ارائه یک دسته‌بند مقاوم به منظور بازشناسی گفتار مبتنی بر هم‌افزایی خوشه‌بندی و فراوانی مشاهدات

محمد مصلح[1]، محمد خیراندیش[2]،مهدی مصلح [3]، نجمه حسین‌پور[4]
[1]گروه مهندسی کامپیوتر، واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران، Mosleh@iaud.ac.ir
[2] گروه مهندسی کامپیوتر، واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران
[3] گروه مهندسی کامپیوتر، واحد اندیمشک، دانشگاه آزاد اسلامی، اندیمشک، ایران
[4] باشگاه پژوهشگران جوان و نخبگان، واحد اندیمشک، دانشگاه ازاد اسلامی، اندیمشک، ایران

چکیده

تکنولوژی بازشناسی گفتار به عنوان یکی از مهمترین شاخه‌های پردازش گفتار از دیر باز مورد توجه پژوهشگران و محققین بوده است. این تکنولوژی قادر است کلمه(کلمات) اداء شده را که با یک سیگنال آکوستیک نمایش داده می‌شود، معین نماید. پیچیدگی سیستم‌های بازشناسی گفتار به ویژگی‌های استخراج شده، ُبعد آنها و نیز دسته‌بند بکار گرفته شده بستگی دارد. در این مقاله، یک دسته‌بند جدید پیشنهاد می‌شود که قادر است در فاز استخراج دانش، از طریق هم‌افزایی خوشه‌بندی و فراوانی مشاهدات، یک مدل مناسب برای هر کلمه مرجع، در قالب دو ماتریس "برنده" و "حداقل فاصله"، محاسبه نماید. روش پیشنهادی در مرحله بازشناسی قادر است با استفاده از یک مکانیزم جریمه-پاداش، میزان شباهت بین گفتار ورودی ناشناخته و مدل‌های مرجع کلمات را معین نماید. به منظور ارزیابی از پایگاه داده فارس‌دات استفاده شده است. نتایج حاصل از آزمایشات متعدد بر روی سیگنال‌های تمیز و نویزی نشان می‌دهند روش ارائه شده در مقایسه با مدل‌های مخفی مارکوف، از دقت بازشناسی بالاتر، مقاوم پذیری بهتر در برابر نویز و نیز پیچیدگی محاسباتی کمتری برخوردار است.

کلیدواژه

بازشناسی گفتار، دسته‌بندی، مدل‌های مخفی مارکوف، خوشه‌بندی، استخراج ویژگی، مقاوم‌پذیری

مقدمه

استفاده از گفتار به عنوان رابطی جهت ارتباط با کامپیوتر/ماشین در سال‌های اخیر محبوبیت ویژه‌ای یافته است و موجب گردیده تحقیقات گسترده‌ای در حوزه پردازش گفتار صورت پذیرد. بازشناسی گفتار، به عنوان یکی از مهمترین شاخه‌های پردازش گفتار از اهمیت ویژه‌ای برخوردار است. بازشناسی گفتار دارای کاربردهای متعدد در حوزه‌های گوناگون است که از جمله آن می‌توان به کابردهای کنترلی فعال شونده با صدا، همچون سیستم‌های مراقبت بهداشتی، سیستم‌های نظامی، جستجوی صوتی در پایگاه داده و غیره اشاره نمود. هدف اصلی سیستم‌های بازشناسی گفتار تبدیل سیگنال آکوستیک به متن می‌باشد. بازشناسی گفتار اساسا یک مسئله چند وجهی و پیچیده است چرا که از یک سو می‌توان آن را به بازشناسی آوا، کلمه، جمله و یا گفتار عامیانه نسبت داد و از طرف دیگر نوع و تعداد گویندگان، لهجه، محیط بازشناسی و غیره، عوامل مهم دیگری بشمار می‌آیند. بنابراین بهبود عملکرد سیستم‌های بازشناسی گفتار از دیر باز بعنوان یکی از چالش-های پیش‌روی محققین در حوزه پردازش گفتار بوده و با وجود تلاش‌های بسیاری که در این خصوص صورت گرفته است، متاسفانه هنوز این سیستم‌ها نتوانسته‌اند کارایی معادل با انسان در شرایط واقع‌بینانه از خود نشان دهند.

پارامتری کردن سیگنال آنالوگ گفتار بعنوان اولین گام در فرآیند بازشناسی گفتار بشمار می‌رود. تکنیک‌های متعددی با الهام از سیستم شنوایی انسان برای این منظور ارائه شده است که از جمله آنها می‌توان به MFCC[1]، LPCC[2, 3] و PLP[4] اشاره نمود. این تکنیک‌ها قادرند یک نمایش پارامتری مفید از سیگنال گفتار ایجاد نمایند. گام بعدی، دسته‌بندی پارامترهای استخراج شده است که در آن بهترین مدل مرجع موجود به الگوی استخراج شده معین گردیده و به عنوان خروجی نمایش داده می‌شود. برای این منظور تکنیک‌های متعددی ارائه شده است که در ادامه به برخی از مهمترین آنها اشاره خواهد شد.

پیچش زمانی پویا(DTW)، که در سال ۱۹۷۸ توسط H.Sakoe و S.Chiba ارائه گردید، از جمله اولین دسته‌بندهایی است که در سیستم های بازشناسی گفتار بکار گرفته شد. این روش قادر است بهترین مسیر انطباق بین دو الگوی گفتاری را بر اساس برنامه‌نویسی پویا محاسبه نماید[۵]. از جمله مشکلات اصلی این روش پیچیدگی زمانی بسیار بالا است چرا که ماهیت آن مبتنی بر برنامه‌نویسی پویا است. نسخه‌های بهبود یافته‌ای از روش DTW به کمک الگوریتم‌های تکاملی ژنتیک[۶] و بهینه سازی اجتماع ذرات[۷] ارائه شده است.

شبکه‌های عصبی مصنوعی(ANN) به دلیل قابلیت‌های ویژه‌ای که در دسته‌بندی غیرخطی الگوها دارند، از اواخر دهه ۸۰ میلادی مورد توجه محققان در زمینه بازشناسی گفتار بوده‌اند.

