ارائه الگوریتمی مبتنی بر روش انرژی– انتروپی برای تشخیص محدوده گفتار در محیط‌های نویزی

حمید دهقانی[۱]، علی مهاجری[۲]
۱- دانشگاه صنعتی مالک اشتر، hamid_deh@yahoo.com
۲- دانشجوی دانشگاه آزاد اسلامی واحد بوشهر، ali.mohajeri@gmail.Com

چکیده

در یک سیستم بازشناسی گفتار کلمات گسسته یا پیوسته، بازدهی تشخیص محدوده گفتار، تاثیر مستقیمی روی کل سیستم بازشناسی دارد. معمولاً انرژی، و نرخ عبور از سطح صفر، به‌عنوان دو پارامتر در تشخیص نقاط ابتدا و انتهای گفتار استفاده می‌شوند. در محیط‌های نویزی پارامتر نرخ عبور از سطح صفر، غیرقابل‌اعتماد است و باعث خطا در کل سیستم بازشناسی می‌شود. در اکثر روش‌ها، نویز زمینه از روی فریم‌های اولیه گفتار تخمین زده می‌شود. در الگوریتم ارائه شده با استفاده از تغییراتی که در انرژی اصلاح شده تیگر می‌دهد، از آن در تشخیص نقاط اولیه ابتدا و انتهای گفتار استفاده می‌شود، سپس با استفاده از اصلاحات مؤثری که در الگوریتم انرژی- انتروپی انجام شده، نقاط دقیقتر گفتار را می‌توان تشخیص داد. در این الگوریتم، حتی با وجود نویز در ابتدای گفتار، محدوده گفتار تشخیص داده می‌شود. از این الگوریتم، برای تشخیص محدوده گفتار های فارسی استفاده شده است. نتایج آزمایشات نشان‌دهنده بازدهی مناسب و قابل قبول روش ارائه شده در زمینه تشخیص گفتار از زمینه نویز و سکوت می‌باشد.

واژه‌های کلیدی

انرژی اصلاح شده تیگر، تشخیص نقاط ابتدا و انتهای گفتار، ویژگی انرژی- انتروپی، محیط‌های نویزی، نویز نفس کشیدن

۱– مقدمه

اولین گام در یک سیستم بازشناسی گفتار گسسته، آشکارسازی دقیق نقاط ابتدا و انتهای گفتار، از زمینه سکوت و نویز است. در سیستم‌های بازشناسی گفتار، تشخیص نادرست محدوده گفتار، اثرات منفی مانند افزایش خطای بازشناسی و افزایش حجم محاسبات در کل سیستم بازشناسی دارد. اکثر الگوریتم‌های تعیین محدوده گفتار، مبتنی بر اندازه‌گیری دو ویژگی نرخ عبور از سطح صفر و انرژی سیگنال صوتی است [۱،۳]. سایشی‌های ضعیف بدون صدا، نظیر/ ف، س، ش، ه / انفجاری‌های ضعیف مانند / پ ، ت ، ک / و خیشومی‌های انتهایی مانند/ م، ن / به‌عنوان بعضی واج‌های کم انرژی در زبان فارسی هستند که بازشناسی آنها ممکن است همراه با خطا باشد. نرخ عبور از سطح صفر، در مورد حروف واکه دار، مقدار کمتری نسبت به سایشی‌های ضعیف بدون واکه و یا انفجاری‌ها دارد، که پس از تعیین نقاط اولیه گفتار، پارامتر مناسبی در تشخیص دقیقتر شروع و پایان گفتار است [۱]. در محیط‌های نویزی پارامتر نرخ عبور از سطح صفر، نمی‌تواند خیلی قابل اعتماد باشد. تیگر الگوریتم جدیدی را برای محاسبه کردن انرژی سیگنال پیشنهاد کرد که این الگوریتم تحت عنوان الگوریتم انرژی تیگر ارائه شد. با اصلاحاتی که در این الگوریتم انجام شد، جهت تشخیص محدوده

گفتار مورد استفاده قرار گرفت. در الگوریتم انرژی تیگر، سیگنال گفتار نه تنها از دامنه، بلکه از فرکانس هم تاثیر می‌پذیرد و در قسمت‌هایی از گفتار که ضعیفتر هستند، ولی مؤلفه فرکانسی متفاوتی نسبت به نویز زمینه دارند، مؤثر عمل می‌کند [۲].

