

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشکده فنی و مهندسی

پایان نامه 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 شاهرود
دانشکده فنی و مهندسی

صورت جلسه هیئت داوران رساله کارشناسی ارشد

جلسه دفاعیه پروژه کارشناسی ارشد مربوط به آقای/خانم صادق مسعودی 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 ۱۹۱۵ قابل قبول می باشد.

پروژه نامبرده مردود می باشد.

پروژه نامبرده به شرط انجام اصلاحات جزئی قابل قبول می باشد. نمره دانشجو متقابلاً اعلام می شود.

امضاء	دانشگاه شاهرود	<input checked="" type="checkbox"/> نام استاد راهنمای اول دکتر منصور ولی
سهم استاد (به درصد):		
امضاء	دانشگاه	<input type="checkbox"/> نام استاد راهنمای دوم
سهم استاد (به درصد):		
امضاء	دانشگاه	<input type="checkbox"/> نام استاد مشاور اول
سهم استاد (به درصد):		
امضاء	دانشگاه	<input type="checkbox"/> نام استاد مشاور دوم
سهم استاد (به درصد):		
امضاء	دانشگاه	<input type="checkbox"/> نام داور اول دکتر جهانگوش کوردیان
سهم استاد (به درصد):		
امضاء	دانشگاه شاهرود	<input checked="" type="checkbox"/> نام داور دوم دکتر علی مطیع نصرآبادی
سهم استاد (به درصد):		
امضاء	دانشگاه	<input type="checkbox"/> نام داور سوم
سهم استاد (به درصد):		
امضاء	دانشگاه	<input type="checkbox"/> نام داور چهارم
سهم استاد (به درصد):		
امضاء		<input checked="" type="checkbox"/> نام نماینده معاونت پژوهشی محمد پوریا
سهم استاد (به درصد):		

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

تقدیم ہے:

مادر

و پدر

آن دو کہ ہر آنچہ دارم بواسطہ وجود آنهاست.

حمد و سپاس خداوند بلند مرتبه را که رحمت و سعادتش همواره شامل حال عالمیان بوده است. بر خود لازم میدانم از زحمات و راهبانی های دلسوزانه استاد راهبانی خود جناب آقای دکتر ولی شکر نمایم که در طول مدت آشناییم با ایشان در عرصه زندگی نیز از راهبانی های سودمند ایشان بهره گرفتم. ضمناً از اساتید محترمی که زحمت دآوری این پایان نامه را به عهده گرفتند شکر 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 پسین¹ 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 ۰.۵٪ 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.....	ر.....
فهرست شکل‌ها.....	ز.....
فصل ۱- بازشناسی خودکار گفتار.....	۱.....
۱-۱- مقدمه.....	۱.....
۲-۱- استخراج ویژگی.....	۳.....
۳-۱- ماژول‌های طبقه‌انتهایی.....	۶.....
۱-۳-۱- لغت نامه.....	۷.....
۲-۳-۱- مدل زبانی.....	۷.....
۳-۳-۱- مدل آکوستیکی.....	۸.....
۴-۳-۱- دی‌کدینگ.....	۱۰.....
۴-۱- مقاومت در برابر نویز.....	۱۰.....
۱-۴-۱- مقاومت آکوستیکی.....	۱۱.....
۲-۴-۱- تکنیک‌های جبران نویز در سیستم‌های بازشناسی خودکار گفتار.....	۱۲.....
فصل ۲- مفاهیم و اصطلاحات فازی.....	۱۴.....
۱-۲- منطق شفاف و منطق فازی.....	۱۴.....
۲-۲- مبانی مجموعه‌های فازی.....	۱۵.....
۱-۲-۲- مجموعه‌های فازی.....	۱۵.....
۲-۲-۲- توابع عضویت.....	۱۷.....
۳-۲- سیستم‌های فازی.....	۲۲.....
۱-۳-۲- فازی‌ساز.....	۲۲.....
۲-۳-۲- ماشین استنتاج فازی.....	۲۲.....
۳-۳-۲- غیر فازی‌ساز.....	۲۲.....
فصل ۳- روشهای ویژگی-مفقود در بازشناسی مقاوم گفتار.....	۲۴.....
۱-۳- مقدمه.....	۲۴.....
۲-۳- اندازه‌گیری‌های طیفی و ماسک‌های طیف‌نگاری.....	۲۵.....
۳-۳- پیش‌زمینه اضافی.....	۲۹.....
۱-۳-۳- بازشناسی گفتار با استفاده از مدل‌های مخفی مارکوف:.....	۲۹.....
۲-۳-۳- به حاشیه بردن تحدید شده چگالی‌های گوسی:.....	۳۰.....
۳-۳-۳- تخمین MAP محدود شده برای متغیرهای تصادفی گوسی:.....	۳۰.....
۴-۳- بازشناسی با اسپکتروگرام‌های غیر قابل اطمینان:.....	۳۱.....
۱-۴-۳- جایگذاری بردار ویژگی:.....	۳۲.....

