به نام خدا

دانشکده مهندسی برق و رباتیک
گروه الکترونیک

استخراج ویژگی مبتنی بر پردازش در حوزهٔ اتوکرولیشن جهت بازشناخت گفتار با استفاده از HTK

دانشجو: **سید حمید اخلاق**

استاد راهنما:
دکتر حسین مروی

استاد مشاور:
دکتر امید رضا معروضی

پایان نامه ارشد جهت اخذ درجهٔ کارشناسی ارشد

بهمن ۱۳۸۸

شماره : ۰۷۱۵/آ.ت.ب
تاریخ : ۱۳۸۸/۱۱/۰۵
ویرایش : - - - -

دانشگاه صنعتی شاهرود

مدیریت تحصیلات تکمیلی

فرم شماره (۶)

بسمه تعالی

فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای **سید حمید اخلاق** رشته **بـــرق** گرایش **الکترونیک**

تحت عنوان : **استخراج ویژگی مبتنی بر پردازش در حوزه اتوکرولیشن جهت بازشناخت گفتار با استفاده از HTK**

که در تاریخ **۱۳۸۸/۱۱/۰۵** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه : عالی امتیاز ۱۹/۸۵) ☐ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹) ۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)

۳ـ خوب (۱۷/۹۹ ـ ۱۶) ۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنما	دکتر مروی	استادیار	
۲ـ استاد مشاور	دکتر معروفی	استادیار	
۳ـ نماینده شورای تحصیلات تکمیلی	دکتر حداد ظریف	استادیار	
۴ـ استاد ممتحن	دکتر زاهدی	استادیار	
۵ ـ استاد ممتحن	دکتر خسروی	استادیار	

تأیید رئیس دانشکده :

تقدیر و تشکر

در ابتدا خدا را به خاطر محبت عظیمش سپاس می‌گویم سپس از زحمات استاد راهنما جناب آقای دکتر حسین مروی و همچنین خانواده و برادرانم تشکر می‌کنم.

تعهد نامه

اینجانب ..... دانشجوی دوره کارشناسی ارشد / دکتری رشته .....

دانشکده ..... دانشگاه صنعتی شاهرود نویسنده پایان نامه/ رساله .....

..... تحت راهنمائی ..... متعهد می شوم .

- تحقیقات در این پایان نامه/ رساله توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه/ رساله تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه / رساله تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه/ رساله رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه/رساله ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه/رساله ، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ : ۸۸/۱۱/۷

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه / رساله بدون ذکر مرجع مجاز نمی باشد.

* متن این صفحه نیز باید در ابتدای نسخه های تکثیر شده پایان نامه/ رساله وجود داشته باشد .

چکیده

یکی از مهمترین مسایل مطرح در زمینهٔ سیستم‌های بازشناخت گفتار تأثیر نویز بر سیگنال گفتار و کاهش دقت بازشناسی است، لذا بازشناخت گفتار مقاوم در برابر نویز یکی از موارد مورد مطالعه محققان می‌باشد. در سالهای اخیر تحقیقات بسیاری در این زمینه به عمل آمده و روش‌های مختلفی ارائه شده است. یک دسته از روش‌های موجود در زمینهٔ بازشناخت گفتار مقاوم، استخراج ویژگی‌هایی از سیگنال گفتار می‌باشد که نسبت به نویز مقاوم باشند. در این پایان نامه هدف معرفی روش‌هایی جهت مقاوم سازی ویژگی‌های استخراج شده از سیگنال گفتار در برابر نویز جمع شونده می‌باشد. حوزه‌ای که در این پایان نامه از آن برای استخراج ویژگی مقاوم استفاده شده است، حوزهٔ اتوکرولیشن می‌باشد. برای این منظور در ابتدا بعد از بررسی مهمترین اجزاء تشکیل دهندهٔ سیستم‌های بازشناخت گفتار و مرور بعضی از کارهای انجام شده در زمینهٔ استخراج ویژگی مقاوم در حوزهٔ اتوکرولیشن، روش‌های پیشنهادی معرفی شده‌اند. از روش‌های موجود در حوزهٔ اتوکرولیشن RAS، DAS، AMFCC و PAC می‌باشند. در این پایان نامه ایده‌هایی جدید جهت بهبود روش‌های AMFCC و PAC بر اساس استفاده از مشتق طیف توان و پنجره گذاری مناسب پیشنهاد شده است. نتایج بدست آمده از پیاده سازی روش‌های پیشنهادی بر روی دیتابیس TIMIT بیانگر بهبود در نرخ بازشناخت گفتار پیوسته نسبت به برخی از روش‌های گذشته است.

واژگان کلیدی: بازشناخت گفتار - استخراج ویژگی - اتوکرولیشن - بستهٔ نرم افزاری مدل مخفی مارکوف

لیست مقاله‌های استخراج شده از پایان نامه

1. S. H. Akhlagh and H. Marvi (2009), “Features Derived from Differential Power Spectrum in Autocorrelation Domain for Continuous Speech Recognition”, ICSV16, Kracow, Poland.

2. S. H. Akhlagh and H. Marvi (2010), “Robust Features Derived from Differentiated Phase Autocorrelation Spectrum for Speech Recognition”, 16th National Conference on Communications, Co-sponsored by IEEE Comm. Society, India.

3. S. H. Akhlagh, H. Marvi and O. R. Ma'rouzi, “Model based and cepstral based features obtained from differentiated autocorrelation spectrum for robust speech recognition”, submitted to DSP Elsevier, under review.