روش انرژی–انتروپی، روی قسمت‌هایی از گفتار که تغییرات طیفی بالایی دارند، تاکید می‌کند. انتروپی، با اندازه‌گیری اطلاعات ناخواسته، شیوه مؤثری در جایگزینی پارامتر نرخ عبور از صفر می‌باشد. در این مقاله با استفاده از روش مبتنی بر انرژی اصلاح‌شده تیگر، نقاط اولیه شروع و انتهای گفتار تعیین می‌شوند. سپس با تغییراتی در الگوریتم انرژی–انتروپی و نرمالیزه‌کردن منحنی‌های اصلاح‌شده انرژی تیگر و انرژی–انتروپی و تعیین آستانه‌های مناسب، می‌توان محدوده گفتار را بدون تخمین زدن از روی فریم‌های اولیه سیگنال به‌صورت دقیقتر تعیین نمود. در این مقاله این نوع اندازه‌گیری جدید انرژی در گفتار فارسی بررسی می‌شود. در قسمت دوم روی حذف نویز ناشی از بکارگیری مبدل A/D و تجهیزات ضبط صدا و غیره، همچنین استفاده از یک الگوریتم مؤثر و بهینه در پیاده‌سازی آن بحث می‌شود. در قسمت سوم روی توصیف الگوریتم، محاسبات الگوریتم اصلاح شده انرژی تیگر، الگوریتم انرژی- انتروپی و همچنین پیشنهادهای انجام شده در بهبود الگوریتم و نحوه تعیین محدوده گفتار بحث شده است. بخش چهارم نتایج قابل قبول و ارزیابی پیاده سازی این الگوریتم‌ها را روی سیگنال گفتار فارسی مورد بررسی قرار داده است. در بخش پنجم نتیجه‌گیری‌های انجام شده، ارائه شده است.

۲– فیلتر پیش تاکید

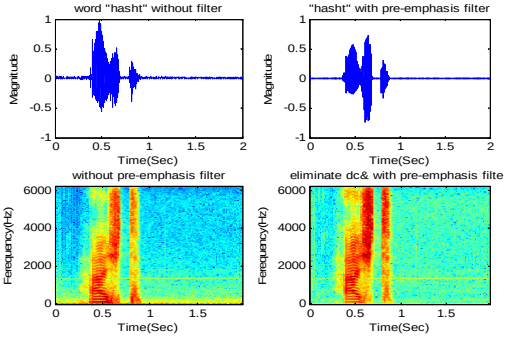
ابتدا جبران‌سازی DC را که ممکن است از طریق میکروفن یا هر مبدل دیگری ایجاد شده باشد، با گرفتن میانگین سیگنال گفتار و کم کردن آن از کل سیگنال حذف می‌شود. جهت حذف نویز زمینه تداخلی که فرکانس آن زیر ۱۰۰ هرتز می‌باشد و همچنین پهن کردن طیف فرکانس سیگنال گفتار، از یک فیلتر دیجیتال FIR مرتبه اول استفاده شده می‌شود [۲]. یک روش بهینه و مؤثر در حجم محاسبات آن به‌صورت معادلات می‌باشد. محاسبات فیلتر پیش تأکید مرتبه اول با انجام تغییرات اندک، ساده‌تر می‌شود، ضریب a معمولاً عددی بین ۹۰. و ۱ انتخاب می‌شود، این ضریب در شبیه‌سازی‌های انجام شده در این مقاله ۹۷. (معادل با ۳۱/۳۲) درنظر گرفته شده است. معادله (۱) یک فیلتر دیجیتال مرتبه اول را نشان می‌دهد که در آن s_n نمونه n ام در یک دنباله می‌باشد. این معادله را می‌توان با معادله (۲) جایگزین نمود که در آن عملیات ضرب با یک جمع ساده و شیفت جایگزین شده است. تقسیم بر ۳۲ با شیفت دادن ۵ بیت به سمت راست ایجاد می‌شود[۶].

$$H(z) = 1-az^{-1} \quad (۱)$$

$$S_n = s_n - \frac{31}{32} s_{n-1}$$

$$S_n = s_n - (s_{n-1} - \frac{1}{32} s_{n-1}) \quad (۲)$$

شکل (۱) شکل موج نطق هشت و طیف فرکانسی آن را قبل و بعد از حذف DC و استفاده از فیلتر پیش تأکید نشان می‌دهد، واضح است که نویز زمینه مخصوصاً نویز ۵۰-۶۰ هرتز حذف شده است [۲].