۴۰.....	۲-۴-۳	روش‌های اصلاح تفکیک کننده.....
۴۲.....	۵-۳	شناسایی مؤلفه‌های غیر قابل اطمینان :.....
۴۳.....	۱-۵-۳	تخمین ماسک‌های اسپکتروگرافیک بر مبنای نسبت سیگنال به نویز :.....
۴۴.....	۲-۵-۳	شناسایی ماسک‌های اسپکتروگرافیک با استفاده از نویز تخمینی حاصل از سری تیلور برداری.....
۴۵.....	۳-۵-۳	تخمین بیز ماسک‌های اسپکتروگرافیک:.....
۴۸.....	۴-۵-۳	مواجهه با عدم قطعیت در تخمین ماسک :.....
۵۲.....	۵-۵-۳	پردازش‌های اضافه‌تر روی طیف ورودی:.....
۵۴.....	۶-۳	خلاصه:.....
۵۶.....	۴	فصل ۴ پیاده‌سازی و ارزیابی روش‌ها.....
۵۷.....	۱-۴	دادگان گفتار.....
۵۷.....	۱-۱-۴	دادگان تعلیم و دادگان تست.....
۵۸.....	۲-۱-۴	برچسب دهی دادگان.....
۵۹.....	۲-۴	بردارهای بازنمایی.....
۶۱.....	۱-۲-۴	استخراج بردارهای بازنمایی LFBE.....
۶۲.....	۲-۲-۴	بردارهای بازنمایی MFCC.....
۶۳.....	۳-۲-۴	تشکیل بردارهای بازنمایی با اضافه کردن مشتقات اول و دوم پارامترها.....
۶۳.....	۳-۴	مدل‌های بازشناسی.....
۶۳.....	۱-۳-۴	مدل شبکه عصبی TDNN.....
۶۶.....	۲-۳-۴	مدل مخفی مارکف.....
۶۶.....	۴-۴	ارزیابی دقت بازشناسی روی دادگان تست تمیز و نویزی.....
۶۶.....	۵-۴	اصلاح بردارهای بازنمایی لگاریتم طیفی با استفاده از روش دادگان مفقود.....
۶۷.....	۱-۵-۴	تعیین ماسک‌های معین و بازسازی سلول‌ها با کمک روش‌های مبتنی بر خوشه.....
۷۱.....	۲-۵-۴	شناسایی سلول‌های غیرقابل اطمینان با تخمین نویز به روش سری تیلور برداری.....
۷۵.....	۳-۵-۴	بازسازی فازی مبتنی بر خوشه دادگان مفقود.....
۸۲.....	۵	فصل ۵ نتیجه‌گیری و پیشنهادات.....
۸۲.....	۱-۵	مزایا و معایب روش‌های بازسازی ویژگی‌های مفقود.....
۸۲.....	۲-۵	مقایسه کلی نتایج.....
۸۴.....	۳-۵	پیشنهادات برای ادامه کار.....
		مراجع..... ۸۵

فهرست جداول

صفحه	عنوان
۵۸	جدول ۱-۴ - نحوه نمادگذاری آواها.....
۵۹	جدول ۲-۴ دسته‌بندی و کدگذاری آواهای فارسی
۶۰	جدول ۳-۴- درصد حضور دسته‌های مختلف آوایی در جملات ۴۰۰ و ۵۰۰ دادگان گفتار نویزی و تمیز
	جدول ۴-۴- صحت بازشناسی دادگان تست گفتار تمیز و نویزی توسط مدل شبکه عصبی TDNN و مدل مخفی مارکف
۶۷	جدول ۱-۵ درصد بهبود نتایج بازشناسی با به کارگیری روش های مختلف تخمین ماسک و بازسازی های مبتنی برخوشه یابی در سطوح بازشناسی فریم برای شبکه عصبی و واج برای مدل مخفی مارکف نسبت به حالت پایه
۸۳	

