
بسم الله الرحمن الرحیم

دانشکده برق و رباتیک

رشته‌ی مهندسی برق گرایش مخابرات سیستم

پایان نامه کارشناسی ارشد

تشخیص مرز کلمات درسیگنال گفتار با استفاده از روش زمان ـ فرکانس

نگارنده:

فاطمه تاجیک ایجدان

استاد راهنما:

دکتر حسین مروی

شهریور ۱۴۰۰

تقدیم به

پدر و مادر عزیز و مهربانم که در سختی‌ها و دشواری‌های زندگی همواره یاوری دلسوز و فداکار و پشتیبانی محکم و مطمئن برایم بوده‌اند.

سپاس بی کران پروردگار یکتا را که هستی مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه چینی از علم و معرفت را روزیمان ساخت.

بدینوسیله از زحمات و تلاش بی دریغ خانواده عزیز و استاد محترم جناب آقای دکتر حسین مروی که در تهیه این مجموعه با این جانب همکاری داشته اند، تشکر و مراتب سپاس قلبی خود را اعلام نموده و موفقیت ایشان را ازخداوند متعال خواهانم.

تعهدنامه

اینجانب **فاطمه تاجیک ایجدان** دانشجوی دوره کارشناسی ارشد رشته مخابرات سیستم دانشکده برق و رباتیک، دانشگاه صنعتی شاهرود نویسنده پایان نامه با موضوع **تشخیص مرز کلمات درسیگنال گفتار با استفاده از روش زمان ـ فرکانس** تحت راهنمایی دکتر **حسین مروی** متعهد می‌شوم:

- 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 صنعتی شاهرود می‌باشد و مقالات مستخرج با نام **دانشگاه صنعتی شاهرود** و یا **Shahrood University of Technology** 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 می‌گردد.
- 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 :

امضاء دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن(مقالات، مستخرج، کتاب، برنامه‌های رایانه ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.

- استفاده از اطلاعات نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

تشخیص مرز کلمه یعنی شناسایی آغاز و پایان کلمه در یک سیگنال گفتاری که نقش مهمی برای بسیاری از کاربردها مانند تشخیص کلمه کلیدی، تشخیص گفتار، سیستم تبدیل متن به گفتار و ... دارد. در سیستم تشخیص گفتار بیشتر خطاها ناشی از تشخیص اشتباه مرزهای کلمات است. تشخیص صحیح مرز کلمه، خطاها را کاهش می‌دهد و عملکرد سیستم‌ها را بهبود می‌بخشد. الگوریتم‌های زیادی برای تشخیص مرز کلمات استفاده می‌شود که برخی از آن‌ها به دلیل نویز محیط و یا عدم مکث خوب بین کلمات عملکرد چندان خوبی برای تشخیص صحیح مرز کلمات ندارند. در صورتی که برخی از روش‌ها با هم ترکیب شوند بهتر می‌توانند مرزهای کلمات را تشخیص دهند.

در این تحقیق هدف تشخیص مرزهای کلمه با استفاده از روش‌های زمان-فرکانس از جمله موجک و ویگنر- ویل است. روش‌های زمان-فرکانس سیگنال را در هر دو حوزه زمان و فرکانس بررسی می‌کنند. چون سیگنال گفتار یک سیگنال ناایستا است و مشخصات آن با زمان تغییر می‌کند بهترین روش برای تحلیل این سیگنال و استخراج ویژگی از آن روش‌های زمان-فرکانس می‌باشد. روش‌های مختلف استخراج ویژگی مانند MFCC، LPC، PLP، تحلیل موجک و روش ویگنر-ویل جهت تشخیص مرز کلمات مورد بررسی قرار می‌گیرند. از دونوع داده برای بررسی نتایج استفاده شده است. در پایگاه‌داده TIMIT زمان بین کلمات خیلی کم است و به طور طبیعی و تقریبا سریع کلمات تلفظ می‌شوند، در حالی که در CADLab کلمات به صورت شمرده بیان می‌شوند. روش ویگنر-ویل قادر به تشخیص مرزهای کلمات با ۸۳/۵٪ برای داده اول، و ۶۸٪ برای داده دوم بهترین روش برای شناسایی مرزهای کلمات می‌باشد. آنتروپی نیز با ۲۵٪ تشخیص برای داده اول و نرخ عبور از صفر با ۳۰٪ تشخیص برای داده دوم کمترین میزان تشخیص صحیح مرزهای کلمات را در میان روش‌های انجام شده دارد.

واژگان کلیدی: تشخیص مرز کلمه، استخراج ویژگی، زمان-فرکانس، MFCC، PLP، LPC

لیست مقالات مستخرج از پایان‌نامه:

...

فهرست

فهرست شکل‌ها

فهرست جداول

فصل اول

کلیات پژوهش

۱-۱- مقدمه

گفتار طبیعی‌ترین و ساده‌ترین روش ارتباط بین انسان‌ها و به عنوان یک وسیله ارتباط جمعی می‌باشد که عامل پیدایش زبان‌های مختلف شده است. گفتار از مهم‌ترین سیگنال‌های پردازشی و ارتباطی است و از یک رشته از صداها تشکیل شده است. صداها نماینده اطلاعاتی هستند که بین انسان‌ها مشترک اند. اغلب پدیده‌های فیزیکی در جهان دارای صدا بوده که تحلیل سیگنال این اصوات ممکن است اطلاعات بسیار مفیدی را ارائه دهند. پردازش گفتار یعنی ارتباط بین انسان و ماشین در اساسی‌ترین شکل طبیعی آن. هدف از پردازش گفتار، تحلیل گفتار برای تشخیص اتوماتیک و استخراج اطلاعات، دریافت برخی از اطلاعات فیزیولوژی گوینده، ارتباط گفتاری بین انسان و ماشین (کامپیوتر) در اساسی‌ترین شکل آن[۱] و پردازش انواع مختلف اصوات جهت کسب یکسری اطلاعات مفید و مورد نظر آن‌ها (از سیگنال صوت) توسط ماشین. از کاربردهای پردازش گفتار می‌توان به شناسایی گفتار و گوینده، تبدیل متن به گفتار، فشرده سازی و بهسازی گفتار و... اشاره کرد.

سیگنال صوتی و شنیداری[۲] احساس ارتعاشات و نوسانات هوا توسط گوش انسان است. اگر این صوت در محدوده فرکانسی ۲۰ هرتز تا ۲۰ کیلوهرتز باشد آکوستیک[۳] نام دارد که با رسیدن به گوش و سپس انتقال به مغز و پردازش روی آن باعث درک مفهوم در ذهن انسان می‌شود. سیگنال گفتار[۴] زیر مجموعه‌ای از سیگنال صوتی و شنیداری می‌باشد که توسط انسان ایجاد می‌شود.

۱-۲- استخراج ویژگی[۵]

استخراج ویژگی یک مرحله مهم در تشخیص مرز کلمات در یک فایل گفتار است. فرآیندی است که بخش‌های مهم اطلاعات را نگه داشته و بخش‌های اضافی را حذف می‌کند. هدف از استخراج ویژگی

[1] Human machine Interaction (HMI)
[2] Audio
[3] acoustic
[4] speech signal
[5] Feature Extraction

نمایش هر سیگنال گفتاری با تعداد محدودی ویژگی است. دلیل این امر آن است که تمام اطلاعات موجود در سیگنال صوتی برای پردازش بسیار زیاد است و همه اطلاعات برای کارهای خاص مناسب نیستند. پس تبدیلاتی روی سیگنال گفتار انجام می‌شود تا بتوان ویژگی مورد نظر را استخراج کرد [۱]. برای مثال، ممکن است برخی ویژگی‌ها در محیط‌های نویزی جواب بهتری دهند یا اینکه حجم محاسبات کمتری را طلب کنند. این ویژگی‌ها گاه در کل سیگنال گفتار و گاه در یک پنجره از آن محاسبه می‌شوند، ولی در تحلیل سیگنال گفتار به‌طور معمول ویژگی‌های سیگنال پنجره‌بندی شده تحلیل می‌شود. این ویژگی‌ها گاه یک عددند، نظیر انرژی یا میانگین پنجره سیگنال گفتار و گاه یک بردارند. برای مثال، خود اندازه سیگنال گفتار در حوزه‌ی زمان را می‌توان به صورت برداری یک‌بعدی با طول تعداد نقاط پنجره گفتار در نظرگرفت. ازاین‌رو، گاه از این ویژگی‌ها با نام ضرایب ویژگی یا بردارهای ویژگی نیز یاد می‌شود. از معروف‌ترین این ویژگی‌ها می‌توان به ضرایب کپسترال فرکانس مل (MFCC[۱])، تعداد عبور از صفر(ZRC[۲])، ضرایب پیشگویی خطی (LPC[۳]) وتحلیل ادراکی پیشگویی خطی(PLP[۴]) اشاره کرد، که رایج ترین ویژگی که در تشخیص گفتار استفاده می‌شود MFCC است [۲].

الگوریتم‌های استخراج ویژگی به سه روش دسته بندی می‌شوند:

۱)تحلیل زمانی

۲) تحلیل طیفی

۳) تحلیل زمان-فرکانس

[1] Mel Frequency Cepstral Coefficients
[2] Zero Crossing
[3] Linear Prediction Coefficients
[4] Perceptual linear predictive

۱-۳- تشخیص مرز کلمه[1]

تشخیص مرز کلمه به شناسایی آغاز و پایان هر کلمه در یک عبارت گفتاری دارای نویز گفته می‌شود. این نویز ممکن است نویز پس زمینه یا مصنوعات تولید شده توسط گوینده باشد. نویز پس زمینه مشکل تشخیص مرز کلمات را بطور قابل توجهی پیچیده می کند. به عنوان مثال، اغلب ابتدای یا انتهای یک کلمه توسط گوینده مانند صدای دهان، لب زدن و تنفس شدید پنهان می شود [۳]. ویژگی های مورد نیاز یک آشکارساز مرز کلمه ایده آل عبارتند از: قابل اطمینان بودن، مقاومت در برابر نویز، دقت بالا، سازگاری، سادگی. در میان این ویژگی‌ها، استحکام در برابر نویز بهترین و سختترین ویژگی است. تشخیص مرز کلمه یک مسئله پیچیده است. تشخیص سکوت کافی نیست، زیرا همه کلمات مکث‌های واضحی بین خود ندارند. علاوه بر این، برخی از واج‌ها ویژگی‌هایی شبیه به نویز دارند. مدت زمان تلفظ کلمات تا حدی توسط تعداد و طول واج‌هایی که کلمه را تشکیل می‌دهند وچگونگی تلفظ گوینده تعیین می‌شود [۴]. تشخیص مرز کلمه می‌تواند وابسته به گوینده و مستقل از گوینده طبقه‌بندی شود. سیستم‌های وابسته به گوینده در مقایسه با سیستم‌های مستقل از گوینده به راحتی توسعه می‌یابند، زیرا می توان اطلاعاتی در مورد ویژگی‌های گفتار یک گوینده مثل گام بدست آورد، اما سیستم‌های مستقل از گوینده باید براساس بسیاری از گویندگان ایجاد شوند که به دلیل تفاوت در نوای گفتار، گام، سن، جنسیت، گویش و .. تشخیص مرز را به چالش می‌کشند.

تشخیص دقیق مرز کلمه به دو دلیل اهمیت دارد:

۱) صحت و اعتبار شناسایی گفتار به تشخیص دقیق مرز کلمات بستگی دارد.

۲) تفکیک کلمات از سکوت، پیچیدگی سیستم را کاهش می دهد [۵].

تشخیص مرز کلمه معمولا با استفاده از اطلاعات صوتی گفتار انجام می‌شود و در بسیاری از کاربردها مانند تشخیص کلمه کلیدی، سیستم تبدیل متن به گفتار و سیستم سنتز و تشخیص گفتار به کار

[1] Word Boundary Detection (WBD)

می‌رود. در سیستم تبدیل متن به گفتار مکث‌های به جا و درست در هنگام تولید گفتار تولید شود تا شنونده معنی را به درستی درک کند [۶]. در هنگام استفاده از اطلاعات مرز کلمات، مهم‌ترین نکته دقیق بودن این اطلاعات می‌باشد. اگر مرزها به درستی شناسایی نشوند، اطلاعات غلط باعث افزایش خطا می‌شود.

تشخیص مرز کلمه یک بخش مهم در سیستم تشخیص خودکار گفتار[1] است و برای عملکرد کارآمد و دقیق سیستم تشخیص خودکار گفتار ضروری است. دلیل اصلی خطاها در سیستم‌های تشخیص خودکار گفتار، تشخیص نادرست مرزهای شروع و پایان کلمات است. برای الگوریتم‌های تشخیص خودکار گفتار ضروری است که بخش های گفتار از غیر گفتاری جدا شوند. فناوری تشخیص گفتار می تواند سیگنال گفتار را به کمک ماشین مخصوص به متن یا دستور تبدیل کند و این یک تحقیق بسیار فعال در پردازش سیگنال و هوش مصنوعی است. تشخیص دقیق مرز کلمات در یک سیستم تشخیص گفتار به دلیل کاهش محاسبات در مراحل شناسایی و افزایش دقت در سیستم تشخیص گفتار از اهمیت ویژه ای برخوردار است [۷].

۱-۴- تقسیم‌بندی[2] گفتار به بخش‌های واکدار[3] و بی‌واک[4]

سیگنال گفتار از سه بخش واکدار، بی‌واک و سکوت تشکیل شده است. دانش ویژگی‌های صوتی به ویژه در بخش‌های واکدار و بی‌واک نقش مهمی را در بسیاری از سیستم‌های تجزیه و تحلیل گفتار ایفا می‌کند. تقسیم بندی گفتار به بخش‌های واکدار، بی‌واک و سکوت[5] برای پردازش ابتدایی لازم است و در بسیاری از موارد مانند سنتز گفتار، تشخیص گفتار و تقویت گفتار کاربرد دارد [۸] .گفتار واکدار

[1] Automatic Speech Recongnition System (ASRS)
[2] Segmentation
[3] voiced
[4] unvoiced
[5] silence

تقریباً از یک طیف[1] ثابت تشکیل شده است. فرکانس های ثابت معمولاً هنگام گفتن مصوت ها[2] ساخته می شوند. ناحیه صوتی گفتار هنگامی ایجاد می شود که چاکنای[3] لرزان از طریق مجرای صوتی[4] طنین انداز شود و سبب بیرون رانده شدن توده‌ی هوا توسط این تارها می‌گردد و پالس‌های متناوب در هوا تولید می‌شود. گفتار واکدار برای فهم گفتار از اهمیت بیشتری برخوردار است. در گفتار بی‌واک تارهای صوتی کاملا از هم باز هستند و هوا به راحتی از میان آن‌ها عبور می‌کند و پالس های غیر متناوب تولید می‌کند. بنابراین سیگنال های صوتی را می توان به دلیل تناوب آن تشخیص و جدا کرد.

براساس ویژگی‌های گفتار واکدار و بی‌واک ذکر شده در جدول(۱-۱)، الگوریتم‌های زیادی برای تشخیص گفتار آوایی و غیر آوایی وجود دارد. مشکلاتی که درهنگام قطعه بندی با آن‌ها مواجه هستیم:

۱) همپوشی گفتاری واکدار و بی‌واک با یکدیگر، شناسایی نقاط پایانی دقیق آن‌ها را دشوار می‌سازد.

۲) گفتار با انرژی کم مانند زمزمه می‌تواند به راحتی با سکوت اشتباه گرفته شود [۹].

جدول ۱-۱: تفاوت بین گفتار بی‌واک و واکدار

گفتار بی‌واک	گفتار واکدار
انرژی پایین در هر فریم	انرژی بالا در هر فریم
نرخ بالا عبور ازصفر	نرخ پایین عبور از صفر
شبیه به نویز	دارای تناوب

۱-۵- ضرورت انجام پژوهش

کلمه قدیمی‌ترین واحد تشخیص مورد استفاده است. زمانی از کلمه به عنوان واحد تشخیص استفاده

[1] spectrum
[2] vowels
[3] glottis
[4] vocal tract

می‌شود که تعداد کلمات محدود باشد و بطور مجزا تلفظ شوند. تشخیص مرز کلمات باعث می‌شود تا در بسیاری از کاربردها مانند بازشناسی گفتار، تبدیل متن به گفتار و ... حجم کمتری از اطلاعات وجود داشته باشد و سرعت افزایش می‌یابد. همچنین قسمت‌های مفید و اصلی گفتار در اختیار می‌باشد که اطلاعات اضافی آن حذف شده در نتیجه دقت نیز افزایش می‌یابد.

۱-۶- روش انجام کار

در این پایان‌نامه از ۹ روش استخراج ویژگی برای تشخیص مرز کلمات استفاده می‌شود. از دو پایگاه داده که دارای سرعت متفاوتی در بیان کلمات هستند، استفاده می‌شود تا تأثیر سرعت تلفظ عبارت گفتاری روی نتیجه نهایی مشخص شود. ابتدا با استفاده از روش مورد نظر ویژگی استخراج می‌شود. سپس برای کاهش پردازش، مقدار DC سیگنال حذف می‌شود. با استفاده از فیلتر باتروورث هموارسازی انجام می‌شود و سپس مقدار آستانه اعمال می‌شود. محدوده‌ای از سیگنال که از مقدار آستانه بیشتر باشد به عنوان مرز کلمه در نظر گرفته می‌شود.

۱-۷- ساختار پایان‌نامه

این پایان نامه از پنج فصل تشکیل شده است که به شرح زیر می‌باشد:

فصل اول مقدمه‌ای در مورد پردازش گفتار، تشخیص مرز کلمه، استخراج ویژگی و قطعه بندی سیگنال گفتار به بخش‌های آوایی و غیرآوایی می‌باشد و توضیحاتی راجع به هر کدام ارائه شده است.

در فصل دوم روش‌های متداولی که برای تشخیص مرز کلمه در کارهای پیشین استفاده شده است، مورد بررسی قرار گرفته و نتایج آن نمایش داده شده است.

در فصل سوم روش‌های استخراج ویژگی با استفاده از آنالیز طیفی و آنالیز زمان-فرکانس در سیگنال-های گفتاری مورد بررسی قرار می‌گیرند.

در فصل چهارم ابتدا پایگاه داده‌ای که در این پایان‌نامه مورد استفاده قرار می‌گیرد، معرفی می‌شود و سپس روش‌های پیشنهادی برای تشخیص مرز کلمات معرفی شده است و همچنین نتایج شبیه سازی

حاصل از استخراج ویژگی‌های مختلف نمایش داده شده و تجزیه و تحلیل و نتیجه‌گیری می‌شود.

در فصل پنجم نتیجه‌گیری کلی آورده شده و همچنین برای ادامه کار روش‌هایی پیشنهاد می‌شود.

فصل دوم

مروری بر تحقیقات پیشین

۲-۱- مقدمه

در فصل قبل راجع به تشخیص مرز کلمه و کاربردهای آن صحبت شد. تشخیص دقیق مرزهای آغاز و پایان کلمه، کیفیت پردازش اولیه را بهبود می بخشد و به کاهش هزینه‌های محاسباتی برنامه‌های گفتاری کمک می‌کند. امروزه مسئله تقسیم‌بندی گفتار و تشخیص نقطه آغازین و پایانی کلمه به کمک روش‌های زمانی و فرکانسی انجام می‌شود. در این فصل روش‌هایی که در کارهای انجام شده برای تشخیص مرز کلمه استفاده شده‌اند، تحلیل و بررسی شده و توضیحاتی راجع به هرکدام ازآن‌ها داده می‌شود.

۲-۲- انرژی زمان-کوتاه[1]

انرژی یکی از ویژگی‌های صوتی در گفتار و یک پارامتر مهم در شناسایی مرزها و فعالیت واکداری می‌باشد. در میان نویز، بخش واکدار و بخش بیواک، نویز کمترین انرژی را دارد. این روش به طور عمده در تشخیص مناطق آوایی مورد استفاده قرار می‌گیرد. این روش توانایی تشخیص وجود صدا در محیط هایی با SNR پایین را دارد [۱۰]. انرژی زمان-کوتاه، انرژی بخش کوتاهی از سیگنال گفتار است. انرژی زمان- کوتاه یک پارامتر برای طبقه بندی ساده و مؤثر بخش‌های صوتی و بدون صدا است. تعریف انرژی طولانی مدت سیگنال به شرح زیر است [۱۱]:

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \qquad (۱-۲)$$

در عبارت فوق E انرژی سیگنال $x(m)$ را نشان می‌دهد. این تعریف برای سیگنال‌های متغیر با زمان مانند گفتار مناسب نیست. چون گفتار یک سیگنال غیرایستا[2] است و مشخصات آن با زمان تغییر می‌کند. پس سیگنال گفتار پنجره‌گذاری شده، به فریم‌های کوتاه مدت تقسیم‌بندی می‌شود تا مشخصات سیگنال به سیگنال ایستا[3] نزدیک شود.

[1] Short-Time Energy (STE)
[2] non-stationary
[3] Stationary

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) = x^2(n-N+1) + \ldots + x^2(n) \qquad (۲-۲)$$

$$x_n(m) = x(m)w(n-m) \qquad n-N+1 \le m \le n \qquad (۲-۳)$$

انرژی کوتاه مدت از رابطه (۲-۴) به دست می‌آید:

$$E_n = \sum_{m=n-N+1}^{n} \left[x(m)w(n-m)\right]^2 \qquad (۲-۴)$$

$w(n-m)$ پنجره، و n شماره فریم است.

۲-۲-۱- الگوریتم کلی از انرژی زمان-کوتاه

تغییراتی در انرژی مرزهای کلمات مشاهده می‌شود که به عنوان نشانه‌ای برای شناسایی مرزهای کلمه است. در مرزهای کلمه، دامنه به طور ناگهانی بالا یا پایین می‌رود [۱۲]. برای هر سیگنال گفتاری یک آستانه مشخص می‌شود که برای تعیین مرزهای کلمه استفاده می‌شود [۶].

برای مشخص کردن مرزهای کلمه، ابتدا سیگنال گفتار فریم بندی و پنجره‌گذاری می‌شود و سپس انرژی در هر فریم گفتار محاسبه می‌شود. برای محاسبه انرژی از رابطه (۲-۴) استفاده می‌شود. در یک پنجره هر نمونه به توان دو می‌رسد، سپس با هم جمع می‌شوند. محدوده‌ای از سیگنال که انرژی آن از انرژی آستانه بیشتر باشد، به عنوان مرز کلمه در نظر گرفته می‌شود.

۲-۲-۲- بررسی کارهای گذشته

در مرجع [۱۱] انرژی یکی از روش‌هایی است، که برای تشخیص قسمت‌های واکدار و بیواک استفاده شده است. در این آزمایش از دو گوینده مرد و زن استفاده شده است که هر کدام جمله‌ای را به زبان انگلیسی بیان می‌کنند. این پژوهش بیشتر به تأثیر اندازه پنجره پرداخته است. پنجره‌ای که در این آزمایش استفاده شده همینگ می‌باشد. مشاهده شده است هرچه طول پنجره کوچکتر باشد، سیگنال نوسان بیشتری دارد و زمان پردازش نیز کمتر است. ولی اگر طول پنجره بزرگتر باشد سیگنال هموارتر

است و زمان پردازش نیز افزایش پیدا می‌کند. پس انتخاب طول پنجره مناسب، مهم است. شبیه‌سازی با استفاده از نرم‌افزار متلب و ضبط صدا با استفاده از نرم‌افزار پرات[1] انجام شده است. قسمت‌هایی که انرژی بیشتر است نشان‌دهنده قسمت‌های واکدار بیشتر می‌باشد و قسمت‌های بیواک انرژی کمتری دارند.

در مرجع [۱۰] نیز از انرژی برای تشخیص نقطه پایانی استفاده شده است. سیگنال‌های گفتاری ضبط شده افراد مختلفی هستند که به زبان ماندارین[2] صحبت می‌کنند. این سیگنال دارای ۲۸ کلمه است و هر کلمه هفت بار گفته می شود، پس تعداد کل کلمات ۱۹۶ کلمه می‌باشد. تعداد تشخیص اشتباه با استفاده از روش انرژی و نرخ عبور از صفر ۱۸ مورد است. پس به میزان ٪۹۰/۸ درست تشخیص داده شده است.

