پیشنهاد یک روش آشکارساز صوت دو مرحله‌ای مبتنی بر مدل مخفی مارکوف

محمد مهدی فارسی‌نژاد[1]، بهزاد زمانی دهکردی[1]، احمد اکبری[1]، بابک ناصر شریف[1و2]
[1] آزمایشگاه پردازش صدا و گفتار، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران
[2] گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه گیلان
mm_farsinejad@comp.iust.ac.ir, {bzamani, akbari, nasser_s}@iust.ac.ir

چکیده: آشکارساز صوت (Voice Activity Detection) ابزار مهمی برای افزایش کارائی روش‌های کدکردن گفتار، بهبود کیفیت گفتار و بازشناسی گفتار محسوب می‌شود. آشکارسازها به روشهای آستانه‌گذاری و روشهای مبتنی بر مدل تقسیم می‌شوند. روشهای آستانه‌گذاری کارائی ضعیفی در محیط نویزی دارند. از اینرو در مقاله حاضر یک الگوریتم VAD مبتنی بر مدل مخفی مارکوف پیشنهاد شده است که در دو مرحله عمل می‌کند. نخست با یک دسته‌بند (مدل مخفی مارکوف)، نوع نویز تشخیص داده می‌شود. در مرحله دوم، آشکارساز صوت مرتبط با آن نویز بکار می‌رود تا عملکرد بالاتری در محیط نویزی داشته باشد. ویژگیهای مورد استفاده در این روش، بردار ۳۹ بعدی شامل لگاریتم انرژی، ۱۲ ضریب MFCC و مشتقات مرتبه اول و دوم آنها می‌باشد. عملکرد الگوریتم پیشنهادی بر روی دادگان TIMIT مورد ارزیابی قرار گرفته است. بر اساس نتایج بدست آمده روش پیشنهادی نسبت به روشهای دیگر عملکرد قابل قبولی از خود نشان داده است.

واژه های کلیدی: آشکارساز صوت، مدل مخفی مارکوف، تشخیص-دهنده نوع نویز، استخراج ویژگی، آستانه‌گذاری.

۱- مقدمه

یکی از مشکلات اصلی در اکثر سیستم‌های پردازش گفتار، محیط-های نویزی و تأثیر مخرب آنها بر روی کارائی سیستمهای پردازش گفتار است [4, 5] . در اکثر کاربردها نیاز است که محدوده حضور گفتار مشخص باشد. بعنوان مثال در سیستم بازشناسی گفتار در صورتی که ابتدا و انتهای گفتار مشخص باشد، عملکرد آن بهتر می‌شود. همچنین در سیستم‌های بهبود کیفیت گفتار برای تخمین نویز زمینه نیاز است که از نواحی‌ای که فقط نویز وجود دارد استفاده شود. [6] الگوریتم‌های تشخیص حضور گفتار با چالش‌های فراوانی مانند حضور نویز و بار محاسباتی تحمیلی به سیستم روبرو هستند. روشهای متعددی جهت کاهش تأثیر نویز روی کارایی آنها ارائه شده است که اغلب نیازمند یک تخمین آماری از نویز می‌باشند که این نیز به عملکرد خود آشکارساز حضور گفتار بر می‌گردد.

در دهه‌های اخیر روشهای مختلفی برای آشکارسازی صوت به ویژه در شرایط نویزی شدید مطرح شده‌اند که می‌توان آنها را به دو دسته کلی تقسیم نمود: روشهای آستانه‌گذاری و روشهای مبتنی بر شیوه‌های آماری و مدل کردن سیگنال و نویز عمل می‌کنند.

بطور معمول الگوریتم VAD از دو قسمت استخراج ویژگی و تصمیم‌گیری تشکیل می‌شود. در بخش استخراج ویژگی از سیگنال گفتار ویژگیهایی مانند انرژی، نرخ عبور از صفر و ضرایب کپسترال استخراج می‌شود و سپس واحد تصمیم‌گیری با استفاده از یک روش دسته‌بندی یکی از دو حالات گفتار یا سکوت را به سیگنال نسبت می-دهد. در مقالات از آستانه‌گذاری [3]، نسبت درست‌نمائی (Likelihood Ratio) و مدل مخفی مارکوف [10] بعنوان کلاس‌بند بهره جسته‌اند.

