تقطیع هجایی گفتار پیوسته فارسی با استفاده از آستانه‌گذاری ضرایب موجک و نرم‌سازی فازیِ تابع انرژی

غزال شیخی[۱] – سید حمید محمودیان[۲]

(۱) کارشناس ارشد - مهندسی پزشکی، عضو هیات مدیره شرکت جویشگر ریزگستر، شهرک علمی و تحقیقاتی اصفهان

(۲) استادیار - دانشکده برق، دانشگاه آزاد اسلامی واحد نجف‌آباد

تاریخ دریافت: بهار ۱۳۹۲ تاریخ پذیرش: بهار ۱۳۹۳

خلاصه: امروزه در تحقیقات حوزه پردازش و بازشناخت گفتار، هجا به دلیل ارتباط قوی آن با تولید و ادراک گفتار در انسان، به عنوان یک واحد زیرکلمه‌ای هر روز بیشتر مورد توجه قرار می‌گیرد. آشکارسازی خودکار مرزهای هجایی گامی مهم در تحقیقات مرتبط با نوای گفتار، تولید گفتار طبیعی و حتی بازشناسی گفتار است. در این مقاله روش جدیدی برای آشکارسازی خودکار مرزهای هجایی در سیگنال گفتار پیوسته فارسی با تکیه بر اطلاعات صوتی ارائه شده است. تحقیقات قبلیِ نویسندگان این مقاله، کارآیی نرم‌سازی فازیِ تابع انرژی را در مقایسه با سایر روش‌های به کار رفته در این زمینه نشان می‌دهد. در این تحقیق، پیشنهاد شده است که از روشی مشابه روش‌های متداول حذف نویز از گفتار به وسیله آستانه گذاری ضرایب موجک برای بهبود خطای درج مرز اضافه استفاده شود. این روند، انرژی همخوان‌های بی‌واکی را که در تابع انرژی قله‌های اضافه ایجاد می‌کنند، به شدت کاهش می‌دهد. نتایج نشان می‌دهند با استفاده همزمان از این روش و روش نرم‌سازی فازی تابع انرژی، خطای درج مرز اضافه در حدود ٪8 کاهش می‌یابد؛ بدون آنکه سایر معیارهای کارآیی تحت تأثیر قرار گیرند. با استفاده از روش پیشنهادی بیش از ٪94 از هجاها با خطایی کمتر از 50 میلی‌ثانیه تقطیع می‌شوند.

کلمات کلیدی: تقطیع هجایی،تبدیل موجک، آستانه‌گذاری ضرایب موجک، واکه، همخوان، فیلتر فازی، انرژی زمان کوتاه.

Syllable Segmentation of Farsi Continuous Speech Using Wavelet Coefficients Thresholding and Fuzzy Smoothing of Energy Contour

Ghazaal Sheikhi[1] - Hamid Mahmoodian[2]

(1) MSc - Jooyeshgar Rizgostar Co., Isfahan Science and Technology Town

ghazaal.sheikhi@gmail.com

(2) Assistant Professor - Department of Electrical Engineering, Najafabad Branch, Islamic Azad University

mahmoodian_hamid@yahoo.com

Syllable, as a sub-word unit, nowadays plays an active role in the field of speech processing and recognition research according to its robust relation to human speech production and cognition. Automatic syllable boundaries detection is an important step forward in the areas of speech prosody, natural speech synthesis and speech recognition. In this paper, a novel method in automatic syllabification of Farsi continuous speech based on acoustic structure is proposed. Our previous studies, showed the proficiency of energy contour fuzzy smoothing method, compared with other prominent works in this area. This paper suggests that the conventional methodology-used in speech enhancement based on wavelet coefficient thresholding would improve syllable segmentation by decreasing insertion error. This process declines the energy in high energy consonants which are responsible for extra peaks in short term energy contour. Experimental results showed that utilizing proposed method along with fuzzy smoothing would diminish insertion error about 8% with no reasonable effect on other efficiency criteria. More than 94% of syllables are automatically segmented using presented technique with less than 50ms error.

Index Terms: Syllable segmentation, wavelet transform, wavelet coefficient thresholding, vowel, consonant, fuzzy filter, short term energy.

۱- مقدمه

بازشناسی گفتار یکی از زمینه‌های تحقیقاتی در علوم مهندسی و گفتار است که به سرعت در حال پیشرفت می‌باشد. این تحقیقات در بسیاری از زمینه‌ها کاربرد دارد. از جمله سامانه‌های کنترل شونده با گفتار[۱] در صنایع و کاربردهای روزمره، ابزارهای موردنیاز افراد معلول، تبدیل متن به گفتار، مترجم‌های خودکار و غیره. امروزه سامانه‌های بازشناسی خودکار گفتار[۲] نه تنها برای زبان انگلیسی بلکه در بسیاری دیگر از زبان‌های دنیا از جمله فارسی توسعه یافته‌اند و توانایی بازشناسی هزاران کلمه را دارند.

سامانه‌های بازشناسی گفتار را می‌توان از جهات مختلف طبقه‌بندی کرد: وابسته به گوینده یا مستقل از آن، حساس یا مقاوم به نویز، دامنه لغات بزرگ، متوسط یا کوچک، پیوسته یا تقطیع شده[۳] و سایر طبقه‌بندی‌ها. با وجودی که کارآیی سامانه‌های بازشناس گفتار به راحتی با عواملی چون نویز و تغییرات کانال، گوینده، میکروفن، لهجه، سرعت گفتار و غیره افت می‌کند، انسان تحت تمامی این شرایط عمل بازشناسی را به راحتی انجام می‌دهد. بنابراین آنچه امروزه در این زمینه مورد توجه محققین قرار گرفته است الگوبرداری از نحوه تولید و ادراک گفتار در انسان است.

ادراک گفتار در انسان از طریق سطوح مختلفی از دانش انجام می‌گیرد. از جمله اطلاعات صوتی[۴]، دانش واژگانی[۵] و دانش زبانی[۶]. هرچند پیاده سازی سامانه‌ای با این قابلیت امکان‌پذیر نیست، اما تقریبا تمامی تلاش‌ها در جهت نزدیک شدن به نحوه ادراک گفتار در انسان، نتایج قابل قبولی داشته‌اند. آنچه در تمامی این سامانه‌ها مشترک است آنست که بازشناسی از واحدهای کوچک مثل آوا، دوآوا، سه آوا یا هجا شروع می‌شود و به بازشناسی واحدهای بزرگتر مثل کلمه می‌انجامد.

انتخاب واحد گفتاری که سامانه براساس آن عمل می‌کند بسیار مهم است. یکی از اولین و پر استفاده‌ترین این واحدها واج است. واج‌ها از نظر زبانی به خوبی تعریف شده‌اند و تعداد واج‌ها در هر زبان بسیار محدود است. با این حال این واحدها بسیار وابسته به مفهوم هستند و مدل‌های واجی اثرات هم‌تلفظی[۷] را در مرز آواها در بر نمی‌گیرند [۱]. از طرفی سیگنال گفتار یک سیگنال غیرایستا است و برخی مرزهای واجی در آن به خوبی قابل تعریف نیستند [۲]. به همین دلیل سامانه‌های بازشناسی واج که به صورت لغزنده[۸] روی مسیر سیگنال گفتار حرکت می‌کنند[۳،۴]. نسبت به سامانه‌های مبتنی بر تقطیع اولیه عمومیت بیشتری یافته‌اند [۵،۶]. در هر صورت هر دو نوع این سامانه‌ها در مدل کردن دینامیک‌های سیگنال گفتار دچار محدودیت هستند.

یک راه‌حل استفاده از مجموعه‌ای از آواها مثل دوآوا است. اما این واحدها بیش از آنکه به تولید و ادراک طبیعی گفتار در انسان مرتبط باشند، راه‌حلی نظری برای مدل کردن اثرات هم‌تلفظی و دینامیک‌های گفتار هستند. در صورتی که اگر از هجا به عنوان واحد بازشناسی استفاده شود این ویژگی‌ها به نحو بسیار طبیعی‌تری مدل می‌شوند. هرچند در هنگام استفاده از هجا تعداد واحدها به نحو قابل ملاحظه‌ای از واحدهای واجی بیشتر می‌شود اما هجا بسیار کمتر از واج به مفهوم وابسته است و اثرات هم تلفظی در مرز آواها به نحو مطلوبی در مدل‌های هجایی لحاظ می‌شوند [۷]. در داخل واحدهای هجایی ویژگی‌هایی مثل واکدار بودن و سایشی بودن به آواهای مجاور گسترش می‌یابند و نیازی به مرزبندی‌های سخت[۹] نیست. ضمناً در این حالت مرزهایی مثل واکه-شبه واکه[۱۰] که به سختی قابل آشکارسازی هستند، در داخل ساختار هجا مدل می‌شوند.