فهرست شکل‌ها

عنوان	صفحه
شکل ۱-۱- ساختار کلی یک سیستم بازشناسی خودکار گفتار.....	۳
شکل ۲-۱- نمایش حوزه زمان کلمات تلفظ شده "missing data" و نمای نزدیکتری از آن روی قسمت‌های 'i' و 's'.....	۳
شکل ۳-۱- مازول طبقه ابتدایی شامل استخراج ویژگی.....	۵
شکل ۴-۱- اسپکتروگرام و نمایش لگاریتم انرژی مقیاس مل کلمات "missing data".....	۶
شکل ۵-۱- نمای یک مدل مخفی مارکف سه حاته با توپولوژی چپ به راست.....	۹
شکل ۶-۱- نحوه مدلسازی تاثر نویز و کانال روی گفتار تمیز.....	۱۲
شکل ۱-۲- اختلاف بین درجه درستی در (الف) منطق دو سطحی {۰،۱} و (ب) منطق فازی [۰،۱].....	۱۵
شکل ۲-۲- توابع عضویت مجموعه شفاف C و مجموعه فازی F.....	۱۶
شکل ۳-۲- درجه عضویت x در مجموعه‌های A و B. $\mu_{Ax0} = 0.75$ و $\mu_{Bx0} = 0.25$	۱۷
شکل ۴-۲- برخی از مشخصات عمومی یک تابع عضویت.....	۱۸
شکل ۵-۲- تابع عضویت گوسی نرمال.....	۱۹
شکل ۶-۲- تابع عضویت جفت تابع سیگموئید.....	۱۹
شکل ۷-۲- توابع عضویت فازی نرم افزار Matlab.....	۲۱
شکل ۱-۳- (الف) اسپکتروگرام "مل" برای یک نطق از گفتار تمیز. (ب) اسپکتروگرام مل برای همان نطق وقتی که با نویز سفیدی با نسبت سیگنال به نویز ۱۰db تخریب شده است. (ج) ماسک اسپکتروگرافیک برای نطق نویزی با استفاده از یک حد آستانه.....	۲۶
شکل ۲-۳- شماتیک روش بازسازی مبتنی بر خوشه.....	۳۵
شکل ۳-۳- بلوک دیاگرام نحوه بازسازی مؤلفه‌های مفقود یک بردار.....	۳۶
شکل ۴-۳- بلوک دیاگرام نحوه بازسازی بردارهای شامل مؤلفه‌های غیر قابل اطمینان با روش چند خوشه.....	۳۸
شکل ۵-۳- مؤلفه‌های نشان داده شده با رنگ خاکستری غیرقابل اطمینان در نظر گرفته می‌شوند. بلوک‌های با خطوط قطری نمایش دهنده مؤلفه‌های با کوواریانس نرمال ۰.۵، یا بزرگتر نسبت به S(2,2) یا S(2,3) می‌باشند. برای شناسایی مؤلفه‌های غیر قابل اطمینان بردار دوم با استفاده از بازسازی مبتنی بر کوواریانس، Xu(2) از روی S(2,2) و S(2,3) و Xu(2) از روی S(1,2)، S(1,4)، S(2,1)، S(2,4)، S(3,2) و S(3,3) ساخته می‌شوند.....	۴۱
شکل ۶-۳- (الف) یک ماسک اسپکتروگرافیک ایده آل برای یک نطق آلوده به نویز سفید با نسبت سیگنال به نویز 10 dB. مؤلفه‌های زمان-فرکانسی قابل اطمینان بر مبنای نسبت سیگنال به نویزشان شناسایی شده‌اند. (ب) ماسک اسپکتروگرافیک برای همان نطق با تخمین محلی نسبت سیگنال به نویز به دست آمده است. (ج) ماسک اسپکتروگرافیک حاصل از تفکیک کننده بیز.....	۴۶
شکل ۷-۳- دقت بازشناسی گفتار آلوده به نویز سفید به صورت تابعی از نسبت سیگنال به نویز با استفاده از دو روش ویژگی مفقود مختلف. (الف) دقت بازشناسی به دست آمده با استفاده از روش حاشیه‌ای. بازشناسی با استفاده از بردارهای لگاریتم طیف انجام شده است. (ب) دقت بازشناسی به دست آمده با استفاده از روش بازسازی مبتنی بر خوشه. بازشناسی با استفاده از بردارهای کپستروم مشتق از اسپکتروگرام‌های بازسازی شده انجام شده است. در هر دو پنجره، سمبل‌های مثلثی عملکرد بازشناسی حاصل با ماسک‌های اسپکتروگرافیک ایده آل را نشان می‌دهند، در حالیکه لوزی‌ها دقت بازشناسی با ماسک‌های	