Kristjansson و همکارانش از ویژگیهای مبتنی بر تابع خودهمبستگی و ضرایب مل کپسترال [7] استفاده کرده‌اند. از دیگر ویژگیها می‌توان انرژی [10]، گام(Pitch) صدا [9]، پریودیک بودن و ویژگیهای آماری با درجه بالا [8] را نام برد. البته ویژگی‌های متعارف مانند انرژی و نرخ عبور از صفر، در حضور نویز مقاوم نیستند. [2] روشهای فوق نویز زمینه را ایستان فرض می‌کنند به همین دلیل در حضور نویزهای غیر ایستان عملکرد ضعیفی از خود نشان می‌دهند.

در مقاله حاضر، یک روش جدید برای تشخیص نویز و گفتار بر اساس مدل مخفی مارکوف ارائه شده است. روش پیشنهادی با بکارگیری تشخیص‌دهنده نوع نویز، ابتدا نوع نویز زمینه را تعیین می-نماید و سپس آشکارساز صوت مرتبط با آن نویز را انتخاب می‌کند. به این ترتیب مدلهای سیستم متناسب با ماهیت و رفتار نویز انتخاب می-شوند. بر همین اساس انتظار می‌رود که این روش عملکرد بالاتری در محیطهای نویزی متنوع با نرخهای سیگنال به نویز مختلف داشته باشد.

ساختار ادامه مقاله به این شکل است. بخش دوم به بررسی روشهای آشکارسازی مبتنی بر مدل و آستانه‌گذاری می‌پردازد. بخش سوم مراحل روش پیشنهادی را شرح می‌دهد. در بخش چهارم دادگان بکار رفته در آزمایشات و روشهای ارزیابی توضیح داده شده‌اند. همچنین نتایج آزمایشات بیان شده و مورد بررسی قرار گرفته‌اند. در انتها نیز نتیجه‌گیری کلی آورده شده است.

۲- روشهای آشکارسازی صوت متداول

در این بخش به بررسی اجمالی روشهای متداولی می‌پردازیم که برای ارزیابی و مقایسه در این مقاله مورد استفاده قرار گرفته‌اند. ادامه این بخش به معرفی ساختار VAD مبتنی بر مدل آماری و سپس VAD با آستانه‌گذاری اختصاص دارد.

۲-۱ VAD مبتنی بر مدل آماری

در VAD مبتنی بر مدل مخفی مارکوف، جداسازی فریم‌های گفتار و غیرگفتار بر اساس آزمون نسبت شباهت (LRT) می‌باشد[2]. اگر هر یک از دو کلاس گفتار و غیرگفتار را بوسیله یک دنباله از بردارهای ویژگی یا مشاهدات مانند $O_{0:t} = \{o_0,.....o_t\}$ نشان دهیم. آنگاه حل مسأله آشکارسازی گفتار/غیرگفتار با استفاده از مدل مخفی مارکوف در واقع محاسبه رابطه (۱) می‌باشد.

$$\arg\max_i \{ p(M_i \mid O_{0t})\} \qquad (۱)$$

که M_i یکی از دو کلاس گفتار یا غیرگفتار می‌باشد. از آنجایی که احتمال رابطه (۱) به صورت مستقیم قابل محاسبه نمی‌باشد، با استفاده از قانون Bayes به صورت رابطه (۲) بازنویسی می‌گردد.

$$p(M_i \mid O) = \frac{P(O \mid M_i)\, P(M_i)}{P(O)} \qquad (۲)$$

در این حالت برای هر یک از دو کلاس گفتار و غیرگفتار یک مدل مارکوف هستند. در این مدل احتمال توأم که بوسیله دنباله o_t و با حرکت در میان دنباله حالتها، ایجاد می‌شود؛ با حاصلضرب احتمالات گذار در احتمالات خروجی به صورت رابطه (۳) محاسبه می‌شود.

$$p(O, X \mid M) = a_{12} b_2(o_1) a_{22} b_2(o_2) a_{23} b_3(o_3) .... \qquad (۳)$$

که در این مدل a_{ij}، $b_j(o_t)$ و o_t به ترتیب احتمال گذار از حالت i به حالت j، احتمال خروجی در حالت j و بردار L بعدی از سیگنال مشاهده شده در فریم tام می باشند. بدلیل نامعلوم بودن حالتهای X، احتمال هر کلاس با مجموع همه دنباله حالتهای ممکن، X =x(1),x(2),x(3),…,x(T) به صورت رابطه (۴) محاسبه می‌شود.[2]

$$p(O \mid M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t)\, a_{x(t)x(t+1)} \qquad (۴)$$