۲-۳- نرخ عبور از صفر[3]

یکی از الگوریتم‌های دیگری که برای تشخیص مرزها استفاده می‌شود نرخ عبور از صفر می‌باشد. نرخ عبور از صفر یعنی تعداد دفعاتی که سیگنال گفتار از صفر عبور می‌کند و تغییر علامت می‌دهد. این معیار یک اندازه گیری ساده از محتوای فرکانسی یک سیگنال می‌باشد. نرخ عبور از صفر سیگنال گفتار در رابطه (۲-۵) نشان داده می‌شود [۱۱]:

(۵-۲)

$$Z_n = \sum_{-\infty}^{\infty} \left| \operatorname{sgn}\left[x(m)\right] - \operatorname{sgn}\left[x(m-1)\right] \right| w(n-m) \qquad (۶-۲)$$

$$\operatorname{sgn}\left[x(n)\right] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$$

(۷-۲)

[1] Praat
[2] mandarin
[3] Zero Crossing Rate

$w(n)=\begin{cases}1/2(2N) & 0\le n\le N-1\\ 0 & otherwise\end{cases}$ در این روابط بالا x(m) سیگنال گفتار، w(n) پنجره مورد استفاده و N طول پنجره می‌باشد.

۲-۳-۱- نشانه تشخیص مرزها با استفاده از نرخ عبور از صفر

نرخ عبور از صفر برای تشخیص گفتار از نویز و تعیین قسمت شروع و پایان گفتار بسیار مفید است. در حالت سکوت نرخ عبور از صفر سیگنال صفر است. در حالت نویزی نرخ عبور از صفر نسبت به بخش‌های گفتاری بیشتر می‌باشد. با توجه به رابطه (۲-۵) تعداد تغییر وضعیت از مثبت به منفی یا برعکس را شمرده و آن مقدار را به عنوان ZCR در هر فریم در نظر گرفته می‌شود. محدوده‌ای از نمونه‌ها که نرخ عبور از صفر آن‌ها از مقدار آستانه بیشتر است، مرز کلمه در نظر گرفته می‌شود.

۲-۳-۲- بررسی کارهای گذشته

در مقاله [۱۱] علاوه بر انرژی، از روش نرخ عبور از صفر نیز برای تشخیص قسمت‌های واکدار و بیواک استفاده شده است. همانطور که گفته شد در این پژوهش بیشتر به تأثیر اندازه پنجره پرداخته شده است. پنجره مورد استفاده برای بررسی این معیار نیز همینگ می‌باشد. مانند روش انرژی مشاهده شده است هرچه طول پنجره کوچکتر باشد، سیگنال نوسان بیشتری دارد ولی اگر طول پنجره بزرگتر باشد سیگنال هموارتر است. طول پنجره تأثیر مهمی در نتیجه نهایی دارد.

در مرجع [۱۳] نرخ عبور از صفر یکی از روش‌هایی است که با روش‌هایی دیگر از جلمه انرژی آنتروپی و خطای پیش‌بینی خطی[1] ترکیب شده و باعث بهبود پاسخ سیستم برای تشخیص مرزها می‌شود. سیستم پیشنهادی از ۱۰ گوینده (۵ مرد و ۵ زن) و ۷۰ عبارت گفتاری از ۷ متن مجزا استفاده کرده است. جدول (۲-۱)، نتایج حداقل دقت به‌دست و جدول (۲-۲)، نتایج حداکثر دقت بدست آمده از این ارزیابی ها را نشان می دهد.

[1] Linear Prediction Error

جدول ۲-۱: مقایسه حداقل دقت به دست آمده در سه الگوریتم [۱۳]

Features	Minimum Accuracy
ZCR + Energy + LPE + Spectral Entropy (Proposed Algorithm)	90.35%
ZCR + Energy + Spectral Entropy	89.33%
ZCR + Energy + LPE	89.58%

جدول ۲-۲: مقایسه حداکثر دقت بدست آمده در سه الگوریتم [۱۳]

Features	Maximum Accuracy
ZCR + Energy + LPE + Spectral Entropy(Proposed Algorithm)	97.22%
ZCR + Energy + Spectral Entropy	95.44%
ZCR + Energy + LPE	96.20%

همانطور که در جدول (۲-۱) و (۲-۲) دیده می‌شود، روش پیشنهاد شده در مرجع[۱۳] با حداکثر دقت ٪۹۷/۲۲ و حداقل دقت ٪۹۰/۳۵ بهترین الگوریتم برای شناسایی مرزها بوده است. روش ZRC در ترکیب با سایر روش‌ها باعث بهبود نتایج می‌شود. برخی از خطاهایی که در سیستم رخ می دهد، به دلیل کاهش انرژی بخش گفتار و نویز در سیگنال گفتار است.

۲-۴- گام[1]

فرکانس ارتعاش تارهای صوتی فرکانس پایه[2] یا گام گفته می‌شود و آن را با F_0 نشان می‌دهند. فرکانس پایه تنها برای صداهای آوایی وجود دارد و گام از روش‌هایی است که برای تشخیص مرز کلمه استفاده می‌شود. هدف یافتن دوره تناوب ارتعاش تارهای صوتی در هر فریم از فایل گفتار است. محدوده فرکانس گام، در مردها در حدود ۵۰ تا ۲۵۰ هرتز، در زنان این فرکانس حدود ۱۵۰ تا ۴۵۰ هرتز و در کودکان بین ۳۰۰ تا ۷۰۰ هرتز است. تشخیص گام نقش بسیار مهمی در فشرده‌سازی گفتار، کدگذاری گفتار، شناسایی گوینده، تشخیص و ترکیب گفتار دارد. الگوریتم‌های تشخیص گام زیادی براساس طیف،کپستروم، شبکه‌های عصبی و تبدیل موجک وجود دارد، اما تعداد کمی از روش‌ها مانند تابع میانگین تفاوت اندازه (AMDF)[3] و تابع همبستگی خودکار(ACF)[4] در تشخیص پیچ کاربرد دارد. AMDF با گرفتن قدر مطلق بین اختلاف سیگنال گفتار اصلی و سیگنال گفتار شیفت یافته، در هر مقدار تأخیر محاسبه می شود. پیاده سازی AMDF محاسبات نسبتاً پایینی دارد و پیاده سازی آن آسان است. رابطه (۲-۸) تابع میانگین تفاوت اندازه را نشان می‌دهد.

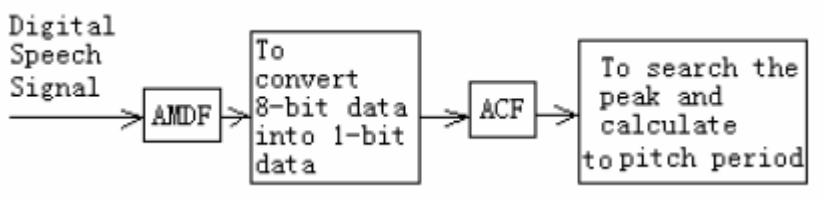
$$x_w(m) = \frac{1}{N-m-1} \sum_{n=0}^{N-m-1} \left| s_w(n+m) - s_w(n) \right| \qquad (۲-۸)$$

$s_W(n)$ سیگنال گفتار و N طول فریم سیگنال گفتار است. محدوده m بین ۰ تا N می‌باشد. یکی از معایب عمده AMDF برای تشخیص گام، شدت تغییرات سیگنال گفتار و نویز پس زمینه است. به منظور کاهش خطا، پیشنهاد شده است که همراه با AMDF از ACF استفاده شود که ACF در رابطه (۲-۹) نشان داده شده است:

$$R(k) = \sum_{n=0}^{N-k-1} x(n)x(n+k) \qquad (۲-۹)$$

[1] Pitch
[2] Fundamental Frequency
[3] Average Magnitude Difference Function (AMDF)
[4] Auto Correlation Function (ACF)

ابتدا مقادیر AMDF برای هر فریم سیگنال محاسبه می شود. مقادیر AMDF به سیگنال‌های یک بیتی تبدیل می شوند تا از پیچیدگی محاسبات در ACF کم شود و همچنین از اثرات دامنه و ساختار سیگنال در تشخیص گام کاسته شوند. سپس ACF مقادیر AMDF محاسبه می‌شود.



شکل ۲-۱: بلوک دیاگرام تشخیص گام با استفاده از AMDF و ACF [۱۴]

میزان خطا با استفاده از الگوریتم تشخیص گام بر اساس AMDF و ACF در مقایسه با AMDF کاهش می یابد، که در جدول (۲-۳) مقایسه شده‌اند [۱۴].

جدول ۲-۳: مقایسه خطای سه روش برای تشخیص گام [۱۴]

AMDF	ACF	ترکیب هر دو روش
۱۵٪	۹٪	۸٪

۲-۴-۱- الگوریتم تشخیص مرز کلمه با استفاده از روش گام

ابتدا با استفاده از روش‌های موجود برای تشخیص گام، گام هر فریم محاسبه می‌شود (در این پایان-نامه از روش خودهمبستگی[1] استفاده شده است). سپس محدوده‌ای از سیگنال که از مقدار آستانه بیشتر است به عنوان مرزکلمات انتخاب می‌شود.

۲-۴-۲- انجام کارهای گذشته

یکی از مرجع‌هایی که از روش گام برای شناسایی مرزهای کلمه استفاده کرده است، مرجع [۱۵] می‌باشد که از زبان‌های هندی و آلمانی به عنوان پایگاه‌داده استفاده کرده است. در زبان هندی مطالعات بر روی گفته‌های ده گوینده انجام شده و هر گوینده ۵۰ جمله را بیان کرده است. جملات دارای طول

[1] autocorrelation

متفاوت از ۵ کلمه تا ۱۵ کلمه بودند. برای زبان آلمانی، در مجموع از پنج گوینده استفاده شده که هر کدام ده جمله بیان کرده‌اند و به طور متوسط ۱۰ کلمه در هر جمله وجود دارد. نتایج مطالعه به وضوح نشان می دهد که زبان‌های هندی برخلاف زبان‌های اروپایی دارای عروضی مشابه هستند. در زبان‌های هندی، اشتباه در تشخیص مرزهای کلمات به دلیل وجود جملات مرکب ایجاد شده است. از نتایج می‌توان دریافت که بیش از ٪۸۵ از مرزهای کلمات در زبان‌های هند به درستی شناسایی شده‌اند، و بیش از ٪۶۵ از مرزهای کلمات با استفاده از روش گام در زبان آلمانی شناسایی می‌شوند.

۲-۵- آنتروپی[1]

آنتروپی فعالیت‌های تصادفی یک سیستم را اندازه‌گیری می‌کند. مشخص شده است که مقدار آنتروپی در تمایز بخش‌های گفتار در یک گفتار پیوسته ضبط شده از قسمت‌های غیرگفتاری، به ویژه در محیط-های پر سر و صدا، بسیار مفید است. وقتی یک سیستم منظم باشد، آنتروپی آن پایین‌تر است و برعکس، هرچه آشفته تر سیستم باشد، آنتروپی بالاتر است. در سال ۱۹۹۸، شن[2] و همکاران [۱۶] اولین نفراتی بودند که مفهوم آنتروپی را در تشخیص نقطه پایانی گفتار استفاده کردند. ابتدا طیف هر فریم با استفاده از تبدیل فوریه به دست می‌آید. از آنجا که این ویژگی فقط با تصادفی بودن انرژی ارتباط دارد و مستقل از دامنه انرژی سیگنال است، می تواند از محاسبات زیادی جلوگیری کرده و در برابر نویز قوی باشد.

اگر $x(n)$ سیگنال گفتار و $x_i(m)$ سیگنال گفتار مربوط به فریم iام بعد از اضافه کردن پنجره باشد، تبدیل فوریه آن از رابطه (۲-۱۰) به دست می‌آید:

$$X_i(k) = \sum_{m=0}^{N-1} x_i(m) \exp(-i\,2\pi km / N) \qquad (۲-۱۰)$$

و انرژی تبدیل فوریه $x_i(m)$ برابر است با:

$$Y_i(k) = |X_i(k)|^2 \qquad (۲-۱۱)$$

[1] Entropy
[2] Shen

تابع چگالی احتمال[1] طیف برای هر فریم از سیگنال گفتار برآورد می‌شود تا بر اساس آن آنتروپی طیفی تعریف و اندازه‌گیری شود. تابع چگالی احتمال برای طیف را می توان با نرمال‌سازی مؤلفه‌های فرکانس تخمین زد:

$$p(k,i)=\frac{Y_i(k)}{\sum_{l=0}^{N/2} Y_i(k)} \qquad k=0,1,....,N/2 \qquad (۱۲-۲)$$

آنتروپی طیفی هر فریم از رابطه (۲-۱۳) به دست می‌آید:

$$H(i)=-\sum_{k=0}^{N/2} p(k,i)\log p(k,i) \qquad (۱۳-۲)$$

۲-۵-۱- نشانه‌های تشخیص مرز کلمه با استفاده از روش آنتروپی

آنتروپی گفتار به دلیل ویژگی ذاتی طیف‌های گفتاری با نویز متفاوت است. مزیت اصلی آنتروپی کاهش حساسیت آن نسبت به تغییرات دامنه سیگنال گفتار است، در حالی که حساسیت بیشتری نسبت به وجود یا عدم وجود سیگنال نشان می دهد [۱۷]. آنتروپی در سیگنال‌های حاوی نویز دارای بیشترین مقدار، در فرکانس خالص مقدار صفر و در سیگنال گفتار حاوی نویز مقداری بین این دو خواهد داشت [۲]. در فرآیند تشخیص نقطه پایانی، مجموع مقادیر آنتروپی طیفی فریم‌ها ابتدا توسط یک فیلتر هموار می‌شود. مقدار آستانه‌ای در نظر گرفته می‌شود. در محدوده‌ای از سیگنال که مقدار آنتروپی از مقدار آستانه بیشتر باشد مرز شناخته می‌شود.

۲-۵-۲- بررسی کارهای گذشته

فی وو[2]برای بهبود آنتروپی طیفی، آنتروپی طیفی زیر باند را پیشنهاد کرد و دریافت که این پارامتر مرز بین گفتار و نویز را به وضوح در SNR های ضعیف تقویت می کند و برای محیط های پر سر و صدا قوی تر است. ایده آنتروپی طیفی زیر باند این است که یک فریم به تعدادی زیر باند تقسیم می شود،

[1] probability density function (pdf)
[2] Fei Wu

سپس آنتروپی طیفی زیر باند را بدست می آورد، بنابراین مشکل دامنه که تحت تأثیر نویزاست را از بین می برد [۱۸]. هر باند فرکانسی فریم را می توان به M زیر باند تقسیم کرد، سپس انرژی دامنه فرکانس هر زیر باند را محاسبه کرد. انرژی زیر باند mام از فریم iام به شرح است:

$$E_b(m,i)=\sum_{k=(m-1)*4}^{(m-1)*4+3} Y_i(k) \qquad 1\le m\le N_b \qquad (۱۴-۲)$$

وتابع چگالی احتمال هر زیر باند برابر است با:

$$p_b(m,i)=\frac{E_b(m,i)}{\sum_{k=1}^{N_b} E(m,i)} \qquad 1\le m\le N_b \qquad (۱۵-۲)$$

آنتروپی طیفی هر زیرباند از رابطه‌ی (۲-۱۶) به دست می‌آید:

$$H_b(i)=\sum_{m=1}^{N_b} p_b(m,i)\log p_b(m,i) \qquad (۱۶-۲)$$

روش پیشنهادی برای تقسیم زیرباندها باعث کاهش خطا و بهبود پاسخ شده است.

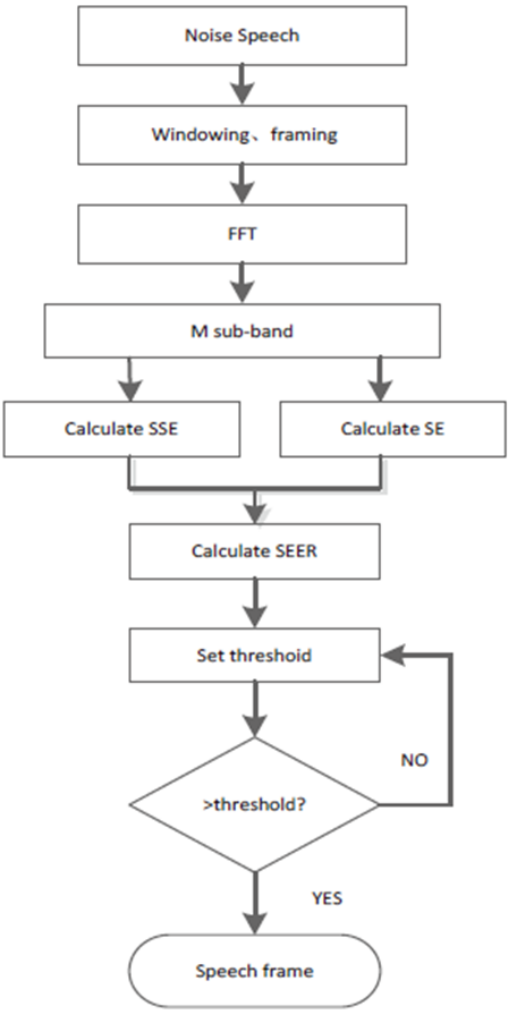
آزمایشات نشان می‌دهد، انرژی زیر باندکلمات زیاد است و آنتروپی طیفی زیر باند کوچک است. انرژی زیر باند محدوده نویز کوچک است و آنتروپی طیفی زیر باند بزرگ است. پس تفاوت دو محدوده کلمات و نویز باعث می‌شود تشخیص نقطه پایانی آسان تر باشد. ولی در محیط‌هایی که نویز زیاد است باعث افزایش اشتباه در تشخیص مرزها می‌شود. بنابراین در مقاله [۱۹] برای بهبود آنتروپی طیفی زیرباند، الگوریتم تشخیص نقطه پایانی نرخ انرژی-آنتروپی زیرباند[1] را پیشنهاد می‌کند که در رابطه (۲-۱۷) نشان داده شده است.

$$SEH(i)=\sqrt{1+\left|SE(i)/H_b(i)\right|} \qquad (۱۷-۲)$$

ابتدا سیگنال گفتار فریم‌بندی می‌شود. در این مقاله از فایل‌هایی با نرخ نمونه برداری ۸ کیلوهرتز استفاده می‌کنند، اندازه فریم ۲۵ میلی ثانیه یعنی ۲۰۰ نمونه، شیفت فریم ۱۰ میلی ثانیه، یعنی ۸۰ نمونه است. سپس انرژی زیر باند، آنتروپی طیفی زیر باند و نرخ انرژی-آنتروپی زیرباند محاسبه می‌شود.

[1] sub-band energy and sub-band spectral entropy (SEH)

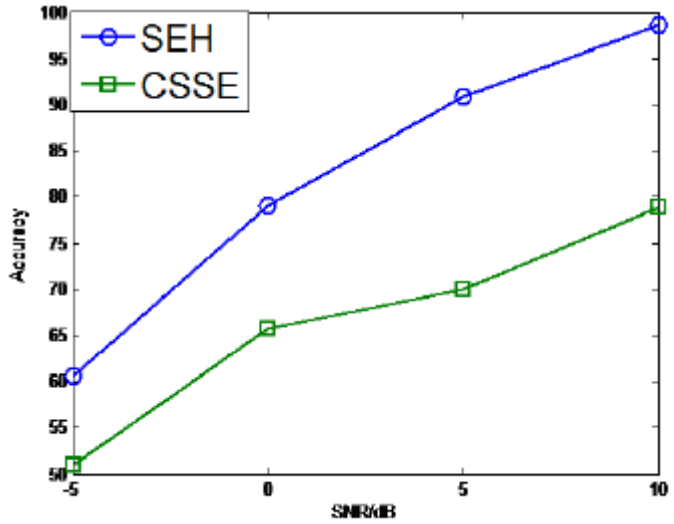
با استفاده از فیلتر برای بهبود سیستم هموار سازی انجام می‌شود. فرایند الگوریتم در شکل (۲-۲) نشان داده شده است. این مقاله با استفاده از Cool Edit Pro پایگاه داده صوتی ایجاد می کند، ۵۰ نمونه گفتار بدون نویز را برای آزمایش انتخاب می‌کند، سپس با نویزهای مختلف در SNR های متفاوت آزمایش شده است. این صداها از کتابخانه استاندارد نویز NOISEX92 گرفته شده است. نتایج آزمایش های شبیه سازی به مقایسه دقت نقطه پایانی گفتار آنتروپی طیفی زیر باند معمولی[1] و الگوریتم بهبود یافته تحت شرایط نویز مختلف و SNR های مختلف می پردازد. سپس نمونه‌های گفتار به ترتیب با الگوریتم ارائه شده در این مقاله به تشخیص نقطه پایانی می‌پردازند.



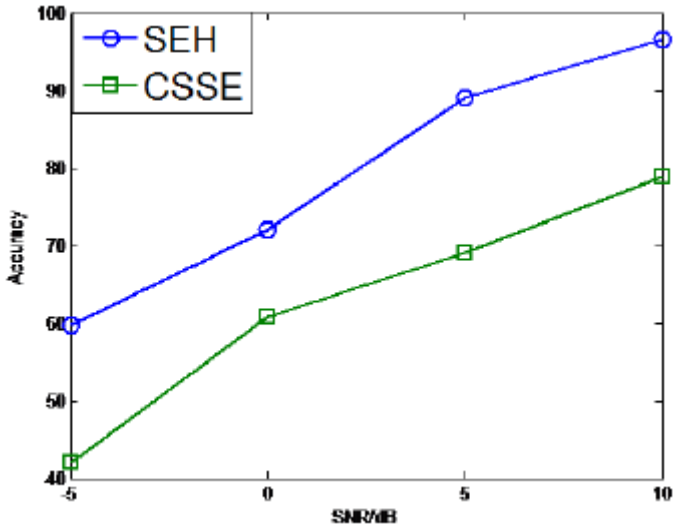
شکل ۲-۲: الگوریتم تشخیص گفتار با استفاده از روش آنتروپی و انرژی زیرباند [۱۹]

[1] conventional sub-band spectral entropy (CSSE)

از شکل (۲-۳) و (۲-۴) مشاهده می‌شود که الگوریتم پیشنهاد شده نسبت به روش آنتروپی طیفی تحت SNR های مختلف برتر است. روش آنتروپی طیفی مرسوم در حین تشخیص گفتار در SNR های کم دقت کمتری دارد ولی روش بهبود یافته آنتروپی طیفی می تواند نقاط پایانی را در SNR های بالاتر تشخیص دهد و برای بدست آوردن بخش گفتاری از نویز مفیدتر است، در نتیجه، دقت تشخیص افزایش می‌یابد.



شکل ۲-۳: تشخیص مرزها تحت نویز سفید[۱۹]



شکل ۲-۴: تشخیص مرزها تحت نویز همهمه [۱۹]

فصل سوم

استخراج ویژگی

۳-۱- مقدمه

در فصل قبل روش‌هایی که تاکنون برای تشخیص مرز کلمه استفاده شدند، بررسی شدند. سیگنالی که اطلاعات فرکانسی آن با گذشت زمان تغییر نمی‌کند سیگنال ایستا نامیده می‌شود. در مقابل سیگنال غیرایستا سیگنالی است که اطلاعات فرکانسیش با زمان تغییر می‌کند. نمایش سیگنال در حوزه زمان اطلاعاتی از فرکانس ندارد و زمان اتفاق افتادن فرکانس را نشان نمی‌دهد. همچنین تبدیل فوریه یک سیگنال فقط دارای محتوای فرکانسی است و هیچ اطلاعات زمانی‌ای ندارد. در نمایش سیگنال در حوزه زمان و فرکانس، متغیرهای زمان و فرکانس به صورت مجزا مورد بررسی قرار می‌گیرند. سیگنال گفتار یک سیگنال غیرایستا است و برای تحلیل بهتر باید در حوزه زمان- فرکانس[1] بررسی شود. نمایش زمان-فرکانس یعنی نمایشی از سیگنال که در هر دو حوزه زمان و فرکانس می‌باشد. نمایش سیگنال در حوزه فرکانس به دلیل عدم توانایی نمایش تغییرات فرکانسی در طول زمان، برای تحلیل سیگنال‌های غیرایستا مناسب نیست. پس نیاز به یک توزیع با توانایی بازنمایی تغییرات فرکانسی نسبت به زمان وجود دارد. توزیع‌های زمان-فرکانس به خوبی این تغییرات سیگنال‌ها را نشان می‌دهند. نمایش زمان- فرکانس هم اطاعات زمانی دارد و هم اطلاعات فرکانسی و برای سیگنال‌هایی که دارای چندین فرکانس متغیر در زمان هستند، مفید هستند [۲۰]. به طور کلی نمایش زمان-فرکانس یک ابزار قوی جهت پردازش سیگنال‌ها است. در ادامه برخی از روش‌های زمان- فرکانس و روش‌های طیفی بررسی می‌شود.

