ارائـه الگـوریتـم نقاط با حداکثر اطلاعات MIP برای‌تشخیص ابتدا و انتهای دستورات گفتاری

ابوالقاسم صیادیان
دانشیار
دانشکده برق، دانشگاه صنعتی امیرکبیر

کامبیز بدیع
دانشیار
مرکز تحقیقات مخابرات ایران

محمد معین
استادیار
مرکز تحقیقات مخابرات ایران

نصرا... مقدم
استادیار
دانشکده فنی و مهندسی، دانشگاه تربیت مدرس

چکیده

آشکارسـازی کلمـات و دستورات گفتاری در هنگام ضبط و بازشناسی در حضور انواع وقایع اکوستیکی غیرگفتاری (پف، کلیک، تـه سـرفه، بـازدم صـدا، ...) و همچنیـن نویـزهای زمیـنه تداوم دار یا غیرتداوم دار را SBED یا VAD می‌گویند. در یک تحقیق گسترده نشـان داده شده است که حدوداً پنجاه درصد خطاهای یک سیستم بازشناسی گفتار، ناشی از خطای آشکارسازی دقیق مرز ابـتدا و انـتهای کلمـات تلفظ شده (بویژه در حضور نویز) می‌باشد. در این تحقیق ضمن بررسی و پیاده سازی اغلب روشهای موفق، روشـی ارائه نمودیم که ضمن استفاده از نقاط قوت روشهای رایج، دارای قابلیت پیاده سازی زمان زنده نیز می‌باشد. در این تحقیق ابـتدا نقاطی از گفتار بنام نقاط MIP (نقاط با حداکثر اطلاعات) شروع و پایان کلمه تشخیص داده میشود. آنگاه با استفاده از تخمین مشخصـات مدل نویز زمینه (به کمک فیلتر غیر خطی مرتب کننده توان)، نقاط شروع و پایانی کلمه و همچنین، فریمهای سکوت بین سـیلابهای درون کلمه‌ای نیز تشخیص داده میشوند. روش فوق همراه با دو روش موفـق دیگر را در بازشناسی ۱۵۰۰ کلمه پرمصرف (در کاربـردهای IT و سرویـسهای مخابراتـی) مورد آزمون‌های مختلف قرار دادیم. در این آزمونها هم دقت روش‌ها وهم مقاومت آنهـا در مقـابل انواع نویزهای زمینه (پریودیک و غیرپریودیک، ایستان و غیر ایستان، رنگی و غیررنگی و...) تحت SNR های (dB ۲۰و۱۵و۱۰و۵) را مـورد بررسـی قـرار دادیـم. در تمامی این آزمونها، روش MIP بعلت استفاده از حداکثر اطلاعات قابل وصول در گفتار موفق‌تر بوده است.

کلمات کلیدی

بازشناسی گفتار، تشخیص ابتدا و انتهای گفتار، آشکارسازی فعالیت گفتار

Use of Maximum Information Point (MIP) for High Precision and Reliable Begin and Endpoint Detection of Speech Command

A. Sayadiyan
Associate Professor

K. Badi
Associate Professor

M. Moin
Assistant Professor

N. Moghadam
Assistant Professor

Abstract

Many speech detection methods have been proposed for the use in speech recognition systems. These methods can be broadly classified into three approaches according to their interaction with the pattern matching paradigm: (1) the explicit approach (2) the implicit approach (3) the hybrid approach . In this research we implement a new explicit method, which is reliable and very prease for real time applications. This new speech detection method (with name of maximum information point MIP) has been proposed and evaluated on a 1500 words discrete utterance speech command system. MIP method use of all low and high level information of speech signal and background noises. The evalution results at (5 to 30 dB S/ N ratio), and all variant of noise (white, color, periodic, stationary, non stationary, variable power level, ...), shows that the MIP method performance is better than the other classical methods.

Keywords

Speech Recognition, Speech Begin and End Point Detection, Voice Activity Detector

مقدمه

برای بازشناسی گفتار یا گوینده در دنیای واقعی، اغلب نیازمند هستیم که گفتار تلفظ شده در یک محیط آغشته به نویزهای محیطی مختلف را مورد پردازش و بازشناسی قرار دهیم. آشکارسازی گفتار در هنگام ضبط در حضور انواع وقایع اکوستیکی غیر گفتاری[۱] و نویزهای زمینه[۲] را SBED[۳] و یا ED[۴] و یا SAD[۵] و یا VAD[۶] می‌گویند. الگوریتمهای مرتبط با VAD نوعاً با محدودیت تأخیر کمتر از چند فریم(۲ تا ۳ فریم) مواجه هستند، در نتیجه معمولاً برای فشرده‌سازی و کدینگ گفتار استفاده میشوند. در الگوریتمهای مرتبط با SBED، تأخیری در حدود نیم تا حداکثر یک ثانیه قابل تحمل می باشد. سیستمهای پیشرفته بازشناسی گفتار یا گوینده، هم از روشهای SBED و هم از روشهای VAD استفاده مینمایند.

در یک تحقیق گسترده [۱] نشان داده شده است که حدود پنجاه درصد خطاهای یک سیستم بازشناسی گفتار، ناشی از خطای آشکارسازی دقیق مرز ابتدا و انتهای کلمات تلفظ شده می‌باشد. عدم دقت کافی در تشخیص فواصل سکوت سیلابهای بین کلمه‌ای یا جمله‌ای نیز می‌تواند موجب کاهش دقت سیستمهای بازشناسی گفتار و گوینده گردد. بنابراین وظیفه روشهای SBED تشخیص دقیق و مطمئن ابتدا و انتهای گفتار تلفظ شده با فرض اینکه در فواصل ما قبل و ما بعد آن در هنگام ضبط نویزهای زمینه یا نویزهای اکوستیکی دیگری وجود دارد میباشد.

همچنین وظیفه روشهای VAD، تشخیص فریمهای گفتار از فریم‌های غیر گفتاری می‌باشد. نکته قابل توجه این است که، فریمهائی در فواصل ابتدا و انتهای یک تلفظ گفتاری می‌توانند، سکوت یا نویز زمینه باشند که تشخیص این مسئله بعهده بخش VAD میباشد.

وقایع اکوستیکی غیر گفتاری (یا نویز) که سیگنال گفتار را در بر گرفته‌اند را میتوان به دو گروه کلی ؛ الف ـ نویزهای ضربه‌ای ب ـ نویزهای غیر ضربه‌ای یا تداوم دار[۸] تقسیم نمود. نویزهای ضربه‌ای توسط طول دوره آن نسبت به سیگنال گفتار قابل تشخیص می‌باشند[۲]. از جمله این نویزها، پف دهان، ته سرفه برای صاف کردن حنجره، کلیک و ضربه میکروفون و موس در هنگام جابجایی می‌باشد. نویزهای تداوم‌دار (غیر ضربه‌ای) معمولاً ما قبل و مابعد گفتار نیز موجودیت خود را حفظ مینمایند. روشهای عمده برای تشخیص SBED وVAD نوعاً مبتنی بر فرض وجود نویزهای تداوم‌دار می‌باشند. نویزهای زمینه، نویزهای مربوطه به صدای همهمه، نویزهای مربوط به صدای پنکه، فن، صدای موتور خانه، صدای موتوریخچال، صدای موتورجارو برقی، صدای کولر، صدای ماشین لباسشوئی، صدای موتور انواع ماشینها (در داخل یا خارج ماشین) و ... همگی جزء نویزهای تداوم‌دار محسوب میشوند. نویزها از نظر هموار بودن طیف به دو گروه؛ نویزهای سفید[۹] و نویزهای رنگی[۱۰] نیز دسته‌بندی میگردند. در نویزهای سفید، پوش طیف تقریباً هموار و یکنواخت می‌باشد. ولی نویزهای رنگی دارای قطب و صفرهائی در پوش طیف می‌باشند. نویزها از نظر ثابت بودن یا متغیر بودن مشخصه‌های آماری و طیفی به دو گروه کلی نویزهای ایستان[۱۱] و غیر ایستان[۱۲] نیز تقسیم‌بندی میگردند. تکنیکهای SBED و VAD برای نویزهای غیر ایستان معمولاً پیچیده‌تر از نویزهای ایستان می‌باشند. نویزهای تداوم‌دار از نظر تکراری بودن نیز به دو گروه نویزهای پریودیک و غیر پریودیک نیز تقسیم‌بندی میشوند.

یکـی دیگر از مشخصه نویزها، ثابت بودن یا متغیر بودن سطح انرژی نویز در فواصل ماقبل و ما بعد گفتار می‌باشد. به نویزی که سـطح انـرژی آن در فواصـل ماقـبل و مـا بعـد تقریـباً ثابت باشد نویز FLN[۱۳] و در غیر اینصورت VLN [۱۴]می‌گویند. تکنیکهای SBED و VAD بـرای نویـزهای بـا سـطح انرژیVLN معمولاً پیچیده‌تر از نویزهای با سطح انرژی تقریباً ثابت می‌باشد [۲]. در ایـن تحقـیق فرض برای این است که سطح نویز در فاصله تلفظ یک کلمه یا فرمان تقریباً ثابت بوده ولی در طول ضبط میتواند تغـیر نماید. روشهای SBED معمولاً به دو گروه کلی BM[۱۵] وRTM[۱۶] قابل تفکیک می‌باشند[۳]. در روشهای پردازش دسته‌ای BM، فـرض بـر ایـن اسـت کـه اطلاعـات تمامـی فـریمهای گفتار و مقدار مناسبی از فریمهای سکوت ماقبل و مابعد در اختیار مـی‌باشـد[۳]. ایـن روشها مـــعمولاً در تـصدیق و تعیین هویت گوینده‌ها و سیستمهای شماره‌گیری صوتی، دستورات صوتی، و ... مورد استفاده قرار می گیرند. در این روشها، پس از ضبط مقدار مناسبی گفتار (همراه با سکوت ماقبل و مابعد) فرآیند ضبط قطـع شـده و سیسـتم بـه فرآیـند پـردازش و بازشناسـی می‌پردازد. یعنی فرآیند پردازش و بازشنای پس از خاتمه ضبط انجام می‌پذیرد. چون تمامی اطلاعات گفتار یکجا در اختیار می باشد، انتظار این است که روشهای BM دارای دقت و عملکرد بالاتری نسبت به روشهای RTM باشند.

در روشـهای زمـان زنـده RTM، نتـیجه پاسخ SBED و بازشناسـی گفتار، سریعتر یا همزمان با ختم فرآیند ضبط یا تلفظ گفـتار انجـام مـی پذیـرد. (با یک تأخیر ثابت و کم). منطقی است که روشهای مورد استفاده در کاربردهای RTM می‌بایستی از نظـر محاسـباتی سـریعتر باشـند. روش مورد نظر در این تحقیق جزء روشهای RTM بوده اگرچه برای کابردهای BM نیز قابل استفاده است. روشهای تشخیص ابتدا و انتهای گفتار SBED از نظر وابستـگی بـه نـوع مـدلسـازی بازشناسی گفتار نیز به سه دسته کلی؛ الف ـ روشهای صریح[۱۷] ب ـ روشهای ضمنی[۱۸] ج ـ روشهای تلفیقی[۱۹] تقسیم بندی میگردند [۴].

