مقایسه و ارزیابی کارآیی انواع روش‌های استخراج پارامترهای بازنمایی و هنجارسازی در بازشناسی مستقل از گوینده گفتار

مهدی رحیمی‌نژاد
کارشناسی ارشد

سید علی سیدصالحی
استادیار

دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر

چکیده

فرآیند بازشناخت گفتار در اولین گام خود نیازمند یک نمایش مناسب از اطلاعات سیگنال گفتاری است. این بازنمایی باید شامل ویژگی‌هایی از سیگنال گفتار باشد که حاوی اطلاعات مفیدی جهت متمایزکردن آواهای مختلف باشند. برای استخراج این بیان مناسب اطلاعات سیگنال یا پارامترهای بازنمایی، روش‌های متفاوتی ارائه شده‌است که در سیستم‌های بازشناخت مختلفی مورد استفاده قرارگرفته‌اند. از رایج‌ترین این روش‌ها، روش‌های طیفی می‌باشند که در آنها تحلیل طیفی از سه طریق تبدیل فوریه زمان کوتاه، مدل پیشگویی خطی و تبدیل ویولت صورت می‌گیرد. در این مقاله متداول‌ترین روش‌های استخراج پارامترهای بازنمایی مورد مقایسه قرارگرفته‌اند و با اعمال تغییراتی در الگوریتم برخی از آنها و نیز استفاده از روش‌های نرمالیزهٔ متفاوت، مناسب‌ترین آنها در بکارگیری برای بازشناسی مستقل از گویندهٔ گفتار عاری از نویز، به کمک یک شبکه عصبی جلوسو با تأخیر زمانی، معرفی گردیده‌اند. در ارزیابی‌های انجام شده روش‌های طیفی الهام گرفته از سیستم شنوایی انسان و مبتنی بر تبدیل فوریه بالاترین صحت بازشناخت را ارائه داده‌اند.

کلمات کلیدی

پارامترهای بازنمایی، هنجارسازی به گوینده، هنجارسازی به بلندی صدا، LHCB، MFCC، PLP، LPCC، ویولت، بازشناخت گفتار مستقل از گوینده.

A Comparative Study of Representation Parameters Extraction And Normalization Methods for Speaker Independent Recognition of Speech

M. Rahiminejad
M.Sc.

S. A. Seyyed Salehi
Assistant Professor

Biomedical Engineering Department,
Amirkabir University of Technology

Abstract

An automatic speech recognition system at The first step for a proper representation of speech input which contains all useful information of the signal. These representations must be the voice features, which may distinguish different phonemes of a language. There are several methods for feature extraction that are used in ASR systems. The most common methods are spectral ones. The spectral analysis is based on Short Time Fourier Transform (STFT), Linear Prediction (LP), or Discrete Wavelet Transform (DWT). In this paper, these feature extraction methods, their modified versions, and some normalization algorithms were studied and the most suitable feature extraction methods for time delay neural network classifier were identified. It was shown that the perceptual

STFT based features are the best, among the features studied in this paper.

Keywords

Representation parameters, speaker normalization, loudness normalization, LHCB, MFCC, PLP, LPCC, wavelet, speaker independent

مقدمه

فرآیند بازشناخت گفتار نیازمند به بیان مناسب سیگنال گفتار ورودی است که اطلاعات مفید آنرا دربر داشته باشد. این اطلاعات گفتار شامل ویژگی‌هایی از سیگنال گفتار می‌باشد که دارای اطلاعات مفیدی جهت متمایزکردن آواهای مختلف می‌باشند. از آنجا که در بازشناخت گفتار، اولین گام، استخراج پارامترهای بازنمایی می‌باشد و چنانچه اطلاعات مفیدی در این مرحله از دست برود دیگر در مراحل بعدی ممکن است قابل جبران نباشد، استخراج پارامترهای بازنمایی را می‌توان یک مرحله زیربنایی در سیستم بازشناخت گفتار به حساب آورد.

تاکنون روشهای مختلفی جهت استخراج پارامترهای بازنمایی از سیگنال گفتار معرفی و مورد استفاده قرارگرفته‌اند. اکثر روشهای متداول استخراج ویژگی، روشهای طیفی می‌باشند یعنی پارامترهای بازنمایی در حوزهٔ فرکانس و از روی طیف سیگنال بدست آورده می‌شوند. برخی نتایج حاکی از این است که برای بازشناخت مستقل از گویندهٔ آواها به کمک شبکهٔ عصبی بازگشتی[1]، بازنمایی در دامنه طیفی مستقل از اینکه به چه روشی بدست آمده باشد، مناسب‌تر می‌باشد[۳]. روشهای طیفی خود انواع مختلف دارند و از یک دیدگاه می توان آنها را به دو دسته کلی تقسیم کرد؛ روشهایی که از آنالیزهای خطی استفاده می‌کنند و روشهایی که از سیستم شنوایی انسان الهام می‌گیرند. در سالهای اخیر گرایش به سمت ایده گرفتن از سیستم شنوایی انسان برای استخراج پارامترهای بازنمایی بوجود آمده است. چراکه سیستم شنوایی انسان یک سیستم بازشناخت با عملکردی بسیار مطلوب می‌باشد. دنگ[2] معتقد است که موفقیت نهایی در ساخت سیستمهای بازشناخت با توانایی‌های مشابه مغز انسان به فهم کامل پردازشهای موجود در سیستم طبیعی شنوایی انسان وابسته است. بدین منظور وی یک مدل غیر خطی برای گوش داخلی پیشنهاد داده و با آزمایش آن به این نتیجه رسیده است که پردازش‌های غیرخطی و الهام گرفته از مدل گوش داخلی، توانایی های قوی تری در پردازش گفتار نسبت به روشهای آنالیز خطی فراهم نموده‌اند[۴]. همچنین هالتون[3] با انجام یک سری آزمایشات جهت آشکارسازی و طبقه بندی ویژگیهای گفتاری آواهای واکدار به این نتیجه رسیده است که الگوریتم‌های بازشناخت گفتار برمبنای مدل شنوایی جایگزین‌های مؤثری نسبت به روشهای طیفی معمولی جهت استخراج فورمنت‌ها هستند [۵].

تحلیلهای طیفی که در روشهای طیفی استخراج پارامترهای بازنمایی مورد استفاده قرار می‌گیرند عبارتند از تبدیل فوریه زمان کوتاه، مدل پیشگویی خطی و تبدیل ویولت. تبدیل فوریه گسستهٔ زمان کوتاه یک روش کلاسیک در تحلیل طیفی سیگنال گفتار می باشد. با استفاده از رابطه تبدیل فوریه از روی نمونه‌های زمانی سیگنال، نمونه‌های فرکانسی آن ساخته می شود. در مدل پیشگویی خطی، یک مدل ریاضی تمام قطب برای سیستم تولید گفتار فرض می‌گردد و ضرایب این تابع تمام قطب که حاوی اطلاعات مربوط به فورمنت‌ها هستند تخمین زده می‌شوند. تبدیل ویولت نیز چون برای سیگنال‌های غیرایستان مناسب است و خاصیت چند رزولوشنی دارد، در سالهای اخیر توجه محققان را جهت استفاده از آن در تحلیل طیفی سیگنال گفتار و برای استخراج پارامترهای بازنمایی به خود جلب کرده است.

با توجه به این مطلب که در طبقه‌بندی آواها، میزان صحت طبقه‌بندی به این نکته بستگی دارد که روش طبقه‌بندی تا چه حد با روش بازنمایی سیگنال انطباق داشته باشد[۶]، بررسی روش‌های بازنمایی مختلف برای یک طبقه‌بندی کنندهٔ خاص مورد نیاز است. در این مقاله هدف مقایسه ای بین روش‌های مختلف استخراج پارامترهای بازنمایی در شرایطی یکسان در بکارگیری برای بازشناسی مستقل از گویندهٔ گفتار عاری از نویز می‌باشد. در این سیستم از یک شبکه عصبی جلوسو با تأخیر زمانی به عنوان طبقه‌بندی کننده استفاده می‌شود. علاوه بر روش‌های بازنمایی متداول، با بهبود الگوریتم برخی از آنها و نیز استفاده از روش‌های هنجارسازی متفاوت جهت هنجارساختن پارامترهای بازنمایی، مناسب‌ترین روش‌های بازنمایی جهت استفاده در سیستم فوق معرفی می‌گردند.

در ادامه مقاله، دادگان مورد آزمایش و شبکه عصبی که به عنوان طبقه‌بندی کننده در آزمایشات مورد استفاده قرار گرفته است در بخش ۲ معرفی می‌گردند و در بخش ۳ به معرفی روش‌های هنجارسازی پارامترهای بازنمایی که در [۱] برای پارامترهای مورد استفاده توسط شبکه عصبی مناسب تشخیص داده‌شده‌اند پرداخته می‌شود. در بخش ۴ روش‌های رایج استخراج پارامترهای بازنمایی که با توجه به گزارشات مقالات و کارهای قبلی انتخاب شده‌اند، معرفی می‌شوند. در بخش ۵ نتایج حاصل از پارامترهای مختلف بیان شده با هم مقایسه می‌گردند و جمع بندی و نتیجه‌گیری نهایی در بخش ۶ آورده شده است.

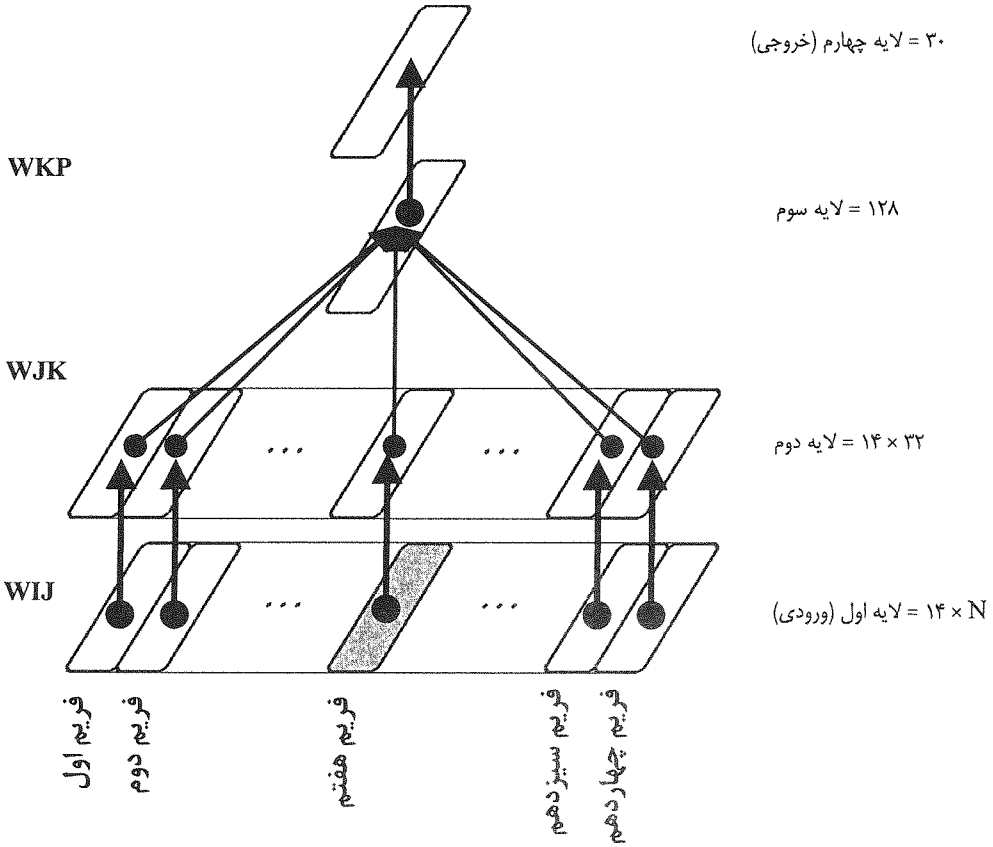
۱ـ مدل شبکه عصبی بازشناس و دادگان آزمایش

