# روشی جدید در تشخیص گوینده مستقل از متن در محیط‌های نویزی

**نونا حیدری اصفهانی[۱] - حمید محمودیان[۲]**
(۱) کارشناس ارشد - شرکت پرشیان فولاد، اصفهان
(۲) استادیار - دانشکده برق، دانشگاه آزاد اسلامی، واحد نجف آباد

تاریخ دریافت: ۱۳۹۲/۵/۱ تاریخ پذیرش: ۱۳۹۳/۳/۱

**خلاصه:** در این مقاله بازشناسی مقاوم به نویز گوینده در حالت مستقل از متن مورد توجه قرار گرفته است. روش پیشنهادی بر مبنای حذف سکوت از جملات و تقطیع آنها به واحدهای کوچک‌تر شامل چند آوا و حداقل یک واکه برای استخراج ویژگی‌های زمان‌بلند از جمله آنتروپی عمل می‌کند. یک واکه پرانرژی در هر قطعه گفتاری برای استخراج فرکانس پایه و فرمنت‌ها شناسایی می‌شود. با اعمال یک روش خوشه‌بندی، ویژگی‌های زمان-کوتاه یعنی ضرایبِ MFCC با ویژگی‌های زمان‌بلند ترکیب می‌شوند. نتایج آزمایشات با استفاده از طبقه‌بندی کننده از نوع MLP نشان می‌دهد که میانگین نرخ بازشناسی گوینده با روش پیشنهادی در حالت بدون نویز ۹۷/۳۳٪ و در نسبت سیگنال به نویز ۲- دسی‌بل ۶۱/۳۳٪ است که نسبت به روش‌های متداول بهبود نشان می‌دهد.

**کلمات کلیدی:** بازشناسی گوینده، ضرایب MFCC، فرکانس پایه، فرمنت، آنتروپی شانون، MLP.

# A Novel Approach in Text-Independent Speaker Recognition in Noisy Environment

**Nona Heydari Esfahani[1] - Hamid Mahmoodian[2]**
(1) MSc – Persian Foolal Company, Isfahan
nh_2004_ee@yahoo.com
(2) Assistant Professor - Department of Electrical Engineering, Najafabad Branch, Islamic Azad University
h_mahmoodian@pel.iaun.ac.ir

**I**n this paper, robust text-independent speaker recognition is taken into consideration. The proposed method performs on manual silence-removed utterances that are segmented into smaller speech units containing few phones and at least one vowel. The segments are basic units for long-term feature extraction. Sub-band entropy is directly extracted in each segment. A robust vowel detection method is then applied on each segment to separate a high energy vowel that is used as unit for pitch frequency and formant extraction. By applying a clustering technique, extracted short-term features namely MFCC coefficients are combined with long term features. Experiments using MLP classifier show that the average speaker accuracy recognition rate is 97.33% for clean speech and 61.33% in noisy environment for -2db SNR, that shows improvement compared to other conventional methods.

**Index Terms:** Speaker identification, MFCC coefficients, pitch ferequency, formants, Shannon entropy, MLP

---

**نویسنده مسئول**: نونا حیدری اصفهانی، دانشگاه آزاد اسلامی واحد نجف‌آباد، nh_2004_ee@yahoo.com

## ۱- مقدمه

امروزه در کاربردهای مختلف نیاز گسترده‌ای به تشخیص هویت افراد به وجود آمده و صوت به دلیل ویژگی‌های خاص خود، کاربرد ویژه‌ای در تشخیص هویت یافته است. هدف از انجام این تحقیق تشخیص افراد از روی گفتار آنها در محیطِ نویزی می‌باشد که یکی از روش‌های بیومتریک در تشخیص هویت افراد است. با توجه با اینکه علت اصلی بروز خطا در سیستم‌های خودکار شناسایی گوینده، عدم تطابق بین الگوهای آموزشی و عملی است، دستیابی به روش‌های مقاوم در مقابل عوامل گوناگون شامل تغییرات محل میکروفون‌ها، ویژگی‌های کانال انتقال، نویزهای محیطی، گفتارهای همزمان و غیره، یکی از موضوعات کلیدی در مبحث شناسایی گوینده محسوب می‌شود.

تاکنون فعالیت‌های زیادی مبتنی بر ویژگی‌های ضرایب مل کپسترال فرکانسی[۱] در حوزه شناسایی گوینده انجام شده است که در شرایط بدون نویز کارآیی قابل قبولی دارند [۱]. با این حال چون برای استخراج ضرایب MFCC از یک بانک فیلتر مبتنی بر مقیاس مل استفاده می‌شود، خروجی‌های فیلترها در این بانک فیلتر به صورت غیریکنواخت تحت تأثیر نویز قرار می‌گیرند. بنابراین این ویژگی‌ها از مقاومت بالایی در مقابل نویز برخوردار نیستند و باعث افت عملکرد سیستم در محیط‌های نویزی می‌شوند.

یکی از روش‌های متداول برای غلبه بر اثر نویز استفاده از تبدیل موجک به شکل‌های مختلف است. در ۲۰۱۲ یک سیستم تشخیص گوینده با استفاده از متوسط‌گیریِ فریم‌ها برای ضرایب چندجمله‌ای و ترکیب آن با تبدیل موجک، پیشنهاد شده است که تا حدودی نسبت به نویز مقاوم است اما با کاهش نسبت سیگنال به نویز به مقادیر کمتر از ۵ دسی‌بل، کارآیی آن به شدت افت می‌کند [۲]. در سال ۲۰۰۹ یک روش بازشناسایی گوینده مقاوم به نویز پیشنهاد شده است که در آن علاوه بر استخراج ضرایب MFCC و ضرایب چند جمله‌ای، از تبدیل موجک گسسته (DWT) سیگنال نویزی نیز استفاده شده است. این ضرایب با یکدیگر ترکیب شده‌اند و سپس سیستم تشخیص گوینده بر روی ۱۵ گوینده با استفاده از شبکه عصبی MLP پیاده‌سازی شده است [۳]. نتایج بهبود نرخ تشخیص گوینده را در برابر نویز سفید گوسی و نویز رنگی با این ترکیب از ویژگی‌ها نسبت به ضرایب MFCC و ضرایب چند جمله‌ای نشان می‌دهد. در سال ۲۰۱۰ نیز سیستم تشخیص گوینده با استفاده از ضرایب MFCC استخراج شده از کانال‌های موجک سیگنال صحبت و توسط طبقه‌بندی کننده HMM پیاده‌سازی شد. نتایج افزایش نرخ تشخیص به میزان ۰/۶٪ را در شرایط بدون نویز و افزایش ۴٪ را در شرایطی با نویز سفید گوسی جمع شونده در نسبت سیگنال به نویز ۲۰ دسی بل نشان می‌دهند [۴]. در سال ۲۰۱۱ Khaled Daqrouq، از ۳۰ ضریب آنتروپی شانون به دست آمده از تبدیل موجک پاکت[۲] سیگنال به همراه فرکانس پایه و ۴ فرکانس فرمنت اول استخراج شده از سیگنال صحبت به عنوان ورودی به شبکه عصبی MLP جهت شناسایی گوینده مستقل از متن استفاده کرد. نرخ تشخیص جملات گفتاری به دست آمده توسط روش ارائه شده در شرایط بدون نویز ۹۱/۰۹٪ است. در پایان تبدیل موجک گسسته[۳] از سیگنال گرفته شده تا در برابر نویز سفید گوسی جمع شونده با نسبت سیگنال به نویز ۲- دسی بل مقاوم شود که طبق نتایج میانگین نرخ تشخیص با موجک سطح ۴ به ازای ۶ گوینده ۳۹/۵۸٪ به دست آمده است [۵]. هدف این مقاله، ترکیب ویژگی‌های مبتنی بر فریم‌بندی با ویژگی‌های مستقل از متنِ استخراج شده از یک بخش گفتاری است تا تأثیر نویز بر کارآیی سیستم تشخیص هویت گوینده در مجموعه بسته و مستقل از متن کاهش یابد. استفاده از یک واحد گفتاری که هم بتوان ویژگی‌های MFCC را از آن استخراج کرد و هم برای استخراج ویژگی‌های مستقل از متن مناسب باشد، در اینجا مورد توجه قرار گرفته است. این در حالی است که در تحقیقات این حوزه اغلب یا بر فریم‌بندی تاکید شده است و یا از کل یک جمله یا عبارت برای استخراج ویژگی استفاده می‌شود. در اینجا واحد گفتاریِ مورد استفاده قطعه گفتاری یا صوتی نامیده می‌شود که از تقطیع یک جمله و حذف سکوت‌ها به دست می‌آید. با استفاده از این واحد گفتاری، ضرایب MFCC که از روش مبتنی بر فریم‌بندی به دست می‌آیند با ویژگی‌های به دست آمده از قطعه صوتی (بدون فریم‌بندی) شامل ضرایب آنتروپی شانون در گره‌های موجک پاکت، فرکانس پایه و فرکانس چهار فرمنت اول با یکدیگر ترکیب شده‌اند تا برای هر بخش گفتاری یک بردار ویژگی تشکیل شود. به منظور امکان ترکیب ویژگی‌های مبتنی بر فریم‌بندی و مبتنی بر فایل صوتی از الگوریتم خوشه‌بندی Kmeans بر روی ضرایبِ MFCC هر گوینده استفاده شده است. در نهایت بردارهای ویژگی جهت طبقه‌بندی به شبکه عصبی پرسپترون چندلایه (MLP) با الگوریتم پس انتشار خطا داده می‌شوند.