بی‌شک هجا مفهومی مهم در توصیف نظری زبان است و از سوی دیگر از نظر شهودی برای زبان‌شناسان مفهومی کاملاً شناخته شده است. به علت ارتباط قوی هجا با ادراک گفتار در انسان، حتی شنوندگان عادی نیز درکی شهودی از هجا دارند و می‌توانند تعداد هجاهای کلمات را بدون دشواری تشخیص دهند[۸]. مرزهای هجا بسیار دقیق‌تر از مرزهای آوا قابل شناسایی هستند و افزودن اطلاعات این مرزها می‌تواند کارآیی سامانه‌های بازشناس گفتار را افزایش دهد [۹-۱۲]. در بسیاری از زبان‌ها ثابت شده است اگر واحد بازشناسی از واج به هجا تغییر یابد صحت بازشناسی افزایش می‌یابد [۱، ۱۳-۲۱].

از سوی دیگر پروزودی یا نوای گفتار که نقش مهمی در طبیعی‌سازی گفتار دارد و لحن[۱۱]، ریتم[۱۲]، آهنگ[۱۳] و لهجه[۱۴] را می‌سازد، در ارتباط مستقیم با هجا است. در ساختار سلسله مراتبیِ پروزودی، هجا پایین‌ترین سطح است. به عبارت دیگر کمیت‌های مرتبط با پروزودی مثل انرژی، مدت و فرکانس پایه در سطح هجا محاسبه و بررسی می‌شوند. بنابراین می‌توان گفت بخش عمده‌ای از تحقیقات در زمینه پروزودی از واحدهای هجایی استفاده می‌کنند [۲۲-۲۷]. استفاده از واحدهای هجایی حتی در بازشناسی زبان نیز نتایج موفقیت‌آمیزی در برداشته است [۷]. در [۲۸] نشان داده شده‌است که ویژگی‌های طیفی استخراج شده از واحدهای هجایی برای تشخیص زبان مناسب هستند.

این نتایج اهمیت واحدهای هجایی را در تحقیقات گفتار نشان می‌دهند. با این حال تمامی این حوزه‌ها به نحوی وابسته به تقطیع هجایی سیگنال گفتار هستند. با وجودی که ساختار هجاها و مرزهای آنها از نظر شهودی کاملاً مشخص هستند، استخراج این مرزها با دقت زیاد از روی سیگنال گفتار کار ساده‌ای نیست. کارهای انجام گرفته در این زمینه را می‌توان در دو دسته‌بندی عمده قرار داد. دسته اول تحقیقات مبتنی بر مدل هستند که از ساختار هجایی زبان موردنظر و گاه محتوای آوایی استفاده می‌کنند. یکی از اولین کارهای انجام شده در این زمینه استفاده از شبکه‌های عصبی مصنوعی است [۲۹]. در [۳۰] نیز از مدل‌های پنهان مارکوف برای این منظور استفاده شده است. دسته دوم روش‌هایی هستند که تنها از اطلاعات سیگنال گفتار و پردازش آن به این منظور استفاده می‌کنند. این روش‌ها به دلیل آنکه به اطلاعات زبانی، آوایی و مدل کردن نیاز ندارند کاربرد بیشتری یافته‌اند.

اگر از دید سیگنال گفتار به هجا نگاه کنیم می‌توان گفت هجا قله انرژی مربوط به یک واکه است که با تعدادی همخوان[۱۵] احاطه شده است. از نظر ساختاری هسته هجا یک واکه است و تعداد همخوان‌های

دوطرف آن بسته به ساختار زبان تغییر می‌کند. بنابراین تابع انرژی زمان کوتاه[۱۶] گفتار دارای اطلاعات زیادی در سطح هجا است. در [۳۱] تابع انرژی از روش‌های مختلف محاسبه شده و برای تقطیع هجایی زبان تایلندی به کار رفته است. در [۱۳] نیز به منظور تقطیع هجایی دادگان TIMIT و NTIMIT از تابع انرژی زمان کوتاه و تاخیرگروهیِ[۱۷] آن استفاده شده است. همچنین نشان داده شده است اگر تقطیع هجایی با استفاده از تاخیر گروهیِ استخراج شده از تابع انرژی زمان کوتاه انجام گیرد، نسبت به نویز مقاوم خواهد بود [۱۳، ۱۸، ۳۲].

تابع انرژی زمان کوتاه به راحتی قابل محاسبه است. هسته‌های هجا بر روی قله‌های این تابع و مرزهای هجا در دره‌های آن قرار دارند. با این حال آشکارسازی دقیق مرزها به دلیل نوسانات زیاد تابع انرژی کار ساده‌ای نیست و لازم است نوسانات اضافه تا حدامکان حذف شوند. برای انجام این کار با استفاده از فیلترها، انتخاب فرکانس قطع و یا طول قاب به صورت بهینه بسیار مشکل است. انتخاب نادرست منجر به حذف برخی مرزها، درج مرزهای اضافه و افزایش خطا می‌شود.

یک روش مؤثر برای حل این مشکل، استفاده از فیلتر فازی است [۳۳]. فیلتر فازی توانایی حذف نوسانات کوچک و ناخواسته را دارد و در عین حال شکل تابع انرژی را حتی در تیزترین نقاط حفظ می‌کند. با این حال خطای درج مرز اضافه در این روش زیاد است و به نظر می‌رسد این خطا بیشتر به دلیل وجود همخوان‌های پرانرژی باشد. در این مقاله روش جدیدی برای بهبود تقطیع هجایی سیگنال گفتار پیشنهاد شده است که علاوه بر کارآیی مناسب، خطای درج مرز اضافه را به نحو چشمگیری کاهش می‌دهد. در این روش از تبدیل موجک[۱۸] برای حذف قله‌های اضافه در تابع انرژی زمان کوتاه استفاده شده است.

تبدیل موجک در بهسازی سیگنال گفتار کاربردهای بسیاری دارد. در روش‌های مبتنی بر تبدیل موجک از پنجره‌های زمانی با طول متفاوت در باندهای فرکانسی مختلف استفاده می‌شود که برای سیگنال غیرایستای گفتار مناسب است. برای حذف نویز، استفاده از آستانه گذاری ضرایب موجک متداول‌ترین روش است [۳۴، ۳۵]. اما در نواحی بی‌واک کارآیی آن به شدت کاهش می‌یابد. روش‌های مختلفی از آستانه‌گذاری نیز برای جلوگیری از کاهش کیفیت سیگنال گفتار در نواحی بی‌واک، معرفی شده است [۳۶، ۳۷]. در این مقاله از روش آستانه‌گذاری، برای کاهش انرژی در نواحی بی‌واک استفاده شده است تا خطای درج مرز اضافه در هنگام تقطیع هجایی کاهش یابد.

در ادامه، در بخش دوم ابتدا روش تقطیع هجایی سیگنال گفتار با استفاده از تابع انرژی زمان کوتاه و فیلترسازی فازی توضیح داده شده است. در بخش سوم جزئیات روش پیشنهادی آمده است. بخش چهار به بررسی نتایج بر روی دادگان فارس‌دات می‌پردازد و در پایان بحث و نتیجه‌گیری آمده است.

۲- آشکارسازی مرزهای هجایی با استفاده از تابع انرژی زمان کوتاه

برای تقطیع هجایی سیگنال گفتار ابتدا لازم است تابع انرژی زمان کوتاه به روش مناسب محاسبه شود. پس از آن نوسانات ناخواسته به روش فازی حذف شده و در پایان با استفاده از معیارهای آستانه مرزهای هجایی آشکار می‌شوند.

۲-۱- محاسبه تابع انرژی زمان کوتاه

توابع مختلفی برای محاسبه تابع انرژی زمان کوتاه پیشنهاد شده است [۲۶]. نشان داده شده است که در تقطیع هجایی سیگنال گفتار پیوسته[۱۹] فارسی استفاده از فرمول انرژی تیگر[۲۰] بهترین نتایج را دارد [۳۸]. فرض کنید s[i] یک عبارت گفتاری پیوسته شامل تعدادی جمله باشد. ابتدا برای حذف DC یک فیلتر پیش‌تاکید[۲۱] با فاکتور ۰/۹۵ به s[i] اعمال می‌شود. عملکرد فیلتر به صورت زیر است.

$$S[i] = s[i] - 0.95s[i] \quad (۱)$$

سپس انرژی تیگر از خروجی فیلتر یعنی S[i] به روش زیر محاسبه می‌شود [۳۹]:

$$E[i] = s_n^2[i] - s_n[i+1]s_n[i-1] \quad (۲)$$
$$E_n = \sum_{i=1}^{W} E[i]$$

که در آن $S_n[i]$ نمونه‌های سیگنال گفتار در قاب nام هستند، W طول قاب موردنظر و E_n انرژی قاب nام است. انرژی در قاب‌هایی به طول ۲۵۶ نمونه (۱۱ میلی‌ثانیه) با همپوشانی ۰/۷۵ محاسبه می‌شود. به منظور حذف انفجارات کوتاه[۲۲] یک فیلتر میانه با طول ۱۱ قاب به تابع انرژی اعمال می‌شود. سپس تابع انرژی به مقدار بیشینه خود هنجارسازی[۲۳] می‌شود. نحوه عملکرد فیلتر میانه به صورت زیر است.

$$S[i] = \text{median}(\{s[i-5], s[i-4], \ldots, s[i+5]\}) \quad (۳)$$

قبل از پردازش لازم است سکوت‌های بین جملات و سکوت‌های طولانی بین کلمات در یک جمله حذف شوند. برای این کار با استفاده از آستانه‌گذاری[۲۴] شروع و پایان بخش‌های گفتاری در هر رکورد مشخص می‌شوند [۳۸]. این روش سکوت‌های مورد بحث را حذف می‌کند. بنابراین می‌توان مرزهای هجایی را در هر فایل گفتاری پیوسته آشکارسازی کرد و نیازی به تقطیع اولیه به جملات مجزا نیست.