که x(0) به عنوان حالت ورودی مدل و x(T+1) به عنوان حالت خروجی مدل در نظر گرفته می‌شوند. بنابراین با استفاده از این مدل، جداسازی فریم های گفتار از غیرگفتار با تخمینی از بیشترین نرخ شباهت دنباله حالتها، با رابطه (۵) قابل محاسبه می‌باشد.

$$\hat{P}(O \mid M) = \max_X \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t)\, a_{x(t)x(t+1)} \right\} \qquad (۵)$$

که احتمال خروجی $b_j(o_t)$ بوسیله توزیع گوسین چند متغیره به فرم رابطه (۶) بدست می‌آید.[2]

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm}\, N(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \qquad (۶)$$

که M_s تعداد مؤلفه‌های مخلوط گوسین و c_{jsm} وزن مؤلفه mام و $N(.;\mu,\Sigma)$ یک گوسی چند متغیره با بردار میانگین μ و ماتریس کواریانس Σ می‌باشد. در این مقاله دو VAD مبتنی بر مدل مخفی مارکوف آموزش داده شده با دادگان تمیز و نویزی مورد ارزیابی قرار گرفته‌اند.

۲-۲ VAD مبتنی بر حد آستانه

یکی دیگر از روش های آشکارساز حضور گفتار روش آستانه‌گذاری می‌باشد. در این روش ویژگیها با حدهای آستانه مقایسه شده و تصمیم-گیری انجام می‌گیرد. حدهای آستانه بطور معمول تجربی بدست می-آیند و با تغییر محیط کارائی خود را از دست می‌دهند. ویژگیهایی که در روشهای حد آستانه مورد استفاده قرار گرفته‌اند اغلب انرژی و نرخ عبور از صفر می‌باشند.[3]

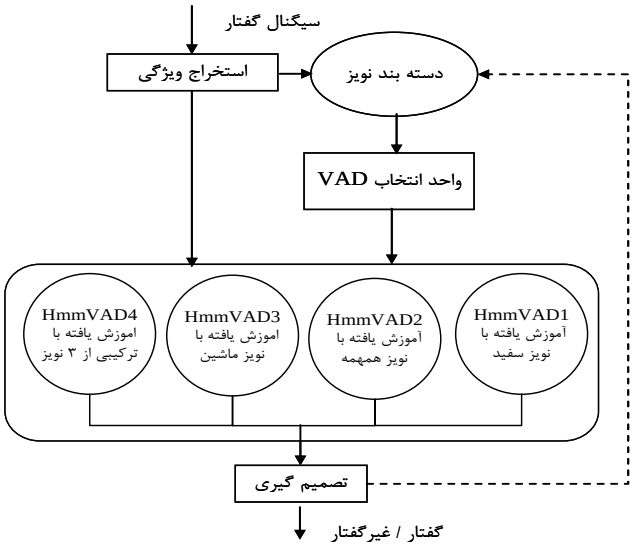
در [3] با کمک بانک فیلتر ویولت، تخمینی از سیگنال گفتار در تمام زیرباندها بدست آورده می‌شود. سپس انرژی سیگنال گفتار تخمین زده شده برای تمام زیرباندها را با هم جمع کرده و بدین ترتیب تخمینی از انرژی سیگنال گفتار در سطح فریم بدست می‌آوریم. اگر این مقدار تخمین زده شده از یک حد آستانه بیشتر باشد، فریم جاری به عنوان گفتار و در غیر اینصورت به عنوان نویز، در نظر گرفته می‌شود. از این VAD نیز به عنوان VAD پایه برای مقایسه عملکرد روش پیشنهادی استفاده می‌گردد.

۳- الگوریتم آشکارساز صوت پیشنهادی

در بخش‌های قبل به این نکته اشاره شد که روش‌های آشکارساز حضور گفتار به حضور نویز حساس هستند. از اینرو روش‌هایی مانند آموزش با دادگان نویزی یا استفاده از پیش پردازش حذف نویز در مقالات متعددی مطرح گردیده است. این روش‌ها نیز اکثراً در محیط-هایی که نویز غیر ایستان است، عملکرد ضعیفی دارند. الگوریتم پیشنهادی در این مقاله سعی کرده است تا با مکانیزم تعیین نوع نویز عملکرد VAD را در این شرایط بالا ببرد.