در ادامه این مقاله مطالب زیر ارائه خواهند شد. ابتدا در بخش (۲) سیستم تشخیص گوینده بررسی خواهد شد. در بخش (۳) ویژگی‌های به کار رفته مختصراً توضیح داده شده‌اند. بخش (۴) به بررسی روش طبقه‌بندی با شبکه MLP پرداخته است. روش پیشنهادی در بخش (۵) توضیح داده شده است و در بخش (۶) داده‌ها و نتایج آزمایشات انجام شده مرور خواهند شد. بخش (۷) نیز به بحث و نتیجه‌گیری و ارائه پیشنهادات می‌پردازد.

## ۲- سیستم تشخیص گوینده

تشخیص گوینده، روند خودکار تشخیص شخصی است که صحبت می‌کند و بر اساس اطلاعات منحصربه فرد موج سیگنال گفتار، گوینده را شناسایی می‌کند. سیستم تشخیص گوینده به دو دسته تشخیص هویت گوینده و تصدیق هویت گوینده تقسیم می‌شود [۶]. سیستم‌های تشخیص هویت به سؤال"چه کسی صحبت می‌کند؟" پاسخ می‌دهند. در حالی که در سیستم‌های تصدیق هویت سوال "آیا گوینده همان کسی است که ادعا می‌کند یا خیر؟" مطرح می‌شود.

سیستم‌های تشخیص گوینده از دیدگاهی دیگر به دو دسته گروه باز[۴] و گروه بسته[۵] تقسیم می‌شوند [۶]. در صورتی که سیستم بتواند

گوینده‌ای را که از پیش نمی‌شناسد به عنوان ناشناس معرفی کند گروه باز است و سیستمی که در آن گفتارهای ادا شده محدود به مجموعه گویندگان شناخته شده است، گروه بسته نامیده می‌شود. همچنین بر اساس آنکه گفتار ادا شده از پیش معلوم است یا خیر تشخیص گوینده به ترتیب به دو بخش وابسته به متن و مستقل از متن تقسیم می‌شود [۶]. سیستم تشخیص گوینده دارای دو مرحله آموزش و تشخیص می‌باشد. در مرحله آموزش سیگنال صحبت گویندگان دریافت شده و بعد از استخراج ویژگی از آن، مرجعی برای هر گوینده ساخته می‌شود. این مرجع می‌تواند یک مدل آماری و یا مجموعه‌ای از الگوها باشد. در مرحله تشخیص، بر اساس مقایسه الگوهای مرجع که از مرحله آموزش حاصل شده‌اند با الگوهای به دست آمده از سیگنال صحبت ورودی، گوینده مورد نظر شناسایی می‌شود.

## ۳- استخراج ویژگی

استخراج ویژگی به منظور به دست آوردن مشخصات پایداری از سیگنال صحبت است که بتواند به خوبی گویندگان را از هم متمایز سازد. ویژگی‌های استخراج شده از سیگنال صحبت برای تشخیص هویت گوینده، باید برای گوینده مورد نظر دارای تغییرات جزئی بوده و در عین حال فاصله زیادی با سایر گویندگان داشته باشند. علاوه بر آن برای شناسایی مستقل از متن، باید از ویژگی‌هایی استفاده کرد که کمترین حساسیت به محتوای آوایی[۶] سیگنال صحبت را داشته باشند. در این مقاله از ویژگی‌های مختلف مانند ضرایب کپسترال مبتنی بر معیار مل[۷]، ضرایب آنتروپی شانون در گره‌های موجک پاکت، فرکانس پایه و فرکانس فرمنت‌ها استفاده شده است که در ادامه توضیح داده می‌شوند.

### ۳-۱- ضرایب کپسترال مبتنی بر معیار مل (MFCC)

ضرایب MFCC از ویژگی‌های رایج در تشخیص الگو در مباحث گفتار است. ایده اصلی در ضرایب کپسترال مبتنی بر معیار مل از خواص گوش انسان در دریافت و درک گفتار گرفته شده است. استفاده از خاصیت مل ارزش بیشتری به اطلاعات محدوده پایین فرکانس نسبت به اطلاعات محدوده بالای فرکانس می‌دهد. این مقیاس دقت تشخیص را در سیستم‌های تشخیص گوینده افزایش می‌دهد [۳]. MFCC در اصل نوع بهبود یافته ضرایب کپسترال می‌باشد و با توجه به اینکه این ضرایب نوعی ویژگی زمان کوتاه و مبتنی بر فریم[۸] است، تغییرات زیادی در زمان کوتاه از خود نشان می‌دهد و برای تعیین واج، کلمه و گوینده مناسب است.

برای محاسبه ضرایب MFCC ابتدا سیگنال صحبت فریم‌بندی شده و پنجره‌گذاری می‌شود تا گسستگی در شروع و انتهای هر فریم کاهش یابد. در گام بعدی تبدیل فوریه روی نتیجه مرحله قبلی اعمال می‌شود. بنابراین هر فریم در حوزه زمان به حوزه فرکانس تبدیل می‌گردد. سپس از فیلتر بانک برای محاسبه انرژی طیف پیرامون فرکانس‌های مشخص استفاده می‌شود. این فیلتر بانک شامل تعدادی فیلتر میان گذر است که به صورت خطی در راستای مقیاس مل چیده شده‌اند و پهنای باند هر فیلتر برابر است با فرکانس مرکزی فیلتر مربوطه در مقیاس مل. در مقیاس مل تا حدود یک کیلو هرتز نگاشت‌ها تقریباً خطی است و در فرکانس‌های بالاتر لگاریتمی می‌باشد. مقیاس هرتز با استفاده از رابطه (۱) به مقیاس مل تبدیل می‌شود.

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \qquad (۱)$$

فیلتر بانک سیگنال صوتی را از یک فرکانس گذر مثلثی عبور می‌دهد. بعد از گذراندن ضرایب از فیلتر بانک، ضرایب MFCC از رابطه (۲) محاسبه می‌شوند.

$$c_i = \sqrt{\frac{2}{N}} \sum_{m=1}^{N} \log\big(S(m)\big) \cos\left(\frac{\pi i}{N}(m - 0.5)\right) \qquad (۲)$$

که در آن i=1,2,…,P، P تعداد ضرایب MFCC، N تعداد فیلترها در فیلتر بانک و S(m) خروجی mامین فیلتر در فیلتر بانک مقیاس مل است. در این مقاله برای به دست آوردن ضرایب CCMF از جعبه ابزار voicebox نوشته شده در نرم افزار MATLAB استفاده شده است. پنجره‌های اعمال شده بر فریم‌ها به صورت همینگ و پنجره‌های اعمال شده بر فیلترها در بانک فیلتر مبتنی بر معیار مل به صورت مثلثی در نظر گرفته شده‌اند.

### ۳-۲- ضرایب آنتروپی شانون

تبدیل موجک پاکت تعمیم یافته تجزیه موجک است که امکان بیشتری برای تحلیل سیگنال ایجاد می‌کند. در تحلیل موجک، سیگنال به دو شاخه تقریب[۹] و جزئیات[۱۰] تقسیم می‌شود و این روند بر روی شاخه تقریب تکرار می‌شود. برای n سطح، تعداد n+1 مسیر ممکن برای تجزیه یا کدگذاری سیگنال وجود دارد. در تحلیل موجک پاکت، جزئیات را نیز می‌توان همانند تقریبات تجزیه کرد. حاصل این کار دسترسی به بیش از $2^n$ روش متفاوت برای کدگذاری سیگنال است. در حقیقت تجزیه موجک گوشه‌ای از تجزیه به روش موجک پاکت است.

بنابراین تبدیل موجک پاکت نمایش بهتری برای آنالیز کردن سیگنال نسبت به تبدیل موجک ارائه می‌دهد. بنابراین می‌توان برای استخراج ویژگی‌های اضافی به منظور نرخ تشخیص بالاتر به کار رود. اما با افزایش عمق درخت موجک پاکت، زمان مورد نیاز برای طبقه‌بندی مناسب پایگاه داده غیرخطی می‌شود و در نتیجه کاهش بعد ضرایب به دست آمده از موجک پاکت امری ضروری است. بدین منظور ضرایب آنتروپی شانون در گره‌های موجک پاکت که یکی از روش‌های مناسب برای کاهش ابعاد ویژگی‌ها است [۷،۵] محاسبه می‌شوند.

به منظور استخراج ضرایب آنتروپی شانون، ابتدا سیگنال صحبت به وسیله تبدیل موجک پاکت در سطح مشخصی تجزیه شده و سپس آنتروپی شانون برای همه گره‌ها در همان سطح با استفاده از معادله (۳) محاسبه می‌شود.

$$El(s) = -\sum_i S_i^2 \log(S_i^2) \qquad (۳)$$

که در آن S سیگنال صحبت و $S_i$ ضرایب تبدیل موجک پاکت می‌باشند. این ویژگی جزو ویژگی‌های زمان بلند سیگنال صحبت است و معمولاً در فواصل زمانی کوتاه، محاسبه نمی‌شود.

