قابل شناسایی است. هدف از سنتز گفتار ارایه سیستمی با این قابلیت است که یک فایل متنی را به گفتاری با صدای طبیعی تبدیل نماید. در کدینگ گفتار سعی می‌شود تعداد بیت‌های مورد نیاز جهت نمایش گفتار با بهره گیری از اطلاعات اضافی[1] گفتار کاهش داده شوند. این مورد هنگامی که لازم است گفتار از طریق کانال‌های مخابراتی انتقال یافته و یا روی یک حافظه دیجیتال ذخیره شود، اهمیت می‌یابد. در بازشناسی خودکار گفتار[2]، هدف دستیابی به ماشینی با این قابلیت است که قادر باشد گفتار طبیعی انسان را از هر گوینده از یک زبان خاص حتی در شرایط محیطی نامناسب بازشناسی نماید. این کار در واقع فرآیند تبدیل

[1] Redundancy
[2] ASR

یک سیگنال آکوستیک گفتار گرفته شده توسط یک میکروفن به دنباله‌ای از کلمات توسط یک برنامه کامپیوتری می‌باشد. در این گزارش به زمینه سوم تکنولوژی گفتار یعنی بازشناسی خودکار گفتار و به طور خاص بهبود مقاوم سازی سیستم بازشناسی گفتار در برابر نویز پرداخته می‌شود.

از سیستم‌های بازشناسی گفتار در زمینه‌های متنوع زیادی می‌توان استفاده نمود. در ارتباط انسان با کامپیوتر مثل اتوماسیون فرمان-کنترل[۱]، ناوبری صدا فعال[۲](اتومبیل و هواپیما)، فعال‌سازی وب از طریق صدا[۳] نمونه‌هایی از این کاربردها هستند. امروزه نرم افزارهای مبدل گفتار به متن به بازار وارد شده و قابل دسترسی هستند. ارتقاء ادوات پردازشی موبایل‌ها استفاده از بازشناسی گفتار در تلفن‌های همراه را در کاربردهایی مثل شماره گیری صوتی[۴] امکان پذیر ساخته است. بازشناسی خودکار گفتار در رابطه با یادگیری زبان و کسانی که در این زمینه نقیصه‌ای دارند هم کارایی خود را نشان داده است. از جمله دیگر زمینه‌هایی که استفاده از سیستم‌های بازشناسی گفتار در آنها رایج است روباتیک، بازی‌های کامپیوتری، دسترسی امنیتی و ترجمه خودکار می‌باشد[۱,۲ ,۳].

در سیستم‌های بازشناسی گفتار رایج استخراج ویژگی از سیگنال گفتار ورودی توسط طبقه ابتدایی[۵] انجام می‌شود و بازشناسی ویژگی‌های استخراج شده توسط طبقه انتهایی[۶] انجام می‌شود. خروجی این سیستم محتمل‌ترین فرض درباره کلمات تلفظ شده می‌باشد. این احتمال با ترکیب سه منبع اطلاعاتی محاسبه می‌شود: مدل آکوستیکی[۷]، مدل زبانی[۸]، واژه نامه[۹]. یک نمایش شماتیک از ساختار کلی یک سیستم بازشناسی گفتار خودکار در شکل ۱-۱ نشان داده شده است. در بخش‌های بعدی ماژول‌های اصلی یک سیستم بازشناسی توصیف خواهند شد.

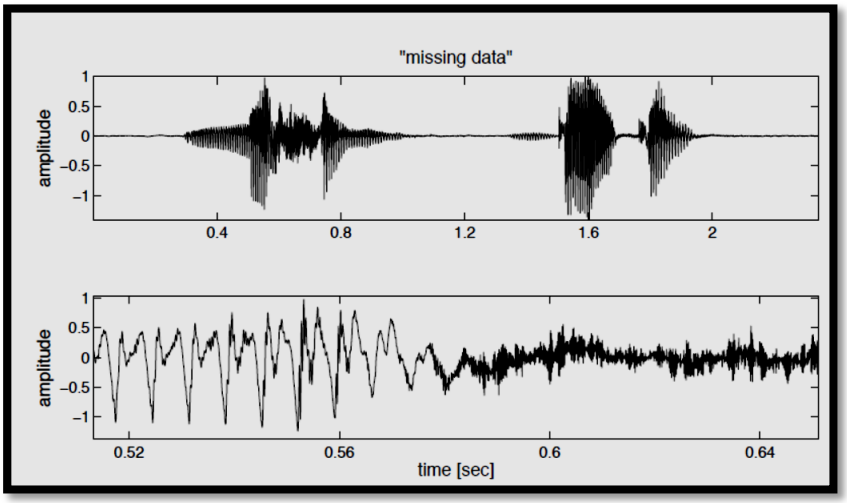


[1] command-control
[2] Active voice navigation
[3] Web enabling via voice
[4] Voice dialing
[5] Front end
[6] Back end
[7] Acoustic model
[8] Language model
[9] lexicon