طیف تخمینی نمایش می‌دهند. سمبل‌های مربعی دقت بازشناسی کننده منطبق را نشان می‌دهد که در محیط تست آموزش داده شده است در حالیکه نمادهای دلتای پایین دقت حاصل را هنگامی که با گفتار تمیز آموزش داده شده نشان می‌دهد. ... ۴۹ شکل ۱-۴-۱- پنجره های همینگ با ۲۵۶ نمونه و ۵۰ درصد همپوشانی ۶۱

شکل ۲-۴-۲- نحوه استخراج پارامترهای LFBE ۶۱

شکل ۳-۴-۳- بانک فیلتر P تایی مثلثی مقیاس مل برای استخراج پارامترهای بازنمایی LFBE از گفتار ۶۲

شکل ۴-۴-۴- مدل بازشناسی مبتنی بر شبکه عصبی TDNN ۶۴

شکل ۵-۴-۵- درصد بازشناسی دسته‌های آوایی مختلف توسط شبکه عصبی برای دادگان تست گفتار تمیز ۶۵

شکل ۶-۴-۶- به ترتیب از بالا به پایین: (۱) نمایش زمان-فرکانسی یا اسپکتروگرام ویژگی های لگاریتم فیلتر بانک یک گفتار تمیز (۲) همان نمایش وقتی گفتار با نسبت سیگنال به نویز صفر dB نویزی شده است (۳) سلول‌های ماسک اسپکتروگرام گفتار نویزی. در این شکل مؤلفه‌های با SNR پایینتر از حد آستانه با رنگ قرمز مشخص شده اند. ۶۸

شکل ۷-۴-۷- مقایسه صحت بازشناسی گفتار نویزی و گفتار بازسازی شده توسط روش مبتنی بر تک خوشه با ماسک معین حاصل از مدل شبکه عصبی ۶۹

شکل ۸-۴-۸- مقایسه صحت بازشناسی گفتار نویزی و گفتار بازسازی شده توسط روش مبتنی بر تک خوشه با ماسک معین حاصل از مدل مخفی مارکف ۶۹

شکل ۹-۴-۹- اثر افزایش تعداد خوشه ها هنگام بازسازی ویژگی های نامطمئن معین در بهبود نتایج بازشناسی توسط شبکه عصبی ۷۰

شکل ۱۰-۴-۱۰- اثر افزایش تعداد خوشه ها هنگام بازسازی ویژگی های نامطمئن معین در بهبود نتایج بازشناسی توسط مدل مخفی مارکف ۷۱

شکل ۱۱-۴-۱۱- بلوک دیاگرام روش بازسازی تک خوشه ای سلول‌های مفقود تیلوری جهت بازشناسی گفتار نویزی ۷۱

شکل ۱۲-۴-۱۲- بلوک دیاگرام روش بازسازی چند خوشه ای سلول‌های مفقود تیلوری جهت بازشناسی گفتار نویزی ۷۲

شکل ۱۳-۴-۱۳- مقایسه اثر حد آستانه تخمین ماسک تیلوری روی نتایج بازشناسی در نسبت های مختلف سیگنال به نویز ۷۳

شکل ۱۴-۴-۱۴- نحوه یافتن بهترین حد آستانه جهت تفکیک سلول‌های اسپکتروگرام به دسته های قابل اطمینان و غیرقابل اطمینان. بازشناسی برای دادگان با نسبت سیگنال به نویز صفر dB انجام شده است. ۷۳

شکل ۱۵-۴-۱۵- نتیجه بازشناسی دادگان بازسازی شده با روش تک خوشه با ماسک‌های واقعی تخمینی از روش VTS حاصل از مدل شبکه عصبی ۷۴

شکل ۱۶-۴-۱۶- نتیجه بازشناسی دادگان بازسازی شده با روش تک خوشه ای با ماسک‌های واقعی تخمینی از روش VTS حاصل از مدل مخفی مارکف ۷۴