۳-۲- مقیاس مل[2]

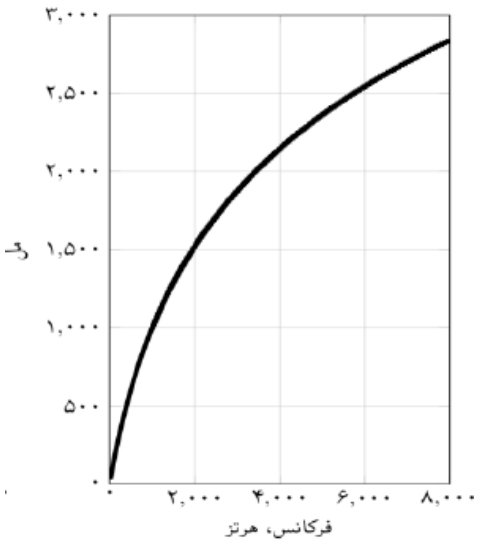
یکی از مقیاس‌های محبوب فرکانس ادراکی، مقیاس مل (مخفف ملودی[3]) است. این مقیاس برگرفته از اثر استونز و همکاران در سال ۱۹۳۷ می‌باشد [۲۱]. شنوایی انسان به اندازه برابر نسبت به همه باندهای فرکانسی یکسان نیست. عملکرد گوش برای فرکانس های کمتر از یک کیلوهرتز، خطی و برای

[1] Time-frequency
[2] Mel-scale
[3] Melody

فرکانس های بالاتر لگاریتمی است. برای تبدیل بین مقیاس ادراکی مل و فرکانس رابطه های متعددی استفاده می‌شود، اما رابطه‌ی رایج برای تبدیل فرکانس به مقیاس مل عبارت است از:

$$m(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \qquad (۱-۳)$$

شکل (۳-۱) باتوجه به رابطه (۳-۱)، مقادیر مقیاس مل نسبت به فرکانس را نشان می‌دهد.



شکل ۳-۱: تغییرات مقیاس مل برحسب رابطه (۳-۱) [۲۵]

۳-۲-۱- استخراج ویژگی ضرایب کپسترال در مقیاس مل[1]

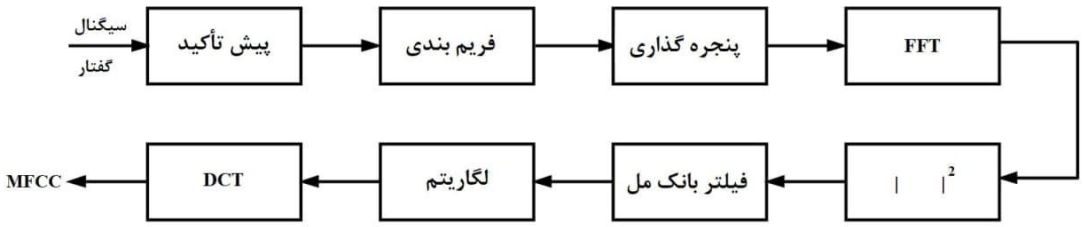
پردازش MFCC به عنوان پرکاربردترین روش استخراج ویژگی در سیستم تشخیص خودکار گفتار کاربرد دارد که در سال ۱۹۸۰ توسط دیویس[2] و مرملستین[3] ابداع شد [۲۱]. MFCC نوعی کپستروم حقیقی است که رفتار سیستم شنوایی انسان را شبیه سازی می‌کند. مراحل به دست آوردن ضرایب در شکل (۳-۲) نمایش داده شده است [۲۲].

گوش انسان به اطلاعات فرکانس پایین ارزش بیشتری می‌دهد. روش MFCC نیز نواحی فرکانس پایین را بهتر از نواحی فرکانس بالا نشان‌می‌دهد [۲۳]. بدترین ایراد این روش این است که در مورد

[1] Mel Frequency Cepstral Coefficients (MFCC)
[2] Davis
[3] Mermelstein

نویز ایمن نیست. بهتر از روش انرژی عمل می‌کند ولی در واقع نتایج MFCC در حضور نویز پس زمینه، خیلی قوی نمی‌باشد [۲۴] و [۲۵].

مراحل به دست آوردن ضرایب در شکل (۳-۲) نمایش داده شده است، که هر مرحله توضیح داده می‌شود:



شکل ۳-۲: مراحل استخراج ویژگی MFCC [۲۲]

- **پیش تأکید[1]:**

در مرحله پیش تأکید از فیلتر بالاگذر برای تقویت فرکانس های بالا و حذف اثرات طیفی حنجره و لب‌ها استفاده می‌شود [۲۴].

$$H(Z) = 1 - \mu Z^{-1} \qquad (۲-۳)$$

μ پارامتر پیش تأکید و مقدار آن بین ۰/۹ تا ۱ است.

- **فریم بندی[2]:**

پس از پیش تأکید سیگنال، سیگنال را فریم بندی می‌شود. طول فریم‌ها برای کاربردهای مختلف متفاوت است (بین ۱۰ تا ۵۰ میلی ثانیه). این فریم‌ها می‌توانند همپوشانی[3] داشته باشند، که این هم پوشانی بین ۲۵ درصد تا ۷۵ درصد طول فریم است.

- **پنجره گذاری[4]:**

برای تبدیل سیگنال غیرایستا به ایستا سیگنال فریم بندی شده در یک پنجره ضرب می‌شود،که

[1] pre-emphasis
[2] Frame Blocking
[3] overlap
[4] Windowing

معمولا این پنجره، همینگ می‌باشد. بعد از استفاده از پنجره ناپیوستگی بین مرز فریم‌ها از بین برود.

- **آنالیز طیفی:**

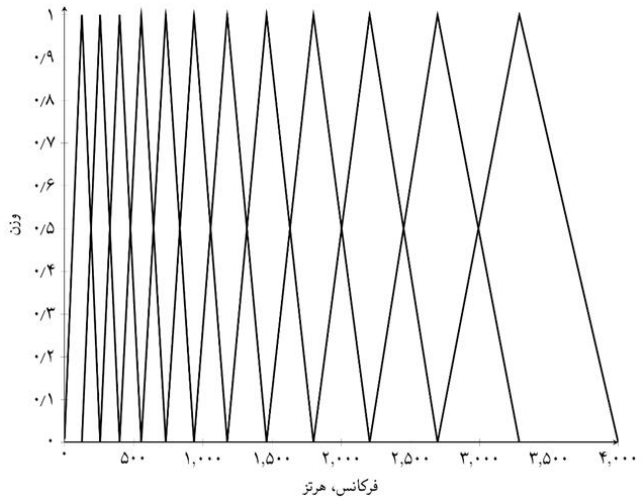
برای انتقال هر فریم گفتار به حوزه فرکانس از تبدیل فوریه گسسته[1] استفاده می‌شود آنالیز طیفی نام دارد. چون سیگنال حقیقی و زوج است، به دلیل تقارن نصف سیگنال نگه داشته می‌شود.

- **فیلتر بانک مل:**

در مرحله بعد، پردازش فیلتر بانک[2] مل انجام می‌شود. فیلتر بانک در مقیاس مل در واقع براساس شنیداری و مدل کردن حساسیت گوش انسان نسبت به حوزه های مختلف فرکانس است که این فرکانس ها با استفاده از رابطه (۳-۱) منتقل می‌شود. شکل (۳-۳) فیلتر بانک‌های مثلثی مل را نشان می‌دهند و در رایج‌ترین شکل، مجموعه‌ای از فیلترهای میان گذر (معمولا۱۲تا۳۰فیلتر) که دارای فرکانس‌های مختلف است تا محدوده مورد نظر از طیف سیگنال را پوشش دهد [۲۵].

- **لگاریتم و تبدیل کسینوسی گسسته:**

سپس لگاریتم انرژی فیلتر بانک‌ها محاسبه می‌شود و در نهایت با اعمال تبدیل کسینوسی گسسته[3] به اندازه تعداد فیلتر بانک‌ها ضرایب کپسترال مقیاس مل به دست می‌آید.



شکل ۳-۳: پاسخ فرکانسی فیلتربانک مل[۲۵]

[1] Discrete Fourier Transform (DFT)
[2] Filter Bank
[3] Discrete Cosine Transform (DCT)

۳-۳- تحلیل پیش‌بینی خطی[1]

تحلیل پیش بینی خطی یکی از قدرتمندترین و پرکاربردترین روش های تحلیل گفتار است. اهمیت این روش هم در توانایی آن در ارائه برآورد دقیق از پارامترهای گفتار و هم در سرعت نسبی محاسبه آن است [۲۶]. مشخصات پیش بینی خطی مزایای زیادی در گفتار و تشخیص خودکار گوینده دارد و این ضرایب به راحتی از طریق سیگنال گفتار تعیین می شوند. ضرایب پیش بینی کننده، اطلاعاتی در مورد فرکانس فرمنت‌ها، پهنای باند آن‌ها و شکل موج چاکنای را نشان می دهد. سرانجام، می توان از ضرایب پیش بینی برای بهبود روش های موجود در تشخیص خودکار گوینده و تشخیص مرزها بر اساس اطلاعات استفاده کرد [۲۷].

LP یک روش صحیح و قابل اعتماد برای تخمین پارامترهای سیستم خطی تولید گفتار برای بخش های واکدار می‌باشد. در واقع چون واکه‌ها دارای قطب هستند، فقط از مدل تمام قطب استفاده می‌شود.

$$H(z)=\frac{S(z)}{E(z)}=\frac{1}{1-\sum_{K=1}^{P}a_k z^{-k}} \qquad (۳-۳)$$

$S(z)$ سیگنال گفتار، $E(z)$ سیگنال تحریک[2] و P تعداد ضرایب پیش‌بینی خطی است. به ضرایب a_k ضرایب پیش‌بینی خطی یا ضرایب LP می‌گویند. ایده اولیه پیش‌بینی خطی براین اساس است که نمونه گفتار فعلی را می‌توان تقریبا به عنوان ترکیب خطی از نمونه‌های گذشته فرض نمود. با استفاده از فیلتر در رابطه (۳-۳) می‌توان نمونه زمانی لحظه n سیگنال خروجی یعنی s(n) را به دست آورد و با استفاده از ترکیب خطی P نمونه قبلی s را پیشگویی کرد که با استفاده از رابطه (۳-۴) نشان داده می‌شود:

$$\tilde{s}[n]=\sum_{k=1}^{p}a_k s[n-k] \qquad (۳-۴)$$

[1] Linear predictive analysis
[2] excitation

تفاوت بین نمونه‌های پیش‌بینی شده و نمونه‌های واقعی را خطای پیشگویی می‌گویند که با رابطه (۳-۵) نشان داده می‌شود:

$$e[n]=s[n]-\tilde{s}[n]=s[n]-\sum_{k=1}^{p}a_k s[n-k] \qquad (۳-۵)$$

ضرایب پیش‌بینی خطی را می‌توان با به حداقل رساندن مجموع تفاوت مربع بین نمونه‌های گفتار واقعی و آن‌هایی که به صورت خطی پیش‌بینی شده‌اند، به دست آورد که با رابطه (۳-۶) تعریف می‌شود:

$$E=\sum_{n=0}^{N-1}e^2[n]=\sum_{n=0}^{N-1}\left(s[n]-\sum_{k=1}^{p}a_k s[n-k]\right)^2 \qquad (۳-۶)$$

منجر به مجموعه ای از p معادله با p مجهول می‌شود که در رابطه (۳-۷) نشان داده شده است و می‌توان با روش‌های مختلفی آن را حل کرد.

$$\sum_{k=1}^{p}a_k\phi(i,k)=\phi(i,0) \quad i=1,2,...,p \qquad (۳-۷)$$

$\phi(i,k)$ تابع همبستگی سیگنال‌های تأخیر یافته توسط i وk می‌باشد. مقدار a_k به دست آمده فرض می‌شود همان مقدار واقعی است که در مدل تولید گفتار به کار می‌رود [۲۸].

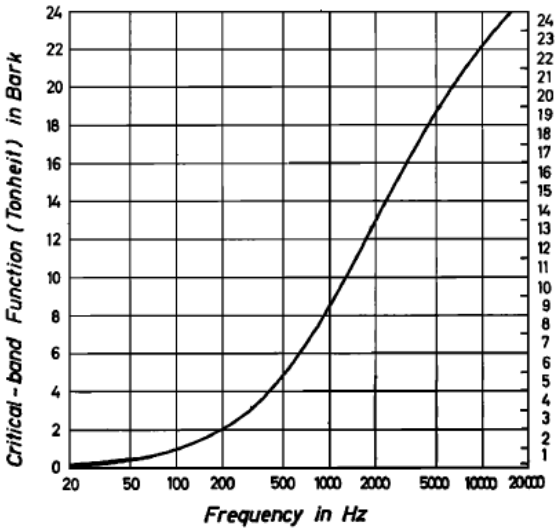
۳-۴- مقیاس بارک[1]

برای تغییر مقیاس فرکانس از مقیاس بارک استفاده شده است که فرکانس صوتی به یک فرکانس غیر خطی تبدیل می‌شود. مقیاس بارک نمایانگر نمایش بهتری از رزلوشن شنوایی انسان است. این مقیاس یک مقیاس آوایی ذهنی است که آقای زویکر[2] در سال ۱۹۶۰ با رابطه‌ی (۳-۸) نشان داده است [۲۸].

$$Bark(f)=13a\tan\left(\frac{0.76f}{1000}\right)+3.5a\tan\left(\frac{f^2}{(7500)^2}\right) \qquad (۳-۸)$$

شکل (۳-۵) باتوجه به رابطه (۳-۸)، مقادیر مقیاس بارک نسبت به فرکانس را نشان می‌دهد.

[1] Bark Scale
[2] Zwicker



شکل ۳-۴: تغییرات مقیاس بارک نسبت به رابطه (۳-۸) [۲۸]

برای مقیاس بارک رابطه‌های متعددی وجود دارد که رابطه‌های (۳-۹) و(۳-۱۰) از رابطه‌های دیگر این مقیاس می‌باشد.

$$Bark(f) = 25 + 75\left(1 + 1.4\left(\frac{f}{1000}\right)^2\right)^{0.69} \qquad (۳-۹)$$

$$Bark(f) = 650 \times \sinh\left(\frac{f}{7}\right) \qquad (۳-۱۰)$$

معیار مقیاس مل لگاریتمی و معیار مقیاس بارک عکس تانژانت است.

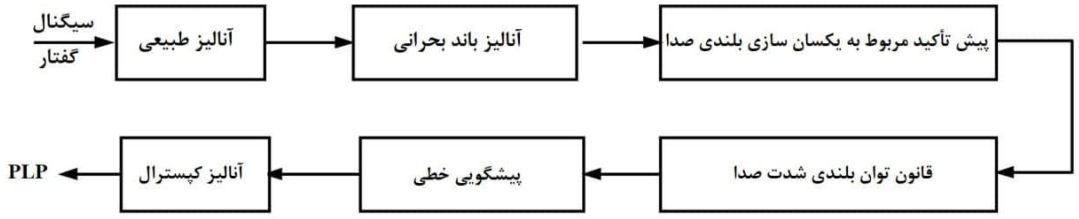
۳-۴-۱- تحلیل ادراکی پیشگویی خطی[1]

روش PLP تغییر و دگرگونی روش LPC است. از معایب روش LPC این است که مکانیزم سیستم شنوایی انسان را در نظر نمی‌گیرد و در تمامی فرکانس‌ها سیگنال گفتار را به یک صورت تحلیل می‌کند. در سال ۱۹۸۹ هرمانسکی[2] تحلیل ادراکی پیشگویی خطی که مبتنی بر درک انسان است را پیشنهاد کرد. ایده اصلی روش PLP این است که قدرت تفکیک فرکانسی و حساسیت نسبت به تغییر فرکانس در گوش انسان در فرکانس‌های مختلف، یکسان نیست و به علاوه حساسیت گوش نسبت به شدت و

[1] Perceptual linear predictive (PLP)
[1] Hermansky

انرژی صوت در فرکانس‌های مختلف متفاوت است [۳۰].

مراحل استخراج ویژگی با استفاده از روش PLP در شکل (۳-۶) آورده شده است که به طور خلاصه هر محله توضیح داده می‌شود:



شکل ۳-۵: مراحل استخراج ویژگی PLP

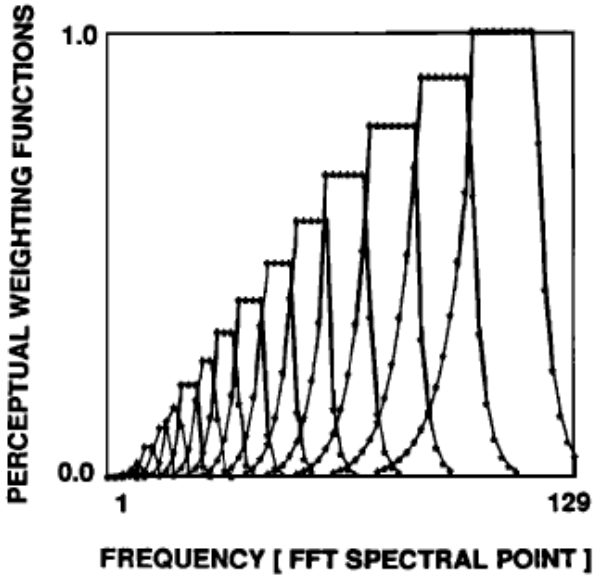
- **آنالیز باند طیفی:**

مانند روش MFCC ابتدا سیگنال را فریم بندی کرده و سپس در پنجره مناسب ضرب می‌شود. در مرحله آنالیز طیفی طیف توان زمان کوتاه هر سیگنال را با استفاده از تبدیل فوریه گسسته محاسبه می‌شود.

- **آنالیز باند بحرانی[1]:**

در این مرحله طیف‌های توان به دست آمده با استفاده از تابع انتقال یعنی رابطه (۳-۸) به حوزه بارک برده می‌شوند تا بعضی از فرکانس ها بر اساس شنوایی انسان اهمیت بیشتری پیدا کنند در واقع رزولوشن فرکانسی گوش را شبیه‌سازی می‌کند. فیلتربانک‌ها در فرکانس‌های پایین باریک‌تر و در فرکانس‌های بالا وسیع‌ترند و مجموعه‌ای از فیلترهای ذوزنقه‌ای به شکل (۳-۶) هستند.

[2] Critical band analysis



شکل ۳-۶: فیلتربانک بارک [۳۰]

- **پیش تأکید مربوط به یکسان سازی بلندی صدا[1]:**

این تابع احساس یکسان نبودن بلندی صدا در فرکانس‌های مختلف را جبران می‌کند و رابطه آن به شکل زیر می‌باشد:

$$E(f) = \left(\frac{f^2}{f^2 + 1.6e^5}\right) \cdot \left(\frac{f^2 + 1.44e^6}{f^2 + 9.6e^6}\right) \qquad (۱۱-۳)$$

- **قانون توان شدت بلندی صدا [2]:**

این تابع رابطه‌ای غیر خطی بین شدت صدا وبلندی صدای دریافت شده را شبیه سازی می‌کند. براین اساس نشان داده شده است که در گوش انسان میزان احساس بلندی با ریشه سوم انرژی آن متناسب است در این بلوک رابطه (۳-۱۲) به شکل زیر می‌باشد:

$$\phi(f) = \psi(f)^{0.33} \qquad (۱۲-۳)$$

- **محاسبه ضرایب پیشگویی خطی:**

در این مرحله استفاده از روشLP ضرایب پیشگویی خطی محاسبه می‌شود.

[1] Equal-loudness pre-emphasis
[2] Intensity to loudness

- **آنالیز کپسترال:**

بعد از به دست آوردن ضرایب LP، با عکس تبدیل فوریه گرفتن از لگاریتم ضرایب پیشگویی ضرایب PLP محاسبه می‌شود.

برخی آزمایش‌ها نشان می‌دهد که روش MFCC نتایج دقیق‌تری نسبت به روش PLP دارد ولی PLP مقاومت بیشتری نسبت به نویز دارد [۳۱].

۳-۵- تبدیل موجک[1]

تبدیل نگاشتی است که یک فضای ورودی را به یک فضای خروجی می‌برد. هدف این است که اطلاعات بیشتری برای تحلیل سیگنال دریافت شود که به صورت عادی قابل نمایش نیست. تبدیل موجک از مهم‌ترین روش‌های زمان-فرکانسی برای تجزیه و تحلیل سیگنال می‌باشد [۳۲]. موجک با بکارگیری توابعی، بر برخی از محدودیت‌های تحلیل فوریه غلبه می‌کند. در واقع آنالیز فوریه حالت بسیار خاصی از موجک است ولی ضعف فوریه در وهله اول عدم قابلیت نمایش زمان-فرکانس است و نمی‌تواند وضوح [2]را بالا ببرد و هر جزئیاتی را نمی‌بیند که ناشی از این است که عملکرد آن سرتاسری[3] است و نمی‌تواند به خوبی وفق[4] شود در حالی که تبدیل موجک به دلیل خاصیت محلی[5] بودن می‌تواند به خوبی سیگنال را تحلیل کند. همانطور که گفته شده سیگنال‌های غیرایستا سیگنال‌هایی هستند که فرکانس‌ها و سایر خصوصیات آماری آن‌ها با گذشت زمان تغییر می‌کند. تبدیل موجک برای سیگنال‌های غیرایستا که متناوب نیستند، بسیار مناسب است. تابع موجک قسمتی از شکل را بررسی می‌کند یعنی علاوه بر فرکانس، موقعیت و محل نیز در نظر گرفته می‌شود [۳۳]. یک سیگنال را با استفاده از تبدیل موجک می‌توان به مجموعه‌ای از توابع پایه که به توابع مادر معروف هستند، به صورت پی در پی به

[1] Wavelet Transform
[2] resolution
[3] global
[4] Adapt
[5] local

سیگنال‌های تقریب[1] و جزئیات[2] تقسیم کرد. اطلاعات فرکانس سیگنال در فرآیند تجزیه به وسیله ضرایب تقریب و ضرایب جزئیات بیان می‌شود. رابطه (۳-۱۳) رابطه بین موجک مادر و فرزندان می‌باشد:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \qquad (۱۳-۳)$$

خواص اصلی یک تابع موجک:

۱) از نظر زمانی، به طرز مناسبی باید محدود باشد. یعنی:

بزرگترین مقدار LB وجود داشته باشد، به نحوی که:

$$\forall t \le \mathrm{LB} \;:\; \psi(t) = 0 \qquad (۱۴-۳)$$

و کوچکترین مقدار UB وجود داشته باشد، به نحوی که:

$$\forall t \ge \mathrm{UB} \;:\; \psi(t) = 0 \qquad (۱۵-۳)$$

۲) باید دارای میانگین صفر باشد:

$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0 \qquad (۱۶-۳)$$

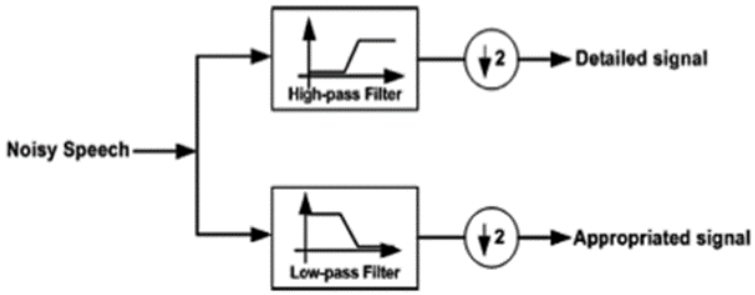
۳) نرم غیر صفر داشته باشد:

$$0 < \int_{-\infty}^{\infty} |\psi(t)|^2 \, dt < \infty \qquad (۱۷-۳)$$

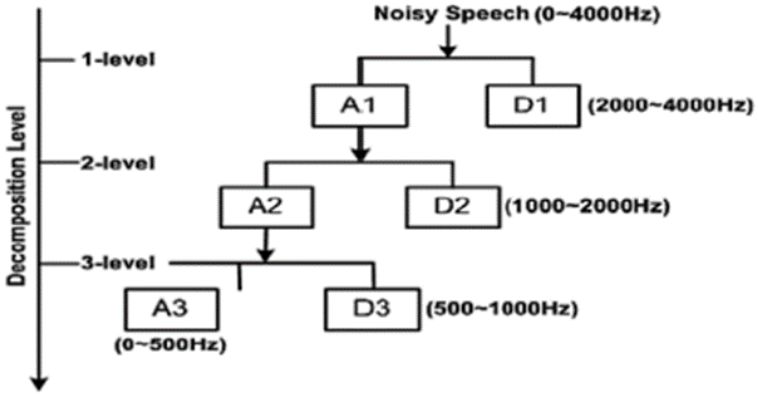
صدای انسان از دو مؤلفه فرکانس پایین و فرکانس بالا تشکیل شده است. با اعمال فیلترهای پایین گذر[3] و بالاگذر[4] بر روی سیگنال اصلی و در مرحله بعد بر روی سیگنال تقریب، سیگنال‌های تقریب و جزئیات به دست می‌آیند. خروجی فیلتر پایین‌گذر مؤلفه تقریب و خروجی فیلتر بالاگذر مؤلفه جزئیات است. شکل (۳-۷) تبدیل موجک گسسته[5] با اعمال فیلتربانک‌ها نشان می‌دهد. پس از اعمال

[1] Approximating
[2] Detail
[3] Low-pass filter
[4] High-pass filter
[5] Discrete Wavelet Transform (DWT)

فیلتر، نمونه برداری با فاکتور دو انجام می‌شود و حجم اطلاعات کاهش می‌یابد. آن‌قدر این کار را ادامه داده می‌شود تا سیگنال تقریب صفر یا نزدیک صفر شود. تعداد سطح‌ها[1] می‌تواند متغیر باشد. هرچه سیگنال ساده‌تر باشد مؤلفه تقریب زودتر به صفر می‌رسد. همچنین می‌توان با اعمال یک بلوک بازسازی کننده[2] سیگنال اصلی یا تقریبی از آن را بدست آورد. در شکل (۳-۸) نمودار مراحل تجزیه یک سیگنال به سه سطح نشان داده شده است. [۳۴].



