
بسم الله الرحمن الرحيم

دانشکده برق و رباتیک

گروه رباتیک

پایان‌نامه کارشناسی ارشد

استخراج ویژگی‌های زمانی- فرکانسی از سیگنال‌های گفتار فارسی جهت بهبود کارایی یک سیستم VAD برای ارتباط انسان و ربات

فهیمه جمهوری

استاد راهنما :

حسین مروی

استاد مشاور:

علیرضا احمدی فرد

بهمن ۱۳۹۳

شماره: ۱۷۴۸/ت ب
تاریخ: ۹۳/۱۱/۱۹
ویرایش: ——

بسمه تعالی

مدیریت تحصیلات تکمیلی
فرم شماره (۶)

فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

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

فهیمه جمهوری شوکت آباد رشته: **برق - رباتیک** گرایش: **بینایی ربات**

تحت عنوان: **استخراج ویژگیهای زمانی- فرکانسی از سیگنالهای گفتار فارسی جهت بهبود کارایی یک سیستم VAD برای ارتباط انسان و ربات**

که در تاریخ **۹۳/۱۱/۱۹** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است:

قبول (با مرجه: عالی امتیاز ۱۹٫۸۵) ☑ دفاع مجدد ☐ مردود ☐

۱- عالی (۲۰ - ۱۹)
۲- بسیار خوب (۱۸/۹۹ - ۱۸)
۳- خوب (۱۷/۹۹ - ۱۶)
۴- قابل قبول (۱۵/۹۹ - ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیات داوران	نام ونام خانوادگی	مرتبه علمی	امضاء
۱- استاد راهنما		استادیار	
۲- استاد مشاور		دانشیار	
۳- نماینده شورای تحصیلات تکمیلی		استادیار	
۴- استاد ممتحن		دانشیار	
۵- استاد ممتحن		استادیار	

رئیس دانشکده:

تقدیم

به پدر و مادرم که همواره یاری رسانم بودند،

به معلمانم که چون شمع روشنگر راهم بودند،

و به تمام علاقه‌مندان علم پردازش گفتار.

سپاس‌گذاری

سپاس خدایی را که با گشودن درهای علم و دانش خود به رویم، مرا بیش از پیش به سمت یگانگی‌اش رهنمون ساخت و با آفرینش انسان‌هایی چون پدر و مادر و معلمان موجبات رشد و شکوفایی‌ام را فراهم آورد. در سایه الطاف او و یاری بندگانش این دفتر به پایان آمد.

از استاد گرامی جناب آقای دکتر مروی که مرا با دنیای شیرین پردازش گفتار آشنا کردند و در مراحل انجام کار با راهنمایی‌های صبورانه و نقطه نظرات خویش روشنگر راهم بودند صمیمانه سپاسگزارم.

از استاد ارجمند جناب آقای دکتر احمدی‌فرد که مشاوره این پایان‌نامه را پذیرفتند کمال سپاس را دارم.

از اساتید گرانمایه‌ای که با تشویق‌هایشان انگیزه ادامه تحصیل را در من ایجاد نمودند بی‌نهایت سپاسگزارم.

از تمام معلمان و اساتید و همه کسانی که به نحوی تا کنون به من آموختند متشکرم.

و

سپاس آخر را به مهربان‌ترین همراهان زندگی‌ام، پدر و مادر عزیزم تقدیم می‌کنم که حضورشان در فضای زندگیم مصداق بارز سخاوت بوده است.

فهیمه جمهوری

۱۳۹۳

تعهد نامه

اینجانب **فهیمه جمهوری شوکت آباد** دانشجوی دوره کارشناسی ارشد رشته **مهندسی رباتیک** دانشکده **برق و رباتیک** دانشگاه صنعتی شاهرود نویسنده پایان نامه **استخراج ویژگی‌های زمانی- فرکانسی از سیگنال‌های گفتار فارسی جهت بهبود کارایی یک سیستم VAD برای ارتباط انسان و ربات** تحت راهنمائی **جناب آقای دکتر حسین مروی** متعهد می شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد 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 می گردد.
- در کلیه مراحل انجام این پایان نامه، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

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

چکیده

استفاده از ربات‌های اجتماعی در زندگی انسان‌ها افزایش یافته است و اصلی‌ترین راه ارتباط انسان‌ها با آنها ارتباط کلامی است. ربات‌های اجتماعی دارای میکروفونی هستند تا بتوانند سیگنال گفتار را برای ارتباط با انسان دریافت کنند. با توجه به دریافت نویز محیطی به هنگام ضبط گفتار انسان، نیاز به سیستمی برای تشخیص قسمت‌های گفتار در سیگنال‌های صوتی ضبط شده می‌باشد. هدف این پایان‌نامه طراحی یک سیستم آشکارساز فعالیت گفتاری است که بتواند با مشخص نمودن بخش‌های گفتار در یک محیط نویزی کارایی یک سیستم پردازش گفتار مورد استفاده برای یک ربات اجتماعی را افزایش دهد.

در این پایان‌نامه، ویژگی‌های مختلفی جهت استخراج از سیگنال گفتار برای سیستم آشکارساز فعالیت گفتاری پیشنهاد شده است. این ویژگی‌ها با استفاده از ترکیب ویژگی انرژی با هر یک از ویژگی‌های ضرایب کپسترال فرکانس مل ریشه، ضرایب کپسترال فرکانس بارک، ضرایب پیش‌بینی ادراکی خطی و ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده ارائه شده‌اند.

روش پیشنهادی دیگری که در این پایان‌نامه مطرح شده بر مبنای ویژگی ویگنر ویل به عنوان یک روش استخراج ویژگی زمانی-فرکانسی است. این روش در مقایسه با روش‌های پیشنهادی مبتنی بر ویژگی انرژی و ویژگی‌های کپسترال، کارایی بهتری دارد. بنابراین برای افزایش کارایی روش‌های مبتنی بر ویژگی‌های کپسترال علاوه بر ویژگی انرژی، با ویژگی ویگنر ویل نیز ترکیب شده‌اند.

برای ارزیابی کارایی روش‌های پیشنهادی، دادگان گفتاری فارس‌دات که یک پایگاه داده استاندارد و به زبان فارسی می‌باشد، به کار رفته است. با افزودن چند نوع متفاوت نویز به فایل‌های این پایگاه داده در شرایط مختلف نسبت سیگنال به نویز، مقاومت روش‌های پیشنهادی نسبت به نویزهای متفاوت سنجیده شده‌اند.

آزمایشات نشان داده‌اند که برخی از روش‌های استخراج ویژگی پیشنهادی در این پایان‌نامه در مقایسه با روش متداول استخراج ویژگی ضرایب کپسترال فرکانس مل در محیطهای نویزی کارایی بهتری دارند.

کلمات کلیدی: آشکارساز فعالیت گفتاری، توزیع ویگنر ویل، ضرایب کپسترال فرکانس مل ریشه، ضرایب کپسترال فرکانس بارک، ضرایب پیش‌بینی ادراکی خطی، ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده، انرژی، استخراج ویژگی، ارتباط بین انسان و ربات، ربات‌های اجتماعی.

مقالات مستخرج از پایان نامه

۱- فهیمه جمهوری، حسین مروی، علیرضا احمدی‌فرد، "استفاده از ویژگی پیش‌بینی ادراکی خطی جهت افزایش کارایی یک سیستم تشخیص‌دهنده گفتار از غیرگفتار"، ارائه شده در چهارمین کنفرانس بین‌المللی آکوستیک و ارتعاشات ISAV2014، تهران، دانشگاه علم و صنعت ایران، پاییز ۹۳.

۲- فهیمه جمهوری، حسین مروی، علیرضا احمدی‌فرد، " معرفی یک سیستم تشخیص‌دهنده گفتار از غیرگفتار بر اساس ضرایب کپسترال فرکانس بارک برای تشخیص هویت گوینده"، پذیرفته شده در بیستمین کنفرانس انجمن کامپیوتر ایران CSICC2015، مشهد، دانشگاه فردوسی مشهد.

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

لیست علائم اختصاری

AMR: Adaptive Multirate

BFCC: Bark Frequency Cepstral Coefficients

DFT: Discrete Fourier Transform

ELRA: European Language Resources Association

EM: Expectation Maximization

FFT: Fast Fourier Transform

HRI: Human Robot Interaction

IPA: International Phonetic Association

LLR: Log Likelihood Ratio

LPC: Linear Predictive Coefficients

LSF: Line Spectral Frequencies

MFCC: Mel Frequencies Cepstral Coefficients

MMSE: Minimum Mean Square Error

PLP: Perceptual Linear Prediction

RMFCC: Root Mel Cepstral Coefficients

RPLP: Revised Perceptual Linear Prediction

SNR: Signal to Noise Ratio

SVM: Support Vector Machine

VAD: Voice Activity Detector/Detection

WVD: Wigner Ville Distribution

فصل ۱

مقدمه

۱-۱ پیشگفتار

توانایی انسان‌ها برای انتقال ایده‌ها و مفاهیم از طریق زبان نوشتاری و گفتاری یکی از بزرگترین ویژگی‌هایی است که آنها را از سایر حیوانات و جانوران متمایز می‌کند. گفتار در نتیجه یک سری انتقالات پیچیده در سطوح مختلف معنایی تولید شده است. اطلاعات موجود در سیگنال گفتار دربردارنده پیام، زبان، شناسه، حالت فیزیکی و عاطفی گوینده می‌باشند. به دلیل شباهت زیاد بین ویژگی‌های ادراکی سیگنال گفتار و سیگنال‌های مشابه آن، تعیین خودکار الگویی مشخص از سیگنال گفتار، همچنان به عنوان مسئله‌ای چالش برانگیز مطرح است. توسعه ماشین هایی که اطلاعات کلمات بیان شده را پردازش می‌نمایند، دامنه‌ای از تحقیقات علمی را در چند دهه اخیر به وجود آورده است. همچنین با گسترش علم و تکنولوژی، ساخت ربات‌هایی که بتوانند از طریق تشخیص گفتار با انسان‌ها ارتباط برقرار کنند، یکی از موضوعات مهم دنیای رباتیک را شکل داده است.

از آنجا که امروزه ربات‌ها ازکارخانه‌ها وارد زندگی روزانه افراد در سراسر جهان شده‌اند، بدون تردید مهارت‌های مورد نیاز آنها بسیار متفاوت از مهارت‌هایی است که در قرن بیستم برای آنان برنامه‌ریزی شده بود. به این دلیل که این دسته از ربات‌ها به زندگی اجتماعی انسان‌ها راه پیدا کرده‌اند، آن‌ها را ربات‌های اجتماعی[۱] می‌نامند. در آینده نزدیک حضور ربات‌های اجتماعی در جوامع بشری به سرعت افزایش خواهند یافت. ایده اصلی در طراحی این ربات‌ها، داشتن ارتباط با انسان‌هاست. پدیده ارتباط بین انسان و ربات[۲] یکی از بزرگترین مشکلاتی است که بشر تاکنون در طراحی ربات‌ها با آن مواجه بوده است. ارتباط بین انسان و ربات یک زمینه تحقیقاتی جدید است که تعداد علاقه‌مندان به آن رو به افزایش هستند. مبحث ارتباط بین انسان و ربات خود مجموعه‌ای از سایر مباحث علمی است. در این رابطه می‌توان به ارتباط بین انسان و ماشین یا کامپیوتر، رباتیک[۳] و هوش مصنوعی[۴] اشاره کرد. هدف

---

[۱] social robots

[۲] Human Robot Interaction (HRI)

[۳] Robotics

[۴] Artificial Intelligence

اصلی سیستم‌های محاوره خودکار در رباتیک برقراری یک رابطه طبیعی بین انسان و ربات است، مشابه روابطی که بین انسان‌ها وجود دارد.

از مزایای این ارتباط می‌توان به حذف نیاز به محصولات مصنوعی ورود اطلاعات مانند موس و کیبورد اشاره کرد. همچنین روش مناسبی را برای استفاده کاربران غیرحرفه‌ای یا دارای معلولیت فراهم می‌آورد. از این ربات‌ها می‌توان به منظور انجام امور خانه‌داری، برقراری امنیت، مراقبت‌های پزشکی از بیماران، ارتباط در یک شبکه خانگی و ایجاد سرگرمی برای کودکان و نوجوانان استفاده کرد.

به منظور داشتن رباتی که بتواند به صورت طبیعی با انسان ارتباط برقرار کند، انتظار داریم که قادر باشد ما را ببیند، صدایمان را بشنود، دارای حس لامسه باشد و همه اطلاعاتی که از این حس‌ها دریافت می‌کند را بفهمد. امروزه پس از طراحی و ساخت یک ربات اجتماعی، پیکره فیزیکی آن به همراه موتورهایش همگذاری[۱] می‌شوند، سپس درایورهای سخت‌افزاری و کنترل کننده‌های نرم‌افزاری برای تولید حرکت در آن آغاز به کار می‌کنند. علاوه بر آن ربات نیازمند یک سیستم بینایی کامپیوتری است تا بتواند اهداف مطلوب مانند چهره انسان را تشخیص داده و محل آنها را به درستی پیدا کند. برای یک ارتباط مؤثر بین انسان و ربات، لازم است که انسان بتواند از طریق گفتار با ربات ارتباط برقرار کند. بنابراین ربات باید قادر به شناسایی گوینده از بین گروهی از افراد باشد و سیگنال گفتار را در محیط واقعی تشخیص دهد. لذا ربات نیازمند یک سیستم تشخیص گفتار[۲] و تشخیص گوینده[۳] نیز می‌باشد [۱].

تاکنون محققان بسیاری تحقیقات علمی خود را به مبحث بینایی ربات اختصاص داده‌اند. اما در توسعه تکنولوژی ارتباط با ربات از طریق کانال صوتی کارهای اندکی انجام شده است. بدین منظور اطلاعات سیستم تشخیص گفتار یا تشخیص گوینده از طریق میکروفونی که به ربات متصل است تهیه می‌شود

[۱] assemble
[۲] Speech Recognition
[۳] Speaker Recognition

[۲]. این میکروفون علاوه بر صدای افراد در معرض کلیه صداهای نویز موجود در محیطهای بی‌درنگ قرار دارد لذا انواع منابع نویز به سیگنال گفتار خسارت وارد می‌کنند و باعث کاهش کیفیت و قابلیت اطمینان سیستم تشخیص گفتار یا تشخیص گوینده می‌شوند. هر سیستم پردازش گفتار قبل از شروع به پردازش داده‌ها، باید دارای داده‌های ورودی معتبر باشد. بنابراین لازم است که به عنوان یک مرحله پیش پردازش حضور و عدم حضور گفتار در سیگنال ورودی بررسی شود. اگر این واحد درست عمل نکند ممکن است عملکرد سیستم پردازش گفتار در محیطهای نویزی با شکست مواجه شود. پس هر سیستم پردازش گفتار نیاز به یک مؤلفه آشکارساز فعالیت گفتاری[۱] دارد و اگر این مؤلفه قابلیت اعتماد کافی را نداشته باشد، دقت سیستم به شدت پایین می‌آید [۳, ۴].

آشکارسازی فعالیت گفتاری، مسأله تشخیص حضور گفتار انسان در یک سیگنال صوتی است. یک سیستم آشکارساز فعالیت گفتاری، داده‌های گفتار را از یک سیگنال ورودی استخراج می‌کند و قسمت‌های غیرگفتار آن را نادیده می‌گیرد. در یک سیگنال خالص و عاری از نویز یا در سیگنالی که نسبت سیگنال به نویز[۲] آن بسیار بالا است، این مسأله به سادگی و با استفاده از یک آستانه[۳] انرژی حل خواهد شد. به عنوان مثال در یک اتاق ساکت و آرام، سطح انرژی گفتار انسان در مقایسه با سطح انرژی غیرگفتار بسیار بالاست. بنابراین استفاده از یک آستانه انرژی می‌تواند برای تشخیص مناسب باشد [۵]. اما هنگامی که سیگنال توسط نویز آسیب می‌بیند، تمایز بین گفتار و غیرگفتار می‌تواند بسیار سخت باشد. بسته به محیطی که گفتار در آن تولید می‌شود یا صدا در آن ضبط می‌گردد، سیگنال غیرگفتار موجود در محیط می‌تواند سکوت، نویز، موسیقی یا انواع سیگنال‌های صوتی دیگر مانند صدای در زدن، سرفه کردن، صدای دستگاه تهویه هوا یا حتی صدای همهمه ناشی از گفتگوی افراد باشد. برخی از انواع نویز به راحتی قابل حذف شدن هستند و می‌توان آنها را کنترل کرد. به عنوان مثال نویزهای ایستا[۴]

---

[۱] Voice Activity Detection

[۲] Signal to Noise Ratio (SNR)

[۳] threshold

[۴] stationary

مانند صدای دستگاه تهویه هوا از این قبیل هستند. این دسته از نویزها را می‌توان با استفاده از روش‌های بهبود گفتار و حذف نویز به سادگی فیلتر کرد [۶]. اما در محیطهای پیچیده‌تر مانند خیابان یا در مراکز خرید یا در ایستگاه‌های قطار، از آنجا که سطح نویز می‌تواند حتی بالاتر از گفتار هدف باشد تشخیص گفتار بسیار مشکل است. به خصوص اگر نویز پس‌زمینه شامل صدای همهمه باشد که مشخصات آماری آن بسیار مشابه مشخصات گفتار است [۷]. انتخاب یک سیستم آشکارساز فعالیت گفتاری تحت تأثیر محیط عامل قرار دارد. فاکتورهایی مانند نیاز به انجام انطباق نویز، تاخیر وارد شده و هزینه محاسباتی باید قبل از تصمیم‌گیری برای یک الگوریتم خاص مورد توجه قرار گیرند.

هدف از این پایان نامه طراحی یک سیستم آشکارساز فعالیت گفتاری است که بتواند با تامین ورودی معتبر برای یک سیستم پردازش گفتار، کارایی آن را افزایش دهد. سیستم پردازش گفتار مورد نظر برای ارتباط با رباتی تعریف شده است که صداهای محیط را از طریق میکروفون متصل به آن دریافت می‌کند. این ربات یک ربات اجتماعی است و محیطهایی که ربات در آن قرار دارد محیط خانه می‌باشد. واضح است که وجود نویز یکی از اجزاء جدا نشدنی محیط خانه است و انواع مختلف نویز در این محیط وجود دارند. این نویزها عبارتند از صدای ماشین لباس‌شویی، صدای دستگاه تهویه هوای مطبوع، صدای همهمه ناشی از سخن گفتن اشخاص در محیط، صدای سرفه و عطسه، صدای خندیدن، صدای باز و بسته شدن در و پنجره و سایر صداهایی که در یک محیط خانگی مرسوم می‌باشند. بنابراین هدف اصلی ما در این پایان‌نامه طراحی سیستم آشکارساز فعالیت گفتاری است که بتواند در برابر انواع نویزهای موجود در محیط مقاوم باشد.

۱-۲ ساختار پایان‌نامه

محتوای این پایان‌نامه در پنج فصل تنظیم شده است. در فصل دوم به مرور سیستم‌های آشکارساز فعالیت گفتاری می‌پردازیم. مراحل طراحی این سیستم‌ها، معیارهای ارزیابی کارایی آنها و دو روش استاندارد در این فصل بیان می‌شوند. در فصل سوم مراحل استخراج ویژگی‌های استفاده شده در این

پایان نامه شرح داده می‌شوند. این روش‌ها شامل چندین ویژگی حوزه کپسترال و توزیع ویگنر ویل می‌باشند. فصل چهارم روش‌های آشکارساز فعالیت گفتاری پیشنهادی را توصیف می‌کند. در فصل پنجم ابتدا ویژگی‌های دادگان گفتاری فارس‌دات بیان می‌شوند، سپس رابطه‌ای برای محاسبه خطا ارائه می‌گردد، پارامترهای استفاده شده در هر کدام از روش‌ها تعیین می‌شوند و سپس آزمایشات انجام شده و نتایج آنها ارائه می‌گردند.

فصل ۲

مروری بر کارهای گذشته

۲-۱ پیشگفتار

آشکارسازی فعالیت گفتاری، مشکل شناسایی نواحی گفتار در یک سیگنال صوتی داده شده است. قبل از اینکه یک سیستم پردازش گفتار شروع به پردازش داده‌ها کند، باید بتوانیم داده‌های معتبر را برای این سیستم فراهم کنیم. یک سیستم آشکارساز فعالیت گفتاری، داده‌های گفتار را از یک سیگنال داده شده استخراج می‌کند و قسمت‌های غیرگفتار آن را نادیده می‌گیرد. تقریبا همه سیستم‌های پردازش گفتار نیاز به یک سیستم آشکارساز فعالیت گفتاری دارند. زیرا اولین مرحله پردازش گفتار در همه این سیستم‌ها، جداسازی قابل اطمینان قسمت‌های گفتار و غیرگفتار از یکدیگر است که این مطلب توسط یک سیستم آشکارساز فعالیت گفتاری به دست می‌آید. این سیستم یک مؤلفه ضروری در فرآیندهایی مانند انتقال گفتار، کاهش نویز، تشخیص گفتار و کدکردن نرخ متغیر گفتار می‌باشد.

وظیفه آشکارساز فعالیت گفتاری تفکیک دو کلاس گفتار (که می‌تواند همراه با نویز باشد) و غیرگفتار از یکدیگر است. در محیطهای عاری از نویز، بیشتر الگوریتم‌ها به خوبی کار می‌کنند اما در حضور نویز کارایی خود را به صورت قابل توجهی از دست می‌دهند و با خطاهای بسیار زیادی مواجه خواهند شد. هنگامی که سیگنال گفتار توسط نویز یا انعکاس صدا به هم می‌ریزد، تشخیص‌های نادرستی در خروجی آشکارساز فعالیت گفتاری رخ می‌دهد. لذا مقاوم بودن نسبت به نویز یکی از نیازهای ضروری یک سیستم آشکارساز فعالیت گفتاری است. علاوه بر نیاز به مقاوم بودن نسبت به نویز، بسته به کاربرد آشکارساز فعالیت گفتاری ممکن است محدودیت‌های دیگری نیز روی الگوریتم وجود داشته باشد. به عنوان مثال فرآیندهای بی‌درنگ[۱] نیاز به الگوریتم‌هایی با تاخیر کمتر دارند.

اکثر سیستم‌های آشکارساز فعالیت گفتاری موجود از یک ساختار پیمانه‌ای[۲] خاص [۸] مانند آنچه در شکل (۲-۱) وجود دارد پیروی می‌کنند. در این ساختار یک سیگنال صوتی ضبط شده از طریق یک

[۱] real time

[۲] modular

میکروفون، ابتدا از یک واحد حذف نویز عبور می‌کند تا نویزها را فیلتر نماید و نسبت سیگنال به نویز را بهبود بخشد. پس از آن واحد بعدی ویژگی‌های صوتی را استخراج می‌کند و یک لیست از پارامترهای سیگنال را برای استفاده مرحله بعد تهیه می‌کند. در مرحله تصمیم‌گیری مشخص می‌شود که هر فریم سیگنال متعلق به کلاس گفتار یا غیرگفتار است. سپس یک روش بهبود و اصلاح تصمیمات[۱] برای افزایش تشخیص‌های درست و کاهش هشدارهای نادرست به کار گرفته می‌شود. در نهایت ممکن است برخی پردازش‌های دلخواه برای نمایش بهتر نتایج تشخیص به کار روند. از بین این واحدها تنها دومین و سومین واحد آن در هسته یک سیستم آشکارساز فعالیت گفتاری قرار دارند که در اینجا به آنها بیشتر خواهیم پرداخت. در حالت کلی یک الگوریتم آشکارساز فعالیت گفتاری معمولی شامل مراحل زیر می‌باشد:

۱- بهبود گفتار و حذف نویز[۲] : برای فیلتر کردن و حذف نویز از سیگنال گفتار، از روش‌های بهبود گفتار استفاده می‌شود که قسمت ۲-۲ به توضیح این مرحله می‌پردازد.

۲- استخراج ویژگی[۳] : از سیگنال صوتی یک سری ویژگی‌هایی استخراج می‌شوند که به عنوان پارامتر در مراحل بعدی و برای تشخیص فریم‌های گفتار و غیرگفتار به کار می‌روند. در قسمت ۲-۳ درباره استخراج ویژگی توضیح بیشتری داده خواهد شد.

۳- تصمیم‌گیری اولیه : در مورد اینکه فریم جاری سیگنال صوتی گفتار یا غیر گفتار است تصمیم‌گیری می‌شود. برای این منظور از قوانین تصمیم‌گیری، آستانه‌های تطبیقی[۴]، مدل‌های آماری، روش‌های یادگیری ماشین[۵] و شناسایی آماری الگو استفاده می‌شود که این مورد در قسمت ۲-۴ توصیف می‌شود.

---

[۱] hangover

[۲] Speech Enhancement and Noise Reduction

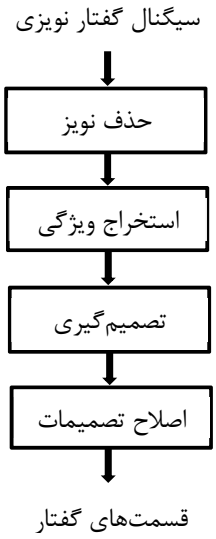
[۳] Feature Extraction

[۴] Adaptive Thresholds

[۵] Machine Learning

۴- بهبود و اصلاح تصمیمات : از آنجا که سیگنال گفتار یک سیگنال همبسته است، در صورتی که فریم جاری را به عنوان فریم گفتار تشخیص دهیم فریم بعدی نیز به احتمال بسیار بالایی فریم گفتار می‌باشد. الگوریتم‌های عادی آشکارساز فعالیت گفتاری، به منظور جلوگیری از انتقال‌های سریع از گفتار به غیرگفتار، تصمیمات اولیه خود را اصلاح می‌کنند. و سپس تخمین‌های نسبت سیگنال به نویز و سایر میانگین‌های در حال اجرا را روی لبه‌ها به روزرسانی می‌نمایند. قسمت ۲-۵ به شرح این مرحله می‌پردازد.