۲-۲- فیلتر فازی

مشکل عمده در روش‌های مبتنی بر انرژی زمان کوتاه حذف نوسانات اضافی است. استفاده از میانگین‌گیر، به علت حذف برخی دره‌های اصلی و باقی گذاشتن برخی نوسانات اضافی با مشکل روبرو است. ضمن آنکه انتخاب مناسب‌ترین طول پنجره کار ساده‌ای نیست. روش آستانه متغیر[۲۵] [۴۰] نیز که در تحقیقات قبلی نویسنده این مقاله، پیشنهاد شده است، گرچه نسبت به روش آستانه ثابت دارای بهبود است، با این حال طول پنجره مناسب برای تعیین مداوم مقدار آستانه حائز اهمیت است.

در روشی که در [۳۳] توسط نویسنده همین مقاله ارائه شده است، برای غلبه بر این مشکل فیلترسازی فازی به کار رفته است. با توجه به آنکه روند تغییرات تابع انرژی زمان کوتاه از ابتدا تا انتهای جمله به صورت موضعی با آهنگ جمله تغییر می‌کند، روش مبتنی بر قواعد فازی برای نرم‌سازی[۲۶] تابع انرژی، این امکان را ایجاد می کند که انرژی هر قاب با توجه به تغییرات موضعی انرژی در همان نقطه فیلتر شود. جزئیات کامل این روش در [۳۳] آمده است. در این روش انرژی هر قاب با توجه به مقادیر انرژی در ۷ قابِ قبلی نرم‌سازی می‌شود. عدد ۷ طبق تحقیقات قبلی با روش سعی و خطا به دست آمده‌است. خلاصه قواعد فازی به صورت زیر است:

۱) اگر بیشتر ورودی‌ها بزرگ باشند آنگاه خروجی بزرگ است.
۲) اگر بیشتر ورودی‌ها متوسط باشند آنگاه خروجی متوسط است.
۳) اگر بیشتر ورودی‌ها کوچک باشند آنگاه خروجی کوچک است.
۴) در غیر این صورت خروجی صفر است.
۵) علامت خروجی مشابه با بیشترین تعداد ورودی‌ها است.

ورودی این قواعد فازی $x_i=E_i-\hat{E}_i$ برای $i=1,2,\ldots,7$ است که در آن E_i و $\hat{E}_i$ به ترتیب انرژی و انرژی فیلترشده قاب iام به روش فازی هستند. برای اولین نمونه انرژی فیلتر شده صفر در نظر گرفته شده است. تابع عضویت برای هریک از قواعد فازی به صورت زیر است:

$$\mu_A=\begin{cases}+\dfrac{2(x_i-c_A)}{\omega}+1 & c_A-\dfrac{\omega}{2}<x_i<c_A\\ -\dfrac{2(x_i-c_A)}{\omega}+1 & c_A\le x_i<c_A+\dfrac{\omega}{2}\\ 0 & \text{Otherwise}\end{cases} \qquad (۴)$$

که در آن A هر یک از قانون‌های فازی، c_A و ω به ترتیب مرکز و عرض تابع عضویتِ قانون A هستند. عرض همه توابع عضویت یکسان است. مرکز قانون چهارم برابر صفر است و این توابع به اندازه $\omega/2$ همپوشانی دارند. عرض توابع عضویت با توجه به هنجارسازی تابع انرژی به صورت تجربی برابر با ۰/۲ انتخاب شده است. همچنین عبارت «بیشترین» در قانون‌های فازی، تابع S شکلِ معمول است که با عبارت زیر توصیف می‌شود:

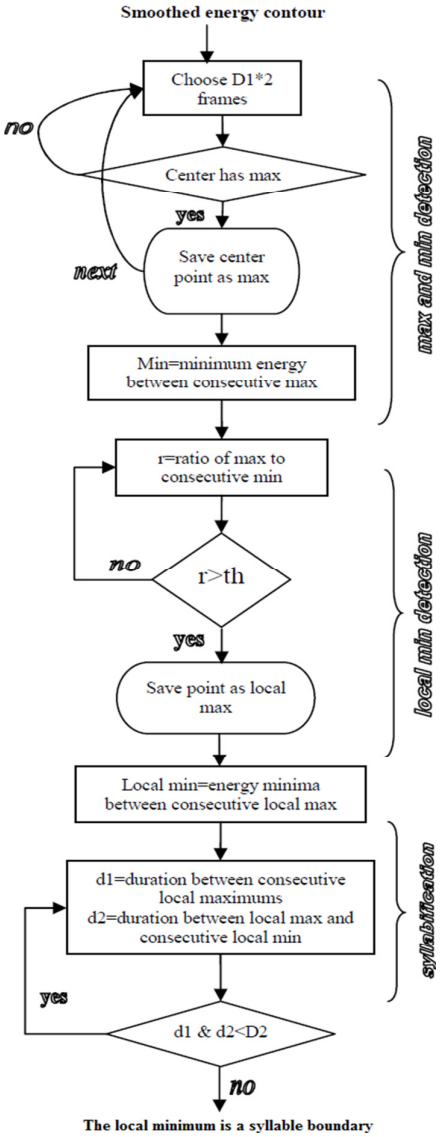
$$\mu_A=\begin{cases}0 & z<0.1\\ 0.5(1-\cos[\dfrac{\pi(z-1)}{0.8}]) & 0.1\le z<0.9\\ 1 & z\ge 0.9\end{cases} \qquad (۵)$$

در نهایت درجه فعالیت یا وزن هریک از گروه‌های فازی به صورت زیر محاسبه می‌شود:

$$\lambda_A=\text{median}[\,\mu_A(x_i):x_i\in A]\times\mu_{most}\left[\frac{(\text{number of } x_i\in A)}{(\text{total number of } x_i)}\right] \qquad (۶)$$

خروجی فیلتر فازی به صورت ضریب همبستگی هر چهار قانون به شکل زیر تعریف می‌شود:

$$\Delta E=\sum_{A=1}^{4}c_A\lambda_A \qquad (۷)$$



شکل (۱): الگوریتم استخراج مرزهای هجایی [۳۳]
Fig. (1): Syllable boundary detection algorithm [33]

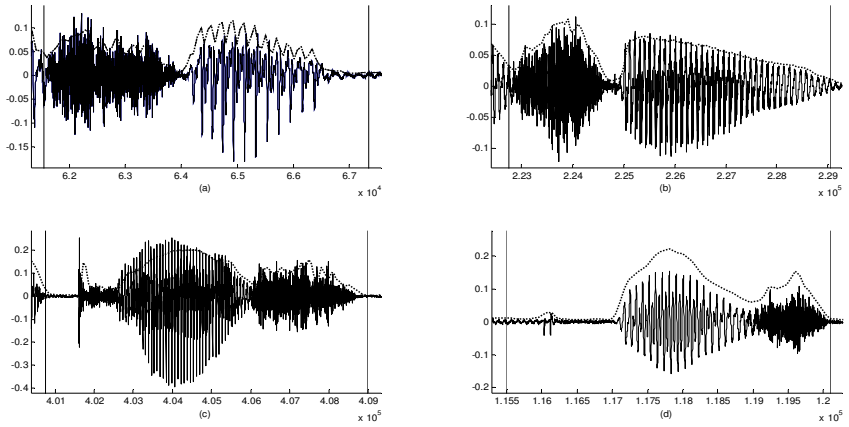
و انرژی فیلترشده عبارت است از:

$$E_{i+1}=\hat{E}_i+\Delta E \qquad (۸)$$

۲-۳- آشکارسازی مرزهای هجایی

پس از فیلتر کردن تابع انرژی باید مرزهای هجایی آشکار شوند. متداول‌ترین روش در شناسایی مرز هجاها، تعیین کمینه‌های تابع انرژی زمان کوتاه با استفاده از مقادیر آستانه است [۳۱]. برای این کار می‌توان از مقادیر آستانه ثابت و یا متغیر استفاده کرد. در روش‌های مبتنی بر آستانه ثابت، تعیین مقادیر آستانه بسیار مهم و وقت‌گیر است. با این حال آزمایشات قبلی نشان داده‌اند که استفاده از آستانه متغیر در تابع فیلترشده به روش فازی نتایج مناسبی در برندارد [۳۸].