### ۳-۳- ویژگی‌های فرکانس پایه و فرکانس فرمنت‌ها

فرکانس پایه فرکانس ارتعاش تارهای صوتی در طی تولید گفتار می‌باشد. اگر کشش تارهای صوتی هنگامی که جریان هوا از دهانه حنجره عبور می‌کند فقط سبب نوسانات آرام شود، می‌تواند پالس هوای شبه پریودیک تولید کند. وجود این پالس سبب می‌شود که مجرای صوتی بتواند حروف صدادار را تولید کند. فرکانس پایه، خصوصیات پریودیک نوسانات تارهای صوتی گوینده را هنگام بیان حروف صدادار نشان می‌دهد و برای شناسایی گوینده مؤثر واقع می‌شود. خاصیت منحصر بفرد بودن فرکانس پایه، به دلیل ساختار آناتومی متفاوت و منحصر بفرد هر گوینده است. فرکانس پایه نه تنها یک ویژگی مستقل از متن است بلکه در تحقیقات گفتاری به عنوان یک ویژگی مؤثر در تشخیص جنسیت گوینده بسیار پرکاربرد است [۸ و ۹].

انتشار امواج پریودیک در سیگنال صحبت، در طیف برخی آواها به ویژه واکه‌ها[۱۱] دیده می‌شود. اندام‌های گفتاری شکل‌های معینی را برای تولید حروف صدادار ایجاد می‌کنند. بنابراین مناطقی از تشدید و ضد تشدید در مجرای صوتی شکل می‌گیرد. مکان این تشدیدها در طیف فرکانسی بستگی به نوع و شکل مجرای صوتی گوینده دارد. فرکانس‌های تشدید شده صوتی در طول گفتار واکدار، فرکانس‌های فرمنت نامیده می‌شوند. فرمنت‌ها هم می‌توانند مانند فرکانس پایه، برای تشخیص جنسیت به کار روند [۱۰] همچنین حالت ملودیک، فیزیولوژی و حتی حالات روحی گوینده را نیز نشان می‌دهند [۱۱]. لازم به ذکر است که فرکانس پایه و فرمنت‌ها جزو ویژگی‌های زمان بلند سیگنال صحبت هستند و معمولاً در ترکیب با سایر ویژگی‌ها استفاده می‌شوند.

روش‌های مختلفی برای به دست آوردن فرکانس پایه و فرکانس‌های فرمنت ارائه شده است. روش‌های کپستروم [۱۲]، تابع خود همبستگی [۱۳] و طیف حاصلضرب هارمونیک [۱۴] از جمله پرکاربردترین‌ها هستند. در این مقاله از روش الگوریتم چگالی توان طیفی (PSD[۱۲]) برای تخمین فرکانس پایه و فرکانس فرمنت استفاده شده است.

الگوریتم PSD شامل دو مرحله است [۵]: ابتدا PSD ($P_{xx}$) با استفاده از روش Yule-Walker خود بازگشتی[۱۳] (AR) تخمین زده می‌شود سپس در مرحله دوم ماکزیمم‌های محلی تشخیص داده می‌شوند. روش Yule-Walker یک روش تخمین طیف پارامتری است که خروجی آن ($P_{xx}$) تخمینی از PSD سیگنال x می‌باشد. این روش مدل فیلتر پیشگوئی خطی خود بازگشتی را توسط مینیمم کردن خطای پیشگوئی پیشرو به شکل حداقل مربعات به سیگنال نسبت می‌دهد. طیف تخمینی به دست آمده به وسیله این روش، مربعات دامنه پاسخ فرکانسی این مدل خود بازگشتی است. پس از اینکه PSD تخمین زده شد، تفاضل آن با استفاده از رابطه (۴) محاسبه می‌شود.

$$D = P_{xx}(i+1) - P_{xx}(i) \qquad (۴)$$

که در آن $P_{xx}(i)$ و $P_{xx}(i+1)$ مقدار چگالی طیف فرکانسی در دو فریم متوالی هستند. سپس محل قرارگیری ماکزیمم‌های محلی، یعنی جایی که D از مثبت به منفی تغییر علامت می‌دهد، شناسایی می‌شود. اولین قله مشخص کننده فرکانس پایه و بقیه قله‌ها به ترتیب نشان‌دهنده فرمنت‌های اول تا چهارم هستند.

### ۴- طبقه بندی با استفاده از MLP

شبکه‌های MLP به دلیل انعطاف‌پذیری و پایداری، توانایی زیادی در طبقه‌بندی الگوها دارند. شبکه MLP استفاده شده در این مقاله یک شبکه چهار لایه، شامل یک لایه ورودی، دو لایه مخفی و یک لایه خروجی می‌باشد. استفاده از دو لایه پنهان به دلیل افزایش توانایی یادگیری شبکه بر روی الگوهای ترکیبی از ویژگی‌های مختلف است. این شبکه برمبنای الگوریتم پس انتشارخطا آموزش داده می‌شود. در این روش خروجی‌های واقعی (خروجی‌های شبکه) باخروجی‌های مطلوب مقایسه می‌شوند و وزن‌ها به وسیله الگوریتم پس انتشار، به صورت تحت نظارت تنظیم می‌گردند تا الگوی مناسب به وجود آید.

مجموع مربع خطا، E، بین خروجی مطلوب و خروجی واقعی برای تمامی نورون‌های لایهٔ خروجی شبکه به صورت رابطه (۵) بیان می‌شود [۱۵].

$$E = \frac{1}{2}\sum_{n=1}^{N}(D_n - Y_n)^2 \qquad (۵)$$

که $D_n$ و $Y_n$ به ترتیب خروجی‌های مطلوب و واقعی نورون شماره nام خروجی هستند و N تعداد نورون های خروجی است. وزنها با هدف کاهش E به مقدار مینیمم به روش گرادیان نزولی تنظیم می‌گردند. معادلهٔ به روز درآوردن وزن‌ها بعد از هر دوره آموزش به صورت روابط (۶) و (۷) است [۱۵]:

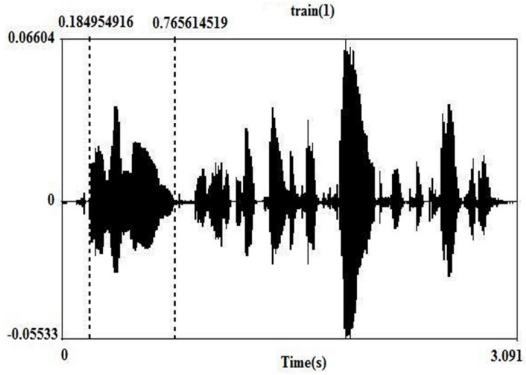
$$w_{ij}(t+1) = w_{ij}(t) + \eta\Delta w_{ij}(t) \qquad (۶)$$

$$\Delta w_{ij}(t) = -\left(\frac{\partial E_n}{\partial w_{ij}(t)}\right) \qquad (۷)$$

که η نرخ یادگیری است و معمولا بین ۰ و ۱ انتخاب می‌شود، $w_{ij}(t+1)$ وزن جدید و $w_{ij}(t)$ وزن قبلی می‌باشد و $E_n$ مجموع مربع خطای خروجی برای الگوی ورودی nام است. روند یادگیری هنگامی متوقف می‌شود که مجموع کل خطا، E، برای همه الگوها از مقدار آستانهٔ تعیین شده کمتر شود یا تعداد کل دوره‌های آموزش به پایان برسد [۱۵].

### ۵- روش پیشنهادی

راه حل پیشنهادی در این تحقیق، استفاده از ترکیب مناسب ویژگی‌های مبتنی بر فریم با ویژگی‌های مبتنی بر قطعات معنی‌دار گفتاری است. در اینجا پیشنهاد شده است که برای استخراج ویژگی‌های زمان بلند، به جای تکیه بر کل جمله یا عبارت از قطعات گفتاری استفاده شود که شامل تعدادی آوا هستند. هر قطعه می‌تواند بسته به محتوای آوایی و سرعت صحبت گوینده شامل یک یا دو هجا باشد. در مواردی که واکه‌ها خیلی کوتاه بیان شده باشند طول قطعه تا یک کلمه نیز افزایش می‌یابد. تقطیع جملات به صورت دستی و با تکیه بر محتوای صوتی و شکل ظاهری سیگنال صحبت انجام گرفته است. شکل (۱) یک جمله و یک بخش جدا شده را نشان می‌دهد.




شکل (۱): یکی از جملات موجود در دادگان و نحوه تقطیع: خط چین‌ها یک ناحیه تقطیع شده را نشان می‌دهند.

Fig. (1): A sample utterance of dataset and segmentation process: dot lines correspond to a segment

ویژگی‌های اولیه مورد استفاده در این تحقیق ضرایب MFCC هستند. این ضرایب به عنوان پرکاربردترین ویژگی‌ها در تشخیص گوینده شناخته می‌شوند و پس از فریم‌بندیِ هر قطعه سیگنال صحبت ضرایب MFCC برای هر فریم محاسبه می‌شوند. اینکه چه تعداد از این ضرایب برای تشخیص گوینده مناسب است نیز در این مقاله مورد بررسی قرار گرفته است. تعداد ضرایب از ۱۲ به تدریج افزایش داده شده و کارآیی سیستم مورد بررسی قرار گرفته است.

