
بسم الله الرحمن الرحيم

دانشکده فنی و مهندسی

پایان نامه کارشناسی ارشد مهندسی برق الکترونیک

عنوان :

تأیید هویّت با استفاده از پردازش سیگنال گفتار

استاد راهنما :

دکتر ساجدی

نگارش :

سید مصطفی موسوی بورا

آبان ماه ۱۳۸۹

تقدیم به

پدر و مادرم

آن آموزگاران نخستین

و

تمامی آموزگارانم

تقدیم به

همسر مهربانم

که در انجام مراحل مختلف پروژه مرا یاری نمود

تقدیر و سپاس :

در انجام مراحل مختلف این پروژه از راهنماییهای ارزشمند و حمایت بی‌دریغ استاد ارجمند و گرامی جناب دکتر ساجدی که همواره تعیین کننده بستر حرکت، مسیر پیشرفت کار و مشوق من بوده است، بر خود لازم می دانم که بدین وسیله زحمات ایشان را ارج نهم و صمیمانه تشکر و قدردانی نمایم.

چکیده

یکی از روشهای تأیید و تعیین هویّت افراد، استفاده از صدای فرد می باشد که مقولهٔ شناسایی گوینده و تصدیق و تعیین هویت گوینده نام دارد. از جمله روشهایی که در پردازش گفتار و بطور خاص در شناسایی گوینده کارآیی بسیار خوبی از خود نشان داده است، مدل چندی‌سازی برداری و مدل شبکه عصبی می‌باشد که وسیله‌ای بسیار قوی برای مدل‌کردن سیگنال‌های تصادفی و ایستا می‌باشند. در این پروژه ما از مدل چندی‌سازی برداری و هم از مدل شبکه عصبی از نوع پرسپترون چند لایه برای تصدیق و تعیین هویت گوینده‌های از روستای بالابورا از توابع منطقه بندپی‌غربی شهرستان بابل که می‌توانند به سه زبان انگلیسی،فارسی و زبان مازندرانی(مازنی) صحبت کنند، بکار گرفته شده است. با توجه به تمهیدات بکار گرفته شده در این پروژه، بر روی سه پایگاه داده از ارقام برای هر سه زبان متشکل از ۲۰ نفر (۱۲ مرد و ۸زن) که در سنین مختلف ۱۲ تا ۶۱ سال بوده‌اند و ارقام صفر تا نُه را بیان کرده‌اند، کارآیی سیستم شناسایی گوینده طراحی شده در این پروژه برای تعیین هویت گوینده در هر کدام از این سه زبان مشخص شده است. بطوریکه نتایج حاصل نشان می‌دهد که درصد شناسایی این سیستم توسط این دو روش در هر سه زبان متفاوت بوده، بطوریکه در مرحله اول دقت سیستم توسط روش VQ تأیید گوینده به زبانهای انگلیسی، فارسی و مازنی بترتیب با MFCC۲۶ برابر با % 94.7، % 92.6 و % 91.3 شده است که در مرحله دوم با استفاده از یک نوع شبکه عصبی پیشنهادی توسطLPCC۲۶ بترتیب برابر با 99.6%، 98.7% و 97.9% بهبود یافته است.

در مراحل مختلف این پروژه تأثیر تعداد دفعات ارائه داده های آموزشی به شبکه عصبی، تاثیر تعداد ویژگیها و نوع ویژگی از نقطه نظر ایستا و گذرا بودن و نیز تاثیر ایجاد تغییرات در پارامترهای یادگیری شبکه، مورد بررسی قرار گرفت. همچنین طی آزمایشهایی کارائی سیستم تصدیق هویت پیشنهاد شده ارزیابی و نتایج حاصل از بازشناسی ارقام و تصدیق هویت با استفاده از روشهای گفته شده با نتایج حاصل از روش کلاسیک چندی سازی برداری مقایسه گردید.