شکل ۱-۱-ساختار کلی یک سیستم بازشناسی خودکار گفتار

۱-۲- استخراج ویژگی

ورودی سیستم بازشناسی یک سیگنال گفتار است که صورت فیزیکی آن یک شکل موج فشار بوده که از شخص در حال صحبت به یک یا چند شنونده منتقل می‌شود. برای تولید این شکل موج، شار هوایی بازدم شش‌ها از میان لوله و حنجره صوتی عبور می‌کند. محفظه لوله صوتی توسط اجزاء زیر کنترل می‌شود: زبان، لب‌ها، دندان‌ها، آرواره[1] و سق دهان. این سیستم مسئول تولید قسمت‌های خاصی از گفتار که در هر بازه خاص زمانی دارای ویژگی‌های آکوستیکی و بیانی[2] منحصر به فردی هستند یعنی واجها[3] می‌باشد. این واجها تحقق آکوستیکی واحدهایی تحت عنوان فونم‌ها یعنی واجهای زبانی که کلمات یک زبان را به وجود می‌آورند، می‌باشند. در شکل ۱-۲ سیگنال حوزه زمان S(n) کلمات تلفظ شده "missing data" نشان داده شده است. سیگنال گفتار به دلیل حالت ثابت تغییر لوله صوتی و حنجره صوتی در طول زمان و دامنه پیوسته می‌باشد. در همین شکل نمایش جزیی تر از یک تکه زمانی کلمه "missing" نشان داده شده است. بخش اول (متناظر با حرف صدادار 'i')دارای ویژگی تناوبی اما بخش آخر(ثابت 's') غیر پریودیک است. متناوب بودن قسمت اول از لرزش تارهای صوتی ناشی می‌شود که شار هوایی را منقطع و عبور آن را دچار وقفه‌هایی نموده و به تولید یک واکه یا حرف صدادار منجر شده است. فرکانس لرزش تارها، فرکانس پایه یا پیچ[4] نامیده می‌شود. اگر هنگام تولید واجها تارهای صوتی دچار لرزش نشوند صدای غیر واکه یا واج بی صدا تولید می‌شود.



شکل ۱-۲- نمایش حوزه زمان کلمات تلفظ شده "missing data" و نمای نزدیکتری از آن روی قسمت‌های 'i' و 's'

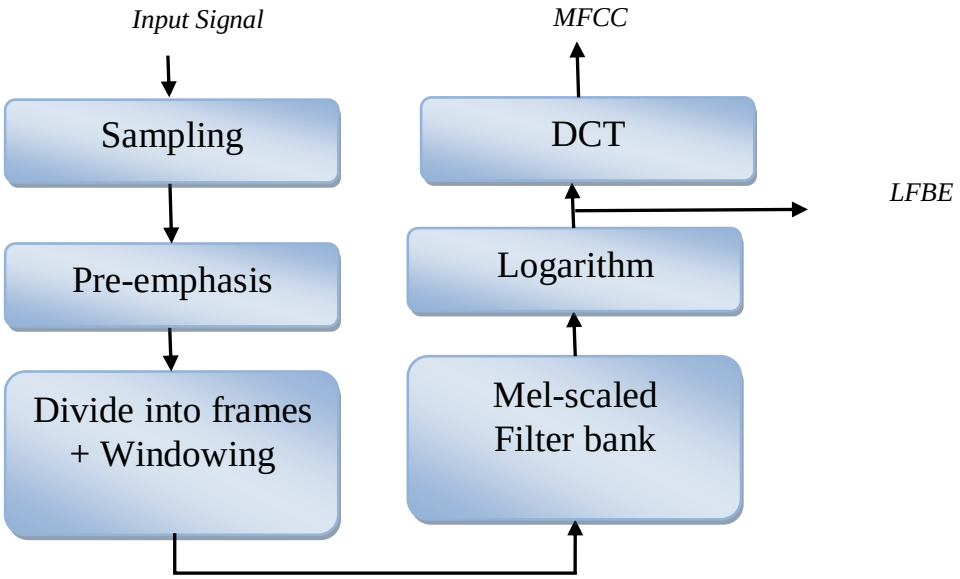
[1] jaw
[2] articulator
[3] phone
[4] pitch