در قسمت ۲-۶ دو روش استاندارد سیستم‌های آشکارساز فعالیت گفتاری معرفی می‌شوند. در قسمت ۲-۷ معیار ارزیابی کارایی یک سیستم آشکارساز فعالیت گفتاری ارائه می‌گردد و در نهایت در قسمت ۲-۸ جمع‌بندی کلی فصل انجام می‌شود.



شکل (۲-۱) ساختار پیمانه‌ای یک سیستم آشکارساز فعالیت گفتاری

۲-۲ بهبود گفتار و حذف نویز

نویز پس‌زمینه عمومی‌ترین فاکتور کاهش کیفیت و قابلیت فهم گفتار است. هدف واحد بهبود گفتار و حذف نویز این است که بدون آسیب رساندن به کیفیت سیگنال گفتار سطح نویز را کاهش دهد. یکی

از روش‌های بهبود گفتار روش تفاضل اسپکترال[1] است که به عنوان یکی از روش‌های کاهش نویز جمع‌شونده کاربرد گسترده‌ای دارد. در ادامه توضیح مختصری درباره آن بیان می‌شود.

۲-۲-۱ تفاضل اسپکترال

روش تفاضل اسپکترال یک روش ساده و کارا برای کاهش نویز جمع‌شونده در سیگنال گفتار است. تکنیک‌های مربوط به تخمین نویز و تفاضل اسپکترال در حقیقت قلب مسائل بهبود گفتار می‌باشند. این روش‌ها در سیستم‌های آشکارساز فعالیت گفتاری می‌توانند به دو صورت مطرح شوند : یا به صورت صریح و در یک مرحله پیش‌پردازش [۹] و یا به صورت ضمنی و در مرحله استخراج ویژگی [۶, ۸].

ایده اصلی این روش برگرفته از این امر می‌باشد که اگر نویز محیط به صورت جمع‌شونده فرض شود، می‌توان اسپکتروم نویز را از اسپکتروم گفتار نویزی طبق رابطه (۲-۱) کم کرد [۱۰]. بدین ترتیب گفتار خالص بدون نویز به دست می‌آید.

$$|\hat{X}_k|^2 = |X_k|^2 - |\hat{N}|^2 \qquad (۲-۱)$$

در رابطه (۲-۱)، $|\hat{X}_k|^2$ اسپکتروم توان گفتار خالص تخمینی در فریم k ام، $|X_k|^2$ اسپکتروم توان گفتار نویزی و $|\hat{N}|^2$ اسپکتروم توان نویز تخمینی را نشان می‌دهند. با توجه به رابطه (۲-۱) در این روش نیاز به تخمین اسپکتروم نویز می‌باشد که این تخمین باید حول نواحی غیرگفتار صورت گیرد. با این فرض که اولین فریم‌های یک فایل گفتاری، فریم‌های سکوت هستند لذا میانگین تعدادی از فریم‌های اولیه می‌تواند تخمین خوبی برای نویز باشد. هر چه از تعداد فریم‌های بیشتری برای تخمین استفاده شود، تخمین بهتری حاصل خواهد شد [۶]. این روش ساده‌ترین فرم روش تفاضل اسپکترال است. گرچه این پردازش را می‌توان در حوزه فرکانس نیز بیان کرد.

[1] Spectral Subtraction

با فرض اینکه $|X|^2$ و $|\widehat{N}|^2$ به ترتیب اسپکتروم توان گفتار نویزی و اسپکتروم توان نویز تخمینی باشند، روش دیگر محاسبه تفاضل اسپکترال، از ضرب کردن اندازه گفتار نویزی $|X|$ در فاکتور بهره g حاصل می‌شود که این فاکتور بهره طبق رابطه (۲-۲) به دست می‌آید [۱۱].

$$g = max\left\{\left(1-\left(\alpha\frac{|\widehat{N}|^2}{|X|^2}\right)^{\gamma/2}\right)^{e/\gamma}, min\left(g_h, \left(\beta\frac{|\widehat{N}|^2}{|X|^2}\right)^{e/2}\right)\right\} \quad (۲-۲)$$

در رابطه (۲-۲)، α فاکتور تفاضل، γ ناحیه تفاضل، e توان بهره، g_h ماکزیمم برای نویز محدوده زمین و β ماکزیمم کاهش نویز در ناحیه توان را مشخص می‌کنند.

۲-۳ استخراج ویژگی

مهم‌ترین مرحله در هر سیستم آشکارساز فعالیت گفتاری، استخراج ویژگی‌های صوتی از سیگنال گفتار است [۱۲]. در بسیاری از روش‌ها فرض می‌شود که سیگنال گفتار در طول یک بازه حدود چند میلی ثانیه ایستا است، بنابراین سیگنال گفتار به قسمت‌هایی با طول ثابت و هم پوشانی کوتاه به نام فریم تقسیم می‌شود. سپس به منظور کاهش اثر ناپیوستگی در ابتدا و انتهای فریم، هر فریم حوزه زمان در یک تابع پنجره ضرب می‌گردد. انواع توابع پنجره معروف عبارتند از: پنجره مستطیلی، پنجره همینگ[۱]، پنجره بارتلت[۲] یا مثلثی و پنجره هنینگ[۳]. فرمول هر کدام از این توابع به ترتیب در روابط (۲-۳)، (۲-۴)، (۲-۵) و (۲-۶) نشان داده شده است [۱۳]:

$$w(n) = \begin{cases} 1 & 0 \le n \le M \\ 0 & otherwise \end{cases} \quad (۲-۳)$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2n\pi/M) & 0 \le n \le M \\ 0 & otherwise \end{cases} \quad (۲-۴)$$

$$w(n) = \begin{cases} 2n/M & 0 \le n \le M/2 \\ 2 - 2n/M & M/2 < n \le M \\ 0 & otherwise \end{cases} \quad (۲-۵)$$

[۱] Hamming Window

[۲] Bartlett Window

[۳] Hanning Window

$$w(n) = \begin{cases} 0.5 - 0.5\cos(2n\pi/M) & 0 \le n \le M \\ 0 & otherwise \end{cases} \quad (۶-۲)$$

انتخاب ویژگی یک مسأله حساس و بحرانی در هر مسأله کلاسه‌بندی[۱] است. در سیستم آشکارساز فعالیت گفتاری ویژگی‌های خوب باید دارای خواص زیر باشند:

۱- قدرت جداکنندگی خوب: این مورد به معنای میزان جداکنندگی بین توزیع فریم‌های گفتار نویزی و فریم‌های فقط نویز می‌باشد. به صورت تئوری یک ویژگی خوب نباید بین کلاس‌های نویز و گفتار همپوشانی داشته باشد.

۲- مقاوم بودن به نویز: مقاومت در برابر نویز، کارایی کلاسه‌بند را در برنامه‌های کاربردی دنیای واقعی تضمین می‌کند. در شرایطی که نویز پس‌زمینه سیگنال گفتار را به هم بریزد، قدرت جداکنندگی ویژگی‌های استخراج شده پایین می‌آید.

ویژگی‌های بسیاری در متون علمی پیشنهاد شده‌اند که می‌توانیم آنها را بدین طریق دسته‌بندی نماییم: ویژگی‌های مبتنی بر انرژی، ویژگی‌های مبتنی بر اسپکترال، ویژگی‌های مبتنی بر کپسترال و ویژگی‌های بلند مدت. هیچ‌کدام از ویژگی‌ها یا ترکیبی از آنها به صورت یکنواخت و در همه شرایط کارایی خوبی ندارند. به عنوان مثال روش‌های مبتنی بر انرژی در شرایطی که نسبت سیگنال به نویز کم باشد کارایی بسیار کمی دارند و سایر روش‌ها نیازمند تخمین این مقدار می‌باشند که این تخمین خود به تنهایی یک مسأله چالش برانگیز در محیطهای غیرایستاست. در محیطهایی که به سرعت تغییر می‌کنند پارامترهای الگوریتم نیازمند بروزرسانی تطبیقی می‌باشند. در بخش‌های بعدی هر دسته از این ویژگی‌ها با جزئیات بیشتری بیان می‌شوند.

[۱] classifying

۲-۳-۱ ویژگی‌های مبتنی بر انرژی

انرژی یک محاسبه ساده از بلندی سیگنال است. در یک روش معمولی برای آشکارساز فعالیت گفتاری می‌توان فرض کرد که گفتار همیشه بلندتر از نویز پس‌زمینه است و سپس فریم‌های با انرژی بالا را به عنوان فریم‌های گفتار و فریم‌های با انرژی پایین را به عنوان فریم‌های غیرگفتار یا نویز در نظر گرفت. اگرچه هنگامی که بلندی گفتار و نویز در سطوح یکسانی قرار دارند، ویژگی انرژی ساده برای جدا کردن گفتار و نویز با شکست مواجه خواهد شد. مانند حالتی که نویز پس‌زمینه در محیط در حال افزایش است یا قسمت‌های گفتار دارای انرژی کمی هستند. در کارهای اولیه در زمینه آشکارساز فعالیت گفتاری، از انرژی در زیرباندهای متفاوت به منظور افزایش قدرت جداکنندگی استفاده شده است. به عنوان مثال، آزمایش اسپکتروم گفتار و نویز نشان می‌دهد هنگامی که انرژی نویز سفید به صورت کلی در طول اسپکتروم گسترش می‌یابد، گفتار صدادار[۱] در باندهای فرکانسی پایین (کمتر از ۲ کیلو هرتز) انرژی بالایی دارد و همچنین گفتار بی‌صدا[۲] در باندهای فرکانسی بالا (هم در ۲ تا ۴ کیلو هرتز و هم بالاتر از ۴ کیلو هرتز) فعال است. روش دیگر برای بهبود مقاومت نویز، ترکیب کردن ویژگی‌های مبتنی بر انرژی با سایر ویژگی‌ها مانند نرخ عبور از صفر [۱۲] یا فرکانس اسپکترال خطی [۱۴] است. معمولا این ویژگی‌ها با گفتار خالص یا شرایط نسبت سیگنال به نویز بالا خوب کار می‌کنند ولی تحت شرایطی که سطح نویز بالا و نسبت سیگنال به نویز کمتر از ۱۰ دسی‌بل است، قدرت جداکنندگی آنها به شدت پایین می‌آید. با این حال به دلیل پیچیدگی محاسباتی کم آنها، ویژگی‌های مبتنی بر انرژی هنوز توسط برخی استانداردها و برنامه‌های کاربردی مختلف دنیای واقعی به کار می‌روند. یک نمونه عمومی آن استاندارد G.729 Annex B [۱۴] است که برداری از ویژگی‌ها را به کار می‌برد. این بردار شامل انرژی باند کامل، نرخ عبور از صفر و انرژی باند پایین (از ۰ تا ۱ کیلوهرتز) می‌باشد. این استاندارد به عنوان یک سند و یک سیستم پایه جهت مقایسه کارایی در بسیاری از تحقیقات به کار می‌رود.

---

[۱] Voiced Speech

[۲] Unvoiced Speech

۲-۳-۲ ویژگی‌های مبتنی بر اسپکترال

یکی از عمومی‌ترین تکنیک‌های پردازش گفتار آنالیز اسپکتروم[۱] فرکانس است که محتوای فرکانس سیگنال در زمان را توصیف می‌کند [۱۳]. این عمل توسط توسعه الگوریتم تبدیل فوریه سریع[۲] امکان می‌یابد که اجرای تبدیل فوریه را با پیچیدگی زمانی $O(n \log n)$ به جای $O(n^2)$ امکان‌پذیر می‌نماید.

با توجه به متون علمی در سیستم‌های آشکارساز فعالیت گفتاری، ویژگی‌های زیادی از حوزه اسپکترال مشتق می‌شوند که بسیاری از آنها مبتنی بر تفاضل و تخمین توان نویز طبق رابطه (۲-۱) می‌باشند.

برای افزایش قدرت جداکنندگی ویژگی‌ها تحت شرایط نویزی بسیاری از روش‌ها به توان نسبی گفتار روی نویز تخمینی در باندهای فرکانسی متفاوت از اسپکتروم توجه دارند [۶]. این روش به صورت مؤثری معادل با نسبت سیگنال به نویز زیر باند سیگنال گفتار است. در حقیقت حتی تحت شرایطی که نویز بسیار زیاد باشد و نسبت سیگنال به نویز ۰ دسی‌بل باشد، می‌توان از روی اسپکتروم گفتار مشاهده کرد که هارمونیک‌های سیگنال گفتار هنوز در برخی از باندهای فرکانسی از نویز قابل تشخیص هستند.

۲-۳-۳ ویژگی‌های مبتنی بر کپسترال[۳]

دسته دیگر ویژگی‌ها از یک مجموعه از تکنیک‌های غیرخطی که به عنوان آنالیز کپستروم[۴] شناخته می‌شوند، استفاده می‌کند. توان کپستروم به صورت رابطه (۲-۷) تعریف می‌شود [۱۳]:

$$c_k = |DFT(\log |X_k|^2)|^2 \qquad (۲-۷)$$

[۱] spectrum
[۲] Fast Fourier Transform (FFT)
[۳] cepstral
[۴] cepstrum

از توان کپستروم در آنالیز توان اسپکتروم سیگنال گفتار استفاده می‌شود. بنابراین ویژگی‌های کپسترال به صورت گسترده در تشخیص گفتار کاربرد دارند. در یک سیستم آشکارساز فعالیت گفتاری قله‌های کپسترال می‌توانند برای تشخیص فرکانس اساسی سیگنال گفتار و فرکانس گام[۱] استفاده شوند [۱۳].

برخی از محققان از ویژگی ضرایب کپسترال فرکانس مل[۲] به عنوان ویژگی ورودی به یک کلاسه‌بند برای تشخیص گفتار و غیرگفتار استفاده می‌کنند [۱۵, ۱۶]. به عنوان مثال در [۱۵] از یک بردار ویژگی ۳۹ بعدی و مدل مخفی مارکوف استفاده شده است. این بردار ویژگی شامل مقادیر لگاریتم انرژی، ۱۲ ضریب کپسترال فرکانس مل و مشتقات مرتبه اول و دوم این ضرایب می‌باشد. در حالی که در [۱۶] استفاده از مشتق ضرایب کپسترال فرکانس مل به تنهایی برای آشکارساز فعالیت گفتاری پیشنهاد داده شده است. دلتا کپستروم به عنوان مشتق مرتبه اول دنباله کپسترال تعریف می‌شود و گاهی اوقات به عنوان یک ویژگی پویا مورد ارجاع قرار می‌گیرد تا بتواند تغییرات پویای بین فریم‌های کپسترال را دنبال کند. دلتا کپستروم با استفاده از رابطه (۲-۸) تعریف می‌شود [۱۳]:

$$\Delta c_k = \sum_{i=1}^{M} k(c_{k+i} - c_{k-i}) \Big/ \left(2\sum_{i=1}^{M} i^2\right) \qquad (۲-۸)$$

در این رابطه یک پنجره دلتا به طول $2M+1$ فریم در زمان k (Δc_k) برای استخراج بردار کپستروم استفاده می‌شود. تنظیم این پارامتر مبتنی بر سایز فریم، نرخ فریم و سایر پارامترهای آن است. مقدار M بزرگتر، اطلاعات موقتی گسترده‌تری را در طول فریم‌های کپسترال پی‌درپی گرفته شده نتیجه می‌دهد.

[۱] pitch frequency

[۲] Mel Frequency Cepstral Coefficients (MFCC)

۲-۳-۴ ویژگی‌های بلند مدت

سیگنال گفتار یک سیگنال غیر ایستاست و مشخصات آماری آن با زمان تغییر می‌کند. یک انسان با نرخ میانگین تولید گفتار، تقریبا حدود ۱۰ تا ۱۵ واج[۱] را در ثانیه تولید می‌کند که هر کدام از واج‌ها توزیع اسپکترال متفاوتی دارند و این امر باعث می‌شود که احتمالات گفتار به صورت گسترده با زمان تغییر کند [۱۳]. از سوی دیگر بیشتر نویزهایی که در مکالمات روزانه با آنها مواجه هستیم ایستا هستند مانند نویز سفید یا نویز ماشینی. و یا حداقل درجات تغییر آنها در مقایسه با تغییرات گفتار بسیار کمتر است. لذا با توجه به درجه غیرایستا بودن سیگنال گفتار، آنالیز حول یک پنجره طولانی‌تر می‌تواند برای تشخیص گفتار از نویز مفید باشد.

برای آشکارساز فعالیت گفتاری، بیشتر تکنیک‌ها با استفاده از گسترش پنجره‌های پردازش [۶] یا انجام پردازش روی ویژگی‌های استخراج شده از مجموعه فریم‌های پیوسته [۱۵, ۱۷] از اطلاعات زمان طولانی استفاده کرده‌اند. همچنین در [۱۸] نشان داده شده است که استفاده از اطلاعات بلند مدت طیف سیگنال گفتار در بهبود عملکرد آشکارساز فعالیت گفتاری مؤثر است.

۲-۴ قوانین تصمیم‌گیری

پس از اینکه مجموعه‌ای از بردارهای ویژگی از سیگنال اصلی استخراج شدند مرحله بعدی تشخیص گروه‌ها یا دسته‌بندی فریم‌های گفتار و غیرگفتار است. در یک مسأله کلاسه‌بندی تشخیص الگو مانند یک سیستم آشکارساز فعالیت گفتاری، یک قانون تصمیم توسط تعریف مجموعه‌ای از مرزهای تصمیم‌گیری به دست می‌آید که فضای ویژگی را به دو دسته مختلف تقسیم‌بندی می‌کند. هر دسته به یکی از کلاس‌ها اختصاص دارد.

[۱] phoneme

در دهه‌های اخیر روش‌های زیادی در مرحله تصمیم‌گیری یک سیستم آشکارساز فعالیت گفتاری در متون علمی پیشنهاد شده‌اند که می‌توانند به گروه‌های زیر تقسیم‌بندی شوند:

- روش آستانه‌گذاری
- روش‌های مدل‌سازی آماری
- روش‌های شناسایی آماری الگو و یادگیری ماشین

در ادامه به مروری بر این روش‌ها پرداخته می‌شود.

۲-۴-۱ روش آستانه‌گذاری

آستانه‌گذاری ساده‌ترین روش برای تعیین مرزهای تصمیم‌گیری است که در آن از یک خط یا مجموعه‌ای از خطوط برای تعیین نواحی هر کلاس در فضای ویژگی استفاده می‌شود. اگر ویژگی استخراج شده از فریم k ام، یک اسکالر مانند $y_k \in \mathbb{R}$ باشد، یک آستانه η (یک عدد اسکالر) برای تقسیم‌بندی فضای اعداد حقیقی به دو بخش استفاده می‌شود. اگر در یک فریم $y_k \geq \eta$ باشد، فریم به کلاس گفتار و در غیر این صورت به کلاس غیرگفتار اختصاص می‌یابد [۱۷] در حالتی که هر یک از ویژگی‌ها به صورت یک بردار $y_k \in \mathbb{R}^N$ باشند، یک بردار از آستانه‌ها می‌تواند برای تقسیم کردن هر بعد فضای $\mathbb{R}^N$ مشابه روش قبل استفاده شود. رابطه (۲-۹) این مطلب را به خوبی نشان می‌دهد.

$$y_k \underset{noise}{\overset{speech}{\gtrless}} \eta \qquad (۹-۲)$$

در برخی از متون بیش از یک آستانه به کار می‌رود که هر دوی آنها برای تشخیص فعال‌سازی و غیرفعال‌سازی انتقالات می‌باشند [۱۹]. یک آستانه η_0 مقادیر پارامتر غیرگفتار ماکزیمم را نشان می‌دهد و انتقالات غیرگفتار به گفتار را تشخیص می‌دهد. در حالی که آستانه دیگر η_1 مقادیر گفتار مینیمم را نشان می‌دهد و انتقالات گفتار به غیرگفتار را تشخیص می‌دهد. سپس طول بین این دو انتقال به عنوان گفتار تخصیص می‌یابد.

آستانه‌ها اغلب از روی مشاهده و بر مبنای توزیع ویژگی‌های استخراج شده در مجموعه داده‌های آموزشی [۲۰] از قبل تخمین زده می‌شوند. هنگامی که شرایط نویز در حال تغییر است یک مقدار برای آستانه به خوبی عمل نخواهد کرد [۱۷]. بنابراین نیاز است که در اینگونه شرایط آستانه‌ها به صورت تطبیقی تعیین شوند. در [۶, ۲۱] یک منحنی کالیبراسیون خطی برای تنظیم آستانه به کار گرفته شده است که به عنوان تابعی از انرژی سیگنال E است. مقادیر انرژی بین دو مقدار E_0 و E_1 که به ترتیب بیشترین و کمترین مقدار انرژی در داده‌های آموزشی هستند، تغییر می‌کنند. اگر آستانه‌های تخمین زده شده در این دو شرط نهایی به ترتیب η_0 و η_1 باشند، آستانه برای یک انرژی معین E از رابطه (۲-۱۰) به دست می‌آید:

$$\eta = \begin{cases} \eta_0 & E \le E_0 \\ \dfrac{\eta_0 - \eta_1}{E_0 - E_1} E + \eta_0 - \dfrac{\eta_0 - \eta_1}{1 - E_1/E_0} & E_0 < E < E_1 \\ \eta_1 & E \ge E_1 \end{cases} \qquad (۲-۱۰)$$

در [۲۲] از یک فاکتور یادگیری α برای کنترل نرخ بروزرسانی آستانه استفاده می‌شود. برای یک پارامتر نویز داده شده n_k در فریم k ام، مقدار آستانه با استفاده از رابطه (۲-۱۱) بروزرسانی می‌شود:

$$\hat{\eta} = \alpha\eta + (1-\alpha)n_k \qquad (۲-۱۱)$$

در رابطه (۲-۱۱) مقادیر α کمتر باعث می‌شوند که مقدار آستانه به نوسانات نویز بسیار حساس باشد. این مورد در حالتی که نویز پس زمینه به صورت ثابت تغییر می‌کند می‌تواند مفید باشد. از طرف دیگر مقادیر α بزرگتر باعث می‌شود که آستانه نسبت به تغییرات بسیار محافظه کار باشد و در شرایطی که تغییرات زمینه کم است می‌تواند مناسب باشد. در [۱۷] یک نسخه تغییر یافته از رابطه (۲-۱۱) پیشنهاد شده است:

$$\hat{\eta} = \alpha \min(\mathcal{Y}) + (1-\alpha)\max(\mathcal{N}) \qquad (۲-۱۲)$$

در رابطه (۲-۱۲) آستانه بر مبنای جدیدترین مقادیر نویز و گفتار استخراج شده از سیگنال صوتی که به ترتیب در $\mathcal{N}$ و $\mathcal{Y}$ ذخیره می‌شوند، بروزرسانی می‌شود. فاکتور یادگیری نیز مانند قبل برای کنترل

نرخ یادگیری به کار می‌رود. این روش از این حقیقت ناشی شده است که مقادیر ویژگی گفتار و نویز دارای یک ناحیه همپوشانی معین می‌باشند که به عنوان مقادیر گفتار مینیمم و ماکزیمم علامت‌گذاری شده‌اند. یعنی فرض می‌شود که مقادیر گفتار بالاتر از مقادیر نویز هستند.

اگر ویژگی‌هایی که در مرحله قبل استخرج شدند دارای قدرت جداکنندگی خوبی باشند و فضای ویژگی آنها به صورت خطی جداپذیر باشد، استفاده از آستانه‌ها می‌تواند مناسب باشد. گرچه هنگامی که سیگنال گفتار توسط نویز معیوب شده باشد، میزان جداپذیری خطی آن کاهش می‌یابد. برای برخی از درجه‌های معین نویز، مرزهای تصمیم‌گیری خطی دیگر نمی‌توانند نواحی کلاس‌ها را به خوبی از هم جدا کنند. بنابراین در چنین شرایطی لازم است که از مدل‌های مبتنی بر آمار و روش‌های مبتنی بر شناسایی آماری الگو و یادگیری ماشین استفاده کرد.

۲-۴-۲ روش‌های مدل‌سازی آماری

این دسته از روش‌ها، مدل‌های آماری را برای تفکیک گفتار از غیرگفتار به کار می‌گیرند. یک قانون تصمیم‌گیری آماری از تست نسبت احتمال تعمیم یافته مشتق می‌شود که برای اولین بار در [۲۳] پیشنهاد شده است. در این روش گفتار نویزی و نویز پس‌زمینه، به صورت گوسین‌های تصادفی مستقل فرض می‌شوند. بنابراین ضرایب تبدیل فوریه گسسته[۱] آنها می‌توانند به صورت متغیرهای تصادفی گوسی مدل شوند که از یکدیگر مستقل هستند. در این حالت دو فرضیه H_0 و H_1 که به ترتیب مربوط به نویز طبق رابطه (۲-۱۳) و گفتار نویزی طبق رابطه (۲-۱۴) می‌باشند عبارتند از:

$$H_0 : X = N \quad (۱۳-۲)$$

$$H_1 : X = N + S \quad (۱۴-۲)$$

[۱] Discrete Fourier Transform (DFT)

در این رابطه S، N و X به ترتیب بردارهای ضرایب تبدیل فوریه گسسته L بعدی گفتار، نویز و گفتار نویزی هستند. عناصر k ام آنها به ترتیب به صورت S_k، N_k و X_k می‌باشند. توابع چگالی احتمال شرطی روی این فرضیه‌ها از روابط (۲-۱۵) و (۲-۱۶) به دست می‌آیند:

$$p(X|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \, exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\} \tag{۱۵-۲}$$

$$p(X|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \, exp\left\{-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\right\} \tag{۱۶-۲}$$

در این روابط $\lambda_S(k)$, $\lambda_N(k)$ به ترتیب واریانس‌های N_k و S_k هستند. نسبت احتمال برای k امین باند فرکانسی از رابطه (۲-۱۷) به دست می‌آید:

$$\Lambda_k = \frac{p(X|H_1)}{p(X|H_0)} = \frac{1}{1+\xi_k} \, exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{۱۷-۲}$$

در رابطه فوق $\xi_k = \lambda_S(k)/\lambda_N(k)$ و $\gamma_k = |X_k|^2/\lambda_N(k)$ به ترتیب نسبت‌های سیگنال به نویز استقرایی و قیاسی می‌باشند. قانون تصمیم‌گیری از میانگین هندسی نسبت‌های احتمال در باندهای فرکانسی مجزا از رابطه (۲-۱۸) به دست می‌آید:

$$\log\Lambda = \frac{1}{L}\sum_{k=0}^{L-1} \log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{۱۸-۲}$$

به صورت قراردادی توزیع گفتار خالص و اسپکتروم نویزی توسط چگالی احتمال‌های گوسی طبق رابطه (۲-۱۹) مدل می‌شوند [۲۴].