برای طبقه‌بندی آواها در این مقاله از یک شبکه عصبی جلوسوی زمانی استفاده شده است. این شبکه براساس نگاشت مسیر بردار بازنمایی به فضای آواها طراحی شده است و در واقع توسط این شبکه درصد صحت بازشناخت فریم‌ها یا به بیان دیگر نمونه‌های مسیر بردارهای بازنمایی، استخراج شده توسط هرکدام از روش‌ها، تعیین می‌گردد. توسط تکنیک‌هایی می توان از بازشناخت فریم‌ها به بازشناخت آواها رسید اما چون هدف مقایسه عملکرد روش‌های مختلف استخراج پارامترهای بازنمایی است و مهم یکسان بودن شرایط آزمایش برای تمام پارامترها است، لذا مقایسه صحت بازشناخت فریم‌ها نیز خود یک معیار مناسب جهت مقایسه عملکرد روش‌های استخراج پارامترهای بازنمایی می‌باشد. همانطور که در شکل ۱ نشان داده شده است، شبکه عصبی مورد استفاده دارای ۱۴ تأخیر زمانی در لایه ورودی می‌باشد، یعنی در ورودی به ۱۴ فریم کنار هم نگاه می‌کند و دارای دو لایه پنهان به ابعاد ۳۲×۱۴ و ۱۲۸ می‌باشد، لایه خروجی نیز دارای ۳۰ نورون، به تعداد آواهایی که باید بازشناسی شوند، می‌باشد. تابع تصمیم گیری هر نورون یک تابع غیر خطی از نوع سیگموئید است که باعث تصمیم گیری نرم در خروجی هر نورون می‌گردد. این شبکه بر اساس طی مسیر بازنمایی بر روی دادگان آزمون لغزیده و همواره یکی از خروجیها که بیانگر یکی از آواها است فعال می‌شود. در تعلیم شبکه و برای بدست آوردن وزنهای مطلوب بین لایه‌ها از روش پس انتشار خطا استفاده می‌گردد و وزنها در جهتی اصلاح می‌شوند که مجموع مربعات خطای خروجی (اختلاف بین خروجی شبکه و خروجی مطلوب) به حداقل برسد. درآموزش شبکه ضریب یادگیری شبکه و ضریب ممنتوم به ترتیب ۰/۱ و ۰/۷ انتخاب شده‌اند و برای اینکه شرایط آموزش و شرایط کار برای تمام حالت‌ها و برای تمام ویژگیهای مختلف یکسان باشد، ضریب یادگیری شبکه در طول آموزش ثابت می‌ماند.

با توجه به اینکه سیستم مورد نظر برای بازشناخت گفتار پیوسته طراحی شده است، لذا دادگانی که برای آموزش و آزمون سیستم انتخاب می‌شوند، باید جملات پیوسته باشند. یعنی گوینده یک جمله را به طور معمولی و بدون ایجاد وقفه‌های عمدی در بین کلمات ادا می‌کند. از طرفی چون سیستم بازشناخت مستقل از گوینده است لذا برای آموزش مدل بازشناخت تا آنجا که ممکن است باید از تنوع گویندگان مختلف استفاده کرد تا مدل بازشناخت در هنگام آموزش فقط به یک سری جملات بیان شده توسط یک یا چند گوینده خاص وابسته نشود و تا آنجا که ممکن است گوینده‌های متفاوت را ببیند. برای این منظور از دو جملهٔ ۴۰۰ و ۵۰۰ از ۱۰۱ گوینده از مجموعه دادگان فارس‌دات که در محیطی عاری از نویز ضبط شده و با فرکانس ۴۴۱۰۰ هرتز نمونه برداری شده‌اند، استفاده می‌شود[۱۳]. جملات ۷۱ گوینده برای تعلیم شبکه و جملات ۳۰ گوینده دیگر به عنوان دادگان آزمون مورد استفاده قرار می‌گیرند. در این دادگان آواها تقطیع نشده‌اند و صرفاً برحسب شماره نمونه‌ها برچسب‌دهی گردیده‌اند. لذا سیگنال‌های گفتاری و برچسب‌دهی آن بدون تقطیع در دسترس می‌باشند.

۲ـ روش‌های هنجارسازی پارامترهای بازنمایی

در بسیاری از سیستمهای بازشناخت گفتار، استفاده از روشهای استخراج پارامترهای بازنمایی به تنهایی پارامترهای مناسبی در اختیار قرار نمی‌دهند و پس از حصول‌بردارهای ویژگی، نیاز است که این پارامترها نسبت به عوامل مختلفی به هنجار[۴] شوند [۷]-[۱۰]. با توجه به اینکه مدل بازشناخت مورد استفاده یک شبکه عصبی جلو سو با تأخیر زمانی می‌باشد، در این قسمت به‌معرفی برخی از روشهای نرمالیزهٔ پارامترهای بازنمایی که طی یک سری آزمایشات برای پارامترهای مورد استفاده برای شبکه عصبی مفید شناخته شده‌اند [۱]، می‌پردازیم.



شکل(۱) ساختار شبکه عصبی جلوسوی زمانی.

الف ـ هنجارسازی عرضی پارامترهای بازنمایی

هنگامیکه یک گوینده، یک آوای خاص را چندبار با بلندی صدای متفاوت بیان کند، پارامترهایی که از این آوا بدست می‌آیند، با هم یکسان نیستند. علت این است که بلندی صدای گوینده در انرژی کل طیف تأثیر مستقیم می‌گذارد و لذا هنگامیکه انرژی طیفی در باندهای فرکانسی مختلف به عنوان پارامترهای بازنمایی محاسبه می‌شوند، این انرژی کل در این مقادیر پارامترها نمود پیدا می‌کند و لذا صحت بازشناسی سیستم بازشناخت گفتار کاهش پیدا می‌کند. لذا کاهش حساسیت پارامترهای بازنمایی نسبت به بلندی صدا می‌تواند باعث بهبود عملکرد سیستم بازشناخت گفتار شود.

در این روش هنجارسازی، میانگین N پارامتر بردار بازنمایی بدست آمده از هر فریم حساب شده وسپس این مقدار میانگین از هر یک از پارامترها کاسته می‌شود [۱]. چون این روش بر روی پارامترهای حاصله از بانک فیلترها اعمال می‌شود، و پارامترها لگاریتم انرژی داخل هر فیلتر هستند، میانگین پارامترها با انرژی کل فریم متناسب می‌باشد و کم کردن این مقدار از پارامترهای بردار بازنمایی مانند آن است که پارامترهای بازنمایی به تقریبی از انرژی کل فریم به هنجار شده باشند. اگر $\bar{x}$ بردار بازنمایی یک فریم باشد و x_i پارامتر i ام (لگاریتم انرژی فیلترiام) این بردار باشد، می‌توان این نحوهٔ هنجارسازی را به صورت زیر بیان کرد.

$$\bar{x} = [x_1, x_2, \mathrm{K}, x_N]$$

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{x}_{new} = \bar{x} - m = [x_1 - m, x_2 - m, \mathrm{K}, x_N - m] \qquad (۱)$$

که m میانگین پارامترهای بردار بازنمایی یک فریم و $\bar{x}_{new}$ بردار به هنجار شده می‌باشند.

ب ـ هنجارسازی طولی پارامترهای بازنمایی

در ایـن روش، هـر پارامتر بازنمایی به میانگین آن در کل دادگان به هنجار می‌شود. این روش هنجارسازی باعث می‌شود که دامنه تغییرات همه پارامترها تقریباً به محدودهٔ بین ۱- تا ۱+ منتقل گردد و در دو طرف صفر تقریباً متوازن باشند [۱]. در این روش، ابتدا بردار میانگین پارامترها روی کل دادگان محاسبه می‌شود.

$$\bar{k}_1 = \frac{1}{M}\sum_{i=1}^{M}\bar{x}_i \qquad (۲)$$

که در این رابطه M برابر تعداد بردارهای بازنمایی در کل دادگان می‌باشد و $\bar{x}_i$، i امین بردار بازنمایی است. پس از بدست آمدن بردار میانگین $\bar{k}_1$، این بردار از تمام بردارهای بازنمایی کم می‌گردد.

$$\bar{q}_i = \bar{x}_i - \bar{k}_1 \qquad (۳)$$

سـپس بردار میانگین نرم اول (میانگین قدر مطلق) [۱] و یا بردار میانگین نرم دوم (میانگین مجذور) [۱۲] پارامترهای $\bar{q}_i$ روی کل دادگان محاسبه می‌گردد. در استفاده از میانگین نرم اول داریم

$$\bar{k}_2 = \frac{1}{M}\sum_{i=1}^{M}|\bar{q}_i| \qquad (۴)$$

و برای بدست آوردن پارامترهای به هنجار شده، تمامی بردارهای $\bar{q}_i$ عنصربه‌عنصر به بردار $\bar{k}_2$ تقسیم می‌شوند.

$$\bar{x}_i^{new} = \left\{\frac{q_{in}}{k_{2n}} \middle| \; n = 1,2,K,N\right\} \qquad (۵)$$

در رابطـه فوق، $\bar{x}_i^{new}$ بردار بازنمایی به هنجار شده و N تعداد پارامترهای هر بردارباز‌نمایی است. q_{in} و k_{2n} به ترتیب n امین مؤلفه از بردارهای $\bar{q}_i$ و $\bar{k}_2$ هستند. اما در استفاده از میانگین نرم دوم داریم

$$\bar{k}_2 = \frac{1}{M}\sum_{i=1}^{M}\bar{q}_i^{\,2} \qquad (۶)$$

و برای بدست آوردن پارامترهای به هنجار شده، تمامی مؤلفه‌های بردارهای $\bar{q}_i$ به جذر مؤلفه‌های بردار $\bar{k}_2$ تقسیم می‌شوند.

$$\bar{x}_i^{new} = \left\{\frac{q_{in}}{\sqrt{\alpha k_{2n}}} \middle| \; n = 1,2,K,N\right\} \qquad (۷)$$

در رابطـه فوق α ضریبی است که بواسطه آن واریانس پارامترهای به هنجار شده تعیین می‌گردد. و مقدار α برابر عکس واریانس پارامـترهای بـه هنجار شده می باشد. به عنوان مثال اگر می خواهیم واریانس پارامترهای به هنجار شده برابر ۰/۵ باشد، بایـد α=۲ انتخاب شود.