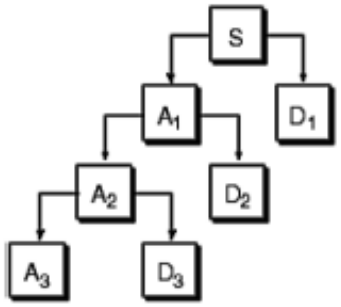
شکل ۳-۷: تبدیل موجک گسسته با استفاده از فیلتر بانک [۳۴]



شکل ۳-۸: نمودار مراحل تجزیه سیگنال به سه سطح [۳۴]

تعداد نمونه‌ها پس از تجزیه ثابت می‌باشد که در شکل (۳-۹) نشان داده شده است.

[1] level
[2] reconteractor



$$S = A_1 + D_1$$
$$= A_2 + D_2 + D_1$$
$$= A_3 + D_3 + D_2 + D_1$$

شکل ۳-۹: مراحل تجزیه سیگنال به سه سطح [۳۴]

در حقیقت، یک گفتارآوایی دارای انرژی بیشتری نسبت به گفتار غیرآوایی در سیگنال است. بنابراین، می توان از انرژی سیگنال زیر باند برای استخراج فعالیت گفتاری استفاده کرد [۳۴].

۳-۶- توزیع ویگنر[1]

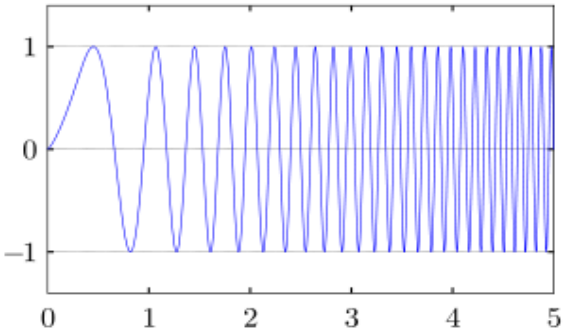
تابع توزیع ویگنر اولین بار توسط ایگن ویگنر[2] در سال ۱۹۳۲ برای توضیح مکانیک کوانتوم مورد استفاده قرار گرفت [۳۵] و در ۱۹۴۸ توسط ویل[3] در پردازش سیگنال مورد استفاده قرار گرفت [۳۶]. این تابع یکی از توزیع‌های مهم در حوزه زمان-فرکانس است و جزو توزیع‌های درجه دوم می‌باشد. یعنی توزیعی که یک سیگنال را از حوزه زمان به صفحه چگالی انرژی نگاشت می‌دهد و این توزیع‌ها همزمان تفکیک پذیری فرکانسی و زمانی بالایی دارند [۳۷].

توزیع ویگنر-ویل برای یک سیگنال به صورت تبدیل فوریه تابع همبستگی وابسته به زمان که به اندازه τ اختلاف زمانی دارند به صورت رابطه (۳-۱۸) می‌باشد:

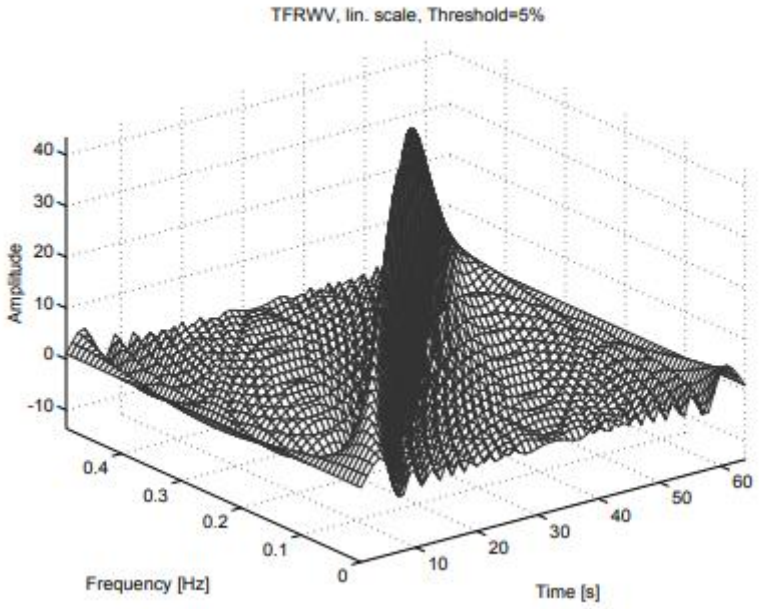
$$W_x(t,v) = \int_{-\infty}^{+\infty} x\left(t+\tau/2\right) x^*\left(t-\tau/2\right) e^{-j2\pi v\tau} d\tau \qquad (۱۸-۳)$$

سیگنال چیرپ[4] در شکل (۳-۱۱) نشان داده شده است. توزیع ویگنر این سیگنال شکل (۳-۱۱) است. ویگنر یک روش تحلیل سیگنال در حوزه زمان-فرکانس است. مشاهده می‌شود که شکل (۳-۱۱) هم اطلاعات زمانی وهم اطلاعات فرکانسی دارد، با افزایش زمان فرکانس نیز افزایش می‌یابد.

[1] Wigner Distribution
[2] Eugene Wigner
[3] ville
[4] chirp



شکل ۳-۷: سیگنال چیرپ [۴۰]



شکل ۳-۸: تبدیل ویگنر- ویل سیگنال چیرپ [۴۰]

خصوصیات مختلف و مطلوب توزیع ویگنر مانند تفکیک پذیری زمانی و فرکانسی بالا و موارد دیگر، توزیع ویگنر را به ابزاری مفید برای تجزیه و تحلیل سیگنال تبدیل می‌کند. اشکال اصلی این توزیع درجه دوم بودن آن و ایجاد عبارات متقاطع[1] بین فرکانس‌های مثبت و منفی سیگنال است که باعث می‌شود تفسیر فضای تبدیل را دشوار می کند. از راه‌هایی که می‌توان توزیع ویگنر را بهبود بخشید از

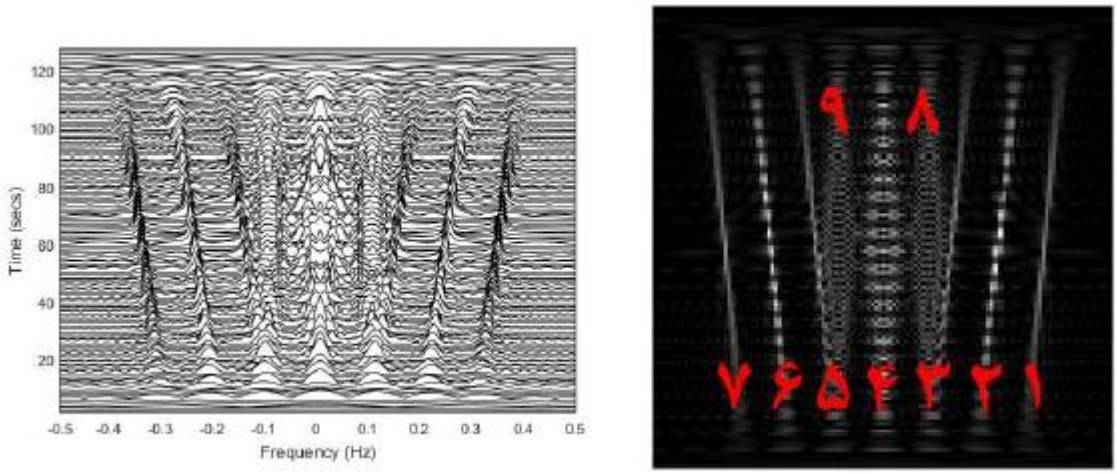
[1] cross terms

بین بردن فرکانس‌های منفی است تا عبارات متقاطع ایجاد نشود. عبارات متقاطع با مخلوط کردن، فیلتر کردن و یا اضافه کردن تابعی از بین می‌روند. به سیگنالی که در آن عبارات متقاطع وجود ندارد، سیگنال تحلیلی گفته می‌شود. توزیع ویگنر-ویل همان توزیع ویگنر است با این تفاوت که به جای سیگنال اصلی از سیگنال تحلیلی استفاده می‌شود. برای یک سیگنال که از M مؤلفه تشکیل شده است تبدیل ویگنر به صورت زیر می‌باشد [۳۸]:

$$x(t)=\sum_{i=1}^{M} x_i(t) \qquad (۱۹-۳)$$

$$W_x(t,\omega)=\sum_{i=1}^{M} W_{x_i}(t,\omega)+\sum_{k=1}^{M-1}\sum_{l=k+1}^{M} 2\Re\left[W_{x_k x_l}(t,\omega)\right] \qquad (۲۰-۳)$$

در رابطه (۳-۲۰) عبارت دوم جمله تداخلی می‌باشد. در واقع مجموع ویگنر دو سیگنال دقیقا برابر با ویگنر مجموع دو سیگنال نمی‌باشد.



شکل ۳-۹: وجود عبارات متقاطع در توزیع ویگنر [۳۷]

در شکل (۳-۱۲) توزیع ویگنر یک سیگنال دو مؤلفه‌ای می‌باشد. بخش‌های ۲ و ۶ به ترتیب عبارات متقاطع بین مؤلفه‌های سیگنال ۱و۳ و همچنین ۵و۷ هستند. بخش ۴ عبارت متقاطع مضاعف حاصل از مؤلفه‌های ۳و۵ و همچنین ۱و۷ است. یعنی بخش ۴ دارای دو عبارت متقاطع است. بخش ۸ عبارت متقاطع بین مؤلفه‌های ۱و۵ و بخش ۹ عبارت متقاطع بین مؤلفه‌های ۳و۷ است. پس عبارت متقاطع بین دو مؤلفه سیگنال اصلی قرار می‌گیرد [۳۷].

۳-۷- ویژگی‌های توزیع ویگنر- ویل

توزیع ویگنر دارای خواص و ویژگی‌هایی می‌باشد که باعث شده این توزیع در برخی از کاربردها مفید باشد. چند نمونه از مهم‌ترین ویژگی‌ها دز زیر بیان شده است [۴۰]:

۱) نگه داری انرژی[1]:

با انتگرال‌گیری از توزیع ویگنر-ویل تابع x روی کل صفحه زمان- فرکانس انرژی کل سیگنال به صورت رابطه نشان (۳-۲۱) داده می‌شود:

$$E_x = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W_x(t,v)\,dtdv \qquad (۲۱-۳)$$

۲) شرایط حاشیه‌ای[2]:

چگالی طیفی انرژی و توان لحظه ای سیگنال x را می توان به عنوان شرایط حاشیه ای توزیع ویگنر- ویل از روابط (۳-۲۲) و(۳-۲۳) به دست آورد. شرایط حاشیه ای یعنی انتگرال در راستای فرکانس صفحه‌ی زمان-فرکانس که چگالی انرژی است و انتگرال در راستای زمان صفحه‌ی زمان-فرکانس که برابر با طیف توان سیگنال است، وجود داشته باشد.

(۲۲-۳)

$$\int_{-\infty}^{+\infty} W_x(t,v)\,dt = \left|X(v)\right|^2$$

$$\int_{-\infty}^{+\infty} W_x(t,v)\,dv = \left|X(t)\right|^2 \qquad (۲۳-۳)$$

۳) مقادیر حقیقی[3]:

تابع توزیع ویگنر- ویل به ازای هر مقدار t وv دارای مقادیر حقیقی است.

$$W_x(t,v) \in \mathbb{R}, \quad \forall t,v \qquad (۲۴-۳)$$

۴) شیفت دادن:

[1] Energy conservation
[2] Marginal properties
[3] Real-valued

از دیگر خواص توزیع ویگنر-ویل شیفت دادن است. یعنی اگر سیگنال اصلی در زمان یا فرکانس شیفت پیدا کند توزیع ویگنر-ویل نیز در همان زمان یا فرکانس همانند روابط (۳-۲۵) و (۳-۲۶) شیفت پیدا می‌کند.

$$y(t)=x(t-t_0)\Rightarrow W_y(t,v)=W_x(t-t_0,v) \qquad (۳-۲۵)$$

$$y(t)=x(t)e^{j2\pi v_0 t}\Rightarrow W_y(t,v)=W_x(t,v-v_0) \qquad (۳-۲۶)$$

۵) فیلتر کردن:

اگر سیگنال y برابرکانوولوشن x و h باشد، بنابراین توزیع ویگنر-ویل سیگنال y برابر با کانوولوشن زمانی سیگنال‌های h و x است که در رابطه (۳-۲۷) نشان داده شده است.

$$y(t)=\int_{-\infty}^{+\infty}h(t-s)\,x(s)ds\Rightarrow W_y(t,v)=\int_{-\infty}^{+\infty}W_h(t-s,v)\,W_x(s,v)ds \qquad (۳-۲۷)$$

۶) فرکانس لحظه‌ای[1]:

یعنی فرکانس سیگنال در یک لحظه خاص را نشان می‌دهد و از گشتاور مرتبه اول در فرکانس توزیع ویگنر- ویل به دست می‌آید که در رابطه (۳-۲۸) نشان داده شده است:

$$f_x(t)=\frac{\int_{-\infty}^{+\infty}vW_{xa}(t,v)dv}{\int_{-\infty}^{+\infty}W_{xa}(t,v)dv} \qquad (۳-۲۸)$$

۷) تأخیر گروهی[2]:

این کمیت از گشتاور مرتبه اول در زمان توزیع ویگنر-ویل به دست می‌آید که در رابطه (۳-۲۹) نشان داده شده است.

$$t_x(v)=\frac{\int_{-\infty}^{+\infty}tW_{xa}(t,v)\,dt}{\int_{-\infty}^{+\infty}W_{xa}(t,v)\,dt} \qquad (۳-۲۹)$$

۸) خاصیت پارسوال:

[1] Instantaneous frequency
[2] Group delay

اندازه انرژی در حوزه زمان با انرژی در حوزه فرکانس برابر است، که در رابطه (۳-۳۰) نشان داده شده است:

$$\left|\int_{-\infty}^{+\infty} x(t)y^{*}(t)dt\right|^{2} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W_{x}(t,v)W_{y}^{*}(t,v)dtdv \qquad (۳۰-۳)$$

فصل چهارم

روش پیشنهادی و نتایج شبیه سازی

۴-۱- مقدمه

در این فصل ابتدا پایگاه داده مورد استفاده معرفی می‌شود و سپس نتایج شبیه‌سازی الگوریتم‌های استخراج ویژگی گوناگون برای تشخیص مرز کلمه که در این پایان نامه آورده شده است، نمایش داده می‌شود. نتایج تحلیل می‌شوند و بهترین نتیجه مشخص می‌گردد.

۴-۲- پایگاه داده[1]

پایگاه داده TIMIT یکی از معروفترین و پرکاربردترین پایگاه داده انگلیسی مورد استفاده در پردازش گفتار است که نتیجه تلاش مشترک محققان MIT، SRI و TI است. همانطور که در جدول (۴-۱) نشان داده شده است، این پایگاه داده از ۲۳۴۲ جمله مجزا از سه مجموعه مختلف تشکیل شده است. ضبط داده‌ها توسط محققان TI انجام شده است. در مجموع ۶۳۰۰ جمله جمع‌آوری شد که از هر ۶۳۰ گوینده، ده جمله ضبط شده است. تقسیم‌بندی جملات در جدول (۴-۱) نشان داده شده است.

جدول ۴-۱: تقسیم بندی جملات پایگاه داده TIMIT [۴۱]

Sentence Type	Sentences	Talker	Sentences/Talker
Dialect (SIR)	2	1260	2
Compact (MIT)	450	3150	5
Diverse (TI)	1890	1890	3
Total	2342	3600	10

جملات مجموعه Dialect مربوط به لهجه می‌باشند. جملات مجموعه Compact از لحاظ واج متراکم می‌باشند و جملات مجموعه Diverse دارای واج‌های مختلف هستند و تقریبا دارای همه‌ی واج‌ها می‌باشد. این پایگاه داده به دو دسته‌ی آزمایش[2] و آموزش[3] تقسیم بندی می‌شود که حدود ۲۰ تا ۳۰ در صد مربوط به آزمایش و حدود ۷۰ تا ۸۰ درصد مربوط به آموزش می‌باشد. هیچ‌کدام از گویندگان در

[1] data base
[2] test
[3] train

هر دو دسته عضویت ندارند. تعداد جملات، تعداد کلمات و بعضی اطلاعات دیگر در جدول (۴-۲) نشان داده شده است[۴۱].

جدول ۴-۲: شرح پایگاه‌داده TIMIT [۴۱]

تعداد جملات	۵۰۴۰
تعداد کلمات	۴۱۱۶۱
میانگین تعداد کلمات خاص	۵۱۰۷
میانگین تعداد کلمات در هر جمله	۸/۲
میانگین تعداد هجا در هر کلمه	۱/۵

۴-۳- پیش پردازش[1]

به منظور استخراج ویژگی‌ها از سیگنال گفتار، باید ابتدا یک سری پیش‌پردازش‌های اولیه روی سیگنال انجام شود. در این پایان‌نامه برای شبیه‌سازی نتایج الگوریتم‌های مورد نظر از نرم‌افزار متلب استفاده شده است. سعی شده است که موارد مورد استفاده در تمامی الگوریتم‌ها یکسان باشد تا بهتر بتوان نتایج را مقایسه کرد. پنجره مورد استفاده همینگ و طول پنجره ۳۲۰ نمونه است. میزان شیفت پنجره ۲۰ نمونه می‌باشد. همچنین به منظور هموارسازی سیگنال و حذف فرکانس‌های بالا از فیلتر باترورث[2] استفاده می‌شود. در پردازش سیگنال برای هموارسازی شکل‌ها معمولا از این فیلتر استفاده می‌شود. برای شبیه‌سازی الگوریتم‌ها از دونوع سیگنال گفتار استفاده شده است:

داده اول مربوط به CADLab است که جملات به طور واضح و شمرده گفته شده است [۴۲]. داده دوم از پایگاه‌داده TIMIT است که گوینده به طور طبیعی و تقریبا سریع صحبت کرده و و زمان بین کلمات کم است.

[1] Preprocessing
[2] Butterworth

مقادیر آستانه در هر مرحله به صورت تجربی و با سعی و خطا بدست می‌آید تا نتیجه بهتری حاصل شود. یعنی مقدارهای مختلفی برای آستانه در نظر گرفته می‌شود و نتیجه شبیه‌سازی با توجه به آن مقادیر آستانه نمایش داده می‌شود. مقدار آستانه‌ای که بهترین نتیجه را ارائه می‌دهد، انتخاب می‌شود.

نتیجه شبیه‌سازی روش‌های مختلف برای تشخیص مرز کلمات تصویری است ک دارای ۴ بخش می‌باشد. بخش اول مرزهای کلمه روی سیگنال گفتار با استفاده از اطلاعات موجود در پایگاه داده و به صورت دستی مشخص شده است. در بخش دوم با استفاده از روش های مختلف ویژگی‌های مورد نظر استخراج می‌شوند. در بخش سوم با استفاده از فیلتر باترورث هموار سازی انجام می‌شود تا تغییرات سیگنال نرم‌تر شود و راحت‌تر بتوان مرزها را تشخیص داد و مقدار آستانه اعمال می‌شود. در بخش آخر مرزهای کلمه با استفاده از روش استخراج ویژگی مورد نظر به دست آمده‌اند.

لازم به یادآوری است به دلیل این‌که پایگاه‌داده مورد استفاده در این تحقیق با پایگاه‌داده مقالات یکی نبودند، ابتدا روش‌های متداول شبیه‌سازی شده‌اند و نتایج آورده شده است.

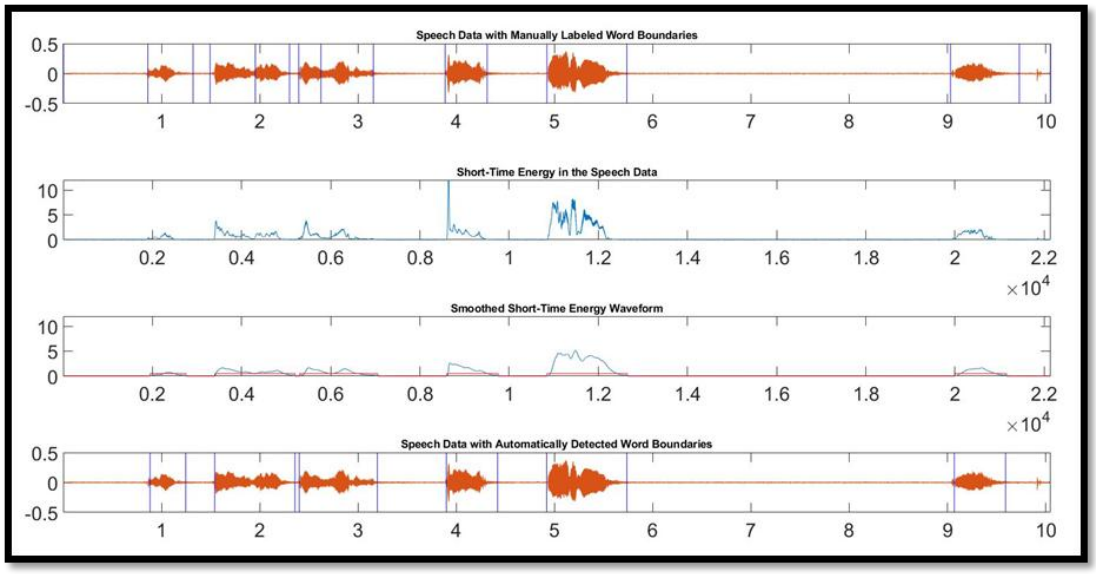
۴-۴- نتایج شبیه‌سازی روش‌های متداول

در ابتدا نتایج روش‌های متداولی که از نوع تحلیل زمانی هستند و برای تشخیص مرزهای کلمه استفاده می‌شدند، بررسی می‌شود.