ویژگی دیگری که در این مقاله مورد استفاده قرار گرفته است ضرایب آنتروپی است. ضرایب آنتروپی شانون در گره‌های موجک پاکت بدون فریم‌بندی سیگنال برای هر قطعه گفتاری محاسبه می‌شوند. این ضرایب به راحتی از هر قطعه استخراج می‌شوند و چون هر کدام از قطعات شامل دسته‌ای از آواهای واکدار و بی‌واک هستند، این ضرایب خصوصیات صحبت گوینده را در گذر از یک آوا به آوای بعدی به خوبی نمایش می‌دهند. برای محاسبه ضرایب آنتروپی شانون، ابتدا سیگنال صحبت با تبدیل موجک پاکت در سطح چهار و با نوع Daubechies (db1) تجزیه می‌شود. با تجزیه سیگنال در سطح ۴، ۳۰ گره به دست می‌آید که ضرایب آنتروپی شانون در تمامی این گره‌ها محاسبه می‌شوند.

**۵-۱- استخراج فرکانس پایه و فرمنت‌ها**

فرکانس پایه و فرکانس چهار فرمنت اول از دیگر ویژگی‌های مورد استفاده دیگر در این روش هستند که برای محاسبه آنها از الگوریتم PSD استفاده می‌شود. این الگوریتم فرکانس پایه و فرکانس فرمنت‌ها را برای هر قطعه گفتاری (بدون فریم‌بندی) استخراج می‌کند. به این منظور ابتدا در هر قطعه گفتاری یک واکه بلند پرانرژی با روشی مقاوم به نویز شناسایی می‌شود تا فرکانس پایه و فرمنت‌ها از آن استخراج شوند.

معمولاً برای شناسایی نواحی واکدار از نوعی آستانه گذاری بر روی ویژگی‌های زمان فرکانس استفاده می‌شود. استفاده از تابع انرژی زمان کوتاه، نرخ عبور از صفر، دامنه پیک کپسترال و یا بعد فرکتال سیگنال گفتار برای این کار متداول است. این روش‌ها مبتنی بر آستانه ثابت هستند و تعیین دقیق مقادیر آستانه و انتخاب پارامترهای مورد نظر جهت آستانه گذاری بسیار حائز اهمیت است.

در این مقاله از ترکیب چهار پارامتر یعنی انرژی، دامنه پیک کپسترال، نرخ عبور از صفر و بعد فرکتال، جهت بازشناسی نواحی واکدار از بی‌واک استفاده شده است. مقادیر آستانه نیز به توجه به پارامترهای آماری دادگان تعلیم محاسبه شده‌اند. ضمن آنکه آستانه مربوط به پیک کپسترال به صورت وقفی[۱۴] تنظیم می‌شود. از آنجایی که دامنه پیک کپسترال در یک بخش واکدار ابتدا به صورت صعودی و سپس به صورت نزولی تغییر می‌کند، این کمیت با حرکت بر روی نواحی واکدار، دائماً به روزرسانی می‌شود تا تغییرات گذرا در بخش‌های حاوی فرکانس پایه در آن لحاظ شود.

برای هر فریم انرژی زمان کوتاه از رابطه (۸) محاسبه می‌شود که در آن $E_n$ انرژی قاب nام، W طول پنجره و s(i) نمونه گفتاری پنجره شده است. معمولاً سطح انرژی در نواحی واکدار بالاتر از نواحی بی‌واک است.

$$E_n = \sum_{i=1}^{W} s(i)^2 \qquad (۸)$$

کمیت مهم دیگر بعد فرکتال است، ورود به نواحی بی‌واک گفتار با افزایش این کمیت همراه است. برای محاسبه بعد فرکتال از روش Kat'z [۱۶] استفاده شده است. در این روش بعد فرکتال از فرمول (۹) محاسبه می‌شود:

$$FD_n = \frac{\log_{10}(L)}{\log_{10}(d)} \qquad (۹)$$

که در آن $FD_n$ بعد فرکتال قاب nام، L طول کل منحنی سیگنال پنجره شده یا همان مجموع فاصله بین نقاط متوالی است و d فاصله اقلیدسی بین اولین نمونه سیگنال با نمونه‌ای است که بیشترین فاصله را دارد. L و d به صورت روابط (۱۰) و (۱۱) محاسبه می‌شوند:

$$L = \sum_{i=1}^{W-1} \sqrt{[(s(i+1) - s(i))^2 + 1]} \qquad (۱۰)$$

$$d = \max\{\text{distance}(s(1), s(i)), i = 2, \dots, W\} \qquad (۱۱)$$

کمیت بعدی، نرخ عبور از صفر است که به صورت تعداد دفعات تغییر علامت سیگنال در پنجره تعریف می‌شود. برای محاسبه پیک کپسترال، ابتدا کپستروم سیگنال پنجره شده از روش متداول محاسبه می‌شود. سپس دامنه پیک کپسترال در فاصله ۲/۵ تا ۱۵ میلی‌ثانیه شناسایی می‌شود. مقادیر آستانه به صورت روابط (۱۲) تا (۱۵) تعریف می‌شوند:

$$Th1 = k1.\,mean(E) \qquad (۱۲)$$

$$Th2 = k2.\,mean(FD) \qquad (۱۳)$$

$$Th3 = k3.\,median(ZCR) \qquad (۱۴)$$

$$Th4 = k4.\,median(CP) \qquad (۱۵)$$

که در آن mean و median به ترتیب نشان‌دهنده میانگین و میانه آماری هستند. نحوه محاسبه این مقادیر آستانه و مقادیر k1 تا k4 به صورت تجربی است و پس از انجام تحلیل‌های آماری بر روی داده‌ها به دست آمده است. برای تشخیص نواحی واکدار/ بی‌واک، مقادیر CP, ZCR, FD, E محاسبه شده و با استفاده از قوانین زیر با مقادیر آستانه مقایسه می‌شوند. خروجی این قوانین برچسبی است که مشخص می‌کند هر فریم مربوط به ناحیه سکوت، گفتار واکدار و یا بی‌واک است.

• If ($E_n$<0.5.Th1) ∧ ($FD_n$>Th2) ∧ ($ZCR_n$>Th3)→ Silence

• If (0.5.Th1<$E_n$<Th1) ∧ ($FD_n$>Th2) ∧ ($CP_n$<Th4)→Unvoiced Speech

• If ($E_n$>Th1) ∧ ($FD_n$<Th2) ∧ ($ZCR_n$<Th3)∧ ($CP_n$>Th4)

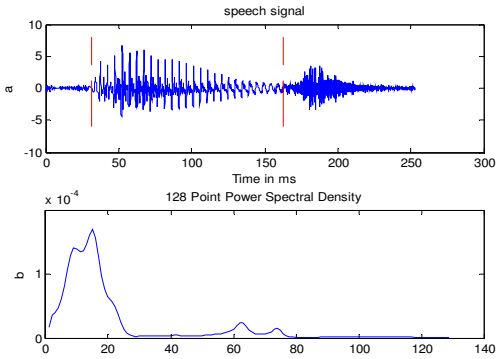
✓ If ($CP_{n-1}$<Th4) ∧ ($CP_{n+1}$>Th4)→ Start of Voiced Speech, Th4=2.Th4

✓ If ($CP_{n-1}$>Th4) ∧ ($CP_{n+1}$>Th4)→ Voiced Speech

✓ If ($CP_{n-1}$>Th4) ∧ ($CP_{n+1}$<Th4)→ End of Voiced Speech, Th4=0.5.Th4

✓ If ($CP_{n-1}$<Th4) ∧ ($CP_{n+1}$<Th4)→ Unvoiced Speech

خروجی این قوانین بخش‌های واکدار با انرژی بالا را مشخص می‌کند که نشان دهنده واکه‌ها هستند و می‌توان فرکانس پایه و فرکانس چهار فرمنت اول را با استفاده از الگوریتم PSD برای واکه‌ها محاسبه کرد. در مواردی که قطعه مورد نظر بیش از یک واکه داشته باشد، واکه با طول بیشتر مبنای محاسبه قرار گرفته است. شکل (۲) یک قطعه گفتاری و واکه استخراج شده از آن با روش پیشنهادی را نشان می‌دهد که با خطچین‌های عمودی مشخص شده است. قسمت پایین شکل، PSD این واکه را نمایش می‌دهد. موقعیت پیک اول برای محاسبه فرکانس پایه و موقعیت بقیه پیک‌ها برای محاسبه فرمنت‌ها مورد استفاده قرار گرفته است.




شکل (۲) یک قطعه گفتاری و واکه استخراج شده از آن (a) به همراه PSD آن واکه (b)

Fig. (2): A sample speech segment and the extracted vowel (a) with vowel's PSD (b)

### ۵-۲- ترکیب ویژگی‌ها و طبقه‌بندی

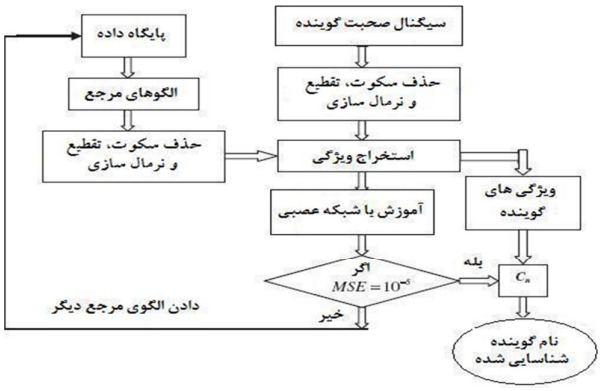
به منظور ترکیب دو دسته ویژگی، یعنی ویژگی‌های مبتنی بر فریم‌بندی و ویژگی‌های مبتنی بر قطعات صوتی از الگوریتم خوشه‌بندی Kmeans بر روی ضرایب MFCC استفاده می‌شود. برای هر قطعه گفتاری یک مرکزِ خوشه در نظر گرفته می‌شود. بنابراین بردار ویژگی نهایی حاوی ۳۵ ویژگی مبتنی بر قطعه صوتی شامل ۳۰ ضریب آنتروپی شانون، فرکانس پایه در یک واکه و فرکانس چهار فرمنت اول است که با ۱۲ ضریب MFCC مربوط به مرکز خوشه آن قطعه صوتی ترکیب می‌شود. در مجموع بردار ویژگی نهایی که از هر بخشِ دادگان استخراج می‌شود یک بردار ۴۷ بعدی خواهد شد. مشخص است که در صورت استفاده از تعداد بیشتری ضریب MFCC بعد بردار ویژگی افزایش خواهد یافت.