روشـهای صـریح، بـه روشهائی گفته میشود که فرآیند SBED مستقل ازنوع مدلسازی کلاسهای اکوستیکی (واج یا کلمه) و مسـتقل از نـوع مدلسازی زبان و لهجه گوینده انجام می‌پذیرد. روشهای صریح، از نظر کاربردی دارای انعطاف‌پذیری بیشتر و از نظـر پـیاده‌سـازی سـریعتر مـی‌باشـند. روشـهای ضـمنی، به روشهائی گفته میشود که فرآیند SBED همراه و جزئی از فرآیند مدلسـازی کلاسـهای اکوسـتیکی مورد توجه و استفاده می‌باشد. این روشها نوعاً به نوع مدلسازی کلاسهای اکوستیکی و زبان و لهجـه و نوع گوینده وابسته می‌باشند. مهمترین روشهای ضمنی، روشهای HMMA[۲۰] و DTWA[۲۱] [۴] [۵] می‌باشند. دقت این روشـها نوعـاً بالاتـر از روشـهای صـریح است. انعطاف‌پذیری این روشها پائین بوده و هزینه محاسباتی پیاده‌سازی آنها بسیار بالا مـی‌باشـد. در روشـهای تلفیقی از تلفیق مناسب روشهای صریح و ضمنی استفاده میشود. بدین گونه که توسط روشهای صریح، تعدادی کاندید برای ابتدا و انتهای کلمه یا جمله گفتاری پیشنهاد نموده و سپس به کمک روشهای ضمنی، ابتدا و انتهای اصلی گفـتار را بادقـت بالاتـر تعیین مینمایند. بطور منطقی این روش از مزایای هر دو روش استفاده نموده و معایب هر دو روش را تا حدودی تعدیل مینماید.

روشـهائی که در این تحقیق مورد توجه و مطالعه قرار گرفته‌اند، روشهای صریح می‌باشند. بر طبق نظر آقای ساوجـی [۱] و [۶] ویژگـیها و مشخصـه‌هـای بـارز روشـهای موفق SBED عبارتند از ؛ اطمینان‌پذیری[۲۲]، مقاوم بودن در مقابل نویز[۲۳]، دقت[۲۴]، قابلیـت تطبـیق بـا شـرایط جدید[۲۵]، سادگی و سهولت تحلیل[۲۶]، امکان پیاده‌سازی زمان زنده[۲۷] و عدم نیاز به اطلاعات قبلـی از نویز[۲۸] . بنابراین وقتی یک روش جدید ارائه و مورد مطالعه قرار می‌گیرد، می بایستی با توجه به شاخصهای فوق محک زده شده و مـورد ارزیابـی قرار گیرد. ویژگیها و پارامترهای عمده‌ایکه برای تشخیص ابتدا و انتهای گفتار مورد استفاده قرار می‌گیرند، نوعاً از جـنس پارمترهائی است که در مرحله کدینگ یا بازشناسی گفتار از آنها استفاده میشود. مهمترین این پارامترها، گین یا توان (نمایـنده انـرژی یا توان سیگنال در عرض یک فریم). نرخ عبور از صفرZCR [۲۹] ـ انرژی باقیمانده از پیش بینی خطی LPC [۳۰] ـ طـول دوره گفـتار TD [۳۱] ـ اندازه پریودیک بودنPM [۳۲] پارامترهای پوش طیف شامل MFCG[۳۳] یا MFCC[۳۴] می‌باشند. پارامتر نـرخ عـبور از صفر ZCR و انرژی باقیمانده ازپیش بینی خطی LPC نوعاً برای SNR >۲۰ dB کارائی دارند[۱ الی ۳]. معمولاً از دو پارامتر فوق برای تشخیص ابتدا و انتهای کلمات گفتاری تقریباً تمیز(صرفاً وجود نویزهای زمینه) استفاده میشود. پارامترهای عمـده درتشـخیص ابـتدا و انـتهای گفـتار یـا سیستمهای VAD برای SNR < ۲۰ dB نوعاً گین و طول دوره گفتار TD، اندازه پریودیک بودن PM و ضرایب طیفی و MFCG و MFCC می‌باشند.

۱ـ بررسی و تحلیل الگوریتمهای SBED موجود

۱-۱ـ الگوریتمهائی که از انرژی و نرخ عبور از صفر ZCR استفاده مینمایند

در تمامی این الگوریتمها [۱] و [۴ الی ۹]، فرض بر این است که محدوده مشخصی در ابتدای فایل ضبط شده، سیگنال غیر گفتاری (سکوت یا نویز زمینه و ...) می‌باشد. سپس چندین سطح آستانه از انرژی و نرخ عبور از صفر را از روی فریمهای ابتدای سیگنال بدست آورده و از آنها بعنوان نماینده نویز زمینه استفاده مینمایند. رایج‌ترین روش، بدست آوردن مقدار متوسط و انحراف معیار دو پارامتر انرژی و نرخ عبور از صفر می باشد. آنگاه با استفاده از یک دیاگرام گذرای چند حالته[۳۵]، که تغییر حالت توسط سطوح آستانه فوق الذکر کنترل میشود[۳]، وجود سیگنال گفتار (ابتدا و انتهای کلمه) را تشخیص میدهند. برای تشخیص وقایع اکوستیکی غیر گفتاری پالسی (کلیک، پف و ...) از کوتاهی طول دوره آنها در مقایسه با گفتار استفاده مینمایند. در بعضی از این روشها [۱]، ابتدا فریمی از سیگنال که دارای بیشترین انرژی است، بعنوان نقطه قابل اعتماد برای شروع جستجو تعیین میگردد. آنگاه با استفاده از سطوح آستانه مربوط به انرژی و نرخ عبور از صفر نویزهای زمینه، ابتدا و انتهای گفتار را تشخیص میدهند.

این روشها با چند مشکل اساسی مواجه هستند. فرض میکنیم در چند فریم ابتدای فایل یا بافر ضبط شده صدا، وقایع اکوستیکی غیر مطلوب (غیر از نویزهای تداوم دار) مانند پف یا کلیک و یا ... باشد. پارامترهای آماری (مقادیر متوسط و انحراف معیار) استخراج شده از آنها برای نویزهای تداوم‌دار غیر معتبر بوده وموجب خطای زیادی در تشخیص مکان ابتدا و انتهای گفتار و خطا در کلاس‌بندی فریمهای گفتار به نویز میگردد. برای پرهیز از وقوع چنین وضعیت ناخوشایندی، آقای لی در مرجع [۳]، پیشنهاد استفاده از مدلسازی GMM،به منظور جداسازی مقاوم نویزهای تداوم دار زمینه از وقایع اکوستیکی دیگر نموده است. روش پیشنهادی آقای لی مبتنی بر این فرضیه است که، سیگنال گفتار و نویز هر کدام توسط یک pdf گوسی تک مخلوط متمایز قابل نمایش هستند. فرضیه فوق نیز دو اشکال دارد، اولاً اینکه بازنمائی آماری سیگنال گفتار توسط یک مخلوط گوسی دارای خطای بسیار زیادی است (برای مدلسازی مناسب حداقل به سه مخلوط یکی برای فریمهای باصدا، دومی برای فریمهای بی‌صدای پائین‌گذر و سومی برای فریمهای بی‌صدای بالا گذر نیازمند هستیم). ثانیاً، مدل پیشنهادی آقای لی فریمهای بی‌صدای بالا گذر با انرژی پائین مانند ف، س، ح، ... را در SNR< ۲۰ dB جزء نویز زمینه مدلسازی مینماید. همچنین این روشها در مقابل نویزهای پریودیک، نویزهای غیر ایستان و نویزهائی که واریانس تغیرات انرژی آنها بالا باشد، دارای عملکرد ضعیفی می‌باشند.

۱-۲ـ الگوریتمهائی که از انرژی و اندازه پریودیک بودن PM استفاده مینمایند

فریمهای با صدای گفتار دارای یک ویژگی مهم هستند که از آنها میتوان برای شناسائی گفتار، با وجود نویزهای زمینه استفاده نمود. این ویژگی مهم آن است که، متوسط انرژی فریمهای با صدا، معمولاً حدود ۳۰ dB تا ۱۰ بیشتر از متوسط انرژی فریمهای بی صدای گفتار و نویزهای زمینه می باشد (حتی برای SNR های کمتر از ۱۰ dB). تعدادی از الگوریتمهای VAD و SBED از این خاصیت برای تشخیص یک ناحیه مطمئن[۳۶] گفتار از آن بهره میبرند [۱] و [۱۰ الی ۱۲]. پس از تعیین ناحیه مطمئن، به کمک پارامترهای دیگر(انرژی و نرخ عبور از صفر و یا خروجی فیلترهای تشخیص لبه و ...)، ابتدا و انتهای گفتار را با دقت بیشتری تخمین می‌زنند. این روشها برای تعیین نواحی باصدا، از الگوریتمهای تخمین فرکانس پیچ استفاده مینمایند. بنابراین دقت و عملکرد این روشها، به مقدار زیادی به عملکرد الگوریتمهای تخمین فرکانس پیچ وابسته است. بنابراین جهت استفاده از کارآئی این روشها، می‌بایستی از الگوریتمهای دقیق و مطمئن تخمین فرکانس پیچ استفاده نمائیم. اغلب روشهای پیشنهادی موجود که از خاصیت پریودیک بودن PM استفاده مینمایند در مقابل نویزهای پریودیک و نویزهای غیر ایستان همهمه[۳۷]، دارای عملکرد ضعیفی هستند.

۱-۳ـ الگوریتمهائی که از تغییرات انرژی طیفی استفاده مینمایند

[۲] و [۹] و [۱۲ الی ۱۵]؛ اگر نویزهای تداوم دار تقریباً ایستان باشند (پریودیک یا غیرپریودیک)، در این صورت

پارامترهای پوش طیف آنها در طی فریمهای متوالی نوعاً ثابت می باشند (یا دارای تغییرات کمتری نسبت به سیگنال گفتار هستند). دراغلب این روشها می بایستی تخمینی از میانگین و واریانس پارامترهای پوش طیف نویز تداوم دار در اختیار باشد. روشهای ارائه شده در مراجع ذکر شده، با متوسط گیری از پارامترهای طیفی چند فریم آغازین فایل یا بافر ضبط شده، متوسط و انحراف معیار پارامترهای طیفی نویزهای تقریباً ایستان مانند MFCG و MFCC را بدست می‌آورند. آنگاه قدر مطلق اختلاف (یا توان دوم اختلاف) پارامترهای طیفی کلیه فریمها و مدل نویز را بدست می آورند. اگر این اختلاف از یک سطوح آستانه (که بصورت تجربی تعیین میشوند) بیشتر باشد، وجود سیگنال گفتار اعلام شده و سپس به کمک انرژی و سایر پارامترها، نقاط دقیق ابتدا و انتهای گفتار را تعیین مینمایند. این روشها برای نویزهای تقریباً ایستان (پریودیک و یا غیر پریودیک) دارای کارائی مناسبی می‌باشند. مشکل عمده این روشها، کاهش عملکرد با تغییر تدریجی سطحی انرژی نویزهای تداوم‌دار و یا تغیر مشخصه طیفی نویزهای غیر ایستان میباشد.