تاریخ دریافت : ۱۳۹۵/۰۷/۲۷ تاریخ پذیرش : ۱۳۹۶/۰۵/۱۸

از جمله اولین تلاش‌های موفق در خصوص بکارگیری شبکه-های عصبی به منظور بازشناسی گفتار، می‌توان به شبکه عصبی تاخیر زمانی(TDNN) اشاره کرد که توسط K.J Lang و همکارانش در سال ۱۹۹۰ پیشنهاد شد[۸]. شبکه TDNN قادر بود ساختار موقتی رخدادهای آکوستیک و روابط بین چنین رخدادهایی را یاد بگیرد. T.Lee و همکارنش در سال ۱۹۹۷ شبکه عصبی بازگشتی(RNN) را به منظور بازشناسی گفتار بکار گرفتند[۹]. در این روش هر کلمه کتابخانه توسط یک شبکهRNN مدل‌سازی می‌شود. در مرحله بازشناسی، بهترین کلمه منطبق شده بر اساس پاسخ خروجی موقتی شبکه تعیین می‌گردد. H-N. Ting و همکارانش در سال ۲۰۱۳ از یک شبکه عصبی خودتنظیم(SANN) به منظور بازشناسی گفتار استفاده کردند[۱۰]. شبکه SANN قادر است به طور خودکار ساختارش را مطابق با اندازه ورودی منطبق نماید. اخیرا شبکه‌های عصبی جدیدی به نام‌های شبکه‌های عصبی عمیق(DNN) و شبکه‌های عصبی کانولوشن(CNN) معرفی شده‌اند که می‌توانند به منظور کاهش واریانس‌های طیفی و مدل‌سازی همبستگی‌های طیفی موجود در سیگنال‌ها استفاده شوند. از این شبکه‌ها به منظور مدل‌سازی سیگنال آکوستیک گفتار استفاده شده است [۱۱-۱۴].

ماشین‌های بردار پشتیبان(SVM) یکی از ابزارهای پیشرفته در یادگیری ماشین به شمار می‌روند که برای حل مسایل دسته-بندی بسیار کارآمد هستند. از جمله اولین تلاش‌ها در خصوص بکارگیری ماشین بردار پشتیبان، جهت بکارگیری در سیستم-های بازشناسی گفتار، می‌توان به کار ارائه شده توسط A. Ganapathiraju و همکارانش اشاره نمود[۱۵]. در سال ۲۰۰۷، R. Solera-Urena و همکارانش از یک ماشین بردار پشتیبان با هسته Dynamic Time Alignment به منظور بازشناسی گفتار استفاده کردند[۱۶]. S-X. Zhang و M.J.F. Gales از ماشین بردار پشتیبان ساخت یافته(SSVM) به منظور بازشناسی گفتار بهره گرفتند[۱۷].

بازشناسی گفتار از علم فازی نیز بی بهره نبوده است و در این خصوص چندین کار گزارش شده است[۱۸-۲۱].

در دهه‌های اخیر مدل‌های مخفی مارکوف(HMM) [۲۲] بعنوان یکی از موفق‌ترین رویکردها جهت مدل‌سازی آکوستیک، در سیستم‌های بازشناسی گفتار بکار گرفته شده است[۲۳-۳۱]. دلیل اصلی این محبوبیت را می‌توان توانایی بسیار بالای آنها در مدل‌سازی آماری تغییرات موجود در سیگنال غیرایستان گفتار متصور شد چرا که این مدل‌ها از یک چارچوب ریاضیاتی مستحکم بهره می‌گیرند. یک مدل HMM یک ماشین حالت متناهی است که هر حالت آن توسط مدل گوسی منفرد و یا مخلوط مدل گوسی مدل‌سازی می‌شود. پارامترهای مدل HMM عموما مبتنی بر روش بیشترین درستنمایی(ML) یا الگوریتم‌های آموزش تفکیک پذیر[۳۲, ۳۳]، با استفاده از داده های آموزشی کافی، معین می‌شوند. الگوریتم Viterbi به عنوان هسته کدگشایی این مدل‌ها، قادر است به صورت بازگشتی، مبتنی بر برنامه‌نویسی پویا، بهترین مسیر انطباق بین دنباله مشاهده ورودی و مدل‌های مرجع را، با پیچیدگی زمانی $O(N^2T)$، محاسبه نماید. از جمله مشکلات اصلی مدل‌های مخفی مارکوف می‌توان به بهینه نبودن رویه‌های آموزش جهت تخمین پارامترها و نیز پیچیدگی محاسباتی فرآیندهای آموزش و کدگشایی اشاره کرد که موجب می‌گردد کارایی این سیستم-ها تا حدودی کاهش یابد. به منظور بهبود این مدل‌ها مقالات مختلفی ارائه شده است. از جمله این روش‌ها می‌توان به بکارگیری روش‌های تکاملی، همچون الگوریتم ژنتیک و الگوریتم بهینه‌سازی اجتماع ذرات، جهت بهبود رویه آموزش اشاره کرد[۲۴, ۳۴]. در خصوص تسریع فرآیند کدگشایی نیز تلاش‌هایی با استفاده از الگوریتم‌های تکاملی ژنتیک و اجتماع ذرات ارائه شده است[۳۵, ۳۶].

در این مقاله یک سیستم بازشناسی گفتار مبتنی بر هم‌افزایی خوشه‌بندی و فراوانی مشاهدات ارائه می‌شود که قادر است با سرعت و دقت مناسبی عملیات بازشناسی کلمات را انجام دهد. رویکرد پیشنهادی در فاز استخراج دانش، قادر است با آنالیز دنباله‌های مشاهده مرتبط با هر کلمه، یک مدل مرجع شامل دو ماتریس "برنده" و "حداقل فاصله" استخراج نماید. در فاز بازشناسی، با استفاده از یک رویکرد جریمه-پاداش فازی قادر است میزان شباهت الگوی ورودی ناشناخته را به مدل‌های مرجع ذخیره شده محاسبه نماید.

ساختار مقاله بصورت پیرو سازماندهی می‌شود. در بخش دوم با ارائه رویکرد پیشنهادی پرداخته می‌شود. در بخش سوم نتایج آزمایشات و مقایسه‌ها ارائه می‌شود. نهایتا مقاله با بخش نتیجه-گیری پایان می‌یابد.

سیستم بازشناسی خودکار گفتار پیشنهادی

دیاگرام بلوکی سیستم بازشناسی گفتار پیشنهادی در شکل (۱) نشان داده شده است. همانطوری که ملاحظه می‌گردد، رویکرد ارائه شده دارای مراحل پیش‌پردازش، استخراج ویژگی، کوانتیزه کردن برداری، استخراج دانش، انطباق الگو و فرآیند تصمیم-گیری است. جزئیات هر یک از این بخش‌ها در ادامه به طور کامل تشریح می‌گردد.