واژه‌های کلیدی :

SPEAKER VERIFICATION–NEURAL NETWORKM–VECTOR QUANTIZATION

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل ۱

مقدمه

۱-۱- تعریف مسئله

یکی از مباحث مطرح در پردازش سیگنال گفتار تعیین و یا تصدیق هویّت گوینده بر اساس سیگنال گفتار می‌باشد. در تعیین گوینده هویّت فرد مورد نظر بعنوان عضوی از گوینده‌های شناخته شده سیستم شناسایی می‌گردد و در تصدیق گوینده ادعای هویّت گوینده بعنوان یکی از گوینده‌های ناشناخته شده سیستم، تصدیق یا رد می‌شود. کاربردهای بسیاری برای بازشناسی گوینده وجود دارد که شامل کنترل دسترسی به اطلاعات محرمانه توسط صدا، ارائه اطلاعات و سرویس‌های شخصی توسط صدا نظیر نامه‌نگاری و خرید تلفنی، برچسب‌دهی گوینده‌ها در گفتگوهای ضبط شده، تحقیقات دادگاهی و جستجوی مجرمین با نمونه صداهای ضبط شده و .... می‌باشد.

بازشناسی گوینده جزء روش‌های تشخیص هویت شخص می‌باشد که بر اساس ویژگی‌های فیزیولوژیکی یا رفتاری شخص عمل می‌کنند و عبارتست از فرآیند خودکار شناسایی شخص گوینده با استفاده از اطلاعات فردی او که از سیگنال گفتار آن شخص استخراج می‌شود. تأیید هویّت گوینده ارتباط تنگاتنگی با تشخیص گوینده دارد که در آن یک فرد با هویّت مجهول ادعای داشتن هویّت یکی از اعضای مجموعه را می‌کند. داده‌های ورودی از گفتار این گوینده تنها با الگوهای گفتار همان عضو از مجموعه گویندگان مقایسه می‌شود و در صورتی که فاصله داده‌های ورودی از الگوی مرجع از یک

آستانه کمتر باشد هویّت فرد تأیید و در غیر اینصورت رد می‌شود. همچنین بازشناسی گوینده برای کنترل دسترسی از راه دور مناسبتر است زیرا صدا از طریق خطوط تلفن قابل انتقال است.

۱-۲- سیستم شنوایی در انسان

گفتار تولید شده توسط اندام‌های گویایی بصورت یک موج صوتی از طریق هوا منتقل و به گوش شنونده می‌رسد. این موج وارد گوش شده و باعث تحریک پردهٔ صماخ می‌گردد. این ارتعاشات از طریق استخوانهای لایه گوش به درون حلزونی منتقل شده و باعث تحریک سلسله اعصاب شنوای که به حلزونی متصل هستند، می‌گردد. در عمل حلزونی به صورت یک بانک فیلتر در محدودهٔ HZ ۲۰۰۰۰ - ۲۰ عمل می‌کند[1]. برخی از خصوصیات گوش عبارتند از :

الف – منحنی Mel

فرکانس یک دانگ صدای[1] خاص که توسط گوش دریافت می‌شود با فرکانس واقعی آن برابر نیست و رابطهٔ خطی میان این دو فرکانس وجود ندارد. منحنی mel رابطهٔ میان این دو فرکانس را نشان می‌دهد. این منحنی در فرکانس‌های کمتر از ۱کیلوهرتز تقریباً خطی است و بالای آن بصورت لگاریتمی افزایش می‌یابد[1],[2].

ب) فیلترهای باند بحرانی[2]

گوش دارای خاصیت حذف نویز در پهنای باند خاصی حول یک فرکانس می‌باشد. این پهنای باند برای فرکانس‌های مختلف مقدار متفاوتی دارد که برای فرکانس‌های کمتر از ۱کیلوهرتز تقریباً برابر ۱۰۰ هرتز و برای فرکانس‌های بالاتر بصورت لگاریتمی افزایش می‌یابد. پهنای باند بحرانی بدین معنی است که اگر انرژی نویز در این باند حول فرکانس f_0 بزرگتر از مقدار نباشد آن فرکانس توسط گوش قابل تشخیص خواهد بود و در غیر اینصورت آن فرکانس قابل شناسایی توسط گوش نیست[1],[2].

[1] Tone
[2] Critical band filters

ج) منحنی‌های بلندی هم تراز[1]

میـزان حساسـیت گـوش بـه فرکـانس‌هـای مختلـف صـدا یکسـان نیسـت. بیشـترین حساسـیت در فرکـانس ۱کیلوهرتز است[2]، که از این خاصیت در استخراج بعضی از پارامترها استفاده می‌شود.