۲ـ ارائه الگوریتم جدید MIP_SBED

الگوریتم MIP_SBED الگوریتمی صریح برای تخمین ابتدا و انتهای کلمات گفتاری هم در محیط‌های تمیز و همچنین در محیطهای نویزی می‌باشد. الگوریتم MIP_SBED، مبتنی بر حداکثر استفاده از اطلاعات ذیل می‌باشد:

الف ـ استفاده از خاصیت پریودیک بودن حدود ۷۰-۵۰ درصد طول کلمات گفتاری

ب ـ استفاده از بالا بودن انرژی نواحی باصدا (صدای پریودیک) نسبت به نواحی بی‌صدا و نویزهای تداوم‌دار زمینه.

ج ـ استفاده از خاصیت حداکثر غیر ایستان بودن نواحی مرزبین واجهای باصدا و بی‌صدا در ابتدا و انتهای کلمات

د ـ استفاده از فیلتر غیرخطی مرتب کننده[۳۹]، به منظور حداکثر جداسازی اطلاعات فریمهای نویز تداوم‌دار زمینه از فریمهای ایمپالسی (مانند پف، کلیک و ...) و فریمهای بی‌صدا و باصدای گفتار.

ه ـ استفاده از طول حداکثر واج بی‌صدا در ابتدای کلمات (حدود ۳۰۰ میلی ثانیه) و طول حداکثر مجموعه واجهای بی‌صدا در انتهای کلمات (حدود ۶۰۰ میلی ثانیه).

ر ـ استفاده از این نکته که طول دوره سکوت بین سیلابهای درون کلمه‌ای حداکثر ۳۰۰ میل ثانیه می‌باشد.

ز ـ استفاد از این نکته که دوره سکوت بین دو کلمه متوالی حداقل می‌بایستی نیم ثانیه باشد.

ی ـ استفاده از حداقل و حداکثرطول دوره کلمات گفتاری

پارامترهای اصلی مورد استفاده در روش MIP-SBED عبارتند از: توان برحسب (dB)، اندازه درست نمائی پریودیک بودن[۴۰] PLL ، ویژگیهای پوش طیف (MFCG یا MFCC).

پارامتر توان و ضرایب MFCG یا MFCC در تمامی روشهای آنالیز سیگنال گفتار (سیستم‌های بازشناسی گفتار و گوینده) استخراج شده و در اختیار می‌باشند. پارامتر درست نمائی پریودیک بودن موقعی استخراج میشود که بخواهیم فرکانس پیچ را استخراج نموده و یا وضعیت باصدا یا بی صدا بودن فریمهای گفتار را تعیین نمائیم. خوشبختانه روش آنالیز این تحقیق برای استخراج ویژگیهای طیفی (ضرائبMFCG یا MFCC)، روش همزمان با پیچ[۴۱] می‌باشد. در نتیجه استخراج فرکانس پیچ و اندازه درست نمائی پریودیک بودن جزئی از فرآیند مرحله آنالیز گفتار می‌باشد. بنابراین برای تشخیص ابتدا و انتهای گفتار بروش MIP-SBED صرفاً از پارامترهای کلاسیک استخراج شده در مرحله آنالیز استفاد مینمائیم.

هدف این تحقیق در ارائه روش MIP-SBED عبارت است از؛

۱ـ مقاوم بودن روش در مقابل انواع نویزهای پریودیک و غیرپریودیک (SNR $\geq$ 5 dB).

۲- دقیق بودن تشخیص مرزهای ابتدا و انتهای گفتار

۳- امکان پیاده‌سازی بصورت دسته‌ای BM و همچنین زمان زنده RTM بر روی کامپیوترهای PC قابل دسترس.

در بخش‌های زیر به شرح جزئیات الگوریتم پیشنهادی می‌پردازیم.

۲-۱ـ روشهای دقیق و مقاوم اندازه‌گیری پریودیک بودن PM

در مراجع [۱۰ الی ۱۲] روشهائی برای اندازه‌گیری پریودیک بودن فریمهای متوالی گفتار ارائه شده است. این روشها اگر چه

دارای عملکـرد مناسبی برای گفتار تمیز(یا برای SNR های بالا و متوسط) می‌باشند، اما برای SNR های پائین (بویژه بین صفر تـا ده) ضـعیف مـی‌باشند. عملکرد پائین این روش موقعی تشدید میشود که نویزهای محیطی غیر ایستان و تقریباً رنگی باشند. (مانـند نویـز همهمـه). مـرجع[۱۶] منبعـی غنی از انواع روشهای تخمین فرکانس پیچ و باصدا بودن فریمهای گفتاری است که میـتوان از ایـن روشـها برای اندازه‌گیری پریودیک بودن گفتار نیز استفاد نمود. بررسیها و مطالعات علمی و عملی محققین این نوشـتار نشـان مـیدهد کـه محاسـبه NCC[۴۲] همبسـتگی مـتقابل نرمالیزه شده [۱۷] بر روی سیگنال گفتار میان گذر شده[۴۳]، پارامتری بسیار ارزشمند جهت اندازه‌گیری مقدار پریودیک بودن فریمهای متوالی سیگنال گفتار می‌باشد. مطالعات مقایسه‌ای و نـتایج حاصـل از آن در مرجع[۱۸] برای انتخاب روش دقیق و مقاوم تخمین فرکانس پیچ، نیز مؤید انتخاب ما میباشد. با توجه به اینکه روش ارائه شده در مرجع [۱۷] و[۱۸] از سیگنال پائین گذر شده استفاده مینماید و همچنین اصلاحات دیگری که در ایـن تحقـیق بـرای بهـبود عملکـرد اندازه‌گیری پریودیک بودن انجام پذیرفته، به شرح مختصر اندازه‌گیری پریودیک بودن PM فریم‌های گفتار در طی این تحقیق می‌پردازیم. مراحل استخراج اندازه پریودیک بودن PM بشرح زیر می‌باشد:

الـف ـ سیگنال گفتار را از یک فیلتر میان گذر FIR [۴۴] از نوع پارکس مک کلان[۴۵] [۱۹] که مشخصات طیفی آن بشرح شکـــل (۱) مـی‌باشـد عـبور میدهـیم. این فیلتر دارای فاز خطی است، لذا هیچ گونه اعوجاجی در فاز سیگنال گفتار اصلی ایجاد نمی‌نماید.

با توجه باینکه انرژی اصلی فریمهای باصدای گفتار زیر KHz ۱ بوده و همچنین حداکثر فرکانس پیچ قابل قبول سیستمHz ۵۰۰ مـی باشـد، فـرکانس توقف فیلتر را KHz ۱ در نظر گرفتیم. بدین ترتیب ضمن حفظ انرژی قسمتهای پریودیک فریمهای باصدا، سیگنال گفتار رانسبت به مؤلفه‌های غیر پریودیک سیگنال تحریک و همچنین نویزهای محیطی دارای مؤلفه‌های نزدیک D.C، بهـبود بخشـیده و مقـاوم مـی‌نمـاید. سـیگنالهای گفـتار، در اغلـب مواقـع بعلت غیر ایده‌آل بودن میکروفون و یا بردهای دیجیـتایزر صـوتی آغشـته به هارمونیکهای اول و دوم برق شهر (۶۰Hz-۵۰) و همچنین مؤلفه‌های D.C بایاس شده می‌باشند. استفاده از فیلتر میان گذر با مشخصه ارائه شده موجب حذف مؤلفه اول و دوم برق شهر و مقادیر D.C مزاحم خواهد شد. شکل (۲) یک نمونه از سیگنال باصدای اصلی و میان گذر شده آنرا نمایش میدهد.

ب ـ فـرض مـیکـــنـیم xb ســیگنال مـیان گـذر شده و(n) xb یک نمونه آن در فریـــم n ام بـاشد. همچنین فرض می‌کنیم PP_min و PP_max پریودپـیچ حداقـل و حداکـثر سیستم باشد. چون فرکانس نمونه‌برداری سیگنال گفتار برای دیتاهای ایـن تحقـیق KHz ۸ بوده و پریودپیچ حداقل را ۲ و پریودپیچ حداکثر را ۱۶ میل ثانیه در نظر گرفتیم، در نتیجه محدوده حداقل و حداکثر تغییرات پریودپیچ برابر PP_min = ۱۶ و PP_max=۱۲۸ نمونه خواهد بود.

فرض میکنیم NCC(n,l) ضریب همبستگی متقابل نرمالیزه شده در فریم nام و برای برداری از سیگنال بطول l باشد.

اگـر xb (n, n + ۱... n+l, n+l+۱, ... x+2l) یـک بـردار بطـول (l ۲) از نقطه nام سیگنال باشد، آنگاه ضریب NCC (n,l) بصورت زیر محاسبه میگردد.

$$NCC(n,l)=\frac{\sum_{i=0}^{\tau} xb(n+i).xb(n+l+i)}{\sqrt{\sum_{i=0}^{\tau} xb^{2}(n+i)}.\sqrt{\sum_{i=0}^{\tau} xb^{2}(n+l+i)}} \qquad (1)$$

در مـرجع [۱۷]، τ طـول بـردارهای متوالـی برابر l لحاظ شده است. انتخاب τ باندازه l موجب میشود که فرآیند تخمین پـریود پـیچ و مقدار پریود یک بودن فریم آنالیز دو مرحله‌ای گردد. برای جلوگیری از اتفاق ذکر شده و کاهش حجم محاسبات، τ را بصورت زیر انتخاب میکنیم:

$$\tau=\begin{cases}100 & \text{if} \quad \tau \le 100\\ l & \text{if} \quad \tau \ge 100\end{cases} \qquad (2)$$

طول ۱۰۰ نمونه معادل ۱۲/۵ میلی ثانیه در فرکانس نمونه‌برداری KHz ۸ می‌باشد. انتخاب طول ۱۰۰ نمونه موجب کاهش

اثـرات فورمنـت اول در تخمیـن فرکانس پیچ و اندازه پریودیک بودن میگردد. همچنین با چند به یک کردن سیگنال میان گذر شـده xb بـه نسـبت $\frac{1}{3}$ تـا $\frac{1}{4}$، میـتوان حجـم محاسـبات را نیز بیشتر کاهش داد. اندازه پریودیک بودن فریم nام بصورت زیر محاسبه میشود:

$$PM(n) = Max \{ NCC(n,l) \} \qquad (۳)$$

$l = pp\_min$ تا pp_max

ج - بر روی کانتور اندازه پریودیک بودن فریم‌ها یعنی PM(n) ها، دو اندازه هموار شده بشرح زیر تعریف مینمائیم:

$$RC\_PM(n) = \frac{1}{L}\sum_{i=0}^{L} PM(n+i) \qquad (۴)$$

$$LC\_PM(n) = \frac{1}{L}\sum_{i=0}^{L} PM(n-i) \qquad (۵)$$

RC_PM(n) را انـدازه پـریودیک بـودن مـتأثر از بافت راست[۴۶] و LC_PM(n) را اندازه پریودیک بودن متأثر از بافت چپ[۴۷] مـی‌نامـیم. L عمـق نفوذ بافت سمت چپ یا راست فریم nام می‌باشد. تجربیات محققین این نوشتار نشان میدهد که انتخاب L می‌بایستی بقسمی انجام پذیرد که طول بافت چپ یا راست کمتر از ۶۰ میلی ثانیه و بزرگتر از ۴۰ میلی ثانیه گردد.