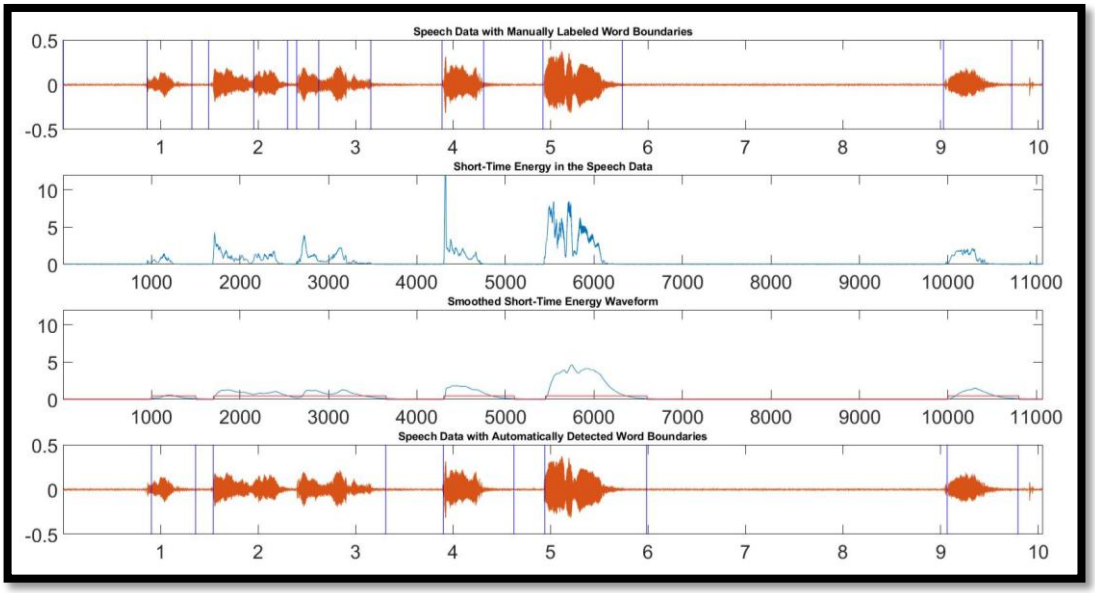
۴-۴-۱- نتایج شبیه‌سازی روش انرژی زمان-کوتاه

الگوریتم جداسازی مرز کلمات با استفاده از انرژی زمان-کوتاه بدین صورت است که ابتدا به منظور بررسی زمان-کوتاه سیگنال گفتار، سیگنال ورودی با ضرب در پنجره‌هایی به طول ۳۲۰ نمونه و با شیفت‌های زمانی ۲۰ ، ۴۰ و ۸۰ نمونه، فریم بندی و پنجره‌گذاری می گردد و سپس انرژی هر فریم محاسبه می شود. برای محاسبه انرژی، نمونه‌های هر پنجره را به توان دو رسانده و باهم جمع می‌شوند. در مرحله بعد به منظور نرم‌ترکردن تغییرات منحنی مربوط به انرژی، مقادیر بدست آمده برای انرژی زمان کوتاه توسط فیلتر باتروث فیلتر می گردد. در نهایت با توجه به میزان آستانه تعریف شده، محدوده

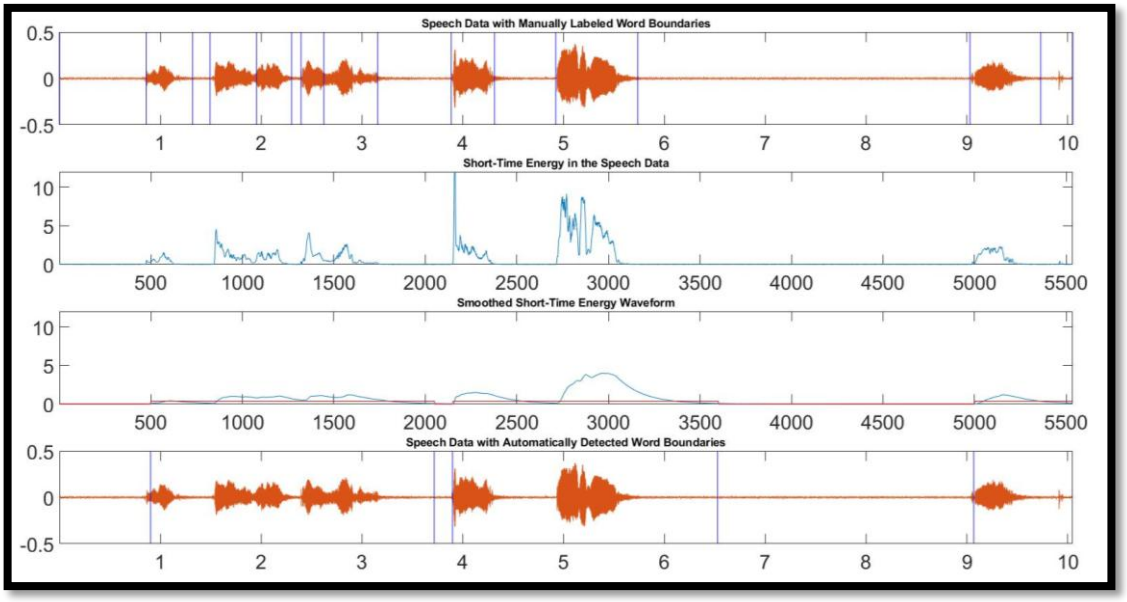
هایی از انرژی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردد.



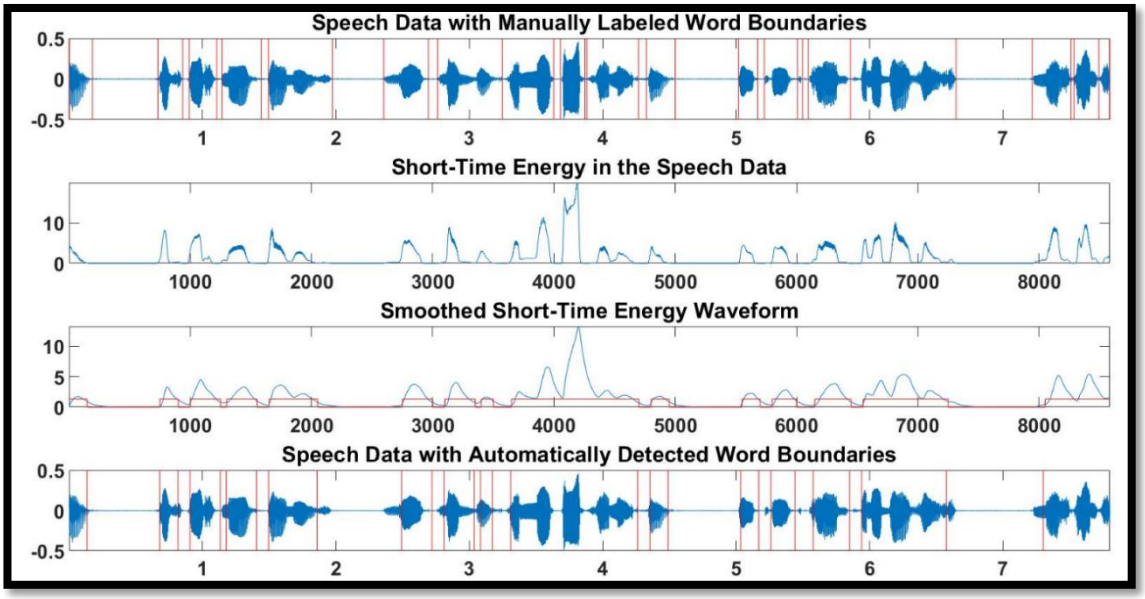
شکل ۴-۱: تشخیص مرز کلمه با استفاده از روش انرژی برای داده اول با طول پنجره ۳۲۰ نمونه و شیفت زمانی ۲۰ نمونه



شکل ۴-۲: تشخیص مرز کلمه با استفاده از روش انرژی برای داده اول با طول پنجره ۳۲۰ نمونه و شیفت زمانی ۴۰ نمونه



شکل ۴-۳: تشخیص مرز کلمه با استفاده از روش انرژی برای داده اول با طول پنجره ۳۲۰ نمونه و شیفت زمانی ۸۰ نمونه



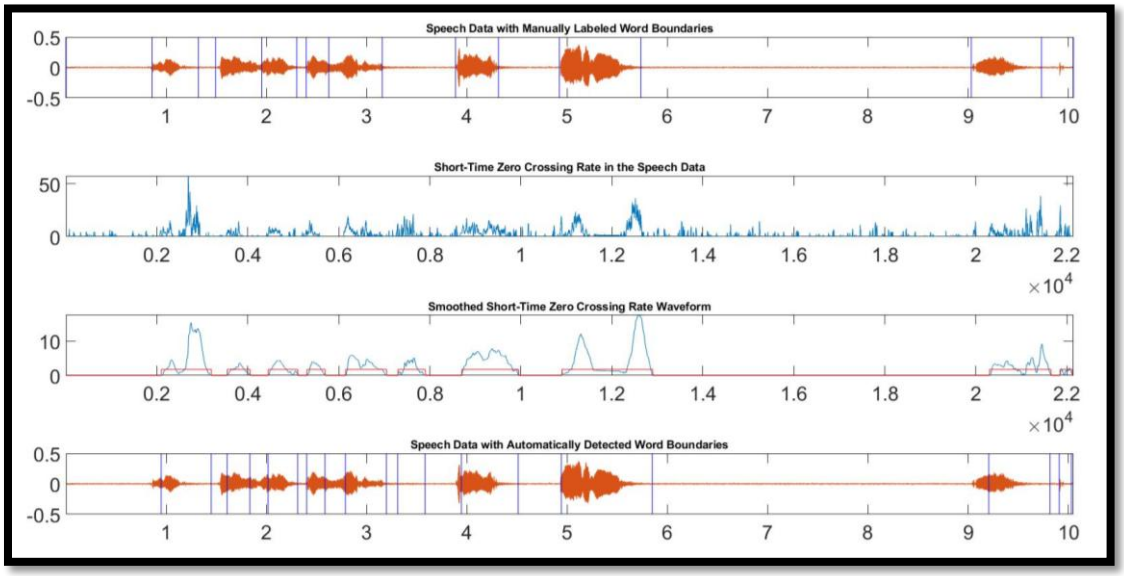
شکل ۴-۴: تشخیص مرز کلمه با استفاده از روش انرژی برای داده دوم با طول پنجره ۳۲۰ نمونه و شیفت زمانی ۲۰ نمونه

با مقایسه نتایج بدست آمده در اشکال (۴-۱) تا (۴-۴)، مشاهده می‌شود که میزان شیفت زمانی در دقت جداسازی کلمات بسیار تاثیرگذار است. همانطور که در فصل دوم بررسی شد، اگر ابعاد پنجره کوچک باشد سیگنال نوسان زیادی دارد ولی اگر ابعاد پنجره افزایش یابد سیگنال هموارتر است. اگر ابعاد پنجره بدون آن‌که میزان شیفت زمانی تغییر کند، افزایش یابد، زمان پردازش بیشتر می گردد بدون آنکه دقت جداسازی تغییر کند. اگر شیفت زمانی خیلی کوچک انتخاب شود، زمان پردازش به

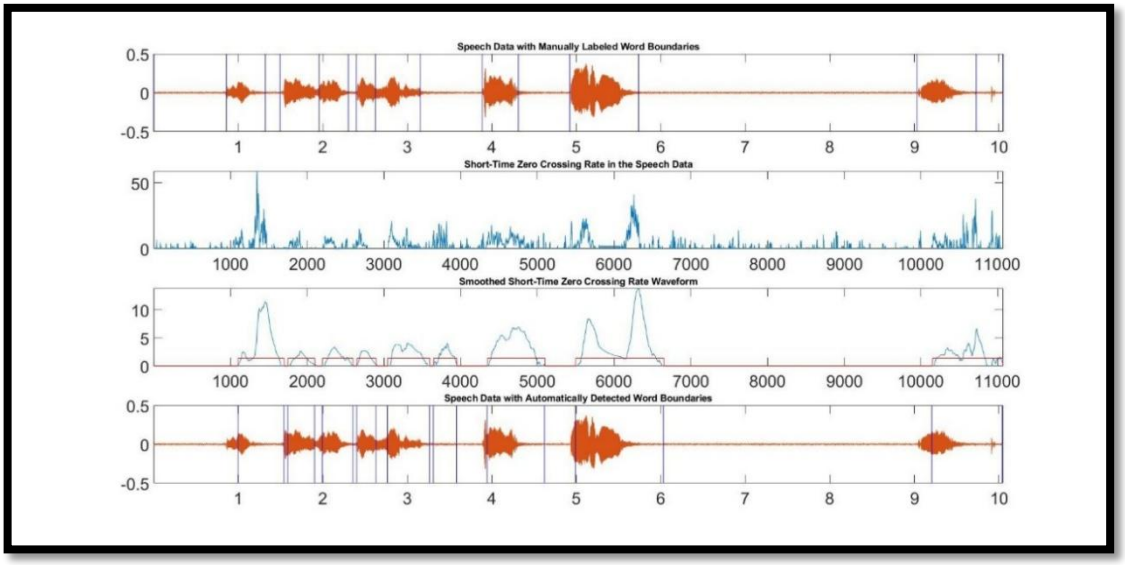
شدت افزایش یافته و دقت جداسازی کلمات نیز کاهش می یابد. بدین صورت که مرزهای بیشتری برای کلمات پیدا می کند. از طرف دیگر اگر میزان شیفت زمانی خیلی بزرگ (به نسبت ابعاد پنجره) انتخاب گردد، در اینصورت دقت جداسازی کلمات کاهش می یابد در این حالت هر چند کلمه را به عنوان یک بخش در نظر می‌گیرد. به همین خاطر انتخاب ابعاد پنجره و مهم تر از آن میزان شیفت زمانی در دقت جداسازی مرز کلمات تاثیر بسیار زیادی دارد.

۴-۴-۲-نتایج شبیه سازی روش نرخ عبور از صفر

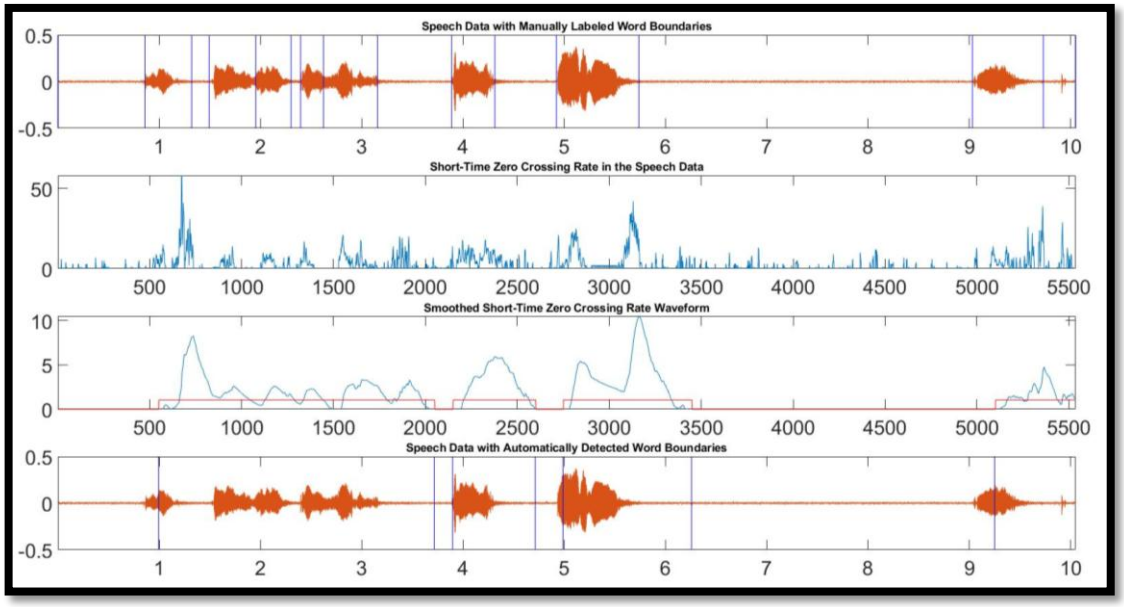
برای تشخیص مرز کلمات با استفاده از معیار گذر از صفر زمان-کوتاه ابتدا به منظور بررسی زمان-کوتاه سیگنال گفتار، سیگنال ورودی با ضرب در پنجره هایی به طول ۳۲۰ نمونه و با شیفت‌های زمانی ۲۰، ۴۰ و ۸۰ نمونه، فریم بندی می گردد و سپس نرخ عبور از صفر هر فریم محاسبه می شود. چون تغییرات خروجی سیگنال خروجی زیاد است، مقدار DC خروجی این مرحله حذف می شود تا حول صفر نوسان کند. در مرحله بعد به منظور نرم‌تر کردن تغییرات منحنی مربوط به این معیار، مقادیر بدست آمده برای مقادیر گذر از صفر زمان-کوتاه توسط فیلتر باترورث فیلتر می گردد و در نهایت با توجه به میزان آستانه تعریف شده، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردد.



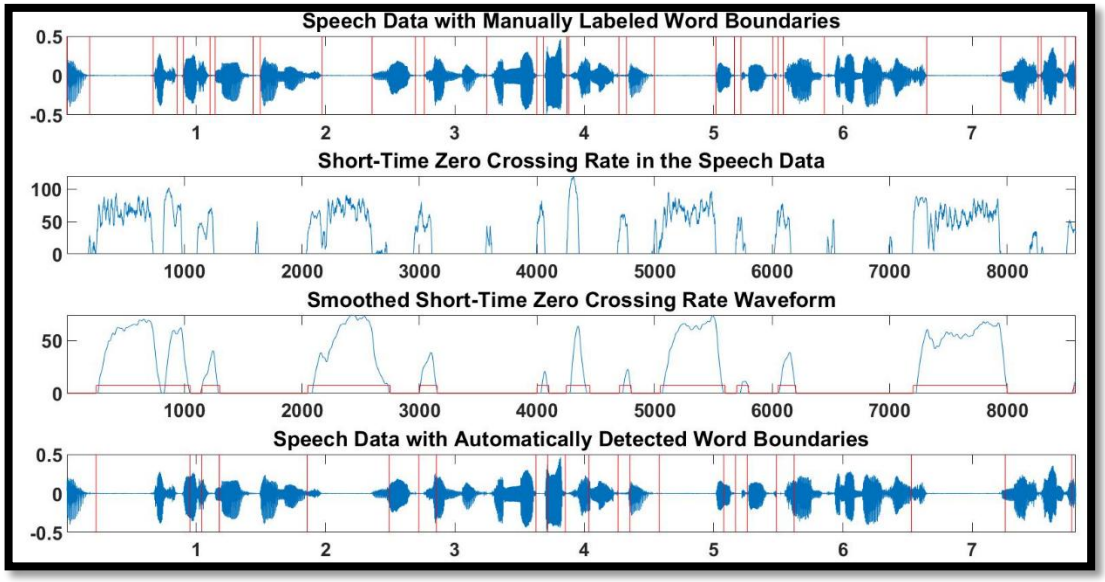
شکل ۴-۵: تشخیص مرز کلمه با استفاده از نرخ عبور از صفر برای داده اول و طول پنجره ۳۲۰ نمونه و شیفت زمانی ۲۰ نمونه



شکل ۴-۶: تشخیص مرز کلمه با استفاده از نرخ عبور از صفر برای داده اول و طول پنجره ۳۲۰ نمونه و شیفت زمانی ۴۰ نمونه



شکل ۴-۷: تشخیص مرز کلمه با استفاده از نرخ عبور از صفر برای داده اول و طول پنجره ۳۲۰ نمونه و شیفت زمانی ۸۰ نمونه



شکل ۴-۸: تشخیص مرز کلمه با استفاده از نرخ عبور از صفر برای داده دوم با طول پنجره ۳۲۰ نمونه و شیفت زمانی ۲۰ نمونه

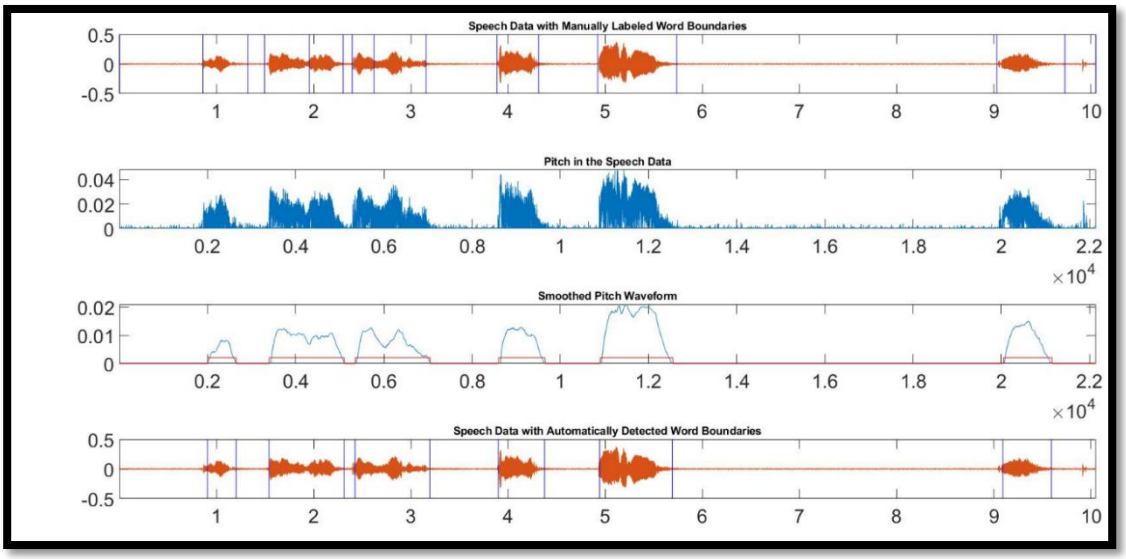
با مقایسه نتایج بدست آمده در اشکال (۴-۵) تا (۴-۸)، مشاهده می‌شود که ابعاد پنجره و میزان شیفت زمانی در دقت جداسازی کلمات بسیار تأثیرگذار است. با بررسی ابعاد پنجره در فصل دوم مشخص شد ابعاد پنجره در نتیجه نهایی تأثیر دارد. اگر ابعاد پنجره بدون آن‌که میزان شیفت زمانی تغییر کند، افزایش یابد، زمان پردازش بیشتر می گردد بدون آن‌که دقت جداسازی تغییر کند. اگر شیفت زمانی خیلی کوچک انتخاب شود، زمان پردازش به شدت افزایش یافته و دقت جداسازی کلمات نیز کاهش می یابد بدین صورت که مرزهای بیشتری برای کلمات پیدا می کند. از طرف دیگر اگر میزان شیفت زمانی خیلی بزرگ (به نسبت ابعاد پنجره) انتخاب گردد، در اینصورت دقت جداسازی کلمات کاهش می یابد و هر چند کلمه را به عنوان یک بخش در نظر می‌گیرد. به همین خاطر انتخاب ابعاد پنجره و مهم تر از آن میزان شیفت زمانی در دقت جداسازی مرز کلمات تاثیر بسیار زیادی دارد. در شکل (۴-۸) برای داده دوم مشاهده می‌شود.