د) قانون توان شنوایی[2]

شدت صوتی که توسط گـوش شـنیده مـی‌شـود بـا تـوان واقعـی سـیگنال برابـر نیسـت بلکـه یـک رابطـهٔ غیـر خطـی میـان ایـن دو وجـود دارد. ایـن رابطـه بـه صـورت ریشـهٔ سـوم دامنـهٔ سـیگنال مـدل مـی‌شـود. از خصوصیات فـوق در اسـتخراج پارامترهـای ادراکـی[3] از سـیگنال گفتـار اسـتفاده مـی‌شـود کـه مهمتـرین آنهـا پارامترهای Mel-Cepstral و PLP هستند[2],[3].

۱-۳-واحدهای گفتار برای بازنمایی گوینده‌ها

یکی از مهمترین ملاحظات در طراحی سیستم بازشناسی گوینده‌ها انتخاب واحدهای گفتاری برای مـدل کـردن گوینده‌ها است. انتخاب واحدها شامل واحدهای آوایی یا زبانشناختی به عنوان جملات کامل، کلمـات و هجاهـا و واحدهای شبه آوا و نیز واحدهای صوتی نظیر واحدهای جدا شده از گفتار بر اساس مـدل‌هـای صـوتی بـه جـای معیارهای آوایی می‌باشد.

برای بازشناسی گوینده ضرورتی برای بازنمایی گفتار بصورت واحدهای واجی نیست و چون سیسـتم‌هـای کـه از واحدهای واجی استفاده می کنند پیچیده تر ار سیستم‌هایی هستند که تنها از مدل‌های صوتی استفاده می‌کنند. لذا بیشتر سیستم‌های بازشناسی گوینده از نوع دوم هستند[4] .

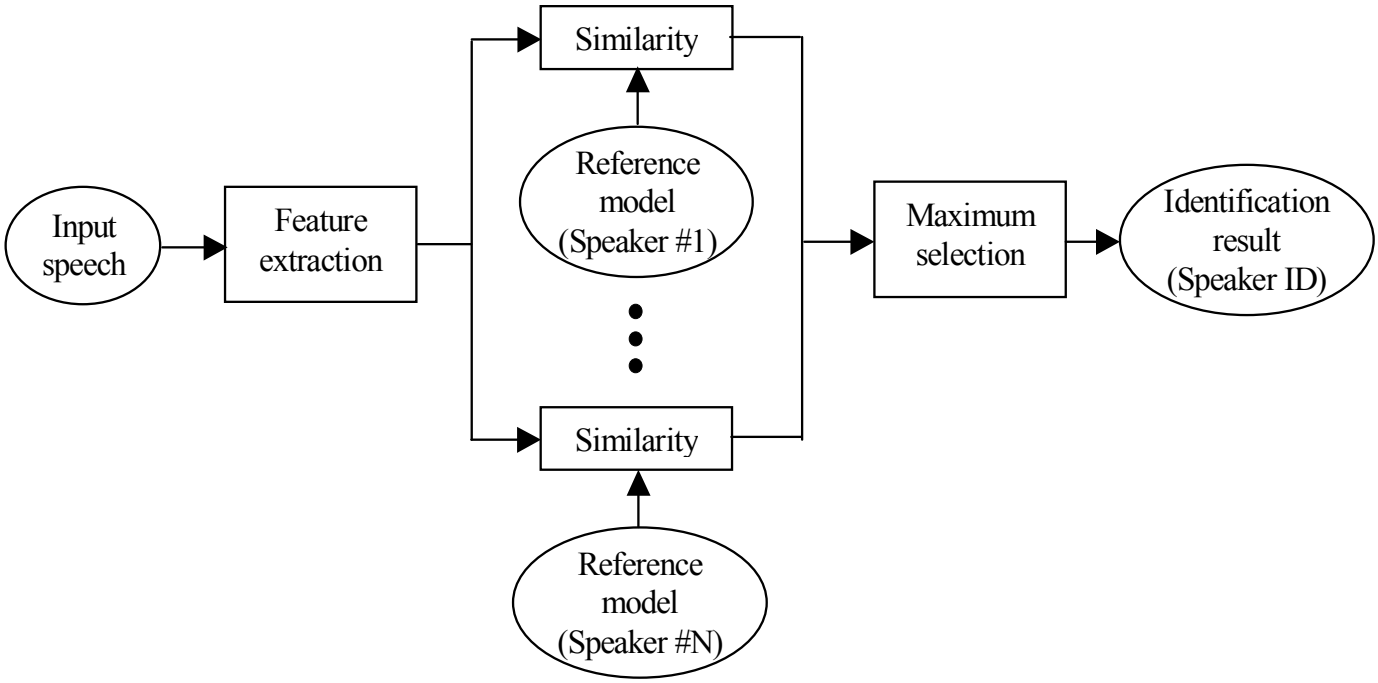
[1] Equal Loudness curves
[2] Power loudness curves
[3] perceptual