در اینجا برای تقطیع هجایی از روش آستانه ثابت پیشنهاد شده در [۳۳] استفاده شده است. برای این کار لازم است کمینه‌های محلی[۲۷] تابع انرژی شناسایی شوند. این روش در شکل (۱) نشان داده شده است. در این روند ابتدا تعداد D1*2 نمونه از انرژی زمان کوتاه، در نظر گرفته می‌شود. هر نمونه مربوط به یک فریم از سیگنال گفتار است. سپس بیشینه‌های[۲۸] تابع انرژی با جستجو در این نمونه‌ها آشکار می‌شوند. مابین بیشینه‌های متوالی نقاط کمینه نیز شناسایی می‌شوند. با استفاده از مقدار آستانه th بیشینه‌های محلی از بین نقاط آشکار شده، انتخاب می‌شوند. به این صورت که اگر نسبت یک بیشینه به کمینه پس از آن از این مقدار آستانه بیشتر باشد، این نقطه به عنوان یک پیک محلی در نظر گرفته می‌شود.



شکل (۲): تعدادی از هجاهای دارای دو قله انرژی. (a): /kö/، (b): /ʃæ/، (c): /öz/، (d): /sæ/

Fig. (2): Some of syllables with two energy peaks. (a)/kö/, (b)/ʃæ/, (c)/öz/, (d)/sæ/

کمینه‌های محلی یا همان مرزهای احتمالی هجا مابین بیشینه‌های محلی قرار می‌گیرند. با شناسایی دره‌های تابع انرژی این نقاط نیز ذخیره می‌شوند. سپس با استفاده از معیار آستانه D2، محدودیت‌های طول زمانی اعمال شده و تعدادی از کمینه‌های محلی حذف می‌شوند.

روند کار به این صورت است که طول زمانی بین هر دو پیک متوالی با معیار آستانه مقایسه می‌شود. پیک‌هایی که فاصله آنها با پیک متوالی‌شان از D2 کمتر باشد، از لیست ذخیره حذف می‌شوند. سپس همین روند برای باقیمانده کمینه‌های ذخیره شده انجام می‌شود. در نهایت کمینه‌های باقیمانده به عنوان مرزهای هجایی درنظرگرفته می‌شوند.

مقدار این سه کمیت برای تابع انرژی تیگر عبارتست از D2=12, th=0.015, D1=20. همان طور که پیشتر اشاره شد، تعیین مقادیر آستانه D1، th و D2 روندی زمان‌بر است و با روش سعی و خطا انجام می‌گیرد. جزئیات کامل مربوط به این روند در [۳۳] آمده است.

۳- روش پیشنهادی

همانطور که پیشتر اشاره شد مشکل عمده در روش مبتنی بر فیلتر فازی، وجود مقدار زیادی خطای درج مرز اضافه است. به منظور یافتن راه حلی برای این مشکل، ویژگی‌های صوتی و محتوای واجی سیگنال گفتار در محل درج این مرزهای اضافه مورد بررسی قرار گرفته است.

مشاهده می‌شود این خطا عمدتاً به دلیل همخوان‌های پرانرژی با طول نسبتاً زیاد رخ می‌دهد مثلا «ش» یا «س». همچنین در گفتار برخی از گویندگان همخوان «ز» در بعضی جملات شبیه «س» تلفظ شده است و باعث ایجاد مرز اضافه می‌شود. این مسأله در مواردی که همخوان «چ» شبیه «ش» تلفظ شده‌است نیز رخ می‌دهد.

در اینجا هدف بررسی دقیق محتوای واجی نیست. با این حال این همخوان‌ها علیرغم تفاوت‌هایشان، دارای ویژگی‌های مشترکی نیز هستند؛ اغلب این آواها نسبتا طولانی بوده و سیگنال گفتار تقریبا در تمامی این قسمت‌ها ساختاری نویزگونه دارد. شکل (۲) سیگنال گفتار (خط پر) و تابع انرژی زمان کوتاه آن (خط‌چین) را در تعدادی از هجاهای حاوی این آواها نشان می‌دهد. هجای مورد نظر با خطوط عمودی مشخص شده است. به ساختار نویزگونه سیگنال گفتار در هجای مشخص شده و قله‌های انرژی آن توجه شود.

در هنگام قرارگرفتنِ این همخوان‌ها در مجاورت واکه‌ها یا همخوان‌های واکدار، در دامنه سیگنال گفتار به طور طبیعی افت و خیز ایجاد می‌شود. از نظر ویژگی‌های صوتی، تابع انرژی زمان کوتاه در محل این همخوان‌ها دارای قله و در مرزهای واکداری/بی‌واکی یا بی‌واکی/ واکداری دارای دره‌های اضافه است. بنابراین هجای حاوی این واج‌ها، دارای دو قله است و در روند تعیین مرزها، یک مرز اضافه آشکارسازی می‌شود. هرچند قله انرژی مربوط به این همخوان‌ها اغلب دارای دامنه کمتری نسبت به واکه‌های مجاور است، اما با روش‌های مبتنی بر آستانه نمی‌توان این قله‌ها را حذف کرد. به دلیل تغییرات انرژی از ابتدا تا انتهای جمله این کار به حذف برخی واکه‌های کم انرژی منجر می‌شود و خطای حذف مرزهای هجایی را افزایش می‌دهد.

بنابراین مشخص است که با پردازش تابع انرژی زمان کوتاه، این همخوان‌ها از واکه‌ها یا هسته‌های هجا قابل تشخیص نیستند و لازم است از اطلاعات دیگری برای حذف این مرزهای اضافه بهره برد. راه‌حل پیشنهادی در این مقاله استفاده از تبدیل موجک است. این روش بر اساس تفاوت مهم این همخوان‌ها با نقاط واک‌دار عمل می‌کند. واکه‌ها حاوی فرکانس پایه[۲۹] هستند و ساختاری شبه‌متناوب[۳۰] دارند، در حالی که همخوان‌های بی‌واک، علیرغم ایجاد قله انرژی، ساختاری نویزگونه دارند. بنابراین می‌توان با استفاده از تبدیل موجک و با روشی مشابه روش‌های متداولِ حذف نویز از گفتار، این همخوان‌ها را حذف کرد.

روند کار به این صورت است که ابتدا تبدیل موجک سیگنال گفتار محاسبه می‌شود. سپس با استفاده از آستانه‌گذاری، سیگنال جدیدی به دست می‌آید که انرژی همخوان‌های بی‌واک در آن به طرز محسوسی کاهش یافته است. در نهایت روند محاسبه انرژی زمان‌کوتاه بر روی این سیگنال جدید اعمال می‌شود. در تابع انرژی نهایی، قله‌های اضافه مربوط به همخوان‌های مشکل ساز وجود ندارند.

۳-۱- تبدیل موجک

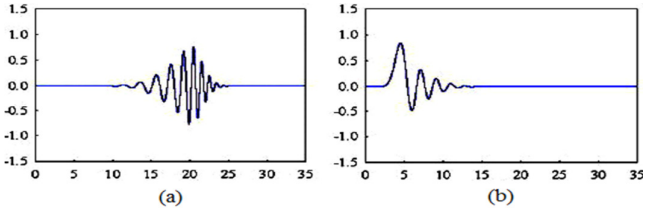
تبدیل موجک در چند سال گذشته به عنوان ابزاری قدرتمند در پردازش سیگنال گفتار شناخته شده‌است. از جمله موارد استفاده این تبدیل، حذف نویز به روش آستانه‌گذاری ضرایب موجک[۳۱] می‌باشد. روش‌های مختلفی از آستانه‌گذاری ضرایب موجک برای بهسازی سیگنال آلوده به نویز پیشنهاد شده است [۳۶]، [۳۷]، [۴۱]. مشکل عمده در آستانه‌گذاری‌های متداول، افت کارآیی و تغییر شکل سیگنال در نواحیِ بی‌واک سیگنال گفتار است. با این حال وقتی این ابزار را برای حذف همخوان‌های پرانرژی به کار می‌بریم، این عیب به یک مزیت تبدیل می‌شود.

بخش عمده‌ای از انرژی سیگنال گفتار در تعداد کمی از ضرایب موجک متمرکز است. قسمت اصلی این انرژی مربوط به نواحی واکدار سیگنال گفتار است. بنابراین از طریق مقایسه ضرایب کوچکتر با یک حد آستانه، می‌توان انرژی بخش‌های بی‌واک را به طرز محسوسی کاهش داد بدون آنکه در شکل کلی تابع انرژی تغییر عمده‌ای ایجاد شود.