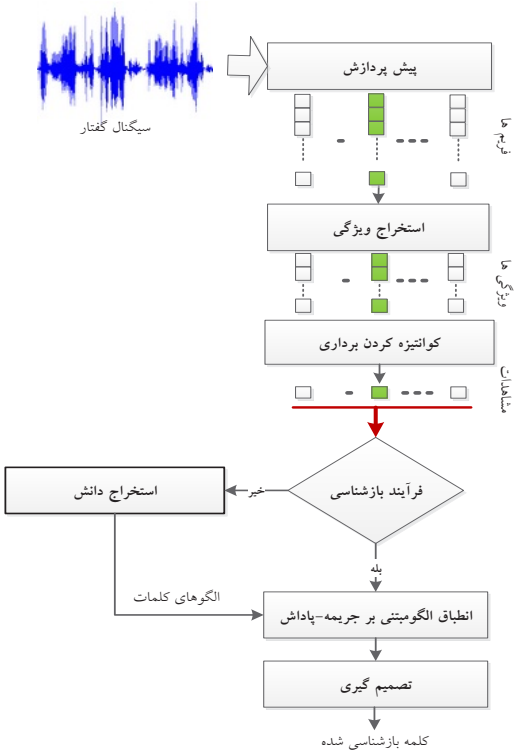
پیش پردازش

اولین بخش در سیستم بازشناسی گفتار پیشنهادی، مرحله

گام، فیلتر کردن پیش‌تاکید است که با استفاده از رابطه (۱) صورت می‌گیرد. فیلتر پیش‌تاکید به منظور هموارسازی فرکانس‌های بالای طیف سیگنال گفتار بکار گرفته می‌شود.

$$H_{pre}(z)=1-0.95z^{-1} \quad (۱)$$

پس از فیلتر کردن پیش‌تاکید، گام بعدی پردازش کوتاه-زمان است. با توجه به اینکه سیگنال گفتار یک سیگنال ناایستان می‌باشد، بنابراین لازم است به مجموعه‌ای ازفریم‌های همپوشان تقسیم گردد به طوری که در هر یک از این فریم‌ها فرض می‌شود سیگنال شبه‌ایستان است. با توجه به اینکه طول سیگنال‌های گفتاری کلمات متفاوت است، بنابراین استفاده از یک پنجره ثابت جهت بخش‌بندی آنها موجب ایجاد الگوهایی با طول متغیر خواهد شد. از آنجایی‌که در رویکرد پیشنهادی فرض بر آن است که طول الگوها یکسان است، بنابراین می‌بایست از تکنیک‌های نرمال‌سازی جهت هم‌طول کردن الگوها



شکل ۱. ساختار کلی سیستم بازشناسی گفتار پیشنهادی

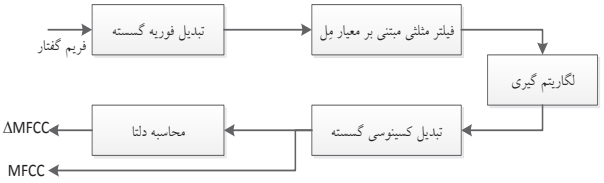
استفاده شود. J.M. Garcia-Cabellos و همکارانش در مقاله خود سه رویکرد مختلف به منظور یکسان‌سازی طول الگوها ارائه نمودند[۳۷]. در رویکرد پیشنهاد شده، از روش پنجره با طول متغیر استفاده شده است. در این روش برای هر سیگنال طول پنجره به گونه‌ای انتخاب می‌شود که تعداد فریم‌ها تعداد مشخصی باشد.

در نهایت به منظور تخریب حداقلی طیف سیگنال پنجره شده این فریم‌ها از یک پنجره همینگ عبور داده می‌شوند. پنجره همینگ دارای تابعی به صورت زیر می‌باشد:

$$w[n]=0.54-0.46cos[2\pi n/(N-1)], 0\le n\le N-1 \quad (۲)$$

استخراج ویژگی

بعد از عملیات پیش‌پردازش، مرحله استخراج ویژگی است. در این مرحله، از هر فریم، ضرایب کپسترال فرکانس مل(MFCC) به همراه مشتق مرتبه اول‌اشان استخراج می‌گردد [۳۸]. در شکل (۲) نحوه تولید ضرایب MFCC نشان داده شده است.



شکل ۲. رویه تولید ضرایب MFCC به همراه مشتق مرتبه اول

کوانتیزه کردن برداری

سومین مرحله در رویکرد پیشنهادی، مرحله کوانتیزه‌کردن برداری است. همانطوری که در شکل (۱) مشاهده می‌شود، سیگنال ورودی پس از طی مراحل پیش‌پردازش و استخراج ویژگی به یک ماتریس ویژگی تبدیل می‌گردد به طوری که هر ستون آن متناظر با ویژگی‌های یک فریم از گفتار ورودی است. کوانتیزه‌کردن برداری موجب می‌شود تا بُعد ماتریس ویژگی کاهش یافته و به یک بردار تبدیل گردد. به این بردار حاصل، "دنباله مشاهده" گفته می‌شود.

قابل ذکر است، قبل از عملیات کوانتیزه‌کردن برداری، لازم است ماتریس‌های ویژگیِ سیگنال‌های آموزشی کلمات، توسط الگوریتم خوشه‌بندی K-Means به K خوشه مجزا تقسیم شده و مراکز این خوشه‌ها، که "کتاب‌کد" نامیده می‌شوند، ذخیره گردند.

به منظور انجام عملیات کوانتیزه‌کردن برداری، در ابتدا لازم است فواصل اقلیدسی، بین ماتریس ویژگی و مراکز خوشه‌ها محاسبه شده و سپس ستون‌های ماتریس ویژگی با شماره خوشه‌هایی که نزدیکترین فاصله را به آنها دارند، جایگزین شوند.