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \, exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{۱۹-۲}$$

به منظور بهبود کارایی روش‌های مدل‌سازی آماری آشکارساز فعالیت گفتاری، محققان سعی کرده‌اند مدلی را پیدا کنند که با مدل توزیع گفتار سازگاری بیشتری داشته باشد. در [۲۵] عنوان شده است که

توزیع‌های لاپلاس[۱] و گاما[۲] که طبق روابط (۲-۲۰) و (۲-۲۱) می‌باشند، می‌توانند به ترتیب مدل بهتری را برای ضرایب تبدیل فوریه گسسته گفتار خالص و نویز ارائه دهند.

$$l(x) = \frac{1}{2a}\, exp\left(-\frac{|x|}{a}\right) \qquad (۲-۲۰)$$

$$\gamma(x) = |x|^h\, exp\left(-\frac{|x|}{a}\right) \frac{1}{2h!\ a^{h+1}} \qquad (۲-۲۱)$$

همچنین به جای استفاده از توزیع‌های مختلف برای گفتار خالص و نویز، در [۲۶] از توزیع لاپلاس پیچیده و در [۲۷] از توزیع گوسی تعمیم یافته[۳] طبق رابطه (۲-۲۲) برای مدل‌سازی ضرایب تبدیل فوریه گسسته گفتار نویزی به صورت مستقیم استفاده شده است که در هر دو حالت نتایج بهتری حاصل شده است.

$$G(x) = \frac{\nu\alpha(\nu)}{2\sigma\Gamma(1/\nu)}\, exp\left(-\left[\alpha(\nu)\frac{|x|}{\sigma}\right]^{\nu}\right) \qquad (۲-۲۲)$$

نتایج نشان داده‌اند که سیگنال گفتار در طول بازه‌های فعال، توسط توزیع لاپلاس در حوزه زمان بهتر می‌تواند مدل شود. در حالی که در فریم‌های لبه قسمت‌های گفتار، توزیع گاما تأثیر بهتری دارد.

در [۲۸] فرض می‌شود که احتمالات غیرگفتار در طول یک پنجره بلندتری نسبت به گفتار، ایستا می‌باشند. برخی از روش‌های آماری از مشاهداتی استفاده می‌کنند که در آن احتمال بالایی که گفتار دارد باعث ایجاد تفاوت از غیرگفتار می‌شود. در [۲۹] احتمال غیرگفتار به صورت گوسی فرض شده و احتمال بالای گفتار برای آشکارسازی فعالیت گفتاری به کار می‌رود. هنگامی که سیگنال گفتار توسط یک نویز غیرگوسی خراب شود یا هنگامی که نواحی غیرصدادار در گفتار مشاهده شوند، کارایی به شدت پایین می‌آید. در [۳۰] روشی برای رفع این مشکل پیشنهاد شده است که در آن از ترکیب احتمالات بالا با نسبت انرژی باند پایین به باند کامل استفاده شده است.

---

[۱] Laplac Distribution

[۲] Gamma Distribution

[۳] Generalized Gaussian Distribution

۲-۴-۳ روش‌های شناسایی آماری الگو و یادگیری ماشین

اخیرا علاقه زیادی در به کارگیری روش‌های یادگیری ماشین به عنوان قوانین تصمیم‌گیری برای آشکارساز فعالیت گفتاری به وجود آمده است [۳۱, ۳۲]. روش‌های بسیاری در سیستم آشکارساز فعالیت گفتاری استفاده شده است مانند ماشین بردار پشتیبان[۱]، کلاسه‌بندی حاشیه ماکزیمم[۲]، شبکه‌های عصبی[۳]، آنالیز جداکنندگی خطی، الگوریتم k-means و الگوریتم‌های ژنیتک. در این قسمت به مرور برخی از این روش‌ها برای آشکارساز فعالیت گفتاری می‌پردازیم.

یکی از اولین کاربردهای یک روش شناسایی آماری الگو برای کلاسه‌بندی گفتار و غیرگفتار و سکوت، در [۳۳] توضیح داده شده است. در این روش یک بردار ویژگی ۵ بعدی از سیگنال گفتار استخراج شده است که این ۵ بعد شامل این ویژگی‌ها می‌باشند : نرخ عبور از صفر، انرژی زمان کوتاه، همبستگی بین نمونه‌های گفتار مجاور، اولین ضریب پیش‌بینی از یک آنالیز کدینگ پیش‌بینی خطی ۱۲ قطبی و انرژی در خطای پیش‌بینی. سپس از یک مجموعه آموزشی که به صورت دستی برچسب گذاری شده است، پارامترهای توزیع احتمال گوسی در این فضای ویژگی برای هر کلاس تخمین زده می‌شوند. قسمت گفتار داده شده به کلاسی تخصیص داده می‌شود که کمترین خطای احتمال را دارد.

در استاندارد G.729 (مراجعه به قسمت ۲-۶-۱) از یک جداکننده خطی تکه‌ای مبتنی بر فرکانس‌های اسپکترال خطی، انرژی باند بالا و باند پایین و نرخ عبور از صفر به عنوان یکی از روش‌های شناسایی آماری الگو استفاده می‌شود.

در [۳۲, ۳۴] از ماشین بردار پشتیبان برای سیستم آشکارساز فعالیت گفتاری روی مجموعه ویژگی‌های متفاوتی استفاده شده است. در هر دو حالت کارایی سیستم آشکارساز فعالیت گفتاری در مقایسه با سایر روش‌های تصمیم‌گیری روی مجموعه ویژگی‌های یکسان، بهبود می‌یابد. یکی از نقاط ضعف این روش

---

[۱] Support Vector Machine (SVM)

[۲] Maximum Marginal Clustering

[۳] Neural Network

حساسیت آن به نویز می‌باشد. به صورت تئوری اگر داده‌های آموزشی برای همه شرایط نویزی در دسترس باشند و شرایط تست با آن تنظیم شده باشد، ماشین بردار پشتیبان می تواند خطای کلاسه‌بندی را حداقل نماید. اما در عمل در شرایط نویزی این امر غیرممکن است و کارایی روش را پایین می‌آورد [۳۵].

تکنیک کلاستربندی حاشیه ماکزیمم [۳۶] ایده ماشین بردار پشتیبان را گسترش می‌دهد. در این روش وابستگی به برچسب‌های آموزشی با پیدا کردن ابرصفحه حاشیه ماکزیمم در فضای ویژگی و همچنین بردار برچسب بهینه‌ای که حاشیه بین همه بردارهای برچسب ممکن را ماکزیمم کند، حذف می‌شود.

در [۳۷] یک آنالیز جداکنندگی خطی پیشنهاد شده است که هدف آن پیدا کردن یک ترکیب خطی از ویژگی‌های بردار ویژگی برای توصیف بهتر سیگنال گفتار است. این روش به دو صورت عمل می‌کند: یکی فرآیند کاهش ابعاد و دیگری روش ترکیب. این روش هنگامی می‌تواند مفید باشد که چند بردار ویژگی با هم ترکیب شوند. در [۳۸] از پرسپترون چند لایه‌ای[۱] و در [۳۹] از تطبیق مبتنی بر قانون فازی استفاده شده است. در [۱۱] الگوریتم k-means برای دسته‌بندی بردارهای ویژگی به کار برده شده است. این الگوریتم فضای ویژگی را به k گروه یا کلاستر تقسیم می‌کند و هر کلاستر را توسط بردار میانگین داده‌های موجود در آن (مرکز کلاستر) نمایش می‌دهد. سایر روش‌های یادگیری ماشین از طبیعت الهام می‌گیرند مانند شبکه‌های عصبی [۴۰] و الگوریتم‌های ژنتیک [۳۱] که در مرحله تصمیم‌گیری سیستم‌های آشکارساز فعالیت گفتاری نیز به کار رفته‌اند.

۲-۵ روش بهبود و اصلاح تصمیمات

برخی از الگوریتم‌ها ممکن است در نقاط شروع و پایان مرزهای قطعه گفتار به خوبی کار نکنند. با این حال این امکان برای الگوریتم وجود دارد که قبل از اینکه تصمیم تعیین کننده فریم جاری پردازش

[۱] Multi-layer Perceptron

شود، چند فریم منتظر بماند تا به بالا یا پایین سطح آستانه برسد. مجموعه‌ای از قوانین تجربی برای بهبود تصمیمات نهایی مبتنی بر تصمیمات قبلی در واحدی به نام روش بهبود و اصلاح تصمیمات قرار دارند. بسیاری از الگوریتم‌ها عموما برای بهبود تصمیمات گرفته شده یک برنامه بهبود و اصلاح تصمیمات را به کار می‌گیرند. این عمل برای جلوگیری از برش دنباله‌های گفتار ضعیف و با توجه به تصمیمات قبلی انجام می‌شود. این برنامه انتقال‌های سریع از گفتار به غیرگفتار را به تاخیر انداخته و باعث کاهش خطای تشخیص می‌گردد. برنامه بر مبنای این ایده خلق شده است که با فرض وجود یک همبستگی قوی بین رویدادهای متوالی فریم‌های گفتار، از تغییرات سریع آن باید جلوگیری کرد.

به عنوان مثال روش بهبود و اصلاح تصمیمات در G.729 شامل ۴ مرحله زیر می‌باشد:

۱- اگر انرژی فریم بیشتر از تفاوت انرژی باند کامل باشد، یک فریم به عنوان گفتار تعیین می‌شود.

۲- اگر دو فریم قبلی نیز به عنوان گفتار تعیین شده باشند و اختلاف انرژی بین فریم‌های جاری و قبلی زیر یک آستانه ثابت باشد، فریم جاری نیز به عنوان گفتار تعیین می‌شود.

۳- اگر ۱۰ تا فریم قبلی یک فریم به عنوان غیرگفتار تعیین شده باشند و اختلاف انرژی بین فریم‌های قبلی و جاری کمتر از یک آستانه ثابت باشد، فریم جاری به عنوان غیرگفتار تعیین می‌شود.

۴- یک تصمیم گفتار در صورتی به غیرگفتار تغییر می‌کند که انرژی فریم جاری زیر سطح نویز و در حد یک آستانه ثابت باشد.

۲-۶ روش‌های استاندارد

الگوریتم‌های استاندارد آشکارسازی فعالیت گفتاری شامل G.729 [۴۱] (توصیه شده توسط واحد بین‌المللی ارتباط از دور) و AMR[۱] نوع ۱ و ۲ [۴۲] (توصیه شده توسط موسسه استانداردهای اروپایی ارتباط از دور) می‌باشند که به اختصار در قسمت‌های بعدی توضیح داده شده‌اند.

۲-۶-۱ استاندارد G.729

در این الگوریتم [۴۱] با استفاده از ۴ پارامتر زیر، هر ۱۰ میلی ثانیه یک تصمیم گرفته می‌شود:

۱- اعوجاج اسپکترال[۲] طبق رابطه (۲-۲۳)

$$\Delta LSF = \sum_{i=1}^{p} (LSF_i - \overline{LSF_i})^2 \qquad (۲۳-۲)$$

۲- تفاوت انرژی باند کامل طبق رابطه (۲-۲۴)

$$\Delta E_f = \overline{E_f} - E_f \qquad (۲۴-۲)$$

۳- تفاوت انرژی باند پایین طبق رابطه (۲-۲۵)

$$\Delta E_l = \overline{E_l} - E_l \qquad (۲۵-۲)$$

۴- تفاوت عبور از صفر طبق رابطه (۲-۲۶)

$$\Delta ZC = \overline{ZC} - ZC \qquad (۲۶-۲)$$

در روابط فوق LSF_i ، E_f ، E_l و ZC به ترتیب فرکانس‌های اسپکترال خط[۳]، انرژی فریم جاری، انرژی باند پایین فریم جاری و نرخ عبور از صفر می‌باشند. علاوه بر آن میانگین پارامترهای متناظر در نویز پس

[۱] Adaptive Multirate (AMR)

[۲] Spectral Distortion

[۳] Line Spectral Frequencies (LSF)

زمینه با علامت ¯ روی متغیر نمایش داده شده‌اند. LSF از دهمین ضریب پیش‌بینی خطی[1] متناظر به دست می‌آید.

این ۴ پارامتر از یک فریم گفتار استخراج می‌شوند و به عنوان یک نقطه در یک فضای ۴ بعدی می‌باشند. نواحی گفتار و غیرگفتار به عنوان نقاطی در این فضا، خوشه‌های جداگانه‌ای را تشکیل می‌دهند (همپوشانی در بین خوشه‌ها وجود دارد). ۱۴ تا ابرصفحه خطی تکه‌ای برای دسته‌بندی یک فریم به کلاس گفتار یا غیرگفتار به کار می‌روند. پارامترهای این ابرصفحه‌ها به صورت تجربی تعیین می‌گردند. تصمیمات گرفته شده توسط توابع مجزاساز با استفاده از یک روش بهبود و اصلاح تصمیمات چهار مرحله‌ای بهبود می‌یابند.

۲-۶-۲ استاندارد AMR نوع ۱ و ۲

در [۴۲] از روش‌های مبتنی بر شکل اسپکترال بهره گرفته شده است. AMR نوع ۱ و ۲ به ترتیب از انرژی‌های زیرباند و کانال، برای تصمیم‌گیری یک سیستم آشکارساز فعالیت گفتاری استفاده می‌کنند. بخش تصمیم‌گیری این سیستم در ارتباط با یک بخش بهبود و اصلاح تصمیمات گسترده می‌باشد. تصمیمات در این سیستم هر ۲۰ میلی‌ثانیه یک بار گرفته می‌شوند.

AMR نوع ۱ انرژی سیگنال گفتار ورودی را از ۹ زیرباند فرکانسی محاسبه می‌کند. پهنای باند این زیرباندها در واقع غیریکنواخت هستند، زیرباندهای با فرکانس پایین‌تر، پهنای باند کمتری دارند. سپس نسبت سیگنال به نویز برای زیرباندها تخمین زده می‌شود. برای این منظور تخمینی از انرژی نویز پس‌زمینه لازم است. برای انجام آن از یک مدل خودکاهنده مرتبه اول استفاده می‌شود. سپس تصمیم‌گیری بر اساس مقایسه نسبت سیگنال به نویز زیرباند تخمینی با یک آستانه سازگار صورت می‌گیرد. و در نهایت تصمیم گرفته شده با استفاده از یک روش اصلاح تصمیمات، بهبود می‌یابد.

[1] Linear Predictive Coefficients (LPC)

AMR نوع ۲ نیز نسبت سیگنال به نویز زیرباند را تخمین می‌زند. تفاوت آن در تعداد زیرباندها (۱۶) و مقیاس غیرخطی به کار رفته در گروه‌بندی باند است. تخمین انرژی نویز پس‌زمینه در هر زیرباند شبیه به مورد AMR نوع ۱، با استفاده از مدل خودکاهنده مرتبه اول می‌باشد. در AMR نوع ۲ نیز تصمیمات با استفاده از یک آستانه سازگار گرفته می‌شوند. نویز غیرایستا توسط محاسبه واریانس نسبت سیگنال به نویز آنی که برای هر فریم تخمین زده می‌شود، پشتیبانی می‌گردد. سپس تصمیمات گرفته شده توسط یک روش بهبود و اصلاح تصمیمات هموار می‌شوند.

۲-۷ ارزیابی کارایی ویژگی‌ها

در متون علمی روش‌های بیشماری برای آشکارساز فعالیت گفتاری پیشنهاد شده‌اند. لذا واضح است که استفاده از یک استراتژی مناسب برای ارزیابی آنها بسیار مهم می‌باشد. از آنجا که یک سیستم آشکارساز فعالیت گفتاری دارای چندین واحد است، یک روش ارزیابی منصفانه باید به مقایسه هر واحد به صورت مستقل توجه نماید. به عنوان مثال مقایسه دو الگوریتم آشکارساز فعالیت گفتاری که از ویژگی‌های مختلف و قوانین تصمیم‌گیری متفاوت استفاده می‌کنند، نمی‌تواند مشخص نماید که بهبود کارایی کلی توسط ویژگی جدید حاصل شده یا به دلیل استفاده از قانون جدید بوده است. لذا کارایی قدرت جداکنندگی ویژگی‌ها به صورت مستقل از کارایی واحد تصمیم‌گیری و واحدهای دیگر مقایسه می‌شوند.

یک امر ضروری و کلیدی دیگر برای الگوریتم‌های ارزیابی آشکارساز فعالیت گفتاری، محاسبه کارایی تحت انواع مختلف نویز و شرایط مختلف نسبت سیگنال به نویز است. این امر باعث می‌شود ناحیه‌ای که در آن الگوریتم آشکارساز فعالیت گفتاری بهترین عملکرد را دارد مشخص شود. روش نمایش این نتایج به صورت متراکم نیستند. به عنوان مثال یک مقایسه نتایج روی N نوع مختلف نویز در K سطح مختلف نسبت سیگنال به نویز ، $N \times K$ تا تنظیمات مختلف را نتیجه می‌دهد.

۲-۷-۱ آماده‌سازی داده‌ها

پایگاه داده‌های بسیاری در متون علمی برای ارزیابی آشکارساز فعالیت گفتاری استفاده شده‌اند. انتخاب پایگاه داده گاهی اوقات به زبان صحبت کردن نویسنده‌ها و ماهیت پروژه تحقیقاتی که آنها روی آن کار می‌کنند وابسته است. به عنوان مثال SpeechDat-Car یک پایگاه داده به زبان اسپانیایی [۴۳] و AURORA-2J یک پایگاه داده به زبان ژاپنی [۴۴] است. برای زبان انگلیسی پایگاه داده‌های TIMIT [۴۵] و AURORA-2 [۴۶] بیشترین کاربردها را دارند. از بین این دو، پایگاه داده TIMIT برای ارزیابی کارایی آشکارساز فعالیت گفتاری مناسب‌تر می‌باشد زیرا در آن آوانویسی دستی در سطح کلمه و واج انجام شده است. با این فرض که آوانویسی دستی آن از دقت قابل قبولی برخوردار باشد، این پایگاه داده در ارزیابی توزیع ویژگی‌ها بسیار مفید است. به عنوان مثال آوانویسی در سطح کلمه می‌تواند برای ترسیم توزیع گفتار در مقابل غیرگفتار استفاده شود. همچنین آوانویسی در سطح واج برای گفتار صدادار در مقابل گفتار بی‌صدا به علاوه نویز کاربرد دارد.

به طور معمول برای مقایسه قدرت جداکنندگی ویژگی‌های آشکارساز فعالیت گفتاری، سیگنال‌های نویز از پایگاه داده NOISEX92 [۴۷] انتخاب می‌شوند. نویزهای انتخاب شده به صورت مصنوعی به هر سیگنال گفتاری در نسبت‌های مختلف سیگنال به نویز اضافه می‌گردند. این فرآیند بدین طریق صورت می‌گیرد:

- محاسبه میانگین توان فریم‌های سیگنال گفتار نرمال‌سازی شده P_x
- محاسبه میانگین توان فریم‌های نویز نرمال‌سازی شده P_n
- محاسبه بهره مطلوب برای گفتار طبق رابطه (۲-۲۷) که در آن توان گفتار مطلوب از رابطه (۲-۲۸) به دست می‌آید:

$$g = \sqrt{\frac{\hat{P}_x}{P_x}} \qquad (۲-۲۷)$$

$$\hat{P}_x = 10^{SNR/10} P_n \qquad (۲۸-۲)$$

- اضافه کردن نویز به سیگنال خالص طبق رابطه (۲-۲۹)

$$\hat{x} = gx + n \qquad (۲۹-۲)$$

توجه به این نکته حائز اهمیت است که یک سطح نویز ثابت به همه سیگنال‌های گفتار و به صورت یکسان اضافه شود. این امر بدین صورت میسر خواهد شد که از یک مقدار بهره g یکسان استفاده نموده و به جای افزودن به سیگنال نویز به سیگنال گفتار افزوده شود. این امر باعث می‌شود که توزیع نویز برای تمام ویژگی‌های حساس به بهره در همه سیگنال‌های گفتاری یکسان باشد.

۲-۷-۲ استانداردهای ارزیابی

به عنوان یک مسأله تشخیص گفتار، یک الگوریتم آشکارساز فعالیت گفتاری با دو نوع خطا می‌تواند مواجه شود :

۱- هنگامی که یک فریم گفتار به عنوان غیرگفتار تشخیص داده شود.

۲- هنگامی که یک فریم غیرگفتار به عنوان گفتار تشخیص داده شود.

همانند یک کلاسه‌بند دودویی، افزایش کارایی یکی از خطاها منجر به افزایش هزینه دیگری خواهد شد. استانداردهای استفاده شده برای محاسبه این خطاها بدین صورت تعریف می‌شوند [۴۸]:

P_{cs} : درصد تشخیص گفتار درست که طبق رابطه (۲-۳۰) محاسبه می‌شود.

رابطه (۲-۳۰)
$$P_{cs} = \frac{\text{تعداد فریم‌های گفتار تشخیص داده شده توسط الگوریتم} \times 100}{\text{تعداد کل فریم‌های گفتار}}$$

P_{cn} : درصد تشخیص نویز درست که طبق رابطه (۲-۳۱) محاسبه می‌شود.

رابطه (۲-۳۱)
$$P_{cn} = \frac{\text{تعداد فریم‌های غیرگفتار تشخیص داده شده توسط الگوریتم} \times 100}{\text{تعداد کل فریم‌های غیرگفتار}}$$

P_f : درصد خطا که طبق رابطه (۲-۳۲) محاسبه می‌شود.

رابطه (۲-۳۲)

$$P_f = \frac{\text{تعداد فریم‌های اشتباه تشخیص داده شده توسط الگوریتم}}{\text{تعداد کل فریم‌ها}} \times 100$$

۲-۸ جمع‌بندی

سیستم آشکارساز فعالیت گفتاری یک مؤلفه مهم در سیستم‌های پردازش گفتار مقاوم است. طراحی این سیستم شامل چهار مرحله می‌باشد. در مرحله اول نویز از سیگنال گفتار حذف می‌شود. در مرحله بعد ویژگی‌هایی که بتوانند قسمت‌های گفتار و غیرگفتار را از یکدیگر جدا کنند، استخراج می‌شوند. پس از آن در مرحله تصمیم‌گیری فریم‌های گفتار و غیرگفتار از یکدیگر تشخیص داده می‌شوند. سپس برای کاهش اشتباهات در تصمیم‌گیری از روشی برای بهبود و اصلاح تصمیمات استفاده می‌شود. برای ارزیابی کارایی یک روش پیشنهادی باید از استانداردهای معینی استفاده کرد.

فصل ۳

روش‌های استخراج ویژگی مورد استفاده

۳-۱ پیشگفتار

این فصل به بررسی روش‌های استخراج ویژگی مورد استفاده در این پایان‌نامه می‌پردازد و مراحل استخراج هر یک از ویژگی‌ها را شرح می‌دهد. قسمت ۳-۲ ویژگی انرژی را توصیف می‌کند. قسمت ۳-۳ به بیان مراحل استخراج ویژگی ضرایب کپسترال فرکانس مل اختصاص دارد. در قسمت ۳-۴ ویژگی ضرایب کپسترال فرکانس مل ریشه[۱] توضیح داده می‌شود. قسمت ۳-۵ مراحل استخراج ویژگی ضرایب پیش‌بینی ادراکی خطی[۲] را بیان می‌کند. قسمت ۳-۶ ضرایب کپسترال فرکانس بارک[۳] را توصیف می‌کند. در قسمت ۳-۷ ویژگی ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده[۴] بررسی می‌شوند. قسمت ۳-۸ به بیان توزیع ویگنر ویل[۵] با استفاده از تبدیل هارتلی می‌پردازد.

۳-۲ انرژی

ویژگی انرژی نشان می‌دهد که دامنه سیگنال گفتار به صورت محسوس با زمان تغییر می‌کند. معمولا در یک سیگنال گفتار دامنه قسمت‌های بی‌صدا کمتر از دامنه قسمت‌های صدادار است. انرژی زمان کوتاه سیگنال گفتار نمایش مناسبی را برای انعکاس تغییرات دامنه سیگنال فراهم می‌آورد [۴۹]. به طور کلی انرژی زمان کوتاه را می‌توان به صورت رابطه (۳-۱) نمایش داد. در این رابطه N تعداد نمونه‌های موجود در هر فریم و $S_i(n)$ ، n امین نمونه از i امین فریم می‌باشد.

$$E_i = \frac{1}{N}\sum_{n=1}^{N} S_i^2(n) \qquad (۳-۱)$$

[۱] Root MFCC (RMFCC)
[۲] Perceptual Linear Prediction (PLP)
[۳] Bark Frequency Cepstral Coefficients (BFCC)
[۴] Revised PLP (RPLP)
[۵] Wigner Ville Distribution (WVD)

۳-۳ ضرایب کپسترال فرکانس مل

مهم‌ترین و رایج‌ترین روش مورد استفاده برای استخراج ویژگی‌های اسپکترال، محاسبه ضرایب کپسترال فرکانس مل است. این ویژگی یکی از عمومی‌ترین تکنیک‌های استخراج ویژگی در زمینه تشخیص گفتار است که به عنوان یک ویژگی مبتنی بر فرکانس مورد توجه بوده و بسیار دقیق‌تر از ویژگی‌های حوزه زمان می‌باشد. در این ویژگی از مقیاس مل که منطبق بر مقیاس گوش انسان است استفاده می‌شود.

مراحل استخراج ویژگی ضرایب کپسترال فرکانس مل بدین صورت می‌باشند [۵۰]:

۱- ابتدا با استفاده از یک فیلتر، سیگنال پیش‌تأکید می‌شود. تابع انتقال این فیلتر طبق رابطه (۳-۲) است.