پارامـتر LC_PM , RC_PM, PM در محـدوده ۱ و -۱ تغییر مینماید. هر قدر پارامترهای فوق به یک نزدیکتر باشند، فریم مـورد نظـر دارای اندازه پریودیک بودن بیشتری خواهد بود در کاربردهای عملی یک سطح آستانه th_PM برای فریمهای باصدا در نظر میگیرند یعنی وقتی مقدار PM از سطح آستانه th_PM بزرگتر باشد، فریم آنالیز را باصدا و در غیر اینصورت بی صدا یا سـکوت در نظـر میگـیرند. در کاربـردهای عملـی سـطح آستانه th_PM در محدوده ۰/۸۵ تا ۰/۸۲۵ (بستگی به سطح نویزهای زمینه و فرکانس نمونه برداری) انتخاب میشود.

۲-۲ـ محاسبه توان میان گذر شده سیگنال گفتار

در مـرجع [۱] و[۲]، دو نـوع انـرژی بـرای هـر فـریم سیگنال گفتار پیشنهاد شده است، اول، انرژی سیگنال در حوزه زمان دوم؛ انـرژی سـیگنال مـیانگذر شـده در حـوزه فرکانس. سپس با ترکیب وزنی این دو نوع انرژی ، یک پارامتر بنام انرژی TF[۴۸] معرفـی و از آن به عنوان نماینده انرژی سیگنال گفتار جهت متمایز سازی نویزهای زمینه از سیگنال گفتار استفاده شده است. تحقـیقات محققیـن این نوشتار نشان میدهد که تلفیق دو نوع انرژی بصورت ذکر شده چندان ضروری نمی‌باشد. اما برای اینکه حتـی الامکـان انـرژی نویـزهای زمیـنه نسـبت به انرژی سیگنال گفتار تضعیف گردند، میان گذر کردن سیگنال گفتار و سپس محاسـبه انـرژی آن ضروری بنظر میرسد. برای انجام این ایده فرض میکنیم x(n+I) نمونه‌های سیگنال در فریم nام باشد. توان سیگنال خام را بصورت زیر تعریف مینمائیم:

$$P(n) = \frac{\sum_{i=-l}^{i=+l} x^2(n+i).w^2(i+l)}{\sum_{i=0}^{2l+1} w^2(i)} \qquad (۶)$$

که درآن W(i) (i=0 تا 2 l+1) وزن پنجره همنیگ و L =2 l+1 طول پنجره همینگ در مرحله آنالیز سیگنال گفتار می‌باشد. حال فرض می‌کنیم X(n+k) نمونه‌های FFT[۴۹] سیگنال پنجره شده برای فریم nام باشند.

نمونه‌های فوق را در یک تابع وزنی بصورت شکل (۳) ضرب مینمائیم.

$$WX(n+k) = FW(k) \cdot X(n+k) \qquad (۷)$$

حال نسبت توان سیگنال اصلی در حوزه فرکانس به نسبت توان سیگنال وزن داده شده در حوزه فرکانس را مطابق رابطه زیر بدست می‌آوریم:

$$\gamma = \frac{\sum_{k=0}^{N/2} WX^2(n+k)}{\sum_{k=0}^{N/2} X^2(n+k)} \qquad (۸)$$

که در آن N طول نمونه‌های FFT می‌باشد(عددی بین ۲۵۶ تا ۱۰۲۴ برای فرکانس نمونه‌برداری ۸KHz) آنگاه توان سیگنال میان گذر شده را بصورت زیر بدست می‌آوریم:

$$BP(n) = \gamma \cdot P(n) \qquad (۹)$$

توان سیگنال میان گذرشده بر حسب dB نیز بصورت زیرتعریف و مورد استفاده قرار میدهیم:

$$LBP(n) = 10 \log_{10}(BP(n)) \qquad (۱۰)$$

به منظور کاهش واریانس توان نویز، برروی کانتور پارامتر توان LBP(n) یک فیلتر پائین گذر هموار ساز مناسب اعمال مینمائیم. توان هموار شده فریم n ام را اصطلاحاً با SLBP(n) نمایش میدهیم. از روی کانتور توان هموار شده SLBP(n)، دو پارامتر جدید توان، بنامهای توان مرتب شده متأثر از بافت چپ و راست بشرح زیر تعریف مینمائیم:

فرض میکنیم:

$$\{SLBP(n), \ldots SLBP(n+i), \ldots S\,LBP(n+L)\}$$

تعداد L نمونه از کانتور توان (سمت راست) فریم n ام باشند. این نمونه‌ها را بصورت صعودی مرتب مینمائیم و نتیجه را $SRP_R(n,i)$ می‌نامیم.

$$SRP_R(n,i) = \{SLBP(n, j_i)\} \quad that \quad SLBP(n, j_i \le SLBP(n, j_{i+1}) \quad و \quad i = 0 \text{ تا } L \qquad (۱۱)$$

به طریق مشابه L نمونه از کانتور توان (سمت چپ) فریم nام را بصورت صعودی مرتب نموده و نتیجه را $S\,LP_L(n,i)$ می‌نامیم.

$$SLP_L(n,i) = \{SLBP(n-j_i)\} \quad that \quad SLBP(n-j_i \le S\,LBP(n-j_{i+1}) \qquad (۱۲)$$

بر روی بردار مرتب شده توان (متأثراز بافت راست)، دو پارامتر میانگین و واریانس توان پایین و بالا $LP_{L,L}(n)_1$ و $HP_{L,L}(n)_2$ بشرح زیر تعریف مینمائیم:

$$RC\text{-}LP_L(n) = \frac{1}{L_1} \sum_{i=0}^{L_1} [SRP_L(n,i)] \qquad (۱۳)$$

$$RC\text{-}LP\text{-}SD_L(n) = \left\{\frac{1}{L_1}\sum_{i=0}^{L_1}\left[SRP_L(n,i) - (RC\_LP_{L_2}(n)^2\right]\right\}^{\frac{1}{2}} \quad (14)$$

$$RC\text{-}HP_L(n) = \frac{1}{L_2}\sum_{i=L-L_2}^{L} SRP_L(n,i) \quad (15)$$

$$RC\text{-}LP\text{-}SD_L(n) = \left\{\frac{1}{L_2}\sum_{i=L-L_2}^{L}\left[SRP_L(n,i) - (RC\_HP_{L_2}(n)^2\right]\right\}^{\frac{1}{2}} \quad (16)$$

بطریق مشابه پارامترهای فوق را برای کانتور توان متاثر از بافت چپ نیز تعریف مینمائیم:

$$LC\_LP_L(n) = \frac{1}{L_1}\sum_{i=0}^{L_1} SLP_L(n,i) \quad (17)$$

$$LC\_LP\_SD_L(n) = \left\{\frac{1}{L_1}\sum_{i=0}^{L_1}\left[SLP_L(n,i) - (LC\_LP_{L_1}(n)^2\right]\right\}^{\frac{1}{2}} \quad (18)$$

$$LC\_HP_L(n) = \frac{1}{L_2}\sum_{i=L-L_2}^{L} SLP_L(n,i) \quad (19)$$

$$LC\_HP\_SD_L(n) = \left\{\frac{1}{L_2}\sum_{i=L-L_2}^{L}\left[SLP_L(n,i) - (LC\_HP_{L_2}(n)^2\right]\right\}^{\frac{1}{2}} \quad (20)$$

L را عمق نفوذ توان بافت چپ یا راست می‌گوئیم. L_1 و L_2 طول فریم‌ها برای متوسط‌گیری توان پائین و بالا بر روی داده‌های مرتب شده می‌باشند.

۲-۳ـ محاسبه نقاط با حداکثر اطلاعات (M IP)

الف ـ انتخاب کاندید اولیه

فرض میکنیم نقطه A اولین فریم بر روی کانتور اندازه پریودیک بودن و کانتور توان هموار شده باشد(وقتی از سمت چپ به سمت راست حرکت می‌کنیم) که هر دو شرط زیر را دارا می‌باشد:

$$RC\text{-}PM(n=A) \geq th_{PM} \quad \text{و} \quad RC\text{-}HP_L(n=A) \geq th_{SP} \quad (21)$$

نقطه (A) را بعنوان کاندید اولیه نقطه شروع با حداکثر اطلاعات (BMIP)[53] در نظر می‌گیریم. L را طوری انتخاب می‌نمائیم که عمق نفوذ سمت راست حدود ۱۵۰ تا ۲۰۰ میلی ثانیه گردد. مقدار L_1 را برابر $L/3$ در نظر میگیریم. در این مرحله لازم است که تست غیر ایستانی[54] را انجام دهیم.

ب ـ اندازه گیری غیر ایستانی

فرض میکنیم ($LC\text{-}LP\text{-}SP_{L_1}(n=A)$, $LC\text{-}LP_L(n=A)$) توان پائین (نماینده نویز زمینه تداوم‌دار) و انحراف معیار آن برای بافت چپ نقطه n=A باشند. همچنین فرض میکنیم($cg(n,k)$ [55] (۱۲تا $k < 1$)، گین کانالهای طیفی در مقیاس mel،

بعنوان نماینده اطلاعات پوش طیف فریم nام باشند. مقادیـــر میانگین و واریانس پارامترهای طیـــفی کلیه فریمهای متولی از n =A تا n = A-L را که توان هموار شده آنها از سطح آستانه th_{NP} کمتر باشد محاسبه مینمائیم.

$$th_{NP} = LC\_LP_{L_1}(n = A) + LC\_LP\_SD_{L_1}(n = A) \qquad (۲۲)$$

$$NCG(n = A,k) = \frac{\sum_{i=0}^{s} SP(n - j_i).cg(n - j_i, k)}{\sum_{i=0}^{s} SP(n - j_i)} \quad ; \quad k = ۱ \text{ تا } ۱۲ \qquad (۲۳)$$

s برابر مجموع تعداد فریمهائی است که توان هموار شده آنها (SP) از سطح آستانه th_{NP} کمتر می‌باشد.

$$NCG\text{-}VAR(n = A,k) = \left\{ \frac{\sum_{i=0}^{s} S^2 P(n - j_i)[cg(n - j_i, k) - NCG(n = A, k)]^2}{\sum_{i=0}^{s} S^2 P(n - j_i)} \right\} ; \quad k = ۱ \text{ تا } ۱۲ \qquad (۲۴)$$