۱-۴- مسائل مطرح در بازشناسی هویّت گوینده

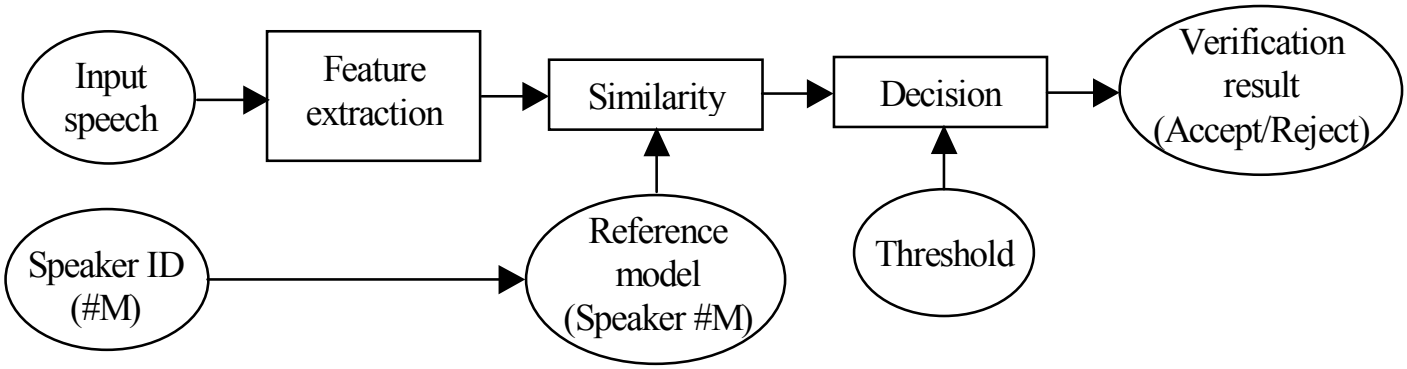
هر عبارتی که توسط شخصی بیان می‌شود حداقل دو نوع اطلاعات را در بر دارد که عبارتند از اطلاعات خود پیام و اطلاعاتی در مورد گوینده پیام می‌باشد. عناصر اصلی یک سیستم بازشناسی هویّت گوینده در شکل (۱-۲) نشان داده شده است. در تعیین هویّت گوینده[۱] سیگنال گوینده ناشناس آنالیز شده و با مدل گوینده‌های شناخته شده مقایسه می‌گردد. گوینده ناشناس بعنوان گوینده‌ای که مدلش بیشترین تطابق را با سیگنال گفتار ورودی دارد معرفی می‌گردد. در مورد تشخیص مجموعهٔ بسته[۲] تصمیمات ممکن برابر تعداد گوینده‌های شناخته شده است ولی در مورد تشخیص مجموعهٔ باز[۳] یک انتخاب دیگر تحت عنوان عدم تطا بق سیگنال گفتار با مدل-های شناخته شده نیز وجود خواهد داشت.

در تصدیق گوینده[۴]، ادعای هویت برای یک سیگنال گفتار بعنوان یکی از مدل‌های شناخته شده، تصدیق و یاد رد می‌شود. در این روش گفتار شخص ناشناس یا مدل گوینده ادعا شده مقایسه می‌گردد و اگر تطابق به اندازه کافی باشد ادعای مزبور تصدیق می‌گردد[5].در شکل صفحه بعد بلوک دیاگرام مربوط به سیستم تعیین و تأیید گوینده نمایش داده شده است.

[1] Speaker Identification
[2] Close Set
[3] Open Set
[4] Speaker Verification



الف) بلوک دیاگرام تعیین گوینده



ب) بلوک دیاگرام تأیید گوینده

شکل (۱-۱) عناصر اصلی سیستم بازشناسی گوینده[5] : الف) تعیین گوینده
ب) تأیید گوینده

۱-۵- روش‌های کلّی بازشناسی گوینده

سیستم‌های بازشناسی گوینده به دو روش وابسته به متن[1] و مستقل از متن[2] عمل می‌کنند. در حالت وابسته به متن، گوینده باید همان جملاتی که برای یادگیری سیستم استفاده شده است را ادا کند ولی در حالت مستقل از متن، گوینده می‌تواند هر جمله‌ای را در زمان تشخیص ادا نماید. از آنجائیکه در حالت وابسته به متن می توان ویژگی‌های مربوط به هر واج یا هجا را بررسی کرد لذا دقت‌های بالاتری نسبت به حالت مستقل از متن بدست می‌آید[6].