سپس بردارهای ویژگی به عنوان ورودی به شبکه عصبی وارد می‌شوند. شبکه عصبی به کار رفته جهت طبقه‌بندی، شبکه پرسپترون سه لایه، با دو لایه مخفی می‌باشد. تعداد گره‌های لایه اول برابر با ابعاد بردارهای ویژگی است که در اینجا ۴۷ می‌باشد. تعداد نورون‌های لایه مخفی اول و دوم به تنوع گویندگان بستگی دارد و برای ۱۵ گوینده مورد آزمایش در این تحقیق با روش تجربی به دست می‌آید. پس از چندین مرحله آزمایش، بهترین عملکرد با ۵۰ نورون در لایه مخفی اول و ۴۰ نورون در لایه مخفی دوم است. نتایج آزمایشات تجربی برای دستیابی به بهترین عملکرد در شبکه عصبی در جدول (۱) آمده است. بررسی‌ها مربوط به بردار ویژگی ۴۸ بعدیِ پیشنهادی در حالت بدون نویز است. همانطور که مشاهده می‌شود افزایش بیشتر تعداد نورون‌ها منجر به تغییر محسوسی در نرخ بازشناسی نمی‌شود و تنها زمان یادگیری را افزایش می‌دهد. ضمناً استفاده از تعداد بیشتری از لایه‌های مخفی به دلیل حجم انبوه محاسبات در مرحله آموزش و تست، معمول نیست. خروجی شبکه عصبی چهار نورون دارد که می‌توانند ۱۶ حالت مختلف را ایجاد کند. بلوک دیاگرام سیستم تشخیص هویت گوینده شبیه‌سازی شده مطابق شکل (۳) است. ساختار شبکه عصبی استفاده شده در این بلوک دیاگرام در شکل (۴) نشان داده شده است که در آن NH تعداد نورون‌های لایه مخفی را بیان می‌کند. نمودار خطای آموزش شبکه عصبی نیز در شکل (۵) نشان داده شده است.

Table (1): Experimental results for neural network to achieve the best performance

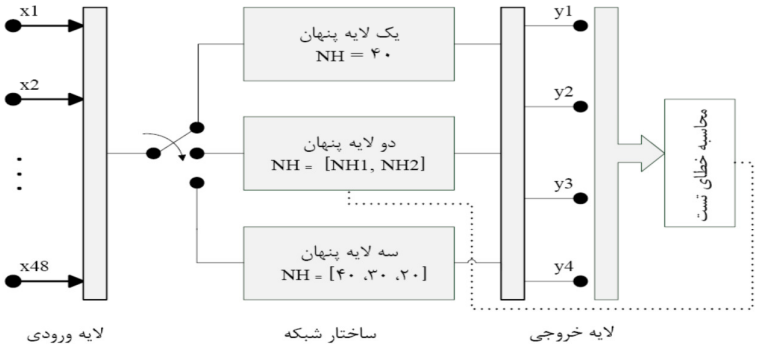
جدول (۱): نتایج آزمون‌های تجربی بر روی شبکه عصبی برای دستیابی به بهترین عملکرد

| پارامترهای شبکه | | سایر پارامترهای شبکه | میانگین نرخ تشخیص گوینده | زمان آموزش بر حسب ثانیه |
|---|---|---|---|---|
| تعداد لایه‌های مخفی | 1 | لایه مخفی اول ۴۰ نورون | 80 | 98.4756 |
| | 2 | لایه مخفی دوم ۳۰ نورون | 85.33 | 242.6515 |
| | 3 | لایه مخفی سوم ۲۰ نورون | 86.66 | 518.0851 |
| تعداد نورون‌ها در لایه مخفی اول | 40 | دارای دو لایه مخفی | 85.33 | 242.6515 |
| | 45 | | 86.66 | 348.6753 |
| | 50 | لایه مخفی دوم ۳۰ نورون | 88 | 384.6829 |
| | 55 | | 89.33 | 501.5637 |
| تعداد نورون‌ها در لایه مخفی دوم | 30 | دارای دو لایه مخفی | 88 | 384.6829 |
| | 35 | | 90 | 412.0913 |
| | 40 | لایه مخفی اول ۵۰ نورون | 97.33 | 498.3161 |
| | 45 | | 98.66 | 617.1348 |



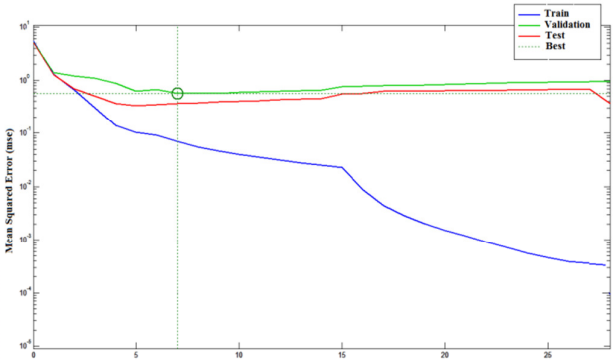

شکل (۳): ساختار سیستم تشخیص گوینده

Fig. (3): The structure of speaker identification system




شکل (۴): ساختار شبکه عصبی

Fig. (4): Neural network structure




شکل (۵): نمودار خطای آموزش شبکه عصبی

Fig. (5): Neural network training error curve

## ۶- نتایج آزمایشات

در این قسمت نتایج آزمایشات و تحلیل آنها با استفاده از روش پیشنهادی و چند روش متداول در تشخیص گوینده آمده است. تمامی تست‌ها هم در حالت بدون نویز و هم در حالت نویزی با نویز سفید گوسی جمع شونده در نسبت‌های سیگنال به نویز مختلف انجام شده است.

### ۶-۱- داده‌ها

داده‌های آزمایش شامل سیگنال‌های گفتار به ازای ۱۵ گوینده شامل ۸ زن و ۷ مرد است که از پایگاه داده TIMIT استخراج شده‌اند. این داده‌ها با فرکانس نمونه‌برداری ۱۶ کیلو هرتز و دقت ۱۶bit به صورت فایل‌های wave جمع‌آوری شده‌اند. هر گوینده ۱۰ جمله را بیان کرده است که ۲ جمله در بین همه گویندگان مشترک بوده و ۸ جمله دیگر متفاوت هستند. ۸ جمله مذکور برای استخراج قطعات گفتاری مربوط به دادگان آموزشی و ۲ جمله مشترک برای تست مورد استفاده قرار گرفته‌اند. استفاده از جملات متفاوت برای تعلیم سیستم و جملات مشترک برای تست، باعث سنجش دقیقی از عملکرد آن در حالت مستقل از متن می‌شود.

در این مقاله برای استخراج ویژگی‌های MFCC که مبتنی بر فریم‌بندی هستند، طول هر فریم ۴۰ میلی‌ثانیه و همپوشانی فریم‌ها ۳۰ میلی‌ثانیه در نظر گرفته شده است. با این حال به منظور افزایش دقت سیستم شناسایی و آماده‌سازی داده‌ها برای استخراج ویژگی‌های دیگر، باید قسمت‌های غیرگفتاری (سکوت) سیگنال صحبت شناسایی شده و حذف شود. علاوه بر آن برای استخراج ویژگی‌های مبتنی بر واکه‌ها از جمله فرکانس پایه و فرمنت‌ها لازم است که هر قطعه گفتاری حاوی حداقل یک واکه باشد. ضمن آنکه استفاده از قطعات بیش از حد کوتاه و یا بلند، کارآیی ویژگی‌های مبتنی بر آنتروپی را در تشخیص گوینده کاهش می‌دهد. بنابراین مرحله آماده‌سازی داده‌ها از اهمیت ویژه‌ای برخوردار است.

در این مقاله سکوت‌های بین کلمات یا مجموعه آواها در هر جمله با استفاده از نرم‌افزار PRAAT[۱۵] حذف و هر یک از قطعات گفتاری در یک فایل صوتی مجزا ذخیره شده است. بررسی بیشتر نشان می‌دهد که بهترین نوع تقطیع، جداسازی تک‌هجایی یا دوهجایی است. البته در برخی موارد که واکه‌ها بسیار کوتاه ادا شده‌اند از تعداد بیشتری آوا و یا کلمه استفاده شده است تا در استخراج فرکانس پایه خطا ایجاد نشود. هر جمله با توجه به محتوای آن و نحوه ادا شدن توسط گوینده به تعدادی قطعه گفتاری تقسیم شده است. جدول (۲) تعداد قطعات گفتاری مربوط به هر گوینده را در دادگان آموزشی و تست نشان می‌دهد.