در محاسبه روابط (۲۳) و (۲۴) فرض بر این است که $SP(n - j_i) \le th_{NP}$ می‌باشد (S تا ۱ = i).
مقدار NCG (n=A,k) و NCG –VAR (n=A,k) بعنوان میانگین و واریانس طیف نویز زمینه تداومدار در نظر گرفته میشود.
حـال بـا فـرض ایـنکه NCG (n=A,k) (۱۲ تـا ۱=k) مقادیـر میانگیـن پارامترهای طیفی نویز زمینه است، واریانس نسبی (پارامترهای فوق) یک قطعه از سیگنال از نقطه n=A تا $n = A+M_1$ بشرح زیر بدست می‌آوریم (M_1 طوری انتخاب می‌شود که طول قطعه حدود ۲۰۰ تا ۱۵۰ میلی ثانیه گردد).

$$CG\_VAR(n = A,k) = \frac{\sum_{i=0}^{i=M_1} S^2 P(n + i)[cg(n + i, k) - NCG(n = A, k)]^2}{\sum_{i=0}^{i=M_1} S^2 P(n + i)} \qquad (۲۵)$$

بـا استفاده از دو دسته واریانس، یکی نماینده سطح تغیرات پارامترهای طیفی نویز زمینه تداومدار و دیگری نماینده انحراف سیگنال تحت بررسی از نویز تداومدار، پارامتری بنام اندازه غیر ایستانی NSM (n=A) بشرح زیر تعریف مینمائیم

$$NSM(n=A) = \frac{۱}{۱۲} \sum_{K=۱}^{۱۲} \frac{CG\_VAR(n = A,k)}{NCG\_VAR(n = A,k)} \qquad (۲۶)$$

چـنانچه مشخصـات طـیف قطعـه مـورد بررسی تقریباً مساوی طیف نویز تداومدار باشد (پریود یک یا غیر پریودیک)، مقدار NSM(n=A) در محـدوده عدد یک در نوسان می‌باشد. ولی چنانچه قطعه مورد نظر مربوط به آوای مصوت[۵۷] یک تلفظ گفتاری باشد، مقدار NSM(n=A بسمت عددی بزرگتر از یک سوق پیدا میکند. بنابراین با طراحی یک سطح آستانه مناسب و بزرگتر از یـک میـتوان مقـدار غیر ایستانی را مشخص نمود. فرض میکنیم سطح آستانه غیر ایستانی برابر th_{NSM} باشد (که بطور تجربی تعیین می‌گردد).

ج ـ تعیین نقطه شروع و پایانی M IP یک تلفظ گفتاری

فـرض میکنـیم کـه نقطـه B فریمـی باشد که شرایط زیر را دارا بوده (سه شرط حداقل اندازه توان، حداقل اندازه پریودیک

بودن و حداقل اندازه غیر ایستانی) و فریم ماقبل آن یکی از شروط فوق را حائز نمیباشد.

$$[(RC\_PM(n=B) \geq th_{PM} \text{ and } RC\_HP_{L_2}(n=B) \geq th_{SP} \text{ and } NSM(n=B) \geq th_{NSM})]$$

$$[RC\_PM(n=B-1) < th_{PM} \text{ or } RC\_HP_{L_2}(n=B-1) < th_{SP} \text{ or } NSM(n=B-1) < th_{NSM})] \quad (۲۷)$$

نقاط فوق بازاء هر سیلاب معمولاً یکبار اتفاق افتاده و در ابتدای واج مصوت سیلاب قرار می‌گیرند. این نقاط را اصطلاحاً نقاط MIP شروع می‌نامیم. هر کلمه بر حسب اینکه از چند سیلاب تشکیل شده باشد، چندین نقطه MIP شروع درآن آشکار سازی خواهد شد. اولین نقطه MIP شروع را اصلاحاً MIP شروع کلمه (یا تلفظ گفتاری) می‌نامیم. چنانچه یک واقعه اکوستیکی (پف، کلیک و ته سرفه و...) زنگ تلفن، صدای موتور کولر، پنکه و فن کامپیوتر و...) داشته باشیم که نقطه MIP شروع برای آن قابل تعیین نباشد(یعنی چهار شرط ذکر شده را بطور همزمان ارضاء ننماید). بعنوان وقایع اکوستیکی غیر گفتاری لحاظ شده و از بازشناسی آن صرفنظر مینمائیم.

برای تعیین نقاط MIP پایانی، بطریق مشابه عمل مینمائیم. البته برای تعیین نقاط MIP پایانی، از اندازه توان، اندازه پریودیک بوده و اندازه غیر ایستانی بافت چپ استفاده مینمائیم. بنابراین برای تعیین نقاط MIP پایانی، از شروط زیر استفاده مینمائیم:

$$[LC\_PM(n=E) \geq th_{PM} \text{ and } LC\_HP_{L_1}(n=E) \geq th_{SP} \text{ and } NSM(n=E) \geq th_{NSM}] \text{and} \quad (۲۸)$$
$$[LC\_PM(n=E+1) < th_{PM} \text{ or } [LC\_HP_{L_2}(n=E+1) < th_{SP} \text{ or } NSM(n=E+1) < th_{NSM}]$$

در محاسبه اندازه غیر ایستانی نقاط MIP پایانی، از مدل نویز زمینه تداومدار اولین نقطه MIP شروع استفاده مینمائیم. شکل(۴) و (۵) نقاط MIP شروع و پایانی یک کلمه (تک سیلابی) و یک کلمه (دو سیلابی) در دو حالت $SNR \approx ۱۰dB$ و $SNR \geq ۳۰dB$ نویز سفید را نشان میدهد. ملاحظه میشود که نقاط فوق در مقابل نویز بسیار مقاوم و مطمئن عمل مینمایند.

۲-٤- تشخیص ابتدا و انتهای تلفظ گفتاری

فرض میکنیم B_1 اولین نقطه MIP شروع یک تلفظ گفتاری باشد. هر کلمه از چندین سیلاب تشکیل شده است. سیلابهای زبان فارسی دارای ساختاری CV[۵۸] و CVC و CVCC[۵۹] میباشند. بنابراین هر کلمه‌ای با یک واج غیر مصوت آغاز میشود. طبق تحقیقات تجربی محققین این نوشتار، واجهای غیر مصوتِ غیر صدادار طویل که در ابتدای کلمات واقع میشوند شامل واجهای(س، ح، ف، ش، خ، چ، ...) می‌باشند. طبق اندازه‌گیری بعمل آمده فاصله فریم شروع واجهای فوق‌الذکر تا نقطه MIP شروع هر کلمه حداکثر ۳۰۰ میلی ثانیه می‌باشد. بنابراین با قطعیت کافی میتوان گفت که در محاوره عادی، ابتدای کلمات حداکثر حدود ۳۰۰ میلی ثانیه ماقبل نقطه MIP شروع کلمه آغاز میشوند. در نتیجه چنانچه نقطه MIP شروع کلمه در اختیار باشد، کافی است جستجوی تشخیص ابتدای کلمه را صرفاً از ۳۰۰ میلی ثانیه ماقبل نقطه MIP شروع کلمه اجرا نمائیم. همچنین طول دوره مجموعه واجهای غیر مصوت غیر صدادار که در انتهای کلمات واقع میشوند حداکثر ۶۰۰ میلی ثانیه بعد از نقطه MIP پایانی کلمه می‌باشند. بنابراین چنانچه نقطه MIP پایانی کلمه در اختیار باشد، کافی است جستجوی تشخیص انتهای کلمه را صرفاً ۶۰۰ میلی ثانیه بعد از نقطه MIP پایانی کلمه اجرا نمائیم. حال برای تشخیص ابتدا و انتهای تلفظ گفتاری، و همچنین متمایزسازی فریمهای گفتاری از نویزهای زمینه (بین سیلابی و غیره و...) از ایده‌های مطرح شده در مراجع[۳] و[۹] بشرح زیر استفاده مینمائیم.

فرآیند کلاس‌بندی گفتار از نویز زمینه را بر روی کلیه فریمهای موجود در بازه زمانی ۳۰۰ میلی ثانیه قبل از نقطه MIP آغازین تا ۶۰۰ میلی ثانیه بعد از نقطه MIP پایانی تلفظ گفتاری اجرا مینمائیم. فرض میکنیم NSM(n) اندازه غیر ایستانی فریم nام نسبت به نویز زمینه تداومدار باشد (با فرض $M_1=1$ در رابطه ۲۵). همچنین NP و NVAR مقادیر متوسط و انحراف معیار توان (بر حسب dB) نویز زمینه تداومدار باشد (که در نقطه MIP آغازین تلفظ گفتاری تخمین زده میشود). البته چنانچه NVAR از

یک سطح آستانه حدود (۰/۵ dB) کمتر باشد آنرا با سطح آستانه فوق جایگزین می‌کنیم. همچنین SP(n) توان هموار شده فریم n ام (بر حسب dB) و NF-COUNT یک کانتر برای شمارش تعداد فریمهای سکوت یا نویز زمینه و SF-COUNT یک کانتر برای شمارش تعداد فریمهای گفتاری باشند.

فریم n ام دربازه زمانی ذکر شده را موقعی بعنوان فریم گفتاری کلاس‌بندی مینمائیم که یکی از شروط زیر را دارا باشد:

الف $SP(n) \geq ۱/۷۵ \quad NVAR + NP$

ب $SP(n) \geq NP+NVAR \quad and \quad NSM(n) \geq ۱/۲۵$

ج $SP(n) \geq NP \text{ and } NSM(n) \geq ۱/۷۵$

کلاس‌بندی فریم‌های سیگنال دریافتی به یکی از دو کلاس گفتاری و غیرگفتاری بر مبنای منطق ذکر شده انجام می‌پذیرد. به منظور تشخیص فواصل سکوت درون کلمه‌ای و همچنین تشخیص انتهای تلفظ گفتاری، فرآیند کلاس‌بندی را در سه ناحیه متوالی بطور متمایز انجام میدهیم.

الف ـ کلاس‌بندی فریمها در ناحیه ماقبل نقطهMIP آغازین کلمه (برای طول دوره‌ای در حدود حداکثر ۳۰۰ میلی ثانیه)

ب ـ کلاس‌بندی فریمها در ناحیه ما بین نقطه MIP آغازین و پایانی کلمه

ج ـ کلاس‌بندی فریمها در ناحیه ما بعد نقطه MIP انتهائی کلمه (برای طول دوره‌ای در حدود حداکثر ۶۰۰ میلی ثانیه).

در ناحیه الف؛ قبل از فرآیند کلاس‌بندی، کانترهای SF-COUNT و NF-COUNT را صفر می‌کنیم سپس کلاس‌بندی فریمها را بر مبنای منطق ذکر شده انجام میدهیم. بازاء هر بار کلاس‌بندی فریمها به گفتار، کانتر SF-COUNT یکی افزود میشود. بعد از اینکه SF -COUNT ≥ ۱ گردید، کانتر NF-COUNT نیز در صورت مشاهده هر بار فریم نویز زمینه یکی افزایش می‌یابد. پس از رسیدن به نقطه MIP آغازین کلمه، چنانچهNF-COUNT > SF-COUNT گردد، فریمهای شروع تا نقطه پیوسته گفتار در نزدیکی نقطه MIP آغازین بعنوان سکوت یا نویزهای اکوستیکی غیر گفتاری (کلیک، پف و ...) در نظر گرفته میشود.