شکل ۱- شکل موج سمت چپ، نطق هشت و اسپکتروگرام آن را بدون فیلتر پیش تأکید نشان می‌دهد. - شکل موج سمت راست، نطق هشت و اسپکتروگرام آن با حذف DCو استفاده از فیلتر پیش تاکید نشان می‌دهد.

۳– توصیف الگوریتم

۳–۱– الگوریتم انرژی تیگر

تیگر در مدل‌سازی تولید گفتار، یک الگوریتم جدید برای محاسبه انرژی یک سیگنال پیشنهاد کرد که الگوریتم انرژی تیگر نامیده می‌شود [۲] . اگر نمونه‌های یک سیگنال به شکل $x_i=A\cos(\Omega i+\Phi)$ باشند. که در آن A دامنه سیگنال، Ω فرکانس گسسته زمان و Φ فاز اولیه سیگنال باشد. انرژی لحظه‌ایی E_i برای نمونه‌های x_i در الگوریتم انرژی تیگر به‌صورت معادله (۳) تعریف شده است:

$$E_i = x_i^2 + x_{i+1}x_{i-1} \quad (۳)$$

$$= A^2 \sin^2(\Omega)$$

$$E_i \cong A^2\Omega^2 \quad (۴)$$

در معادله (۴) واضح است که عبارت انرژی، نه تنها به مربع دامنه سیگنال، بلکه به مربع مؤلفه‌های فرکانس نوسان نیز وابسته

است. یک اصلاح‌سازی روی الگوریتم تیگر انجام شده، که در معادله (۳) بجای محاسبه‌کردن انرژی لحظه‌ایی برای هر نمونه سیگنال، محاسبات برای هر فریم انجام داده می‌شود، که تحت عنوان اندازه‌گیری انرژی تیگر روی فریم بکار برده می‌شود [۲] . با انجام محاسبات زیر :

۱- محاسبه‌کردن طیف توان برای هر فریم
۲- وزن‌دهی هر نمونه از طیف توان هر فریم با مجذورمولفه‌های فرکانس
۳- گرفتن ریشه مربع طیف توان وزن داده شده انرژی برای هر فریم [۲] .

الگوریتم انرژی تیگر درتشخیص نقاط ابتدا و انتهای گفتار بکار برده می‌شود. مخصوصاً وقتی که یک سایشی ضعیف نظیر / ف س ش .../ یا انفجاری مانند / ب پ ت ... / وجود داشته باشد، که دامنه ضعیف اما فرکانس بالاتری دارند.

۳–۲– الگوریتم ویژگی انرژی–انتروپی(EEF)

اگرچه انرژی اصلاح‌شده تیگر در حضور نویزهایی از قبیل نویز سخنان نامفهوم و نویز زمینه موزیک خوب عمل می‌کند، ولی در برابر نویزهای غیرایستان مانند صداهای مکانیکی، از قبیل نویز ناشی از صدای باز و بسته‌شدن درب و نویز ناشی از صدای لرزش موتور، دچار خطا می‌شود [۴]. الگوریتم ویژگی‌های انرژی-انتروپی که به اختصار EEF نامیده می‌شود، قابل اعتمادتر از روش‌های مبتنی بر انرژی خالص است و مقاومت بیشتری در برابر انواع نویزهای با شدت بالا دارد. در این روش در ابتدا، انرژی و انتروپی هر فریم محاسبه می‌شود، سپس با حذف مقدار میانگین از هر فریم، مقادیر تنظیم شده به‌صورت نقطه به نقطه در هم ضرب می‌شود که با عمل ضرب، روی نواحی گفتاری تأکید و نواحی غیرگفتاری تضعیف می‌شود و برای تشخیص محدوده گفتار پارامتر مناسبی می‌باشد [۴].