۳-۲- آستانه‌گذاری ضرایب موجک

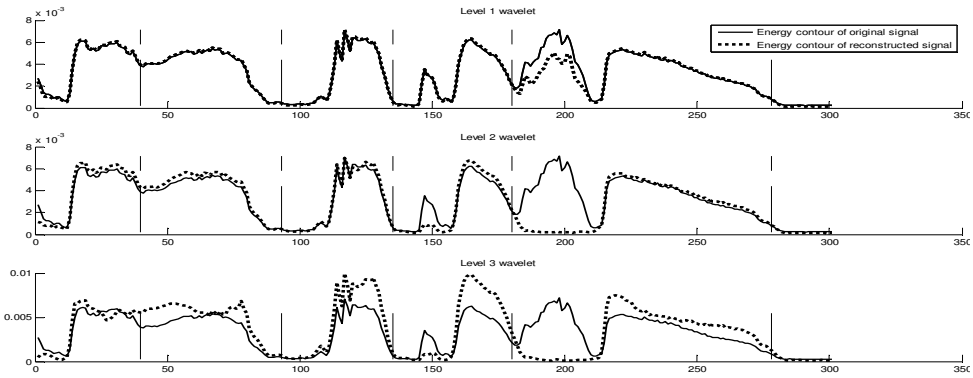
برای تجزیه[۳۲] سیگنال گفتار، از پایه موجک دابوچی[۳۳] درجه ۲۰ (db20) استفاده شده است که در زمینه بهسازی سیگنال گفتار بسیار پرکاربرد است. شکل (۳) تابع موجک مربوطه و تابع مقیاس[۳۴] آن را نشان می‌دهد. عمق پیشروی در درخت موجک[۳۵] برابر ۲ در نظر گرفته شده است. نتایج تجربی نشان می‌دهند استفاده از عمق پیشروی برابر۱ تأثیر چندانی در حذف انرژی همخوان‌های مورد بحث ندارد. افزایش عمق پیشروی به میزان بیشتر از ۲ نیز به نحو نامطلوبی، منجر به تغییر شکل انرژی زمان کوتاه می‌شود. این مساله در شکل (۴) نشان داده شده است. مراحل مربوط به استخراج این نمودارها در ادامه به طور کامل توضیح داده شده است.

این شکل تابع انرژی زمان کوتاه مربوط به نیمه انتهایی جمله «پیپ دلم می‌خواهد بکشم» را نشان می‌دهد. کلیه توابع انرژی در شکل (۴) به روش فازی نرم‌سازی شده‌اند. نمودارها به ترتیب از بالا به پایین مربوط به عمق پیشروی ۱، ۲ و ۳ هستند. انرژی سیگنال اصلی در هر نمودار با خط پر و انرژی سیگنال آستانه‌گذاری شده با نقطه‌چین نشان داده شده است. خطوط عمودی مرزهای هجا را نشان می‌دهند.



شکل (۳): موجک db20: a تابع موجک و b تابع مقیاس

Fig. (3): Db20 wavelet: (a) Daubechies wavelet and (b) scale function



شکل (۴): اثر مقادیر مختلف عمق پیشروی درخت موجک در تابع انرژی

Fig. (4): The effect of different values for wavelet tree decomposition level

به قله اضافه مربوط به همخوان «ش» در هجای آخر و قله مربوط به «ک» در هجای قبلی توجه کنید. در اینجا قله مربوط به «ش» هم از نظر دامنه و هم از نظر طول زمانی کاملاً همرده با واکه‌های موجود در هسته‌های هجاست و مسلماً به عنوان یک هجا آشکارسازی می‌شود. به ازای عمق پیشروی برابر ۱ تنها مقدار کمی از انرژی این قله کم می‌شود. افزایش عمق پیشروی به حذف کامل این قله منجر شده است و نهایتا با افزایش بیش از حد، شکل تابع انرژی تغییر می‌کند. همان طور که از نمودارها مشخص است، بهترین نتایج مربوط به عمق پیشروی ۲ است.

پس از تجزیه سیگنال گفتار، ضرایب موجک آستانه‌گذاری می‌شوند. برای انتخاب مقادیر آستانه از روش اکتشافی ریسک بدون پیش فرض استِین[۳۶] [۴۲] استفاده شده است. فرمول (۹) نحوه محاسبه آستانه را در این روش نشان می‌دهد.

$$thr = \sqrt{(2\ln(n\log_2 n))} \qquad (۹)$$

n تعداد نمونه‌های سیگنال است. بعد از تعیین مقدار آستانه، باید تابع آستانه‌گذاری برای پردازش ضرایب انتخاب شود. پردازش ضرایب با روش آستانه‌گذاری نرم[۳۷] انجام شده است.

طبق [۴۳] در حذف نویز سفید از گفتار بهترین روش، آستانه‌گذاری سخت[۳۸] است. آستانه‌گذاری سخت به این صورت عمل می‌کند که ضرایب موجکی که اندازه آنها بزرگتر از آستانه است نگه داشته می‌شوند و مابقی صفر می‌شوند. آستانه‌گذاری نرم ضرایب موجک زیر آستانه را صفر می‌کند، ضرایب بزرگتر را به اندازه مقدار آستانه به سمت صفر شیفت می‌دهد.

از نظر عملکرد بر روی سیگنال، آستانه‌گذاری سخت انرژی سیگنال و جزئیات آن را حفظ می‌کند. در عوض آستانه‌گذاری نرم باعث از دست رفتن انرژی و ایجاد سیگنال نرم می‌شود. در اینجا هدف حذف نویز نیست بلکه ایجاد تخریب در نواحی بی‌واک گفتار است. بنابراین باید از روش آستانه‌گذاری نرم استفاده شود. مرحله بعدی ساخت مجدد سیگنال گفتار یا همان ترکیب[۳۹] از روی ضرایب آستانه‌گذاری‌شده‌است. ترکیب بر اساس روش متداول مورد استفاده در حذف نویز انجام می‌گیرد.

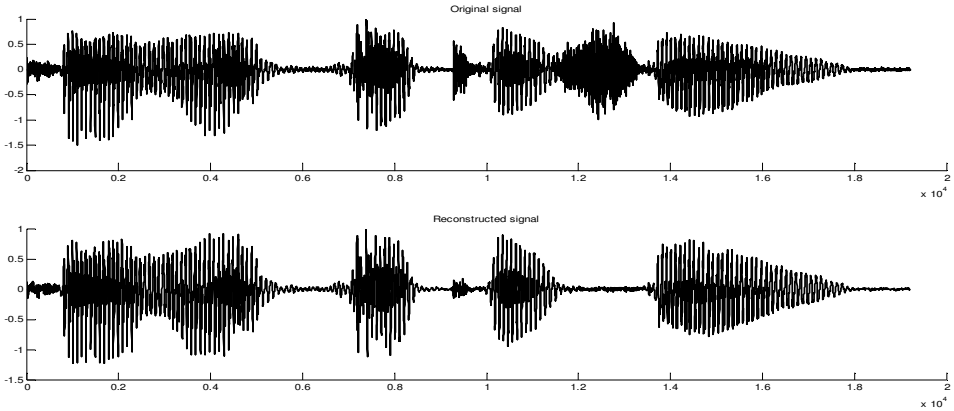
با وجودی که کیفیت سیگنال بازسازی شده به طرز محسوسی کاهش می‌یابد، سیگنالی که در این مرحله به دست می‌آید از نظر شنیداری قابل فهم است. سیگنال به دست آمده در این مرحله به جای سیگنال اصلی برای استخراج تابع انرژی و تعیین مرزهای هجایی مورد استفاده قرار می‌گیرد. شکل (۵) سیگنال گفتار مربوط به نیمه انتهایی جمله «پیپ دلم می‌خواهد بکشم» را به همراه سیگنال حاصل از آستانه گذاری نشان می‌دهد. کاهش شدید دامنه در قسمت انتهای سیگنال مربوط به واج «ش» است. برای مقایسه بهتر، اسپکتروگرام دو سیگنال نیز در شکل (۶) نشان داده شده‌است.

۳-۳- آشکارسازی مرزهای هجایی

در این مرحله انرژی زمان کوتاهِ سیگنال بازسازی شده، با روشی مطابق بخش (۱-۲) محاسبه می‌شود. به دلیل استفاده از آستانه‌گذاری نرم، تابع انرژی به دست آمده در این مرحله حالت نرم‌تری[۴۰] دارد و نوسانات ناخواسته موجود در آن نسبت به انرژی زمان کوتاه سیگنال اصلی بسیار کمتر است. با این حال قبل از آشکارسازی مرزهای هجایی باید نوسانات محلی به روش مناسبی حذف شوند. بنابراین تابع انرژی زمان کوتاه با استفاده از فیلتر فازی توصیف شده در بخش (۲-۲) نرم‌سازی می‌شود.