Table (2): Number of speech segments for each speaker

جدول (۲): تعداد قطعات گفتاری داده‌ها مربوط به هر گوینده

| گویندگان | جنسیت | تعداد قطعات گفتاری دادگان آموزش | تعداد قطعات گفتاری دادگان تست |
|---|---|---|---|
| گوینده شماره ۱ | زن | 32 | 7 |
| گوینده شماره۲ | زن | 46 | 9 |
| گوینده شماره ۳ | زن | 40 | 12 |
| گوینده شماره ۴ | زن | 42 | 8 |
| گوینده شماره ۵ | زن | 37 | 10 |
| گوینده شماره ۶ | مرد | 45 | 9 |
| گوینده شماره ۷ | مرد | 47 | 10 |
| گوینده شماره ۸ | مرد | 45 | 9 |
| گوینده شماره ۹ | مرد | 49 | 11 |
| گوینده شماره ۱۰ | مرد | 38 | 11 |
| گوینده شماره ۱۱ | زن | 44 | 9 |
| گوینده شماره ۱۲ | زن | 50 | 8 |
| گوینده شماره ۱۳ | زن | 36 | 10 |
| گوینده شماره ۱۴ | مرد | 35 | 7 |
| گوینده شماره ۱۵ | مرد | 24 | 6 |

برای حذف مقدار DC فیلتر پیش تاکید بر روی سیگنال‌ها اعمال شده و در نهایت سیگنال‌های صحبت توسط فرمول (۱۶) نرمالیزه می‌شوند [۱۷] تا صرف‌نظر از تفاوت در دامنه قابل مقایسه با یکدیگر شوند.

$$S_{Ni} = \frac{S_i - \mu}{\sigma} \qquad (۱۶)$$

که $S_i$، iامین مولفه سیگنال S است، $\mu$ و $\sigma$ به ترتیب میانگین و انحراف معیار بردار S می‌باشند. $S_{Ni}$، iامین مولفه از سری‌های سیگنال $S_N$ پس از نرمالیزاسیون است.

### ۶-۲- استخراج ویژگی و طبقه‌بندی

در این مرحله ویژگی‌های مورد نظر از هر قطعه سیگنال استخراج می‌شود. استخراج ویژگی از دادگان آموزشی در حالت بدون نویز و در دادگان تست پس از اضافه کردن نویز به ازای مقدار مشخصی از نسبت سیگنال به نویز انجام می‌گیرد. ۱۲ تا ۱۶ضریب MFCC پس از فریم-بندی استخراج می‌شوند. سپس بر روی ضرایب استخراج شده از هر قطعه گفتاری یک خوشه‌بندی Kmeans اعمال می‌شود تا یک مرکز خوشه برای هر فایل صوتی به دست آید. برای هر فایل صوتی ۳۰ ضریب آنتروپی شانون نیز محاسبه شده و به بردار ویژگی اضافه می‌شود. سپس با استفاده از روش مبتنی بر آستانه‌گذاری که در قسمت ۵-۱- بیان شد، محدوده یک واکه پرانرژی در هر قطعه گفتاری شناسایی شده، فرکانس پایه و چهار فرمنت با روش PSD استخراج می‌شوند.

سپس شبکه با استفاده از ویژگی‌های دادگان آموزشی تحت آموزش قرار گرفته و تست می‌شود. با توجه به نحوه طراحی خروجی هدف، در هنگام تست از معیار فاصله اقلیدسی برای تعیین میزان شباهت خروجی شبکه با خروجی هر گوینده استفاده می‌شود. کمترین فاصله به عنوان گوینده شناسایی شده در نظر گرفته می‌شود. نحوه محاسبه فاصله اقلیدسی و تشخیص گوینده به این صورت رابطه (۱۷) است [۵]:

$$C_n = 100 - \left[100\sqrt{\left(\sum(P_n - SR_n)^2 / \sum {P_n}^2\right)}\right] \qquad (۱۷)$$

که در آن $C_n$ معیار شباهت به دست آمده با استفاده از فاصله اقلیدسی، $P_n$ خروجی هدف مرجع گویندگان و $SR_n$ خروجی شبکه می باشد. کمیت n شماره گوینده را نشان می‌دهد و بنابراین $n=\{1,2,\ldots,15\}$ است. منظور از علامت تفریق، تفاضل تک به تک اعضا هر بردار است. به دلیل نحوه تعریف خروجی هدف و ساختار شبکه، بردارها دارای ۴ عضو هستند. فرمول نوشته شده در بین براکت فاصله اقلیدسی را به صورت عددی بین ۰ تا ۱ بیان می‌کند که با ضرب آن در عدد ۱۰۰ فاصله به صورت درصد بیان می‌شود. با کسر این مقدار از ۱۰۰ معیاری به دست می‌آید که نشان دهنده درصد شباهت خروجی با خروجی هدف هر یک از گویندگان است. گوینده شناسایی شده به صورت زیر با رابطه (۱۸) مشخص می‌شود.

$$C = \arg(\max(C_n, n = 1, \ldots, 15)) \qquad (۱۸)$$

به منظور مقایسه و ارزیابی عملکرد سیستم پیشنهادی، در این آزمایش‌ها دو نوع نرخ تشخیص برای سیستم محاسبه می‌شود. یکی از

معیارهای ارزیابی، نرخ تشخیص گوینده است که در اینجا نحوه محاسبه این کمیت به این صورت است که برای هر گوینده اگر بیش از نصف قطعات صوتی درست تشخیص داده شود، گوینده مورد نظر صحیح تشخیص داده شده است.

علاوه بر نرخ تشخیص گوینده، کمیت دیگری با عنوان نرخ تشخیص قطعه گفتاری نیز در نظر گرفته شده است. برای محاسبه این کمیت ابتدا مجموع کل قطعات سیگنال صحبت موجود در دادگان آزمایش محاسبه می‌شود. سپس تعداد قطعات سیگنالی که توسط شبکه به درستی برچسب‌دهی شده‌اند نیز محاسبه می‌گردد. مجموع کل قطعه سیگنال‌هایی که به درستی بازشناسی شده‌اند تقسیم بر کل قطعات سیگنال‌های موجود به صورت درصد، بیانگر نرخ تشخیص سیگنال است. روابط (۱۹) و (۲۰) به ترتیب نحوه محاسبه نرخ تشخیص قطعه صوتی و نرخ تشخیص گوینده را نشان می‌دهند.

$$\text{Segment Accuracy\%} = 100[\sum(\text{Truely Detected Segments})/\sum(\text{All segmants})] \quad (۱۹)$$

$$\text{Speaker Accuracy\%} = 100[\sum(\text{Truely Detected Speakers})/\sum(\text{Speakers})] \quad (۲۰)$$

ضمناً به منظور سنجش صحیح عملکرد شبکه و با توجه به دو لایه بودن شبکه و احتمال یادگیری ناقص، تمامی مراحل تست و آموزش به جای یک بار، پنج بار انجام شده است. در هر مرحله شبکه پس از آموزش با داده‌های بدون نویز، با داده‌های نویزی مورد تست قرار می‌گیرد. میانگین هر دو پارامتر ارزیابی و بهترین نتیجه طی پنج بار آموزش و تست شبکه به عنوان پارامتر ارزیابی مورد استفاده قرار گرفته است.

Table (3): Recognition results in noise-free condition

جدول (۳): نتایج تشخیص در حالت بدون نویز

| بردار ویژگی | بهترین نرخ تشخیص گوینده | میانگین نرخ تشخیص گوینده | بهترین نرخ تشخیص قطعه گفتاری | میانگین نرخ تشخیص قطعه گفتاری |
|---|---|---|---|---|
| ۱۲ ضریب MFCC | 86.66 | 62.66 | 80.88 | 63.08 |
| ۱۲ ضریب MFCC+ انرژی | 100 | 89.33 | 85.29 | 76.32 |
| ۱۲ ضریب MFCC+ فرمنت+ فرکانس پایه+ آنتروپی | 86.66 | 76 | 77.20 | 66.76 |
| ۱۲ ضریب MFCC+ انرژی+ فرمنت+ فرکانس پایه+ آنتروپی | 100 | 97.33 | 83.82 | 82.20 |

Table (4): Recognition results in -2dB SNR

جدول (۴): نتایج تشخیص با نسبت سیگنال به نویز ۲- دسی بل

| بردار ویژگی | بهترین نرخ تشخیص گوینده | میانگین نرخ تشخیص گوینده | بهترین نرخ تشخیص قطعه گفتاری | میانگین نرخ تشخیص قطعه گفتاری |
|---|---|---|---|---|
| ۱۲ ضریب MFCC | 80 | 54.66 | 65.44 | 53.38 |
| ۱۲ ضریب MFCC+ انرژی | 53.33 | 37.33 | 58.08 | 42.79 |
| آنتروپی | 66.66 | 46.66 | 61.74 | 44.85 |
| آنتروپی+ فرکانس پایه+ فرمنت | 60 | 54.66 | 55.88 | 52.05 |
| ۱۲ ضریب MFCC+ فرکانس پایه+ فرمنت+ آنتروپی | 86.66 | 61.33 | 72.79 | 59.70 |

### ۶-۳- نتایج آزمایشات در حالات مختلف ترکیب ویژگی‌ها

در اولین آزمایش از ۱۲ ضریب MFCC و ترکیب‌های مختلفی از ویژگی‌ها استفاده شده است. نتایج تشخیص گوینده هم به صورت میانگین و هم به صورت بهترین نتیجه در حالت بدون نویز و با نویز ۲- دسی‌بل به ترتیب در جداول (۳) و (۴) نشان داده شده است. کلیه نتایج به صورت درصد بیان شده است. در حالت کلی افزایش میانگین تشخیص نوعی بهبود پایدار در نتایج را نشان می‌دهد. در حالی که افزایش بهترین نرخ تشخیص به تنهایی، می‌تواند نشان دهنده ناپایداری شبکه در یادگیری متعادل در هر ۵ مرحله آموزش و تست باشد.