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل اول

مقدمه

فصل اول: مقدمه

۱–۱ مقدمه

با پیشرفت علم و ساخت وسایل الکترونیکی پیشرفته برای انجام کارهای مختلف، بشر این نیاز را در خود احساس کرده است که بتواند راحتتر با این وسایل ارتباط برقرار کند. بدون شک یکی از بهترین و آسانترین راههای برقراری ارتباط از طریق صحبت کردن است که در زندگی روزانه در برخورد با دوستان و همکاران از آن برای بیان منظور خود استفاده می‌کنیم. متاسفانه یا خوشبختانه مسیر پیشرفت علم به سمتی پیش می‌رود که این دوستان و همکاران جای خود را به وسایل الکترونیکی از قبیل کامپیوتر و تلفن همراه داده‌اند. افزایش روز افزون استفاده از این وسایل، دانشمندان را به این فکر انداخت که به جای استفاده از دست برای وارد کردن فرمانها و برقراری ارتباط با این وسایل از صحبت کردن معمولی استفاده شود. عملی کردن این فکر مربوط به شاخه‌ای از علم با عنوان

بازشناخت گفتار اتوماتیک[1] می‌شود. در طول ۶۰ سال گذشته تحقیقات زیادی در این حوزه انجام شده است و پیشرفتهای چشم گیری هم بدست آمده است.

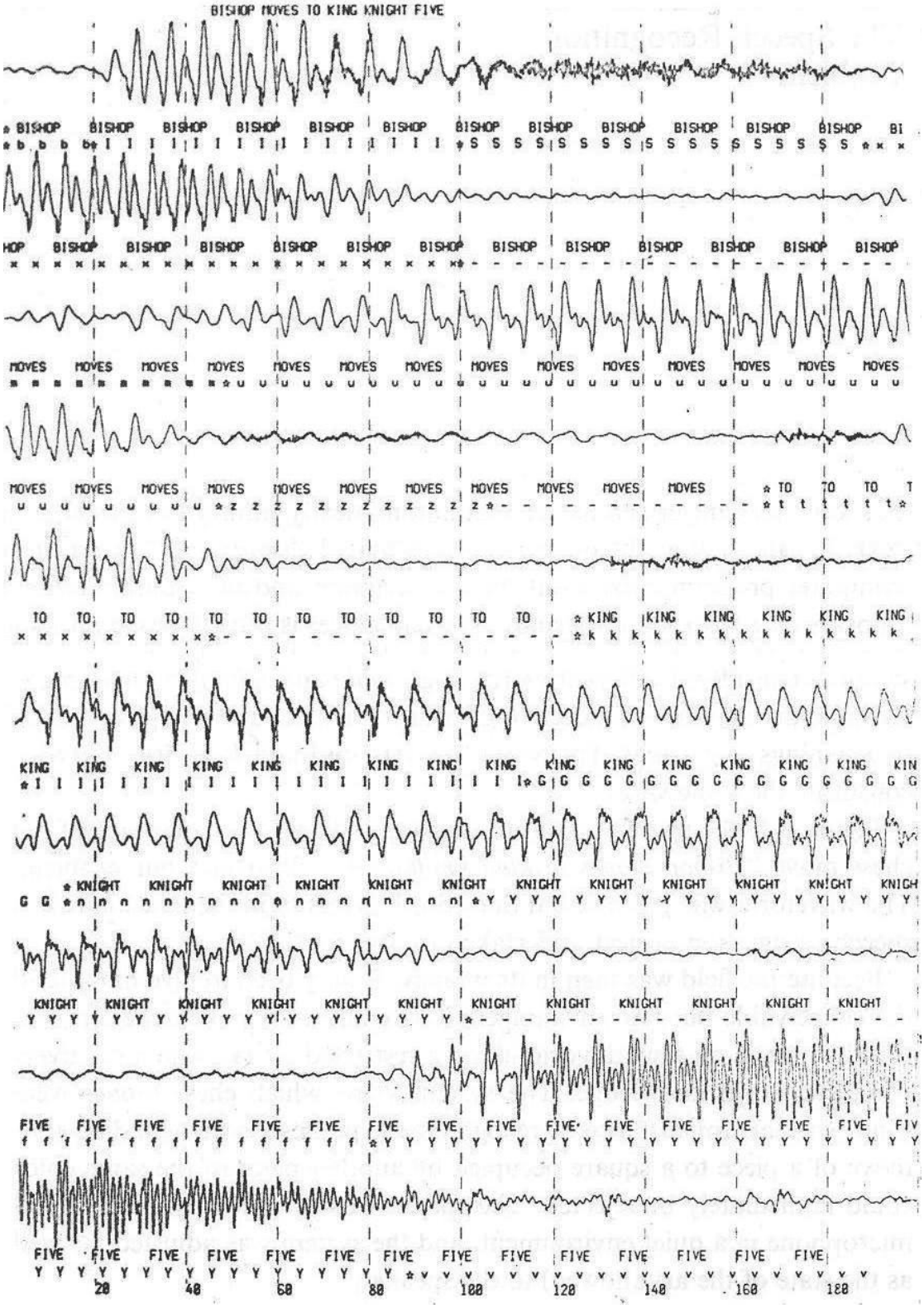
سیستم‌های بازشناخت گفتار اولیه فقط قادر به تشخیص بعضی از حروف یا کلمات محدود و جدا از هم بودند. ولی سیستم‌های بازشناخت گفتار امروزی قادر هستند گفتار پیوسته را با دقت نسبتاً خوبی تشخیص دهند. البته بیان این مطلب نیز لازم است که با وجود تمام پیشرفتهای انجام شده هنوز قادر به ساخت سیستمی که بتواند مستقل از محیط و شخص صحبت کننده کلمات گفته شده را به خوبی انسان تشخیص دهد نیستیم.

مسائل زیادی وجود دارد که ساخت سیستم‌های بازشناخت گفتار را با مشکل مواجه می‌کند. به عنوان مثال شکل موجی که از صوت بدست می‌آید متناسب با شخص گوینده و در شرایط محیطی متفاوت تغییرات زیادی از خود نشان می‌دهد. علاوه بر تغییر شخص و شرایط محیطی بعضی از مواقع حروفی که از یک شخص ثابت و در یک مکان ثابت ادا می‌شوند نیز به خاطر محتوای کلمه ادا شده دارای شکل موجهای متفاوتی هستند. به عبارت دیگر داشتن دو شکل موج یکسان دلیل بر این نمی‌شود که هر دو مربوط به یک صدا می‌شوند. به عنوان مثال همانطور که در شکل ۱-۱ مشاهده می‌شود شکل موج حرف i در کلمه KING بیشتر به شکل موج حرف u در کلمه MOVE شبیه است تا به شکل موج حرف i در کلمه BISHOP [۱].