ایـن نحـوهٔ هنجارسـازی موجـب می‌شود کـه مقادیر مؤلفه‌های مختلف دادگـان در فضای N بعدی ورودی شبکه عصبی در

محلهای مناسبتری برای شکل‌گیری مدل نگاشت فضای بازنمایی به آواها قرار گیرند و از طرفی با هم دامنه شدن پارامترهای بازنمایی نسبت به هم، ارزش همه مؤلفه‌ها از دید شبکه عصبی، تقریباً یکسان گردد و لذا شبکه بهتر می‌تواند عمل طبقه‌بندی را روی پارامترهای به هنجار شده انجام دهد[۱].

پ ـ هنجارسازی طولی پارامترهای بازنمایی، جداگانه برای هر گوینده

این روش نیز مانند روش هنجارسازی قبلی است با این تفاوت که درروش قبلی، بردارهای $\bar{k}_1$ و $\bar{k}_2$ روی تمام دادگان آموزش محاسبه می‌شدند اما در این روش بردارهای $\bar{k}_1$ و $\bar{k}_2$ روی دادگان مربوط به هر گوینده و به صورت جدا محاسبه می‌شوند و دادگان هر گوینده دارای یک بردار $\bar{k}_1$ و یک بردار $\bar{k}_2$ مختص به خود می‌باشد [۱]. مزیت این روش به روش قبل در این است که بردارهای $\bar{k}_1$ و $\bar{k}_2$ تخمینی از مشخصات گویندهٔ مربوطه را دربر دارند و پس از هنجارسازی بردارهای بازنمایی مربوط به دادگان یک گوینده، این پارامترها تا حدودی نسبت به گوینده به هنجار می‌شوند. اما در روش قبل دو بردار $\bar{k}_1$ و $\bar{k}_2$ میانگینی از اطلاعات همه گویندگان را تخمین می‌زنند.

۳ـ روش‌های استخراج پارامترهای بازنمایی

روش‌های طیفی استخراج پارامترهای بازنمایی ازنظر نوع تحلیل طیفی به سه دسته تقسیم می‌شوند که در این قسمت متداول‌ترین روش‌های هر دسته معرفی می‌شوند.

الف ـ روش‌های طیفی مبتنی بر تبدیل فوریه زمان کوتاه

در روش‌های طیفی استخراج پارامترهای بازنمایی بخصوص در روش‌هایی که از تحلیل فوریه جهت بدست آوردن طیف سیگنال گفتار استفاده می‌شود، عموماً از بانک فیلتر جهت محاسبه انرژی طیف حول فرکانس‌هایی مشخص به عنوان پارامترهای بازنمایی استفاده می‌گردد. تعداد این فیلترهای میانگذر در سیستم های مختلف متفاوت است ولی معمولاً بین ۶ الی ۲۰ فیلتر مورد استفاده قرار می‌گیرند. افزایش فیلترها معمولاً موجب بهبود کیفیت بازشناخت می‌شود ولی در صورت کاهش پهنای باند فیلترها تا حد کمتر از فرکانس گام[۵] صدای گوینده، کیفیت بازشناخت افت می کند[۲]. تنظیم فواصل مابین فیلترها به صورت غیرخطی و در مقیاس مل[۶] یا بارک[۷] که مقیاس‌هایی الهام گرفته از سیستم شنوایی انسان می‌باشند، انجام می‌گیرد. روابط تبدیل مقیاس هرتز به این دو مقیاس عبارتند از[۳۰]:

$$f_{mel} = 2595 \log\left(1+\frac{f_{Hz}}{700}\right) \qquad (۸)$$

$$f_{Bark} = 6 \ln\left(\frac{f_{Hz}}{600}+\sqrt{\left(\frac{f_{Hz}}{600}\right)^2+1}\right) \qquad (۹)$$

هردو مقیاس تقریباً شبیه به هم بوده و تا فرکانس یک کیلوهرتز بصورت تقریباً خطی و بالاتر از این فرکانس به صورت لگاریتمی می باشند. پارامترهای طیفی مبتنی بر تبدیل فوریه که در این مقاله مورد مطالعه قرارگرفته اند عبارتند از LHCB[8]، LHCBC[9]، MFCC[10] و PLP[11] که در ادامه به معرفی آنها می پردازیم.

پارامترهای LHCB، لگاریتم انرژی خروجی فیلتر بانک‌های مجذور هنینگ و متساوی الفاصله در مقیاس بارک با پهنای یک و فاصله یک نسبت به هم می‌باشند [۲]. الگوریتم استخراج پارامترهای LHCB بصورت زیر است:

گام اول ـ انتخاب یک فریم از سیگنال به طول N نمونه (در پیاده سازی تمام الگوریتم‌ها در این مقاله طول فریم ۲۳ میلی‌ثانیه درنظرگرفته شده است که با توجه به فرکانس نمونه‌برداری ۴۴/۸ کیلوهرتز از سیگنال گفتار، طول هر فریم N=۱۰۲۴ نمونه خواهدبود.)

گام دوم ـ حذف مقدار dc فریم

گام سوم ـ ضرب فریم در پنجره زمانی همینگ با رابطه زیر

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & ;\ 0\le n\le N-1\\ 0 & ;\ \text{otherwise}\end{cases} \qquad (۱۰)$$

گام چهارم ـ محاسبه تبدیل فوریه زمان کوتاه N نقطه ای از فریم پنجره گذاری شده $X(k)$

گام پنجم ـ محاسبه طیف توان $|X(k)|^2$

گـام ششم ـ اعمال بانک فیلترهای مجذور هنینگ روی طیف توان. پاسخ تبدیل گسسته فوریه یک فیلتر هنینگ به صورت زیر است:

$$\psi(k)=\begin{cases}0.5-0.5\cos\left(\dfrac{2\pi k}{M}\right) & ;\ 0\le k\le M\\ 0 & ;\ \text{Otherwise}\end{cases} \qquad (۱۱)$$

گام هفتم ـ محاسبه لگاریتم خروجی هر فیلتر به عنوان ویژگی

$$E_j=\sum_{k=1}^{N}\left|\psi_j(k)\right|^2|X(k)|^2 \quad ;\ j=0,1,K\ ,M \qquad (۱۲)$$

$$C_j=\log(1+E_j) \qquad (۱۳)$$

که M تعداد فیلترهای بکار رفته در بانک فیلتر می‌باشد.

بـرای بدسـت آمدن پارامترهای LHCBC کافیست که پارامترهای LHCB را به حوزه کپستروم انتقال داد. برای این منظور بـاید از پارامترهای C_j عکس تبدیل فوریه گسسته گرفت. اما چون C_j مقادیری حقیقی و متقارن می‌باشند بجای عکس تبدیل فوریه گسسته می توان از تبدیل کسینوسی گسسته استفاده کرد:

$$c_m=\sum_{j=1}^{M}\ln(1+E_j)\cos\left(m\left(\frac{2j-1}{2}\right)\frac{\pi}{M}\right) \quad ;\ 1\le m\le L \qquad (۱۴)$$

در رابطـه فـوق E_j مجمـوع انـرژی فیلتر i ام است و M تعداد کل فیلترها می‌باشد. L نیز تعداد ضرایب مورد نظر در دامنه کپسـتروم اسـت. مقـدار L همـواره کوچـک‌تـر یا مساوی M می‌باشد. اگر L=M اختیار شود، از تمام ضرایب کپستروم به عنوان پارامـترهای بـاز نمایـی اسـتفاده کـرده‌ایـم. اما به خاطر اینکه آنالیز کپسترال کانوالو دو سیگنال مربوط به تحریک و فیلتر لوله صـوتی را بـه صـورت جمـع در مـی‌آورد و چـون لگاریتم طیف فیلتر لوله صوتی، تغییرات آرام دارد ولی سیگنال تحریک باعث تغیـیرات شدید در سیگنال گفتار می‌شود، لذا ضرایب ابتدایی کپستروم بیانگر طیف فیلتر لوله صوتی می‌باشند و ضرایب بالایی کپستروم مربوط به سیگنال تحریک هستند[۱۴]. به همین خاطر فقط‌چند ضریب اول حفظ می‌شوند و بقیه ضرایب دور ریخته مـی‌شـوند. بـا توجه به مطالعاتی که در[۲] بر روی تعداد فیلترهای فرکانسی برای بدست‌آوردن پارامترهای LHCB و LHCBC انجام‌گرفته‌است، تعداد آنها برابر M=۱۸ اختیار شده‌است. در سیستم‌های بازشناخت گفتار تعداد ضرایب مورد استفاده کپستروم (L) کمـتر یـا مساوی ۱۵ اختیار می‌شودکه در این مقاله نیز L=۱۵ اختیار شده است. همانطور که گفته شد آنالیز کپسترال دو سیگنال مربوط به تحریک و ناحیه صوتی را به‌صورت خطی تفکیک‌پذیر می‌کند اما این تفکیک ایده‌آل نیست و این دو سیگنال بـر همدیگـر اثـر مـی‌گذارنـد و چـند ضـریب اول کپسـتروم علاوه بر اطلاعات ناحیه صوتی و اطلاعات گفتاری، حاوی اطلاعات سـیگنال تحـریک و اطلاعـات گویـنده نـیز مـی‌باشـد. جهت کاهش این اثرات درچند ضریب اول، از اعمال تابع وزن بر ضرایب

کپسـتروم اسـتفاده مـی‌شـود. در این تابع وزن، ارزش کمتری به ضرایبی که متأثر از اطلاعات گوینده می‌باشند، داده می‌شود و ضـرایبی کـه فقـط حـاوی اطلاعـات ناحیه صوتی و اطلاعات گفتاری می‌باشند، ارزش بیشتری پیدا می‌کنند. درمراجع متفاوت توابـع وزن متفاوتی جهت اعمال بر ضرایب کپستروم تعریف شده‌اند. معمول ترین تابع وزن مناسب برای ضرایب کپسترومی که براساس خروجی بانک فیلترها بدست می‌آید به صورت زیر می‌باشد[۱۵] و [۳۱]:

$$g(m) = 1 + \frac{L}{2}\mathrm{Sin}\left(\frac{m\pi}{L}\right) \quad ; 1 \le m \le L \qquad (۱۵)$$

که در رابطه فوق L برابر تعداد ضریب کپستروم انتخاب شده است.