۳–۳– توصیف الگوریتم در تشخیص محدوده گفتار

در این قسمت با استفاده از فیلتر پیش تأکید و فیلتر میان‌گذر و انجام تغییراتی در دو الگوریتم ذکر شده بالا، از آن به‌صورت بهینه، برای سیگنال گفتار فارسی استفاده شده است و در بخش‌های زیر به‌صورت خلاصه به محاسبات آن اشاره می‌شود.

۳–۳–۱– فیلتر پیش تاکید

با حذف DC و استفاده از یک فیلتر پیش تأکید FIR مرتبه اول، ضمن حذف بعضی نویزها، طیف سیگنال نرم تر می‌شود که تأثیر قابل ملاحظه‌ای در بالا بردن دامنه قسمت‌های ضعیف گفتار، در منحنی انرژی اصلاح شده تیگر و منحنی EEF، دارد و قادر است نقاط دقیقتری را در محیط‌های نویزی بدست آورد. عملکرد آن در شکل‌های بخش شبیه‌سازی مشاهده می‌شود. در این مقاله با یک فیلتر میان‌گذر FIR با مرتبه ۵۰ و فرکانس قطع بین ۲۵۰ هرتز و ۳۷۵۰ هرتز که بسیار شبیه باند تلفن است و تقریباً تمامی اطلاعات گفتار را دارد، سیگنال گفتار فیلتر شده است [۴].

۳–۳–۲– محاسبات انرژی اصلاح شده تیگر

این الگوریتم به‌صورت خلاصه در ادامه بیان شده است، ابتدا سیگنال گفتار فیلتر شده با فریم‌های ۲۰ میلی ثانیه و ۵۰ در صد روی‌هم‌افتادگی فریم‌بندی می‌شود. که W به‌عنوان طول وینـدو و S_i به‌عنوان نمونه گفتار i ام نام‌گذاری می‌شود. انرژی اصلاح شده تیگر از لحاظ طیفی قوی‌تر از معادله (۳) است. تبدیل فوریه سریع FFT در هر فریم محاسبه شده است:

$$X(w) = \sum_{i=-\infty}^{\infty} S_i e^{-jwi} \qquad (۵)$$

که در آن X(w) تبدیل دنباله S در حوزه فرکانس می‌باشد. سپس دامنه‌های این طیف، با مجذور مؤلفه‌های فرکانس وزن‌دهی می‌شود. در این قسمت بجای محاسبه طیف توان [۴]، مطابق معادله (۶) از طیف نمونه‌های سیگنال گفتار استفاده شده است، علت استفاده از آن به این دلیل است که باعث تقویت دامنه قسمت‌های ضعیف گفتار می‌شود. این رویکرد یکی از نوآوری‌های این مقاله محسوب می‌شود:

$$f_i = w_i^2 \, X(w_i) \qquad (۶)$$

در انتها انرژی اصلاح‌شده تیگر T_i به‌صورت ریشه جمع مؤلفه‌های فرکانس محاسبه می‌شود(معادله (۷)).

$$T_i = \sqrt{\sum_{k=1}^{K} f_k} \qquad (۷)$$

۳–۳–۳– محاسبات پیشنهادی انرژی– انتروپی

در ادامه محاسبات ویژگی‌های انرژی-آنتروپی خلاصه می‌شود و به بعضی از جزئیات پیاده‌سازی‌های انجام شده، اشاره می‌شود. ابتدا برای هر فریم i، انرژی با جمع مجذور تمام نمونه‌های فریم محاسبه می‌شود:

$$E_i = \sum_{k=1}^{K} (S_{ik}^2) \qquad (۸)$$

که در آن S_{ik} نمونه kام از فریم iام می‌باشد. برای هر فریم، سیگنال از حوزه زمان به حوزه فرکانس نگاشت داده می‌شود که در