شکل ۱۷-۴-۱۷- تغییر دقت بازشناسی مدل شبکه عصبی با افزایش تعداد خوشه ها هنگام استفاده از ماسک‌های تیلوری ۷۵

شکل ۱۸-۴-۱۸- تغییر دقت بازشناسی مدل مخفی مارکف با افزایش تعداد خوشه ها هنگام استفاده از ماسک‌های تیلوری ۷۵

شکل ۱۹-۴-۱۹- بلوک دیاگرام روش فازی پیشنهادی جهت بهبود روش‌های بازسازی مبتنی بر خوشه ۷۶

شکل ۲۰-۴-۲۰- نمونه هایی از توابع عضویت فازی از نوع سیگموئید به همراه نمایش پارامتر های آنها ۷۷

شکل ۲۱-۴-۲۱- مقایسه صحت بازشناسی گفتار بازسازی شده با استفاده از روش بازسازی تک خوشه فازی و غیر فازی با ماسک معین حاصل از مدل شبکه عصبی ۷۸

شکل ۲۲-۴-۲۲- مقایسه صحت بازشناسی گفتار بازسازی شده با استفاده از روش‌های بازسازی تک خوشه فازی و غیر فازی با ماسک معین حاصل از مدل مخفی مارکف ۷۸

- شکل ۴-۲۳- اثر افزایش تعداد خوشه‌ها هنگام بازسازی فازی ویژگی‌های نامطمئن در بهبود نتایج بازشناسی توسط شبکه عصبی ۷۹
- شکل ۴-۲۴- اثر افزایش تعداد خوشه‌ها هنگام بازسازی فازی ویژگی‌های نامطمئن معین در بهبود نتایج بازشناسی توسط مدل مخفی مارکف ۷۹
- شکل ۴-۲۵- تأثیر ناچیز اصلاح فازی روش تک خوشه‌ای با ماسک‌های تیلوری در دقت بازشناسی توسط شبکه عصبی ۸۰
- شکل ۴-۲۶- تأثیر ناچیز اصلاح فازی روش تک خوشه‌ای با ماسک‌های تیلوری در دقت بازشناسی توسط مدل مخفی مارکف ۸۰
- شکل ۴-۲۷- دقت نتایج بازشناسی با افزایش تعداد خوشه‌ها در روش فازی با کمک ماسک‌های تیلوری حاصل از شبکه عصبی ۸۱
- شکل ۴-۲۸- دقت نتایج بازشناسی با افزایش تعداد خوشه‌ها در روش فازی با کمک ماسک‌های تیلوری حاصل از مدل مخفی مارکف ۸۱

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 مهندسیین 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 بهره‌گیری از اطلاعات اضافی^۱ 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^۲، 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

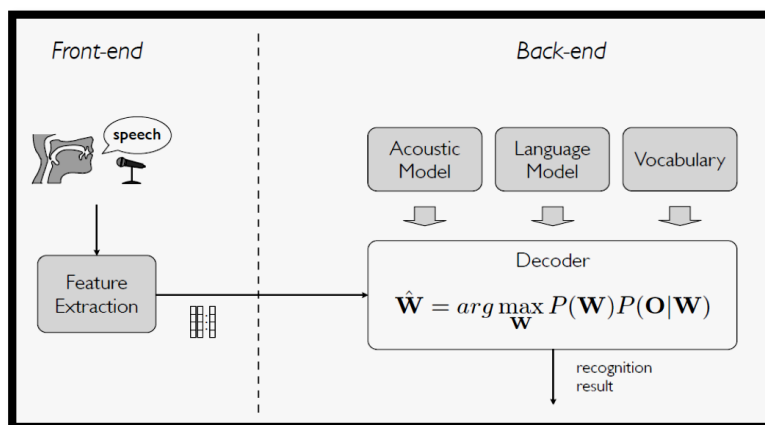
^۱ Redundancy

^۲ 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)، فعال سازی 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].