ضرایب کپستروم در مقیاس مل (MFCC)، در بسیاری از سیستم‌های بازشناخت گفتار به عنوان پارامترهای بازنمایی مطرح اسـت و ایـن روش استخراج ویژگی بیشتر از سایر روش‌ها مورد استفاده قرار می‌گیرد. الگوریتم این روش همانند الگوریتم روش LHCBC اسـت و همانـند آن روش پـس از پـنجره گـذاری و تبدیل فوریه‌زمان کوتاه، یک بانک فیلترها که نحوهٔ توزیع آن به صـورت غـیر خطـی اسـت، بـر طـیف تـوان اعمال می‌گردد و پس از محاسبه لگاریتم مجموع خروجی هر فیلتر با اعمال تبدیل کسینوسـی‌گسسـته، ضرایب کپستروم بدست می‌آیند و در آخر با اعمال تابع وزن بر روی ضرایب‌کپستروم، پارامترهای بازنمایی MFCC حاصل می‌شوند.

یکـی از تفـاوت اساسـی بیـن پارامـترهای MFCC و LHCBC در مقـیاس غیرخطـی است‌که برای محور فرکانس و توزیع فیلـترها، برگـزیده شـده است. همانطور که دیده شد، برای پارامترهای‌LHCB و LHCBC از مقیاس بارک استفاده شده است و بـرای پارامـترهای MFCC از مقـیاس مـل استفاده می‌شود. تفاوت مهم دیگر در شکل فیلترهای استفاده شده در بانک فیلترها مـی‌باشـد. در پارامـترهای MFCC از فیلـترهای مثلثـی استفاده می شود، حال آنکه در پارامترهای LHCB از فیلترهای مجذور هنینگ استفاده می شود.

یکـی دیگـر از روش‌های استخراج پارامترهای بازنمایی که بر اساس سیستم شنوایی‌انسان و الهام گرفته شده از آن می‌باشد، پارامـترهای پیشـگویی خطـی ادراکـی (PLP)است که در سال ۱۹۹۰ و توسط هرمانسکی معرفی شده است. در این روش سه ویژگی از فیزیولوژی شنوایی به منظور رسیدن به تخمینی از طیف شنوایی مورد استفاده قرارگرفته است:

۱ـ رزولوشن طیفی بر مبنای باندهای بحرانی

۲ـ منحنی بلندی یکنواخت[۱۲] صدا

۳ـ قانون توان شدت-بلندی[۱۳] صدا

ویژگـی اول تـبدیل محـور فرکانسـی از مقیاس هرتز به مقیاس بارک است. این تبدیل‌که برگرفته از سیستم شنوایی انسان اسـت در روش LHCB نـیز مورد استفاده قرار گرفته است و دو ویژگی بعدی ویژگیهایی است که هرمانسکی در این روش مورد توجـه قـرار داده اسـت و بـیان نموده است که باعث بهبود عملکرد سیستم‌های بازشناخت گفتار گشته است[۱۶] و [۱۷]. با در نظرگرفتن سه ویژگی فوق و اعمال آنها بر روی سیگنال گفتار، طیف شنوایی به‌کمک یک مدل تمام قطب تخمین زده می‌شود. البـته بـرای اعمـال ایـن سه ویژگی فیزیولوژی، سیگنال گفتار به کمک تبدیل فوریه زمان کوتاه از حوزه زمان به حوزه فرکانس انـتقال مـی‌یابد و از این جهت این روش جزو روش‌های استخراج پارامترهای بازنمایی است که براساس تبدیل فوریه و در حوزه فرکانس انجام می‌گیرند. الگوریتم استخراج پارامترهای‌PLP را می‌توان به صورت زیرخلاصه کرد:

گام اول ـ انتخاب یک فریم از سیگنال به طول N نمونه

گام دوم ـ حذف مقدار dc فریم

گام سوم ـ ضرب فریم در پنجره زمانی همینگ به طول N نمونه

گام چهارم ـ محاسبه تبدیل فوریه N نقطه از فریم پنجره گذاری شده

گام پنجم ـ محاسبه طیف توان

گـام ششم ـ اعمال بانک فیلتر به شکل ذوزنقه نامتقارن بر روی طیف توان به همراه اعمال منحنی بلندی یکنواخت صدا. رابطه هر فیلتر ذوزنقه‌ای نامتقارن در مقیاس بارک به صورت زیر است:

$$\psi_j(z)=\begin{cases}10^{(z-z_j+0.5)} & ;\ z\le z_j-0.5\\ 1 & ;\ z_j-0.5<z<z_j+0.5\\ 10^{-2.5(z-z_j-0.5)} & ;\ z\ge z_j+0.5\end{cases} \qquad (۱۶)$$

کـه در رابطـه فوق z فرکانس در مقیاس بارک است و $z_j=0.9994j$ مرکز فیلتر jام می‌باشد. و منحنی بلندی یکنواخت صدا به صورت زیر تقریب‌زده شده است:

$$E(\omega)=1.151\sqrt{\frac{(\omega^2+144\times10^4)\omega^2}{(\omega^2+16\times10^4)(\omega^2+961\times10^4)}} \qquad (۱۷)$$

کـه ω فـرکانس زاویـه‌ای می‌باشد. اگر خروجی فیلتر jام پس از اعمال منحنی بلندی یکنواخت را F_j بنامیم، این مقدار توسط رابطه زیر حساب می‌گردد.

$$F_j=E(\omega_j)\sum_{k=1}^{N}\psi_j(k)|X(k)|^2 \quad ;\ j=0,1,K\ ,18 \qquad (۱۸)$$

گام هفتم ـ اعمال تبدیل شدت ـ بلندی صدا

$$\Phi_j=\sqrt[3]{F_j} \qquad (۱۹)$$

گام هشتم ـ اعمال مدل تمام قطب از مرتبه L=۵

گام نهم ـ تبدیل ضرایب مدل تمام قطب به ضرایب کپستروم حقیقی

$$\hat{x}(n)=\begin{cases}0 & ;\ n<0\\ \ln(G) & ;\ n=0\\ a_n+\sum_{k=1}^{n}(\frac{k}{n})x(n)a_{n-k} & ;\ 0<n\le L\end{cases} \qquad (۲۰)$$

که a_n ها ضرایب مدل تمام قطب، G بهره مدل تمام قطب و x(n) نیز ضرایب کپستروم مختلط می‌باشند.

گام دهم ـ اعمال وزن بر ضرایب کپستروم

$$\gamma(n)=n \quad ;\ 1\le n\le L \qquad (۲۱)$$

ب ـ روش‌های طیفی مبتنی بر مدل پیشگویی خطی

در ایـن روش‌هـا طـیف توان سیگنال بر اساس مدل پیشگویی خطی بدست می‌آید. یک مدل ریاضی مؤثر ولی در عین حال سـاده بـرای فرایند گفتار درنظر گرفته می‌شود که شامل یک منبع تحریک و یک فیلتر مدل کنندهٔ لوله صوتی است. این فیلتر لولـه صـوتی بـه کمـک یـک تـابع تمام قطب مدل می گردد و به کمک ضرایب این مدل که ضرایب پیشگویی خطی[۱۴] (LPC) مـی‌باشـند، فورمنـتهای طـیف گفتار قابل استخراج هستند. بطور کلی بر اساس این روش طیفی دو نوع پارامتر بازنمایی حاصل مـی‌گـردد. پارامـترهای LPCC[15] کـه بطور مستقیم ضرایب پیشگویی خطی به ضرایب کپستروم تبدیل می شوند و پارامترهای Mel-LPCC که از بانک فیلتر جهت استخراج ویژگی استفاده می‌شود[۲۰].

الگوریتم استخراج پارامترهای LPCC به صورت زیر است:

گام اول ـ انتخاب یک فریم از سیگنال به طول N نمونه

گام دوم ـ حذف مقدار dc فریم

گام سوم ـ ضرب فریم در پنجره زمانی همینگ به طول N نمونه

گام چهارم ـ محاسبه ضرایب پیشگویی خطی(LPC) به کمک روش لوینسن ـ دربین

گام پنجم ـ تبدیل ضرایب پیشگویی خطی (LPC) به ضرایب کپستروم پیشگویی خطی (LPCC)

$$\hat{h}(n)=\begin{cases}0 & ;\ n<0\\ \ln(\xi) & ;\ n=0\\ a(n)+\sum_{k=1}^{n-1}\left(\frac{k}{n}\right)h(n)h(k)a(n-k) & ;\ 0<n\le M\end{cases} \qquad (۲۲)$$

در رابطه فوق، M مرتبه مدل تمام قطب می‌باشد. $\hat{a}(n)$ ها ضرایب مدل تمام قطب و ξ خطای ناشی شده از تخمین‌گر مرتبه M ام در الگوریتم لوینسن ـ دربین می‌باشد که معادل بهره مدل تمام قطب می‌باشد. پارامترهای $\hat{h}(n)$ ضرایب کپستروم مختلط پیشگویی خطی می‌باشند.

گام ششم ـ انتخاب L ضریب اول کپستروم به عنوان پارامترهای LPCC

گام هفتم ـ ضرب تابع وزن در پارامترهای LPCC

در الگوریتم فوق مرتبه مدل تمام قطب متناسب با فرکانس نمونه‌برداری انتخاب می‌شود[۱۸] و برای سیگنالی با فرکانس نمونه برداری ۱۶ کیلوهرتز این مقدار معمولاً برابر ۱۴ یا ۱۶ اختیار می‌گردد و در این حالت از ۱۲ ضریب اول کپستروم به عنوان پارامترهای بازنمایی استفاده می‌نماییم. تابع وزنی که برای پارامترهای LPCC استفاده می‌شود مانند پارامترهای MFCC و LHCBC از رابطه (۱۵) بدست می‌آید.

برای بدست آوردن پارامترهای Mel-LPCC از الگوریتمی مانند LHCBC استفاده می‌شود با این تفاوت که در پارمترهای Mel-LPCC برای بدست آوردن طیف سیگنال از مدل پیشگویی خطی استفاده می‌گردد نه از تبدیل فوریه. و پس از بدست آمدن طیف از بانک فیلتر مجذور هنینگ در مقیاس مل استفاده می‌شود و انرژی هرباند به دامنه کپستروم انتقال می‌یابد. اما برای آنکه بتوان مقایسه بهتری بین پارامترهای LHCB و Mel-LPCC انجام داد، در پارامترهای Mel_LPCC بجای مقیاس مل از مقیاس بارک استفاده می‌کنیم و در ادامه از آن با عنوان پارامترهای Bark_LPCC یاد خواهیم کرد.

پ ـ روش‌های طیفی مبتنی بر تبدیل ویولت

برخلاف تبدیل فوریه، تبدیل ویولت گسسته[۱۶] (DWT) برای سیگنالهای غیرایستان نیز تعریف شده است و همچنین دارای خاصیت چند رزولوشنی[۱۷] می‌باشد[۲۱]. براساس دو خاصیت فوق، در چند سال اخیر گرایش به سمت استفاده از تبدیل ویولت افزایش یافته است [۲۳]-[۲۵].