معادله (۵) محاسبات آن انجام شده است [۴]. سپس بجای تخمین زدن تابع چگالی احتمال، با نرمالیزه‌کردن طیف نمونه‌ها، بهتر است تابع چگالی احتمال با مجذور طیف نمونه‌های سیگنال تخمین زده شود، که تاثیر آن در بالابردن دامنه قسمت‌های ضعیف گفتار در منحنی انرژی-انتروپی است:

$$P_i = \frac{x(w_i)^2}{\sum_{k=1}^{K} x(w_k)^2} \qquad (۹)$$

$x(w_i)$ طیف انرژی مؤلفه فرکانس w_i، p_i چگالی احتمال، k تعداد کل نقاط FFT در هر فریم است. انتروپی هر فریم H_i برای هر فریم i به‌صورت زیر تعریف می‌شود:

$$H_i = \sum_{k=1}^{K} P_k \log(P_k) \qquad (۱۰)$$

در اینجا بنابر الگوریتم ویژگی آنتروپی انرژی[۴] که معمولاً از میانگین ۱۰ فریم اول جهت کاهش نویز زمینه استفاده می‌شد، صرف‌نظر می‌شود. با استفاده از نمودار انرژی تیگر محدوده گفتار، بدون تخمین نویز زمینه محاسبه می‌شود و در انتها ویژگی EE به‌صورت زیر تعریف می‌شود [۵].

$$EEF_i = \sqrt{1 + \left| E_i \times H_i \right|} \qquad (۱۱)$$

۳–۳–۴– مکان یابی مناطق کم انرژی

انرژی تیگر تحت عنوان T و انرژی- انتروپی EEF را در فاصله بین صفر و یک نرمالیزه و به‌ترتیب T1 و EEF1 می‌نامیم. اولین سطح تصمیم‌گیری مربوط به T1 است که با تعریف سطح آستانه Tresh_B1 نقطه ابتدای گفتار و آستانه Tresh_E1 نقطه اولیه انتهای گفتار تعیین می‌شوند:

$$tb1 = \text{argmin}_i \{ T1(i) \geq Tresh\_B1 \}, 1 \leq i \leq N \qquad (۱۲)$$

$$te1 = \text{argmin}_i \{ T1(i) \geq Tresh\_E1 \}, N \leq i \leq 1 \qquad (۱۳)$$

مقداری که برای آستانه‌ها بکار برده می‌شود مقدار ثابتی است که در اینجا۱۱.، ۹. برای آستانه‌های Tresh_B1 و Tresh_E1 بکار رفته است. بنابراین انرژی فریم تیگر، به‌عنوان تخمینی از ابتدا و انتهای گفتار است و از ویژگی انرژی- انتروپی طبق توصیف زیر جهت تعیین نقاط واقعی استفاده شده است.

۳–۳–۵– تشخیص نهایی محدوده گفتار

برای منحنی نرمالیزه شده EEF1 دو آستانه یکی برای محدوده شروع گفتار Tresh_B و دیگری برای محدوده انتهای گفتار Tresh_E درنظر گرفته می‌شود، این دو آستانه مقدار خیلی کمتری دارند [۵]. با جستجو از tb1 طبق تعریف زیر نقطه شروع مشخص می‌شود:

$$tb = \text{argmin}_i \{ EEF1(i) \geq Tresh\_B \}, tb1 \leq i \leq N \qquad (۱۴)$$

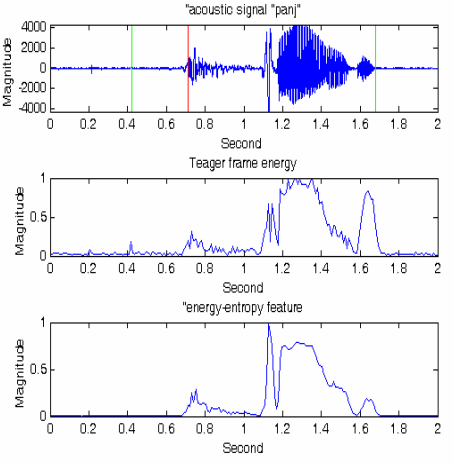
با جستجو از te1 طبق تعریف زیر نقطه انتهای گفتار مشخص می‌شود:

$$te = \text{argmax}_i \{ EEF1(i) \geq Tresh\_E \}, te1 \leq i \leq N \qquad (۱۵)$$