امروزه سیستم‌های بازشناخت گفتار کاربردهای مختلفی پیدا کرده‌اند. از کاربردهای پزشکی گرفته تا کاربردهای نظامی ولی یکی از مهمترین این کاربردها استفاده از سیستم‌های بازشناخت گفتار برای تبدیل سیگنال صحبت به متن است.

در این فصل در ابتدا تاریخچه مختصری از بازشناخت گفتار آورده شده است سپس هدف از این پایان نامه و ساختار فصلهای مختلف آن بیان شده است.

[1] Automatic Speech Recognition



شکل ۱-۱ جمله "Bishop moves to king knight five" به همراه شکل موج مربوط به آن [۱].

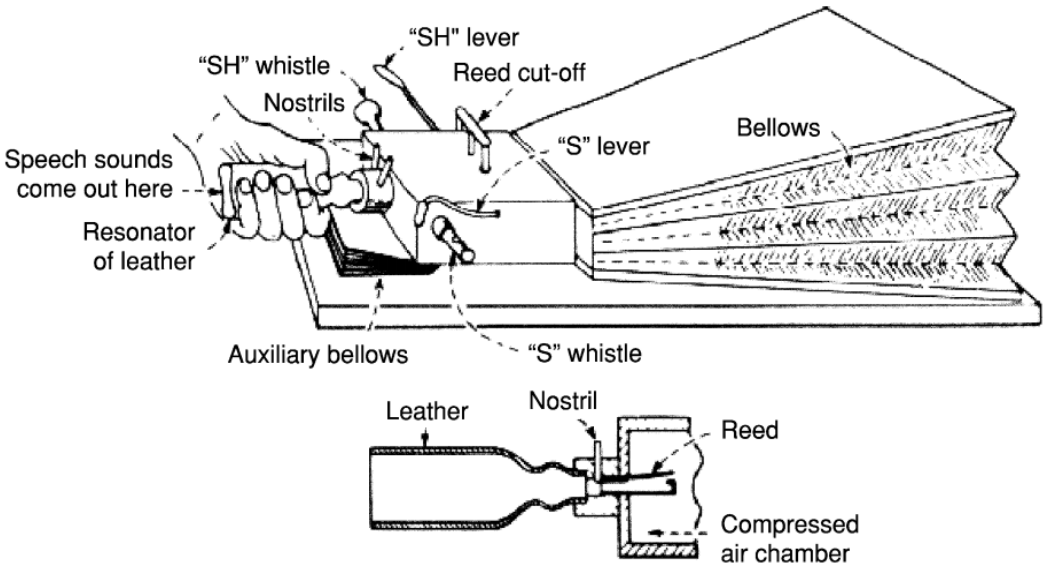
۱-۲ تاریخچه بازشناخت گفتار

در طول ۶۰ سال گذشته پیشرفتهای زیادی در قسمتهای مختلف سیستم‌های بازشناخت گفتار به عمل آمده است. ولی اکثر سیستم‌های مدرنی که امروزه در زمینه بازشناخت گفتار وجود دارند مدیون کارهایی هستند که تقریبا از نیمه دوم قرن هیجدهم میلادی، برای ساخت ماشینی که بتواند مانند انسان صحبت کند، انجام شد. قبل از مطرح شدن مبحث بازشناخت گفتار دانشمندان علاقه‌مند به ساخت ماشینی بودند که بتواند مانند انسان صحبت کند. در سال ۱۷۷۳ میلادی، دانشمند روسی Christian Kratezenstein موفق به ساخت ماشینی برای تولید آواهای صدادار[۲] شد. مدتی بعد در سال ۱۷۹۱ میلادی Wolfgnag von Kempelen موفق به ساخت یک "ماشین گفتار آکوستیکی-مکانیکی[۳]" شد و در سال ۱۸۰۰ میلادی Charles Wheatstone یک نمونهٔ کاملتر ماشین صحبت کننده Kempelen را به کمک تشدید کننده‌هایی که از چرم ساخته شده بود ایجاد کرد. در این ماشین، همانطور که در شکل ۱-۲ نشان داده شده است، به کمک دست، صداهای مختلف شبیه به صدای انسان تولید می‌شد [۲].

بعدها در نیمه اول قرن بیستم میلادی Fletcher و همکارانش در آزمایشگاههای بل موفق شدند که رابطه بین طیف سیگنال گفتار[۴] و خصوصیات صدا را پیدا کنند. در سال ۱۹۳۰ میلادی Homer Dudley تحت تاثیر تحقیقات انجام شده توسط Fletcher قرار گرفت و موفق شد یک ترکیب کننده گفتار[۵] با عنوان VODER[۶] ایجاد کند. این ماشین جدید نمونه الکترونیکی ماشین مکانیکی سخنگوی Wheatstone بود که به صورت مکانیکی کنترل می‌شد [۲].

---

[2] Vowel Sounds
[3] Acoustic-Mechanical Speech Machine
[4] Speech Spectrum
[5] Speech Synthesizer
[6] Voice Operating Demonstrator



شکل ۱-۲ ماشین سخنگو ساخته شده توسط Wheatstone [۲].

افرادی مانند Harvery Fletcher و Homer Dudley به خوبی اهمیت طیف سیگنال را برای شناسایی دقیق آواهای موجود در سیگنال گفتار نشان دادند. متاثر از کارهای انجام شده توسط این دو دانشمند برجسته و همچنین سادگی محاسبه طیف توان به وسیله تکنیکهای موجود در پردازش سیگنال، اغلب سیستم‌های مدرن بازشناخت گفتار امروزی بر پایه اندازه‌گیری طیف توان سیگنال گفتار (یا نمونه‌های تغییر یافته آن مانند کپستروم) قرار گرفته‌اند [۲].