۴-۴-۲- نتیجه شبیه سازی با استفاده از روش گام

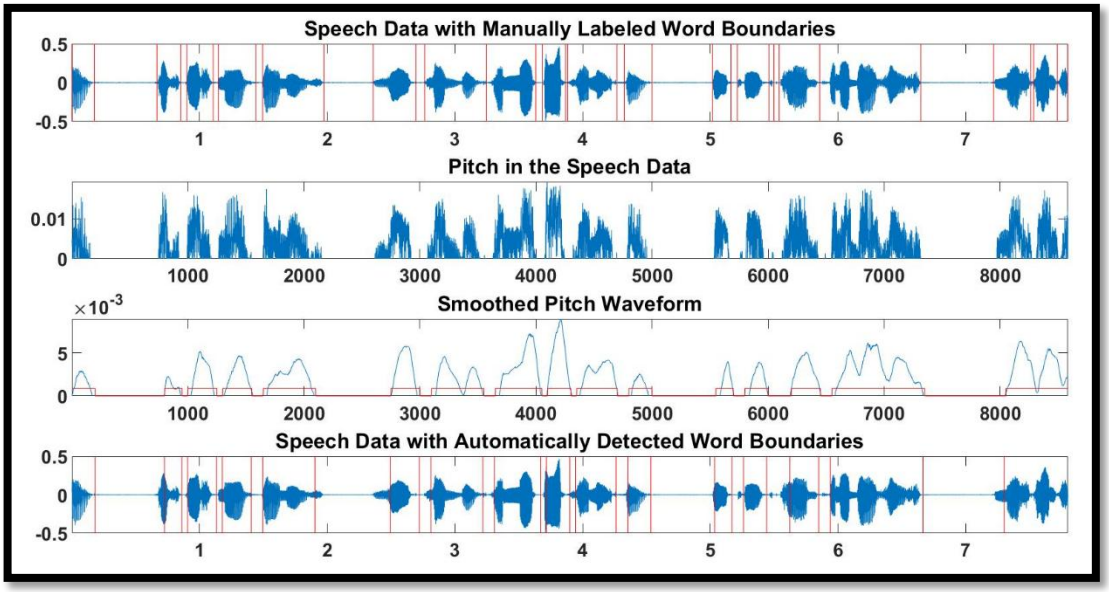
همانطور که در فصل دو گفته شد، روش‌های مختلفی برای یافتن پیچ سیگنال گفتار استفاده می‌شود. در این پایان‌نامه برای به دست آوردن پیچ از خودهمبستگی سیگنال استفاده شده است. ابتدا به منظور

بررسی زمان-کوتاه سیگنال گفتار، سیگنال ورودی با ضرب در پنجره به طول ۳۲۰ نمونه و با شیفت زمانی ۲۰ نمونه، فریم بندی می‌گردد. سپس با استفاده از خودهمبستگی گام فریم‌های واکدار استخراج شده، محاسبه می شود. خود همبستگی سیگنال یعنی محاسبه میزان شباهت سیگنال و مقدار شیفت یافته‌اش. در مرحله بعد به منظور کم‌کردن تغییرات خروجی، مقدار DC خروجی این مرحله حذف می شود. برای نرم‌تر کردن تغییرات منحنی مربوطه، مقادیر بدست آمده توسط فیلتر باتروث فیلتر می گردد و در نهایت با توجه به میزان آستانه‌ای که از روش سعی و خطا به دست آمده، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند

همانطورکه در شکل (۴-۹) و (۴-۱۰) با مقایسه نمودار مربوط به نتایج جداسازی کلمات بصورت دستی و استفاده از الگوریتم فرکانس گام مشاهده می گردد، الگوریتم مبتنی بر فرکانس گام برای هر دو داده توانسته است بیشتر مرزهای کلمات را پیدا کند، اما برخی از مرزها را پیدا نکرده است و در آن قسمت‌ها چند کلمه با هم در یک محدوده قرار گرفته اند.



شکل ۴-۹: تشخیص مرز کلمه با استفاده از روش گام برای داده اول



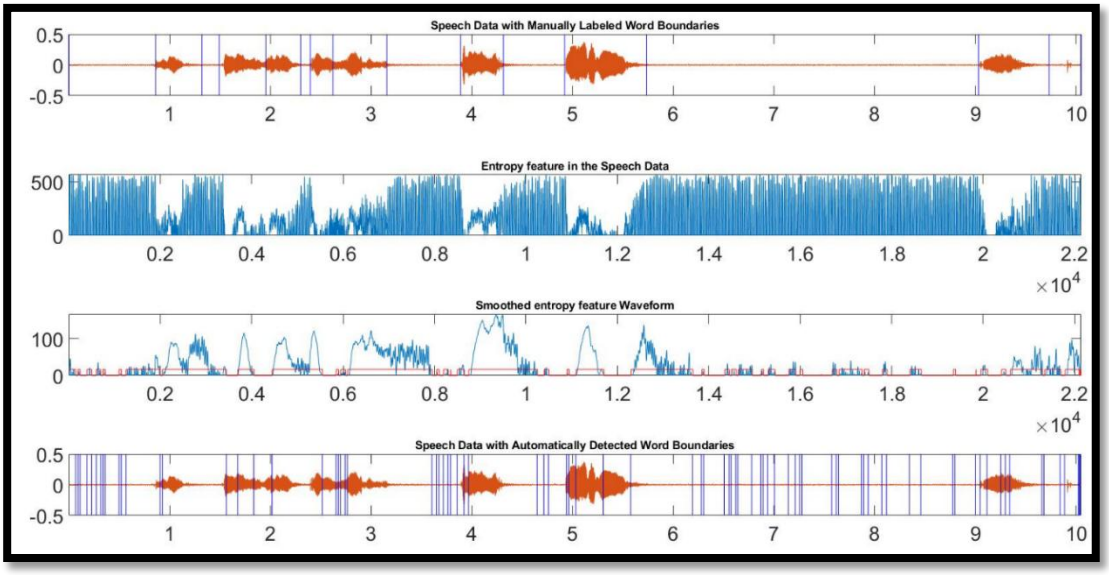
شکل ۴-۱۰: تشخیص مرز کلمه با استفاده از روش گام برای داده دوم

۴-۴-۳- نتیجه شبیه‌سازی روش آنتروپی

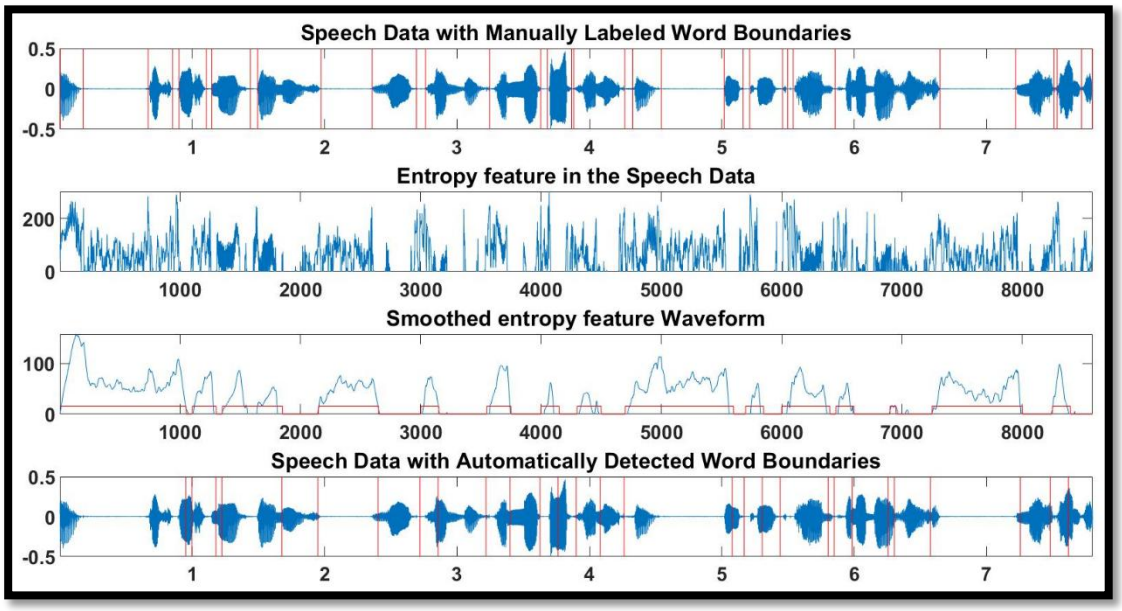
برای تشخیص مرز کلمات با استفاده از آنتروپی هر فریم برای بررسی زمان-کوتاه سیگنال گفتار، سیگنال ورودی با ضرب در پنجره به طول ۳۲۰ نمونه و با شیفت زمانی ۲۰ نمونه، فریم بندی می گردد. سپس مقادیر فریم ها به ماکزیمم مقدار در هر فریم نرمالیزه می گردند. در مرحله بعد مقدار آنتروپی مربوط به هر فریم محاسبه شده و سپس مقدار DC خروجی این مرحله حذف می شود. در اینجا از سیگنال اصلی برای محاسبه آنتروپی استفاده شده است. به منظور نرم تر کردن تغییرات منحنی مربوطه، مقادیر بدست آمده توسط فیلتر باتروث فیلتر می گردد و سپس مقدار DC خروجی این مرحله حذف می شود و در نهایت با توجه به میزان آستانه تعریف شده، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند.

همانطورکه در شکل (۴-۱۱) و (۴-۱۲) با مقایسه نمودار مربوط به نتایج جداسازی کلمات به صورت دستی و استفاده از الگوریتم مربوط به آنتروپی مشاهده می گردد، الگوریتم مبتنی بر آنتروپی در مشخص کردن مرز کلمات برای هر دو داده نتوانسته است موفق عمل کند و به اشتباه مرزهای زیادی را برای کلمات پیدا کرده است. به همین جهت این روش نمی تواند روش قوی‌ای برای شناسایی مرز کلمات

باشد.



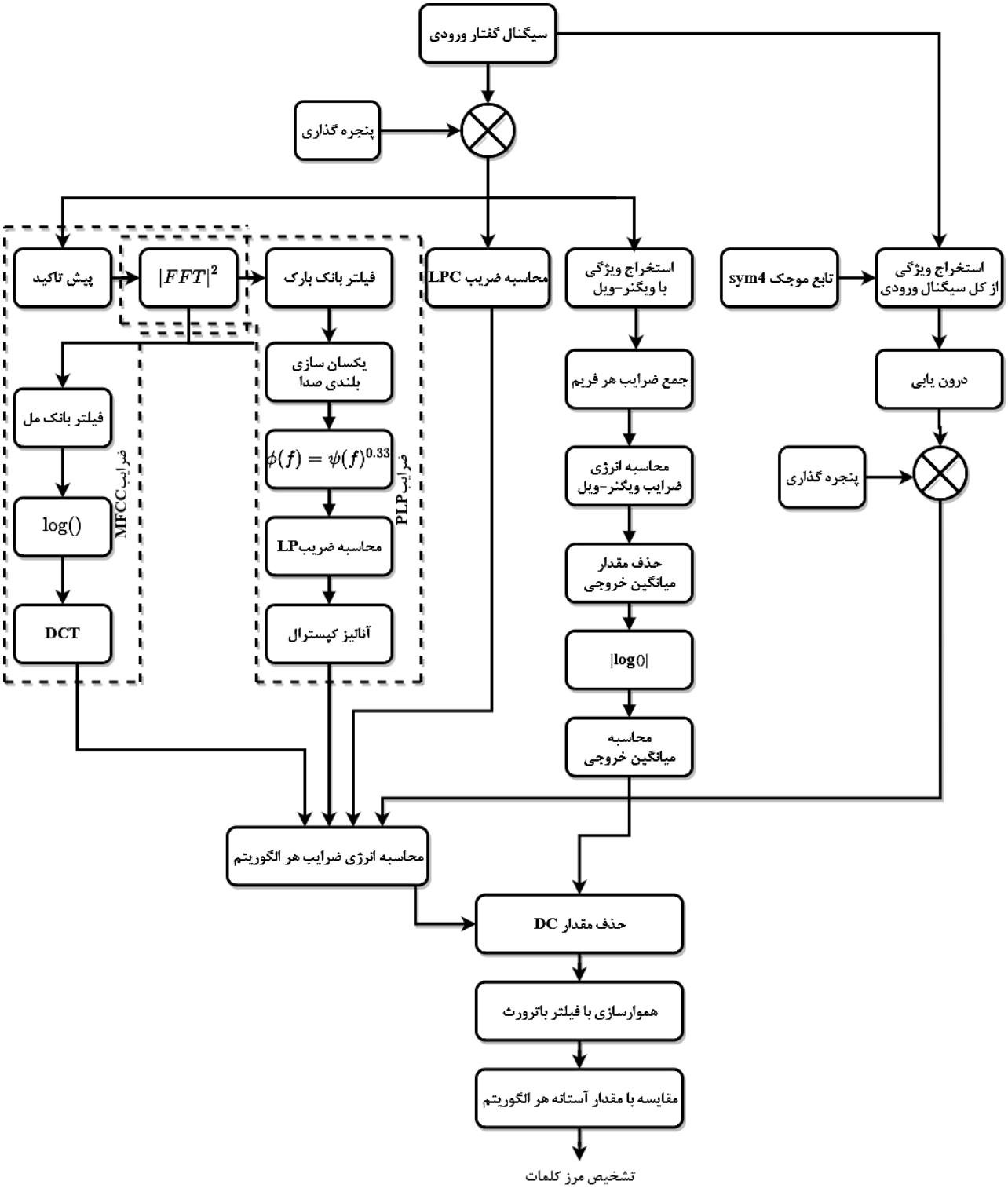
شکل ۴-۱۱: تشخیص مرز کلمه با استفاده از روش آنتروپی برای داده اول



شکل ۴-۱۲: تشخیص مرز کلمه با استفاده از روش آنتروپی برای داده دوم

۴-۵- بلوک دیاگرام کلی روش‌های پیشنهادی

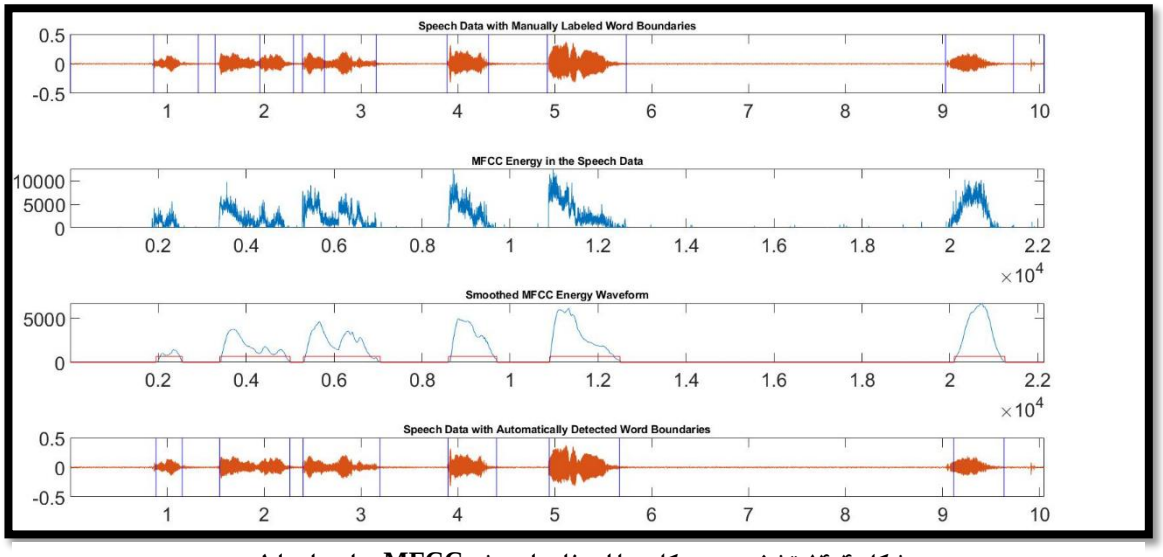
در شکل (۹-۴) بلوک دیاگرام پیشنهادی برای تشخیص مرز کلمات نشان داده شده است.



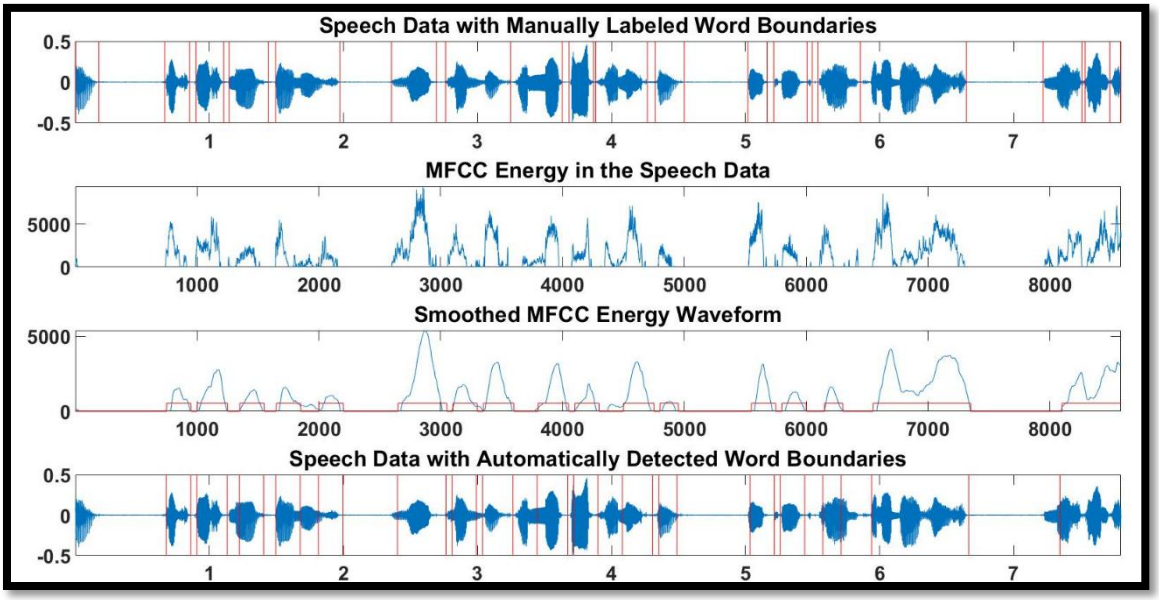
شکل ۴-۱۳: بلوک دیاگرام پیشنهادی برای تشخیص مرز کلمات

۴-۵-۱- نتیجه شبیه‌سازی روش MFCC

در الگوریتم تشخیص مرز کلمات با استفاده از روش استخراج ویژگی MFCC به منظور بررسی زمان-کوتاه سیگنال گفتار، سیگنال ورودی با ضرب در پنجره به طول ۳۲۰ نمونه و با شیفت زمانی ۲۰ نمونه، فریم بندی می گردد. سپس توسط روش MFCC از هریک از فریم ها ۱۳ ویژگی استخراج می گردد و در مرحله بعد انرژی ضرایب MFCC مربوط به هر فریم محاسبه می‌شود. برای محاسبه انرژی ضرایب تک تک ضرایب را به توان دو رسانده باهم جمع می‌شوند. سپس برای کم‌کردن تغییرات خروجی مقدار DC محاسبه می‌شود و از خروجی سیگنال حذف می‌شود. برای به دست آوردن مقدار DC جمع ضرایب ویژگی‌های انرژی تقسیم بر تعداد ضرایب می‌شود. سپس به منظور نرم‌تر کردن تغییرات منحنی مربوطه، مقادیر به دست آمده توسط فیلتر باتروث فیلتر می گردد و در نهایت با توجه به میزان آستانه تعریف شده، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند.



شکل ۴-۱۴: تشخیص مرز کلمه با استفاده از روش MFCC برای داده اول



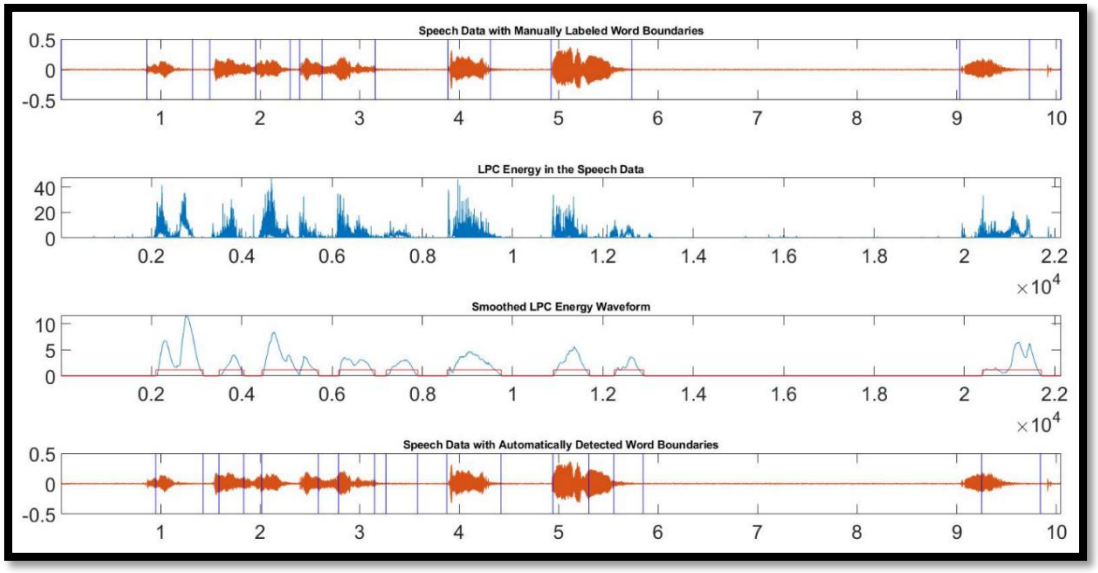
شکل ۴-۱۵: تشخیص مرز کلمه با استفاده از روش MFCC برای داده دوم

همانطورکه در شکل (۴-۱۴) و (۴-۱۵) با مقایسه نمودار مربوط به نتایج جداسازی کلمات به صورت دستی و استفاده از الگوریتم استخراج ویژگی MFCC مشاهده می گردد، الگوریتم مبتنی بر انرژی ضرایب فرکانس مل توانسته است بیشتر مرزهای کلمات را پیدا کند اما برخی از مرزها را پیدا نکرده است و در آن قسمت ها چند کلمه با هم در یک محدوده قرار گرفته اند. اما برای داده اول بسیار بهتر توانسته است مرزهای کلمات را تشخیص دهد.

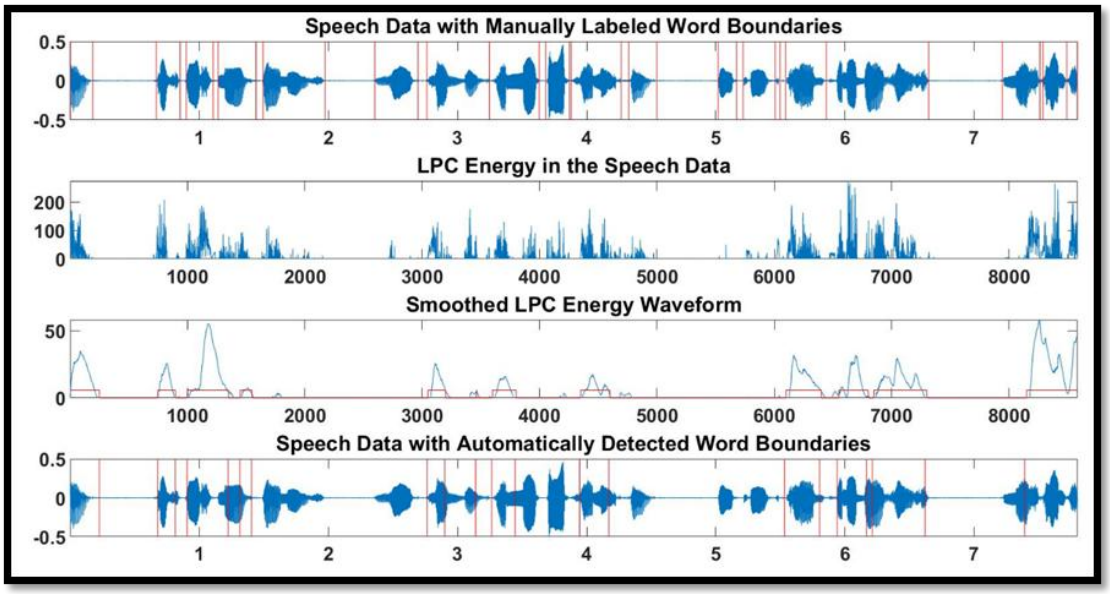
۴-۵-۲- نتیجه شبیه‌سازی روش LPC

الگوریتم جداسازی مرز کلمات با استفاده از روش استخراج ویژگی LPC بدین صورت است که ابتدا به منظور بررسی زمان کوتاه سیگنال-گفتار، سیگنال ورودی با ضرب در پنجره به طول ۳۲۰ نمونه و با شیفت زمانی ۲۰ نمونه، فریم بندی می گردد. سپس توسط روش LPC از هریک از فریم ها ۱۳ ویژگی استخراج می گردد و در مرحله بعد انرژی ضرایب LPC مربوط به هر فریم محاسبه می شود. برای محاسبه انرژی، ضرایب LPC به توان دو می‌رسند و سپس باهم جمع می‌شوند. سپس مقدار DC خروجی این مرحله حذف می شود. برای محاسبه مقدار میانگین ضرایب انرژی باهم جمع می‌شوند و تقسیم به تعداد آن‌ها می‌شوند. به منظور نرم‌تر کردن تغییرات منحنی مربوطه ، مقادیر بدست آمده توسط فیلتر

باترورث فیلتر می گردد و در نهایت با توجه به میزان آستانه‌ای که به صورت تجربی به دست آمده است، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند.