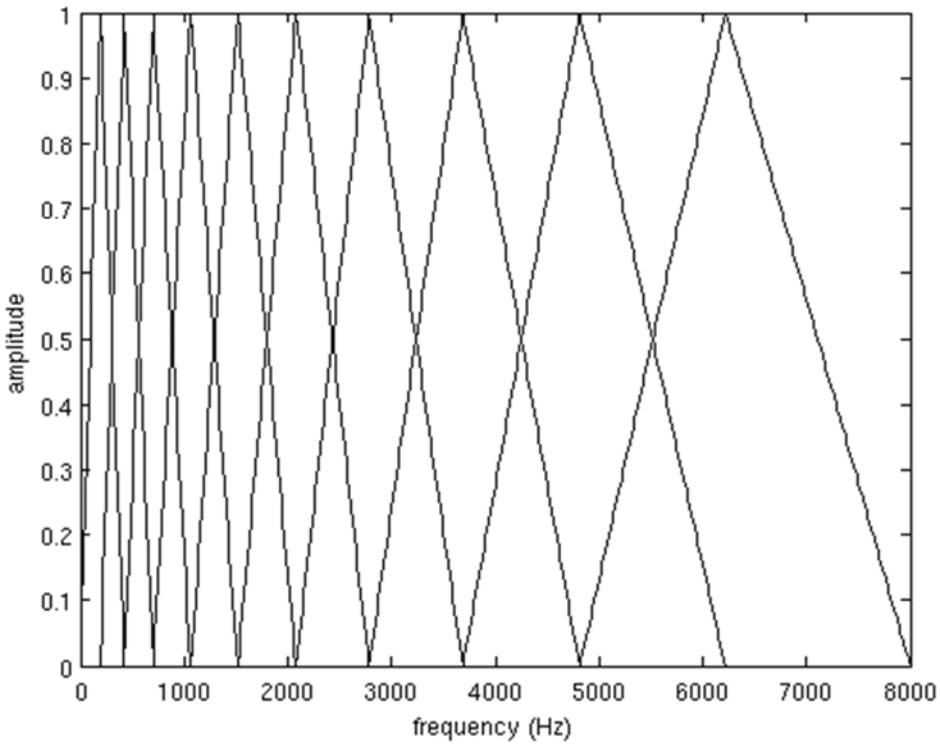
$$H(z) = 1 - \alpha z^{-1} \qquad (۳-۲)$$

۲- سیگنال پیش‌تأکید شده، به یک سری فریم‌های زمانی که تعداد دلخواهی از نمونه‌ها می‌باشند، تقسیم‌بندی می‌شود. در بیشتر سیستم‌ها برای انتقال روان از یک فریم به فریم دیگر از همپوشانی فریم‌ها استفاده می‌شود. سپس هر فریم زمانی برای حذف ناپیوستگی در لبه‌ها از یک پنجره همینگ عبور داده می‌شود.

۳- پس از فریم‌بندی و پنجره کردن، مجذور اندازه تبدیل فوریه هر فریم محاسبه می‌شود تا مؤلفه‌های فرکانسی سیگنال حوزه زمان استخراج شوند. برای سرعت بخشیدن به امر پردازش، الگوریتم تبدیل فوریه سریع به کار می‌رود.

۴- سپس فریم‌های تبدیل شده به حوزه فوریه از مل فیلترها عبور داده می‌شوند. فیلترهای مل فیلترهای مثلثی شکلی هستند که با فاصله‌های مساوی در مقیاس مل از هم قرار دارند. شکل (۳-۱) نمایش ساده‌ای از فیلترهای مل می‌باشد. رابطه (۳-۳) ارتباط بین فرکانس در مقیاس هرتز و فرکانس در مقیاس مل را نشان می‌دهد.

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



شکل (۳-۱) نمایش فیلترهای مل

۵- در مرحله بعد از خروجی مل فیلترها لگاریتم گرفته می‌شود.

۶- در مرحله آخر از تبدیل کسینوسی استفاده می‌شود.

شکل (۳-۳) مراحل استخراج ضرایب کپسترال فرکانس مل را نشان می‌دهد.

۳-۴ ضرایب کپسترال فرکانس مل ریشه

روش استخراج ویژگی ضرایب کپسترال فرکانس مل ریشه نیز مانند روش بیان شده در قسمت ۳-۳ می‌باشد با این تفاوت که در مرحله پنجم آن به جای استفاده از تابع لگاریتم از تابع ریشه استفاده می‌شود [۵۱]. یعنی خروجی مرحله قبل به توان γ می‌رسد. γ عددی بین ۰ و ۱ است. در شکل (۳-۳) مراحل استخراج ویژگی ضرایب کپسترال فرکانس مل ریشه نشان داده شده است.

۳-۵ ضرایب پیش‌بینی ادراکی خطی

مدل پیش‌بینی ادراکی خطی اولین بار توسط هرمنسکی [۵۲] ایجاد شد. پیش‌بینی ادراکی خطی، گفتار انسان را بر مبنای محتوایی از علوم روانشناسی شنوایی انسان مدل می‌کند. این ویژگی سیگنال گفتار را از طریق برخی از فرآیندهای آوایی تبدیل به یک روش ادراکی معنادار می‌کند. در این روش اطلاعات نامرتبط سیگنال گفتار دور ریخته می‌شوند و بنابراین نرخ تشخیص گفتار را بهبود می‌بخشد [۵۳]. مراحل به دست آوردن ضرایب پیش‌بینی ادراکی خطی بدین ترتیب می‌باشند:

۱- فریم‌بندی و پنجره کردن

۲- مجذور اندازه تبدیل فوریه با استفاده از الگوریتم تبدیل فوریه سریع

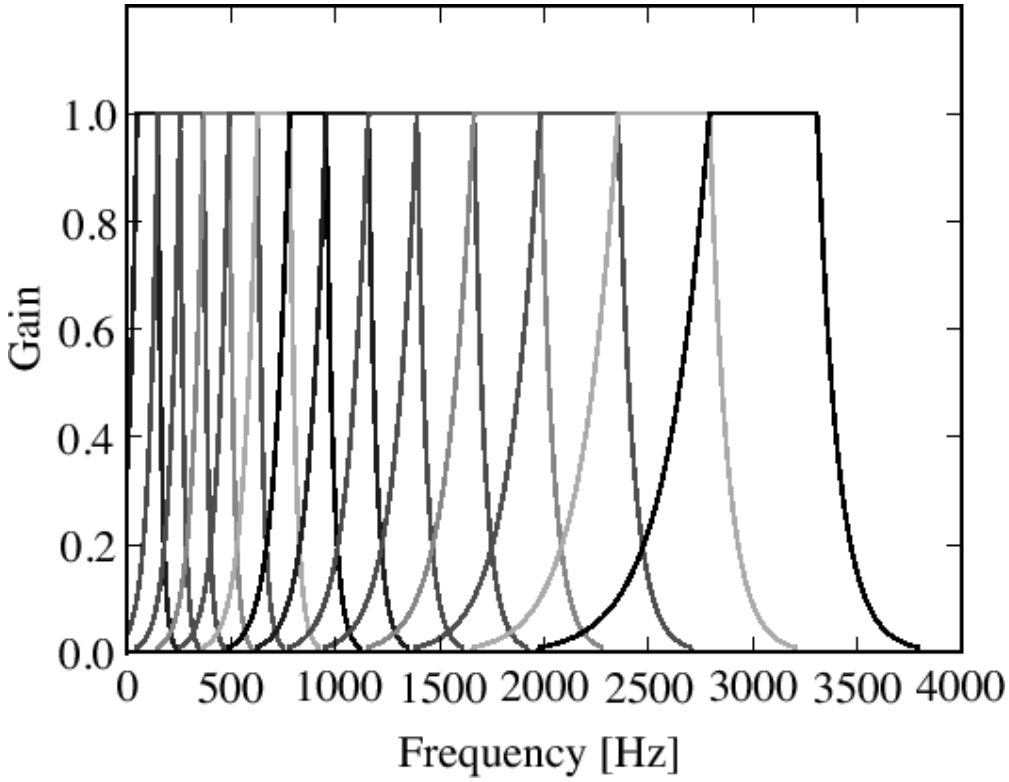
از مرحله سوم به بعد بلوک‌های مهم آنالیز پیش‌بینی ادراکی خطی آغاز می‌گردند.

۳- در این مرحله یک تبدیل فرکانسی صورت می‌گیرد که در آن فرکانس در مقیاس هرتز طبق رابطه (۳-۴) به فرکانس در مقیاس بارک تبدیل می‌شود. این تبدیل نمایش بهتری را از رزولوشن شنوایی انسان در فرکانس ایجاد می‌کند. در واقع سیگنال گفتار از یک سری فیلترهای ذوزنقه‌ای شکل با فاصله‌های مساوی در مقیاس بارک عبور می‌کند. شکل (۳-۲) نمایشی ساده از بارک فیلترها می‌باشد [۵۴ ,۵۲]

$$f_{bark} = 6\ln\left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right] \qquad (۳-۴)$$

۴- پیش‌تأکید بلندی صدای مساوی: در این مرحله مؤلفه‌های فرکانسی اسپکتروم گفتار توسط یک منحنی بلندی صدای مساوی پیش‌تأکید می‌شوند. این عمل حساسیت نامساوی شنوایی انسان در فرکانس‌های متفاوت را در سطح ۴۰ دسی‌بل تخمین می‌زند. تابع انتقال این فیلتر از رابطه (۳-۵) به دست می‌آید [۵۲, ۵۴]

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)\omega^4}{(\omega^2 + 6.3 \times 10^6)^2(\omega^2 + 0.38 \times 10^9)} \quad (۵-۳)$$



شکل (۳-۲) نمایش فیلترهای بارک

۵- تبدیل شدت بلندی صدا: در این مرحله فشرده‌سازی اسپکتروم گفتار تغییر یافته بر طبق قانون توان شنوایی انجام می‌گیرد. این عمل رابطه غیرخطی بین شدت صدا و بلندی دریافتی از آن را مدل می‌کند. رابطه این تبدیل طبق رابطه (۳-۶) می‌باشد [۵۲, ۵۴]

$$l(\omega) = E(\omega)^{0.33} \quad (۶-۳)$$

۶- مدل تمام قطب: در این مرحله، اسپکتروم توسط یک مدل تمام قطب طبق رابطه (۳-۷) و با استفاده از روش اتوکرولیشن تقریب زده می‌شود. قبل از آن باید تبدیل فوریه معکوس روی نمونه‌های اسپکتروم اعمال شود. برای مدل تمام قطب مرتبه M ام فقط $(M + 1)$ مقادیر اولیه اتوکرولیشن لازم است [۱۳].

$$H(z) = \frac{1}{1 - \sum_{i=1}^{M} \alpha_i z^{-i}} \quad (۷-۳)$$

۷- آنالیز کپستروم: آخرین مرحله استخراج ضرایب پیش‌بینی ادراکی خطی، مرحله آنالیز کپستروم است. در این مرحله ضرایب α_i به دست آمده از مدل تمام قطب توسط رابطه بازگشتی (۳-۸) به ضرایب کپسترال تبدیل می‌شوند [۱۳].

$$c_k = \alpha_k + \frac{1}{k}\sum_{i=1}^{k-1} i c_i \alpha_{k-i}\,, \quad 1 \le k \le M \tag{۳-۸}$$

شکل (۳-۴) مراحل استخراج این ضرایب را نشان می‌دهد.

۳-۶ ضرایب کپسترال فرکانس بارک

یکی دیگر از روش‌های استخراج ویژگی، ضرایب کپسترال فرکانس بارک است. این روش از ترکیب دو روش ۳-۳ و ۳-۵ ایجاد شده است. مراحل استخراج این ویژگی بدین ترتیب می‌باشد [۵۵, ۵۶]:

۱- فریم‌بندی و پنجره کردن

۲- مجذور اندازه تبدیل فوریه با استفاده از الگوریتم تبدیل فوریه سریع

۳- سه مرحله آنالیز بیان شده در ضرایب پیش‌بینی ادراکی خطی

۴- اعمال لگاریتم

۵- اعمال تبدیل کسینوسی گسسته

در شکل (۳-۴) مراحل استخراج این ضرایب دیده می‌شود.

۳-۷ ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده

ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده نوع دیگری از ترکیب دو روش ۳-۳ و ۳-۵ است. مراحل استخراج این ویژگی بدین صورت می‌باشد [۵۵, ۵۶]:

۱- پیش‌تأکید کردن

۲- فریم‌بندی و پنجره کردن

۳- مجذور اندازه تبدیل فوریه با استفاده از الگوریتم تبدیل فوریه سریع

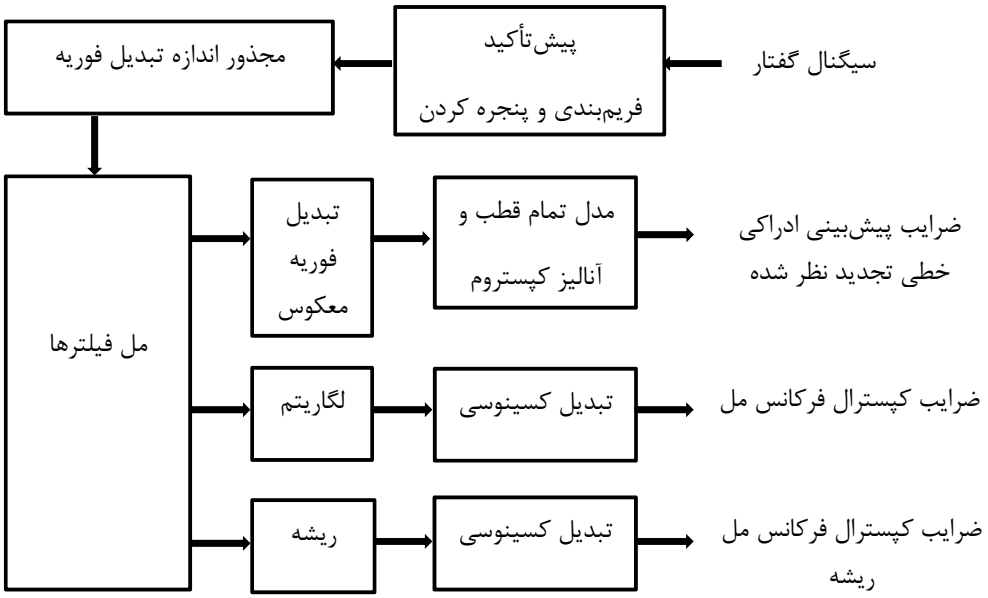
۴- فیلترهای مل

۵- معکوس تبدیل فوریه

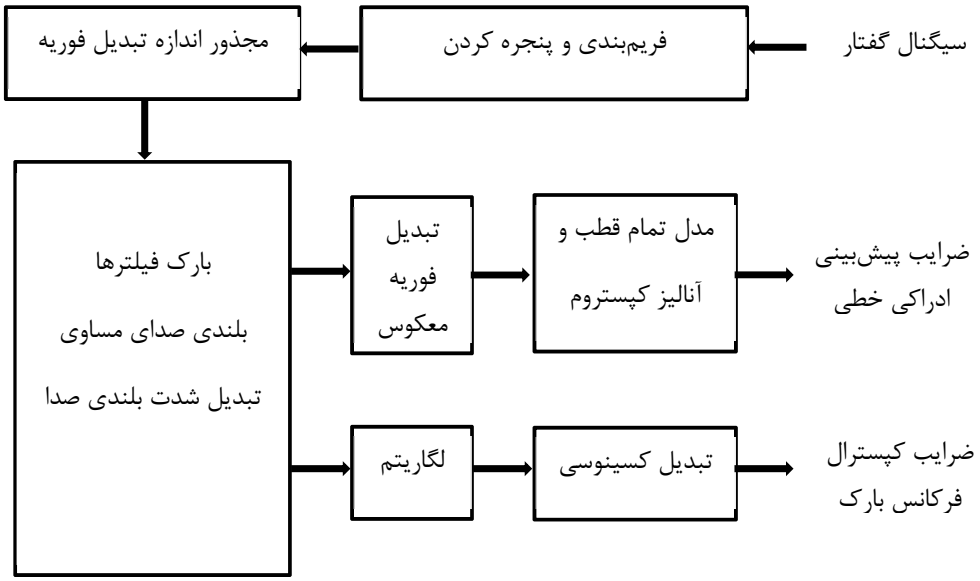
۶- مدل تمام قطب

۷- آنالیز کپستروم

شکل (۳-۳) مراحل استخراج این ویژگی را نمایش می‌دهد.



شکل (۳-۳) مراحل استخراج ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده، ضرایب کپسترال فرکانس مل [۵۵، ۵۶] و ضرایب کپسترال فرکانس مل ریشه [۵۱]



شکل (۳-۴) مراحل استخراج ضرایب پیش‌بینی ادراکی خطی و ضرایب کپسترال فرکانس بارک [۵۵، ۵۶]

۳-۸ توزیع ویگنر ویل

توزیع ویگنر ویل متداول‌ترین نمایش توزیع‌های زمانی فرکانسی است که بسیار ساده و قدرتمند می‌باشد. یک توزیع زمانی فرکانسی تبدیلی است که یک سیگنال یک بعدی را به یک فضای دوبعدی زمان و فرکانس انتقال می‌دهد. توزیع ویگنر ویل ابتدا توسط E.Wigner و درباره مکانیک کوانتوم معرفی شده است [۵۷] و سپس به صورت مستقل توسط J.Ville توسعه یافته است که در آن تبدیلات یکسان برای پردازش سیگنال و آنالیز اسپکترال را به کار برده است. یک خصوصیت مهم این توزیع این است که می‌تواند یک نمایش زمانی فرکانسی با رزولوشن بالا از سیگنال‌های غیرایستا را تامین کند. سیگنال گفتار یک سیگنال شبه پریودیک و غیرایستا است که به یک سیگنال نویز غیرایستا اضافه می‌گردد. با توجه به این خصوصیت استفاده از ابزاری مانند توزیع ویگنر ویل، یک مزیت عمده است و به عنوان انگیزه‌ای برای کاربرد آن در سیستم آشکارساز فعالیت گفتاری مطرح می‌باشد [۵۸].

برای سیگنال زمانی $x(t)$ توزیع ویگنر ویل از رابطه (۳-۹) به دست می‌آید [۷۸،۷۹]. با توجه به رابطه (۳-۹) توزیع ویگنر ویل می‌تواند به عنوان تبدیل فوریه تابع اتوکرولیشن سیگنال $x(t)$ تعریف شود.

$$WV(t,\omega) = \int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right).x^*\left(t-\frac{\tau}{2}\right)e^{-j\omega\tau}d\tau \qquad (۳-۹)$$

در رابطه (۳-۹)، t نماینده زمان، ω فرکانس زاویه‌ای و τ تاخیر زمانی است و $x\left(t+\frac{\tau}{2}\right).x^*\left(t-\frac{\tau}{2}\right)$ تابع اتوکرولیشن لحظه‌ای می‌باشد. همچنین سیگنال $x(t)$ سیگنال تحلیلی متناظر با سیگنال حقیقی $s(t)$ است. سیگنال تحلیلی یک سیگنال مختلط است که هم شامل مؤلفه‌های موهومی و هم شامل مؤلفه‌های حقیقی است. مؤلفه‌های موهومی توسط تبدیل هیلبرت به دست می‌آیند. اگر $s(t)$ یک سیگنال حقیقی باشد، سیگنال تحلیلی آن از رابطه (۳-۱۰) محاسبه می‌شود [۵۹].

$$x(t) = s(t) + js_{ht}(t) \qquad (۳-۱۰)$$

در رابطه (۳-۱۰)، $s_{ht}(t)$ تبدیل هیلبرت $s(t)$ است که با رابطه (۳-۱۱) بیان می‌شود [۵۹].

$$s_{ht}(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty} s(\tau)\frac{1}{t-\tau}\,d\tau \qquad (۳-۱۱)$$

توزیع ویگنر ویل در حالت گسسته از رابطه (۳-۱۲) به دست می‌آید [۵۸].

$$WV(t,\omega) = 2\sum_{n=-\infty}^{+\infty} x\left(t+\frac{n}{2}\right).x^*\left(t-\frac{n}{2}\right)e^{-2j\omega n} \qquad (۳-۱۲)$$

در پردازش گفتار دیجیتال می‌توان توزیع ویگنر ویل گسسته را با پنجره‌ای با طول N محدود کرد. در این صورت رابطه آن به صورت رابطه (۳-۱۳) می‌باشد [۵۸].

$$WV(t,\omega) = 2\sum_{n=-N+1}^{N-1} x\left(t+\frac{n}{2}\right).x^*\left(t-\frac{n}{2}\right)e^{-2j\omega n} \qquad (۳-۱۳)$$

همان‌طور که بیان شد توزیع ویگنر ویل از تبدیل فوریه تابع اتوکرولیشن سیگنال تحلیلی $x(t)$ به دست می‌آید. معمولا برای محاسبه تبدیل فوریه در توزیع ویگنر ویل از الگوریتم تبدیل فوریه سریع استفاده

می‌شود. اما در برخی از متون برای محاسبه تبدیل فوریه در این توزیع تبدیل هارتلی[1] سریع [۵۹] به کار رفته است که در بخش بعد به شرح آن می‌پردازیم.

۳-۸-۱ تبدیل هارتلی

تبدیل هارتلی یکی از تبدیلاتی است که سال‌ها مورد استفاده محققان بوده است اما اخیرا مورد توجه و علاقه بیشتری واقع شده است و این امر به دلیل کشف الگوریتم تبدیل هارتلی سریع می‌باشد. نتایج تحقیقات نشان داده است که الگوریتم تبدیل هارتلی سریع، سریع‌تر از الگوریتم تبدیل فوریه سریع عمل می‌کند.

تبدیل فوریه گسسته دنباله‌ای از اعداد حقیقی را از حوزه زمان به یک دنباله از اعداد مختلط در حوزه فرکانس تبدیل می‌کند. نیمی از اعداد موجود در حوزه فرکانس زائد هستند زیرا اطلاعات موجود در فرکانس‌های منفی، در فرکانس‌های مثبت تکرار می‌شوند. به عبارت دیگر اگر $X_f(k)$ تبدیل فوریه یک دنباله حقیقی از اعداد حقیقی باشد، رابطه‌های (۳-۱۴) و (۳-۱۵) برقرار می‌باشند [۱۳].

$$Real\big(X_f(k)\big) = Real\big(X_f(-k)\big) \qquad (۱۴-۳)$$

$$Imag\big(X_f(k)\big) = Imag\big(-X_f(-k)\big) \qquad (۱۵-۳)$$

حاصل‌ضرب اعداد مختلط در یک کامپیوتر نیاز به چهار حاصل‌ضرب ممیز شناور و دو جمع ممیز شناور دارد. با توجه به میزان حافظه مورد نیاز برای اطلاعات زائد و تعداد محاسبات لازم برای آنها، تبدیل فوریه روش کارایی برای تبدیل اعداد حقیقی به حوزه فرکانس نیست. بنابراین در برخی از مقالات برای رفع این مشکل تبدیل هارتلی معرفی شده است. تبدیل هارتلی گسسته یک دنباله از اعداد حقیقی در حوزه زمان را به یک دنباله از اعداد حقیقی در حوزه فرکانس تبدیل می‌کند. تبدیل هارتلی را می‌توان

[1] Hartley Transform

به عنوان الگوریتمی که افزونگی را در حوزه فوریه حذف می‌کند، بیان کرد. این حذف افزونگی با دسته‌بندی مجدد اعداد با رابطه (۳-۱۶) به دست می‌آید [۶۰].

$$X_h(k) = Real\big(X_f(k)\big) - Imag\big(X_f(k)\big) \quad (۱۶-۳)$$

در رابطه (۳-۱۶) $X_h(k)$ تبدیل هارتلی یک دنباله و $X_f(k)$ تبدیل فوریه همان دنباله است. همان طور که در این رابطه دیده می‌شود تبدیل هارتلی اعداد حقیقی را نتیجه می‌دهد. در تبدیل از حوزه فوریه به حوزه هارتلی هیچ اطلاعاتی از بین نمی‌رود زیرا داده‌ها در حوزه هارتلی می‌توانند با استفاده از روابط (۳-۱۷) و (۳-۱۸) به حوزه فوریه تبدیل شوند [۶۰].

$$Real\big(X_f(k)\big) = [X_h(k) + X_h(-k)]/2 \quad (۱۷-۳)$$

$$Imag\big(X_f(k)\big) = [X_h(k) - X_h(-k)]/2 \quad (۱۸-۳)$$

از آنجا که هیچ اطلاعاتی در این تبدیل از بین نمی‌رود و دنباله حقیقی در حوزه فرکانس از حداقل فضای حافظه مورد نیاز استفاده می‌کند و ضرایب کمتری نیاز دارد، تبدیل هارتلی نسبت به تبدیل فوریه کارایی بیشتری دارد.

تبدیل فوریه N نقطه‌ای از رابطه (۳-۱۹) و تبدیل هارتلی N نقطه‌ای توسط رابطه (۳-۲۰) به دست می‌آید [۶۱].

$$X_f(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)nk} \quad (۱۹-۳)$$

$$X_h(k) = \sum_{n=0}^{N-1} x(n) cas\big((2\pi/N)nk\big) \quad (۲۰-۳)$$

در رابطه (۳-۲۰) مقدار $cas\big((2\pi/N)nk\big)$ از رابطه (۳-۲۱) محاسبه می‌شود [۶۱].

$$cas\big((2\pi/N)nk\big) = \cos\big((2\pi/N)nk\big) + \sin\big((2\pi/N)nk\big) \quad (۲۱-۳)$$

لذا به منظور افزایش کارایی، در توزیع ویگنر ویل به جای تبدیل فوریه تابع اتوکرولیشن می‌توان تبدیل هارتلی تابع اتوکرولیشن را استفاده کرد. بنابراین رابطه توزیع ویگنر ویل با استفاده از تبدیل هارتلی از رابطه (۳-۲۲) به دست می‌آید.

$$WV(t,\omega) = 2\sum_{n=-N+1}^{N-1} x\left(t+\frac{n}{2}\right).x^*\left(t-\frac{n}{2}\right)cas(\omega nk) \qquad (۲۲-۳)$$

۳-۹ جمع‌بندی

روش‌های استخراج ویژگی مورد استفاده در سیستم‌های آشکارساز فعالیت گفتاری پیشنهادی شامل ویژگی‌های انرژی، ویژگی‌های حوزه کپسترال و توزیع ویگنر ویل می‌باشند. ویژگی انرژی نمایشی از تغییرات دامنه سیگنال است. ویژگی‌های حوزه کپسترال مطرح شده پنج نوع می‌باشند که عبارتند از: ضرایب کپسترال فرکانس مل، ضرایب کپسترال فرکانس مل ریشه، ضرایب پیش‌بینی ادراکی خطی، ضرایب کپسترال فرکانس بارک و ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده. از میان این ویژگی‌ها دو ویژگی ضرایب کپسترال فرکانس بارک و ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده ترکیبی از دو ویژگی ضرایب کپسترال فرکانس مل و ضرایب پیش‌بینی ادراکی خطی می‌باشند. ویژگی ویگنر ویل با استفاده از تبدیل هارتلی سریع به دست می‌آید و یک نمایش زمانی فرکانسی با رزولوشن بالا را از سیگنال‌های غیرایستا فراهم می‌نماید.