۱-۲-۱ دهه‌های ۱۹۵۰ و ۱۹۶۰ میلادی

بیشتر سیستم‌های بازشناخت گفتار از محتوای آوایی سیگنال گفتار استفاده می‌کنند. معمولا حدود ۳۰ آوای اصلی در زبان وجود دارد که از هر زبان به زبانی دیگر فرق می‌کند [۳]. در تلاشهای اولیه برای بازشناخت گفتار عمدتاً سعی بر این بود که روشی پیدا شود که بتوان اجزاء آوایی[۷] موجود در سیگنال صحبت را به صورت آکوستیکی شناسایی کرد. این اجزاء اصلی موجود در سیگنال صحبت

[7] Phonetic elements

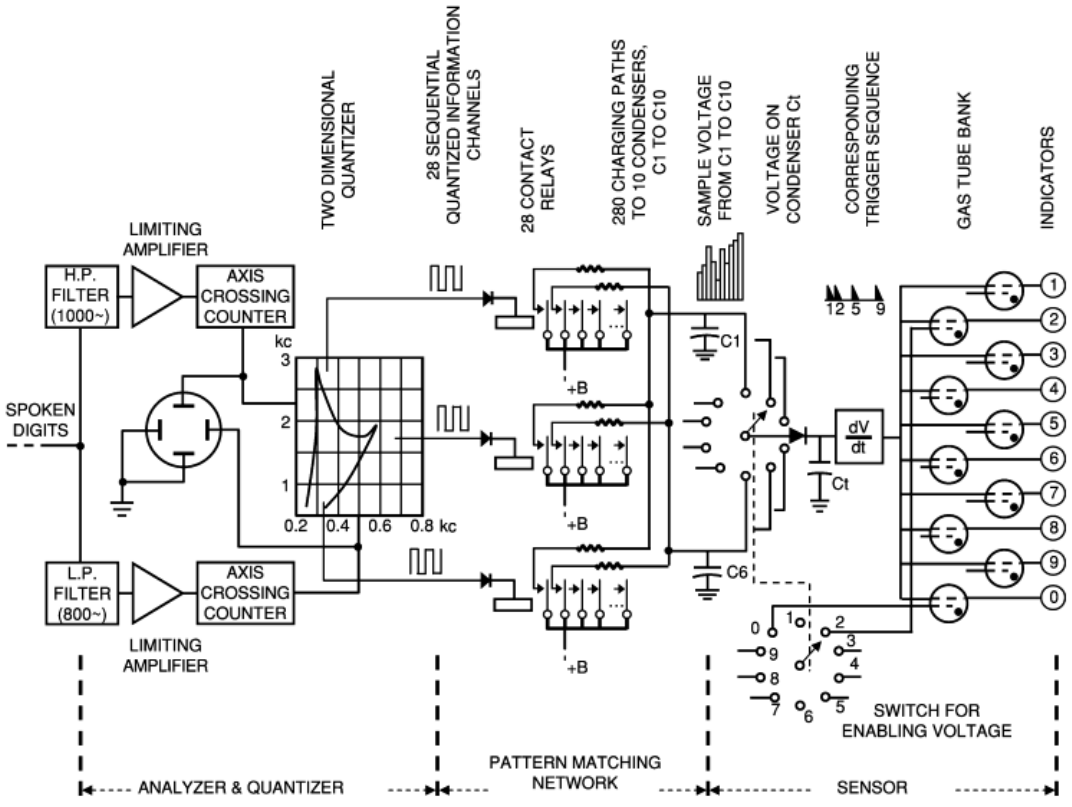
شامل آواها[8] می‌شود که محل و روش ادای آنها باعث تولید یک صدای بخصوص در متن‌های آوایی مختلف می‌شود. برای مثال برای تولید یک آوای صدادار[9]، تارهای صوتی[10] باید به ارتعاش در آیند تا مجرای صوتی[11] را تحریک و باعث انتشار هوا درون آن شوند، که نتیجه آن ایجاد صدای مورد نظر خواهد بود. نوع صدای ایجاد شده بستگی به فرکانسهای تشدید مجرای صوتی دارد. این فرکانسهای تشدید، که نشان دهنده محلهای اصلی تمرکز انرژی در طیف توان سیگنال گفتار هستند، فرکانسهای فرمنت[12] نامیده می‌شوند. در سال ۱۹۵۲ میلادی Davis، Biddulph و Balashek از آزمایشگاههای بل موفق به ساخت سیستمی برای تشخیص اعداد جدا از هم، مربوط به یک گوینده، شدند [۴]. در این سیستم از فرکانسهای فرمنت تخمین زده شده مربوط به ناحیه صدادار، مربوط به هر عدد، استفاده می‌شد. شکل ۱-۳ بلوک دیاگرام تشخیص دهنده عدد، ایجاد شده توسط Davis و همکارانش، و شکل ۱-۴ موقیت فرمنتهای ۱ و ۲ مربوط ده عدد صفر تا نه را نشان می‌دهد. از این نمودارها به عنوان الگوی مرجع برای معین کردن هویت عدد نامعلوم گفته شده استفاده می‌شده است [۲].

از دیگر سیستم‌های بازشناخت گفتاری که در دهه ۱۹۵۰ میلادی ساخته شد می‌توان به سیستم ساخته شده توسط Olson و Belar از آزمایشگاههای RCA، در سال ۱۹۵۶، اشاره کرد که قادر بود ۱۰ سیلاب[13] جدا از هم مربوط به یک شخص صحبت کننده را تشخیص دهد. در سال ۱۹۵۹ در کالج دانشگاهی انگلستان، Fry و Denes سعی در ساختن یک سیستم تشخیص دهنده آوا کردند که قادر بود ۴ حرف صدادار و ۹ حرف بیصدا را تشخیص دهد. در این سیستم با استفاده از اطلاعات آماری مربوط به رشته آواهای مجاز در انگلیسی توانستند دقت تشخیص آوا را، در کلمه‌هایی که دارای ۲ یا تعداد بیشتری آوا هستند، افزایش دهند. از این کار می‌توان به عنوان اولین کار، در