یکی از روشهای استفاده از تبدیل ویولت، بکار بردن لگاریتم انرژی خروجی فیلترهایی است که تبدیل ویولت ایجاد می‌کند [۲۲]، [۲۳]، [۲۷]. تبدیل ویولت را می‌توان با بانک فیلتر و عمل کاهش نرخ نمونه برداری معادل سازی کرد[۲۸]. لذا تبدیل ویولت یک سیگنال، مشابه اعمال یک بانک فیلتر با باند فرکانسهای متفاوت بر روی سیگنال می‌باشد و ضرایب ویولت حاصله، خروجی هر کدام از فیلترها می‌باشند که کاهش نرخ نمونه برداری نیز برروی آنها اعمال شده است و به کمک قضیه پارسوال می‌توان انرژی خروجی هر فیلتر را حساب کرد. در این حالت پس از انتخاب یک فریم پنجره‌گذاری شده از سیگنال با توجه به اینکه در رزولوشن‌های به اندازه کافی بالا نمونه‌های سیگنال خیلی به ضرایب مقیاس نزدیک هستند، می‌توان فرض کرد که نمونه‌های فریم پنجره‌گذاری‌شده، ضرایب مقیاس در مقیاس بالا می‌باشند و از روی آنها ضرایب مقیاس و ویولت درمقیاس پایین‌تر را حساب کرد[۲۱]. معمولاً در سیگنالهای گفتار، پس از ۶ مرحله تقسیم فرکانسی و بدست آوردن ضرایب مقیاس در ۶

مـرحله، ضـرایب مقـیاس مـرحله بعـدی آنقدر کوچک می‌شوند که قابل چشم پوشی می‌باشند. لذا شش باند فرکانسی خواهیم داشـت کـه بـه کمـک قضـیه پارسـوال انـرژی این شش باند حساب شده و لگاریتم آنها به عنوان پارامترهای DWTBF[18] مورد استفاده قرار می‌گیرد.

روش دیگـر جهـت استفادهٔ بهینه از انرژی فیلترهای مختلف، استفاده از تجزیه بسته ویولت(WPD)[19] بجای DWT است. WPD مانـندDWT اسـت بـا ایـن تفـاوت کـه در انـتخاب باندهـای فرکانسی تنوع وجود دارد. با استفاده از این خاصیت، بانک فیلـترهای معـادل مقـیاس بارک یا مل که الهام گرفته از شنوایی انسان می‌باشند، توسط تبدیل فوریه قابل پیاده‌سازی هستنـد [۲۴]،[۲۶]،[۲۹]. الگوریتم استخراج پارامترهای بازنمایی پارامترهای MFWP[20] را می‌توان بصورت زیر خلاصه کرد:

گام اول ـ انتخاب یک فریم از سیگنال به طولN نمونه

گام دوم ـ حذف مقدار dc

گام سوم ـ ضرب فریم در پنجره زمانی همینگ

گام چهارم ـ اعمال WPD بر فریم پنجره‌گذاری شده

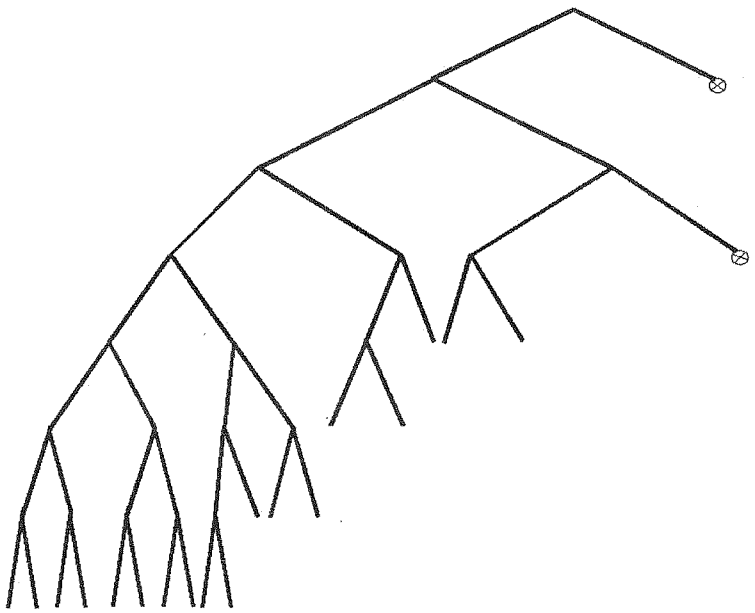
گام پنجم ـ انتخاب باندهای فرکانسی مشابه بانکهای فیلتر مل، مطابق شکل۲

گام ششم ـ محاسبه انرژی داخل هر باند فرکانس با استفاده از قضیه پارسوال

گام هفتم ـ اعمال تبدیل کسینوس گسسته بر لگاریتم انرژی داخل هر باند فرکانس

گام هشتم ـ انتخاب پانزده ضریب اول کپستروم به عنوان پارامترهای بازنمایی

گام نهم ـ اعمال تابع وزن بر ضرایب کپستروم



شکل (۲) درخت انتخابی در WPD جهت استخراج پارامترهای MFWP.

جدول(۱) دسته بندی ۲۹ آوایی زبان فارسی[21].

واکه‌ها	/@/ , /a/ , /e/ , /o/ , /u/ , /i/
شبه واکه‌ها	/y/ , /l/ , /m/ , /n/ , /r/
انفجاری‌ها	/b/ , /d/ , /q/ , /g/ , /?/ , /p/ , /t/ , /k/
سایشی‌ها	/f/ , /v/ , /s/ , /z/ , /$/ , /*/ , /h/ , /x/
سایشی ـ انفجاری‌ها	/j/ , /#/

در ادامـه و در بخـش ۵، نـتایج حاصل از ارزیابی عملکرد پارامترهای بازنمایی برای بازشناخت مستقل از گوینده گفتار مورد

بررسی و مقایسه قرار می‌گیرند.

٤ـ نتایج آزمایشات

در این بخش به نتایج بدست آمده از پارامترهای رایج بازنمایی که در بخش ۴ به معرفی آنها پرداخته شد و نیز پارامترهای بازنمایی که از بهبود آنها و اعمال هنجارسازی های مختلف حاصل‌گردید می‌پردازیم. همانطور که در بخش ۲ اشاره شد از دو جمله ۴۰۰و ۵۰۰ دادگان فارس‌دات به عنوان دادگان آزمایش استفاده شده است. این دو جمله تقریباً ۲۹ آوای فارسی را دربردارند و عملکرد شبکه هم برروی کل آواها و هم بطور جداگانه برای دسته آواهای مختلف طبق جدول۱ بررسی می‌شود.

در اولین آزمایشات پارامترهای LHCB با هنجارسازی‌های مختلف مورد بررسی قرار گرفت تا بهترین نوع هنجارسازی مشخص گردد. همانطور که در جدول۲ مشاهده می‌شود در مورد هنجارسازی عرضی پارامترهای LHCB یکبار نیز از حضور میانگین مولفه‌های یک بردار به عنوان یک پارامتر جدا (پارامتر نوزدهم) استفاده شده است. این پارامتر که تقریبی از انرژی می‌باشد باعث بهبود عملکرد بازشناسی شده است. هنجارسازی‌های طولی به خصوص هنجارسازی طولی نرم دوم با واریانس ۰/۵برای پارامترهای به هنجار شده نسبت به پارامترهای عرضی از عملکرد بهتری برخوردار هستند. این برتری بطور محسوس در واکه‌ها، شبه‌واکه‌ها و سایشی‌ها مشاهده می‌گردد. نتایج هنجارسازی طولی پارامترها بطور جداگانه برای هر گوینده نشان می‌دهد که هنجارسازی طولی پارامترها بطور جداگانه برای هر فرد نسبت به روش هنجارسازی طولی یکجا برای کل افراد عملکرد بهتری دارد. چرا که در این حالت اطلاعات مربوط به هرگوینده از پارامترهای بازنمایی همان گوینده تا حدی حذف

جدول (۲) مقایسه کارآیی روشهای مختلف هنجارسازی در بهبود کیفیت بازشناسی آواها که برروی پارامترهای بازنمایی LHCB اعمال گردیده‌اند.

	درصد صحت بازشناخت فریم روی دادگان آزمون						
	نوع بازشناسی	کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	هنجارسازی عرضی پارامترها + LHCB	۷۹/۶۲	۸۳/۴۲	۶۷/۷۹	۶۴/۹۳	۸۲/۳۸	۶۰/۴۵
۲	پارامترنوزدهم + LHCB + هنجارسازی عرضی پارامترها	۸۰/۵۴	۸۴/۳۶	۶۸/۵۱	۶۷/۱۹	۸۳/۰۷	۶۲/۷۱
۳	LHCB + هنجارسازی طولی نرم۱ پارامترها	۸۱/۰۶	۸۴/۱۹	۷۰/۶۱	۶۴/۵۲	۸۵/۲۸	۵۴/۸۰
۴	LHCB + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۱)	۸۲/۱۱	۸۵/۵۸	۷۱/۶۷	۶۶/۹۹	۸۳/۸۴	۶۵/۸۲
۵	LHCB + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵)	۸۲/۷۶	۸۶/۶۰	۷۱/۳۳	۶۷/۷۴	۸۵/۰۳	۵۸/۴۷
۶	هنجارسازی طولی پارامترها + LHCB برای هر فرد جدا	۸۳/۹۱	۸۸/۴۹	۷۲/۳۸	۷۰/۳۵	۸۴/۰۴	۷۳/۷۳

می‌شود ولی در دو روش هنجارسازی طولی نرم اول و نرم دوم پارامترهای بازنمایی، اطلاعاتی که به عنوان اطلاعات گوینده از پارامترهای بازنمایی کم می‌شود میانگینی از اطلاعات تمام گویندگان دادگان آزمایش می‌باشد.