همانطور که در جدول (۳) مشاهده می‌شود بهترین نرخ تشخیص گوینده به ازای ۱۲ ضریب MFCC به همراه انرژی زمان کوتاه به دست می‌آید. نتیجه مشابهی به ازای ترکیب کلیه ویژگی‌ها حاصل می‌شود. با این حال میانگین نرخ تشخیص گوینده با استفاده از ترکیب ویژگی‌های پیشنهادی بسیار بالاتر است. این مسئله نشان می‌دهد ترکیب این ویژگی‌ها حاوی اطلاعات دقیق‌تری از صحبت هر گوینده است، بنابراین شبکه در هر پنج بار تعلیم، به نحو مناسبی همگرا می‌شود و میانگین نرخ تشخیص در کنار بهترین نرخ افزایش می‌یابد. روند مشابهی در مورد نرخ تشخیص قطعه گفتاری وجود دارد. علیرغم آنکه بهترین نرخ تشخیص قطعه گفتاری در ردیف دومِ جدول از ردیف آخر اندکی بهتر است، سه معیار تشخیصی دیگر در حالت ترکیب ویژگی‌ها به طرز محسوسی نسبت به سایر روش‌ها بالاتر هستند.

به منظور بررسی میزان مقاومت به نویز در روش پیشنهادی، در جدول (۴) معیارهای تشخیص در نسبت سیگنال به نویز ۲- دسی‌بل نشان داده شده است. بر خلاف جدول (۳)، در اینجا نرخ تشخیص به ازای ترکیب MFCC و انرژی به شدت افت کرده و به کمترین مقدار رسیده است. دلیل این مسئله می‌تواند حساسیت زیاد انرژی به نویز باشد. بنابراین روند کاملاً مشابهی برای این دسته از ویژگی‌ها در حالت بدون نویز و نویزی مشاهده نمی‌شود و در شرایط نویزی مناسب‌تر است که بردار ویژگی شامل کمیت انرژی نباشد. ترکیب پیشنهادیِ ضرایب بدون انرژی در حالت نویزی بهترین نرخ تشخیص گوینده را نسبت به سایر روش‌ها به طرز محسوسی افزایش داده است.

با توجه به مقاوم بودن ویژگی آنتروپی و برای تحلیل بهتر نتایج، دو ردیف دیگر به این جدول اضافه شده است که معیارهای تشخیص را بر روی ویژگی آنتروپی و سپس ترکیب آن با فرکانس پایه و فرمنت‌ها نشان می‌دهد. نتایج به دست آمده نشان می‌دهد که تقطیع جملات و ترکیب ویژگی‌های مبتنی بر فریم با ویژگی‌های مبتنی بر قطعات گفتاری که در اینجا پیشنهاد شده است باعث بهبود عملکرد سیستم می‌شود که بیانگر مؤثر بودن تقطیع جملات و استفاده از واحدهای کوچک‌تر از جمله است. میانگین تشخیص قطعه صوتی با استفاده از روش پیشنهادی تا ۵۹/۷۰٪ بهبود یافته است. علاوه بر آن به ازای تمامی معیارها استفاده از روش پیشنهادی نرخ تشخیص را بهبود داده است.

### ۶-۴- تأثیر افزایش تعداد ضرایبِ MFCC بر عملکرد سیستم پیشنهادی

در کارهای گذشته اشاره شده است که بخش عمده اطلاعات گوینده در ضرایب مرتبه پایین MFCC قرار دارند و استفاده از تعداد زیادی از ضرایب منجر به بهبود نرخ تشخیص نمی‌شود [۱۸]. در این آزمایش به بررسی تأثیر افزایش تعداد ضرایب MFCC بر روی نرخ تشخیص پرداخته شده است. تعداد ضرایب به تدریج از ۱۲ افزایش داده شده است و هر چهار معیار مرتبط با نرخ تشخیص محاسبه شده است. نتایج این آزمایشات در دو حالت بدون نویز و به ازای نسبت سیگنال به نویز ۲- دسی‌بل به ترتیب در جداول (۵) و (۶) آمده است. همان طور که مشخص است با افزایش تعداد ضرایب از ۱۵ به ۱۶، کلیه معیارهای نرخ تشخیص افت می‌کنند. بنابراین افزایش بیش از این تعداد منجر به بهبود نرخ تشخیص نمی‌شود.

بررسی جدول (۵) نشان می‌دهد در حالت بدون نویز بهبود تدریجی در سه معیار تشخیص به ازای افزایش تعداد ضرایب تا ۱۵ ضریب به وجود می‌آید. بهترین نرخ تشخیص گوینده در همه حالات یکسان است که البته برآورد دقیقی از عملکرد سیستم نیست. به ازای ۱۶ ضریب عملکرد سیستم افت می‌کند و بنابراین بهترین نرخ تشخیص مربوط به ۱۵ ضریب است. این روند به طرز محسوس‌تری در جدول (۶) به ازای نسبت سیگنال به نویز ۲- دسی‌بل مشاهده می‌شود. با اینکه در حالت نویزی نتایج روال منظمِ جدول (۵) را ندارند، اما همچنان بهترین نتایج به ازای ۱۵ ضریب به دست می‌آید.

Table (5): Recognition results with gradually increased number of MFCC coefficients in noise-free condition

جدول (۵): تأثیر افزایش تعداد ضرایب MFCC بر روی نرخ تشخیص در حالت بدون نویز

| تعداد ضرایب MFCC | بهترین نرخ تشخیص گوینده | میانگین نرخ تشخیص گوینده | بهترین نرخ تشخیص قطعه گفتاری | میانگین نرخ تشخیص قطعه گفتاری |
|---|---|---|---|---|
| 12 | 86.66 | 62.66 | 80.88 | 63.08 |
| 13 | 86.66 | 73.33 | 78.67 | 68.67 |
| 14 | 86.66 | 77.33 | 78.67 | 70.29 |
| 15 | 86.66 | 78.66 | 80.88 | 73.23 |
| 16 | 86.66 | 74.66 | 80.14 | 68.97 |

Table (6): Recognition results for naive and hybrid feature vectors with increased number of MFCC coefficients in -2dB SNR

جدول (۶): تأثیر افزایش تعداد ضرایب MFCC در حالت ترکیبی و غیرترکیبی در نسبت سیگنال به نویز ۲- دسی‌بل

| بردار ویژگی | بهترین نرخ تشخیص گوینده | میانگین نرخ تشخیص گوینده | بهترین نرخ تشخیص قطعه گفتاری | میانگین نرخ تشخیص قطعه گفتاری |
|---|---|---|---|---|
| ۱۲ ضریب MFCC | 80 | 54.66 | 65.44 | 53.38 |
| ۱۳ ضریب MFCC | 73.33 | 54.66 | 66.17 | 51.47 |
| ۱۴ ضریب MFCC | 40 | 32 | 43.38 | 34.70 |
| ۱۵ ضریب MFCC | 80 | 56 | 73.52 | 53.23 |
| ۱۶ ضریب MFCC | 60 | 37.33 | 52.20 | 37.05 |
| ۱۵ ضریبMFCC+ فرکانس پایه+ فرمنت+ آنتروپی | 60 | 50.66 | 55.88 | 53.52 |

Table (7): Performance of speaker identification system for three sets of feature vectors

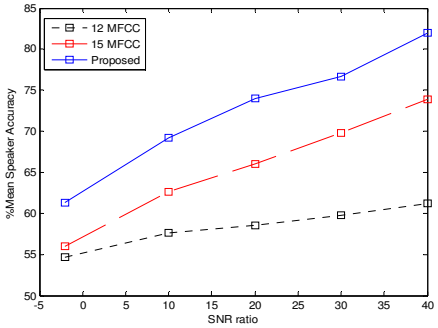
جدول (۷): عملکرد سیستم تشخیص گوینده به ازای سه دسته از بردارهای ویژگی

| بردارهای ویژگی | SNR | میانگین نرخ تشخیص گوینده | میانگین نرخ تشخیص قطعه صوتی |
|---|---|---|---|
| روش پیشنهادی | 40 | 82 | 74 |
| | 30 | 76.66 | 70.58 |
| | 20 | 74 | 67.52 |
| | 10 | 69.23 | 65 |
| | -2 | 61.33 | 59.70 |
| ۱۵ ضریب MFCC | 40 | 73.9 | 68 |
| | 30 | 69.80 | 65.23 |
| | 20 | 66 | 61.88 |
| | 10 | 62.66 | 57.7 |
| | -2 | 56 | 53.23 |
| ۱۲ ضریب MFCC | 40 | 61.20 | 61.77 |
| | 30 | 59.8 | 60.01 |
| | 20 | 58.50 | 58 |
| | 10 | 57.66 | 56.47 |
| | -2 | 54.66 | 53.38 |

با توجه به این نتایج، در مرحله بعدی ترکیب ویژگی‌های پیشنهادی این بار با ۱۵ ضریب MFCC در حالت نویزی مورد آزمایش قرار گرفته است. نتیجه در جدول (۶) آمده است. مقایسه این نتیجه با نتایج جدول (۴) نشان می‌دهد که هرچند استفاده از ۱۵ ضریب MFCC به تنهایی نتایج قابل قبولی دارد اما همچنان در حالت نویزی بهترین نتایج به ازای ترکیب ۱۲ ضریب با فرکانس پایه، فرمنت‌ها و ضرایب آنتروپی به دست می‌آید. دلیل این مسئله می‌تواند افزایش بیش از حد ابعاد بردار ویژگی باشد که باعث می‌شود شبکه توانایی یادگیری و همگرایی مناسب را از دست بدهد. به عبارت دیگر فرمنت‌ها و ضرایب آنتروپی در گره‌های فرکانس بالا می‌توانند حاوی همان اطلاعاتی باشند که با افزایش ضرایبِ MFCC به ویژگی‌ها اضافه می‌شود. در این حالت استفاده از ۱۵ ضریب به تنهایی، تنها در یکی از معیارها منجر به بهبود اندکی نسبت به روش پیشنهادی شده است.