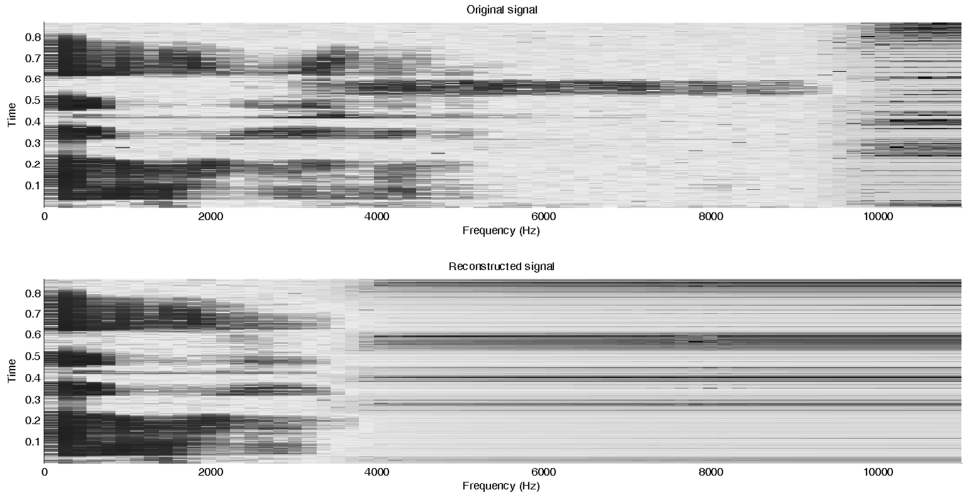
شکل (۷) تابع انرژی زمان کوتاه مربوط به سیگنال گفتار بازسازی شده را در جمله «پیپ دلم می‌خواهد بکشم» قبل و بعد از نرم‌سازی به روش فازی نشان می‌دهد. نمودار نقطه‌چین و نمودار توپر به ترتیب انرژی را قبل و بعد از نرم‌سازی فازی نشان می‌دهند. مرزهای هجا در این شکل با خط‌چین‌های عمودی نشان داده شده‌اند.

از روی شکل مشخص است که استفاده از آستانه‌گذاری ضرایب موجک و فیلتر فازی به طور همزمان باعث شده است که در هر واحد هجایی تنها قله انرژی مربوط به واکه یا هسته هجا حفظ شود و سایر نوسانات و قله‌های ناخواسته حذف شده‌اند. در این مرحله کافی است با استفاده از روند توصیف شده در فلوچارت شکل (۱) مرزهای هجایی آشکار شوند. مراحل روش پیشنهادی در بلوک دیاگرام شکل (۸) خلاصه شده‌اند.



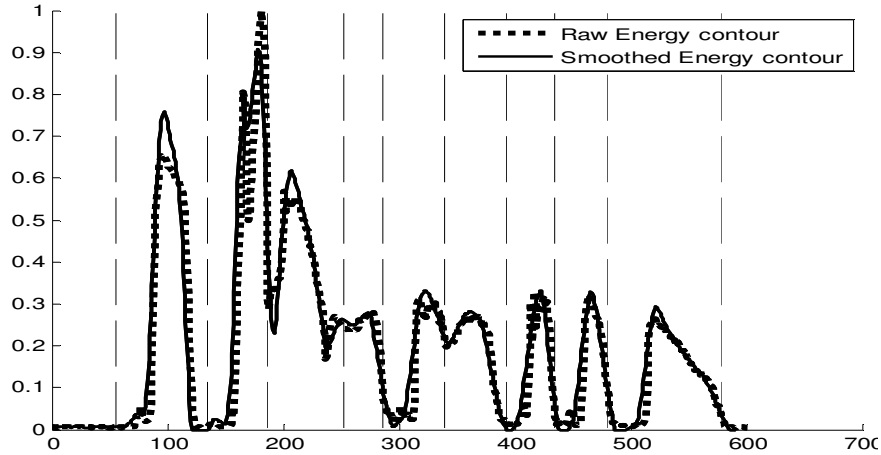
شکل (۵): سیگنال گفتار اصلی و سیگنال بازسازی شده پس از آستانه‌گذاری

Fig. (5): The final half of sentence /pīp delæm mikhähæd bekeʃæm/: original speech signal and reconstructed signal after thresholding



شکل (۶): اسپکتروگرام سیگنال اصلی و بازسازی شده مربوط به شکل (۵)

Fig. (6): Spectrogram for original and reconstructed signal in figure 5



شکل (۷): تابع انرژی زمان کوتاه قبل و بعد از نرم‌سازی فازی

Fig. (7): Short term energy contour before and after fuzzy smoothing

۴- نتایج

کارآیی روش پیشنهادی با انجام آزمایشات بر روی بخشی از دادگان فارس‌دات[۴۱] [۴۴] مورد ارزیابی قرار گرفته است. مجموعه مورد استفاده در این تحقیق شامل ۲۰ فایل صوتی است که هر یک بیان ۱۰ جمله توسط یک گوینده هستند. برای جلوگیری از اثر تغییرات لهجه تمامی فایل‌ها از میان گویندگان با لهجه تهرانی انتخاب شده‌اند. ضمناً سعی شده است در انتخاب فایل‌ها تنوع جملات و کلمات به میزان زیادی حفظ گردد و تا حد امکان از جملات تکراری صرفنظر شده‌است. در کل ۲۰۰ جمله بیان شده توسط ۲۰گوینده مختلف (۱۲ مرد و ۸ زن) دادگان موردنظر را تشکیل می‌دهند. برای ایجاد برچسب هجایی از برچسب‌های آوایی موجود در دادگان فارس‌دات و ساختار هجاهای فارسی استفاده شده است. ساختار اصلی هجا در زبان فارسی به یکی از صورت‌های CV، CVC و CVCC است،که C نشان دهنده همخوان و V نشان دهنده واکه است. با این حال به دلیل حذف برخی واج‌ها در گفتار پیوسته، هجاهایی به صورت Vیا VC نیز پدید می‌آیند. با تکیه بر این اطلاعات به عنوان قانون ساختاری هجا و با استفاده از برچسب واکه/همخوان مرزهای هجاییِ دادگان استخراج شده است.

در مجموع ۱۹۳۱ هجا در کل دادگان وجود دارد که ۱۱۶۵ تای آنها به وسیله گویندگان مرد و بقیه توسط گویندگان زن بیان شده‌اند. میزان فراوانی ساختارهای مختلف هجایی نیز به این صورت است: در حدود ٪۵۰ از هجاها دارای ساختار CV هستند. پس از آن بیشترین فراوانی مربوط به هجاهای CVC است که کمتر از ٪۴۰ هجاهای موجود در دادگان را تشکیل می‌دهند و در نهایت حدود ٪۱۰ هجاهایی با ساختار CVCC هستند. دو ساختار دیگر در کمتر از ٪۰/۵ موارد رخ داده‌اند.در این تحقیق، مرزهای صحیح هجایی با استفاده از برچسب‌های واجی دادگان فارس‌دات و ساختار هجاهای زبان فارسی به دست آمده‌اند. به عبارت دیگر مرزهای واقعی هجا بر اساس معیار انسانی محاسبه شده‌اند.

به منظور مقایسه، تقطیع هجایی سیگنال گفتار با استفاده از چند روش مختلف انجام گرفته است. این روش‌ها عبارتند از آستانه ثابت [۳۱] یا میانگین‌گیر(MA[42])، آستانه متغیر (AT[43]) [۴۰]، تاخیر گروهی (GD[44]) [۱۸]، [۴۳]، نرم‌سازی فازی (FS[45]) [۳۳] و روش پیشنهادی یعنی آستانه‌گذاری ضرایب موجک به همراه نرم‌سازی فازی (WF[46]). جدول (۱) نتایج هر یک از این روش‌ها را نشان می‌دهد. معیارهای خطای درج شده در این جدول با توجه به کارهای قبلی دراین زمینه انتخاب شده‌اند.

ستون اول درصد هجاهایی را نشان می‌دهد که مرزهای آنها با خطایی کمتر از ٪۲۰ طول متوسط هجا آشکارسازی شده‌اند. طول متوسط هجاهای فارسی ۲۵۰ میلی‌ثانیه است. در ستون دوم حداکثر مقدار خطا در تشخیص محل مرز نسبت به طول هجا نشان داده شده است. ستون سوم درصد خطای طول هجا را نشان می‌دهد.

برای محاسبه این معیار ابتدا اختلاف طول هجای واقعی و هجای آشکار شده محاسبه می‌شود. سپس مجموع این‌مقادیر برای همه هجاها به تعداد کل هجاهای آشکار شده تقسیم شده و به صورت درصد بیان می‌شود. این معیار تا حد زیادی نشان دهنده خطای متوسط در آشکارسازی مرزهای هجایی است.