انتقال انرژی بانک فیلترها به حوزه کپستروم و محاسبه ضرایب کپستروم، یکی دیگر از مراحل پردازش سیگنال گفتار است که در برخی بازنمایی‌ها ازجمله MFCC، LHCBC و PLP مورد استفاده قرار می‌گیرد. همانطور که اشاره شد در پارامترهای MFCC از مقیاس مل و در پارامترهای LHCBC از مقیاس بارک استفاده می‌شود و تفاوت دیگری که در روش استخراج این دو نوع پارامتر بازنمایی وجود دارد شکل بانک فیلتر می‌باشد. در MFCC از فیلترهای مثلثی و در LHCBC مانند پارامترهای LHCB از فیلترهای مجذور هنینگ استفاده می‌شود. برای اینکه بتوان یک مقایسه بین اثر مقیاس‌های مل و بارک انجام داد

یکبار نیز پارامترهای MFCC را با استفاده از بانک‌های فیلتر مجذور هنینگ استخراج می‌نماییم. در تعلیم شبکه عصبی با استفاده از هرکدام از سه پارامتر فوق الذکر، شبکه عصبی در کمینه موضعی[۲۲] افتاده بطور کامل تعلیم نمی‌بیند و صحت بازشناسی فریم برای کل آواها در حدود ۷۳٪ بدست می‌آید. شبکه عصبی به دلایل مختلفی در کمینه موضعی قرارمی‌گیرد که یکی از آنها می تواند تغییرات زیاد دامنه برخی از مؤلفه‌های بردار بازنمایی باشد. با اعمال هنجارسازی طولی نرم دوم پارامترهای بازنمایی مشکل تعلیم شبکه عصبی برای هر سه نوع پارامتربازنمایی فوق حل می‌شود. در این حالت میزان صحت بازشناسی فریم‌ها روی کل آواها برای پارامترهای به هنجار شده MFCC با بانک فیلترهای مثلثی و مجذور هنینگ به ترتیب برابر ۸۰/۲۶٪ و ۸۰/۹۱٪ بدست آمد. همین صحت بازشناسی برای پارامترهای LHCBC برابر ۸۱/۴۰٪ حاصل شد. این مقادیر نشان می‌دهند که عملکرد پارامترهای LHCBC از پارامترهای MFCC اندکی بهتر می‌باشد. علت این مطلب را تا حدی می توان در عملکرد بهتر نوع پنجره‌های طیفی انتخابی در LHCBC یعنی پنجره‌های مجذور هنینگ دانست. همچنین اثر مقیاس‌های مل و بارک تقریباً شبیه به هم می‌باشد. در جدول۳ صحت بازشناخت ناشی از این پارامترها نشان داده شده است. با مقایسه این جدول و جدول۲ مشاهده می‌گردد که پارامترهای LHCB با هنجارسازی طولی نرم دوم نسبت به پارامترهای کپستروم MFCC و LHCBC و با همین هنجارسازی عملکرد بهتری برای شبکه عصبی ایجاد کرده است.

پارامترهای طیفی دیگری که براساس تبدیل فوریه زمان کوتاه بدست می‌آیند پارامترهای PLP می‌باشند. پارامترهای PLP یکی دیگر از پرکاربردترین پارامترهای بازنمایی در سیستم‌های بازشناخت گفتار می‌باشد. هنگامیکه درکنار پارامترهای PLP، تقریبی از انرژی فریم نیز به عنوان پارامتری جدا استفاده می‌شود، صحت بازشناسی فریم روی کل آواها ۳٪ رشد کرده و از مقدار ۷۸/۴۰٪ به ۸۱/۴۶٪ می‌رسد. در جدول۴ صحت بازشناخت فریم روی دسته آواهای مختلف ناشی از پارامترهای PLP در مقایسه با پارامترهای LHCB با هنجارسازی های عرضی و طولی نرم دوم آورده شده‌است. مشاهده می‌شود که پارامترهای PLP با احتساب پارامتری که تقریبی از انرژی فریم می‌باشد، عملکرد مناسبی برای شبکه عصبی مهیا ساخته است.

جدول(۳) نتایج حاصل از پارامترهای LHCBC و MFCC.

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	MFCC + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵) ، فیلتر مثلثی	۸۰/۲۶	۸۵/۰۶	۶۵/۴۹	۶۲/۳۵	۸۲/۵۳	۶۴/۱۲
۲	MFCC + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵) ، فیلتر مجذور هنینگ	۸۰/۹۱	۸۵/۴۶	۶۶/۸۸	۶۴/۳۴	۸۳/۰۷	۵۷/۰۶
۳	LHCBC + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵)	۸۱/۴۰	۸۵/۵۴	۶۸/۷۸	۶۳/۸۳	۸۴/۵۱	۶۰/۱۷

جدول(٤) مقایسه نتایج حاصل از پارامترهای PLP و LHCB.

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	پارامتر انرژی + PLP	۸۱/۴۶	۸۲/۹۶	۷۲/۴۷	۶۸/۰۲	۸۶/۲۳	۷۰/۹۰
۲	PLP	۷۸/۴۰	۸۰/۸۴	۷۰/۲۳	۵۸/۹۶	۸۲/۵۵	۶۲/۷۱
۳	پارامترنوزدهم + LHCB + هنجارسازی عرضی پارامترها	۸۰/۵۴	۸۴/۳۶	۶۸/۵۱	۶۷/۱۹	۸۳/۰۷	۶۲/۷۱
۴	LHCB + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵)	۸۲/۷۶	۸۶/۶۰	۷۱/۳۳	۶۷/۷۴	۸۵/۰۳	۵۸/۴۷

در بررسی پارامترهای طیفی مبتنی بر مدل تمام قطب و پیشگویی خطی باید به این نکته دقت داشت که مرتبه مدل متناسب با فرکانس نمونه برداری سیگنال گفتار باید انتخاب گردد. لذا چون دادگان فارس‌دات با فرکانس ۴۴/۱ کیلوهرتز نمونه‌برداری شده‌است و مدل تمام قطب مرتبه بالا تخمین مناسبی از طیف بدست نمی‌دهد، در ابتدا فرکانس نمونه برداری به ۱۶ کیلوهرتز کاهش داده می‌شود. مرتبه مدل تمام قطب برابر ۱۴ و یکبار نیز برابر ۱۶ انتخاب شده است و برای اینکه تغییری در شرایط آزمایش بوجود نیاید، برای این روش‌ها نیز مانند روش‌های قبلی از فیلتر پیش تأکید استفاده نشده است. البته برای اینکه اثر این فیلتر بر روی عملکرد پارامترهای بازنمایی در بازشناخت گفتار مشخص گردد، یکبار نیز برای استخراج پارامترهای LPCC سیگنال گفتار را از یک فیلتر پیش تأکید $P(z)=1-0.97z^{-1}$ عبور داده‌ایم. در این روش‌ها پس از حصول ضرایب مدل تمام قطب (ضرایب پیشگویی خطی) به کمک الگوریتم لوینسن ـ دربین، یا بطور مستقیم این ضرایب به ضرایب کپستروم تبدیل می‌شوند (LPCC) و یا از روی مدل تمام قطب طیف توان تخمین زده شده و با استفاده از بانک فیلتر انرژی باندهای فرکانسی مختلف بدست آمده و به حوزه کپستروم انتقال می‌یابند (Bark-LPCC). همانطورکه در جدول۵ مشاهده می‌شود بجز در سایشی ـ انفجاری‌ها در بقیه موارد وجود فیلتر پیش تأکید باعث بهبود عملکرد پارامترهای LPCC شده است. نتایج جدول۵ نشان می‌دهد که پارامترهای LPCC عملکرد مناسبی درجهت بازشناخت گفتار نداشته است. برای بهبود این پارامترها ازبهره مدل تمام قطب به عنوان تقریبی از انرژی فریم درکنار دیگر پارامترهای LPCC استفاده شد. بهره مدل تمام قطب برای واکدارها با پریود گام متناسب است. نتایج مندرج در جدول۵ مؤید این نکته است که اضافه شدن بهره مدل تمام قطب باعث بهبود عملکرد پارامترهای LPCC شده است.

الگوریتم استخراج پارامترهای Bark-LPCC فقط در نوع تحلیل طیفی با روش استخراج پارامترهای LHCBC متفاوت است و همانطور که در جدول۶ نشان داده شده است، نتایج نسبتاً ضعیفی از این پارامترها نسبت به پارامترهای LHCBC حاصل شده است. این نتایج بیان می‌کند که مدل تمام قطب استفاده شده نمی‌تواند به خوبی تبدیل فوریه زمان کوتاه طیف توان سیگنال را تقریب بزند. صحت بازشناخت فریم روی کل آواها برای پارامترهای Bark-LPCC با هنجارسازی طولی نرم دوم پارامترها برابر ۷۶/۲۰٪ می‌باشد حال آنکه همین صحت بازشناخت برای پارامترهای LHCBC با همین هنجارسازی برابر ۸۱/۴۰٪ بوده است.

جدول (۵) نتایج حاصل از پارامترهای LPCC.

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	LPCC (مرتبه مدل= ۱۴)	۷۵/۶۶	۷۸/۸۶	۶۱/۶۴	۶۱/۶۰	۷۶/۵۹	۶۱/۸۶
۲	LPCC + فیلتر پیش تأکید (مرتبه مدل= ۱۴)	۷۷/۰۰	۸۰/۸۵	۶۵/۸۲	۶۳/۸۰	۷۹/۰۵	۴۷/۷۴
۳	LPCC + بهره مدل تمام قطب (مرتبه مدل= ۱۴)	۷۷/۷۱	۷۹/۷۳	۶۸/۱۵	۶۰/۷۴	۸۳/۳۵	۵۳/۱۱
۴	LPCC (مرتبه مدل= ۱۶)	۷۵/۵۹	۷۷/۷۹	۶۲/۵۶	۶۰/۲۳	۸۱/۳۶	۴۲/۶۶

جدول (٦) نتایج حاصل از پارامترهای Bark-LPCC در مقایسه با LHCBC.

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	Bark-LPCC + بهره مدل تمام قطب + هنجارسازی طولی نرم ۲ پارامترها	۷۶/۲۰	۸۰/۸۰	۶۵/۰۶	۵۶/۲۵	۷۴/۴۱	۴۷/۷۴
۲	LHCBC + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵)	۸۱/۴۰	۸۵/۵۴	۶۸/۷۸	۶۳/۸۳	۸۴/۵۱	۶۰/۱۷

پارامترهای مبتنی بر تبدیل گسسته ویولت[23](DWTBF) و همچنین پارامترهای بازنمایی به کمک تجزیه بسته ویولت و بانک‌های مشابه مقیاس مل[24](MFWP) پارامترهایی هستند که در سال‌های اخیر برای بازشناخت گفتار مورد توجه قرار گرفته‌اند. موارد استفاده این نوع پارامترها بیشتر در بیان ویژگی‌ها و طبقه‌بندی گروهی خاصی از آواها مانند انفجاری‌ها بوده است ویا برای بازشناسی‌های محدود مانند بازشناسی ۹ رقم از یکدیگر و یا بازشناسی آواهای منقطع صورت گرفته است. در این مقاله عملکرد این پارامترها را برای بازشناسی گفتار پیوسته بررسی کرده‌ایم و با اصلاح الگوریتم استخراج این پارامترها باعث بهبود عملکرد آن شده‌ایم.