استخراج دانش

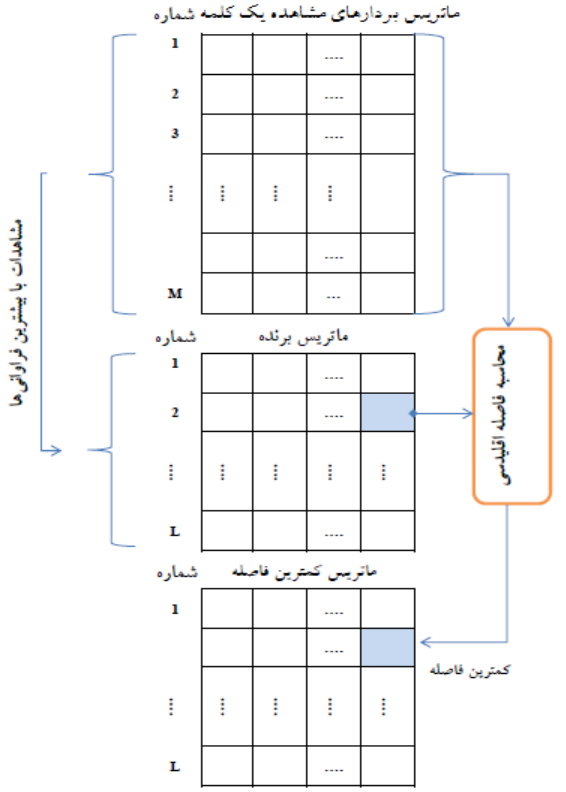
هدف اصلی از مرحله استخراج دانش، تولید الگوهای مناسب به منظور مدل‌سازی کلمات است. رویکرد پیشنهادی قادر است با تحلیل بردارهای مشاهده مرتبط با کلمات مرجع، دو ماتریس "برنده" و "حداقل فاصله" را، به عنوان الگوی هر کلمه مرجع، محاسبه و ذخیره نماید.

ماتریس "برنده" نشان می‌دهد در دنباله‌های مشاهده هر کلمه مرجع، کدامین مشاهدات در هر فریم، بیشترین تکرارها را به

خود اختصاص داده‌اند. بنابراین به منظور تعیین ماتریس "برنده"، دنباله‌های مشاهده کلمات مرجع بررسی شده و سپس L تا از مشاهداتی که بیشترین تکرار را دارند، به عنوان نواحی برنده معین می‌شوند.

ماتریس "حداقل فاصله"، مبین کمترین فاصله اقلیدسی هر فریم، از نواحی برنده‌اش است. بنابراین فریمی که فاصله آن تا ناحیه برنده‌اش، کمترین مقدار را داشته باشد، بیشترین تشابه و فریمی که بیشترین فاصله را داشته باشد، کمترین تشابه را خواهد داشت.

شکل (۳) نمایش تصویری فرآیند استخراج دانش را نشان می-دهد. علاوه بر این، در شکل (۴) شبه کد مربوطه نشان داده شده است.



شکل ۳. نمایش نحوه تولید ماتریس‌های "برنده" و "حداقل فاصله"

```
Procedure Knowlege_Ext traction(Input: Observation Sequences OS, Output: Templates Winner and Min_ Distance )
Begin
 {N is the number of reference words}
{M is the number of word utterances}
 For i=1 to N do
 Begin
   OS[i]=Matrix of observation sequences of i^th word
   Winner[i]=Find_Winners(OS[i]);
   Min_Distance[i]=Find_MinDistance(Winner[i],OS[i]);
 End;
 Return Winner and Min_Distance
End;
{-------------------------------------------------------------------------------}
Function  W=Find_Winners (Input: X)
Begin
 {T is the length of observation sequences}
 {L is the number of frequencies}
 For i=1 to L do
   For  j=1 to T do
     W(i,j)=Find i^th of the most frequency X (:,j);
   End
 End;
 Return W;
End;
{-------------------------------------------------------------------------------}
Function Z=Find_MinDistance (Inputs: W, X)
Begin
 For i=1 to L do
   For  j=1 to T do
     For k=1 to M do
       Distance(k)=| Codebook(X(k, j))- Codebook(W(i,j))|;
     End
     Z(i, j)=Min(Distance)
   End;
 End;
 Return Z;
End;
```

شکل ۴. شبه کد فرآیند استخراج دانش در روش پیشنهادی

انطباق الگو

مرحله انطباق الگو قادر است میزان شباهت سیگنال گفتاری ناشناخته به مدل‌های مرجع ذخیره شده را تعیین نماید. برای این منظور از یک رویکرد جریمه-پاداش فازی استفاده شده است. رویه کار بدین صورت است که در ابتدا فواصل اقلیدسی بین بردار مشاهده آزمایشی و ماتریس‌های "برنده" کلمات محاسبه می‌شوند. سپس، ماتریس‌های فواصل بدست آمده، از ماتریس‌های "حداقل فاصله" کم شده و در نتیجه ماتریس‌هایی تحت عنوان ماتریس‌های "جریمه" ایجاد می‌شوند. بنابراین، متناظر با هر مدل مرجع کلمه، یک ماتریس جریمه تولید می‌گردد.

در صورتی که میزان تفاضل، عددی منفی شود، نشان دهنده آن است که فاصله فریم مربوطه از ناحیه برنده متناظرش، از مقدار کمینه ذخیره شده در مرحله استخراج دانش، کمتر است. بنابراین در چنین حالتی، به جای در نظر گرفتن جریمه منفی، جریمه صفر در نظر گرفته می‌شود.

نهایتا به منظور محاسبه میزان شباهت دنباله مشاهده آزمایشی به مدل‌های مرجع کلمات، لازم است ماتریس‌های "جریمه"، فازی‌سازی شوند. همانطوری که می‌دانیم توابع عضویت فازی به طرق مختلفی قابل تعریف هستند. با توجه به شرایط مسئله، تابع عضویت مثلثی مناسب‌ترین گزینه می‌باشد چرا که این تابع قادر است به جریمه صفر، بیشترین درجه تعلق و به جریمه حداکثر، کمترین درجه تعلق را اختصاص دهد. بنابراین با اعمال تابع عضویت مثلثی بر روی ماتریس‌های "جریمه"، ماتریس-هایی نتیجه می‌شوند که آنها را ماتریس‌های "پاداش" معرفی می‌نامیم. از طریق ماتریس‌های "پاداش" به سادگی می‌توان درجه تعلق یک دنباله مشاهده آزمایشی را به تک تک مدل‌های مرجع کلمات معین نمود.