در ناحیه ب؛ کانترهای SF-COUNT و NF-COUNT را صفر میکنیم. با مشاهده یک فریم گفتاری (بر مبنای منطق کلاس‌بندی ذکر شده)، به کانتر SF-COUNT یکی می‌افزائیم. همچنین با مشاهده یک فریم غیر گفتاری (سکوت یا نویز زمینه)، به کانتر NF-COUNT یکی اضافه میکنیم. در طی انجام فرآیند فوق دو وضعیت ممکن است اتفاق بیافتد؛

۱ـ تعداد فریمهای غیرگفتاری NF-COUNT از یک سطح آستانه بیشتر گردد (حدوداً ۴۰۰ میلی ثانیه، یعنی حدود ۱۰۰ میلی ثانیه بیشتر از حداکثر فاصله سکوت سیلابهای متوالی درون کلمه‌ای). در این صورت آخرین فریمی که گفتار کلاس‌بندی شده است به عنوان انتهای گفتار تلقی شده و فرآیند تشخیص انتهای گفتار خاتمه می‌یابد.

۲- به یک نقطه MIP پایانی می‌رسیم و در این حالت تعداد فریمهای غیر گفتاری NF-COUNT صفر بوده و یا عدد کوچکی می‌باشد. در این صورت کلاس‌بندی ناحیه (ج) را بشرح زیر ادامه میدهیم.

در ناحیه ج؛ مجدداً کانترهای SF-COUNT و NF-COUNT را صفر می‌کنیم. بر مبنای منطق ذکر شده، فریمها را به کلاسهای گفتاری و غیر گفتاری کلاس‌بندی میکنیم و کانترهای متناظر را (با مشاهده هر کلاس) یکی افزایش میدهیم. در طی انجام فرآیند فوق دو وضعیت ممکن است اتفاق بیافتد؛

۱- تعداد فریمهای غیرگفتاری NF-COUNT از یک سطح آستانه بیشتر گردد (حدود ۴۰۰ میلی ثانیه)، در این صورت آخرین فریمی که گفتار کلاس‌بندی شده است بعنوان انتهای گفتار تلقی شده و فرآیند تشخیص انتهای گفتار خاتمه می‌یابد.

۲ـ تعداد فریمهای غیر گفتاری NF-COUNT کمتر از ۳۰۰ میلی ثانیه است، ولی NF-COUNT + SF-COUNT بیشتر از ۶۰۰ میلی ثانیه گردیده است. در این صورت آخرین فریم گفتاری بعنوان فریم انتهای تلفظ گفتاری تلقی شده و فرآیند تشخیص انتهای تلفظ گفتاری خاتمه می‌یابد.

تشخیص نقطه MIP پایانی تلفظ گفتاری

تشخیص نقطه MIP آغازین تلفظ گفتاری نسبتاً ساده است. زیرا اولین نقطه MIP شروع که با آن مواجه شدیم بعنوان نقطه MIP آغازین تلفظ گفتاری در نظر گرفته میشود. اما در طی فرآیند کلاس‌بندی فریمها به دو کلاس گفتار و غیر گفتار

(سکوت، نویز زمینه) با نقاط MIP پایانی متعددی مواجه میشویم (بر حسب اینکه تلفظ گفتاری از چند کلمه یا چندین سیلاب تشکیل شده باشد). در این حالت می‌بایستی آخرین نقطه MIP پایانی را برای آن تلفظ گفتاری با دقت و اطمینان بالائی تعیین نمائیم نقطه MIP پایانی درون کلمه‌ای که حائز یکی از شرایط زیر باشد بعنوان نقطه MIP پایانی تلفظ گفتاری در نظر گرفته میشود:

۱ـ برای یک دوره حدود ۵۰۰ میلی ثانیه بعد از آن، بیشتر از ۸۵ درصد فریمها، غیر گفتاری کلاس‌بندی شده باشند.

۲ـ برای یک دوره حدود ۷۰۰ میلی ثانیه‌ای بعد از آن نقطه، بیشتر از ۹۰ درصد فریمها، غیر گفتاری یا گفتاری ولی غیر پریودیک (بی صدا) کلاس‌بندی شده باشند.

۳ـ نتایج پیاده‌سازی

فرکانس نمونه‌برداری تمامی داده‌های گفتاری مورد استفاده KHz ۸ بوده و نمونه ها بصورت PCM، ۱۶ بیتی ضبط شده‌اند. پایگاه داده مورد استفاده ۱۵۰۰ کلمه پر مصرف در کاربردهای IT [۶۰] و سرویسهای مخابراتی بوده است. داده‌های گفتاری توسط میکروفون جهتی خوب و تحت شرایط SNR> ۳۰ dB ضبط شده‌اند. ۱۰ نفر گوینده (۷ نفر مرد و ۳ نفر زن) فرمانها و تقاضاهای صوتی را (با ۱۰ بار تکرار از هر فرمان) تلفظ نموده‌اند. بین هر دو فرمان یا تقاضای صوتی حداقل یک ثانیه سکوت برقرار بوده است. برای ارزیابی مقاوم بودن الگوریتمها، از دو گروه نویز غیرپریودیک (نویز سفید[۶۱]، نویز همهمه[۶۲]، نویز صورتی[۶۳]، نویز اداری) و نویز پریودیک(زنگ تلفن، صدای موتور کولر، صدای موتور یخچال) با نسبت SNR های مختلف (dB ۲۰و ۱۵و ۱۰و ۵) استفاده نمودیم. برای اضافه کردن نویزها به سیگنال از روشهای مندرج در توصیه نامه‌های ITU استفاده نمودیم. به منظور بررسی عملکرد روش ارائه شده در مقابل روشهای مؤثر دیگر، دو روش جدید و مطرح در مرجع [۱] و[۳] نیز پیاده‌سازی شده‌اند. در بخشهای زیر روش مندرج در مرجع[۱] را BE۱ و روش مندرج در مرجع[۳] را BE ۲ و روش ارائه شده در این تحقیق را MIP نامیدیم. آزمونهای متعددی برای ارزیابی عملکرد روش MIP در تشخیص ابتدا و انتهای گفتار بشرح زیر انجام دادیم:

الف ـ آزمون‌های ارزیابی دقت تشخیص

ابتدا و انتهای کلیه کلمات گفتاری ضبط شده از همه گویندگان بصورت دستی(با استفاده از نرم‌افزارهای بسیار دقیق و بکارگیری افراد خبره) بر چسب زده شده‌اند. در آزمون اول هدف بررسی دقت تشخیص الگوریتم‌ها بوده است. در این آزمون بر چسب دستی بعنوان بر چسب مرجع مدنظر قرار گرفته است. اختلاف زمانی بر چسب الگوریتمها در مقایسه با برچسب دستی بعنوان معیار دقت، مورد توجه قرارگرفته است. جدول (۱) مقادیر متوسط و انحراف معیار اختلاف زمانی برچسبهای اتوماتیک و دستی را برای SRN ≥ ۳۰dB (بدون اضافه کردن نویز)برای تشخیص ابتدای کلمه را (بر حسب میلی ثانیه) نشان میدهد. جدول (۲) این اختلاف زمانی را برای تشخیص انتهای کلمات گفتاری نشان میدهد. با ملاحظه جداول (۱) و (۲) مشاهده میشود که؛ دقت روش MIP برای گفتار تمیز(بدون اضافه کردن نویز) بهتر از روشهای BE ۱ و BE۲ می‌باشد. همچنین مشاهده میشود که دقت هر سه روش برای تشخیص ابتدای کلمه بهتر از انتهای کلمه میباشد. برای ارزیابی بهتر روشها، یک سیستم بازشناسی گفتار گسسته بصورت وابسته به گوینده با استفاده از مدلسازی HMM پنج حالته طراحی و پیاده‌سازی نمودیم. نتایج دقت بازشناسی در جدول (۳) درج گردیده است. در این آزمون از گفتار تمیز (بدون اضافه کردن نویز) استفاده نموده و برای هر حالت مدل، ۴ مخلوط[۶۴] طراحی کرده‌ایم. ۱۲ ضریب MFCC و ۱۲ ضریب مشتق اول آن و مشتق لگاریتم گین با آنالیز PLP استخراج شده‌اند، بعنوان ویژگی طیفی مورد استفاده قرار گرفته‌اند.

از روی جدول (۳) مشاهده میشود که دقت بازشناسی بروش برچسب اتوماتیک MIP در حد روش دستی و بهتر از روشهای اتوماتیک دیگر می‌باشد.

ب ـ آزمونهای ارزیابی مقاوم بودن تشخیص ابتدا و انتهای گفتار در مقابل نویزهای غیرپریودیک

در این آزمون با اضافه کردن انواع نویزهای غیرپریودیک با نسبت SNR مختلف، مقاوم بودن الگوریتمها در مقابل نویز را مورد ارزیابی قرار دادیم. جدول (۴)تا (۷) نتایج این آزمونها را برای انواع نویزهای غیر پریودیک نشان میدهد. با مروری بر نتایج حاصل، مشاهده میشود که روش MIP دارای عملکردی نزدیک برچسب زدن دستی و مقاوم‌تر از روشهای دیگر می‌باشد.

همچنین ملاحظه میشود، بهبودی که حاصل شده است برای انواع نویزهای غیرپریودیک و برای انواع سطوح SNR برقرار است. همچنین مشاهده میشود که بدترین نوع نویز در بازشناسی گفتار، نویز غیر ایستان همهمه می‌باشد.

ج ـ آزمونهای ارزیابی مقاوم بودن تشخیص ابتدا و انتهای گفتار در مقابل نویزهای پریودیک

نویزهای پریودیک که نوعاً در محیطهای اداری و منزل با آن مواجه هستیم، صدای زنگ تلفن، صدای فن، (فن کوئل) صدای موتور کولر می‌باشند. این صداها تلفیقی از نویزهای سفید (با توان کمتر) و نویزهای پریودیک (با توان بیشتر) می‌باشند. جداول (۸ تا ۱۰) نتایج بازشناسی را برای سه نوع آزمون نویز پریودیک و سطوح SNR مختلف نشان میدهد با مروری بر نتایج حاصل مشاهده میکنیم که روش MIP در مقابل نویزهای پریودیک نیز مقاوم تر از روشهای دیگر می‌باشد.