در تجزیه بسته ویولت برخلاف تبدیل ویولت گسسته انتخاب باندهای فرکانسی در اختیار ما می‌باشد و معمولاً باندهایی انتخاب می‌شوند که موجب تقسیم‌بندی محور فرکانسی مشابه مقیاس مل یا بارک گردند. در روش MFWP که در این مقاله مورد بررسی قرارگرفته است، تجزیه بسته ویولت تا هفت مرحله انجام می‌گیرد و سپس باندهای فرکانسی مشابه مقیاس مل انتخاب می‌گردد و سپس مانند روش استخراج پارامترهای MFCC، انرژی هر باند محاسبه شده و به ضرایب کپسترم تبدیل می‌گردند. صحت بازشناخت حاصل از پارامترهای MFWP با پارامترهای MFCC در جدول۷ مقایسه شده است. نتایج نشان می‌دهد که عملکرد پارامترهای MFWP بسیار ضعیفتر می‌باشد. در پارامترهایDWTBF نیز تبدیل ویولت تا مقیاس شش انجام گرفته و سپس لگاریتم انرژی هر باند فرکانسی به عنوان پارامتر بازنمایی مورد استفاده قرار گرفته است. نتایج حاصل از این پارامترها نیز چندان قابل توجه نیست. در روش اصلاح شده که در این مقاله برای استخراج پارامترهای بازنمایی به کمک تبدیل ویولت گسسته پیشنهاد شده است، بجای آنکه تبدیل ویولت گسسته را روی هر فریم اعمال کنیم، از کل سیگنال گفتار تبدیل ویولت گسسته می‌گیریم و سپس ضرایب ویولت متناسب با هر فریم از سیگنال را جدا می‌کنیم. در واقع فریم‌بندی را روی ضرایب ویولت انجام می‌دهیم نه روی سیگنال گفتار. سپس لگاریتم انرژی هر باند فرکانسی را به عنوان پارامترهای بازنمایی مورد استفاده قرار می‌دهیم. نتایج حاصل از آزمایشات نشان می‌دهد که این پارامتر های بازنمایی اصلاح شده (MDWTBF[25]) نسبت به پارامترهای DWTBF عملکرد بهتری دارند و صحت بازشناخت فریم روی کل آواها از ۷۳/۷۹٪ به ۷۷/۸۶٪ رسیده است. در جدول۷ نتایج حاصل از این دو نوع پارامتر نیز مورد مقایسه قرارگرفته‌اند. علت این برتری رفع اثر

جدول(۷) نتایج حاصل از پارامترهای DWTBF و MFWP

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	هنجارسازی طولی نرم۲ + DWTBF (واریانس = ۰/۵)	۷۳/۷۹	۷۹/۲۹	۵۹/۵۴	۵۵/۶۸	۷۴/۸۷	۵۷/۹۴
۲	هنجارسازی طولی نرم۲ + MDWTBF (واریانس = ۰/۵)	۷۷/۸۶	۷۹/۶۸	۶۵/۹۱	۶۱/۸۷	۸۲/۸۵	۷۱/۴۷
۳	هنجارسازی طولی نرم۲ + MFWP (واریانس = ۰/۵)	۷۱/۵۰	۷۶/۳۸	۵۵/۱۰	۴۸/۵۹	۷۱/۵۳	۵۶/۵۰
۴	MFCC + هنجارسازی طولی نرم۲ پارامترها (واریانس = ۰/۵) ، فیلتر مجذور هنینگ	۸۰/۹۱	۸۵/۴۶	۶۶/۸۸	۶۴/۳۴	۸۳/۰۷	۵۷/۰۶

جدول(۸) نتایج حاصل از ترکیب پارامترهای LHCB با DWTBF و MDWTBF.

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	هنجارسازی طولی پارامترها + LHCB	۸۱/۰۶	۸۴/۱۹	۷۰/۶۱	۶۴/۵۲	۸۵/۲۸	۵۴/۸۰
۲	LHCB_DWTBF هنجارسازی طولی پارامترها +	۸۲/۵۲	۸۶/۴۴	۷۱/۶۹	۶۶/۵۴	۸۴/۱۹	۶۷/۲۳
۳	LHCB_MDWTBF هنجارسازی طولی پارامترها +	۸۳/۲۱	۸۵/۷۹	۷۱/۶۹	۷۰/۳۵	۸۷/۹۶	۷۴/۰۱

مخرب پنجره‌گذاری بر طیف توان می‌باشد.

با ترکیب پارامترهای LHCB و MDWTBF یا DWTBF و در کنارهم قراردادن آنها در یک بردار بازنمایی جدید، صحت بازشناخت نسبت به پارامترهای LHCB به تنهایی، رشد نشان داده است و صحت بازشناخت فریم روی کل آواها برای پارامترهای LHCB با هنجارسازی طولی نرم اول که برابر با ۸۱/۰۶٪ می‌باشد در ترکیب با پارامترهای MDWTBF و DWTBF به ترتیب به مقادیر ۸۳/۲۱٪ و ۸۲/۵۲٪ رسیده است. پارامترهای بازنمایی مبتنی بر تبدیل ویولت دارای اطلاعات مفیدی از تغییرات سریع سیگنال گفتار می‌باشند و ترکیب این اطلاعات با اطلاعات پارامترهای LHCB باعث می‌شود تا شبکه عصبی بهتر بتواند آواهای مختلف را طبقه‌بندی کند. نتایج حاصل از این پارامترها برای دسته آواهای مختلف در جدول۸ مورد مقایسه قرارگرفته‌اند.

پارامترهای دیگری که در این مقاله مورد بررسی قرارگرفتند ادغام برخی از پارامترها با ضرایب دلتای مربوطه می‌باشد. درسیستم‌های بازشناخت گفتار که از مدل‌های استاتیک بازشناخت گفتار استفاده می‌کنند، از ضرایب دلتا جهت استفاده از اطلاعات دینامیکی سیگنال گفتار استفاده می‌شود. اما شبکه عصبی جلوسو با تأخیر زمانی به علت اینکه در ورودی به ۱۴ فریم متوالی نگاه می‌کند، مدل دینامیک‌های ورودی را می‌آموزد و اطلاعات دینامیکی گسترده‌تری نسبت به ضرایب دلتا از سیگنال گفتار استخراج می‌کند. به عبارت دیگر ضرایب دلتا ازنظر اطلاعات دینامیکی نمی‌تواند اطلاعات جدیدی در اختیار شبکه عصبی قرار دهد. اما چون اطلاعات گوینده در فریم‌های نزدیک به هم به آرامی تغییر می‌کند، در ضرایب دلتا که پارامترهای دو فریم نزدیک به هم از یکدیگر کم می‌شوند، تاحدودی اطلاعات گوینده حذف می‌گردد. و لذا با اضافه شدن

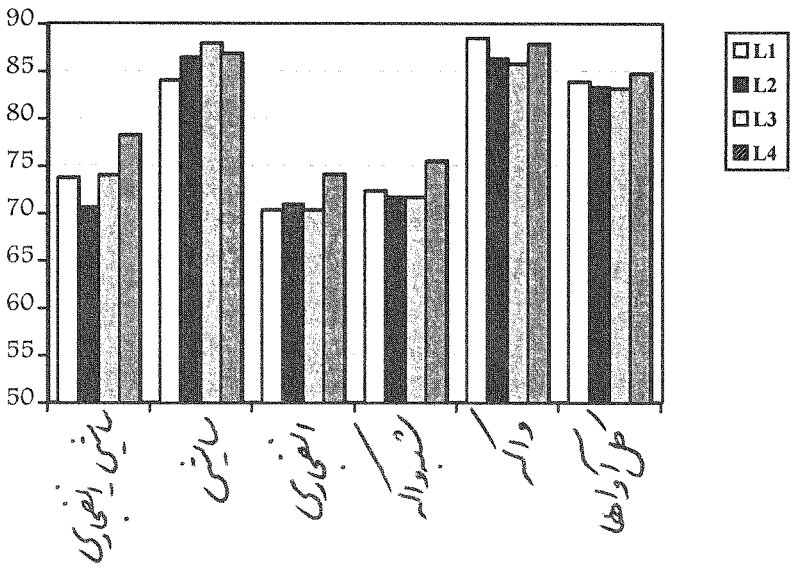
جدول(۹) نتایج حاصل از اضافه کردن ضرایب دلتا.

	نوع بازشناسی	درصد صحت بازشناخت فریم روی دادگان آزمون					
		کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
۱	هنجارسازی عرضی پارامترها + LHCB + Δ	۸۲/۹۸	۸۵/۴۹	۷۲/۳۴	۶۹/۶۶	۸۶/۷۵	۷۴/۰۱
۲	هنجارسازی عرضی پارامترها + LHCB	۸۰/۵۴	۸۴/۳۶	۶۸/۵۱	۶۷/۱۹	۸۳/۰۷	۶۲/۷۱
۳	پارامتر تقریب انرژی فریم + PLP + Δ	۸۳/۳۴	۸۶/۳۴	۷۱/۶۹	۷۰/۹۷	۸۶/۴۷	۷۰/۶۲
۴	پارامتر تقریب انرژی فریم + PLP	۸۱/۴۵	۸۲/۹۶	۷۲/۴۷	۶۸/۰۲	۸۶/۲۳	۷۰/۹۰
۵	DWTBF + Δ	۷۴/۹۱	۷۸/۳۴	۶۰/۰۵	۵۷/۴۱	۷۹/۸۵	۵۹/۰۴
۶	DWTBF	۷۲/۳۱	۷۷/۱۰	۵۶/۹۶	۵۱/۳۷	۷۹/۵۰	۴۴/۰۷
۷	MDWTBF + Δ	۷۸/۵۵	۸۰/۹۸	۶۸/۲۴	۶۲/۲۵	۸۲/۸۰	۶۱/۳۰
۸	MDWTBF	۷۷/۸۶	۷۹/۶۸	۶۵/۹۱	۶۱/۸۷	۸۲/۸۵	۷۱/۴۷
۹	LHCB_MDWTBF + Δ + هنجارسازی طولی نرم ۲ (واریانس = ۰/۵)	۸۴/۷۹	۸۷/۸۹	۷۸/۵۲	۷۴/۱۲	۸۶/۸۷	۷۸/۲۵

ضرایب دلتا به پارامترهای بازنمایی، شبکه عصبی قادر خواهد بود که با دقت بیشتری آواها را طبقه‌بندی کند. آزمایش بر روی چهار نوع پارامتر بازنمایی LHCB، PLP، DWTBF و MDWTBF نشان داد که اضافه شدن ضرایب دلتا به پارامترهای بازنمایی حدود ۲٪ صحت بازشناسی فریم روی کل آواها را افزایش می‌دهد که نتایج آزمایشات مربوطه در جدول۹ درج شده‌است. همانطور که در ردیف آخر این جدول دیده می‌شود در کنار هم قراردادن پارامترهای LHCB و MDWTBF در یک بردار بازنمایی و استفاده از هنجارسازی طولی نرم۲ (واریانس = ۰/۵) و همچنین اضافه کردن ضرایب دلتا باعث رشد قابل توجه درصد صحت بازشناخت کل فریمها تا میزان ۸۴/۷۹٪ شده است.

در بین پارامترهای بازنمایی مختلف که در این مقاله مورد بررسی قرار گرفتند، پارامترهای LHCB با هنجارسازی طولی

پارامـترها بطـور جداگانـه بـرای هر گوینده با صحت بازشناخت فریم ۸۳/۹۱٪، پارامترهای PLP به همراه ضرایب دلتا با صحت بازشـناخت فـریم برابـر ۸۳/۳۴٪، پارامترهای بازنمایی تشکیل شده از پـارامترهای LHCB و MDWTBF با هنجارسازی طولی نـرم اول بـا صـحت بازشـناخت فریم برابر ۸۳/۲۱٪ و پارامترهای بازنمایی تشکیل شده از پـارامترهای LHCB و MDWTBF با هنجارسـازی طولـی نـرم دوم بـه همراه ضرایب دلتا با صحت بازشناخت فریم برابر ۸۴/۷۹٪ باعث بهترین عملکرد شبکه عصبی جلوسو با تأخیر زمانی در بازشناخت فریم شده‌اند. نتایج حاصل از این چهار بازنمایی برتر در شکل۳ درج شده است.