۴– آزمایش‌های انجام شده

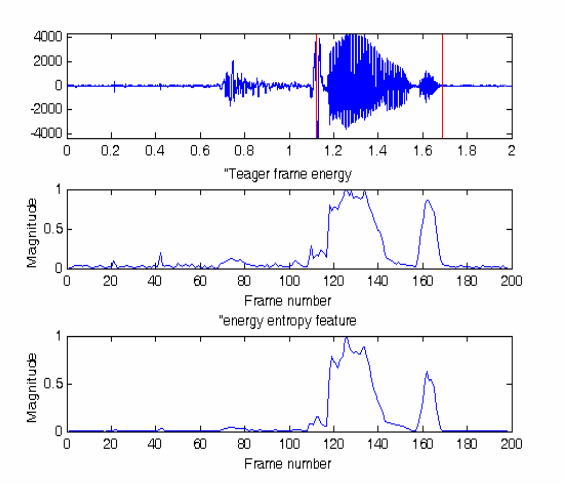
لازم به ذکر است که پیاده‌سازی‌های انجام شده در محیط مطلب انجام گرفته است و داده‌های مورد استفاده توسط تیم مجری جمع‌آوری گردیده است.

الگوریتم توصیف شده بالا با یک منبع اطلاعاتی گفتار تشکیل شده از اعداد صفر تا نه فارسی که توسط ۱۰ گوینده، ۵ گوینده مرد و ۵ گوینده زن، که هر کدام ۲ بار کلمات صفر تا ده را تکرار کرده‌اند تست شده است و در جدول ۱ آورده شده است. میزان دقت الگوریتم با گوش‌دادن متوالی به اجرای گفتار و استفاده از نرم‌افزارها تعیین شده است. فرکانس نمونه‌برداری ضبط شده ۲۲۵۰۰ هرتز و طول فریم ۲۰ میلی ثانیه که شامل ۴۵۰ نمونه در هر فریم همراه با۵۰ درصد روی هم افتادگی است در شکل (۲) تشخیص نادرست محدوده گفتار، درحضور نویز ناشی از نفس‌کشیدن و صدای لب‌ها و کلیک‌های دهان واضح است.

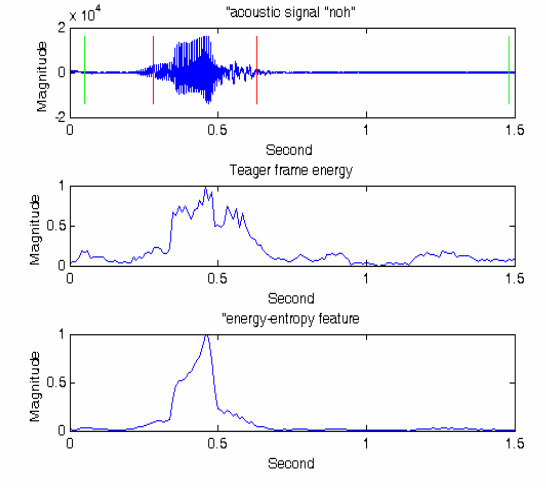


شکل ۲- تشخیص نادرست نقطه شروع نطق "پنج" در حضور نویز نفس‌کشیدن و نویز کلیک دهان، به دلیل عدم استفاده از فیلتر پیش تأکید (محدوده گفتار بین ۶۷۲ . تا ۶۸ . ۱ ثانیه است).

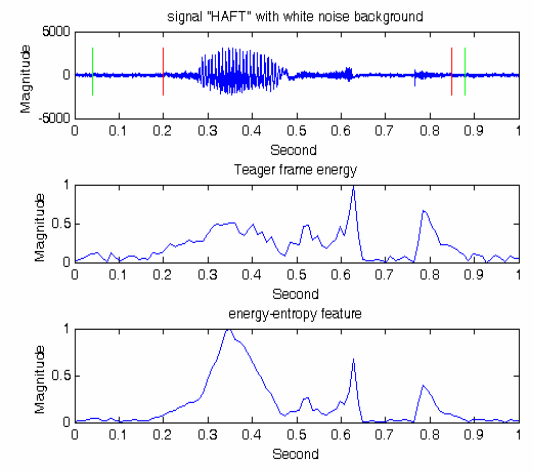
در منحنی نرمالیزه شده انرژی- انتروپی شکل (۳)، حذف نویز ناشی از نفس‌کشیدن را که در فاصله ۷. تا ۱ . ۱ ثانیه است به وضوح دیده می‌شود. شکل (۴) یک نمونه تشخیص صحیح نقطه شروع و پایان نطق "نه" را با وجود نویز در ابتدای گفتار را نشان می‌دهد. شکل (۵) تشخیص صحیح کلمه هفت را در حضور نویزهای غیرایستان نظیر نویز سفید نشان می‌دهد.