روش پیشنهادی مشتمل بر دو مرحله می‌باشد. ابتدا با استفاده از یک تشخیص دهنده نوع نویز، نوع نویزی که سیگنال گفتار را آلوده کرده است، تعیین می‌شود. سپس با توجه به نوع نویز، VAD مربوط به آن برای مرحله بعد که همان آشکارسازی حضور نویز است، انتخاب می-شود. شکل (۱) شمای کلی الگوریتم پیشنهادی را نشان می‌دهد. فیدبک نشان داده شده با خط‌چین در این الگوریتم نشان‌دهنده قابلیت تطبیق-پذیر بودن آن در کاربردهای برخط می‌باشد. نتایج بدست آمده در این مقاله در حالت برون‌خط (Offline) می‌باشد. در مرحله شروع الگوریتم، با فرض فقط نویز بودن چند فریم ابتدایی سیگنال ورودی، دسته‌بند نویز مبتنی بر HMM نوع نویز را تشخیص می‌دهد. سپس در مرحله انتخاب VAD، لگاریتم احتمال درست‌نمائی حاصل از مدل نویز انتخاب شده با یک حد آستانه که به صورت تجربی بر اساس سه نوع نویز مورد آزمایش بدست‌آمده مقایسه می‌گردد. اگر شرط حد آستانه لگاریتم احتمال درست‌نمائی را برآورده نماید، آشکارساز صوت مبتنی بر مدل

مربوط به همان نویز (مدلها با دادگان نویزی آلوده به همان نویز آموزش داده شده‌اند) انتخاب شده و در غیر این صورت آشکارساز صوت مبتنی بر مدلی انتخاب می‌شود که با دادگان آلوده به انواع نویز آموزش دیده است. اضافه کردن این حالت برای برخورد مناسب با شرایط نویزی ناآشنا برای سیستم می‌باشد



شکل۱: ساختار الگوریتم آشکارسازی صوت پیشنهادی

۴- ارزیابی و نتایج

۴-۱ دادگان و معیارهای ارزیابی

برای مقایسه عملکرد الگوریتمهای VAD در شرایط مختلف نویزی از دادگان TIMIT استفاده شده است. دادگان آموزشی شامل ۴۶۲۰ جمله و دادگان تست شامل ۱۳۴۴ جمله می‌باشد. برای بدست آوردن دادگان نویزی، نویزهای سفید، همهمه و ماشین با نرخهای سیگنال به نویز 0dB، 5dB، 10dB و 20dB از دیتابیس Noisex-92 انتخاب و به طور مصنوعی به مجموعه دادگان تمیز TIMIT اضافه می‌شوند. برای بررسی عملکرد الگوریتمهای VAD از معیارهای نرخ اشتباه در رد (False Rejection Rate(FRR) و نرخ اشتباه در اعلام (False Alarm Rate(FAR)) استفاده خواهد شد. روابط (۷) و (۸) نحوه محاسبه این دو معیار را نشان می‌دهند. [6]

$$FRR = \frac{N_{FR}}{N_S} \times 100 \qquad (۷)$$

$$FAR = \frac{N_{FA}}{N_{nS}} \times 100 \qquad (۸)$$

که N_s، N_{ns}، N_{FR} و N_{FA} به ترتیب تعداد کل فریم‌های گفتار، تعداد کل فریم‌های غیرگفتار، تعداد فریم‌های گفتار که غیرگفتار تشخیص داده شده‌اند و تعداد فریم‌های غیرگفتاری که گفتار تشخیص داده شده‌اند، می‌باشند.

۴-۲ نتایج آزمایشات

مدلهای در نظر گرفته شده در کلیه‌ی آزمایشات دارای ۵ حالت و هر حالت دارای ۸ مخلوط گاوسی می باشند. در ضمن طول فریم‌ها ۲۵ میلی ثانیه و میزان شیفت فریم ۱۰ میلی ثانیه در نظر گرفته شده است. ویژگی‌های مورد استفاده شامل لگاریتم انرژی، ۱۲ ضریب MFCC و مشتقات مرتبه اول و دوم آنها می‌باشد

روش پیشنهادی با سه روش پایه (روش آستانه‌گذاری بهمراه تخمین نویز پیوسته[3]، روش VAD مبتنی بر مدل آموزش داده شده با دادگان نویزی و روش VAD مبتنی بر مدل آموزش داده شده با دادگان تمیز) مقایسه شده است. جدول (۱) نحوه نام‌گذاری روشها را نشان می‌دهد.