---

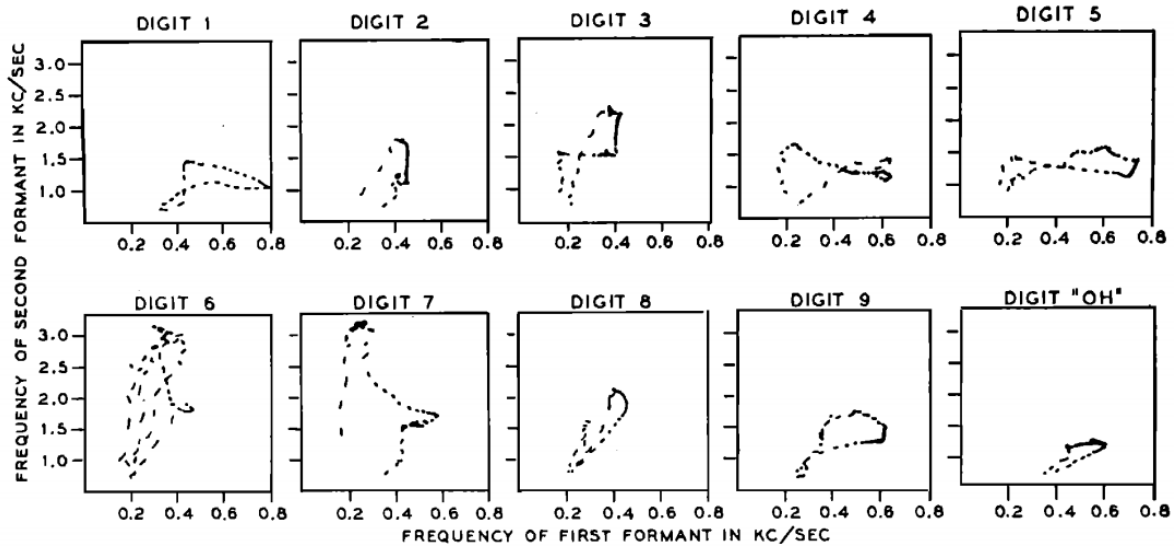
[8] Phoneme
[9] Vowel Sound
[10] Vocal Cords
[11] Vocal Tract
[12] Formant Frequencies
[13] Syllable

زمینه بازشناخت گفتار، نام برد که در آن از اطلاعات آماری مربوط به علم ترکیب[۱۴] (در سطح آوا) استفاده شده است [۵]. در همین سال توسط Forgie از دانشگاه لینکلن یک تشخیص دهنده، مستقل از گوینده، که قادر بود ۱۰ حرف صدادار را تشخیص دهد، ساخته شد [۶]. در دهه ۱۹۶۰ میلادی چندین آزمایشگاه ژاپنی توانایی خود را برای ساخت، سخت افزارهای با کاربرد مشخص، برای بازشناخت گفتار نشان دادند که مهمترین آنها تشخیص دهنده حروف صدادار Suzuki و Nakata از آزمایشگاه تحقیقاتی رادیو در ژاپن، تشخیص دهنده آوا مربوط به Sakai و Doshita از دانشگاه کیوتو [۷] و تشخیص دهنده عدد مربوط به آزمایشگاههای NEC می‌باشد. در سیستم ساخته شده توسط Sakai و Doshita برای اولین بار از جدا کنندهٔ سیگنال گفتار برای آنالیز و تشخیص در قسمتهای مختلف جمله ورودی استفاده شد و از آن می‌توان به عنوان پیشگام سیستم‌های بازشناخت گفتار پیوسته نام برد [۲].



شکل ۱-۳ بلوک دیاگرام مدار تشخیص دهنده عدد ساخته شده توسط Davis و همکارانش [۴].

[14] Statistical Syntax



شکل ۱-۴ تصویر فرمنت ۱ در مقابل فرمنت ۲ مربوط به عددهای صفر تا نه [۴].

یکی از مشکلهای اساسی که در سیستم‌های بازشناخت گفتار وجود داشت مسئله غیر یکنواختی مقیاس زمان برای همتراز کردن الگوهای گفتار بود. در دهه ۱۹۶۰ میلادی، Martin و همکارانش در آزمایشگاههای RCA مجموعه‌ایی از روشهای مقدماتی برای نرمالیزه کردن زمان را توسعه دادند که بر اساس آن قادر بودند نقطه‌های ابتدا و انتهای سیگنال گفتار را پیدا کنند. این امر به طور قابل ملاحظه‌ایی کارآیی سیستم را افزایش می‌داد. تقریبا در همان زمان Vintsyuk در اتحاد جماهیر شوروی استفاده از روش برنامه‌نویسی پویا[۱۵] را، برای همزمان کردن جمله‌های گفته شده، پیشنهاد کرد. اما کار او تا دهه ۱۹۸۰ میلادی در کشورهای دیگر تا حد زیادی ناشناخته ماند. تقریبا در همان زمان در تلاشی مستقل در ژاپن، Sakoe و Chiba، در آزمایشگاههای NEC، نیز شروع به استفاده از برنامه‌نویسی پویا برای حل مشکل غیر یکنواختی کردند [۸]. بعد از این استفاده از برنامه‌نویسی پویا برای حل مشکل غیر یکنواختی گسترش بیشتری پیدا کرد و روشهای مناسبتری پیشنهاد شدند که به طور کلی با عنوان پیچش زمانی پویا (DTW[۱۶])، در تطبیق الگوی گفتار[۱۷]،

[15] Dynamic Programming
[16] Dynamic Time Warping
[17] Speech Pattern Matching

شناخته شدند. از اواخر دهه ۱۹۷۰، برنامه‌نویسی پویا در شکلهای مختلف از جمله الگوریتم ویتربی[۱۸] [۹]، که از نظریه مخابرات می‌آمد، به یکی از تکنیکهای اجتناب ناپذیر در بازشناخت گفتار اتوماتیک تبدیل شد.

۱–۲–۲ دهه‌های ۱۹۷۰ و ۱۹۸۰ میلادی

در دهه‌های ۱۹۷۰ و ۱۹۸۰ میلادی پیشرفتهای چشمگیری در زمینه بازشناخت گفتار اتوماتیک بدست آمد که در ادامه به طور مختصر به آنها اشاره می‌شود.

۱–۲–۲–۱ دهه ۱۹۷۰

در اواخر دهه ۱۹۶۰ میلادی، Atal و Itakura مستقل از هم اساس روش کدینک پیشگویی خطی[۱۹](LPC) را توسعه دادند، که به طور قابل ملاحظه‌ایی محاسبه پاسخ فرکانسی مجرای صوتی را از روی شکل موج سیگنال گفتار راحت می‌کرد. در اواسط دهه ۱۹۷۰، ایده‌های اولیه در مورد اعمال تکنیک‌های تشخیص الگو، بر پایه روش LPC، در بازشناخت گفتار به وسیله افراد مختلفی از جمله Itakura [۱۰]، Rabiner و Levinson [۱۱] ارائه شد.