در طبقه ورودی سیستم بازشناسی خودکار گفتار، سیگنال آنالوگ توسط یک میکروفن دریافت شده و با نمونه برداری و کوانتیزاسیون به شکل دیجیتال در می‌آید. نمونه برداری فرآیند استخراج مقادیر سیگنال آنالوگ در نمونه‌های گسسته زمانی است.

برای بازشناسی گفتار، حذف اطلاعات زاید و حشو در سیگنال گفتار از آن جهت که نمایش موثر جنبه‌های اساسی گفتار را در قالب ویژگی‌های محدودی در پی دارد، مطلوب است. به علاوه ویژگی‌های مرتبط با گفتار باید روی همه گویندگان ثابت باشد، یعنی کمیت‌های مشابه واج‌های یکسانی که توسط گویندگان مختلف ادا شده‌اند باید حفظ شوند. در طی سالیان متمادی، ویژگی‌های بازنمایی متعددی از جمله ویژگی‌های کدینگ پیش‌بینی خطی(LPC)[1]، ضرایب پیش‌بینی خطی مفهومی(PLP)[2]، طیف نگارهای(اسپکتروگرام) مدولاسیون، لگاریتم انرژی بانک فیلتر در مقیاس مل[3] (LFBE) و ضرایب کپسترال فرکانس مل(MFCC)[4] معرفی شده‌اند. در این تحقیق فقط از ویژگی‌های LFBE و MFCC که از تبدیلی خطی کسینوسی روی ویژگی‌های LFBE بدست می‌آید استفاده شده است[۴, ۵,۶,۷].

به منظور استخراج ویژگی‌های لگاریتم انرژی بانک فیلتر از مراحل پیش پردازشی که در شکل ۱-۳ برای یک سیستم نوعی نشان داده شده‌اند، استفاده می‌گردد. بدین ترتیب که پس از نمونه برداری شکل موج حوزه زمان از یک فیلتر پیش تاکید عبور داده می‌شود. این کار بخش‌های بالایی طیف گفتار را که در اثر فیلتر لوله صوتی تضعیف شده است را تقویت می‌کند. در ادامه سیگنال پیش تاکید شده به فریم‌های دارای هم پوشانی تقسیم می‌شود. در صورت استفاده از فریم‌های با طول ثابت، طول مناسب جهت استخراج ویژگی‌های ستبر بین ۱۰ تا ۳۰ میلی ثانیه می‌باشد که در خلال آن گفتار شبه ایستا فرض می‌شود. در بیشتر سیستم‌های بازشناسی گفتار از شیفت با همپوشانی حدود ۵۰ درصد استفاده می‌شود.

[1] Linear Prediction Coefficients
[2] Perceptual Linear Prediction
[3] Log-mel Filter Bank Energies
[4] Mel-scaled Frequency Cepstral Coefficients