جدول ۱: نام‌گذاری روشهای مورد ارزیابی

نام روش	توضیحات
Proposed_VAD	روش پیشنهادی
HmmVAD_NT	VAD مبتنی بر مدل آموزش یافته با دادگان نویزی
HmmVAD_CT	VAD مبتنی بر مدل آموزش یافته با دادگان تمیز
Threshold_VAD	VAD مبتنی بر حد آستانه با تخمین نویز پیوسته

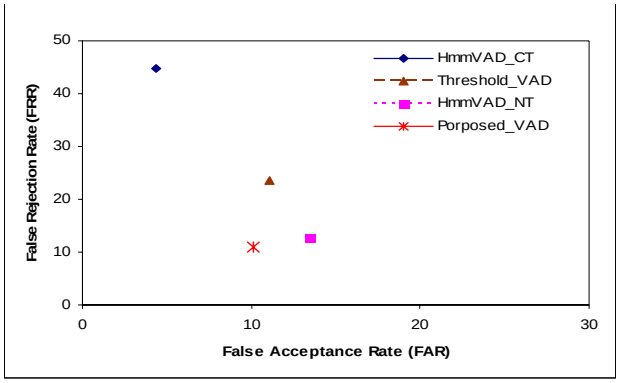
در روش Threshold_VAD، 20dB نویز از سیگنال نویزی حذف کرده سپس با اعمال یک حد آستانه تجربی در مورد سکوت یا عدم سکوت تصمیم‌گیری می‌شود. در روش HmmVAD_CT دو مدل سکوت و گفتار در نظر گرفته شده که با دادگان تمیز آموزش داده می‌شوند. در روش HmmVAD_NT دو مدل نویز و گفتار در نظر گرفته شده که با ترکیبی از دادگان نویزی آموزش داده می‌شوند. در روش پیشنهادی در مرحله اول که شامل شناسایی نوع نویز می‌باشد، سه مدل نویز سفید، همهمه و ماشین تعریف شده و در مرحله بعد توسط الگوریتم انتخاب، VAD متناسب با نوع نویز، یکی از روشهای VAD دو مدله نویز و گفتار انتخاب می‌گردد.

شکل (۲) نمودار ROC (تغییرات FAR را بر حسب FRR) میانگین‌گیری شده روی انواع نرخ سیگنال به نویز و انواع نویز را نشان می‌دهد. شکل۲-الف نمودار ROC را برای روشهای مختلف در حضور نویز سفید میانگین گرفته شده روی انواع نرخ سیگنال به نویز نشان می‌دهد. چنانچه در شکل دیده می‌شود روش پیشنهادی به مرکز نزدیکتر است که کارائی بهتر روش را نشان می‌دهد.

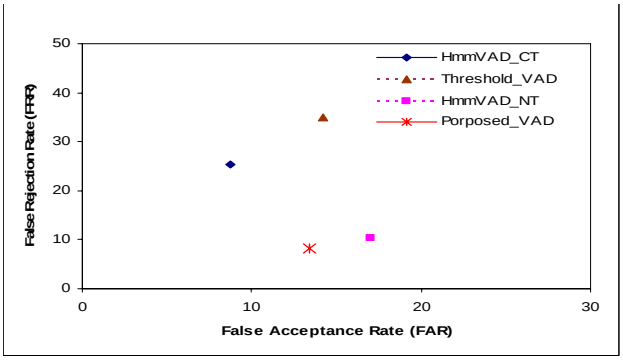
شکل ۲-ب تغییرات FAR بر حسب FRR را در حضور نویز همهمه نشان می‌دهد. طبق شکل روش پیشنهادی در حضور این نویز هم عملکرد بهتری نسبت به دیگر روشها دارد. در حضور این نویز عملکرد روش آستانه‌گذاری، بدلیل تخمین نادرست نویز، رفتار ضعیفتری نسبت به شرایط نویز سفید دارد.