فصل ۴

روش‌های آشکارساز فعالیت گفتاری

پیشنهادی

۴-۱ پیشگفتار

در این فصل به شرح و بررسی روش‌های آشکارساز فعالیت گفتاری پیشنهادی می‌پردازیم.

فرآیند اصلی تشخیص فریم‌ها در همه روش‌های ارائه شده مطابق رابطه (۴-۱) است. طبق این رابطه فریمی که گفتار تشخیص داده شود، به عنوان فریم فعال انتخاب شده و مقدار ۱ به آن تعلق می‌گیرد و فریم غیرگفتار (سکوت یا نویز) به عنوان فریم غیرفعال انتخاب می‌شود و مقدار ۰ به آن اختصاص می‌یابد.

$$Frame_i = \begin{cases} 1, & if \quad speech \\ 0, & if\ nonspeech \end{cases} \tag{۱-۴}$$

۴-۲ روش‌های پیشنهادی

۴-۲-۱ روش اول

این روش بر مبنای روش ارائه شده در مرجع [۱۱] می‌باشد که در آن تغییراتی اعمال شده است. در این روش ابتدا سیگنال گفتار با روش تفاضل اسپکترال که در فصل دوم قسمت ۲-۲-۱ بیان شد، بهبود می‌یابد. سپس سیگنال بهبود یافته فریم‌بندی می‌شود و برای از بین بردن ناپیوستگی در لبه‌های فریم‌ها، هر فریم در یک پنجره همینگ ضرب می‌شود. رابطه این پنجره همان‌طور که در فصل دوم مطرح شد طبق رابطه (۲-۴) می‌باشد.

سپس از فریم‌های پنجره شده مطابق رابطه (۴-۲) مقادیر لگاریتم انرژی استخراج می‌شوند.

$$E_i = 10\log_{10}\left(\frac{1}{N-1}\sum_{n=1}^{N}(x_i[n] - \mu_i)^2 + \varepsilon\right) \tag{۲-۴}$$

در رابطه (۴-۲) $x_i[n]$، n امین نمونه از i امین فریم گفتار است. μ_i، میانگین نمونه‌های فریم i ام می‌باشد که از رابطه (۴-۳) به دست می‌آید. N تعداد نمونه‌های هر فریم و $\varepsilon = 10^{-6}$ مقدار ثابتی است که برای جلوگیری از صفر شدن لگاریتم در نظر گرفته شده است [۱۱].

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} x[n] \qquad (۳-۴)$$

مقادیر انرژی به دست آمده برای هر فریم مرتب می‌شوند. بدین ترتیب فریم‌های غیرگفتار که مقدار انرژی کمتری دارند، در ابتدای آرایه مرتب شده و فریم‌های گفتار که مقادیر انرژی بیشتری دارند در انتهای این آرایه قرار می‌گیرند. برای انتخاب فریم‌های گفتار و غیرگفتار درصد ثابتی از تعداد کل فریم‌های موجود در سیگنال در نظر گرفته می‌شود. که در اینجا این درصد ثابت ۱۰ درصد کل فریم‌ها تنظیم شده است. سپس از ابتدای آرایه میزان ۱۰ درصد کل فریم‌ها به عنوان فریم‌های غیرگفتار و از انتهای آرایه نیز میزان ۱۰ درصد کل فریم‌ها به عنوان فریم‌های گفتار انتخاب می‌شوند. به عنوان مثال اگر در یک سیگنال تعداد ۲۰۰۰ تا فریم وجود داشته باشند، میزان ۱۰ درصد آن تعداد ۲۰۰ فریم را در بر می‌گیرد. لذا از ابتدا تا اندیس ۲۰۰ آرایه انرژی‌های مرتب شده، به عنوان گفتار انتخاب می‌شوند و از اندیس ۱۸۰۰ تا اندیس ۲۰۰۰ (انتهای آرایه) به عنوان غیرگفتار در نظر گرفته می‌شوند.

پس از آن طبق روش مطرح شده در قسمت ۳-۴ ، ضرایب کپسترال فرکانس مل ریشه برای هر فریم پنجره شده از سیگنال بهبود یافته استخراج می‌شوند. سپس ضرایب معادل اندیس فریم‌های گفتار و غیرگفتاری که در مرحله قبل انتخاب شده‌اند، پیدا می‌شوند.

بردارهای انتخاب شده وارد مرحله آموزش می‌شوند. برای مرحله آموزش از الگویتم k-means استفاده شده است. مقدار K معادل ۱۶ تنظیم شده است. این الگوریتم یک بار برای بردارهای گفتار و بار دیگر برای بردارهای غیرگفتار تکرار می‌شود و بدین ترتیب مدل‌های گفتار و غیرگفتار طبق رابطه‌های (۴-۴) و (۴-۵) ساخته می‌شوند.

$$\lambda^{speech} = Train(RMFCC\ of\ speech\ frames) \qquad (۴-۴)$$

$$\lambda^{nonspeech} = Train(RMFCC\ of\ nonspeech\ frames) \qquad (۴-۵)$$

هر دو مدل گفتار و غیرگفتار به صورت مدل‌های مخلوط گوسی به فرم رابطه (۴-۶) در نظر گرفته می‌شوند.

$$p(x|\lambda) = \sum_{k=1}^{K} P_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (۶-۴)$$

در رابطه (۶-۴) P_k وزن‌های ترکیبی، μ_k بردارهای میانگین به دست آمده از الگوریتم k-means و Σ_k ماتریس‌های کواریانس می‌باشند. تعداد گوسین‌ها در هر دو مدل یکسان و به تعداد ۱۶ تا انتخاب شده‌اند. با توجه به [۶۲] الگوریتم k-means می‌تواند یک حالت حدی از الگوریتم ماکزیمم سازی امید ریاضی[۱] باشد. که در این حالت ماتریس‌های کواریانس به صورت ماترس‌های واحد مانند رابطه (۴-۷) می‌باشند.

$$\Sigma_k = \epsilon \mathrm{I}, \quad \epsilon \longrightarrow 0 \quad (۷-۴)$$

با فرض یکسان در نظر گرفتن هزینه کلاسه‌بندی نادرست و احتمالات پیشین برای هر دو مدل گفتار و غیرگفتار، تست نسبت لگاریتم احتمال[۲] برای بردار $x_i[n]$ در حضور گفتار طبق رابطه (۴-۸) به قاعده نزدیک‌ترین همسایه طبق رابطه (۴-۹) ساده می‌شود [۱۱].

$$\log p(x_i|\lambda^{speech}) \geq \log p(x_i|\lambda^{nonspeech}) \quad (۸-۴)$$

$$min_k \left\| x_i - \mu_k^{speech} \right\|^2 \leq min_k \left\| x_i - \mu_k^{nonspeech} \right\|^2 \quad (۹-۴)$$

بدین ترتیب مقدار نسبت لگاریتم احتمال برای هر فریم طبق رابطه (۴-۱۰) محاسبه می‌گردد.

$$LLR = min_k \left\| x_i - \mu_k^{nonspeech} \right\|^2 - min_k \left\| x_i - \mu_k^{speech} \right\|^2 \quad (۱۰-۴)$$

در این صورت فریم‌هایی به عنوان فریم گفتار انتخاب می‌شوند که در آنها این مقدار بزرگتر یا مساوی صفر باشد. علاوه بر این شرط از یک آستانه انرژی نیز استفاده می‌شود. مقدار این آستانه انرژی ۷۵-دسی‌بل در نظر گرفته شده است. بنابراین برای اینکه یک فریم به عنوان فریم گفتار تشخیص داده شود باید رابطه (۴-۱۱) برقرار باشد.

$$VAD = LLR \geq 0 \ \wedge \ E_i > energy\_threshold \quad (۱۱-۴)$$

[۱] Expectation Maximization (EM)

[۲] Log Likelihood Ratio (LLR) test

تفاوت این روش با روش مرجع [۱۱] این است که در این روش از ضرایب کپسترال فرکانس مل ریشه استفاده نموده‌ایم ولی در روش مرجع [۱۱] از ضرایب کپسترال فرکانس مل استفاده شده است. علاوه بر این ما در این روش ضرایب کپسترال فرکانس مل ریشه را از سیگنال بهبود یافته با تفاضل اسپکترال استخراج کرده‌ایم در حالی که در روش مرجع [۱۱] ضرایب کپسترال فرکانس مل از سیگنال اولیه استخراج شده‌اند.

۴-۲-۲ روش دوم

این روش مبتنی بر ضرایب پیش‌بینی ادراکی خطی و مقادیر لگاریتم انرژی می‌باشد. در این روش نیز مانند روش مطرح شده در قسمت ۴-۲-۱ عمل می‌کنیم با این تفاوت که به جای استفاده از ضرایب کپسترال فرکانس مل ریشه از ضرایب پیش‌بینی ادراکی خطی که مراحل استخراج آن در فصل سوم قسمت ۳-۵ تشریح شد، استفاده می‌کنیم. در این صورت رابطه‌های (۴-۴) و (۴-۵) به صورت روابط (۴-۱۲) و (۴-۱۳) به کار می‌روند.

$$\lambda^{speech} = Train(PLP\ of\ speech\ frames) \quad (۱۲-۴)$$

$$\lambda^{nonspeech} = Train(PLP\ of\ nonspeech\ frames) \quad (۱۳-۴)$$

۴-۲-۳ روش سوم

این روش مبتنی بر ضرایب کپسترال فرکانس بارک و مقادیر لگاریتم انرژی می‌باشد. در این روش نیز مانند روش مطرح شده در قسمت ۴-۲-۱ عمل می‌کنیم با این تفاوت که به جای استفاده از ضرایب کپسترال فرکانس مل ریشه از ضرایب کپسترال فرکانس بارک که مراحل استخراج آن در فصل سوم قسمت ۳-۶ تشریح شد، استفاده می‌کنیم. در این صورت رابطه‌های (۴-۴) و (۴-۵) به صورت روابط (۴-۱۴) و (۴-۱۵) به کار می‌روند.

$$\lambda^{speech} = Train(BFCC\ of\ speech\ frames) \quad (۱۴-۴)$$

$$\lambda^{nonspeech} = Train(BFCC\ of\ nonspeech\ frames) \quad (۱۵-۴)$$

۴-۲-۴ روش چهارم

این روش مبتنی بر ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده و مقادیر لگاریتم انرژی می‌باشد. در این روش نیز مانند روش مطرح شده در قسمت ۴-۲-۱ عمل می‌کنیم با این تفاوت که به جای استفاده از ضرایب کپسترال فرکانس مل ریشه از ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده که مراحل استخراج آن در فصل سوم قسمت ۳-۷ تشریح شد، استفاده می‌کنیم. در این صورت رابطه‌های (۴-۴) و (۴-۵) به صورت روابط (۴-۱۶) و (۴-۱۷) به کار می‌روند.

$$\lambda^{speech} = Train(RPLP\ of\ speech\ frames) \quad (۱۶-۴)$$

$$\lambda^{nonspeech} = Train(RPLP\ of\ nonspeech\ frames) \quad (۱۷-۴)$$

۴-۲-۵ روش پنجم

این روش مبتنی بر توزیع ویگنر ویل و مقادیر لگاریتم انرژی می‌باشد. در این روش نیز مانند روش‌های قبل ابتدا سیگنال گفتار با روش تفاضل اسپکترال بهبود می‌یابد. سپس سیگنال بهبود یافته فریم‌بندی می‌شود و برای از بین بردن ناپیوستگی در لبه‌های فریم‌ها، در یک پنجره همینگ که رابطه آن در فصل دوم بیان شد، ضرب می‌شوند.

سپس از فریم‌های پنجره شده مطابق رابطه (۳-۲۰) مقادیر ویگنر ویل را با استفاده از تبدیل هارتلی استخراج می‌نماییم. با توجه به اینکه طول هر فریم را ۳۰ میلی‌ثانیه در نظر گرفته‌ایم، تعداد نمونه‌های هر فریم حدود ۶۰۰ نمونه است. در نتیجه ویژگی ویگنر ویل استخراج شده برای هر فریم سیگنال یک ماتریس با طول و عرضی به ابعاد ۶۰۰ خواهد بود. بنابراین با داشتن تعداد بیش از ۲۰۰۰ فریم، با حجم زیادی از داده‌ها مواجه هستیم. به منظور کاهش این حجم داده‌ها فقط داده‌های سطر اول ماتریس را برای هر فریم انتخاب می‌کنیم. در این صورت با فرض داشتن ۲۰۰۰ فریم در یک سیگنال، یک ماتریس

با طول ۲۰۰۰ و عرض ۶۰۰ داریم. می توان این ماتریس را به صورت یک سیگنال جدید فریم‌بندی شده در نظر گرفت.

با استفاده از رابطه (۴-۱۸) از این سیگنال جدید ویژگی انرژی را استخراج می‌نماییم:

$$E\_wv_i = 10\log_{10}\left(\frac{1}{N-1}\sum_{n=1}^{N}(wv_i[n] - \mu_i)^2 + \epsilon\right) \quad (۱۸-۴)$$

که در آن $wv_i[n]$ ، n امین نمونه از i امین فریم ، μ_i میانگین نمونه‌های i امین فریم، N سایز فریم و ϵ ثابت دلخواهی که برای جلوگیری از ایجاد لگاریتم صفر برابر 10^{-16} در نظر گرفته شده است، می‌باشند.

در این صورت فریم‌هایی به عنوان فریم گفتار انتخاب می‌شوند که در آنها مقدار لگاریتم به دست آمده از رابطه (۴-۱۸) بزرگتر از یک مقدار آستانه تعیین شده باشد. این مقدار آستانه را به صورت تجربی برابر ۵۰- در نظر گرفته‌ایم. بنابراین برای اینکه یک فریم به عنوان فریم گفتار تشخیص داده شود باید رابطه (۴-۱۹) برقرار باشد.

$$VAD = E\_wv_i > -50 \quad (۱۹-۴)$$

۴-۲-۶ روش ششم

این روش نیز مانند روش بیان شده در قسمت ۴-۲-۱ عمل می‌کند که در آن از ضرایب کپسترال فرکانس مل استفاده شده است. اما روش استخراج ضرایب کپسترال فرکانس مل در آن با روش مطرح شده در قسمت ۳-۳ متفاوت می‌باشد. در این روش برای استخراج ضرایب کپسترال فرکانس مل به جای استفاده از تبدیل فوریه از توزیع ویگنر ویل استفاده می‌کنیم. بدین ترتیب که پس از فریم‌بندی و پنجره کردن، مجذور اندازه توزیع ویگنر ویل را از فیلترهای مل عبور داده و سپس از خروجی مل فیلترها لگاریتم و تبدیل کسینوسی گسسته می‌گیریم.

۴-۲-۷ روش هفتم

این روش نیز مانند روش بیان شده در قسمت ۴-۲-۱ عمل می‌کند اما روش استخراج ضرایب کپسترال فرکانس مل ریشه در آن با روش مطرح شده در قسمت ۳-۴ متفاوت می‌باشد. در این روش برای استخراج ضرایب کپسترال فرکانس مل ریشه به جای استفاده از تبدیل فوریه از توزیع ویگنر ویل استفاده می‌کنیم. بدین ترتیب که پس از فریم‌بندی و پنجره کردن، مجذور اندازه توزیع ویگنر ویل را از فیلترهای مل عبور داده و سپس از خروجی مل فیلترها به عنوان ورودی تابع ریشه استفاده می‌کنیم و در نهایت از خروجی تابع ریشه تبدیل کسینوسی گسسته می‌گیریم.

۴-۲-۸ روش هشتم

این روش نیز مانند روش بیان شده در قسمت ۴-۲-۱ عمل می‌کند که در آن از ضرایب پیش‌بینی ادراکی خطی استفاده شده است. اما روش استخراج ضرایب پیش‌بینی ادراکی خطی در آن با روش مطرح شده در قسمت ۳-۵ متفاوت می‌باشد. در این روش برای استخراج ضرایب پیش‌بینی ادراکی خطی به جای استفاده از تبدیل فوریه از توزیع ویگنر ویل استفاده می‌کنیم. بدین ترتیب که پس از فریم‌بندی و پنجره کردن، مجذور اندازه توزیع ویگنر ویل را از فیلترهای بارک عبور داده و دو مرحله دیگر آنالیز را روی آن اعمال می‌نماییم. پس از انجام تبدیل فوریه معکوس و با استفاده از مدل تمام قطب، ضرایب کپسترال را به دست می‌آوریم.

۴-۲-۹ روش نهم

این روش نیز مانند روش بیان شده در قسمت ۴-۲-۱ عمل می‌کند که در آن از ضرایب کپسترال فرکانس بارک استفاده شده است. اما روش استخراج ضرایب کپسترال فرکانس بارک در آن با روش مطرح شده در قسمت ۳-۶ متفاوت می‌باشد. در این روش برای استخراج ضرایب کپسترال فرکانس بارک به جای استفاده از تبدیل فوریه از توزیع ویگنر ویل استفاده می‌کنیم. بدین ترتیب که پس از فریم‌بندی

و پنجره کردن، مجذور اندازه توزیع ویگنر ویل را از فیلترهای بارک عبور داده و دو مرحله دیگر آنالیز را روی آن اعمال می‌نماییم. سپس از خروجی آنالیز شده لگاریتم و تبدیل کسینوسی گسسته می‌گیریم.

۴-۲-۱۰ روش دهم

این روش نیز مانند روش بیان شده در قسمت ۴-۲-۱ عمل می‌کند که در آن از ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده استفاده شده است. اما روش استخراج ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده در آن با روش مطرح شده در قسمت ۳-۷ متفاوت می‌باشد. در این روش برای استخراج ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده به جای استفاده از تبدیل فوریه از توزیع ویگنر ویل استفاده می‌کنیم. بدین ترتیب که پس از فریم‌بندی و پنجره کردن، مجذور اندازه توزیع ویگنر ویل را از فیلترهای مل عبور می‌دهیم. سپس با انجام تبدیل فوریه معکوس و با استفاده از مدل تمام قطب، ضرایب کپسترال را به دست می‌آوریم.

۴-۳ جمع‌بندی

روش‌های پیشنهادی در این پایان‌نامه را می‌توان به سه دسته تقسیم‌بندی کرد. دسته اول روش‌های مبتنی بر ضرایب کپسترال و مقادیر لگاریتم انرژی می‌باشند. چهار نوع از انواع ویژگی‌های کپسترال را در این دسته از روش‌ها استفاده نمودیم.

دسته بعدی روش مبتنی بر ویژگی ویگنر ویل است که به عنوان یکی از ویژگی‌های زمانی-فرکانسی به کار می‌رود.

دسته آخر روش‌های مبتنی بر ویژگی‌های کپسترال و ویژگی ویگنر ویل و مقادیر لگاریتم انرژی هستند. در این دسته از روش‌ها پنج نوع از انواع ویژگی‌های کپسترال را استفاده نمودیم. در محاسبه ویژگی‌های کپسترال استفاده شده در این دسته از روش‌ها، از ویژگی ویگنر ویل استفاده شده است.

فصل ۵

ارزیابی کارایی

۵-۱ پیشگفتار

در این فصل در قسمت ۵-۲ ویژگی‌های دادگان گفتاری مورد استفاده در این کار را بیان می‌نماییم. در قسمت ۵-۳ روش محاسبه خطا را توضیح می‌دهیم. در قسمت ۵-۴ مقداردهی پارامترهای به کار رفته در روش‌های پیشنهادی را عنوان می‌کنیم. در قسمت ۵-۵ نتایج حاصل از چندین آزمایش انجام شده روی روش‌های فصل چهارم ارائه می‌شوند و هر آزمایش مستقل از سایر آزمایشات، تجزیه و تحلیل و ارزیابی می‌گردد. در قسمت ۵-۶ جمع‌بندی کلی و در قسمت ۵-۷ پیشنهادات آینده را خواهیم داشت.

۵-۲ دادگان گفتاری مورد استفاده

یکی از مهم‌ترین ابزارها برای طراحی و ساخت سیستم‌های هوشمند شناسایی گفتار، تهیه دادگان گفتاری برای آموزش مدل‌های صوتی صداها و آواهای تشکیل‌دهنده یک زبان است. برای رسیدن به این هدف تعداد بسیار زیادی نمونه از این آواها که توسط گویندگان مختلف ادا شده‌اند، باید توسط انسان و به صورت دستی و با دقت میلی‌ثانیه از گفتار آنها استخراج شوند.

دادگان فارس‌دات[۱] [۶۳] مجموعه‌ای از عبارات و جملات است که توسط گویندگان فارسی زبان از مناطق مختلف کشور بیان شده است. این دادگان در سطح واج (آوا) با دقت میلی‌ثانیه تقطیع و برچسب‌دهی شده و به صورت فایل‌های مجزا ذخیره گردیده است. این دادگان، به عنوان دادگان استاندارد گفتاری زبان فارسی در داخل و خارج کشور شناخته شده هستند و برای آموزش سیستم‌های هوشمند تشخیص گفتار، تبدیل گفتار به متن فارسی و تعیین هویت گوینده استفاده می‌شود. اهمیت طراحی و ساخت این دادگان در زبان فارسی، همانند اهمیت دادگان TIMIT در زبان انگلیسی است و دارای ساختاری مشابه با آن می‌باشد. در زبان فارسی، دادگانی با تنوع، وسعت و کیفیت برچسب‌زنی دادگان فارس‌دات وجود ندارد. دادگان فارس‌دات دارای ویژگی‌ها و قابلیت‌های زیر است:

[۱] FarsDat

۱- استخراج پرکاربردترین کلمات زبان فارسی از روزنامه‌ها.

۲- طراحی ۳۸۶ جمله با استفاده از ۱۰۰۰ کلمه شامل کلیه دنباله‌های دوآوایی در زبان فارسی.

۳- انتخاب ۳۰۴ گوینده بر حسب جنسیت، سن، میزان تحصیلات و لهجه از نقاط مختلف کشور.

۴- پوشش ۱۰ لهجه رایج فارسی در کشور اعم از تهرانی، ترکی، اصفهانی، جنوبی، شمالی، خراسانی، بلوچی، کردی، لری و یزدی.

۵- تولید ۲۰ جمله در ۲ جلسه توسط هر گوینده با کیفیت صوتی بسیار بالا در اتاقک ضد صدا.

۶- تقطیع دستی جملات به آواهای تشکیل‌دهنده آنها و تهیه برچسب‌های واجی و آوایی برای آن.

۷- ضبط جملات با فرکانس ۲۲۰۵۰ هرتز و با کیفیت بسیار بالا (نسبت سیگنال به نویز ۳۴ دسی‌بل).

۸- استفاده از نشانه‌گذاری بین‌المللی IPA[۱] برای برچسب‌زنی آواها و واج‌های زبان فارسی.

۹- تأیید کیفیت دادگان توسط مؤسسه بین‌المللی ELRA[۲].

با توجه به مزایای بیان شده در مورد دادگان گفتاری فارس‌دات و دقت آن در تقطیع و برچسب‌گذاری و همچنین اهمیت استفاده از دادگان گفتاری به زبان فارسی، به منظور ارزیابی کارایی روش‌های پیشنهادی در این تحقیق از این دادگان استفاده کرده‌ایم.

۵-۳ محاسبه خطا

برای محاسبه خطا از رابطه (۵-۱) استفاده می‌کنیم.

$$error = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T(n)}\sum_{t=1}^{T(n)}\mathcal{L}\left\{\hat{l}_t(n) \neq l_t(n)\right\} \qquad (۱-۵)$$

[۱] IPA: International Phonetic Association

[۲] ELRA: European Language Resources Association

در رابطه (۵-۱) N تعداد فایل‌های انتخابی از دادگان فارس‌دات، $T(n)$ تعداد فریم‌های موجود در فایل انتخابی، $l_t(n)$ برچسب استاندارد مربوط به فایل انتخابی موجود در دادگان و $\hat{l}_t(n)$ برچسب تعیین شده توسط روش پیشنهادی است. در تابع $\mathcal{L}$ هنگامی که دو مقدار $l_t(n)$ و $\hat{l}_t(n)$ مخالف هم باشند مقدار ۱ را نتیجه می‌دهد.

۵-۴ تنظیم پارامترها

پارامترهای موجود در روش‌های مطرح شده را بدین ترتیب مقداردهی می‌نماییم.

۱- پارامترهای موجود در روش تفاضل اسپکترال را با توجه به مرجع [۱۱] مقداردهی می‌نماییم. فاکتور تفاضل α را مقدار ثابت ۱۰ تنظیم می‌کنیم. ناحیه تفاضل γ را برابر ۲ و توان بهره e را نیز برابر ۲ در نظر می‌گیریم. بدین ترتیب با در نظر گرفتن $(\gamma, e) = (۲,۲)$ از فیلتر وینر[۱] برای ناحیه تفاضل استفاده می‌کنیم. همچنین برای ماکزیمم نویز محدوده زمین g_h و ماکزیمم کاهش نویز در ناحیه توان β به ترتیب مقادیر ثابت ۱ و ۰،۰۱ را تعیین می‌نماییم. برای تخمین نویز از روش مینیمم مربع میانگین خطا[۲] استفاده کرده‌ایم.