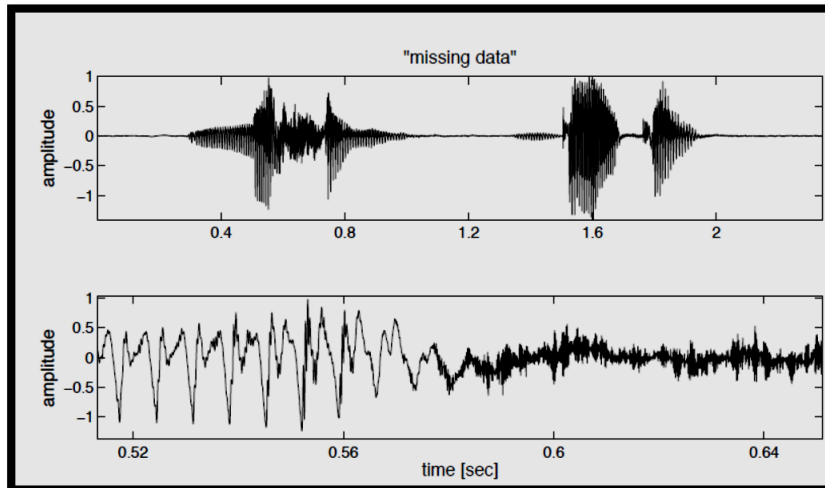
در سیستم‌های بازشناسی گفتار رایج استخراج ویژگی از سیگنال گفتار ورودی توسط طبقه ابتدایی^۵ انجام می‌شود و بازشناسی ویژگی‌های استخراج شده توسط طبقه انتهایی^۶ انجام می‌شود. خروجی این سیستم محتمل‌ترین فرض درباره کلمات تلفظ شده می‌باشد. این احتمال با ترکیب سه منبع اطلاعاتی محاسبه می‌شود: مدل آکوستیکی^۷، مدل 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.



¹command-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

۲-۱- 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^۱ و سق دهان. این سیستم مسئول تولید قسمت‌های خاصی از گفتار که در هر بازه خاص زمانی دارای ویژگی‌های آکوستیکی و بیانی^۲ منحصر به فردی هستند یعنی واج‌ها^۳ می‌باشد. این واج‌ها تحقق آکوستیکی واحدهایی تحت عنوان فونم‌ها یعنی واج‌های زبانی که کلمات یک زبان را به وجود می‌آورند، می‌باشند. در شکل ۲-۱ 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 جزئی 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^۴ نامیده می‌شود. اگر هنگام تولید واج‌ها تارهای صوتی دچار لرزش نشوند صدای غیر واکه یا واج بی صدا تولید می‌شود.



شکل ۲-۱- 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^۱ (LPC)، ضرایب پیش‌بینی خطی مفهومی^۲ (PLP)، طیف نگارهای (اسپکتروگرام) مدولاسیون، لگاریتم انرژی بانک فیلتر در مقیاس مل^۳ (LFBE) و ضرایب کپسترال فرکانس مل^۴ (MFCC) 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].

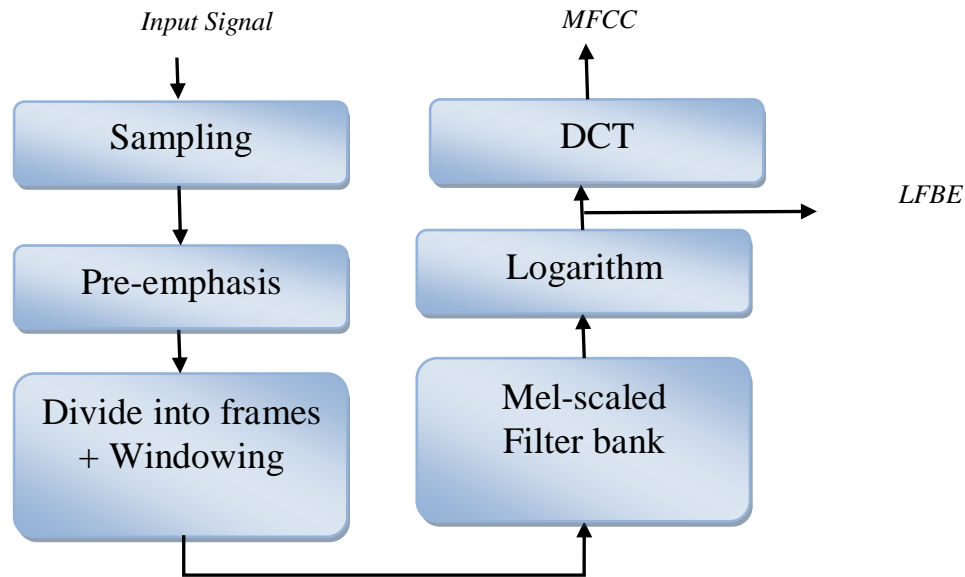
به منظور استخراج ویژگی‌های لگاریتم انرژی بانک فیلتر از مراحل پیش پردازشی که در شکل ۱-۳ برای یک سیستم نوعی نشان داده شده‌اند، استفاده می‌گردد. بدین ترتیب که پس از نمونه برداری شکل موج حوزه زمان از یک فیلتر پیش تاکید عبور داده می‌شود. این کار بخش‌های بالایی طیف گفتار را که در اثر فیلتر لوله صوتی تضعیف 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 ستر 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- مازول طبقه ابتدایی شامل استخراج ویژگی