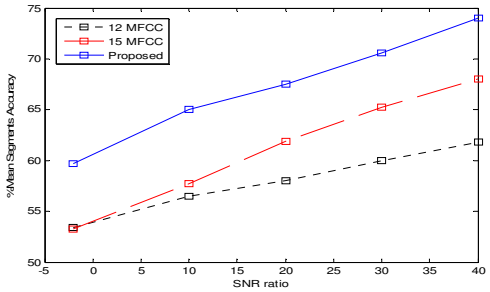
### ۶-۵- بررسی میزان مقاومت به نویز در سیستم پیشنهادی

در انتها نحوه عملکرد سه دسته از بردارهای ویژگی که کارآیی بهتری دارند، به ازای مقادیر مختلف از نسبت سیگنال به نویز مورد آزمایش قرار گرفته و نتایج در جدول (۷) نشان داده شده است. شکل (۶) میانگین نرخ تشخیص گوینده را به ازای نسبت سیگنال به نویز ۲- تا ۴۰ دسی‌بل نشان می‌دهد. نمودار نقطه‌چین مربوط به ۱۲ ضریبِ MFCC، نمودار خط‌چین مربوط به ۱۵ ضریبِ MFCC و خط پر نشان دهنده روش پیشنهادی یعنی حالتی است که ۱۲ ضریبِ MFCC با فرکانس پایه، فرمنت‌ها و آنتروپی ترکیب شده‌اند. همان طور که از شکل مشخص است روش پیشنهادی به ازای مقادیر مختلف از نسبت سیگنال به نویز دارای نرخ تشخیص بهتر و افت کمتری است. بنابراین نسبت به نویز مقاوم‌تر است. شکل (۷) نتایج آزمایشات مشابه را برای معیار میانگین نرخ تشخیص قطعه گفتاری نشان می‌دهد که در این شکل نیز روش پیشنهادی دارای برتری نسبت به دو روش دیگر است.




شکل (۶): میانگین نرخ تشخیص گوینده به ازای سه دسته از بردارهای ویژگی در نسبت‌های سیگنال به نویز مختلف

Fig. (6): Average speaker recognition rate for three sets of feature vectors in different signal to noise ratios




شکل (۷): میانگین نرخ تشخیص قطعه صوتی به ازای سه دسته از بردارهای ویژگی در نسبت‌های سیگنال به نویز مختلف

Fig. (7): Average speech segment recognition rate for three sets of feature vectors in different signal to noise ratios

## ۷- نتیجه گیری

در این مقاله از ترکیب ویژگی‌های مبتنی بر فریم‌بندی با ویژگی‌های مستقل از متنِ استخراج شده از یک بخش گفتاری استفاده شد تا تأثیر نویز بر کارآیی سیستم تشخیص هویت گوینده مستقل از متن کاهش یابد. به جای استفاده از روش‌های متداول مثل میانگین‌گیری، برای کاهش ابعاد ویژگی‌های زمان کوتاهِ مبتنی بر فریم از خوشه‌بندی Kmeans استفاده شد. نتایج آزمایشات کارآیی روش پیشنهادی را هم

در حالت بدون نویز و هم به ازای مقادیر مختلف از نسبت‌های سیگنال به نویز نشان می‌دهند. با توجه به نتایج به دست آمده در حالت نویزی با نسبت سیگنال به نویز ۲- دسی بل بهترین نتایج به ازای ترکیب ۱۲ ضریب با فرکانس پایه، فرمنت‌ها و ضرایب آنتروپی به دست آمد که این ترکیب از ویژگی‌ها میانگین نرخ تشخیص گوینده را به میزان ۶/۶۷٪ و میانگین نرخ تشخیص قطعه گفتاری را به میزان ۶/۳۲٪ نسبت به روش متداول مبتنی بر ۱۲ ضریب MFCC افزایش داده است. از آنجایی که در این مقاله به جای استفاده از فریم از قطعات گفتاری استفاده شده است، تعداد واحدهای گفتاری به شدت کاهش می‌یابد. بنابراین ترکیب ویژگی‌ها و افزایش بعد بردار ویژگی حجم دادگان آموزش و تست را افزایش نمی‌دهد. با این حال پیشنهاد می‌شود در مراحل بعدی به جای تکیه بر تقطیع دستی، از تقطیع اتوماتیک دادگان استفاده شود. با روشی مناسب و مقاوم به نویز می‌توان واحدهای چند آوایی و یا هجایی را در جملات جدا کرده و سپس ویژگی‌های مستقل از متن را استخراج کرد.

**پی‌نوشت:**

1- MFCC
2- WPT
3- DWT
4- Open-Set
5- Closed-Set
6- Phonetic context
7- MFCC
8- Frame Based
9- Approximation
10- Detail
11- Wovels
12- Power spectral density
13- Auto-Regressive
14- Addaptive
15- PRAAT

## References


[1] R. ShanthaSelvaKumari, S. SelvaNidhyananthan, G. Anand, "Fused Mel feature sets based text-independent speaker identification using Gaussian mixture model", Procedia Engineering, Vol. 30, pp. 319-326, 2012.

[2] K. Daqrouq, K.Y. Al Azzawi, "Average framing linear prediction coding with wavelet transform for text-independent speaker identification system", Computers & Electrical Engineering, Vol. 38, No. 6, pp. 1467-1479, Nov. 2012.

[3] A. Shafik, S.M. Elhalafawy, S.M. Diab, B.M. Sallam, F.E. Abd El-samie, "A wavelet based approach for speaker identification from degraded speech", International Journal of Communication Networks and Information Security (IJCNIS), Vol. 1, No. 3, Dec. 2009.

[4] M.I. Abdalla, S.A. Hanaa, "Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov models", JOURNAL OF TELECOMMUNICATIONS, Vol. 1, No 2, March 2010.

[5] K. Daqrouq, "Wavelet entropy and neural network for text-independent speaker identification", Engineering Applications of Artificial Intelligence, Vol. 24, No 5, pp. 796–802, Aug. 2011.

[6] Md. Murad Hossain, B. Ahmed, M. Asrafi, "A real time speaker identification using artificial neural network", 10th international conference on computer and information technology, iccit, pp.1-5, 27-29 Dec. 2007.

[7] E. Avci, "A new optimum feature extraction and classification method for speaker recognition: GWPNN ", Expert Systems with Applications, Vol. 32, No. 2, pp. 485–498, Feb. 2007.

[8] H. Harb, C. Liming, "Gender identification using a general audio classifier", Proceeding of the IEEE/ICME, Vol. 2, pp. II-733-736, July 2003.

[9] H. Harb, L. Chen, "Voice-based gender identification in multimedia applications", Journal of Intelligent Information Systems, Vol. 24, No. 2-3, pp. 179-198, March 2005.

[10] J.A. Bachorowski, M.J. Owren, "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech", Journal of the Acoustical Society of America, Vol. 106, No. 2, pp. 1054–1063, Aug. 1999.

[11] A. Cherif, L. Bouafif, T. Dabbabi, "Pitch detection and formants analysis of arabic speech processing", Applied Acoustcs, Vol. 62, No. 10, pp. 1129–1140, Oct. 2001.

[12] A.M. Noll, "Cepstrum pitch determination", Journal of the Acoustical Society of America, Vol. 41, pp. 293-309, 1967.

[13] W. Yutai, L. Bo, J. Xiaoqing, L. Feng, W. Lihao, "Speaker recognition based on dynamic MFCC parameters", Proceeding of the IEEE/IASP, pp. 406-409, April 2009.

[14] S. Chougule, P.P. Rege, "Language independent speaker identification", Proceeding of the IEEE/ICIT, pp. 364-368, 15-17 Dec. 2006.

[15] S. Haykin, "Neural networks", Macmillan College Publishing Company, Section 5.3: The Steepest Descent Method, 1994.

[16] M. Katz, "Fractals and the analysis of waveforms", Computers in Biology and Medicine, Vol. 18, No. 3, pp. 145-156, 1988.

[17] J.D. Wu, B.F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition", Expert Systems with Applications, Vol. 36, No. 2, pp. 3136–3143, March 2009.

[18] S. Pandiaraj, H.N.R. Keziah, D.S. Vinothini, L. Gloria, "A confidence measure based – score fusion technique to integrate MFCC and Pitch for speaker verification", Proceeding of the IEEE/ICECT, Vol. 3, pp. 317-320, April 2011.