٤ـ جمع‌بندی و نتیجه گیری

در این تحقیق روشی مؤثر و مطمئن و دقیق برای تشخیص ابتدا و انتهای گفتار ارائه نمودیم. از ویژگیهای مهم روش ارائه شده آن است که؛ از تمامی اطلاعات مفید و قابل ارائه سیگنال گفتار شامل؛ پریودیک بودن، بالا بودن سطح انرژی قسمتهای باصدا نسبت به قسمتهای بی صدا، اندازه غیر ایستانی قسمتهای گفتار نسبت به نویزهای تداوم‌دار محیطی و ... به نحو مؤثری برای جداسازی گفتار از نویزهای زمینه استفاده مینماید. یکی از ویژگیهای مؤثر روش ارائه شده آن است که هم برای نویزهای پریودیک و هم برای نویزهای غیر پریودیک دارای دقت لازم می‌باشد. از ویژگیهای دیگر روش ارائه شده آن است، که هم برای سیستم‌های زمان زنده و هم برای سیستم دسته‌ای قابل پیاده‌سازی است. یکی از مشکلات عمده روشهای دیگر، تخمین پارامترهای مدل نویز زمینه تداوم‌دار است. در اغلب روشهای کلاسیک، از اطلاعات چندین فریم آغازین ضبط برای مدلسازی نویز استفاده میشود. در روش MIP، مشخصات نویز زمینه تداوم‌دار با بهترین دقت و مستقل از وقایع اکوستیکی اطراف گفتار تخمین زده میشود. همچنین این تخمین برای کاربردهای زمان زنده که فرآیند ضبط ادامه دارد (با فواصل سکوت بیش ازیک ثانیه) و امکان دارد وضعیت نویززمینه تغییر یابد بسیار مفید و مؤثر است. برای ارزیابی کارائی روش MIP در تشخیص ابتدا و انتهای گفتار، یک سیستم بازشناسی گفتار وابسته به گوینده با دایره لغات ۱۵۰۰ کلمه ای طراحی و پیاده‌سازی گردید. تمامی نتایج حاصل برای انواع نویزهای پریودیک و غیرپریودیک و همچنین برای انواع سطوح توان نویز، مؤید دقت و برتری عملکرد روش MIP نسبت به روشهای کلاسیک موجود می‌باشد.

جدول (۱) اختلاف زمانی (بر حسب میلی ثانیه) بر چسب دستی و روشهای اتوماتیک در تشخیص ابتدای گفتار.

روش	متوسط	انحراف معیار
BE۱	۲۰/۸	۲۱/۵
BE۲	۱۴/۵	۱۵/۶
MIP	۸/۷	۱۰/۴

جدول (۲) اختلاف زمانی (بر حسب میلی ثانیه) برچسب دستی و روشهای اتوماتیک در تشخیص انتهای گفتار.

روش	متوسط	انحراف معیار
BE۱	۳۵/۷	۳۰/۸
BE۲	۲۸/۴	۲۲/۵
MIP	۱۷/۶	۱۵/۴

جدول (۳) نتایج بازشناسی گفتار تمیز (بروشHMM).

روش	متوسط دقت بازشناسی کلمات	انحراف معیاردقت بازشناسی کلمات مختلف
برچسب دستی	۹۷/۳	۲/۴
BE۱	۹۴/۳	۳/۸
BE۲	۹۵/۷	۳/۱
MIP	۹۶/۹	۲/۶

جدول (٤) نتایج دقت بازشناسی برای نویز غیر پریودیک سفید(White Noise).

روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠	SNR = ٥ dB
بر چسب دستی	٩۶/٨	٩۵/٣	٩۴/١	٩١/٨
روش BE١	٩١/٢	٨٨/٧	٨۵/۴	٨١/۶
روش BE٢	٩۴/٣	٩١/۵	٨٨/٣	٨۴/٢
روش MIP	٩۶/٣	٩۴/۵	٩٢/٧	٨٩/٨

جدول (٥) نتایج دقت بازشناسی برای نویز غیر پریودیک صورتی(Pink Noise).

روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠ dB	SNR = ٥ dB
بر چسب دستی	٩۶/٧	٩۵/٢	٩٣/٩	٩١/٢
BE١	٩١/٠	٨٨/٣	٨۵/٠	٨٠/٢
BE٢	٩۴/١	٩١/٢	٨٧/٨	٨٣/٣
MIP	٩۶/٢	٩۴/٣	٩٢/٢	٨٩/۵

جدول (٦) نتایج دقت بازشناسی برای نویز غیرپریودیک محیط اداری(Office Noise).

روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠ dB	SNR = ٥ dB
بر چسب دستی	٩۶/٩	٩۵/۵	٩۴/٣	٩٢/١
BE١	٩١/۴	٨٩/٠	٨۵/۵	٨١/٩
BE٢	٩۴/۵	٩١/٧	٨٨/۶	٨۴/٧
MIP	٩۶/۵	٩۴/٧	٩٣/١	٩٠/١

جدول (٧) نتایج دقت بازشناسی برای نویز غیر پریودیک همهمه(Babble Noise).

روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠ dB	SNR = ٥ dB
بر چسب دستی	٩۶/٢	٩۴/۴	٩٠/٣	٨٧/۶
BE١	٩٠/٢	٨۵/٧	٨٢/۴	٧٨/١
BE٢	٩٣/٣	٨٩/٨	٨٣/٧	٧٩/٩
MIP	٩۵/٨	٩٣/۵	٨٩/۵	٨۵/٧

جدول (٨) نتایج دقت بازشناسی برای نویز پریودیک زنگ تلفن.

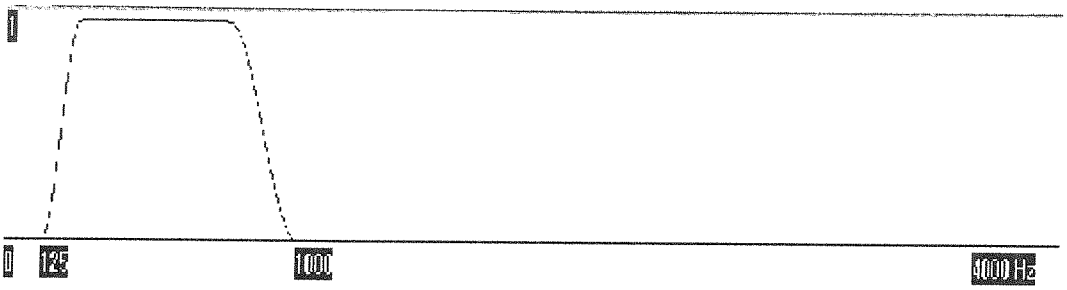
روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠ dB	SNR = ٥ dB
بر چسب دستی	٩۶/٨	٩۵/۴	٩٣/٧	٩٠/١
BE١	٩١/۶	٨٩/٩	٨۴/٢	٨٠/۴
BE٢	٩۴/٧	٩١/٨	٨٧/٣	٨١/۶
MIP	٩۶/۵	٩۴/٩	٩٢/۴	٨٨/٨

جدول (٩) نتایج دقت بازشناسی برای نویز پریودیک فن کوئل.

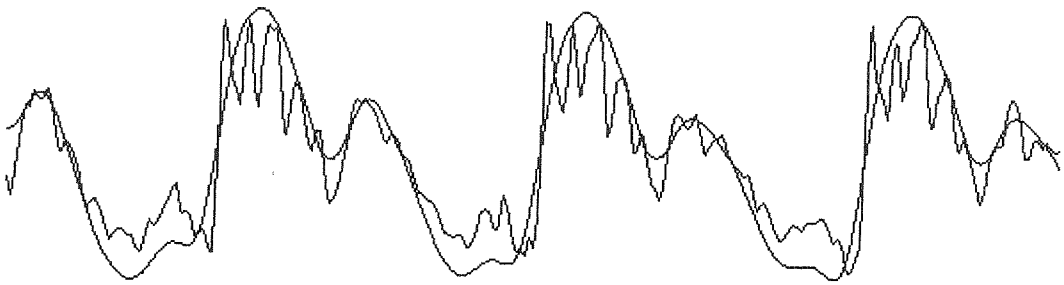
روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠ dB	SNR = ٥ dB
بر چسب دستی	٩۶/۴	٩۵/٢	٩٣/٩	٩١/۴
BE١	٩١/١	٨٩/٢	٨۵/١	٨١/١
BE٢	٩۴/٢	٩١/٧	٨٨/۴	٨٣/٧
MIP	٩۶/١	٩۴/٣	٩٢/٨	٨٩/٧

جدول (١٠) نتایج دقت بازشناسی برای نویز پریودیک موتو کولر.

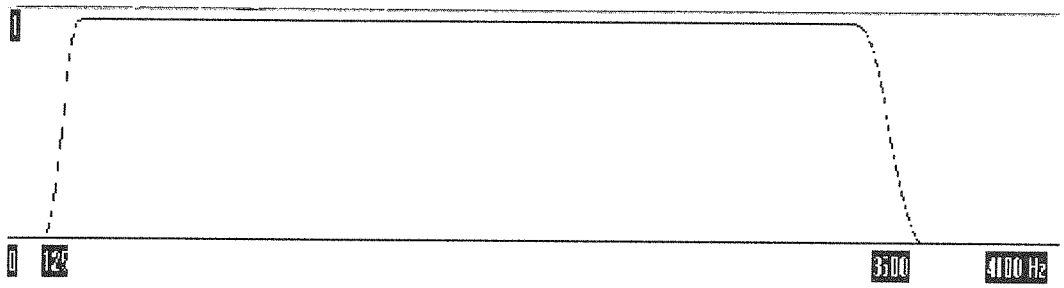
روش	SNR = ٢٠ dB	SNR = ١٥ dB	SNR = ١٠ dB	SNR = ٥ dB
بر چسب دستی	٩۶/٣	٩۵/۴	٩۴/١	٩١/۶
BE١	٩١/٣	٨٩/۶	٨۵/۴	٨١/٩
BE٢	٩۴/۴	٩١/٩	٨٨/٧	٨۴/١
MIP	٩۵/٨	٩۴/۵	٩٣/١	٨٩/٩



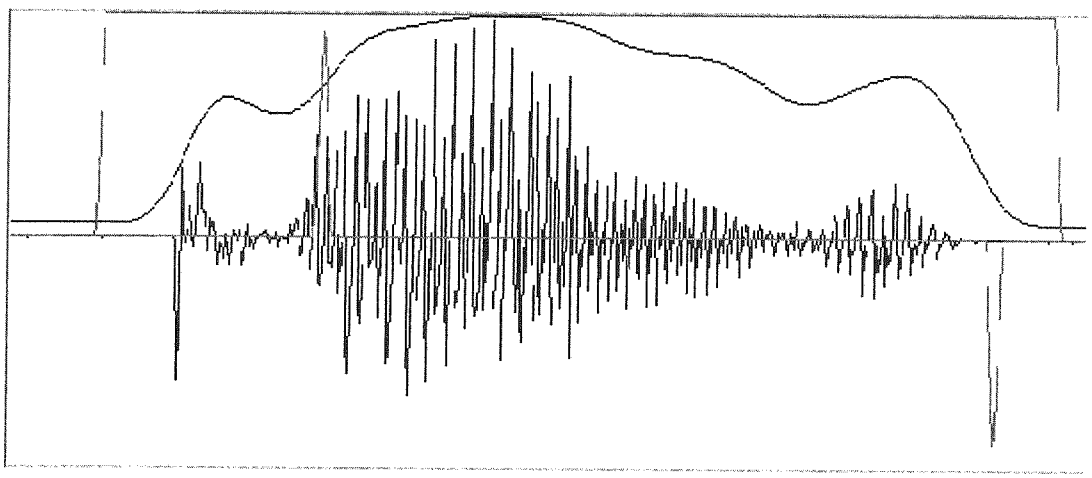
شکل (۱) پاسخ فرکانسی فیلتر FIR میانگذر برای تخمین فرکانس پیچ.