در این زمان بود که Tom Martin، کسی که در حل مشکل غیر یکنواختی جمله‌های گفتار نیز به موفقیت‌هایی دست پیدا کرده بود، اولین شرکت تجاری در زمینه بازشناخت گفتار را با عنوان Threshold Technology Inc. بنیان نهاد. این شرکت موفق شد که اولین سیستم بازشناخت گفتار اتوماتیک واقعی را با عنوان VIP-100 ایجاد کند. از این سیستم فقط در چند کاربرد ساده استفاده شد

[18] Viterbi Algorithm
[19] Linear Predictive Coding

ولی اهمیت آن در آنجا بود که توانست موسسه پروژه‌های پژوهشی پیشرفته دفاعی (DARPA[۲۰]) مربوط به وزارت دفاع آمریکا را تحت تاثیر قرار دهد تا پژوهشی در زمینه درک گفتار (SUR[۲۱])، در اوایل دهه ۱۹۷۰، آغاز کند. از جمله سیستم‌هایی که تحت حمایت DARPA ساخته شد می‌توان به سیستم ساخته شده توسط دانشگاه کارنگی ملون با عنوان "Harpy" نام برد که قادر بود گفتار را، با فرهنگ لغتی بالغ بر ۱۰۱۱ کلمه، با دقت قابل قبولی تشخیص دهد. یکی از دستاوردهای مهم سیستم Harpy انجام عمل جستجو از طریق گراف بود. در این روش برای بازشناخت گفتار، کلام به صورت شبکه متصل بهمی، که از روی نمایش حرفی کلمات[۲۲] و همچنین قواعد نحوی[۲۳] به همراه نقاط ابتدا و انتهای کلمات به دست آمده بود، نشان داده می‌شد. در سیستم پیشنهادی Harpy، گفتار ورودی، بعد از گذشتن از مرحله تجزیه و تحلیل پارامتری به قسمتهای جداگانه تقسیم می‌شد و هر رشته پارامتری جدا شده، مربوط به گفتار، به وسیله فاصله ایتاکورا[۲۴] [۱۰] به الگوی آوایی خاصی نسبت داده می‌شد. سیستم Harpy، اولین سیستمی بود که از مزیت شبکه با حالت محدود (FSN[۲۵])، برای کاهش حجم محاسبات و افزایش کارآیی در پیدا کردن نزدیکترین رشته، استفاده می‌کرد.

سیستم‌های دیگری که در تحت پروژه SUR مربوط به DARPA ساخته شدند عبارتند از Hearsay II از دانشگاه کارنگی ملون و HWIM[۲۶] ساخته شده توسط شرکت BBN[۲۷] که هیچ یک از این دو سیستم در پایان پروژه ۵ ساله DARPA، در سال ۱۹۷۶، به هدف‌های تعیین شده توسط آن دست پیدا نکردند. تقریبا در همان زمان که پروژه DARPA در حال تمام شدن بود کار ساختن سیستم جدیدی در دانشگاه کارنگی ملون توسط Jim Baker با نام سیستم DRAGON آغاز شد

[20] Defense Advanced Research Projects Agency
[21] Speech Understanding Research
[22] lexical
[23] Syntactic
[24] Itakura Distance
[25] Finite State Network
[26] Hear What I Mean
[27] Bolt Beranek and Newman

[۱۲]. Jim Baker در اوایل دهه ۱۹۸۰ به Massachusetts رفت و در آنجا شرکت جدیدی را با نام DRAGON ایجاد کرد.

در دهه ۱۹۷۰ همزمان با DARPA تحقیقات در زمینه بازشناخت گفتار در آزمایشگاههای بل[۲۸] و IBM[۲۹] نیز شروع به شکل گیری کرد. این دو مرکز با دو تفکر متفاوت شروع به تحقیقات در زمینه سیستم‌های بازشناخت گفتار برای مصارف تجاری کردند.

تلاشهای IBM که توسط Fred Jelinek هدایت می‌شد منجر به ساخت یک ماشین تحریر فعال شونده با صدا (VOT[۳۰]) شد. کار اصلی این ماشین تبدیل کردن جمله‌های گفته شده به رشته‌ایی از حروف و کلمات بود که بر روی صفحه نمایش نشان داده می‌شد و یا بر روی کاغذ تایپ می‌شد [۱۳]. این سیستم بازشناخت گفتار، که Tangora خوانده می‌شد، در حقیقت یک سیستم وابسته به گوینده[۳۱] بود (ماشین تحریر باید با صدای هر کاربر به طور جداگانه آموزش داده می‌شد). در این سیستم تمرکز فنی بر روی دو چیز بود یکی اندازه فرهنگ لغت (تا آنجا که امکان دارد بزرگ باشد) و دیگری ساختار مدل زبان[۳۲] (مجموعه‌ایی از قواعد نحوی آماری[۳۳] مدل زبان خوانده می‌شد). در این سیستم مدل زبان نشان می‌داد که با چه احتمالی یک رشته از نمادهای زبان (آواها و کلمه‌ها) در سیگنال صحبت ظاهر می‌شوند. یکی از پر کاربردترین نمونه‌های مدل زبان، مدل n-gram است که در آن احتمال وقوع رشته‌ایی از مرتبه n کلمه تعریف شده است [۲].