۲- طول فریم‌ها را ۳۰ میلی‌ثانیه در نظر می‌گیریم که به اندازه ۱۰ میلی‌ثانیه با یکدیگر همپوشانی دارند [۱۱].

۳- مقدار ضریب α در فیلتر پیش‌تأکید ۰،۹۷ می‌باشد.

۴- تعداد فیلترهای مل و بارک ۲۷ فیلتر هستند [۱۱].

۵- از هرکدام از ضرایب کپسترال استفاده شده فقط تعداد ۱۲ ضریب اولیه استخراج شده را برای هر فریم انتخاب می‌کنیم [۱۱].

---

[۱] Wiener Filter

[۲] Minimum Mean Square Error (MMSE)

۶- مقدار γ در روش ضرایب کپسترال فرکانس مل ریشه ۰,۰۰۱ است.

۷- تعداد قطب‌های مدل تمام قطب در روش‌های پیش‌بینی ادراکی خطی و پیش‌بینی ادراکی خطی تجدید نظر یافته ۱۲ قطب می‌باشند.

۸- در الگوریتم k-means مقدار k را برابر ۱۶ قرار داده و همچنین تعداد گوسین‌های گفتار و غیرگفتار را نیز ۱۶ در نظر گرفته‌ایم [۱۱].

۹- تعداد ۵۰ فایل را به صورت تصادفی از دادگان گفتاری فارس‌دات انتخاب می‌کنیم.

۱۰- برای ارزیابی روش‌های پیشنهادی و محاسبه کارایی آنها انواع مختلف نویز که از مرجع [۶۴] دانلود شده‌اند، به صورت مصنوعی به هر سیگنال گفتاری در نسبت‌های مختلف سیگنال به نویز اضافه می‌گردند. این فرآیند به صورتی که در قسمت ۲-۷-۱ بیان شد صورت می‌گیرد. نویزهای استفاده شده در این تحقیق عبارت از نویز سفید، صدای دستگاه تهویه هوای مطبوع، صدای ماشین لباسشویی و صدای همهمه می‌باشند.

۵-۵ نتایج آزمایشات

در این قسمت نتایج آزمایشات مختلف انجام شده روی روش‌های فصل چهارم را بررسی و تجزیه و تحلیل می‌نماییم.

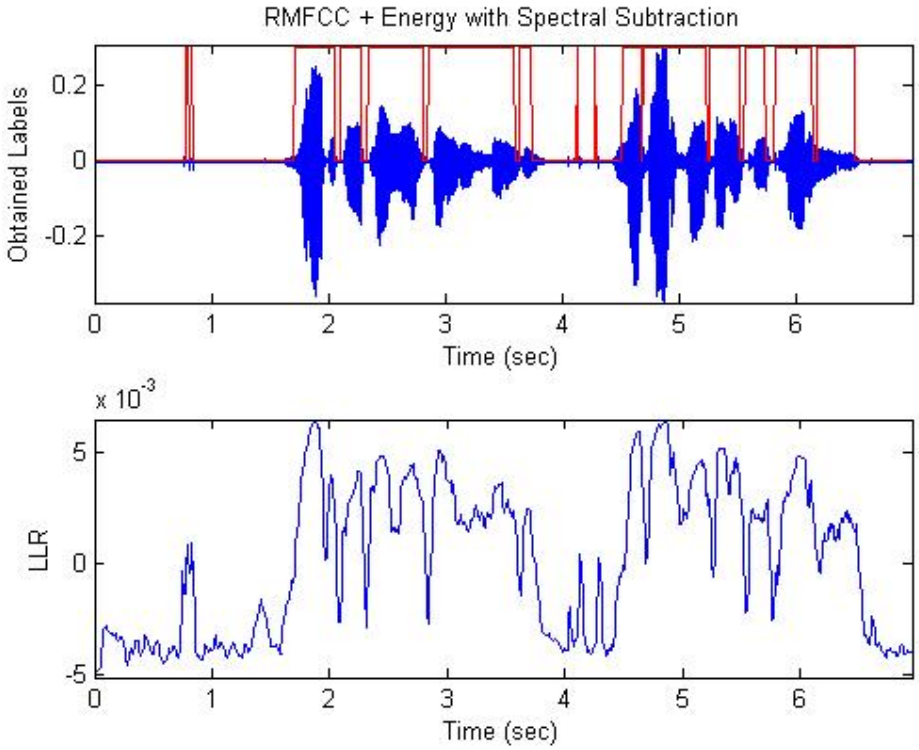
۵-۵-۱ آزمایش اول

در این آزمایش یکی از فایل‌های دادگان گفتاری فارس‌دات را به صورت تصادفی انتخاب کرده‌ایم. شکل (۵-۱) اطلاعات ۷ ثانیه از این فایل را برای روش پیشنهادی اول نمایش می‌دهد. در این شکل دو نمودار مشاهده می‌شوند. نمودار اول آن شکل موج ۷ ثانیه از این فایل انتخابی را نشان می‌دهد. شکل موج این فایل در بازه‌های (۴ – ۱,۸) ثانیه و (۶,۶ – ۴,۵) ثانیه حاوی گفتار و در سایر بازه‌ها حاوی سکوت است. خروجی مطلوب سیستم آشکارساز فعالیت گفتاری می‌بایست در بازه‌های گفتار معادل غیرصفر (۰,۲۵)

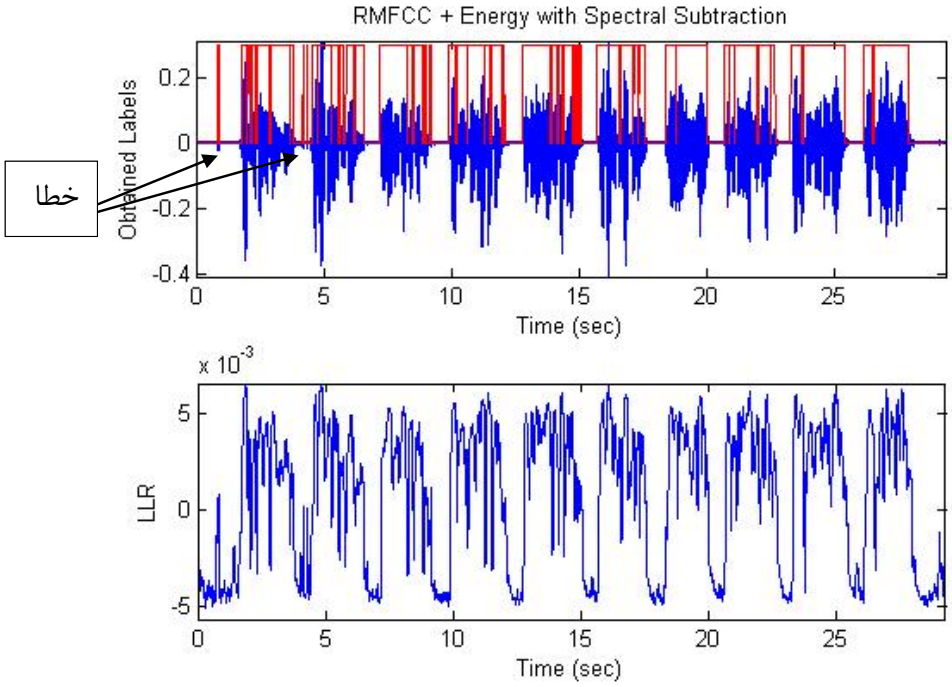
و در بازه‌های غیرگفتار معادل ۰ باشد. در غیر این صورت سیستم دارای خطا خواهد بود. در اولین نمودار شکل (۵-۱) علاوه بر شکل موج ۷ ثانیه از فایل انتخابی دادگان گفتاری فارس‌دات، برچسب‌های خروجی سیستم آشکارساز فعالیت گفتاری به دست آمده از روش پیشنهادی اول نمایش داده شده است. همان‌طور که در این شکل مشاهده می‌شود خروجی حاصل از این روش در بازه‌های گفتار به درستی تعیین شده است ولی در بازه‌های سکوت دارای خطا می‌باشد. این خطا در بازه‌های (۱ – ۰) ثانیه و (۴٫۵ – ۴) ثانیه رخ داده است. این بازه‌ها حاوی سکوت می‌باشند و می‌بایست برچسب ۰ داشته باشند، در حالی که در این روش پیشنهادی برچسب غیر صفر به آنها اختصاص یافته و به عنوان گفتار تشخیص داده شده‌اند. نمودار دوم در شکل (۵-۱) مقدار نسبت لگاریتم احتمال را نشان می‌دهد که این مقدار طبق رابطه (۴-۱۰) محاسبه شده است. از مقایسه این دو نمودار نتیجه می‌شود قسمت‌هایی که در نمودار دوم دارای مقدار نسبت لگاریتم احتمال بزرگتر یا مساوی ۰ می‌باشند، در نمودار اول به عنوان گفتار انتخاب شده‌اند.

شکل‌های (۵-۲) تا (۵-۱۱) نمایش خروجی سیستم آشکارساز فعالیت گفتاری برای این فایل انتخابی از دادگان گفتاری فارس‌دات، با استفاده از هر یک از روش‌های پیشنهادی می‌باشند. در این شکل‌ها (به جز شکل (۵-۶)) نیز مانند شکل (۵-۱) دو نمودار وجود دارد. اولین نمودار آن نمایش شکل موج فایل انتخابی دادگان فارس‌دات را نشان می‌دهد. علاوه بر آن برچسب‌های به دست آمده از هر روش نیز در نمودار اول مشاهده می‌شوند. نمودار دوم هر شکل، مقدار نسبت لگاریتم احتمال در هر روش را نشان می‌دهد. در شکل (۵-۸) مقدار نسبت لگاریتم احتمال در نمودار دوم معادل ۰ می‌باشد. این مطلب نشان می‌دهد که در روش پیشنهادی هفتم این مقدار در تشخیص قسمت‌های گفتار نقش تعیین کننده‌ای ندارد و تنها با استفاده از معیار آستانه انرژی تصمیم‌گیری لازم انجام شده است. در شکل (۵-۶) که مربوط به روش پیشنهادی پنجم است تنها یک نمودار وجود دارد که مانند اولین نمودار سایر شکل‌ها می‌باشد. در این روش پیشنهادی از روش آستانه‌گذاری برای آشکارسازی فعالیت گفتاری استفاده نموده‌ایم، بنابراین مقدار نسبت لگاریتم احتمال در آن وجود ندارد.

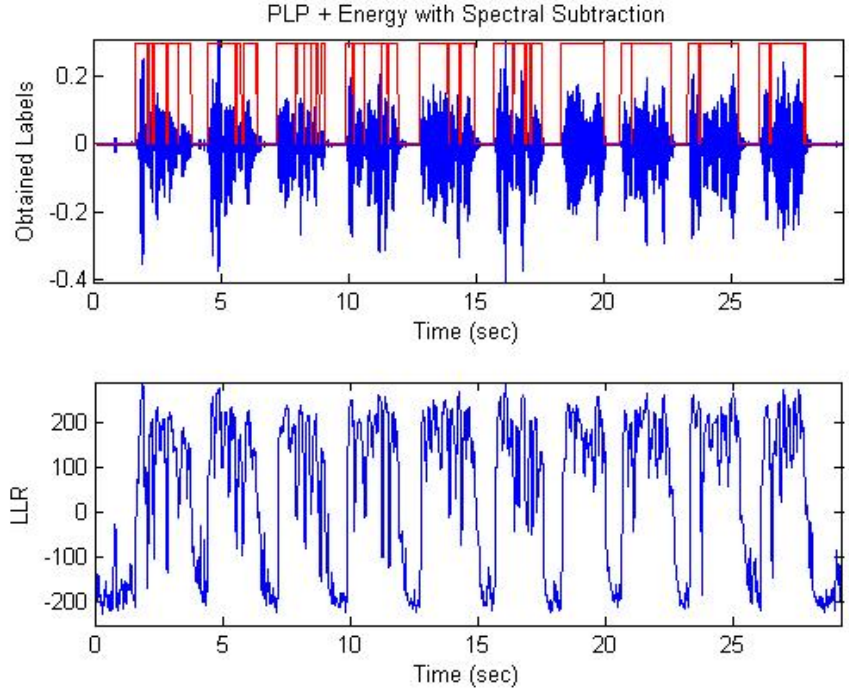
سپس در شکل‌های (۵-۱۲) و (۵-۱۳) به ترتیب خروجی حاصل از روش‌های مراجع [۱۱] و [۴۱] مشاهده می‌شوند. شکل (۵-۱۲) نیز از دو نمودار تشکیل شده است. نمودار اول آن نمایش شکل موج فایل انتخاب شده از دادگان گفتاری فارس‌دات به همراه برچسب‌های ایجاد شده در این روش می‌باشد. نمودار دوم آن مقدار نسبت لگاریتم احتمال را نشان می‌دهد. نمودار شکل (۵-۱۳) مربوط به برچسب‌های تولید شده در روش استاندارد G729 [۴۱] برای فایل منتخب دادگان گفتاری فارس‌دات است. از مقایسه این شکل با سایر شکل‌ها می‌توان نتیجه گرفت که روش مرجع [۴۱] نمی‌تواند به خوبی سایر روش‌های ارائه شده قسمت‌های گفتار را از قسمت‌های غیرگفتار تشخیص دهد، زیرا بسیاری از بازه‌های غیرگفتار را به عنوان گفتار تشخیص می‌دهد. در شکل‌های (۵-۲)، (۵-۵)، (۵-۸) و (۵-۱۲) قسمت‌هایی که با فلش و خطا نمایش داده شده‌اند، قسمت‌های غیرگفتاری هستند که به عنوان گفتار تشخیص داده شده‌اند. اما در مورد کیفیت یا عدم کیفیت سایر روش‌ها نمی‌توان اظهار نظر کرد و نیاز به آزمایشات بیشتری می‌باشد.



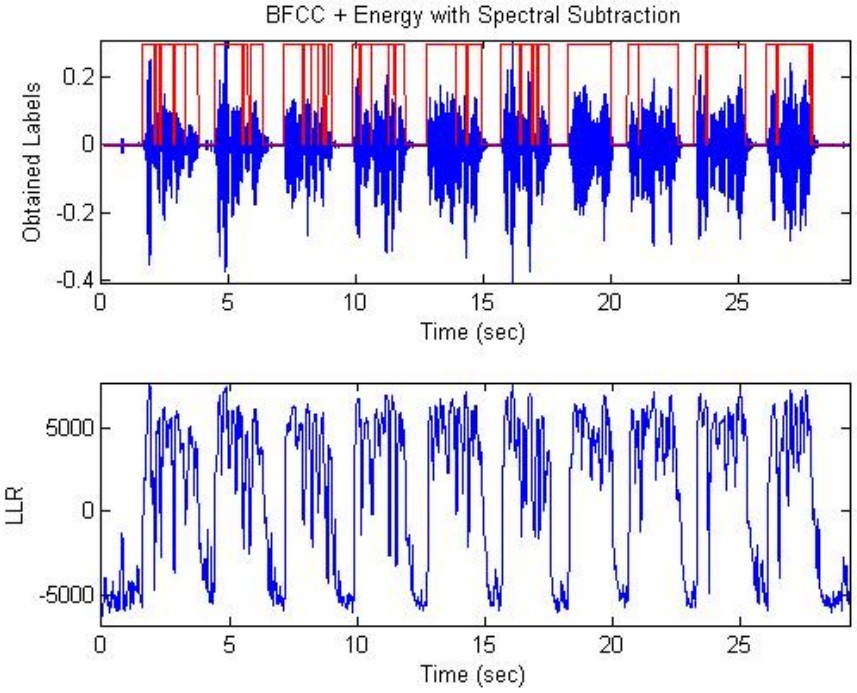
شکل (۵-۱) خروجی روش پیشنهادی اول برای ۷ ثانیه از فایل انتخابی دادگان فارس‌دات



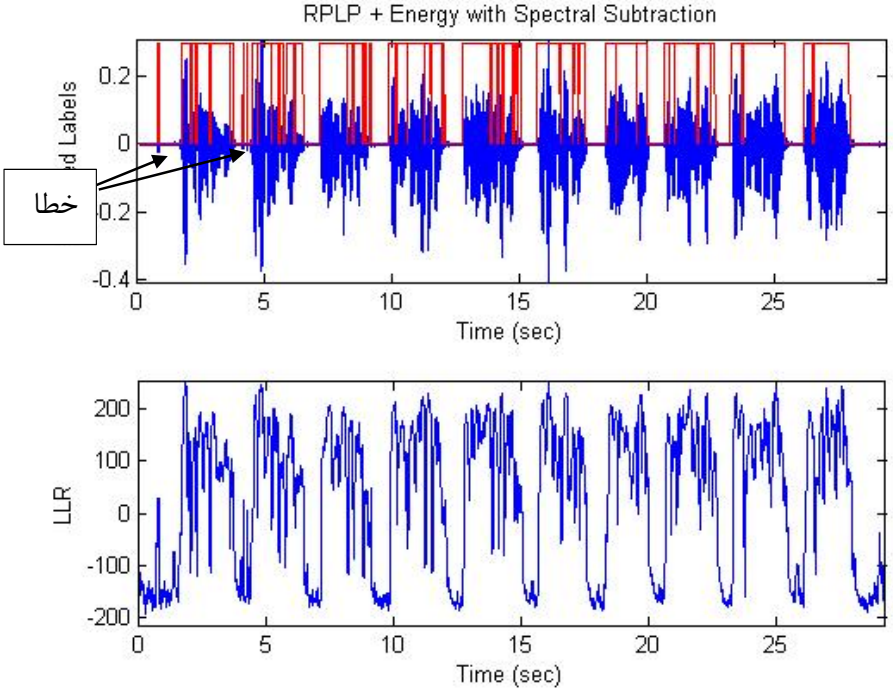
شکل (۵-۲) خروجی روش پیشنهادی اول برای فایل انتخابی دادگان فارس‌دات



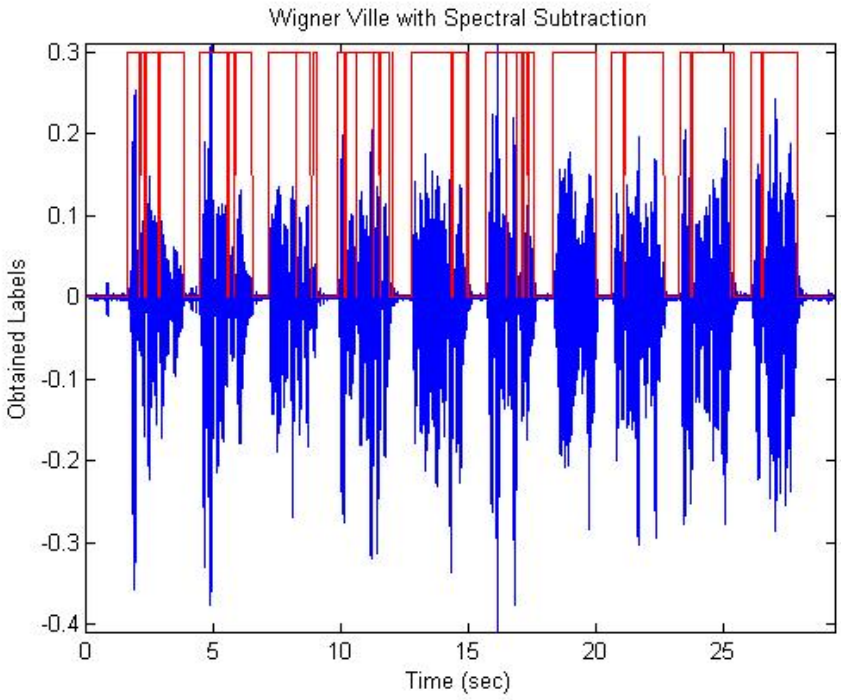
شکل (۵-۳) خروجی روش پیشنهادی دوم برای فایل انتخابی دادگان فارس‌دات



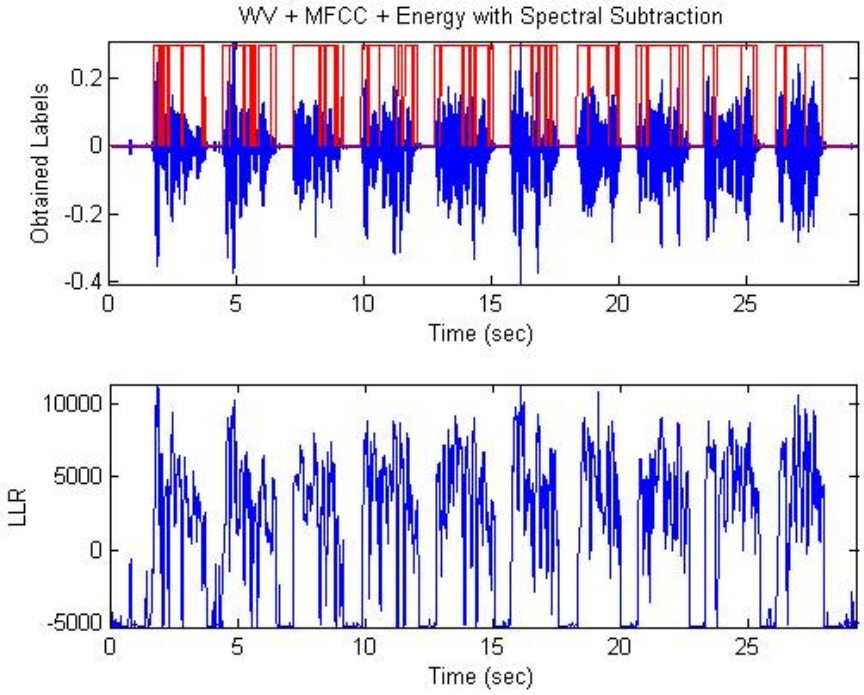
شکل (۵-۴) خروجی روش پیشنهادی سوم برای فایل انتخابی دادگان فارس‌دات



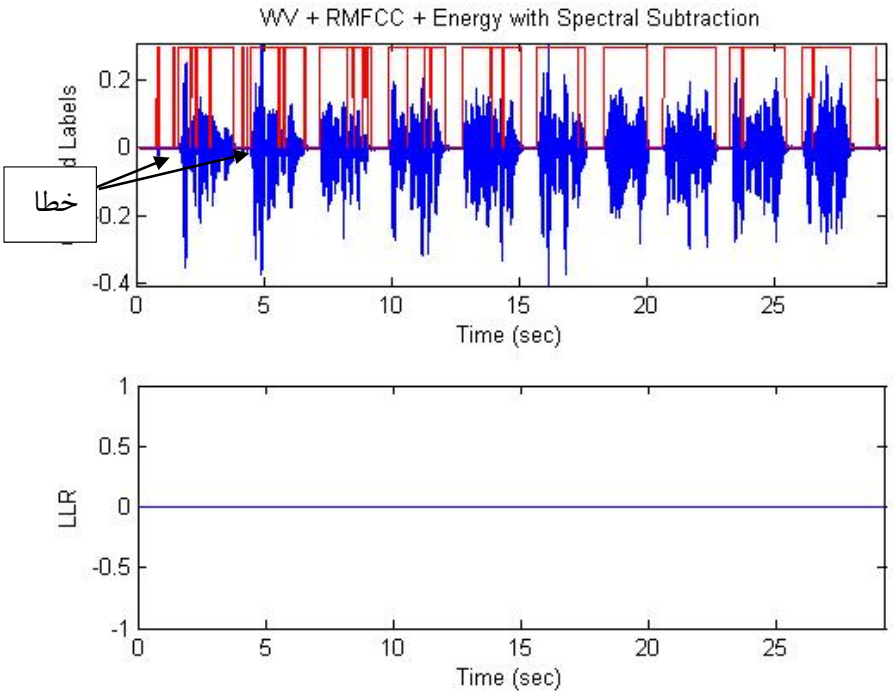
شکل (۵-۵) خروجی روش پیشنهادی چهارم برای فایل انتخابی دادگان فارس‌دات



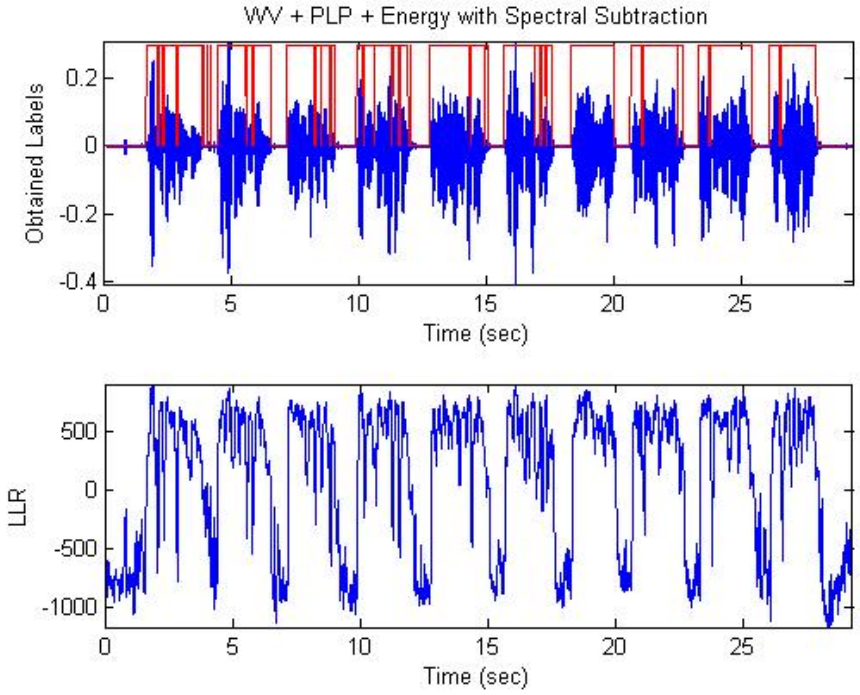
شکل (۵-۶) خروجی روش پیشنهادی پنجم برای فایل انتخابی دادگان فارس‌دات