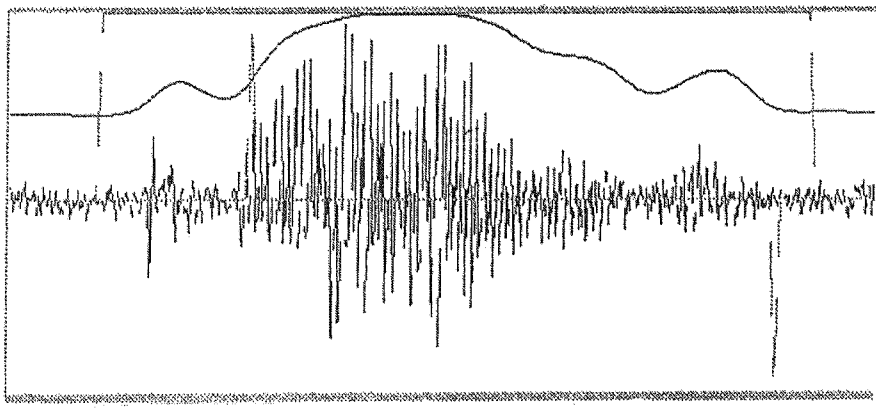
شکل(۲) یک قطعه از سیگنال با صدا، همراه با سیگنال میانگذر شده معادل.



شکل(۳) تابع وزنی در حوزه فرکانس برای محاسبه توان سیگنال.

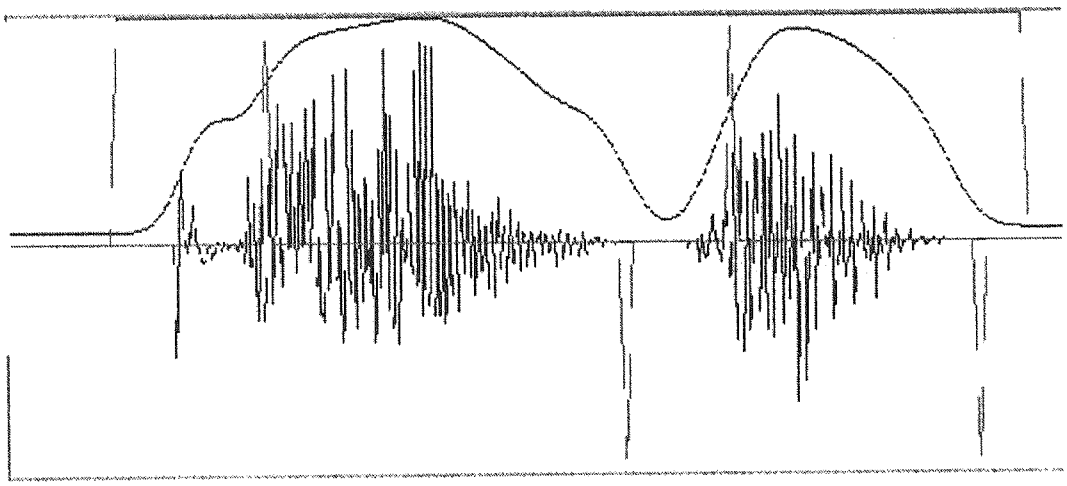


الف ـ کلمه تک سیلابی و $SNR \geq 30dB$

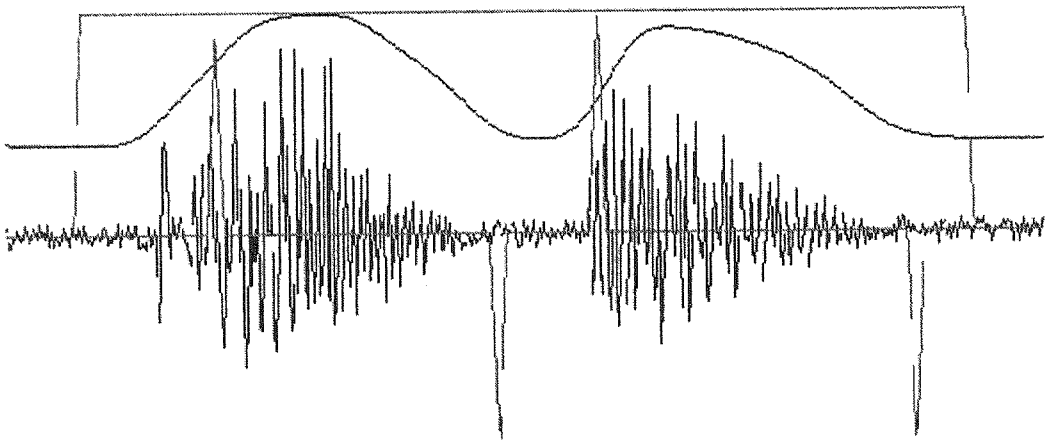


ب ـ کلمه تک سیلابی و SNR ≥ ۱۰ dB

شکل (٤) نقاط MIP شروع و پایانی برای کلمه تک سیلابی.



الف ـ کلمه دو سیلابی و SNR ≥ ۳۰ dB



ب ـ کلمه دو سیلابی و SNR ≥ ۱۰ dB

شکل (٥) نقاط MIP شروع و پایانی برای کلمه دوسیلابی.

واژه‌نامه‌ها

1-Nonspeech Acoustic Events
2- Background noise
3- Speech Begin and Endpoint Detection
4- Endpoint Detection
5- Speech Activity Detection
6- Voiced Activity Detection
7- Impulse Noise
8- Continuant Noise

9- White Noise
10- Color Noise
11- Stationary
12- Nonstationary
13- Fixed Level Noise
14- Variable Level Noise
15- Batch Mode
16- Real Time Mode
17- Explicit
18- Implicit
19- Hybrid
20- Hidden Markov Model Alignment
21- Dynamic Time Warpping Alignment
22- Reliability
23- Robustness
24- Accuracy
25- Adaptation
26- Simplicity
27- Real Time Processing
28- Non Priori Knowledge of Noise
29- Zero Crossing Rate
30- Linear Prediction Error Energy
31- Time Duration
32- Periodicity Measure
33- Mel Frequency Channel Gain
34- Mel Frequency Cepstrum Coefitient
35- Multi State Transition diagram
36- Island of Reliable
37- Babble Noise
38- Maximum Information Point SBED
39- Sorting Filter
40- Periodicity Likelihood Measure
41- Pitch Syncronous Analisis
42- Normalized Cross Correlation
43- Bandpass filtering
44- Finite Impulse
45- Parks Mac Celelan
46- Decimation
47- Right Context Dependent PM
48- Left Context Dependent PM
49- Time Frequency Energy
50- Fast Fourier Transform
51- Log of Bandpass Power
52- Smothed Log of Bandpass Power
53- Sorted Right Context Power
54- Sorsted Left Context Power
55- Low Power
56- High Power
57- Begin of MIP
58- Non Stationatry Measure
59- Vowel
60- Consonant Vowel
61- Consonant Vowel Consonant
62- Inforamation Technology
63- Pink Noise
64- Office Noise
65- Telephone Ring

مراجع

[1] J. C. Junqua, B. Mak, B. Reaves, "A Robust Algorihtm for Word Boundary Detection In The Presence of Noise ", IEEE Transaction On Speech And Audio Processing, Vol.2, No.3, July 1994.
[2] C. T. Ling, J.Y. Lin, G. D. Wu, "A Robust Word Boundary Detection Algorothm For Variable Noise Level Environment In Cars", IEEE Trans. On Intelligent Transportation Systems, Vol.3, No.1, March 2002.
[3] Q. Li, J. Zheng, A. Tsai, Q. Zhou, "Robust Endpoint Detection And Energy Normalization For Real Time Speech And Speaker Recognition" IEEE Trans. On Speech And Audio Processing, Vol.10, No.3, March 2002.
[4] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, "An Improved Endpoint Detector For Isolated word Recognition"IEEE Trans. On Acoustics, Speech And Signal Processing, Vol. Assp –29, No.4, August 1981.
[5] H. Ney, "An Optimization Algorithm For Determining The Endpoints Of Isolated Utterances" Proc. Of ICASSP-81, pp. 720-723,1981.
[6] M. H. Savoji, "A Robust Algorithm For Accurate Endpointing Of Speech Signals" Speech Communication Vol.8, pp.46-60, 1989, North Holland, Elsevier Science Publishers.
[7] C. Tsao, R. M. Gray, "An Endpoint Detector For LPC Speech Using Residual Error Look-Ahead For Vector Quantization Applications" Proc.Of ICASS-82, 1982, pp. 18b-701-4.
[8] B. Reaves, "Comments On An Improved Endpoint Detector For Isolated Word Recgnition" Corresp. Of IEEE Acoust. Speech, Signal Processing, Vol.39, pp.526-527, Feb. 1991.
[9] S. E. B. Ghazale, K. Assleh, "A Robust Endpoint Detection Of Speech For Noisy Environments With Application To Automatic Speech Recognition" In Proc. Of IEEE ICASSP, IV – 3808-3811,2002.
[10] M. Hamada, Y. Takizawa, T. Norimatsu, "A Noise Roboust Speech Recognition System," In Proc. Of ICSLP-90, pp.893-896, 1990.
[11] R. Tucker, "Voice Activity Detection Using A Periodicity Measure" , IEEE Proceedings-1, Vol.139, No.4, August1992.
[12] S. Tanyer, H. Ozer, "Voice Activity Detection In Nonstationary Qausian Noise" In Proc Of ICSLP-98, 1998.
[13] F. Beritelli, S. Casale, A. Cavallaro, "A Robust Voice Activity Detector For Wireless Communications Using Soft Computing" IEEE Journal On Selected Area In Communications, Vol.16, No.9, Decmber 1998.
[14] J. Sohn, N. S. Kim, W. Sung, "A Statistical Model Based Voice Activity Detection" IEEE Signal Processing Letters, Vol.6, No.1, Janurary 1999.
[15] S. G. Tanyer, H. Ozer, "Voice Activity Detection In Nonstationary Noise", IEEE Transction On Speech And Audio Processing, Vol.8, July 2000.
[16] W. Hess. "Pitch Determination Of Speech Signals," New York, Springer, 1983.
[17] Y. Medan, E. Yair, D.Chazan, "Super Resolution Pitch Dtermination Of Speech Signals" IEEE Trans. On Signal Processing, Vol. 39, No. 1, January 1991.
[18] P. Veprek, M. S. Scordilis, "Analysis, Enhancement And Evaluation Of Five Pitch Determination Techniques" Elsevier Trans. On Speech Commumincation Vol.37 , Pp. 249-270, 2002.
[19] A. V. Oppenheim, R. W. Schafer, "Discrete Time Sigmal Processing" Prectice Hall Publisher, 1989.
[20] J. Pencak, D. Nelson, "The NP Speech Activity Detection Algorithm" IEEE Proc. On ICASSP, pp.138-384, 1995.