تصمیم‌گیری

به منظور تعیین برچسب سیگنال گفتار ورودی ناشناخته، کافی است میانگین ماتریس‌های "پاداش" مرتبط با کلمات مرجع را محاسبه نمود. کلمه مرجعی که دارای بیشترین مقدار میانگین پاداش باشد به عنوان کلمه مورد نظر توسط سیستم تشخیص داده می‌شود. شبه کد فرآیند دسته‌بندی دنباله آزمایشی ناشناخته در شکل (۵) نشان داده شده است.

```
Procedure Penalty-Reward Classifier(Inputs: X test sequence, Winner and Min_Distance matrices )
Begin
 {N is the number of reference words}
 {L is number of frequencies}
 For i=1 to N do
 Begin
    W=Winner[i];
   MD=Min_Distance[i];
   For l=1 to L do
   Begin
     Distance (l, :) = | Codebook(X) - Codebook (W (l, :))| ;
   End;
   Penalty=Distance-MD;
   Reward=Fuzzy(Penalty);
   Score(i)=Mean(Reward);
End;
Return recognized word=Argmax(Score);
End;
```

شکل ۵. شبه کد دسته‌بند پیشنهادی مبتنی بر سیستم جریمه-پاداش

نتایج آزمایشات

در بخش قبل به ارائه رویکرد پیشنهادی به منظور بازشناسی گفتار گسسته پرداخته شد. در این بخش به ارزیابی کارایی روش پیشنهادی پرداخته می‌شود. قبل از ارائه نتایج لازم است مجموعه داده استفاده شده، نرم افزار شبیه‌سازی و مشخصات سیستم کامپیوتری استفاده معرفی شوند.

پایگاه داده مورد استفاده فارس‌دات می‌باشد. این پایگاه داده شامل تنوعی از داده‌های صوتی ادا شده توسط ۳۰۴ گوینده بومی می‌باشد که از لحاظ سن، جنس، لهجه و سطح تحصیلات با یکدیگر متفاوت می‌باشند. ۶۰۸۰ عبارت گفتاری به طور دستی بر حسب آوا و کلمه بخش‌بندی و برچسب‌گذاری شده-اند. به منظور ارزیابی روش پیشنهادی یک مجموعه ۲۰ کلمه‌ای تصادفی استفاده شده است که در پنج دسته چهار کلمه‌ای تقسیم شده‌اند.

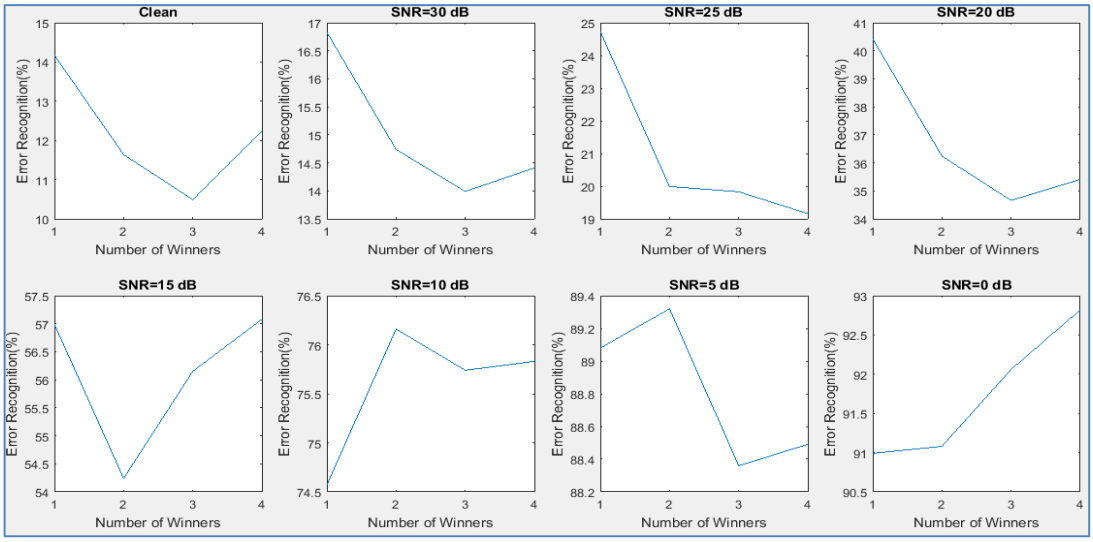
تمامی مراحل سیستم پیشنهادی با استفاده از نرم افزار MATLAB 2015 بر روی یک کامپیوتر با پردازنده Core i3 با ۲ گیگا بایت حافظه RAM پیاده سازی شده است.

جدول (۱) نتایج حاصل از ارزیابی رویکرد پیشنهادی با تعداد فراوانی مختلف، برای سیگنال‌های گفتاری تمیز و نویزی را

نشان می‌دهد. در این آزمایشات از نویز سفید گوسی در هفت سطح سیگنال به نویز(SNR) مختلف ۰، ۵، ۱۰، ۱۵، ۲۰، ۲۵ و ۳۰ دسی‌بل، استفاده شده است. لازم به ذکر است که ۷۵ درصد سیگنال‌های گفتاری جهت تولید مدل‌ها و ۲۵ باقی مانده جهت ارزیابی کارایی رویکرد پیشنهادی بکار گرفته شده‌اند. به منظور قضاوت عادلانه در خصوص نتایج، میانگین وزن‌دار خطای بازشناسی در جدول (۱) محاسبه و ارائه شده است. همانطوریکه در جدول (۱) مشاهده می‌شود، بهترین کارایی سیستم پیشنهادی برای تمامی مجموعه‌های کلمات، با تعداد فراوانی‌های مختلف، به سیگنال‌های تمیز تعلق دارد. علاوه بر این، ملاحظه می‌گردد کاهش مقدار SNR، افزایش خطای بازشناسی را در کلیه حالات بدنبال خواهد داشت. لازم به ذکر است، کارایی رویکرد پیشنهادی بی‌تاثیر از پارامتر تعداد فراوانی‌ها نیست. نتایج حاصله نشان می‌دهد متوسط میزان خطا با افزایش تعداد فراوانی‌ها در ابتدا شروع به کاهش کرده و سپس افزایش می‌یابد به طوری در نظر گرفتن تعداد سه فراوانی، در بسیاری از حالات، مقادیر کمینه خطای بازشناسی را نتیجه می‌دهد. این پدیده را می‌توان به خوبی در شکل (۶) ملاحظه کرد.