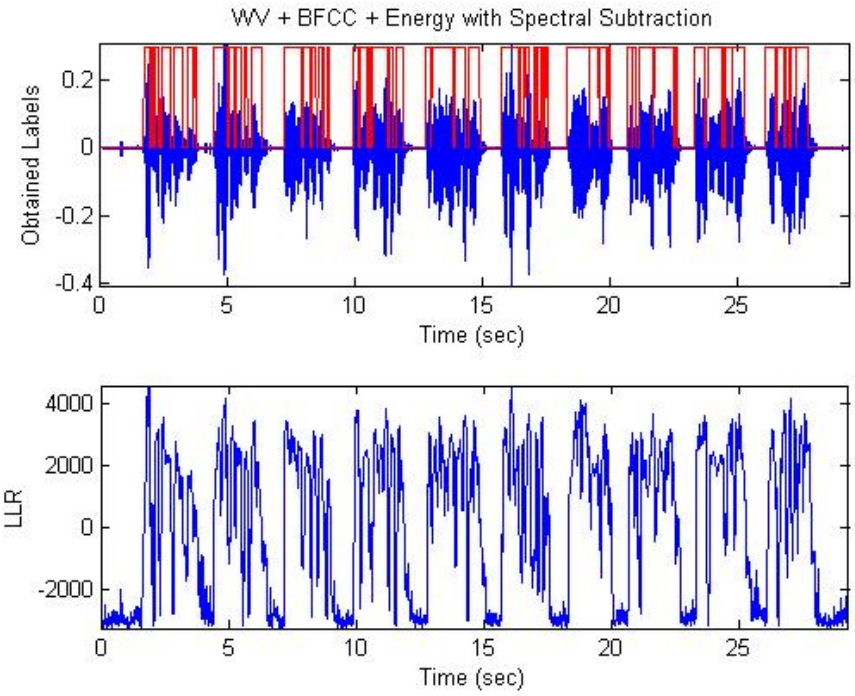
شکل (۵-۷) خروجی روش پیشنهادی ششم برای فایل انتخابی دادگان فارس‌دات



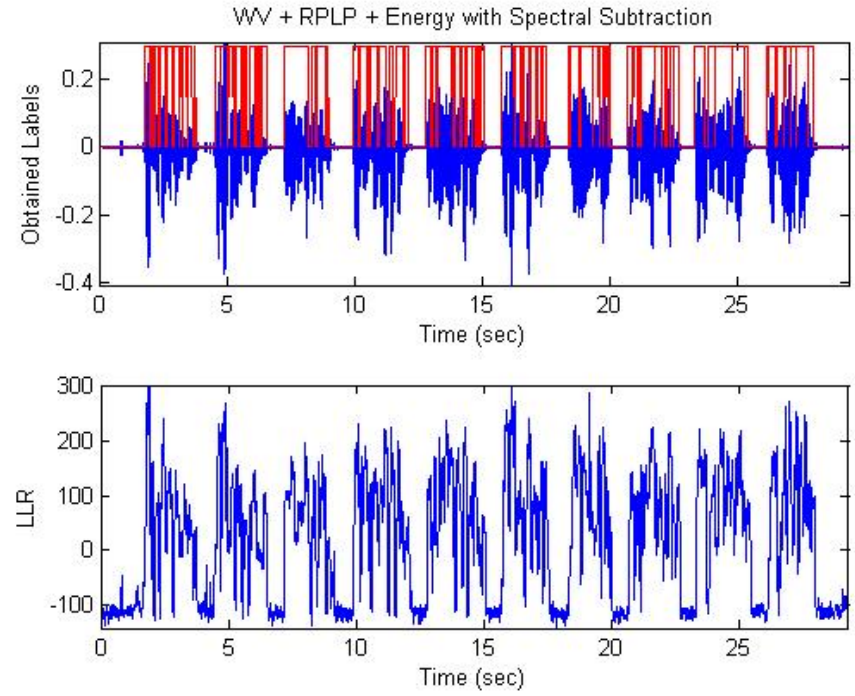
شکل (۵-۸) خروجی روش پیشنهادی هفتم برای فایل انتخابی دادگان فارس‌دات



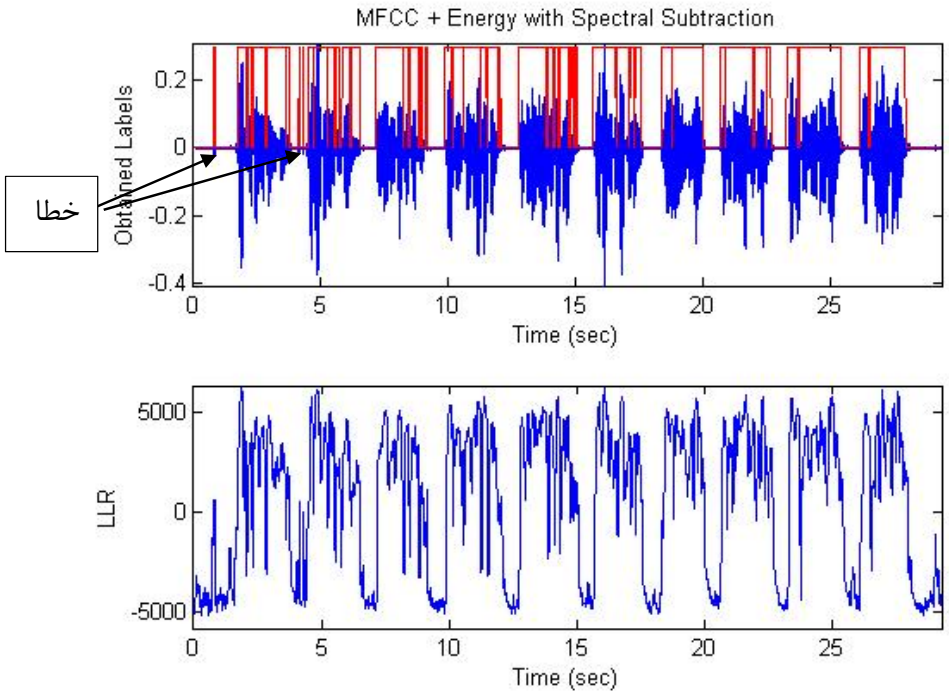
شکل (۵-۹) خروجی روش پیشنهادی هشتم برای فایل انتخابی دادگان فارس‌دات



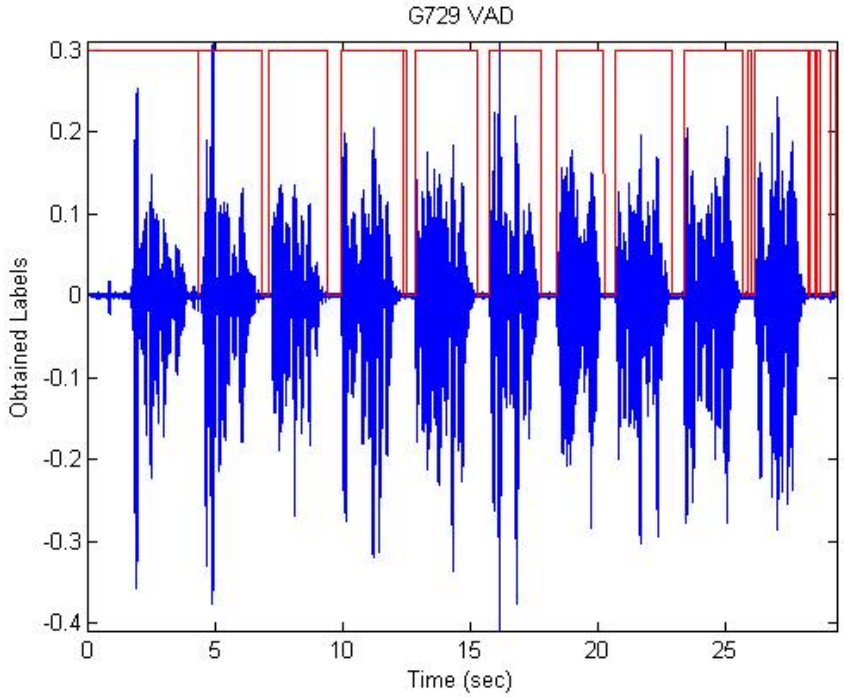
شکل (۵-۱۰) خروجی روش پیشنهادی نهم برای فایل انتخابی دادگان فارس‌دات



شکل (۵-۱۱) خروجی روش پیشنهادی دهم برای فایل انتخابی دادگان فارس‌دات



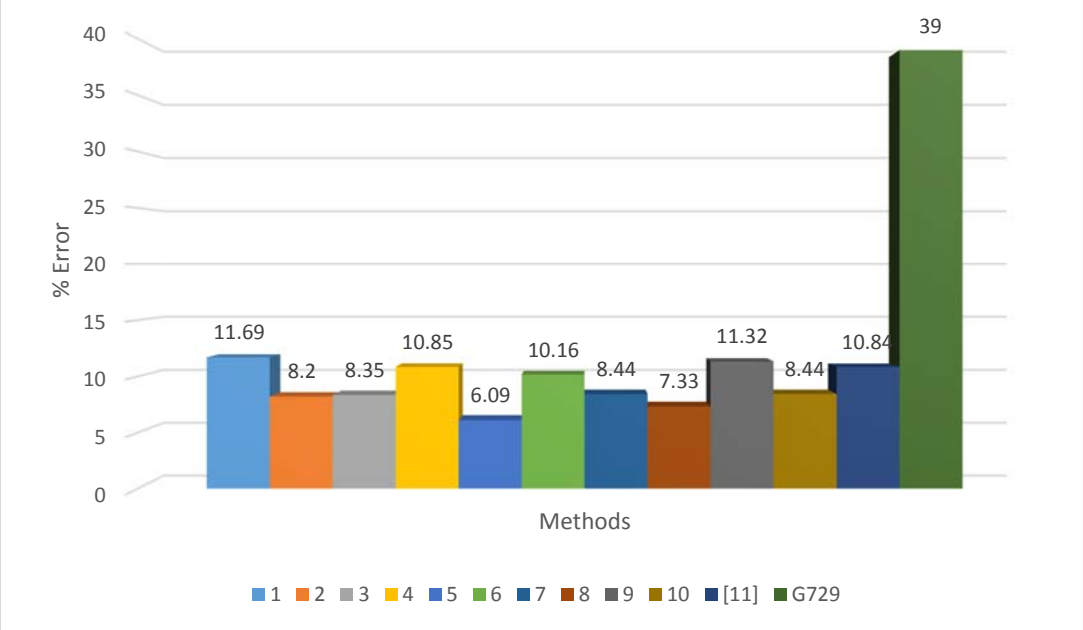
شکل (۵-۱۲) خروجی روش مرجع [۱۱] برای فایل انتخابی دادگان فارس‌دات



شکل (۵-۱۳) خروجی روش استاندارد G729 [۴۱] برای فایل انتخابی دادگان فارس‌دات

۵-۵-۲ آزمایش دوم

در این آزمایش ۵۰ فایل از دادگان گفتاری فارس‌دات را به صورت تصادفی انتخاب می‌کنیم. سپس با استفاده از رابطه (۵-۱) درصد خطای هر کدام از روش‌های مطرح شده را در حالت خالص و بدون نویز ارزیابی می‌کنیم. نتایج این آزمایش در جدول (۵-۱) قرار دارند. همان‌طور که در این جدول مشاهده می‌شود، روش پیشنهادی پنجم در مقایسه با سایر روش‌های پیشنهادی و روش مرجع [۱۱] و روش استاندارد G729 [۴۱] درصد خطای کمتری دارد. در شکل (۵-۱۴) نمودار میله‌ای مربوط به نتایج این آزمایش مشاهده می‌شود.



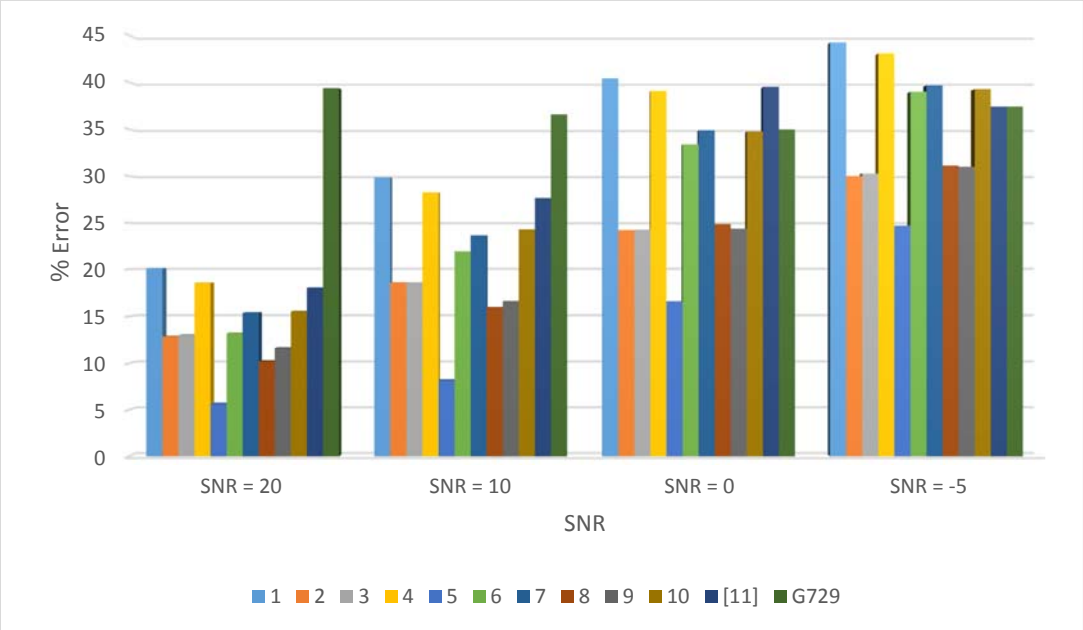
شکل (۵-۱۴) درصد خطای روش‌های ارائه شده در حالت بدون نویز

جدول (۵-۱) درصد خطای روش‌های ارائه شده در حالت بدون نویز

روش	درصد خطا
پیشنهادی ۱	۱۱,۶۹
پیشنهادی ۲	۸,۲۰
پیشنهادی ۳	۸,۳۵
پیشنهادی ۴	۱۰,۸۵
پیشنهادی ۵	**۶,۰۹**
پیشنهادی ۶	۱۰,۱۶
پیشنهادی ۷	۸,۴۴
پیشنهادی ۸	۷,۳۳
پیشنهادی ۹	۱۱,۳۲
پیشنهادی ۱۰	۸,۴۴
مرجع [۱۱]	۱۰,۸۴
G729 [۴۱]	۳۹

۵-۵-۳ آزمایش سوم

در این آزمایش ۵۰ فایل از دادگان گفتاری فارس‌دات را به صورت تصادفی انتخاب می‌کنیم. سپس با استفاده از رابطه (۵-۱) درصد خطای هر کدام از روش‌های مطرح شده را در حضور نویز سفید و تحت نسبت‌های سیگنال به نویز ۲۰ و ۱۰ و ۰ و ۵- دسی‌بل محاسبه می‌نماییم. روش افزودن نویز مطابق روش بیان شده در فصل دوم قسمت ۲-۷-۱ می‌باشد. نتایج حاصل از این آزمایش در جدول (۵-۲) قرار دارد. با توجه به اطلاعات این جدول در همه انواع نسبت‌های سیگنال به نویز آزمایش شده روش پیشنهادی پنجم درصد خطای کمتری را از خود نشان داده است. نمودار میله‌ای حاصل از این نتایج در شکل (۵-۱۵) مشاهده می‌شود.



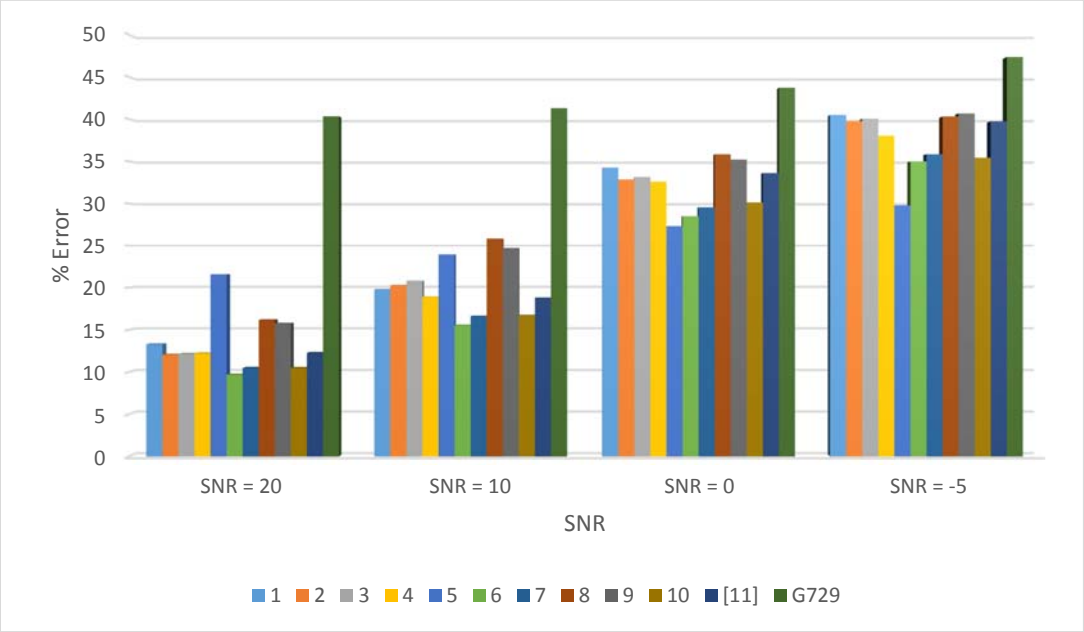
شکل (۵-۱۵) درصد خطای روش‌های ارائه شده در حضور نویز سفید

جدول (۵-۲) درصد خطای روش‌های ارائه شده در حضور نویز سفید

SNR / روش	۲۰	۱۰	۰	-۵	میانگین
پیشنهادی ۱	۲۰,۱۰	۲۹,۸۶	۴۰,۴۱	۴۴,۲۴	۳۳,۶۵
پیشنهادی ۲	۱۲,۸۵	۱۸,۵۹	۲۴,۲۳	۲۹,۹۸	۲۱,۴۱
پیشنهادی ۳	۱۳,۰۲	۱۸,۵۷	۲۴,۲۸	۳۰,۲۶	۲۱,۵۴
پیشنهادی ۴	۱۸,۵۶	۲۸,۲۶	۳۹,۰۸	۴۳,۰۷	۳۲,۲۴
پیشنهادی ۵	**۵,۶۲**	**۸,۱۶**	**۱۶,۵۵**	**۲۴,۶۹**	**۱۳,۷۵**
پیشنهادی ۶	۱۳,۲۰	۲۱,۹۱	۳۳,۳۷	۳۸,۹۸	۲۶,۸۷
پیشنهادی ۷	۱۵,۳۸	۲۳,۶۷	۳۴,۸۶	۳۹,۶۶	۲۸,۳۹
پیشنهادی ۸	۱۰,۲۲	۱۵,۹۳	۲۴,۸۸	۳۱,۱۱	۲۰,۴
پیشنهادی ۹	۱۱,۶۳	۱۶,۵۸	۲۴,۳۷	۳۰,۹۸	۲۰,۸۹
پیشنهادی ۱۰	۱۵,۵۲	۲۴,۳۱	۳۴,۷۴	۳۹,۲۵	۲۸,۴۶
مرجع [۱۱]	۱۸,۰۴	۲۷,۶۵	۳۹,۵۰	۴۴,۸۸	۳۲,۵۲
G729 [۴۱]	۳۹,۳۶	۳۶,۵۶	۳۴,۹۶	۳۷,۳۹	۳۷,۰۷

۵-۵-۴ آزمایش چهارم

این آزمایش نیز مانند آزمایش دوم می‌باشد با این تفاوت که در حضور نویز دستگاه تهویه هوای مطبوع انجام می‌شود. نتایج حاصل از این آزمایش در جدول (۵-۳) قرار دارد. با توجه به مقادیر این جدول در حالت‌هایی که نسبت سیگنال به نویز ۰ و ۵- دسی‌بل هستند روش پیشنهادی پنجم درصد خطای کمتری دارد و در حالت‌های نسبت سیگنال به نویز ۱۰ و ۲۰ دسی‌بل درصد خطای روش پیشنهادی ششم کمتر می‌باشد. به صورت میانگین روش پیشنهادی ششم کارایی بهتری را از خود نشان داده است. نمودار این نتایج در شکل (۵-۱۶) مشاهده می‌شود.



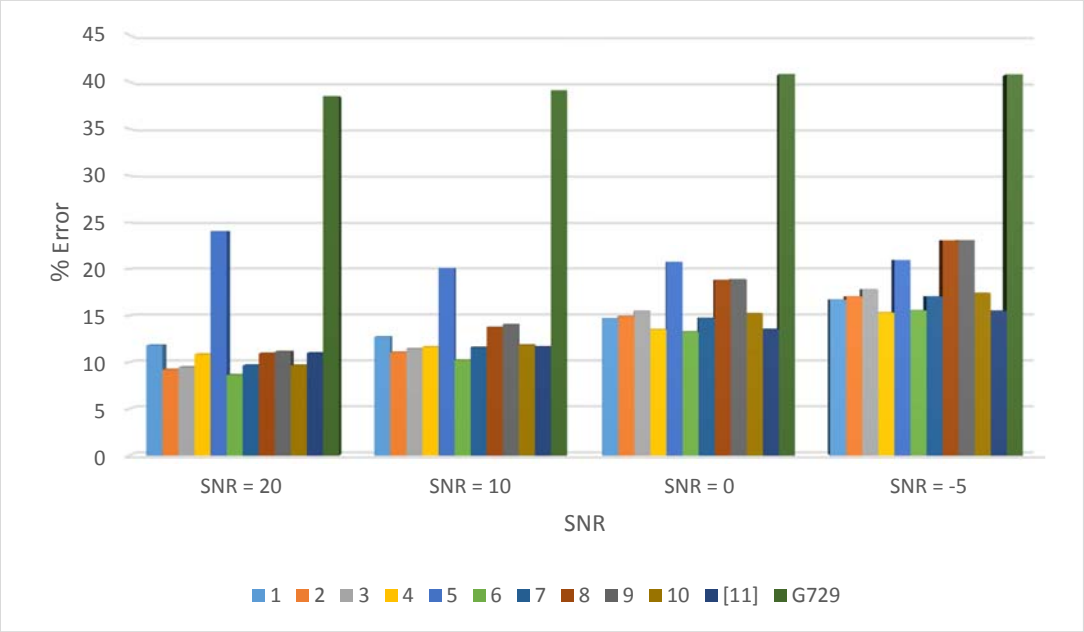
شکل (۵-۱۶) درصد خطای روش‌های ارائه شده در حضور نویز دستگاه تهویه

جدول (۵-۳) درصد خطای روش‌های ارائه شده در حضور نویز دستگاه تهویه

SNR / روش	۲۰	۱۰	۰	-۵	میانگین
پیشنهادی ۱	۱۳,۲۹	۱۹,۷۷	۳۴,۲۶	۴۰,۴۹	۳۰,۷۷
پیشنهادی ۲	۱۲,۰۴	۲۰,۲۶	۳۲,۸۷	۳۹,۷۶	۲۶,۲۴
پیشنهادی ۳	۱۲,۱۵	۲۰,۷۷	۳۳,۱۵	۴۰,۰۵	۲۶,۵۳
پیشنهادی ۴	۱۲,۲۰	۱۸,۸۹	۳۲,۵۹	۳۸,۰۳	۲۵,۴۳
پیشنهادی ۵	۲۱,۵۵	۲۳,۹۱	**۲۷,۳۴**	**۲۹,۷۸**	۲۵,۶۴
پیشنهادی ۶	**۹,۶۷**	**۱۵,۵۲**	۲۸,۵۲	۳۴,۹۹	**۲۲,۱۷**
پیشنهادی ۷	۱۰,۴۸	۱۶,۶۰	۲۹,۵۵	۳۵,۴۸	۲۳,۰۳
پیشنهادی ۸	۱۶,۱۶	۲۵,۸۶	۳۵,۸۴	۴۰,۳۰	۲۹,۵۴
پیشنهادی ۹	۱۵,۷۵	۲۴,۶۹	۳۵,۲۱	۴۰,۶۷	۲۹,۰۸
پیشنهادی ۱۰	۱۰,۴۷	۱۶,۷۱	۳۰,۱۴	۳۵,۴۲	۲۳,۱۹
مرجع [۱۱]	۱۲,۲۶	۱۸,۷۶	۳۳,۶۱	۳۹,۷۳	۲۶,۰۹
G729 [۴۱]	۴۰,۳۲	۴۱,۲۹	۴۳,۷۱	۴۷,۳۴	۴۳,۱۶

۵-۵-۵ آزمایش پنجم

این آزمایش نیز مانند آزمایش دوم می‌باشد با این تفاوت که در حضور نویز ماشین لباس‌شویی انجام می‌شود. نتایج حاصل از این آزمایش در جدول (۵-۴) قرار دارد. با توجه به نتایج این جدول در حالت‌های نسبت سیگنال به نویز ۲۰ و ۱۰ و ۰ روش پیشنهادی ششم و در حالت نسبت سیگنال به نویز ۵- روش پیشنهادی چهارم درصد خطای کمتری دارند. به طور میانگین روش پیشنهادی ششم عملکرد بهتری دارد. نمودار میله‌ای این نتایج در شکل (۵-۱۷) مشاهده می‌شود.