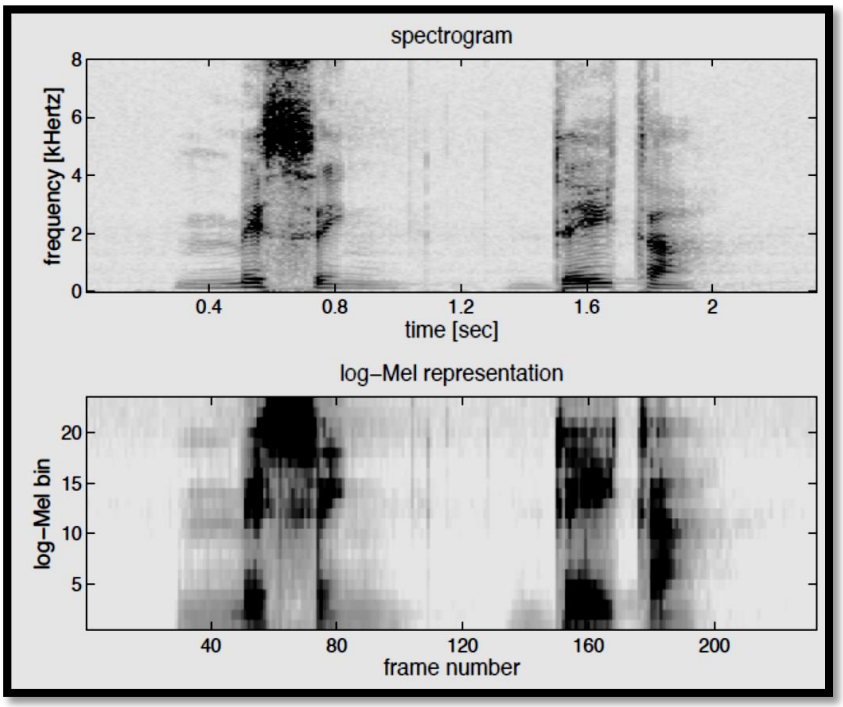
شکل ۱-۳- ماژول طبقه ابتدایی شامل استخراج ویژگی

پس از پنجره گذاری روی فریم‌ها(مثلاً پنجره همینگ)، طیف توان یا اسپکتروگرام به کمک تبدیل فوریه گسسته[1] به دست می‌آید. اسپکتروگرام چگونگی تغییرات انرژی سیگنال را در طول زمان و فرکانس نمایش می‌دهد. در شکل ۱-۴ اسپکتروگرام سیگنال حوزه زمان شکل ۱-۲ نشان داده شده است. در این مثال فرکانس نمونه برداری ۱۶ کیلو هرتز، طول فریم ۲۵ میلی ثانیه و شیفت ۱۰ میلی ثانیه می‌باشد. یک بانک فیلتر مقیاس مل هم روی دنباله فریم‌ها به کار گرفته شده است. مقیاس فرکانسی مل رزولوشن فرکانسی گوش انسان را توسط یک محور فرکانسی غیر خطی که در فرکانس‌های پایین خطی بوده اما در نواحی فرکانسی بالا تقریباً لگاریتمی است نمایش می‌دهد. شکل فیلترهای شنوایی انسان دارای پاسخ‌هایی با دامنه‌های تقریباً مثلثی شکل هستند. از این رو فواصل بانک فیلتری با D کانال به صورت لگاریتمی هستند تا حساسیت لگاریتمی گوش انسان را دنبال کنند.

ویژگی‌های MFCC یا کپسترال با به کار گرفتن ماتریس تبدیل کسینوسی گسسته[2] روی ویژگی‌های LFBEبه دست می‌آید. از آن جایی که شکل طیف حاصل از لوله صوتی هموار است، مؤلفه‌های انرژی طیف گرایش به همبستگی دارند. هدف اصلی از تبدیل DCT، تبدیل این ویژگی‌های طیفی به ضرایب کپسترال غیر هم بسته است. این امر تعداد ویژگی‌های حاوی اطلاعات کم را کاهش می‌دهد. نوعاً تعداد K ویژگی کپسترال استخراجی تقریباً نصف تعداد باندهای فرکانسی است. با کاهش همبستگی بین مؤلفه‌های MFCC می‌توان این ویژگی‌ها را با ماتریس های کوواریانس قطری مدل نمود. این امر به طور چشمگیری بار

[1] Discrete Fourier Transform
[2] Discrete Cosine Transform

محاسباتی در یک فرآیند تطبیق آماری کاهش می‌دهد در حالیکه هنوز دقت تخمین‌های احتمال قابل قبول باشد. از سوی دیگر تخمین احتمال‌ها از روی تعداد دادگان محدودتر ساده‌تر است. بنابراین MFCC به عنوان پرکاربردترین نوع ویژگی بازنمایی در سیستم‌های ASR مطرح است.



شکل ۱-۴- اسپکتروگرام و نمایش لگاریتم انرژی مقیاس مل کلمات"missing data"

در سیستم‌های بازشناسی خودکار گفتار مرسوم است ویژگی‌های MFCC را به منظور به همراه داشتن دینامیک زمانی گفتار با مشتقات زمانی آنها همراه نمایند. در بیشتر سیستم‌ها از مشتقات مرتبه اول و دوم که به ترتیب سرعت و شتاب نامیده می‌شوند، استفاده می‌گردد. افزودن این مشتقات زمانی مدل‌سازی رفتار غیر ایستان ناشی از تغییرات پیوسته لوله صوتی را امکان پذیر می‌سازد.