ستون بعدی درصد حذف هجاها را نشان می‌دهد که نسبت تعداد هجاهای حذف شده به کل هجاهاست. در نهایت ستون آخر درصد درج مرزهای اضافه را نشان می‌دهد که با روشی مشابه به دست می‌آید. کلیه مقادیر تا یک رقم اعشار گرد شده‌اند.

همانطور که قبلاً اشاره شد، جدول (۱) برتری نرم‌سازی فازی نسبت به سایر روش‌های متداول را نشان می‌دهد. از مقایسه نتایج جدول مشخص می‌شود که روش پیشنهادی به دلیلی حذف قله‌های اضافه خطای درج مرز اضافه را به نحو چشمگیری کاهش داده است. علاوه بر آن حذف مرزهای اضافه باعث تشخیص دقیق‌تری از طول دو هجای مجاور شده و در نتیجه درصد هجاهای آشکارشده با خطای کمتر از ۵۰ میلی‌ثانیه اندکی بهبود یافته است.

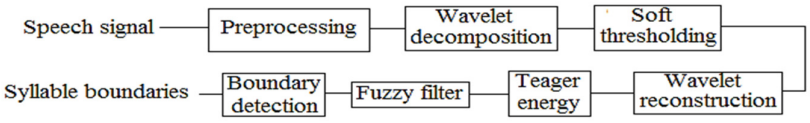
هرچند انتظار می‌رود کاهش خطای درج به بهبود متوسط خطای طول هجا منجر شود، اما از ستون سوم جدول مشخص است که این خطا به میزان بسیار اندکی افزایش یافته و از ۸/۶ به ۹/۰ رسیده است. در توضیح این مسئله می‌توان گفت که حذف قله‌های انرژی اضافه باعث ایجاد محدوده‌ای نسبتا مسطح در انرژی زمان کوتاه می‌شود و درج مرز بین دو هجا در این محدوده دارای خطا خواهد بود.

به عبارت دیگر در محدوده‌ای که قبلاً دو دره و یک قله وجود داشته تنها مرز بین دو هجا باید آشکار شود. این مسئله از طرفی حداکثر خطای طول هجا را کاهش داده و سوی دیگر منجر به قرارگیری غیردقیق مرز می‌شود. هرچند این کاهش ۰/۴ درصدی در مقایسه با بهبودی که در خطای درج رخ داده است قابل صرفنظر است.

به منظور درک بهتر مسئله، مرزهای آشکارشده به روش نرم‌سازی فازی و روش پیشنهادی در شکل (۹) نشان داده شده است. نمودارهای قسمت بالا و پایین شکل به ترتیب انرژی زمان کوتاه جمله «پیپ دلم می‌خواهد بکشم» و جمله «بالاخره رمز قفل را کشف کردم» را نشان می‌دهند. برای نمایش انرژی زمان کوتاه سیگنال اصلی، نقطه‌چین و برای نمایش انرژی به دست آمده از روش پیشنهادی، خطوط پر به کار رفته است.

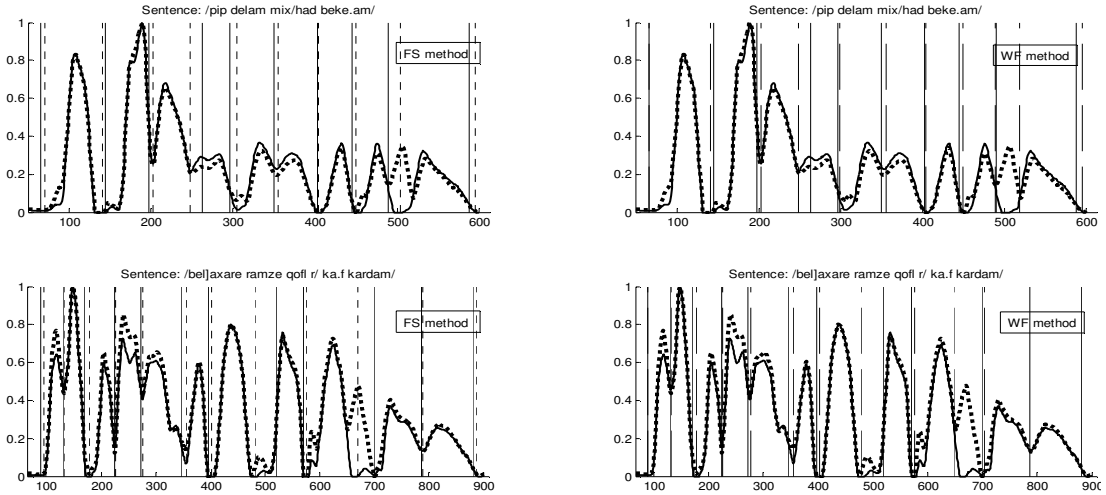
خطوط عمودی پر نشان دهنده مرزهای واقعی هجا هستند. خط‌چین‌ها مرزهای آشکارسازی شده به روش FS و نقطه‌چین‌ها مرزهای آشکارشده با روش WF را نشان می‌دهند. همان طور که در شکل مشخص است، در بعضی از قسمت‌ها مرزهای آشکارشده دقیقاً یکسان هستند. در این دو نمونه حذف مرز اتفاق نمی‌افتد.

شکل نشان می‌دهد آستانه‌گذاری ضرایب موجک نه تنها باعث کاهش درج مرزهای اضافه می‌شود بلکه شکل کلی تابع انرژی را نیز تغییر نمی‌دهد. هجای آخر در جمله بالایی و هجای نهم در جمله پایینی دارای دوقله انرژی هستند که با روش FS این هجاها به عنوان دو هجا تقطیع می‌شوند.



شکل (۸): بلوک دیاگرام روش پیشنهادی

Fig. (8): Block diagram for proposed method



شکل (۹): مرزهای آشکارسازی شده به روش FS و WF

Fig. (9): Detected boundaries using FS and WF method

Table (1): A comparison of different techniques for syllable segmentation

جدول (۱): مقایسه کارآیی روش‌های مختلف در تقطیع هجایی گفتار

Error / method	syllables with less than 50ms error %	Max error relative to syllable duration%	Duration error%	Insertion%	Deletion%
MA method	**86.3**	**33**	**10.2**	**9.3**	**7.5**
AT method	**88.7**	**33**	**9.8**	**7.6**	**6.4**
GD method	**74.5**	**84**	**16.6**	**9.8**	**6.5**
FS method	**93.8**	**27**	**8.6**	**14.2**	**3.4**
WF method	**94.7**	**24**	**9.0**	**6.1**	**3.4**

۵- نتیجه‌گیری

در این مقاله روش جدیدی برای بهبود کارآیی سامانه تقطیع هجایی سیگنال گفتار پیوسته ارائه شده‌است. این روش مبتنی بر آستانه‌گذاری ضرایب موجک در سیگنال گفتار و سپس نرم‌سازی فازی تابع انرژی است. روش نرم‌سازی فازی قبلاً توسط نویسنده مقاله معرفی شده است. با این حال خطای درج مرز اضافه در این روش غیرقابل قبول است. وجود همخوان‌هایی با طول و انرژی زیاد باعث ایجاد قله‌های اضافه در تابع انرژی زمان کوتاه سیگنال گفتار می‌شود. هجاهای حاوی این همخوان‌ها دارای دو قله انرژی هستند که باعث تشخیص یک مرز اضافه در فرآیند تقطیع می‌شود.

برای کاهش خطای درج مرز اضافه، استفاده از آستانه‌گذاری ضرایب موجک پیشنهاد شده است. این روش در تحقیقات حوزه گفتار به منظور بهسازی و حذف نویز از گفتار بسیار متداول است. در اینجا این روند نه برای کاهش نویز بلکه برای کاهش انرژی همخوان‌ها استفاده شده است. آزمایشات انجام شده بر روی بخشی از دادگان فارس‌دات کارآیی روش مذکور را نشان می‌دهد. نتایج نشان می‌دهند با استفاده از این روش خطای درج مرز اضافه به طرز محسوسی کاهش می‌یابد. علاوه بر آن از بین رفتن مرزهای اضافه دقت تشخیص طول هجاهای مجاور را افزایش می‌دهد و باعث بهبود سایر معیارهای خطا می‌شود.

با توجه به اهمیت تقطیع خودکار گفتار پیوسته در تحقیقات گفتاری به خصوص در زمینه نوای گفتار، تعیین مرزهای هجایی با دقت قابل قبول می‌تواند راه را برای انجام پژوهش در حوزه آهنگ، لحن و ریتم گفتار هموارتر سازد. زمانی که این کار در زبان‌های هجامحور مثل فارسی انجام می‌شود، اهمیت موضوع دوچندان می‌شود. هدف آن است که در مراحل بعدی ویژگی‌های نوایی سیگنال گفتار از قطعات هجایی استخراج شود و سپس این ویژگی‌ها در تشخیص آهنگ جمله به کار رود.