جدول ۱. نمایش خطای بازشناسی روش پیشنهادی در حالت بدون نویز و با نویز سفید گوسی با SNR های مختلف

پارامتر تعداد فراوانی	تعداد کلمات	نرخ خطا روش پیشنهادی(٪)							
		Clean	30 dB	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
1	4	1.25	5	8.75	23.75	36.25	52.50	63.75	56.25
	8	6.87	7.50	13.75	31.25	53.12	69.37	86.87	91.25
	12	15	16.25	22	38.33	52.91	73.71	87.5	94.16
	16	15.65	20.31	27.18	44.10	57.5	75.93	92.81	93.12
	20	18	20.50	32	45.75	64.75	80.50	93	94.25
	میانگین	14.17	16.83	24.73	40.42	56.99	74.57	89.08	90.99
2	4	1.25	2.5	6.25	15	35	45	49.95	53.75
	8	6.70	8.12	13.75	30	49.37	72.5	83.75	91.25
	12	11.25	13.33	17.91	33.75	52.91	75.83	92.08	93.75
	16	13.75	16.78	22.5	38.12	53.12	79.37	93.12	93.12
	20	14.25	19.25	24.50	43	61.75	81.50	94.75	95.25
	میانگین	11.64	14.74	19.99	36.24	54.24	76.16	89.32	91.08
3	4	1.25	1.25	1.25	10	21.25	33.75	40	57.5
	8	4.37	6.25	14.37	29.37	53.75	76.87	84.62	91.78
	12	10.41	12.91	18.33	35	56.66	77.91	89.58	92.5
	16	12.18	18.43	23.75	39.06	59.03	77.81	94.37	95.93
	20	13.50	16.75	23.50	38	61.50	80.75	94	95.75
	میانگین	10.49	13.99	19.83	34.66	56.15	75.74	88.36	92.06
4	4	1.25	1.25	1.25	11.25	28.75	42.5	53.75	66.25
	8	6.25	9.37	14.37	33.12	54.37	72.50	85	90.62
	12	12.08	12.91	18.75	35.41	55.83	74.58	88.33	93.33
	16	13.12	14.37	20.93	35.93	57.50	79.68	92.18	95.31
	20	16.25	20	23.50	40.75	64.25	81.50	94	96.75
	میانگین	12.24	14.41	19.16	35.41	57.08	75.83	88.49	92.82

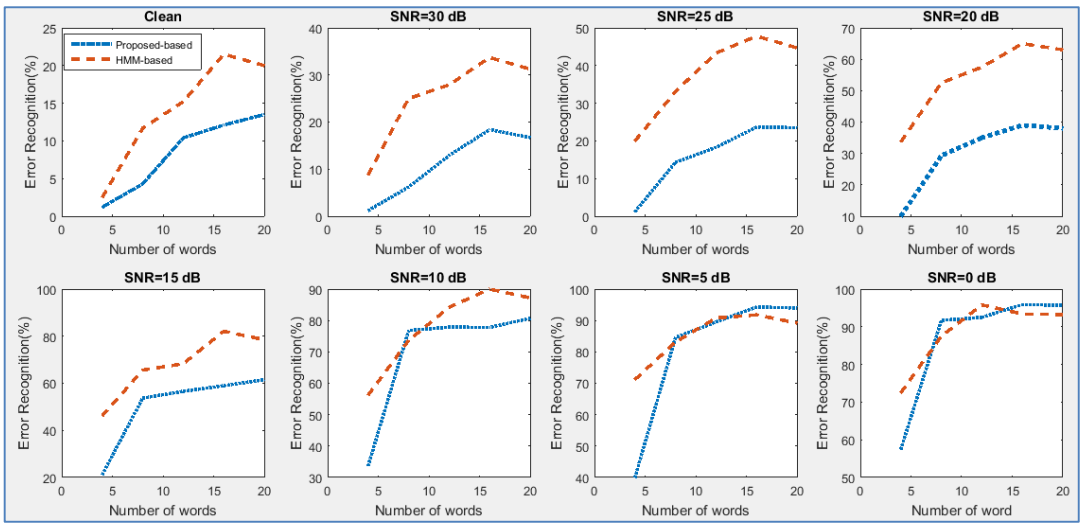


شکل ۶. متوسط خطای بازشناسی روش پیشنهادی با تعداد فراوانی‌های مختلف

همانطوریکه بیان شد، سیستم‌های بازشناسی گفتار مبتنی بر مدل‌های مخفی مارکوف به عنوان یکی از بهترین روش‌ها به شمار می‌روند. برای این منظور رویکرد پیشنهادی با یک سیستم بازشناسی گفتار مبتنی بر مدل‌های مخفی مارکوف مورد ارزیابی و مقایسه قرار گرفته است. نتایج حاصل از ارزیابی ها در جدول (۲) و شکل (۷) نشان داده شده است.

جدول ۲. نتایج حاصل از ارزیابی روش پیشنهادی با مدل مخفی مارکوف

سیگنال‌های گفتاری	نرخ خطای روش پیشنهادی(٪)						نرخ خطای روش مدل مخفی مارکوف(٪)					
	تعداد کلمات					میانگین	تعداد کلمات					میانگین
	4	8	12	16	20		4	8	12	16	20	
Clean	1.2	4.3	10.4	12.1	13.5	10.49	2.5	11.7	15.2	21.5	20	17.16
with SNR 30dB	1.2	6.2	12.9	18.4	16.7	13.99	8.75	25	27.9	33.7	31.2	28.90
with SNR 25 dB	1.2	14.3	18.3	23.7	23.5	19.83	20	33.1	43.3	47.8	44.7	42.05
with SNR 20 dB	10.0	29.3	35.0	39.0	38.0	34.66	33.7	52.5	57.5	65	63	59.08
with SNR 15 dB	21.2	53.7	56.6	59	61.5	56.15	46.2	65.6	68.3	82.1	78.5	73.54
with SNR 10 dB	33.7	76.8	77.9	77.8	80.7	75.74	56.2	73.7	84.2	90	87.2	83.48
with SNR 5 dB	40.0	84.6	89.5	94.3	94.0	88.30	71.2	83.1	90.8	91.8	89.2	88.20
with SNR 0 dB	57.5	91.7	92.5	95.9	95.7	92.03	72.5	87.5	95.8	93.4	93.2	91.63