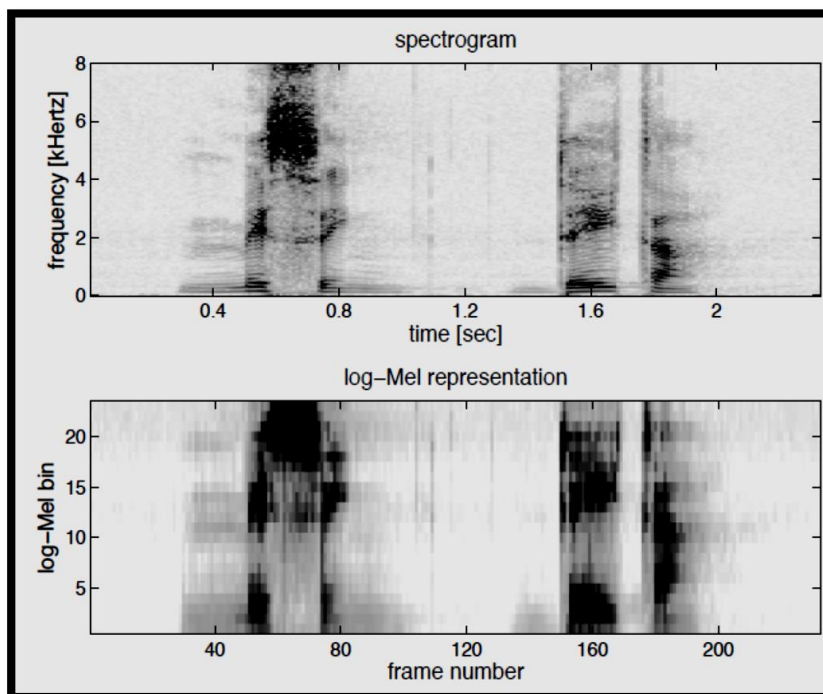
پس از پنجره گذاری روی فریم‌ها (مثلاً پنجره همینگ)، طیف توان یا اسپکتروگرام به کمک تبدیل فوریه گسسته^۱ به دست می‌آید. اسپکتروگرام چگونگی تغییرات انرژی سیگنال را در طول زمان و فرکانس نمایش می‌دهد. در شکل ۱-۴ اسپکتروگرام سیگنال حوزه زمان شکل ۱-۲ نشان داده شده است. در این مثال فرکانس نمونه برداری ۱۶ کیلو 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^۲ روی ویژگی‌های LFBE به 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 $\mathbf{W} = \{\text{word}_1, \text{word}_2, \dots, \text{word}_m\}$ توسط یک ماژول استخراج ویژگی در طبقه ابتدایی به دنباله‌ای از مشاهدات $\mathbf{O} = \{O_1, O_2, \dots, O_T\}$ 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:

$$\begin{aligned}\widehat{\mathbf{W}} &= \operatorname{argmax}_{\mathbf{W}} p(\mathbf{W}|\mathbf{O}) \\ &= \operatorname{argmax}_{\mathbf{W}} \frac{p(\mathbf{W}|\mathbf{O})p(\mathbf{W})}{p(\mathbf{O})} \\ &\propto \operatorname{argmax}_{\mathbf{W}} p(\mathbf{W}|\mathbf{O})p(\mathbf{W}) \quad (1-1)\end{aligned}$$

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 (1-1) 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:

$$p(\mathbf{W}) = \prod_{l=1}^m p(\text{word}_l | \text{word}_{l-N+1} \dots \text{word}_{l-1}) \quad (2-1)$$

¹ Maximum A Posteriori

² A priori

³ lexicon

⁴ phonetic

⁵ semantic

⁶ syntactic

⁷ pragmatic

⁸ deterministic

⁹ Context free