الف) وابسته به متن

از گوینده خواسته می‌شود که کلمه عبور را ادا نماید و این سیگنال با مدل‌های شناخته شدهٔ گوینده‌ها مقایسه و تصمیم گیری گرفته می‌شود.

ب) مستقل از متن

در بسیاری از کاربردها کلمهٔ عبور از پیش تعیین شده، قابل استفاده نیست بلکه سیستم باید توانایی شناسایی گوینده با هر گفتاری را داشته باشد. بعلت کاربردهای فراوان این نوع تشخیص بیشتر مورد توجه است.

ج) وابسته به متن با کلمات عبوری اتفاقی

هر دو روش قبل به آسانی قابل شکست هستند چون اگر صدای گوینده ضبط و دوباره پخش شود بعنوان یکی از گوینده‌ها پذیرفته خواهد شد. برای جلوگیری از این کار یک مجموعه از کلمات نظیر اعداد به عنوان رمز عبور استفاده می‌شوند. و هر بار از کاربر خواسته می‌شود که یک توالی خاصی از این کلمات را ادا نماید. اخیراً یک چنین سیستمی ساخته شده است که کلمات عبور هر بار بطور کامل عوض می‌شود[7].

[1] Text Dependent
[2] Text Independent

۱-۶- بازنمایی مدل‌های گوینده‌ها

معمولترین روش برای بازشناسی خودکار گوینده در حالت وابسته به متن ذخیرهٔ ویژگیهای مقطعی است. در این روش هر گوینده بصورت متوالی از بردارهای ویژگی جملهٔ ادا شده مدل می‌شود و تصمیم بر اساس تطابق الگوها[1] صورت می‌پذیرد.

۱-۷- معیار شباهت ویژگی‌ها

مهمترین وجه تمایز میان سیستم‌های بازشناسی گوینده نحوهٔ ترکیب پارامترها و مقایسه آنها با مدل‌های مرجع می‌باشد. مقایسه به کمک یک معیار شباهت[2] یا معیار فاصله[3] انجام می‌گیرد و با مقایسه ویژگی‌های گویندهٔ ناشناس با مدل‌های شناخته شده نزدیکترین مدل به عنوان گوینده برنده انتخاب می‌شود.

فاصله میان دو بردار V′ و V بطول L بصورت : $\mathrm{d} = \sum_{i=1}^{l} |\mathrm{Vi} - V'i|$

یا فاصله اقلیدوسی : $d^2 = \sum_{i=1}^{l} (|V_\mathrm{i} - V'_i|)^2$

یا فاصله وزن دار : $d^2 = \sum_{i=1}^{l} (Wi(|V_\mathrm{i} - V'_i|))^2$

که در آن $W_i=1/\sigma_i$ که σ_i تخمینی از واریانس ضریب iام بردار ویژگی می‌باشد.

۱-۸- مقاوم[4]سازی سیستم در برابر تغییرات

تغییرات در تکرارهای مختلف عامل مهمی در کاهش کارایی سیستم‌های بازشناسی گوینده می‌باشد. این تغییرات ناشی از تغییرات در خود شخص، تغییرات شرایط ضبط وانتقال و نویزی می‌باشد. مطالعه نشان داده است که جملات ضبط شده در یک جلسه همبستگی بیشتری نسبت به جملات ضبط شده در جلسات دیگر دارند لذا مقابله با این تفاوت ضروری است[8],[9]. عملاً دو روش برای حل این معضل وجود داردکه عبارتند از :

[1] Template
[2] Liklyhod Measure
[3] Distance Measure

[4] Robust

الف) به روز کردن مدل‌ها در هر بار استفاده

در این روش از جملاتی که در مرحلهٔ بازشناسی توسط گوینده ادا می‌شود نیز برای یادگیری سیستم استفاده می‌شود. شرایط یادگیری شامل تعداد جلسات یادگیری، تعداد جملات و شرایط ضبط و انتقال گفتار می‌شود. تغییرات در این شرایط در زمان بازشناسی کارایی سیستم را به شدت تحت تأثیر قرار می‌دهد[9].