شکل ۴-۱۶: تشخیص مرز کلمه با استفاده از روش LPC برای داده اول



شکل ۴-۱۷: تشخیص مرز کلمه با استفاده از روش LPC برای داده دوم

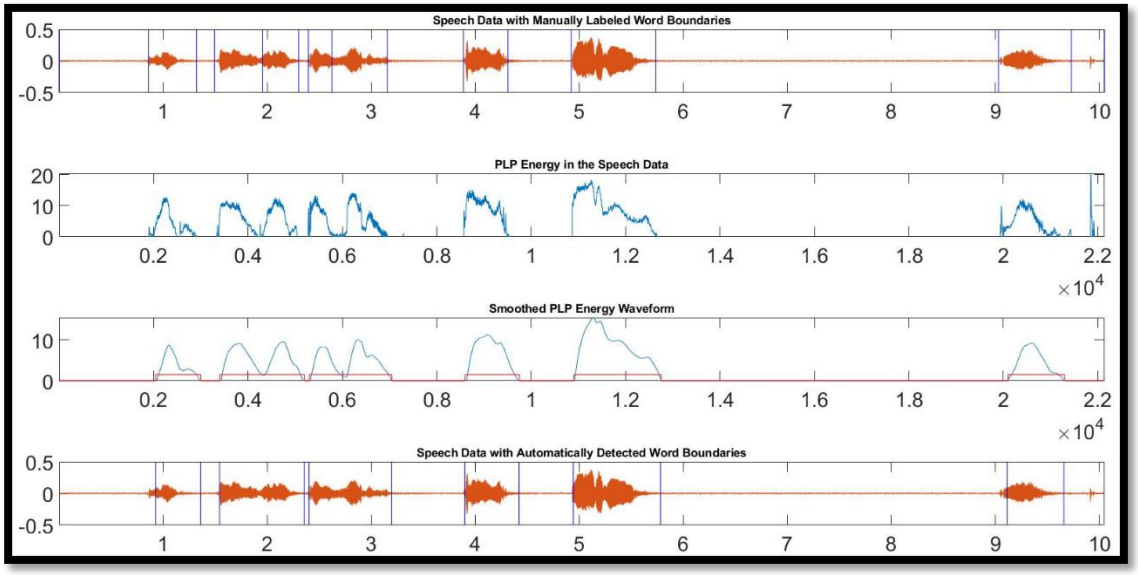
همانطورکه در شکل (۴-۱۶) و (۴-۱۷) با مقایسه نمودار مربوط به نتایج جداسازی کلمات بصورت دستی و استفاده از الگوریتم استخراج ویژگی LPC مشاهده می گردد، الگوریتم مبتنی بر انرژی ضرایب

پیشگویی خطی برخی از مرزها را درست و برخی دیگر را به اشتباه شناسایی کرده است. با مقایسه خروجی های LPC و MFCC می توان اینطور بیان نمود که الگوریتم مبتنی بر MFCC توانایی بیشتری در تخمین مرزهای کلمات در قیاس با روش مبتنی بر LPC دارد.

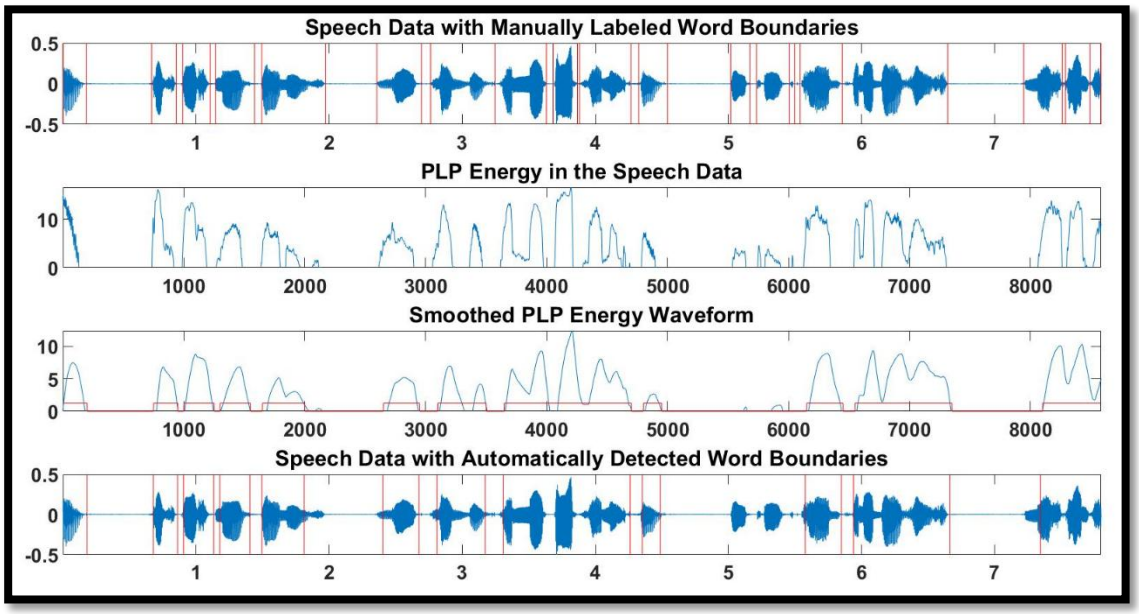
۴-۵-۳- نتیجه شبیه‌سازی روش PLP

الگوریتم جداسازی مرز کلمات با استفاده از روش استخراج ویژگی PLP بدین صورت است که ابتدا به منظور بررسی زمان-کوتاه سیگنال گفتار، سیگنال ورودی با ضرب در پنجره به طول ۳۲۰ نمونه و با شیفت زمانی ۲۰ نمونه، فریم بندی می گردد. سپس توسط روش PLP از هریک از فریم ها ۱۳ ویژگی استخراج می گردد و در مرحله بعد انرژی ضرایب PLP مربوط به هر فریم محاسبه می شود. برای محاسبه انرژی هر فریم ویژگی‌های PLP به دست آمده به توان دو می‌رسند و با هم جمع می‌شوند. سپس مقدار DC خروجی این مرحله حذف می شود. مانند قبل برای به دست آوردن مقدارDC مقدار میانگین ضرایب خروجی محاسبه می‌شود. به منظور نرم تر کردن تغییرات منحنی مربوطه، مقادیر بدست آمده توسط فیلتر باتروث فیلتر می گردد و در نهایت با توجه به میزان آستانه تعریف شده، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند.

همانطورکه در شکل (۴-۱۸) و (۴-۱۹) با مقایسه نمودار مربوط به نتایج جداسازی کلمات بصورت دستی و استفاده از الگوریتم استخراج ویژگی PLP مشاهده می گردد، الگوریتم مبتنی بر انرژی PLP برخی از مرزها را درست و برخی دیگر را به اشتباه شناسایی کرده است. با مقایسه خروجی‌های مربوط به PLP و LPC می توان اینطور بیان نمود که الگوریتم مبتنی بر PLP توانایی بیشتری در تخمین مرزهای کلمات در قیاس با روش مبتنی بر LPC برای هر دو داده دارد.



شکل ۴-۱۸: تشخیص مرز کلمه با استفاده از روش PLP برای داده اول



شکل ۴-۱۹: تشخیص مرز کلمه با استفاده از روش PLP برای داده دوم

۴-۵-۴- تئوری کار روش موجک

همانطور که گفته شد سیگنال‌ها تا این قسمت یا در حوزه زمان یا در حوزه فرکانس بودند. برای تحلیل بهتر سیگنال‌ها، در حوزه زمان-فرکانس بررسی می‌شوند. روش موجک و روش ویگنر از روش‌های زمان-فرکانس هستند که نتایج آن‌ها در این قسمت و قسمت بعد آورده شده است.

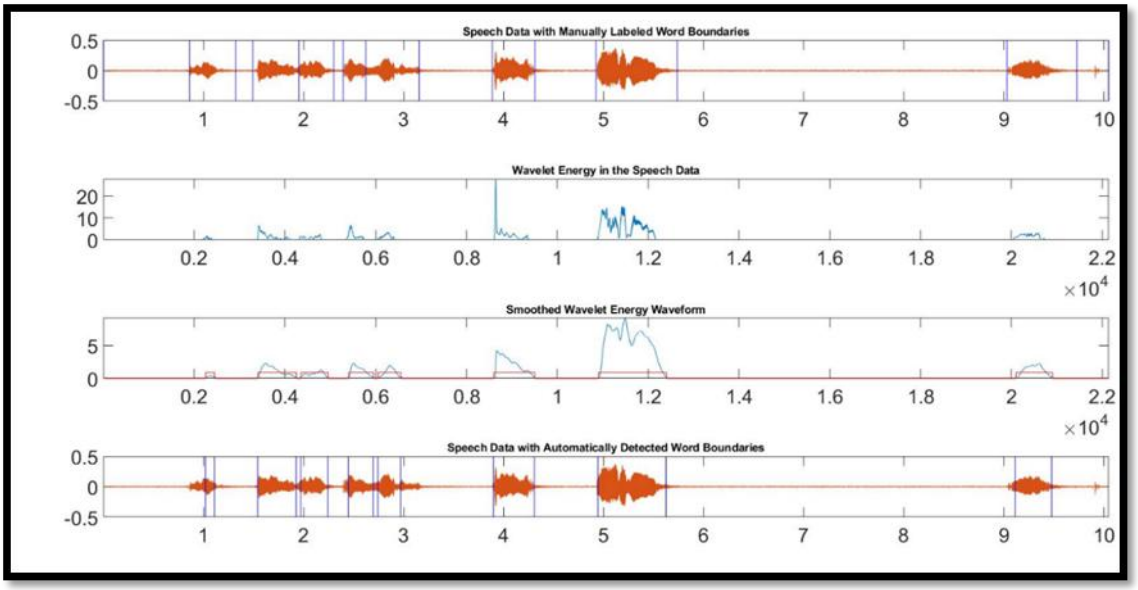
مراحل تشخیص مرز کلمات با استفاده از روش موجک به شرح زیر است:

- ابتدا با استفاده از تبدیل موجک گسسته از کل سیگنال ورودی ویژگی استخراج می‌شود. تابع موجک مورد استفاده sym4 است.
- پس از استخراج ویژگی به این دلیل که فقط از مؤلفه‌های فرکانس پایین یعنی مؤلفه تقریب استفاده می‌شود طول سیگنال نصف شده و نمی‌تواند با روش دستی مقایسه شود. به همین دلیل درون‌یابی صورت می‌گیرد تا طول این سیگنال با سیگنال اصلی برابر شود. برای درون-یابی دو نمونه قبل و بعد را در نظر گرفته و تقسیم به دو می‌شود.
- در مرحله بعد برای فریم‌بندی، پنجره‌هایی با طول ۳۲۰ نمونه در سیگنال ضرب می‌شود. انرژی هر پنجره که انرژی ویژگی‌های موجک است، محاسبه می‌شود. می‌توان به جای انرژی از معیارهای دیگری مانند آنتروپی نیز استفاده کرد.
- اگر خروجی تغییرات زیادی داشته باشد برای کم کردن تغییرات و پردازش سریع‌تر مقدار DC خروجی را حذف می‌شود. برای این کار میانگین انرژی ویژگی‌ها را محاسبه کرده و از خروجی ویژگی‌ها کم می‌شود تا خروجی تغییرات کمتری داشته باشد.
- سپس برای نرم‌تر کردن تغییرات منحنی و کاهش اثر نویز توسط فیلتر باترورث، فیلتر می‌گردد.
- با استفاده از روش سعی و خطا و جای‌گذاری مقادیر مختلف برای آستانه، بهترین مقدار آستانه برای تشخیص مرز کلمات استفاده می‌شود.
- در مرحله آخر مرزهای کلمات به دست آمده با استفاده از روش موجک با مرزهایی که به صورت دستی به دست آمده‌اند، مقایسه می‌شوند و مقدار تشخیص صحیح مشخص می‌شود.

۴-۵-۴-۱- نتایج شبیه‌سازی روش موجک

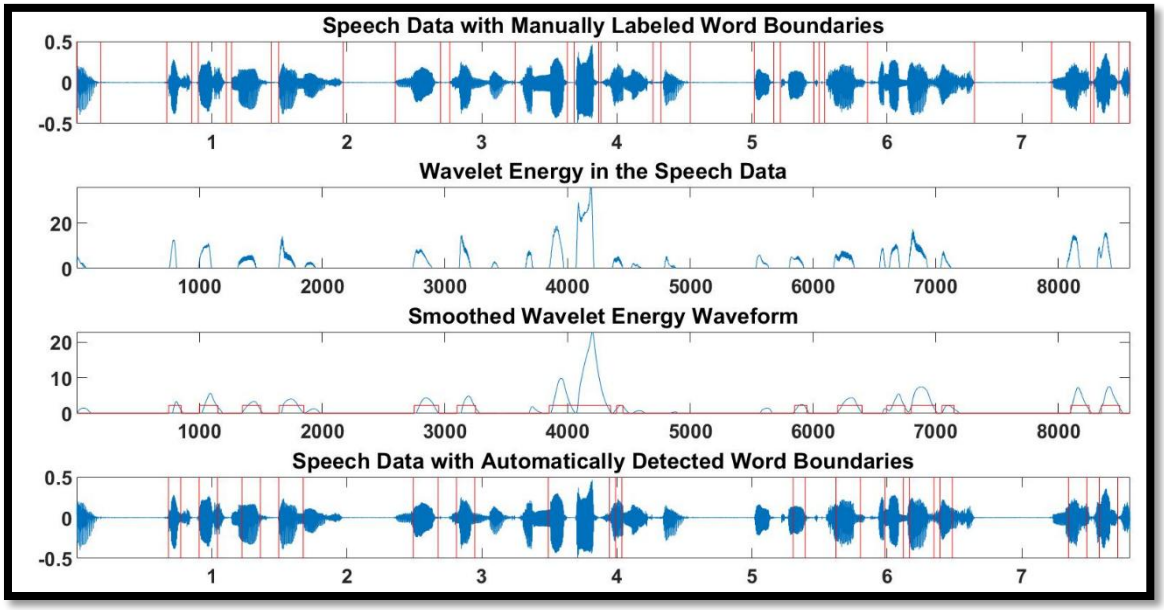
در قسمت قبل الگوریتم جداسازی مرز کلمات با استفاده از روش موجک توضیح داده شد. در این قسمت نتایج شبیه‌سازی با استفاده از تبدیل موجک برای تشخیص مرز کلمات نمایش داده می‌شود. در قسمت اول شکل (۴-۲۰) مرزهای کلمات به صورت دستی مشخص شده‌اند. سپس استخراج ویژگی

انجام شده و درون‌یابی صورت می‌گیرد. در مرحله بعد با فریم بندی خروجی مرحله قبل، انرژی زمان - کوتاه برای هر فریم محاسبه شده و سپس مقدار DC خروجی این مرحله حذف می شود و سیگنال توسط فیلتر باترورث، فیلتر می‌گردد. با توجه به میزان آستانه تعریف شده برای داده اول، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند. در قسمت سوم شکل‌ها بخش‌هایی که بیشتر از مقدار آستانه است با خطوط قرمزی که ۰/۱ ماکزیمم ارتفاع سیگنال است نشان داده شده است. این خطوط همان مرزهای کلمات است که در قسمت چهارم بر روی سیگنال اصلی نشان داده شده است. سپس دو قسمت اول و آخر مقایسه می‌شود و تشخیص‌های صحیح مشخص می‌گردد.



شکل ۴-۲۰: تشخیص مرز کلمات با استفاده از روش موجک برای داده اول

همانطورکه در شکل (۴-۲۰) و (۴-۲۱) با مقایسه نمودار مربوط به نتایج جداسازی کلمات بصورت دستی و استفاده از الگوریتم مبتنی بر موجک مشاهده می گردد، الگوریتم مبتنی بر انرژی موجک برخی برای داده‌های اول بسیار بهتر از داده‌های دوم مرزها را مشخص کرده است.



شکل ۴-۲۱: تشخیص مرز کلمات با استفاده از روش موجک برای داده دوم

۴-۵-۵- تئوری کار روش ویگنر-ویل

مراحل تشخیص مرز کلمات با استفاده از روش ویگنر-ویل به شرح زیر است:

- ابتدا سیگنال ورودی با ضرب در پنجره‌ای به طول ۳۲۰ نمونه و شیفت زمانی ۲۰ نمونه فریم‌بندی و پنجره‌گذاری می‌شود.
- سپس تبدیل ویگنر-ویل همه‌ی فریم‌ها محاسبه می‌شود. برای هر فریم تمام توزیع‌ها باهم جمع می‌شود. یعنی برای هر فریم یک بردار وجود دارد.
- در مرحله بعد انرژی ضرایب ویگنر-ویل محاسبه می‌شود. برای محاسبه انرژی، هر ضریب انرژی به توان دو رسیده سپس تمام ضرایب جمع می‌شوند.
- برای کم کردن تغییرات خروجی، میانگین انرژی ویگنر-ویل محاسبه شده و از مقدار خروجی کم می‌شود. برای محاسبه مقدار میانگین تمام ضرایب انرژی با هم جمع می‌شوند و تقسیم بر تعداد ضرایب می‌شود.
- مشکلی که روش ویگنر-ویل دارد این است که اختلاف بین اعداد در سیگنال زیاد است. برای مثال یکی صفر و دیگری ده‌هزار است. در صورتی که تغییرات داده زیاد باشد، برای

دیده شدن تمام اعداد باید مقیاس لگاریتمی شود. بعد از استخراج لگاریتم چون اعداد مثبت و منفی می‌شوند، برای جلوگیری از این مشکل قدر مطلق اعداد نیز محاسبه می‌شود.

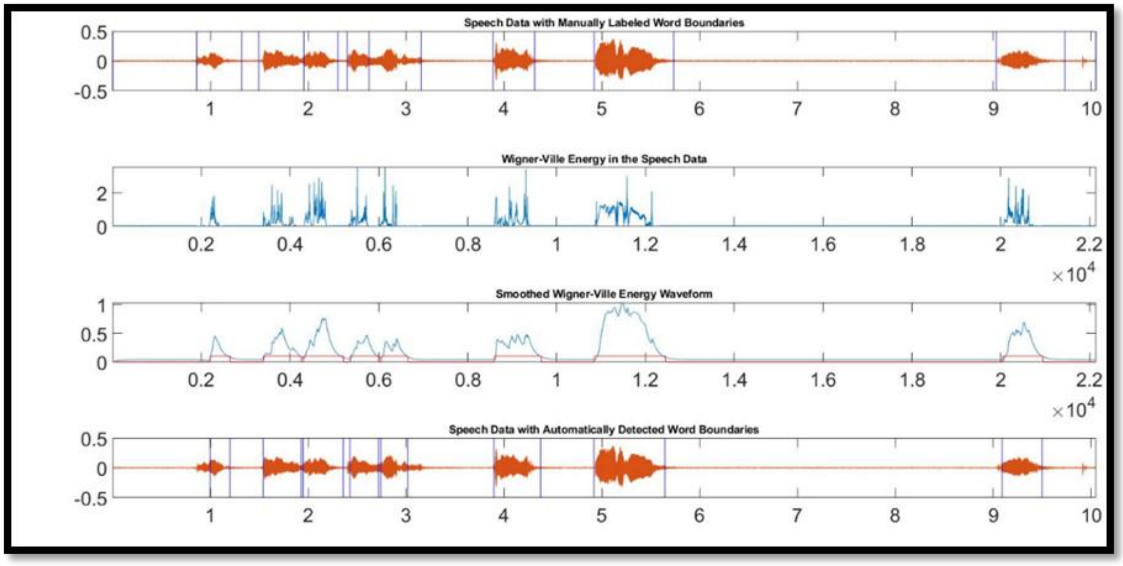
- در مرحله بعد میانگین را از خروجی داده‌های لگاریتم گرفته شده کم می‌شود. به دلیل این‌که ممکن است بعضی از داده‌های لگاریتم گرفته شده از مقدار میانگین کمتر شود بار دیگر قدر مطلق گرفته می‌شود تا نتیجه همیشه مثبت باشد.
- در مرحله بعد برای تشخیص بهتر مرزهای کلمه هموارسازی انجام می‌شود.
- با توجه به مقدار آستانه محدوده‌ای از سیگنال که بیشتر از مقدار آستانه باشد به عنوان مرز کلمه در نظر گرفته می‌شود.

۴-۵-۵-۱- نتیجه شبیه‌سازی روش ویگنر- ویل

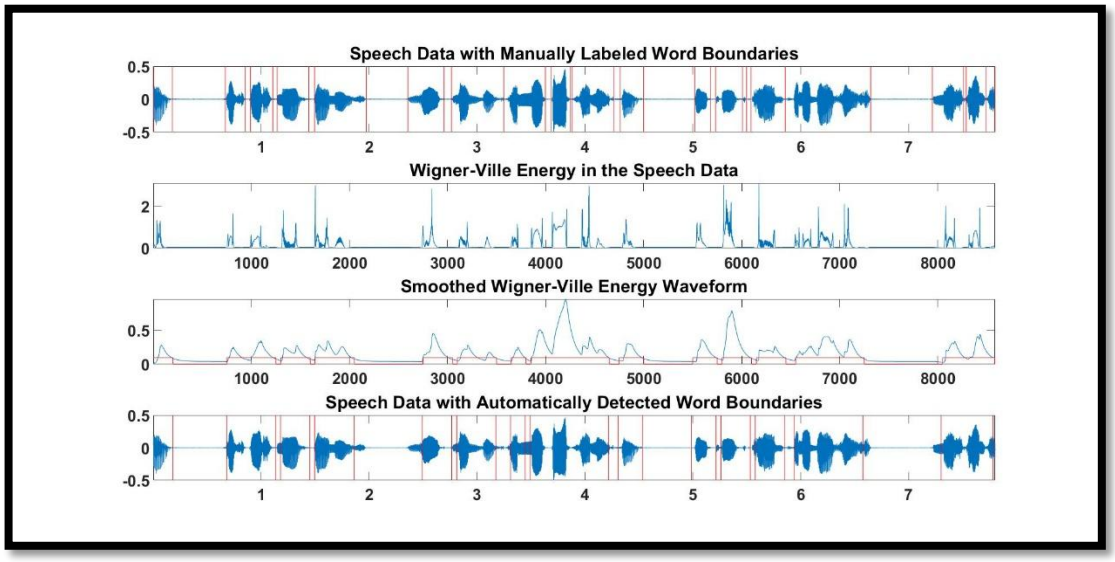
در قسمت قبل مراحل تشخیص مرز کلمات با استفاده از روش ویگنر-ویل توضیح داده شد. در این قسمت نتایج شبیه‌سازی با استفاده از توزیع ویگنر-ویل برای تشخیص مرز کلمات نمایش داده می-شود. در قسمت اول شکل (۴-۲۲) مرزهای کلمات به صورت دستی مشخص شده‌اند. ابتدا سیگنال فریم‌بندی شده و سپس با استفاده از روش ویگنر-ویل استخراج ویژگی انجام می‌شود. مقدار DC خروجی این مرحله حذف می شود و سیگنال توسط فیلتر باترورث، فیلتر می‌گردد. با توجه به میزان آستانه تعریف شده، محدوده هایی از منحنی که دارای مقدار بزرگتری از سطح آستانه است به عنوان مرز کلمات انتخاب می گردند. در قسمت سوم شکل (۴-۲۲) بخش‌هایی که بیشتر از مقدار آستانه است با خطوط قرمزی که ۰/۱ ماکزیمم ارتفاع سیگنال است نشان داده شده است. این خطوط همان مرزهای کلمات است که در قسمت چهارم بر روی سیگنال اصلی نشان داده شده است. سپس دو قسمت اول و آخر مقایسه می‌شود و تشخیص‌های صحیح مشخص می‌گردد.

همانطورکه در شکل (۴-۲۲) و (۴-۲۳) با مقایسه نمودار مربوط به نتایج جداسازی کلمات بصورت دستی و استفاده از الگوریتم مبتنی بر لگاریتم انرژی ضرایب ویگنر-ویل مشاهده می گرددکه الگوریتم پیشنهادی در قیاس با سایر روش ها برای هر دو داده توانسته است تعداد بیشتری مرز را به درستی

تخمین بزند. با توجه به نتایج بدست آمده می توان اینطور بیان نمود که الگوریتم پیشنهادی مناسبترین و قوی ترین روش از بین روش های بررسی شده برای شناسایی مرز کلمات است.