در آزمایشگاههای بل بر خلاف IBM، که هدف ساخت یک ماشین تحریر صوتی وابسته به گوینده بود، هدف ساخت یک سیستم بازشناخت گفتار مستقل از گوینده[۳۴] بود [۱۴] که قادر باشد با استفاده از مخابرات راه دور سرویس‌هایی مثل شماره‌گیری از طریق صدا را در اختیار همه مردم قرار

---

[28] AT&T Bell Laboratories
[29] International Business Machines Corporation
[30] Voice-activated Typewriter
[31] Speaker-dependent System
[32] Language Model
[33] Statistical Syntactical Rules
[34] Speaker-independent System

دهد. در سیستم‌های ساخته شده توسط آزمایشگاههای بل هدف این بود که سیستم بتواند به خوبی برای تعداد وسیعی از گوینده‌ها (ده‌ها میلیون نفر) کار کند، بدون اینکه نیاز باشد دستگاه را برای هر گوینده به صورت جداگانه آموزش داد. به عبارت دیگر تمرکز بر روی ساخت سیستمی بود که بتواند با تغییرات صوتی ذاتی موجود در سیگنال گفتار مربوط به گوینده‌های مختلف با لهجه‌های متفاوت مقابله کند. این امر منجر به ایجاد طیف وسیعی از الگوریتم‌های دسته‌بندی[۳۵] گفتار برای ایجاد الگوهای مرجع صداها و کلمه‌ها (در ابتدا قالبها[۳۶] ولی در نهایت مدل‌های آماری) شد که قادر بودند برای محدوده وسیعی از گوینده‌ها و لهجه‌ها مورد استفاده قرار گیرند. علاوه بر این تحقیقات در زمینه درک و کنترل تغییرات صوتی، مربوط به نمایش‌های مختلف سیگنال گفتار مربوط به گوینده‌های متفاوت، منجر به بررسی تعداد زیادی از روشهای اندازه‌گیری فاصله طیفی (مثل فاصله ایتاکورا [۱۰]) و همچنین تکنیکهای مدل کردن آماری [۱۵] شد که می‌توانستند نمایش غنی از جمله‌های بیان شده فراهم کنند. از آنجایی که معمولا کاربردهایی مثل شماره‌گیری صوتی شامل جمله‌های کوتاه با تعداد لغات کم می‌شود، بیشترین تمرکز در آزمایشگاههای بل بر روی مدل صوتی[۳۷] (نمایش طیفی صداها و کلمه‌ها) بود تا مدل زبان (نمایش نحوی یا گرامری کاری که باید انجام شود). همچنین یکی از مهمترین روشهای مورد استفاده، در آزمایشگاههای بل، برای دستیابی به هدف مورد نظر استفاده از مفهوم معین کردن کلمه کلیدی[۳۸]، برای تشخیص مفهوم جمله گفته شده، بود [۱۶]. هدف این روش پیدا کردن یک کلمه یا عبارت کلیدی، که دارای اهمیت به خصوصی است، در یک جمله بزرگتر بود. سیستم بازشناخت گفتار به وسیله این کلمه یا عبارت کلیدی درخواست گوینده را متوجه می‌شد و متناسب با آن به کاربر جواب و یا سرویس مورد نظر را ارائه می‌داد.

---

[35] Clustering Algorithm
[36] Template
[37] Acoustic Model
[38] Keyword Spotting

۱-۲-۲-۲ دهه ۱۹۸۰

در دهه ۱۹۸۰ بیشتر تمرکز بر روی ساخت سیستمی بود که قادر باشد رشته‌ایی از کلمه‌های متصل به هم (مثل عددها) را که به صورت روان[۳۹] گفته شده است، تشخیص دهد. یکی از مهمترین کارهایی که در این دهه برای رسیدن به این هدف انجام شد جایگزینی روش ساده و مستقیم تطبیق الگو[۴۰]، با روش سخت مدل کردن آماری برای تشخیص الگو بود. دو روش آماری که در این دوره در زمینه بازشناخت گفتار مورد توجه قرار گرفت و در سالهای بعد توسعه پیدا کرد عبارتند از مدلهای مخفی مارکوف (HMM[۴۱]) و شبکه‌های عصبی مصنوعی (ANN[۴۲]).

کاربرد مدلهای مخفی مارکوف در بازشناخت گفتار در ابتدا در اواسط دهه ۱۹۷۰ توسط افرادی مثل Jim Baker از دانشگاه کارنگی ملون [۱۲]، تیم بازشناخت گفتار موجود در IBM [۱۳] و همچنین موسسه IDA[۴۳] در پرینستون معرفی شد. بعد از سمینارهای متعددی که در اوایل دهه ۱۹۸۰ در دانشگاه پرینستون در مورد کاربرد مدلهای مخفی مارکوف در بازشناخت گفتار برگذار شد آزمایشگاههای بل کار خود را در زمینه کاربرد مدلهای مخفی مارکوف در بازشناخت گفتار آغاز کرد، که نتیجه آن انتشار دو مقاله در این زمینه بود [۱۷،۱۸]. در این دو مقاله تئوری پایه مدلهای مخفی مارکوف توضیح و هچنین با یک مثال طریقه اعمال آنها به مسائل مربوط به بازشناخت گفتار شرح داده شده بود. بعد از انتشار این دو مقاله مهم، با حمایت DARPA، استفاده از مدلهای مخفی مارکوف در بازشناخت گفتار به سرعت گسترش پیدا کرد.

مدل مخفی مارکوف یک فرآیند آماری دولایه است که فرآیند آماری زیرین قابل مشاهده نیست (دلیل استفاده از واژهٔ مخفی) ولی به کمک یک فرآیند آماری دیگر که رشته قابل مشاهده

---

[39] Fluent
[40] Template based
[41] Hidden Markov Models
[42] Artificial Neural Networks
[43] Institute for Defense Analyses

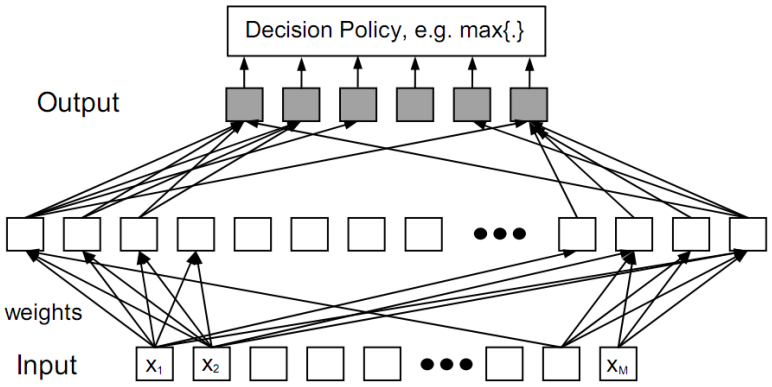
خروجی را ایجاد می‌کند می‌توان در مورد فرآیند آماری مخفی نظر داد. مدل مخفی مارکوف قادر است با قدرت نسبتاً بالایی سیگنال صحبت را که دارای ذاتی تغییر پذیر است مدل کند [۱۹].