ب) هنجارسازی[1] ویژگی‌های صوتی

یکی از روش‌های هنجارسازی پارامترها استفاده از روش CMN[2] است که اثرات خطی کانال و تغییرات بلند مدت طیفی را حذف می‌کند. CMS[3] یکی از روش‌های فوق است که در آن ضرایب کپسترال در طول یک توالی از گفتار میانگین گیری شده و مقدار میانگین از ضرایب کپسترال در طول یک توالی از گفتار میانگین‌گیری شده و مقدار میانگین از ضرایب کپسترال هر فریم کم می‌شود. این روش اثرات خطی کانال را به خوبی حذف می‌کند ولی بعضی از اطلاعات مفید را هم از بین می‌برد لذا برای گفتار کوتاه مدت مناسب نیست[9].

۱-۹- محاسبه آستانه

پس از اینکه در مرحله آموزش الگوهای مرجع را تولید نمودیم می‌بایست معیاری برای تصمیم گیری در مرحله تشخیص(تست) در دست داشته باشیم که به آن آستانه می‌گوییم که نحوه تعیین آن به شرح زیر است :
چون تطبیق الگوها بر اساس کلمات مجزا صورت می‌گیرد، بنابراین می‌بایست مقدار آستانه برای هر کلمه بصورت جداگانه تعیین گردد. مسأله دیگر، نحوه تعیین مقدار آستانه مربوط به هر کلمه می‌باشد که با استفاده از یکی از دو روش زیر صورت می‌گیرد :

الف) برای یک گوینده(Intraspeaker)

ب) برای مجموعه‌ای از گوینده‌گان(Interspeaker)

در حالت اول برای هرکلمه‌ای یک گوینده خاص ادا می‌کند یک آستانه بدست می‌آید ولی در حالت دوم برای هر

[1] Normalization
[2] Cepstral Mean Normalization
[3] Cepstral Mean Subtraction

کلمه‌ای که تمام گوینده‌گان ادا می‌کنند یک آستانه محاسبه می‌گردد. در یک پایگاه داده که برای n نفر تهیه شده باشد که در آن هر گوینده m کلمه را بیان کرده باشد در حالت اوّل لازمست m.n آستانه، ولی در حالت دوّم کافی است که فقط m آستانه محاسبه گردد.

از آنجاییکه در این پروژه از حالت دوم برای تعیین آستانه استفاده شده است، حال به شرح روش‌های مختلفی برای تعیین آستانه می‌پردازیم. معیارهای مختلفی از جمله معیارهای ZFR[1]، ZFA[2] و EER[3] برای تعیین آستانه بکار گرفته شده است که ما در این پروژه از معیارEER استفاده نموده‌ایم و به شرح آن می‌پردازیم.

۱-۹-۱- معیارEER

دو معیار قبلی نگرش بسیار سخت گیرانه‌ای نسبت به پذیرش[4] یا رد[5] یک گوینده ایجاد می‌کنند، بنابراین دیدگاه معتدلتری که در این پروژه مورد استفاده قرار گرفته است معیار EER می‌باشد که در آن هدف این است که آستانه طوری تعیین شود که پذیرش نادرست[6] و ردّ نادرست[7] با یکدیگر برابر شوند.

بعلت اینکه تعداد نمونه‌های مورد بررسی محدود است عملاً امکان رسیدن به احتمال مساوی برای خطاهای فوق کم است به همین دلیل از تخمین رابطهٔ(۱-۳) استفاده می‌کنیم :

$$EE^{8} = \frac{FA+FR}{2} \qquad (۳-۱)$$

با استفاده از مقادیر T_{min} و T_{max} و فاصله بدست آمده از مقایسه دو کلمه مقدار آستانه T طوری تعیین می‌گردد که احتمال‌های پذیرش نادرست و ردّ نادرست تا حد امکان به یکدیگر نزدیک شوند. شکل(۱-۲) نحوه تغییر پارامترهای FA و FR را نسبت به یکدیگر نشان می‌دهد.

[1] Zero False Rejection
[2] Zero False Acceptance
[3] Equal Error Rate
[4] Acceptance
[5] Rejection
[6] False Acceptance
[7] False Rejection
[8] Equal Error