پی‌نوشت:

1. Voice command
2. Automatic Speech Recognition System
3. Segmented
4. Acoustic level
5. Vocabulary level
6. Language level
7. Co-articulation
8. Frame based
9. Strict
10. Semi vowel
11. Accent
12. Rhythm
13. Tone
14. Dialect
15. Consonant
16. Short Term Energy Function

17. Group delay
18. Wavelet transform
19. Continuous
20. Teager
21. Pre-emphasize
22. Short bursts
23. Normalize
24. Thresholding
25. Adaptive threshold
26. Smoothing
27. Local minima
28. Maxima
29. Pitch frequency
30. Quasi periodic
31. Wavelet coefficient thresholding
32. Decomposition
33. Daubechies
34. Scaling function
35. Wavelet tree level
36. Heuristic Stein unbiased risk
37. Soft thresholding
38. Hard thresholding
39. Reconstruction
40. Smoother
41. FarsDat
42. Moving Average
43. Adaptive Threshold
44. Group Delay
45. Fuzzy Smoothing
46. Wavelet based Fuzzy smoothing

References

[1] Z. Hu, J. Sehalkwyk, E. Barnard, R. Cole, "Speech recognition using syllable-like units", ICSLP, Vol. 2, pp.1117-1120, 1996.

[2] R. Cole, B. Oshika, M. Noel et al, "Labeler agreement in phonetic labeling of continuous speech", ICSLP, pp.2131-2134, 1994.

[3] O. Ghitza, M. Sondhi, "Hidden Markove models with templates as non-stationary states: an application to speech recognition", Jou. of Com. Speech and Language, Vol.2, pp.101-119, 1993.

[4] Z. Hu, E. Bernard, R. Cole, "Transition-based feature extraction within frame-based recognition", Eurospeech conf., pp.1555-1558, 1995.

[5] M. Ostendorf, S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition", IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. 37, No. 12, pp.1857-1869, 1989.

[6] J. Sirigos, N. Fakotakis, G. Kokkonakis, "A hybrid syllable recognition system based on vowel spotting", Jou. of Speech Com., Vol. 38, pp.427-440, 2002.

[7] T. Nagarajan, H.A. Murthy, "Language identification using parallel syllable-like unit recignition", IEEE/ICASSP, Vol.1, pp.401-404, 2004.

[8] S. Greenberg, "Speaking in short hand- A syllable centric perspective for understanding pronunciation variation", In. Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, 1998.

[9] W. Reichel, G. Ruske, "Syllable segmentation of continuous speech with artificial neural network", Eurospeech Conf., pp.1771-1774, 1993.

[10] K. Kirchhoff, "Syllable-level desynchronization of phonetic features for speech recognition", ICSLP, Vol. 4, pp.2274-2276, 1996.

[11] M.J. Hunt, M. Lening, P. Mermelstein, "Experiments in syllable-based recognition of continuous speech", IEEE/ICASSP, Vol.3, pp.880-883, 1980.

[12] S.L. Wu, M.L. Shire, S. Greenberg et al, "Integration syllable boundary information into speech recognition", IEEE/ICASSP, Vol.2, pp.987-990, 1997.

[13] R. Janakiraman, J.C. Kumar, H. . Murthy, "Robust syllable segmentation and its application to syllable-centric continuous speech recognition", IEEE/ NCC, India, pp.1-5, 2010.

[14] H. Tolba, M. Azmi, "Comparative experiments to evaluate the use of syllable for large-vocabulary automatic speech recognition", IEEE/CCSIT, pp.250-253, 2009

[15] V. Barkhoda, A. Bahrampour et al, "A comparative study on quality of different text to speech systems based on variant speech units for Kordi language", ISCEE, Tabriz, Iran, 2009.

[16] M. Bacchiani, M. Ostendorf, "Design of a speech recognition system based on acoustically derived segmental units", IEEE/ICASSP, Vol.1, pp.443-446, 1996.

[17] A. Ganapathiraju, J. Hamaker et al, "Syllable-based large vocabulary continuous speech recognition", IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. 9, pp.358-366, 2001.

[18] V.K. Prasad, T. Nagarajan, H.A. Murthy, "Continuous speech recognition using automatically segmented data at syllabic units", ICSP, Vol. 1, pp.235-238, 2002.

[19] H.N. Ting, Y. Jasmy, S. Hossein et al, "Malay syllable recognition based on multilayer perceptron and dynamic time warping", INSSPA, Vol. 2, pp.743-744, 2001.

[20] H. Matsu'ura, T. Nitta, S. Hirai et al, "A large vocabulary word recognition system based on syllable recognition and nonlinear word matching", IEEE/ICASSP, Vol.1, pp.183-186, 1988.

[21] A. Tanaka, S. Kamiya, "A speech processing system based on syllable identification by using phonological patterns", IEEE/ICASSP, pp.2231-2234, 1986.
[22] S. Zhang, Q. Shi, Y. Qin, "Modeling syllable-based pronunciation variation for accented Mandarin speech recognition", IEEE/ICPR, pp.1606-1609, 2010.
[23] N.T Umpon, S. Chansareewittaya, S. Auephanwiriyakul, "Phoneme and tonal accent recognition for Thai speech", Elsevier Jou. Expert Systems with Appl., Vol. 38, pp.13254-13259, 2011.
[24] W. Hu, T. Huang, B. Xu, "Study on prosodic boundary location in Chinese Mandarin", IEEE/ICASSP, Vol. 1, pp. 501-504, 2002.
[25] D. Wang, L. Lu, H.J. Shang, "Speech segmentation without speech recognition", IEEE/ICASSP, Vol. 1, pp.468-471, 2003.
[26] F. Tamborini, "Prosodic prominence detection in speech", ISSPA, Vol.1, pp.385-388, 2003.
[27] S. Kim, "The role of prosodic cues in word segmentation of Korean", Int. Speech Conf., pp.3005-3008, 2004.
[28] K.P. Li, "Automatic language identification using syllabic features", IEEE/ICASSP, pp.297-300, 1994.
[29] A. Noetzel, "Robust syllable segmentation of continuous speech using neural networks", IEEE/CEI, pp.580-585, 1991.
[30] P. Nel, J.D. Preez, "Automatic syllabification using hierarchical hidden Markove models", IEEE/ICASSP, Vol. 1, pp.768-771, 2003.
[31] N. Jittiwarangkul, S. Jitapunkul et al, "Thai syllable segmentation for connected speech based on energy", IEEE/APCCS, pp.169-172, 1998.
[32] V.K. Prasad, T. Nagarajan, H.A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay function", Jou. of Speech Com., Vol. 42, pp.429-446, 2004.
[33] G. Sheikhi, F. Almasganj, "Segmentation of speech into syllable units using fuzzy smoothed short term energy contour", ICBE, pp.195-198, 2011.
[34] D.L. Donoho, "De-noising by soft thresholding", Jou. IEEE Trans. on Inf. Theory, Vol. 41, No. 3, pp. 613-627, 1995.
[35] Y.S. Ing, N.K. Soo, K.Y. Chai, "Wavelet for speech denoising", IEEE/TENCON, Vol.2, pp.479- 482, 1997.
[36] M.T. Johnson, X. Yuan, Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding", Jou. of Speech Com., Vol. 49, No. 2, pp.123-133, 2007.
[37] Y. Ghanbari, M.R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", Jou. of Speech Com., Vol. 48, No. 8, pp. 927-940, 2006.
[38] G. Sheikhi, "Syllable boundary detection and analysis in Farsi connected speech using robust features and prosodic cues", M.Sc. Thesis, Biomedical Engineering Department, Amirkabir University of Technology, March 2007.
[39] H.M. Teager, S.M. Teager, "A phenomenological model for vowel production in the vocal tract", Ch. 3, pp.73-109, College-Hill Press, 1983.
[40] G. Sheikhi, F. Almasganj, "Syllable segmentation of Farsi connected speech using variable threshold", ACCSI, Feb. 2007, Iran.
[41] M. Bahoura, J. Rouat, "Wavelet speech enhancement based on time–scale adaptation", Jou. of Speech Com., Vol. 48, No. 12, pp.1620-1637, 2006.
[42] X.P. Zhang, M.D. Desai, "Adaptive denoising based on SURE risk", Jou. of IEEE Signal Proc. Letters, Vol. 5, No. 10, pp.265-267, 1998.
[43] S. Tabibian, B. Zamani Dehkordi et al, "A proposed wavelet basis matched to speech signal for enhancement and evaluation of effective parameters", ACCSI, Kish, March 2008.
[44] M. Bijankhan, M.J. Sheikhzadegan, "FARSDAT- the Farsi spoken language database", ICSST, Perth, Australia, Vol. 2, pp.826-829, 1994.