۱-۳- ماژول‌های طبقه انتهایی

دنباله کلمات بیان شده از m کلمه $\mathbf{W} = \{word_1, word_2, \ldots, word_m\}$ توسط یک ماژول استخراج ویژگی در طبقه ابتدایی به دنباله‌ای از مشاهدات $\mathbf{O} = \{O_1, O_2, \ldots, O_T\}$ تبدیل می‌شود. در اینجا T تعداد فریم‌های دنباله و O_t یک فریم گفتار در زمان t در حوزه ویژگی انتخابی می‌باشد. هدف از ماژول بخش انتهایی جستجو به منظور یافتن محتمل‌ترین دنباله کلمات $\hat{\mathbf{W}}$ با داشتن دادگان آکوستیکی $\mathbf{O}$ می‌باشد. طبق تئوری تصمیم

گیری بیز، بازشناسی گفتار را می‌توان به صورت یک مسأله کدینگ ماکزیمم احتمال پسین[۱] (MAP) روی همه دنباله‌های کلمات ممکن بر مبنای احتمال پیشین[۲] $P(\mathbf{W}|\mathbf{O})$ مطرح نمود:

$$\widehat{\boldsymbol{W}} = argmax_W p(\boldsymbol{W}|\boldsymbol{O})$$
$$= argmax_{\boldsymbol{W}} \frac{p(\boldsymbol{W}|\boldsymbol{O})p(\boldsymbol{W})}{p(\boldsymbol{O})}$$
$$\propto argmax_{\boldsymbol{W}} p(\boldsymbol{W}|\boldsymbol{O})p(\boldsymbol{W}) \qquad (۱-۱)$$

تاکنون سه منبع اطلاعاتی معرفی شده‌اند: مدل آکوستیکی$P(\mathbf{O}|\mathbf{W})$، مدل زبانی$P(\mathbf{W})$ و لغت نامه[۳]. از آنجا که مشاهدات آکوستیکی $\mathbf{O}$ برای همه فرض‌های $\mathbf{W}$ یکسان هستند، مخرج $P(\mathbf{O})$ را می‌توان از معادله (۱-۱) حذف نمود.

۱-۳-۱- لغت نامه

لغت نامه شامل لیستی از تمامی کلمات قابل شناسایی (لغت‌ها) به همراه توصیف واجی[۴] آنها می‌باشد. توصیف واجی تعیین می‌کند که چه دنباله‌ای از واج‌ها هنگام بازشناسی یک کلمه مجاز هستند. اندازه لغت نامه می‌تواند برای کاربردهایی که فقط به تعداد کمی لغت نیاز است(مثلاً شماره گیر صوتی) کوچک باشد در حالیکه در بازشناسی از میان هزاران کلمه به لغت نامه‌های بزرگ نیاز باشد.

۱-۳-۲- مدل زبانی

مدل زبانی احتمال پیشین $P(\mathbf{W})$ یعنی احتمال ادای $\mathbf{W}$ توسط گوینده را که از دادگان آکوستیکی مستقل است، محاسبه می‌کند. از این مدل برای سنجش کلمات مفروض به منظور کاهش تعداد دنباله‌های کلمات استفاده می‌شود. مدل زبانی قواعد معنایی[۵]، نحوی[۶]و عملی[۷] ذاتی زبان را در بر دارد. مثالی از مدل‌های معین[۸] گرامرهای مستقل از متن[۹] هستند که برای کار با لغت نامه‌های کوچک مناسب هستند. پر کاربردترین مدل‌های آماری n-گرام‌ها هستند که نوعاً هنگام کار با لغت‌نامه‌های بزرگ مورد استفاده قرار می‌گیرند. در یک n-گرام با داشتن (n-1) کلمه ماقبل یک احتمال به کلمه تخصیص داده می‌شود:

$$p(\boldsymbol{W}) = \prod_{l=1}^{m} p(word_l \mid word_{l-N+1} \dots word_{l-1} \quad (۲-۱)$$

[1] Maximum A Posteriori
[2] A priori
[3] lexicon
[4] phonetic
[5] semantic
[6] syntactic
[7] pragmatic
[8] deterministic
[9] Context free