شکل۷. نمایش نرخ خطای روش پیشنهادی با روش مدل مخفی مارکوف در شرایط تمیز و نویزی

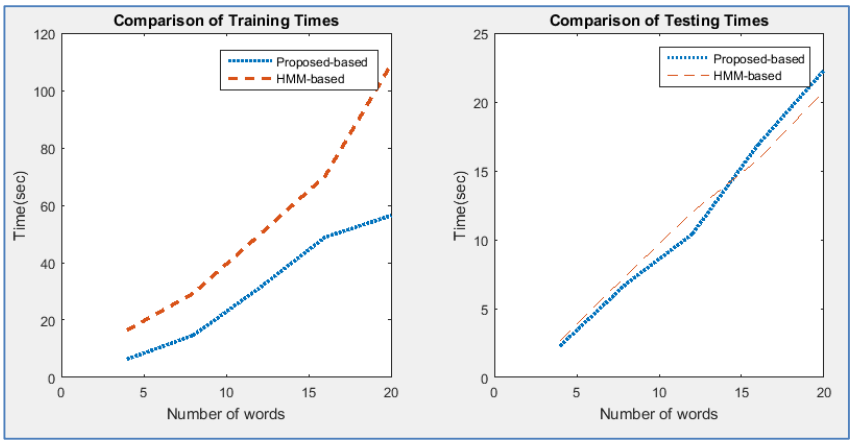
همانطوری که مشاهده می‌شود، در روش مدل مخفی مارکوف به مانند رویکرد پیشنهادی، میانگین نرخ خطا با کاهش نسبت SNR افزایش می‌یابد. در حالت سیگنال‌های گفتاری تمیز، دقت روش پیشنهادی به طور متوسط حدود ۷ درصد از مدل مخفی مارکوف بهتر است. نکته جالب توجه در اینجا آن است که اگرچه دقت روش پیشنهادی در نسبت SNR های کمتر از ۱۰ دسی‌بل بسیار نزدیک به روش مدل مخفی مارکوف است، ولیکن در سیگنال‌های نویزی با SNR های محدوده ۱۰ تا ۳۰ دسی‌بل به طور متوسط بیش از ۱۷ درصد کارایی بالاتری را دارا می‌باشد. این نتیجه نشان می‌دهد روش پیشنهادی از مقاوم پذیری بالایی در برابر نویز برخوردار است.

روش ارائه شده علاوه بر نرخ خطای بازشناسی، از نظر پیچیدگی زمانی نیز با مدل مخفی مارکوف مورد ارزیابی و مقایسه قرار گرفته است. برای این منظور زمان تولید مدل‌های مرجع و نیز زمان بازشناسی برای مجموعه کلماتی با اندازه‌های مختلف در جدول(۳) نشان داده شده است. علاوه بر این، در شکل(۸) مقایسه زمان‌های آموزش و بازشناسی روش پیشنهادی با مدل مخفی مارکوف برای مجموعه کلمات مختلف نشان داده شده است.

همانطوری که در جدول (۳) مشاهده می‌شود، اگرچه روش پیشنهادی از نظر زمان بازشناسی به مدل مخفی مارکوف بسیار نزدیک است ولیکن دارای زمان استخراج دانش(آموزش) بسیار کمتری می‌باشد. این مسئله را می‌توان به وضوح در شکل (۸) ملاحظه کرد که در آن با افزایش تعداد کلمات مرجع، زمان آموزش مدل مخفی مارکوف دارای رشد سریعتری از روش پیشنهادی است. دلیل این امر را می‌توان به آسانی در رویه استخراج دانش روش پیشنهادی مشاهده کرد.

جدول ۳. نتایج حاصل از ارزیابی روش پیشنهادی با مدل مخفی مارکوف از نظر زمان استخراج دانش و زمان بازشناسی

زمان(ثانیه)	روش پیشنهادی						روش مدل مخفی مارکوف					
	میانگین	تعداد کلمات					میانگین	تعداد کلمات				
		20	16	12	8	4		20	16	12	8	4
استخراج دانش	40.4	56.4	48.9	31.1	14.7	6.4	69.98	109	70.2	49.6	29.3	16.5
بازشناسی	15.08	22.3	16.9	10.4	6.8	2.3	14.71	20.8	15.8	12	7.4	2.7



شکل ۸. نمایش مقایسه زمان‌های آموزش و بازشناسی روش پیشنهادی با مدل مخفی مارکوف

نتیجه‌گیری

در این مقاله یک دسته‌بند جدید به منظور بکارگیری در سیستم‌های بازشناسی گفتار مجزا ارائه شده است که قادر است با استفاده از یک رویکرد جریمه-پاداش فازی، عملیات بازشناسی گفتار را بخوبی انجام دهد. به منظور ارزیابی دسته-بند پیشنهادی، آزمایشات متعددی با استفاده از سیگنال‌های گفتاری تمیز و نویزی، بر روی مجموعه کلمات مختلف پایگاه

با سیستم بازشناسی گفتار مبتنی بر مدل مخفی مارکوف، نشان می‌دهد رویکرد پیشنهادی از مقاوم‌پذیری مناسبتری در برابر نویز برخوردار است. علاوه بر این، پیچیدگی زمانی فرآیندهای استخراج دانش و بازشناسی مورد بررسی قرار گرفت. نتایج حاصله نشان می‌دهد اگرچه سرعت بازشناسی روش پیشنهادی بسیار نزدیک به مدل‌های مخفی مارکوف است ولیکن رویکرد ارائه شده در زمان بسیار کمتری قادر است مدل‌های مرجع را استخراج نماید. بنابراین نتیجه گرفته می‌شود که این روش

می‌تواند در محیط‌های نویزی و کاربردهای بلادرنگ بخوبی مورد استفاده قرار گیرد.

تشکر و قدردانی

از حمایت دانشگاه آزاد اسلامی واحد دزفول در انجام این طرح تحقیقاتی با عنوان ارائه یک دسته بند جدید به منظور بازشناسی گفتار گسسته فارسی مبتنی بر روش پاداش و جریمه، کمال تشکر و قدردانی را داریم.

مراجع

[1] S. Davis, and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, vol. 28, no. 4, pp. 357-366.

[2] P. E. Papamichalis, “Practical approaches to speech coding," Prentice-Hall, Inc., 1987.

[3] J. Makhoul, “Linear prediction: A tutorial review,” Procedding of IEEE, vol. 63, isuue 4, 1975, pp. 561-580.

[4] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," IEEE International Conference on Acostics, Speech and Signal Processing, Turkey, 2000, pp. 1635-1638.

[5] H. Sakoe, and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Transactions on Acoustics, Speech, and Signal processing, 1978, vol. 26, no. 1, pp.43-49

[6] S. Kwong, Q. He, and K.-F. Man, “Genetic time warping for isolated word recognition,” International Journal of Pattern Recognition and Artificial Intelligence, 1996, vol. 10, no. 07, pp. 849-865.

[7] S. Rategh, F. Razzazi, A. M. Rahmani et al., "A time warping speech recognition system based on particle swarm optimization,"International Conference on Modeling and Simulation, Malasia, 2008, pp. 585-590.