شکل ۳- تشخیص صحیح نقاط ابتدا و انتهای نطق "پنج" بعد از استفاده فیلتر پیش تاکید و حذف نویز نفس‌کشیدن و نویز دهان و لب‌ها و متعادل‌شدن شکل موج در منحنی انرژی– انتروپی.



شکل ۴- تشخیص نقاط ابتدا و انتهای عدد "نه" با وجود نویز در ابتدای گفتار.



شکل ۵- تشخیص صحیح محدوده گفتار در نطق "هفت" در حضور نویز سفید.

جدول ۱- بازدهی الگوریتم، در تشخیص ابتدا و انتهای اعداد صفر تا نه.

بازدهی	تعداد خطا	تعداد فایلها	گوینده	روش
٪۸۶	۱۶	۱۰۰	مرد	انرژی- ZCR
٪۸۴	۱۸	۱۰۰	زن	
٪۹۲	۸	۱۰۰	مرد	انرژی-انتروپی
٪۹۳	۹	۱۰۰	زن	
٪۹۸	۲	۱۰۰	مرد	انرژی-انتروپی پیشنهادی
٪۹۷	۳	۱۰۰	زن	

۵– نتیجه‌گیری

واج های خیلی ضعیف ابتدا و انتهای گفتار در محیطهای نویزی به سهولت قابل تشخیص نیست. با استفاده از تغییراتی در انرژی اصلاح شده تیگر، و همچنین اصلاحاتی در روش انرژی- انتروپی، واجهای ضعیفی که در ابتدا و یا انتهای گفتار قرار می‌گیرند، بهتر آشکار می‌شوند. نتایج آزمایشات نشان می‌دهد که الگوریتم پیشنهادی در این مقاله، با استفاده از فیلتر پیش تأکید قبل از فیلتر میان‌گذر و تغییراتی در الگوریتم تیگر و الگوریتم اصلاح شده انرژی- انتروپی و استفاده از آستانه‌های ثابت، بازدهی قابل قبولی در تشخیص نقاط ابتدا و انتهای گفتار دارد، مخصوصاً وقتی که واجهای ضعیف در شروع یا انتهای گفتار قرار گیرد. این الگوریتم بدون تخمین نویز، در محیطهایی با نویز بالا، بازدهی مناسبی از خود نشان داده است.

۶– مراجع

[1] L.R. Rabiner, M.R. Sambur; **"An Algorithm for Determining the Endpoints of Isolated Utterances "**, Bell Sjjsl. Tech. J., Vol. 54, pp.297-315, 1975.

[2] [G.S. Ying, C.D. Mitchell, L.H. Jamieson; **"Endpoint Detection of Isolated Utterances Based on A Modified Teager Energy Measurement"**. In Proc. IEEE ICASSP-92, pp.732-735, 1992.

[3] H. Qiang, Z. Youwei; **"On Prefiltering and Endpoint Detection of Speech Signal"**, Proceedings of ICSP 1998 , pp.749-752, 1998.

[4] L.S.Huang, C.H.Yang; "**A Novel Approach to robust speech endpoint detection in car environments**", ICASSP-2000, Vol. 3, PP.1751-1754, 2000

[5] L. Gu and S.A. Zahorian; **"A New Robust Algorithm for Isolated Word Endpoint Detection"**, IV-4161 ICASSP, 2002.

[6] W.Han, C.F.Chan, C.S.Choy, K.P.Pun; **"An Efficient MFCC Extraction Method in Speech Recognition",** ISCAS 2006. Volume , Issue , 4 pp, 2006.