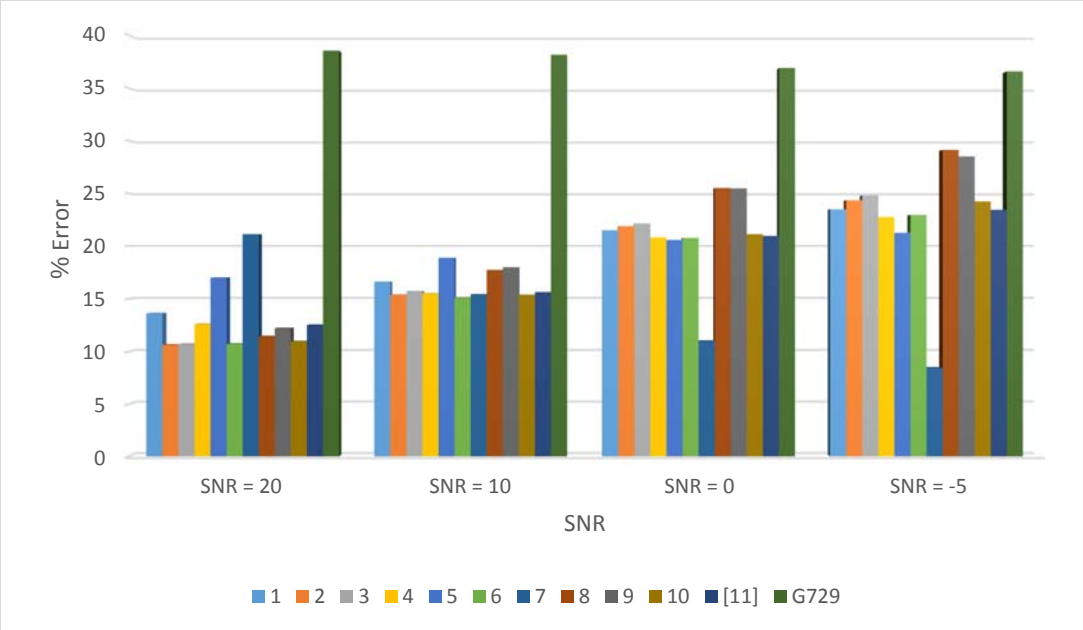
شکل (۵-۱۷) درصد خطای روش‌های ارائه شده در حضور نویز ماشین لباس‌شویی

جدول (۵-۴) درصد خطای روش‌های ارائه شده در حضور نویز ماشین لباسشویی

SNR / روش	۲۰	۱۰	۰	-۵	میانگین
پیشنهادی ۱	۱۱,۸۱	۱۲,۷۰	۱۴,۶۶	۱۶,۶۷	۱۳,۹۷
پیشنهادی ۲	۹,۲۰	۱۱,۰۵	۱۴,۹۰	۱۷,۰۲	۱۳,۰۴
پیشنهادی ۳	۹,۴۷	۱۱,۴۳	۱۵,۴۷	۱۷,۷۸	۱۳,۵۴
پیشنهادی ۴	۱۰,۸۵	۱۱,۶۴	۱۳,۴۹	**۱۵,۳۰**	۱۲,۸۳
پیشنهادی ۵	۲۴,۱۰	۲۰,۰۷	۲۰,۷۳	۲۰,۹۴	۲۱,۴۷
پیشنهادی ۶	**۸,۶۲**	**۱۰,۲۱**	**۱۳,۲۴**	۱۵,۵۴	**۱۱,۹۰**
پیشنهادی ۷	۹,۶۶	۱۱,۵۸	۱۴,۷۱	۱۷,۰۲	۱۳,۲۴
پیشنهادی ۸	۱۰,۹۳	۱۳,۷۳	۱۸,۷۸	۲۳,۱۱	۱۶,۶۴
پیشنهادی ۹	۱۱,۱۳	۱۴,۰۳	۱۸,۸۶	۲۳,۱۴	۱۶,۸۰
پیشنهادی ۱۰	۹,۶۵	۱۱,۸۳	۱۵,۲۱	۱۷,۴۱	۱۳,۵۳
مرجع [۱۱]	۱۰,۹۹	۱۱,۶۳	۱۳,۵۲	۱۵,۴۹	۱۲,۹۰
G729 [۴۱]	۳۸,۴۷	۳۹,۱۰	۴۰,۸۱	۴۰,۷۹	۳۹,۷۹

۵-۵-۶ آزمایش ششم

این آزمایش نیز مانند آزمایش دوم می‌باشد با این تفاوت که در حضور نویز همهمه انجام می‌شود. نتایج حاصل از این آزمایش در جدول (۵-۵) قرار دارد. با توجه به نتایج این جدول در حالت نسبت سیگنال به نویز ۲۰ دسی‌بل روش پیشنهادی دوم، در حالت نسبت سیگنال به نویز ۱۰ دسی‌بل روش پیشنهادی ششم و در حالت‌های نسبت سیگنال به نویز ۰ و ۵- دسی‌بل روش پیشنهادی پنجم درصد خطای کمتری دارند. به طور میانگین روش پیشنهادی ششم در مقایسه با سایر روش‌های پیشنهادی و روش مرجع [۱۱] و روش استاندارد G729 [۴۱] در حضور نویز همهمه کارایی بهتری دارد. همچنین نمودار این نتایج در شکل (۵-۱۸) مشاهده می‌شود.



شکل (۵-۱۸) درصد خطای روش‌های ارائه شده در حضور نویز همهمه

جدول (۵-۵) درصد خطای روش‌های ارائه شده در حضور نویز همهمه

SNR / روش	۲۰	۱۰	۰	-۵	میانگین
پیشنهادی ۱	۱۳,۶۱	۱۶,۵۷	۲۱,۵۳	۲۳,۵۱	۱۸,۸۰
پیشنهادی ۲	**۱۰,۶۳**	۱۵,۳۶	۲۱,۹۱	۲۴,۳۶	۱۸,۰۶
پیشنهادی ۳	۱۰,۷۰	۱۵,۶۸	۲۲,۱۷	۲۴,۸۲	۱۸,۳۴
پیشنهادی ۴	۱۲,۵۹	۱۵,۵۰	۲۰,۸۴	۲۲,۷۷	۱۷,۹۳
پیشنهادی ۵	۱۶,۹۹	۱۸,۸۵	**۲۰,۶۰**	**۲۱,۲۸**	۱۹,۴۳
پیشنهادی ۶	۱۰,۷۲	**۱۵,۰۹**	۲۰,۸۰	۲۲,۹۸	**۱۷,۴۰**
پیشنهادی ۷	۱۱,۰۰	۱۵,۴۰	۲۱,۱۵	۲۴,۰۸	۱۷,۹۰
پیشنهادی ۸	۱۱,۴۲	۱۷,۷۰	۲۵,۵۳	۲۹,۱۳	۲۰,۹۴
پیشنهادی ۹	۱۲,۲۱	۱۷,۹۴	۲۵,۴۸	۲۸,۵۲	۲۱,۰۳
پیشنهادی ۱۰	۱۰,۹۵	۱۵,۳۴	۲۱,۱۵	۲۴,۲۶	۱۷,۹۳
مرجع [۱۱]	۱۲,۵۱	۱۵,۵۹	۲۰,۹۸	۲۳,۴۶	۱۸,۱۴
G729 [۴۱]	۳۸,۵۴	۳۸,۱۵	۳۶,۹۰	۳۶,۵۷	۳۷,۵۴

۵-۵-۷ آزمایش هفتم

در این آزمایش پس از پیاده‌سازی روش‌های مطرح شده در فصل چهارم در نرم‌افزار متلب نسخه ۲۰۱۳، زمان محاسبه هر کدام را سنجیده‌ایم. این سنجش برای ۱۰ ثانیه از فایل انتخابی مربوط به آزمایش اول انجام شده است. نتایج حاصل از این آزمایش در جدول (۵-۶) مشاهده می‌شوند. با توجه به نتایج این جدول، روش‌های مبتنی بر ویژگی‌های کپسترال تاخیر کمتری دارند. روش پیشنهادی چهارم در مقایسه با سایر روش‌های پیشنهادی در زمان کمتری محاسبه می‌شود. این روش در مقایسه با روش مرجع [۱۱] و روش استاندارد G729 [۴۱] نیز زمان تاخیر کمتری دارد.

برای ارزیابی بهتر روش‌های ارائه شده از یک معیار دیگر به نام فاکتور بی‌درنگ استفاده کرده‌ایم. این فاکتور از رابطه (۵-۲) به دست می‌آید [۱۱]. در این رابطه t_{exe} زمان محاسبه هر یک از روش‌ها می‌باشد.

$$rt\_factor = \frac{10}{t_{exe}} \qquad (۵-۲)$$

مقدار فاکتور بی‌درنگ در روش پیشنهادی چهارم بیشترین مقدار را دارد. این مطلب بدین معناست که این روش در مقایسه با سایر روش‌ها زمان تاخیر کمتری دارد. از بین روش‌های مبتنی بر توزیع ویگنر ویل روش‌های پیشنهادی چهارم و پنجم فاکتور بالاتری در مقایسه با روش‌های پیشنهادی ششم تا دهم دارند ولی در مقایسه با روش‌های مبتنی بر کپسترال و روش مرجع [۱۱] و روش استاندارد G729 [۴۱] فاکتور بی‌درنگ آنها بسیار پایین می‌باشد. بنابراین در پردازش‌های بی‌درنگ روش پیشنهادی چهارم به دلیل داشتن زمان تاخیر کمتر، در مقایسه با سایر روش‌های پیشنهادی عملکرد بهتری دارد. لازم به ذکر است که این آزمایش روی یک سیستم با مشخصات زیر انجام شده است:

- CPU : Intel(R) Core(TM) i7-2630QM
- RAM : 6.00 GB

جدول (۵-۶) زمان محاسبه روش‌های ارائه شده برای ۱۰ ثانیه از فایل انتخابی دادگان فارس‌دات

زمان / روش	زمان محاسبه ۱۰ ثانیه از یک فایل انتخابی (ثانیه)	فاکتور بی‌درنگ
پیشنهادی ۱	۰,۴۳	۲۳,۲۶
پیشنهادی ۲	۰,۴۹	۲۰,۴۱
پیشنهادی ۳	۰,۳۷	۲۷,۰۳
پیشنهادی ۴	۰,۳۳	۳۰,۳۰
پیشنهادی ۵	۳۷,۸۶	۰,۲۷
پیشنهادی ۶	۳۷,۴۸	۰,۲۷
پیشنهادی ۷	۳۷,۷۶	۰,۲۶
پیشنهادی ۸	۳۸,۰۸	۰,۲۶
پیشنهادی ۹	۳۷,۹۰	۰,۲۶
پیشنهادی ۱۰	۳۷,۹۹	۰,۲۶
مرجع [۱۱]	۰,۳۵	۲۸,۵۷
G729 [۴۱]	۱,۱۶	۸,۶۲

۵-۶ جمع‌بندی

در ارتباط بین انسان و ربات، به منظور رسیدن به یک ارتباط پیشرفته و طبیعی نیاز به سیستمی مطمئن برای تشخیص گفتار از سایر صداهای موجود در محیطهای نویزی می‌باشد. لذا در این پایان‌نامه سعی ما بر این بوده است تا الگوریتم‌هایی را برای سیستم آشکارساز فعالیت گفتاری پیشنهاد دهیم که در برابر نویز مقاوم باشند.

روش‌های پیشنهادی به چند دسته تقسیم می‌شوند. دسته اول (روش‌های پیشنهادی اول تا چهارم) روش‌های مبتنی بر ضرایب کپسترال و انرژی هستند. در این روش‌ها از تأثیر ترکیب ویژگی‌های ضرایب پیش‌بینی ادراکی خطی، ضرایب کپسترال فرکانس مل ریشه، ضرایب کپسترال فرکانس بارک و ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده با ویژگی انرژی هر کدام به صورت جداگانه استفاده کرده‌ایم. کارایی این روش‌ها را در حالت بدون نویز و نیز در حضور انواع نویزهای سفید، دستگاه تهویه هوای مطبوع، ماشین لباس‌شویی و صدای همهمه با نسبت‌های سیگنال به نویز ۲۰، ۱۰، ۰ و ۵- دسی‌بل و با به کار بردن دادگان گفتاری استاندارد فارس‌دات سنجیده‌ایم. نتایج آزمایشات نشان داده‌اند که روش‌های پیشنهادی دسته اول در مقایسه با روش مرجع [۱۱] و الگوریتم استاندارد G729 [۴۱] درصد خطای کمتری دارند.

دسته دوم (روش پیشنهادی پنجم) روش مبتنی بر ویژگی ویگنر ویل است. این روش در حالت بدون نویز و در حضور نویز سفید گوسی در مقایسه با بقیه روش‌های پیشنهادی و روش مرجع [۱۱] و روش استاندارد G729 [۴۱] کمترین درصد خطا را دارد. همچنین در حضور نویز دستگاه تهویه هوای مطبوع و نویز همهمه در حالت‌های نسبت سیگنال به نویز ۰ و ۵- دسی‌بل درصد خطای کمتری دارد.

دسته سوم (روش‌های پیشنهادی ششم تا دهم) روش‌هایی هستند که از ترکیب روش‌های دو دسته اول و دوم به دست می‌آیند. در این روش‌ها با استفاده از به کارگیری ویژگی ویگنر ویل در ویژگی‌های ضرایب کپسترال و ویژگی انرژی توانستیم نتایج حاصل از روش‌های دسته اول را بهبود ببخشیم.

یکی از نوآوری‌های این پایان‌نامه استفاده از ضرایب پیش‌بینی ادراکی خطی، ضرایب کپسترال فرکانس مل ریشه، ضرایب کپسترال فرکانس بارک و ضرایب پیش‌بینی ادراکی خطی تجدید نظر شده در سیستم آشکارساز فعالیت گفتاری است. علاوه بر این استفاده از ویژگی ویگنر ویل با به کارگیری تبدیل هارتلی در سیستم آشکارساز فعالیت گفتاری برای اولین بار در این پایان‌نامه انجام شده است. همچنین ترکیب ویژگی‌های ذکر شده با ویژگی ویگنر ویل نیز از ابداعات دیگر این کار می‌باشد. همه روش‌های پیشنهادی در این پایان‌نامه در مقایسه با الگوریتم استاندارد G729 [۴۱] کارایی بهتری دارند. استفاده از پایگاه داده گفتاری فارسی زبان در سیستم آشکارساز فعالیت گفتاری برای نخستین بار در این پروژه صورت گرفته است.

۷-۵ پیشنهادات برای آینده

گرچه روش‌های مبتنی بر ویژگی ویگنر ویل در مقایسه با بسیاری از روش‌ها و روش‌های مبتنی بر ضرایب کپسترال ارائه شده در این پایان‌نامه، از کارایی نسبتا خوبی برخوردار هستند اما به دلیل داشتن هزینه محاسباتی بالا، در سیستم‌های بی‌درنگ مانند ارتباط بین انسان و ربات چندان مناسب به نظر نمی‌رسند، زیرا یکی از نیازهای این نوع سیستم‌ها انجام پردازش‌ها در حداقل زمان ممکن می‌باشد، در صورتی که زمان محاسبه توزیع ویگنر ویل برای هر ۱۰ ثانیه از یک فایل صوتی دادگان گفتاری فارس‌دات حدود ۳۸ ثانیه است. بنابراین زمان پاسخ سیستم بالا بوده و ربات عملکرد کندی در ارتباط با انسان خواهد داشت. لذا پیشنهاد ما برای آینده این است که با حفظ کارایی بالای این الگوریتم‌ها، در کاهش هزینه محاسباتی آن کوشش شود. این امر در سایه استفاده از برنامه‌نویسی توزیع شده[۱] یا موازی[۲] تحقق می‌پذیرد.

[۱] distributed

[۲] parallel

علاوه بر این، پیشنهاد دیگر ما به کار بردن سیستم محلی سازی منبع صدا[1] در کنار سیستم آشکارساز فعالیت گفتاری است. با استفاده از این سیستم، یک ربات قادر است بین صدای گوینده‌های محلی که روبه‌روی او قرار دارند و صدای سایر گوینده‌ها تمایز قائل شود. این امر در مورد ارتباط بین انسان و ربات از اهمیت بالایی برخوردار است.

[1] sound source localization

مراجع

[1] H. Kim, J. S. Choi, and M. Kim, "Human-Robot Interaction in Real Environments by Audio-Visual Integration," *International Journal of Control, Automation, and Systems,* vol. 5, pp. 61-69, 2007.

[2] R. Brueckmann, A. Scheidig, and H.-M. Gross, "Adaptive Noise Reduction and Voice Activity Detection for improved Verbal Human-Robot Interaction using Binaural Data," presented at the IEEE International Conference on Robotics and Automation, Roma,Italy, 2007.

[3] P. Y. Shih, P. C. Lin, and J. F. Wang, "Improving real-time voice activity detection for perceptual robotic control in noisy environment," presented at the TENCON 2011, Bali, Indonesia, 2011.

[4] H. D. Kim, K. K, and T. O, "Two-channel-based voice activity detection for humanoid robots in noisy home environments," presented at the IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 2008.

[5] I. Ali and G. Saha, "A Robust Iterative Energy Based Voice Activity Detector," presented at the Third International Conference on Emerging Trends in Engineering and Technology, Goa, India, 2010.

[6] J. Ramirez, J. Segura, C. Benitez, A. D. L. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication,* vol. 42, pp. 271–287, 2004.

[7] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication,* vol. 52, pp. 41-60, 2010.

[8] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *Selected Topics in Signal Processing,* vol. 4, 2010.

[9] J. Ramirez, J. Segura, C. Benitez, D. la, and A. Rubio, "Voice activity detection with noise reduction and long-term spectral divergence estimation," presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP 2004), 2004.

[10] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 27, pp. 113–120, 1979.

[11] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detection for speaker verification with noisy telephone and microphone data," in *ICASSP 2013*, Vancouver, BC, Canada, 2013.

[12] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Transactions on Consumer Electronics,* vol. 57, p. 196, 2011.

[13] L. Rabiner and R. Schafer, *Digital processing of speech signals*: Prentice Hall, 1978.

[14] A. Benyassine, E. Shlomot, H.-Y. Su, and E. Yuen, "A robust low complexity voice activity detection algorithm for speech communication systems," in *iSpeech Coding For Telecommunications Proceeding*, 1997.

[۱۵] م. فارسی‌نژاد، ب. زمانی‌دهکردی، ا. اکبری و ب. ناصرشریف، "پیشنهاد یک روش آشکارساز صوت دو مرحله‌ای مبتنی بر مدل مخفی مارکوف"، چهاردهمین کنفرانس سالانه انجمن کامپیوتر ایران، ۱۳۸۷.

[16] T. Fukuda, O. Ichikawa, and M. Nishimura, "Phone-duration-dependent long-term dynamic features for a stochastic model-based voice activity detection," presented at the Ninth Annual Conference of the International Speech Communication Association, 2008.

[17] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Consumer Electronics Audio, Speech, and Language Processing,* vol. 19, pp. 600-613, 2011.

[۱۸] ع. اسماعیلی، م. احدی، ا. فصیحی، "آشکارسازی فعالیت گفتاری به کمک بهسازی گفتار و استفاده از دیورژانس طیفی بلند مدت سیگنال"، دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران، ۱۳۸۵.

[19] L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Tech,* vol. 54, pp. 297–315, 1975.

[20] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA2006),* pp. 65–70, 2006.

[21] G. Evangelopoulos and P. Maragos, "Speech event detection using multiband modulation energy," presented at the Ninth European Conference on Speech Communication and Technology, 2005.

[22] S. Gerven and F. Xie, "A comparative study of speech detection methods," presented at the Fifth European Conference on Speech Communication and Technology, 1997.

[23] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.

[24] J. Ramirez, J. Segura, J. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEE Transactions on Audio, Speech, and Language Processing,,* vol. 15, p. 2177, 2007.

[25] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002.

[26] J. Chang and N. Kim, "Voice activity detection based on complex Laplacian model," *Electronics Letters,* vol. 39, pp. 632–634, 2003.

[27] J. Chang, J. Shin, and N. Kim, "Voice activity detector employing generalized Gaussian distribution," *Electronics Letters,* vol. 40, pp. 1561–1563, 2004.

[28] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," *IEEE Workshop Speech Coding for Telecommunications,* pp. 85–86, 1993.

[29] R. G. E. Nemer, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process,* vol. 9, pp. 217–231, 2001.

[30] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech Audio Process,* vol. 13, pp. 965–974, 2005.

[31] M. Farsinejad and M. Analoui, "A new robust voice activity detection method based on genetic algorithm," presented at the Telecommunication Networks and Applications Conference ATNAC 2008, 2008.

[32] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," presented at the 6th International Conference on signal processing, 2002.

[33] B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence classification with applications to speech recognition," *IEEE Trans. Acoust. Speech Signal Process,* vol. 24, pp. 201–212, 1976.

[34] T. Kinnunen, E. Chernenko, M. Tuononen, P. Frnti, and H. Li, "Voice activity detection using MFCC features and support vector machine," presented at the International Conerence on Speech and Computer, 2007.

[35] J. Wu and X. Zhang, "Maximum margin clustering based statistical VAD with multiple observation compound feature," *Signal Processing Letters, IEEE,* p. 1, 2011.

[36] F.Wang, B. Zhao, and C. Zhang, "Linear time maximum margin clustering," *IEEE Transactions on Neural Networks,* vol. 21, pp. 319–332, 2011.

[37] A. Martin and L. Mauuary, "Robust speech/non-speech detection based on LDAderived parameter and voicing parameter for speech recognition in noisy environments," *Speech communication,* vol. 48, pp. 191–206, 2006.

[38] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky*, et al.*, "Qualcomm-ICSI-OGI features for ASR," presented at the Seventh International Conference Spoken Language Processing, 2002.

[39] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of ITU-T/ETSI voice activity detectors," presented at the International Conference Acoust., Speech, Signal Process, 2001.

[40] T. Pham, C. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," presented at the International Conference on Computing and Communication Technologies RIVF'09, 2009.

[41] A. Benyassine, E. Shlomot, and H.-Y. Su, "ITU Recommendation G.729 Annex B: A Silence Compression Scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.,* vol. 35, pp. 64–73, 1997.

[42] G. 06.94, "Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector for Adaptive Multi Rate (AMR) Speech Traffic Channels " 1999.

[43] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler*, et al.*, "Speechdat-car: A large speech database for automotive environments," presented at the II LREC Conference, 2000.

[44] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka*, et al.*, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," *IEICE transactions on information and systems,* pp. 535–544, 2005.

[45] J. Garofolo. DARPA TIMIT: Acoustic-phonetic Continuous Speech Corps CD-ROM [Online].

[46] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," presented at the ICSLP00, 2000.

[47] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," 1992.

[48] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Trans. Speech Audio Processing,* vol. 13, pp. 965–974, 2005.

[49] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," *American Society for Engineering Education (ASEE) Conference Proceedings,* pp. 1-7, 2008.

[50] N. Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition," *INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY,* vol. 1, 2013.

[51] X. Wang and Z. Han, "A Novel Acoustic Feature Extraction Algorithm Based on Root Cepstrum Coefficients and CCBC for Robust Speech Recognition," pp. 643-647, 2008.

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

[53] L. Xie and Z.-Q. Liu, "A Comparative Study of Audio Features For Audio to Visual Conversion in MPEG-4 Compliant Facial Animation," *ICMLC, Dalian,* 2006.

[54] J. Najafi and H. Marvi, "PLP Based CELP Speech Coder," pp. 263-267, 2009.

[55] M. G. Sumithra and A. K. Devika, "A Study on Feature Extraction Techniques for Text Independent Speaker Identification," presented at the International Conference on Computer Communication and Informatics (ICCCI -2012) IEEE, Coimbatore, INDIA, 2012.

[56] P. Kumar, A. Biswas, A. N. Mishra, and M. Chandra, "Spoken Language Identification Using Hybrid Feature Extraction Methods," *JOURNAL OF TELECOMMUNICATIONS,* vol. 1, 2010.

[57] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Physical Review,* vol. 40, p. 749, 1932.

[58] L. Kaushik and D. O'Shaughnessy, "Voice activity detection using modified Wigner-Ville distribution," presented at the Interspeech 2008, 2008.

[59] W. J. Staszewski, K. Worden, and G. R. Tomlinson, "Time–frequency analysis in gearbox fault detection using the wigner–ville distribution and pattern recognition," *Mechanical Systems and Signal Processing,* vol. 11, pp. 673-692, 1997.

[60] A. P. Averchenko and B. D. Zhenatov, "Hartley Transform as Alternative to Fourier Transform in Digital Data Processing Systems," Omsk, Russia 2014.

[61] V. K. Singh, S. Gupta, and U. D. Dalal, "Performance Comparison of Discrete Hartley Transform (DHT) and Fast Fourier Transform (FFT) OFDM System in AWGN Channel," *International Journal of Computer Applications,* vol. 70, 2013.

[62] C. M. Bishop, *Pattern Recognition and Machine Learning, Springer Science + Business Media*. LLC, New York, 2006.

[63] M. Bijankhan, J. Sheikhzadegan, M. R. Roohani, and Y. Samareh, "FARSDAT - The Speech Database of Farsi Spoken Language," presented at the 5th Australian international conference, Speech science and technology 1994.
[64] *http://www.freesound.org/*.

Absrtact

The use of social robots in human's life has increased and the main way of Human Robot Interaction (HRI) is based on verbal communication. social robots are endowed with microphones to receive speech signal for interaction with people. In order to receive environmental noise when human's speech is recording, we need to have a system for detecting speech segments in recorded signals. Therefore, the goal of this work is to design a Voice Activity Detection (VAD) system to detect speech segments in noisy environment and increase the performance of the speech processing system in a social robot.

In this work, different features are proposed for extraction from speech for VAD system. These features are presented with combination of energy and other features such as Root Mel Frequency Cepstral Coefficients (RMFCC), Bark Frequency Cepstral Coefficients (BFCC), Perceptual Linear Prediction (PLP) and Revised Perceptual Linear Prediction (RPLP).

Another proposed method is based on Wigner Ville Distribution (WVD) as a time-frequency feature. It has a better performance in compared to other proposed method. Therefore, for improving the performance of cepstral-based methods, we combined them with WVD.

To evaluate the performance of these methods, we utilized FarsDat database as a Persian standard database. To compute robustness of proposed feature extraction methods to noise, we add different kinds of noise in different level of Signal to Noise Ratio (SNR).

Exprimental results show that some of the proposed feature extraction methods has the better performance in comparison with MFCC in different noisy environment.

Key Words: Voice Activity Detection, Wigner Ville Distribution, Root Mel Frequency Cepstral Coefficients, Bark Frequency Cepstral Coefficients, Perceptual Linear Prediction, Revised Perceptual Linear Prediction, energy, feature extraction, Human Robot Interaction, social robots.

Shahrood University of Technology

Faculty of Robotic Engineering

Time and frequency domain feature extraction from persian speech signals to improve performance of a Voice Activity Detection (VAD) system in human robot interaction

Fahimeh Jomhoori

Supervisor:

Hossein Marvi

Adviser:

Alireza Ahmadyfard

Date:

February 2015