[8] K. J. Lang, A. H. Waibel, and G. E. Hinton, “A time-delay neural network architecture for isolated word recognition,” Neural networks, 1990, vol. 3, no. 1 ,pp. 23-43.

[9] T. Lee, P. Ching, and L.-W. Chan, “Isolated word recognition using modular recurrent neural networks,” Pattern Recognition, 1998, vol. 31, no. 6, pp. 751-760.

[10] H.-N. Ting, B.-F. Yong, and S. M. Mirhassani, “Self-adjustable neural network for speech recognition,” Engineering Applications of Artificial Intelligence,2013, vol. 26, no. 9, pp. 2022-2027.

[11] A.-r. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Transactions on Audio, Speech, and Language Processing, 2012, vol. 20, no. 1, pp. 14-22.

[12] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview,"IEEE international Conference on Acoustics,Speech and Signal Processing, Canada, 2013, pp. 8599-8603.

[13] T. N. Sainath, B. Kingsbury, G. Saon et al., “Deep convolutional neural networks for large-scale speech tasks,” Neural networks, 2015, vol. 64, pp. 39-48.

[14] S. M. Siniscalchi, D. Yu, L. Deng et al., “Exploiting deep neural networks for detection-based speech recognition,” Neurocomputing, 2013, vol. 106 ,pp. 148-157.

[15] A. Ganapathiraju, J. E. Hamaker, and J. Picone, “Applications of support vector machines to speech recognition,” IEEE Transactions on Signal Processing, 2004, vol. 52, no. 8, pp. 2348-2355.

[16] R. Solera-Urena, D. Martin-Iglesias, A . Gallardo-Antolin et al., “Robust ASR using support vector machines,” Speech Communication, 2007, vol. 49, no. 4, pp. 253-267.

[17] S.-X. Zhang, and M. J. Gales, “Structured SVMs for automatic speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, 2013, vol. 21, no. 3, pp. 544-555.

[18] Y.-Q. Ying, and P.-Y. Woo, “Speech recognition using fuzzy logic,” International Joint Conference on Neural Networks, USA, 1999, pp. 2962-2964.

[19] E. Avci, and Z. H. Akpolat, “Speech recognition using a wavelet packet adaptive network based fuzzy inference system,” Expert Systems with Applications, 2006, vol. 31, no. 3, pp. 495-503.

[20] R. Halavati, S. B. Shouraki, and S. H. Zadeh, “Recognition of human speech phonemes using a

[30] N. Najkar, F. Razzazi, and H. Sameti, “An evolutionary decoding method for HMM-based continuous speech recognition systems using particle swarm optimization,” Pattern Analysis And Applications, 2014, vol. 17, no. 2, pp. 327-339.

[31] A. Shaukat, H. Ali, and U. Akram, "Automatic Urdu Speech Recognition using Hidden Markov Model, "International Conference on Image, Vision and Computing, UK, 2016, pp. 135-139.

[32] Q. Hong, and S. Kwong, “A training method for hidden Markov model with maximum model distance and genetic algorithm,” International Conference on Neural Networks and Signal Processing, China, 2003, pp. 465-468.

[33] S. Mizuta ,and K. Nakajima, “A discriminative training method for continuous mixture density HMMs and its implementation to recognize noisy speech,” Journal of the Acoustical Society of Japan (E), 1992, vol. 13, no. 6, pp. 389-393.

[34] H. Sajedi, H. Sameti, H. Beigy et al., "Discriminative training of Hidden Markov Model using pso algorithm." 12th Annual Computer Society of Iran, Iran, 2007, pp. 295-302 (in persian).

[35] M. Mosleh, S. Setayeshi, and A. M. Rahmani, “A synergy between HMM-GA based on stochastic cellular automata to accelerate speech recognition,” IEICE Electronics Express, 2009, vol. 6, no. 18, pp. 1304-1311.

[36] N. Najkar, F. Razzazi, and H. Sameti, “A novel approach to HMM-based speech recognition systems using particle swarm optimization,” Mathematical and Computer Modelling, 2010, vol. 52, no. 11, pp. 1910-1920.

[37] J. M. Garca-Cabellos, C. Pelaez-Moreno, A. Gallardo-Antolin et al., “SVM classifiers for ASR: A discussion about parameterization,” 12th European Signal Processing Conference, Austria, 2004, pp. 2067-2070.

[38] X. Huang, A. Acero, H.-W. Hon et al ,.Spoken language processing: A guide to theory, algorithm, and system development: Prentice hall PTR, 2001.

novel fuzzy approach,” Applied Soft Computing, 2007, vol. 7, no. 3, pp. 828-839.

[21] W. Silva, and G. Serra, “Intelligent Genetic Fuzzy Inference System for Speech Recognition: An Approach from Low Order Feature Based on Discrete Cosine Transform,” Journal of Control, Automation and Electrical Systems, 2014, vol. 25, no. 6, pp. 689-698.

[22] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proceedings of the IEEE, 1989, vol. 77, no. 2, pp. 257-286.

[23] C. Pisarn, and T. Theeramunkong, “An HMM-based method for Thai spelling speech recognition,” Computers & Mathematics with Applications, 2007, vol. 54, no. 1, pp. 76-95.

[24] P. Bhuriyakorn, P. Punyabukkana, and A. Suchato, "A genetic algorithm-aided hidden markov model topology estimation for phoneme recognition of thai continuous speech,", International Coference on Software Engineering, Artificial Intelligent, Networking and parallel/Distributed Computing, Thailand, 2008, pp. 475-480.

[25] J. Cai, G. Bouselmi, Y. Laprie et al., “Efficient likelihood evaluation and dynamic Gaussian selection for HMM-based speech recognition,” Computer Speech & Language, 2009, vol. 23, no. 2 ,pp. 147-164.

[26] D. H. Milone, L. E. Di Persia, and M. Torres, “Denoising and recognition using hidden Markov models with observation distributions modeled by hidden Markov trees,” Pattern Recognition, 2010, vol. 43, no. 4, pp. 1577-1589.

[27] V .Radha, "Speaker independent isolated speech recognition system for Tamil language using HMM,"Procedia Engineering, 2012, pp. 1097-1102.

[28] T. Ma, S. Srinivasan, G. Lazarou et al., “Continuous speech recognition using linear dynamic models,” International Journal of Speech Technology, 2014, vol. 17, no. 1, pp. 11-16.

[29] P. Paramonov, and N. Sutula, “Simplified scoring methods for HMM-based speech recognition,” Soft Computing, 2015, vol. 20, no. 9, pp. 3455-3460.