درصد صحت بازشناخت فریم روی دادگان آزمون							
	نوع بازشناسی	کل	واکه	شبه‌واکه	انفجاری	سایشی	سایشی- انفجاری
L1	هنجارسازی طولی پارامترها + LHCB جداگانه برای هر گوینده	۸۳/۹۱	۸۸/۴۹	۷۲/۳۸	۷۰/۳۵	۸۴/۰۴	۷۳/۷۳
L2	پارامتر تقریب انرژی فریم + Δ + PLP	۸۳/۳۴	۸۶/۳۴	۷۱/۶۹	۷۰/۹۷	۸۶/۴۷	۷۰/۶۲
L3	LHCB_MDWTBF هنجارسازی طولی نرم ۱ پارامترها +	۸۳/۲۱	۸۵/۷۹	۷۱/۶۹	۷۰/۳۵	۸۷/۹۶	۷۴/۰۱
L4	LHCB_MDWTBF + Δ + هنجارسازی طولی نرم ۲ (واریانس = ۰/۵)	۸۴/۷۹	۸۷/۸۹	۷۸/۵۲	۷۴/۱۲	۸۶/۸۷	۷۸/۲۵

شکل(۳) نتایج حاصل از چهار بازنمایی برتر.

۵ـ جمع‌بندی و نتیجه‌گیری

در ایـن مقاله روش‌های رایج استخراج پارامترهای بازنمایی مورد بررسی ومقایسه قرارگرفت و با اعمال تغییراتی در الگوریتم برخـی از آنهـا و بهـبود پارامـترهای بازنمایـی و همچنین با استفاده از روش‌های هنجارسازی متفاوت، مناسب‌ترین پارامترهای بازنمایـی در بکارگـیری بـرای بازشناسـی مستقل ازگوینده گفتار عاری از نویز با استفاده از شبکه عصبی جلوسو با تأخیر زمانی معرفـی گـردید. در بیـن پارامـترهای مـورد بررسـی، پارامـترهای LHCB باهنجارسازی طولی پارامترها بطور جداگانه برای هر گویـنده بـا صـحت بازشناخت فریم ۸۳/۹۱٪ ، پارامترهای PLP به همراه ضرایب دلتا با صحت بازشناخت فریم برابر ۸۳/۳۴٪ و پارامـترهای بازنمایـی تشکیل شده از پارامترهای LHCB و MDWTBF با هنجارسازی طولی نرم اول با صحت بازشناخت فریم برابر ۸۳/۲۱٪ باعث بهترین عملکرد شبکه عصبی جلوسو با تأخیر زمانی در بازشناخت فریم‌ها شده‌اند.

در آزمایشات انجام شده مشاهده شد که استفاده از تقریبی از انرژی فریم به عنوان یک پارامتر جدا و در کنار سایر پارامترها

باعث بهبود عملکرد شبکه جلوسوی زمانی در طبقه‌بندی آواها می‌گردد. علت این رشد به این خاطر است که وقتی تقریبی از انرژی فریم بطور مجزا توسط شبکه عصبی مورد توجه قرار می‌گیرد به عنوان یک عامل مهم برای تفکیک فریم‌های کم انرژی و پر انرژی عمل می‌کند.

نتیجه دیگر اثر مثبت فیلتر پیش تأکید در عملکرد پارامترهای بازنمایی می‌باشد. آزمایش انجام شده بر روی پارامترهای LPCC مؤید این نکته است که وجود فیلتر پیش تأکید رشدی حدود ۱/۵٪ در بازشناخت فریم روی کل آواها دارد.

از بین هنجارسازی های بررسی شده هنجارسازی طولی پارامترها با پارامترهای به هنجار کننده جداگانه برای هر گوینده و نیز هنجارسازی طولی نرم دوم پارامترها با واریانس ۰/۵ برای پارامترهای به هنجار شده، بهترین عملکرد را برای شبکه عصبی به همراه داشته‌اند.

۶ـ قدردانی

این تحقیق با حمایت و همکاری پژوهشکده پردازش هوشمند علائم به انجام رسیده است.

زیرنویسها

1- Recurrent Neural Networks
2- Deng
3- Holton
4- Normal
5- Pitch Frequency
6- Mel Scale
7- Bark Scale
8- Logarithm of square Hanning Critical Band filter banks
9- Logarithm of square Hanning Critical Band filter banks Cepstral
10- Mel Frequency Cepstral Coefficients
11- Perceptual Linear Prediction
12- Equal Loudness
13- Intensity-Loudness
14 -Linear Prediction Coefficients
15- Linear Prediction Coefficients Cepstral
16- Discrete Wavelet Transform
17- Muti Resolusion
18- Discrete Wavelet Transform Based Feature
19 -Wavelet Packet Decompose
20- Mel Frequency Wavelet Packet

۲۱- برخی از این علایم که استاندارد نمی‌باشند عبارتند از:

/*/ برای صدای "ژ"، /#/ برای صدای "چ"، /$/ برای صدای "ش"، /x/ برای صدای "خ"، /?/ برای پست چاکنای(حمزه)، /q/ برای صدای "ق"، /@/ برای صدای "آ"

22 -Local minimum
23- Disceret Wavelet Transform Based Features
24- Mel Frequency Wavelet Packet
25-Modified DWTBF

مراجع

[۱] س. ع. سیدصالحی، هنجارسازی خطی و غیرخطی پارامترهای بازنمایی در بازشناسی گفتار.، دردست تدوین، ۱۳۸۲

[۲] س.ع. سیدصالحی، بازشناخت گفتار پیوسته فارسی با استفاده از مدل عملکردی مغز انسان در درک گفتار.، پایان‌نامه دکتری، دانشگاه تربیت مدرس، دانشکده فنی و مهندسی، بهمن ۱۳۷۴

[3] T. Robinson, J. Holdesworth, R. Patterson, F. Fallside, A Comparison of Preprocessors for The Cambridge Recurrent Error Propagation Network Speech Recognition System., Proceedings of ICSLP, pp. 1033-1036, (1990)
[4] L.I. Deng, Processing of Aucostic Signals in a Cochlear Model Incorporating Laterally Coupled Suppressive Elements., Journal of Neural Networks, Vol. 5, (1992)
[5] T. Holton, S.D. Love, S.P. Gill, Formant and Pitch-Pulse Detection Using Models of Auditory Signal Processing., Proceeding of ICSLP'92, pp. 81-84, (1992)
[6] B. Chigier, H.C. Leung, The Effects of Signal Representations, Phonetic Classification Techniques, and The Telephone Network., Proceedings of ICSLP (1992)
[7] R. Chengalvarayan, Robust Energy Normalization Using Speech/Nonspeech Discriminant for German Connected Digital Recognition., Proceedings of Eurospeech'99, Budapest, Hungary, (1999)
[8] M. Naito, L. Deng, Y. Sagisaka, Model-Based Speaker Normalization Methods for Speech Recognition., Proceedings of Eurospeech'99, Budapest, Hungary, (1999)
[9] L. Welling, R. Heab-Umbach, X. Aubert, N. Haberland, A Study on Speaker Normalization Using Vocal Tract Normalization and Speaker Adaptive Training., Proceedings of ICASSP, pp. 797-800, (1998)
[10] C. Furlanello, D. Ggiuliani, E. Trentin, D. Falavigna, Application of Generalized Radial Basis Functions in Speaker Normalization and Identification., IEEE International Symposiom on Circuits and Systems, Vol. 3, pp. 1704-1707, (1995)
[11] S.A. Seyyed Salehi, A Neural Network Speech Recognition Based on the Both Acoustic Steady Portions and Transitions., Proceedings of ICASP, Beijing (2000)
[12] D.E. Olsan, J.C. Harris, M.H. Capps, Computerized Polygraph Scoring System., Journal of Forensic Sciences (1996)
[13] M. Bijankhan and et. al., FARSDAT – The Speech Database of Farsi Spoken Language., Proc. SST-94, Perth, pp. 826-831, (1994)
[14] C. Becchetti, L.P. Ricotti, Speech Recognition, John Wiley & Sons., LTD, (2000)
[15] J.H.L. Hansen, S. Bou-Ghazale, A Computer Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress., IEEE Transaction on Speech and Audio Processing. Vol. 8, No. 4, pp. 429-442, July (2000)
[16] H. Hermansky, Perceptual Linear Predictive (PLP) Analysis of Speech., Journal of Acoustic Society of America, Vol. 87, No. 4, pp. 1738-1752, April (1990)
[17] J.C. Junqua, H. Wakita, H. Hermansky, Evaluation and Optimization of Perceptually Based ASR Front-End., IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 1, pp. 39-48, January (1993)
[18] J.R. Deller, J.G. Proakis, J.H.L. Hansen, Discrete-Time Processing of Speech Signals., Mcmillan Publishing Company, (1993)
[19] Q. Li, F.K. Soong, O. Siohan, An Auditory System-Based Feature for Robust Speech Recognition., Proceedings of Eurospeech (2001)
[20] H. Matsumoto, M. Moroto, Evaluation of MEL-LPC Cepstrum in a Large Vocabulary Continuous Speech Recognition., Proceedings of ICSLP (2000)
[21] C.S. Burrus, R.A. Gopinath, H. Guo, Introduction to Wavelets and Wavelet Transforms., Prentice Hall Inc., (1998)
[22] O. Farooq, S. Datta, Speaker Independent Phoneme Recognition by MLP Using Wavelet Features., Proceedings of ICSLP (2000)
[23] O. Farooq, S. Datta, Dynamic Feature Extraction by Wavelet Analysis., Proceedings of ICSLP (2000)
[24] O. Farooq, S. Datta, Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition., IEEE Signal Processing Letters, Vol. 8, No. 7, pp. 196-198, July (2001)
[25] J.N. Gowdy, Z. Tufekci, Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition., Proceedings of ICASSP (2000)
[26] E. Lukasik, Wavelet Packets Based Features Selection for Voiceless Plosives Classification., IEEE Transactions on Speech and Audio Processing, pp. 689-692, (1999)
[27] C.J. Long, S. Datta, Wavelet Based Feature Extracion for Phoneme Recognition., Proceedings of ICSLP (1996)
[28] P.P. Vaidanathan, Multirate Systems and Filter Banks, Prentice Hall Inc., (1993)
[29] I. Cohen, Enhancement of Speech Using Bark-Scaled Wavelet Packet Decomposition., Proceedings of Eurospeech (2001)
[30] B. Davis, P. Mermelstein, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences., IEEE Transactions on Audio and Speech and Signal Processing, pp. 357-366, July (1987)
[31] B. H. Juang, L. Rabiner, J. Wilpon, On the Use of Bandpass Lifterin in Speech Recognition., IEEE Transactions on Audio and Speech and Signal Processing, pp. 947-954, August (1980)