شکل ۲-ج تغییرات FAR بر حسب FRR را در حضور نویز ماشین نشان می‌دهد. از آنجایی که ماهیت این نویز ایستان می‌باشد،

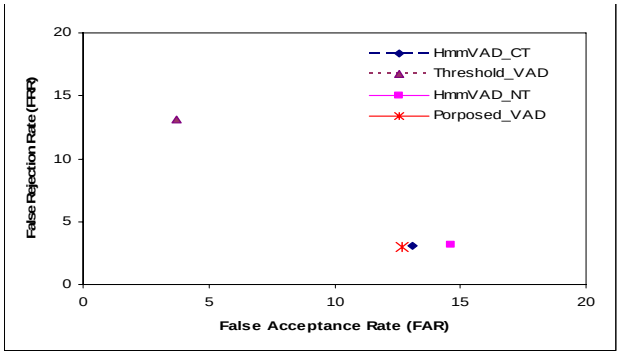
برای همه روشها نتایج بهتری نسبت به نویزهای سفید و همهمه حاصل شده‌است. در حضور این نویز عملکرد بهتر روش پیشنهادی نسبت به دیگر روشها آنچنان محسوس نمی‌باشد.



الف– نویز سفید



ب– نویز همهمه



ج– نویز ماشین

شکل ۲: نمودار ROC میانگین‌گیری شده روی انواع مختلف نویز و SNR

برای بررسی تمامی شرایط نویزی و میزان بهبودی روش پیشنهادی از معیار بهبودی نسبی، که به صورت رابطه (۹) تعریف می‌شود، استفاده می‌گردد[1].

$$\%\,\mathrm{Improvment}=\frac{(FAR+FRR)_{Baseline}-(FAR+FRR)_{method}}{(FAR+FRR)_{Baseline}}\times 100 \qquad (۹)$$

با توجه به رابطه فوق روش پیشنهادی نسبت به روش Threshold_VAD و HmmVAD_NT به HmmVAD_CT ترتیب۸۶ /۴۱ و ۱۷/۹۳ و ۴۱/۱۶ درصد بهبودی داشته‌است.

۵- .نتیجه گیری

در این مقاله یک روش آشکارسازی صوت دو مرحله‌ای ارائه گردید. در این روش پس از تشخیص نوع نویز در مرحله اول، آشکارساز مبتنی بر مدل آن نوع نویز انتخاب می‌شود. نتایج ارزیابی این روش آشکارسازی صوت نشانگر عملکرد مناسب آن نسبت به روشهای آشکارساز آستانه‌گذاری و روشهای VAD مبتنی بر مدل تک مرحله‌ای است. میانگین میزان خطای FRR و FAR بر روی نویزها با نرخ سیگنال به نویزهای مختلف به ترتیب عدد ۷/۴۳ و ۱۲/۰۷ می‌باشد. برای بررسی تمامی شرایط نویزی و میزان بهبودی روش پیشنهادی از معیار بهبودی نسبی استفاده گردید. روش پیشنهادی بر روی مجموع میانگین FAR و FRR در شرایط نویزی با نرخ سیگنال به نویزهای مختلف نسبت به روش Threshold_VAD پایه، ۴۱/۸۶ درصد و نسبت به روش HmmVAD_NT پایه، ۱۷/۹۳ درصد و نسبت به روش HmmVAD_CT پایه، ۴۱/۱۶ درصد بهبودی داشته‌است.

مراجع

[۱] سلیمانی سیده اکرم، احدی سیدمحمد، " **آشکارسازی فعالیت گفتاری با استفاده از وزن‌دهی چندین ویژگی به کمک آنالیز تفکیک خطی(LDA)** "، سیزدهمین کنفرانس انجمن ملی کامپیوتر ایران، ۱۳۸۷

[2] Sohn J., Kim N. S. and Sung W., *"A statistical model-based voice activity detection,"* IEEE Signal Process Letters, vol. 6, pp. 1-3, 1999

[3] Mohammdi M., Nasersharif B., Rahmani M. and Akbari A., "*The New Sub-band Level Voice Activity Detector Based on Wavelet Transform*," ICEE, 2007.

[4] Ishizuka K., Nakatani T., Fujimoto M., and Miyazaki N., "*Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio*," Interspeech, pp.230-233, 2007.

[5] Fujimoto M., Ishizuka K., and Nakatani T., "*A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme*," ICASSP, pp.4441-4444, 2008.

[6] Fujimoto M. and Ishizuka K., "*Noise robust voice activity detection based on switching Kalman filter*," IEICE Transaction on Information and Systems, Vol.E91-D, No.3, pp.447-467, 2008.

[7] Kristjansson T., Deligne S., and Olsen P., *"Voicing features for robust speech detection,"* Proc. Interspeech, pp. 369-372, 2005.

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

[9] ETSI standard document, ETSI ES 202 050 V1.1.3, 2003.

[10] Basu S., *"A linked-HMM model for robust voicing detection and speech detection,"* ICASSP, pp. 816-819, 2003.