شکل ۴-۲۲: تشخیص مرز کلمات با استفاده از روش ویگنر-ویل برای داده اول

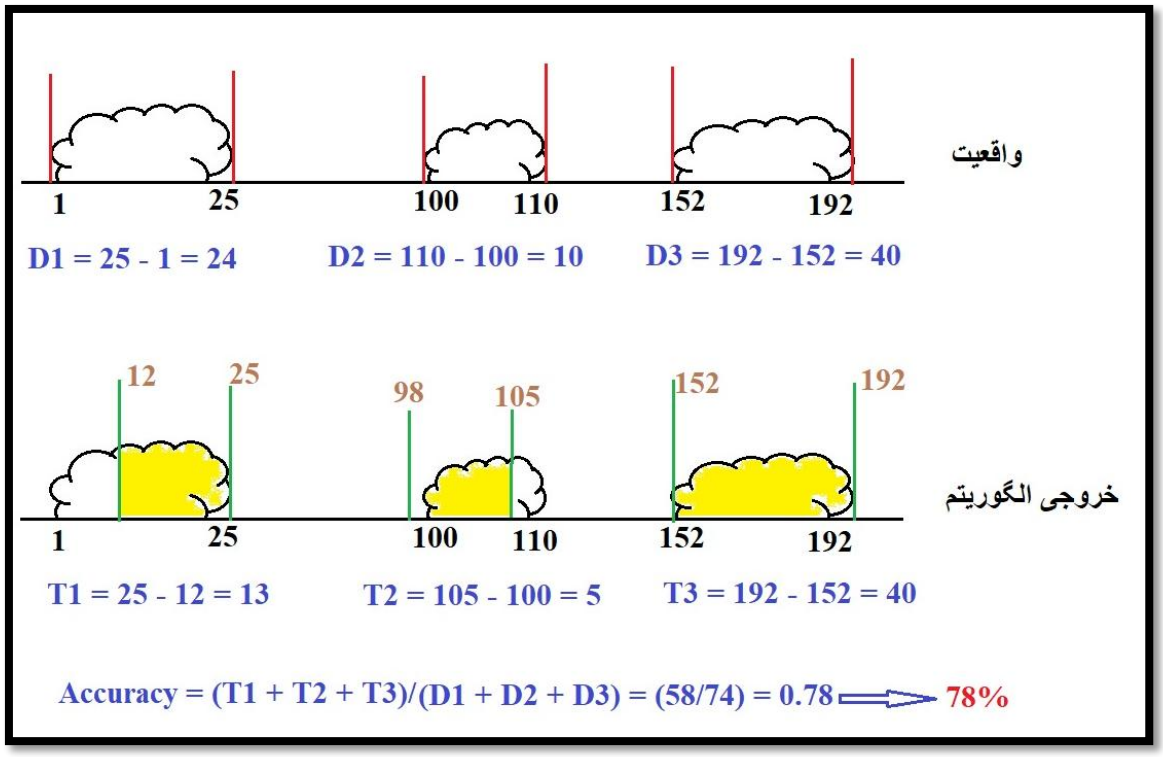


شکل ۴-۲۳: تشخیص مرز کلمات با استفاده از روش ویگنر-ویل برای داده دوم

۴-۶- نتیجه‌گیری

در بخش‌های قبل الگوریتم تشخیص مرز کلمات توضیح داده شده و نتایج شبیه‌سازی روش‌های مختلف نشان داده شده است. برای به دست آوردن دقت تشخیص صحیح در هریک از روش‌ها مانند

شکل (۴-۲۵) عمل می‌کنیم. این روش، شهودی است و دارای دقت خوبی می‌باشد. مرزهایی که با استفاده از اطلاعات موجود در پایگاه داده مشخص می‌شوند در قسمت واقعیت نشان داده شده است. با استفاده از روش های مختلف مرزهای کلمه به دست می‌آیند و در قسمت خروجی الگوریتم نشان داده می‌شود. برای مثال در شکل (۴-۲۴) ۳ مرز مشخص شده است. مرز قسمت اول از نمونه‌ی ۱ تا نمونه ۲۵ است. اما خروجی الگوریتم ابتدای مرز را نمونه ۱۲ و انتهای آن را نمونه ۲۵ تشخیص داده است. یعنی الگوریتم تا حدودی مرز را درست در تشخیص داده است. با توجه به توضیحات داخل شکل دقت تشخیص، مشخص می‌شود. پس در این روش مرز تشخیص داده کلا اشتباه در نظر گرفته نمی‌شود و درصدی از آن صحیح است. در صورتی که بازه‌ای تشخیص داده شود که در محدود صحیح نباشد در محاسبات در نظر گرفته نمی‌شود.



شکل ۴-۲۴: روش به دست آوردن دقت تشخیص مرز کلمات

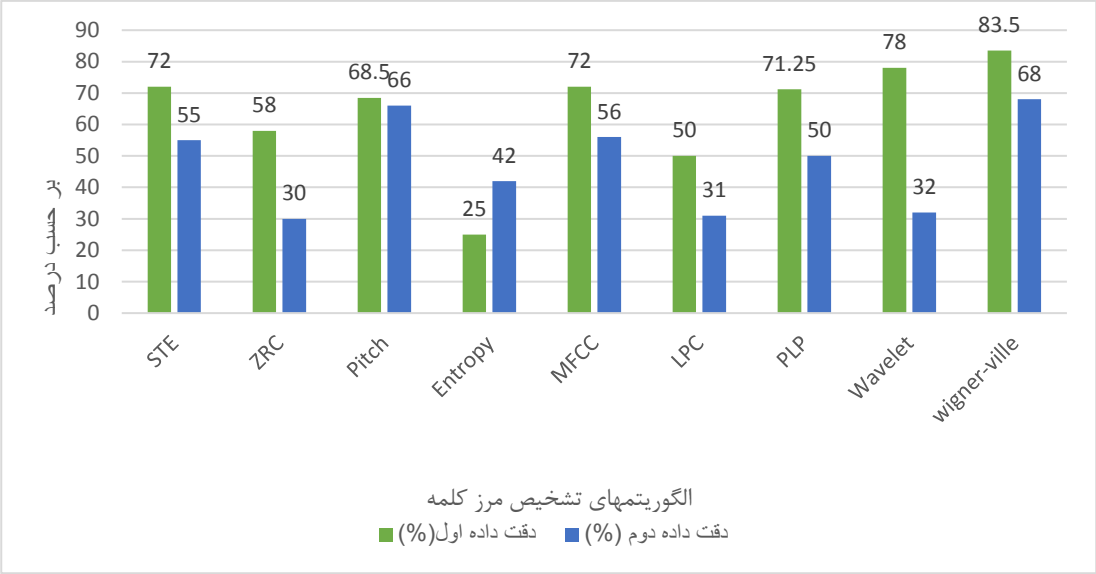
بعد از به دست آوردن دقت تشخیص هر روش مشاهده می‌شود که روش ویگنر-ویل که روش استخراج ویژگی از نوع زمان-فرکانس است بهترین روشی است که برای تشخیص مرز کلمات در سیگنال‌های گفتاری استفاده می‌شود. پس تحلیل سیگنال در هر دو حوزه زمان و فرکانس باعث

تشخیص بهتر مرز کلمات می‌شود. همچنین روش موجک که روش دیگری از روش‌های زمان-فرکانس است دقت بالایی در تشخیص مرز کلمات در داده‌های با پیچیدگی کمتر دارد. در نتیجه سیگنال گفتار که یک سیگنال ناایستا است و مشخصات آماری آن با زمان تغییر می‌کند، تحلیل این سیگنال در هر دو حوزه زمان و فرکانس باعث تشخیص بهتر مرز کلمات می‌شود. آنالیزهای زمانی در تشخیص مرزهای کلمات خیلی قوی نبودند. برای مثال در روش آنتروپی از سیگنال خام استفاده شده است که نتیجه مطلوبی نداشته است. اگر ابتدا از سیگنال تبدیل فوریه گرفته شود و سپس با استفاده از روش آنتروپی مرز کلمات مشخص شود، دقت نتیجه بهتر می‌شد. در جدول (۴-۴) دقت تشخیص هر یک از روش‌ها نشان داده شده است.

جدول ۴-۳: نتایج روش‌های تشخیص مرز کلمه

روش	STE	ZRC	Pitch	Entropy	MFCC	LPC	PLP	Wavelet	Wigner-ville
دقت برای داده اول(٪)	۷۲	۵۸	۶۸/۵	۲۵	۷۲	۵۰	۷۱/۲۵	۷۸	۸۳/۵
دقت برای داده دوم (٪)	۵۵	۳۰	۵۹	۴۲	۵۶	۳۱	۳۲	۴۲	۶۸

طبق نتایج ارائه شده، در داده‌هایی با پیچیدگی کمتر، الگوریتم‌ها عملکرد تقریباً مشابهی دارند. در واقع میزان مکث‌های بین کلمات و نحوه تلفظ گوینده در دقت تشخیص مرزها مؤثر است. با این حال، هنگامی که از داده‌هایی با پیچیدگی بیشتر استفاده می‌شود، شاهد افت کارایی نسبتا زیاد روش‌های قبلی در مقایسه با روش ویگنر-ویل هستیم.



شکل ۴-۲۵: نمودارهای دقت برای داده اول و داده دوم

فصل پنجم:

نتیجه‌گیری و پیشنهادات

۵-۱-جمع‌بندی

در این تحقیق، هدف تشخیص مرزهای کلمه با استفاده از روش‌های مختلف استخراج ویژگی است. تشخیص مرز کلمه یعنی مشخص کردن ابتدا و انتها یک کلمه در یک سیگنال گفتار. تشخیص مرزهای کلمات در بسیاری از کاربرها مانند بازشناسی گفتار، سنتز گفتار، تبدیل متن به گفتار و... کاربرد دارد. از دونوع داده در این پایان‌نامه استفاده شده است. پایگاه داده TIMIT و داده‌ای که کلمات را به طور واضح بیان می‌کند. الگوریتم‌های استخراج ویژگی به سه دسته آنالیز زمانی، آنالیز طیفی و آنالیز زمان-فرکانس تقسیم می‌شوند. در فصل دوم روش‌های مرسومی که برای تشخیص مرز کلمه استفاده شده است و شامل انرژی، نرخ عبور از صفر، گام و آنتروپی، بررسی شدند. ابتدا مراحل پیش پردازش سیگنال گفتار به دلیل تغییر مشخصه‌های سیگنال گفتار در طول زمان و به عبارتی ناایستا بودن آن صورت می‌گیرد. در واقع استخراج ویژگی از یک حوزه زمانی نسبتا بزرگ، اطلاعات معتبری را در اختیار نمی-گذارد. به همین منظور سیگنال گفتار فریم‌بندی و پنجره‌گذاری می‌شود. پنجره مورد نظر همینگ و طول پنجره ۳۲۰ نمونه و مقدار شیفت پنجره ۲۰ نمونه می‌باشد. پنجره بسیار مهم است زیرا که حاشیه-های یک فریم در کم یا زیاد شدن خطا مؤثرند. با استفاده از الگوریتم مورد نظر ویژگی‌ها استخراج می‌شوند. سپس توسط فیلتر باتروورث که یک فیلترپایین گذر و درجه یک است به منظور کم کردن نویز و افزایش کارایی سیستم، هموارسازی سیگنال انجام می‌شود. سپس برای تشخیص مرزها آستانه‌ای در نظر گرفته می‌شود. روش به دست آوردن آستانه به صورت تجربی و سعی و خطا است. اگر مقدار آستانه بالا در نظر گرفته شود باعث از دست رفتن تعداد زیادی از مرزهای کلمات می‌شود. پس چند مقدار آستانه آزمایش می‌شود و بهترین آن که تعداد بیشتر از مرزها را تشخیص انتخاب می‌شود. در روش مورد نظر، اگر ویژگی استخراج شده مقدارش از آستانه بیشتر شود به عنوان مرز کلمات در نظر گرفته می‌شود. سپس مرزهای تشخیص داده شده با الگوریتم مورد نظر با مرزهایی که به صورت دستی و از اطلاعات موجود در پایگاه داده مشخص شده‌اند، مقایسه می‌شود و میزان تشخیص صحیح مرزهای

کلمات مشخص می‌شود. در این روش‌ها به دلیل وجود نویز در محیط، استفاده از برخی از این معیارها چندان مناسب نمی‌باشد. همچنین عدم وجود مکث کافی باعث خطا در شناسایی مرزها می‌شود. گوینده باید کلمات را به وضوح تلفظ کند.

در فصل سوم از روش‌های طیفی و روش‌های زمان - فرکانس برای شناسایی مرزهای کلمات استفاده شده است. MFCC، LPC، PLP، موجک و ویگنر-ویل از جمله‌ی این روش‌هاست. در سه روش اول پس از فریم بندی و پنجره‌گذاری از هر فریم ۱۳ ویژگی استخراج می‌شود و میزان انرژی ضرایب هر الگوریتم محاسبه می‌شود. در الگوریتم موجک با استفاده از تابع موجک مؤلفه تقریب به دست می‌آید و سپس به‌منظور همسان‌سازی طول مؤلفه تقریب با سیگنال اصلی درون یابی صورت می‌گیرد و انرژی ضرایب موجک به دست می‌آید. در ویگنر-ویل نیز ضرایب ویگنر-ویل از هر فریم استخراج می‌شود و ضرایب در هر فریم با هم جمع می‌شوند و مقدار انرژی ضرایب محاسبه می‌شود. پس از محاسبه ضرایب انرژی و حذف مقدار DC سیگنال، مقدار آستانه در هرالگوریتم اعمال می‌شود. در این الگوریتم‌ها نیز مقدار آستانه با سعی و خطا و روش تجربی به دست آمده که ضریبی از ماکزیمم ارتفاع سیگنال هموارشده می‌باشد. مرزها مشخص می‌شود و با مقدار مرزهایی که به روش دستی مشخص شده‌اند مقایسه می-گردند. طبق نتایج به دست آمده روش ویگنر- ویل بیشترین مقدار تشخیص صحیح را برای هر دو داده دارد. در نهایت نتایج شبیه‌سازی در فصل چهارم نشان داده شده است.

۵-۲-پیشنهادات برای ادامه کار

در این قسمت پیشنهاداتی برای تشخیص مرز کلمه و بهبود نتایج داده شده است:

- یکی از چالش‌های اصلی تشخیص مرزکلمات، انتخاب مقدار آستانه می‌باشد. اگر بتوان برای تمام روش‌ها آستانه را با روش مشخصی و به صورت یکسان محاسبه کرد، بهتر می‌توان الگوریتم‌ها را با یک‌دیگر مقایسه کرد. حتی می‌توان دو مقدار برای آستانه در نظر گرفت. یکی برای شناسایی ابتدای کلمات و دیگر برای شناسایی انتهای کلمات.

- همچنین برای هموار سازی سیگنال‌های خروجی می‌توان از فیلترهای دیگر با درجه‌های متنوع استفاده کرد تا تأثیر آن‌ها در نتیجه نهایی مشخص شود.
- می‌توان از الگوریتم های زمان-فرکانس دیگر همچون روش تبدیل فوریه کسری و یا الگوریتم‌های یادگیری همچون شبکه‌های عصبی، فازی و احتمالاتی آزمایش کرد.
- همانطور که در فصل دو مشاهده شده ترکیب روش ZRC با سایر روش‌ها باعث بهبود نتایج شده بود. می‌توان الگوریتم‌های مختلف مانند MFCC، ویگنر-ویل و آنتروپی را باهم ترکیب کرد تا باعث بهبود دقت شود.

منابع

[1] A. Caranica, H. Cucu, A. Buzo and C. Burileanu, "On the design of an automatic speech recognition system for Romanian language," *Journal of Control Engineering and Applied Informatics* , vol. 18, no. 2, pp. 65-76, 2016.

[۲] آیت، مبانی پردازش سیگنال گفتار، انتشارات دانشگاه پیام نور، ۱۳۸۷.

[3] L. Lamel, L. Rabiner and A. Rosenberg, "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 777-785, 1981.

[4] J.-C. Junqua, B. Mak and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise," *IEEE Transactions on speech and audio processing*, vol. 2, no. 3, pp. 406-412, 1994.

[5] A. D. Vijayendra and V. K. Thakar, "Word boundary detection for Gujarati speech recognition using in-ear microphone," in *2016 1st India International Conference on Information Processing (IICIP)*, Delhi, India, 2016.

[6] A. Agarwal, A. Jain and N. Prakash, "Word Boundary Detection in Continuous Speech based on Suprasegmental Features for Hindi Language," in *2010 2nd International Conference on Signal Processing Systems*, Dalian, China, 2010.

[7] Z. Hu, J. Schalkwyk, E. Barnard and R. Cole, "Speech recognition using syllable-like units," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Philadelphia, PA, USA, 1996.

[8] D. Enqing, L. Guizhong, Z. Yatong and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," in *6th International Conference on Signal Processing, 2002.*, Beijing, China, 2002.

[9] D. Arifianto, "Dual parameters for voiced-unvoiced speech signal determination," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Honolulu, HI, USA, 2007.

[10] C. Guanghua, L. Junhai and Y. Jun, "An Improved Method of Endpoints Detection Based on Energy-Frequency-Value," in *Conference on High Density Microsystem Design and Packaging and Component Failure Analysis, 2006. HDP'06*, Shanghai, China, 2006.

[11] M. Jalil, F. A. Butt and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *2013 The International Conference on Technological*

Advances in Electrical, Electronics and Computer Engineering (TAEECE), Konya, Turkey, 2013.

[12] S. Kumar, S. Phadikar and K. Majumder, "Modified segmentation algorithm based on short term energy & Zero crossing rate for maithili speech signal," in *2016 International Conference on Accessibility to Digital World (ICADW)*, Guwahati, India, 2016.

[13] T. H. Zaw and N. War, "The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2017.

[14] L. Hui, B. Dai and L. Wei, "A pitch detection algorithm based on amdf and acf," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, France, 2006.

[15] G. R. Rao and J. Srichland, "Word boundary detection using pitch variations," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Philadelphia, PA, USA, 1996.

[16] J. Shen, J. Hung and L. Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments," *ICSLP*, vol. 98, pp. 232-235, 1998.

[17] L. Jin and J. Cheng, "An improved speech endpoint detection based on spectral subtraction and adaptive sub-band spectral entropy," in *2010 International Conference on Intelligent Computation Technology and Automation*, Changsha, China, 2010 .

[18] B.-F. Wu and K.-C. Wang, "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 762-775, 2005.

[19] Y. Zhang, K. Wang and B. Yan, "Speech endpoint detection algorithm with low signal-to-noise based on improved conventional spectral entropy," in *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, Guilin, China, 2016.

[20] P.-S. Antonia, Applications in Time-Frequency Signal Processing, CRC press, 2018.

[21] R. D. Patterson, "Auditory filter shape," *The Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 802-809, 1974.

[22] M.-A. Nasr, M. Abd-Elnaby, A.-S. El-Fishawy, S. El-Rabaie and F.-E. A. El-Samie, "Speaker identification based on normalized pitch frequency and Mel Frequency Cepstral Coefficients," *International Journal of Speech Technology*, vol. 21, pp. 941-951, 2018.

[23] I. Daly, Z. Hajaiej and A. Gharsallah, "Speech analysis in search of speakers with MFCC , PLP , Jitter and Shimmer," in *2017 International Conference on Advanced Systems and Electric Technologies (IC_ASET)*, Hammamet, Tunisia, 2017.

[24] H. Wang, Y. Xu and M. Li, "Study on the MFCC similarity-based voice activity detection algorithm," in *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, Deng Feng, China, 2011.

[25] G.-D. Wu and C.-T. Lin, "Word boundary detection with mel-scale frequency bank in noisy environment," *IEEE Transactions on Speech and Audio Processing,* vol. 8, no. 5, pp. 541-554, 2000.

[26] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America,* vol. 55, no. 6, pp. 1304-1312, 1974.

[27] L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing, Now Publishers, 2007.

[28] E. Zwicke and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America,* vol. 68, no. 5, pp. 1523-1525, 1998.

[29] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America,* vol. 33, no. 2, pp. 248-248, 1960.

[30] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America,* vol. 874, no. 4, pp. 1738-1752, 1989.

[31] A. Kalia, S. Sharma and S. Pandey, "Comparative analysis of speaker recognition system based on voice activity detection technique, MFCC and PLP Features," in *Intelligent Computing Techniques for Smart Energy Systems*, Singapore, 2020.

[32] "Wavelets in speech segmentation," in *MELECON 2008 - The 14th IEEE Mediterranean Electrotechnical Conference*, Ajaccio, France, 2008.

[33] T. Ro, T. Marwala and S. Chakraverty, "Precise detection of speech endpoints dynamically: A wavelet convolution based approach," *Communications in Nonlinear Science and Numerical Simulation,* vol. 67, pp. 162-175, 2019.

[34] C. Yali, L. Dongsheng, J. Shuo and N. Xuefen, "A speech endpoint detection algorithm based on wavelet transforms," in *The 26th Chinese Control and Decision Conference (2014 CCDC)*, Changsha, China, 2014.

[35] M. Eshaghi and M. .. Mollaei, "Voice activity detection based on using wavelet packet," *Digital Signal Processing,* vol. 20, no. 4, pp. 1102-1115, 2010.

[36] E. P. Wigner, "On the Quantum Correction for Thermodynamic Equilibrium," in *Part I: Physical Chemistry. Part II: Solid State Physics*, Springer, 1197, pp. 110-120.

[37] J. Ville, "Theorie et application dela notion de signal analytique," *Câbles et transmissions,* vol. 2, no. 1, pp. 61-74, 1948.

[۳۸] م. ع. نژادی و ح. حسن پور، "تحلیل و پردازش سیگنال به کمک روش های زمان فرکانس," *مجله علوم رایانشی*, جلد ۲, ۱۳۹۸.

[39] R. B. Pachori and P. Sircar, "A new technique to reduce cross terms in the Wigner distribution," *Digital Signal Processing,* vol. 17, no. 2, pp. 466-474, 2007.

[40] F. Auger and P. Flandrin, "Time – Frequency Toolbox For Use with MATLAB," 1995, pp. 1-150.

[41] S. Group, "SPEECH DATABASE DEVELOPMENT AT MIT: TIMIT AND BEYOND," *Speech Communication,* vol. 2, no. 4, pp. 351-356, 1990.

[42] R, Subramanian, "Automatic Word Boundary Detection Using Adaptive Scaling", Northeastern University, Boston, MA, USA, 2010.

Abstract

Word boundary detection means the beginning and end of a word in a speech signal, which plays an important role for most applications such as speech recognition, text-to-speech system, and so on. In speech recognition system, most errors are due to incorrect recognition of word boundaries. Correct word boundaries reduces detection errors and improves systems performance.There are many algorithms used for detecting the boundary of words, some of which are not good enough to correct the word boundary due to the noise of the environment or a good pause between words. Which can be reduced by noise methods to be able to identify the borders well. While some methods are combined, they can better detection word boundaries.

In this research, the aim is to detection word boundaries using time-frequency methods such as wavelet and Wigner-Ville. Time-frequency methods examine the signal in both time and frequency domains. Because the speech signal is an unstable signal and its characteristics change with time, the best way to analyze this signal is to extract features from those time-frequency methods. Different methods of extracting MFCC, PLP, LPC, wavelet analysis and Wigner-Ville method were examined to identify word boundaries.
Two types of data were used to evaluate the results. In the TIMIT database, time between words is very short and words are pronounced naturally and almost quickly, while in CADLab words are expressed numerically. The Wigner-Will method is able to identify word boundaries with 83.5% for the first data, and 68% for the second data is the best way to identify word boundaries. Entropy with 25% detection for the first data and ZRC with 30% detection for the second data has the lowest rate of correct detection of word boundaries among the methods performed.
Keywords: Word Boundary detection, Feature Extraction, Time-Frequency, MFCC, PLP, LPC

Shahrood University of Technology

Faculty of Electrical and Robatic Engineering

Rock Mechanics MSc Thesis in

Words boundary detection in speech signal using time-frequancy method

By:

Fateme Tajik Ijdan

Supervisor:

Dr Hossein Marvi

October 2021