یکی دیگر از تکنولوژی‌هایی که در دهه ۱۹۸۰ در حوزه بازشناخت گفتار مطرح شد، شبکه‌های عصبی مصنوعی بود که در ابتدا در دهه ۱۹۵۰ معرفی شده بود ولی نتوانسته بود نتایج قابل قبولی را بدهد. در دهه ۱۹۸۰ به خاطر پیدایش مدل پردازش توزیع یافته موازی (PDP[۴۴])، که شامل مجموعه‌ایی انبوه از المانهای پردازشی می‌شد، به همراه روش آموزش مربوط به آن با عنوان انتشار خطا به عقب[۴۵]، علاقه‌ها را برای تقلید از مکانیزم پردازش عصبی موجود در انسان زنده کرد. در آن زمان نوعی از PDP با نام پرسپترون چند لایه[۴۶]، که در شکل ۱-۵ نشان داده شده است، به خاطر توانایی‌اش در تخمین هر تابع ورودی، با دقت دلخواه، بیشترین توجه‌ها را متوجه خودش کرد. به عبارت دیگر اگر به دنبال یک سیستم باشیم که در آن نمونه ورودی توسط یک تابع به کلاس مربوطه‌اش تعلق پیدا کند، پرسپترون چند لایه یکی از کاندیدهای خوب برای این منظور می‌باشد. تلاشهای اولیه برای استفاده از شبکه‌های عصبی در بازشناخت گفتار فقط شامل کارهای ساده مثل تشخیص چند آوا یا چند کلمه می‌شد که اغلب موفقیت آمیز نیز بود. اما از آنجا که در سیستم‌های بازشناخت گفتار به طور اجتناب ناپذیری احتیاج بود که تغییرات زود گذر موجود در سیگنال گفتار مهار شود، شبکه‌های عصبی با شکل اولیه‌شان قادر به انجام این کار نبودند. برای مقابله با این مشکل شبکه عصبی از توانایی مدل مخفی مارکوف در مهار کردن تغییرات زود گذر سیگنال گفتار در شبکه عصبی استفاده شد و سیستم‌های بازشناخت گفتاری ایجاد شد که در آنها هم از شبکه عصبی و هم از مدل مخفی مارکوف استفاده شده بود [۲].

از دیگر کارهایی که در دهه ۱۹۸۰ انجام شد می‌توان به روش پیشنهادی Furui اشاره کرد، که در آن علاوه بر ضرایب لحظه‌ایی کپسترال از مشتق اول و دوم آنها نیز به عنوان ویژگی‌های طیفی

[44] Parallel Distributed Processing
[45] Error Back-propagation
[46] Multilayer Perceptron

برای بازشناخت گفتار استفاده کرده بود [۲۰]. این روش پیشنهادی هم اکنون به صورت گسترده‌ایی در تمام سیستم‌های بازشناخت گفتار استفاده می‌شود.



شکل ۱-۵ پرسپترون چند لایه [۲].

در این دهه نیز DARPA هدایت تحقیقات در زمینه ساخت سیستم‌های بازشناخت گفتار پیوسته با تعداد لغات زیاد را بر عهده گرفت که باعث بدست آمدن پیشرفتهای زیاد و به دست آمدن نتایج خوبی در بازشناخت گفتار با فرهنگ لغت ۱۰۰۰ کلمه‌ایی شد. بیشتر نتایج بدست آمده حاصل تلاش دانشگاهها و آزمایشگاههای زیر بود. دانشگاه کارنگی ملون با سیستم SPHINX [۲۱]، BBN با سیستم BYBLOS [۲۲]، مرکز تحقیقاتی استنفورد (SRI[۴۷]) با سیستم DECIPHER [۲۳]، آزمایشگاههای لینکلن [۲۴]، موسسه فناوری ماساچوست (MIT[۴۸]) [۲۵] و آزمایشگاههای بل [۲۶].

۱-۲-۳ دهه‌های ۱۹۹۰ و ۲۰۰۰ میلادی تاکنون

در دهه ۱۹۹۰ در زمینه تشخیص الگو پیشرفتهایی بدست آمد. مسئله تشخیص الگو که معمولا در چهارچوب تئوری تصمیم‌گیری بیز[۴۹] دنبال می‌شد، که در آن احتیاج به تخمین توزیع داده‌ها بود، تبدیل شد به یک مسئله بهینه‌سازی که در آن هدف مینیمم کردن خطای تجربی بازشناخت بود

[47] Stanford Research Institute
[48] Massachusetts Institute of Technology
[49] Bayes' Decision Theory

[۲۷]. علت به وجود آمدن این تغییر نیز از آنجا بود که تابع توزیع مربوط به سیگنال صحبت با دقت خوبی قابل تخمین نبود و در نتیجه تئوری تصمیم‌گیری بیز غیر قابل استفاده می‌شد. مفهوم حداقل خطای تجربی منجر به بوجود آمدن تکنیک‌های جدیدی شد که از میان آنها آموزش وابسته به تبعیض[۵۰] مثل جداسازی با خطای حداقل (MCE[۵۱]) و روشهای بر پایه کرنل[۵۲] مثل ماشین بردار پشتیبانی (SVM[۵۳]) دارای عمومیت بیشتری بودند [۲].

DARPA در دهه‌های ۱۹۹۰ و ۲۰۰۰ به منظور ایجاد انگیزه در ساخت سیستم‌های جدید که دارای دقت و قدرت بالایی باشند محدوده وسیعی از ماموریتها را در زمینه بازشناخت گفتار و درک زبان طبیعی[۵۴] انتخاب و تعریف کرد. برای ارزیابی سیستم‌های ساخته شده بیشتر از نرخ خطای کلمه به عنوان معیار اندازه‌گیری کارآیی سیستم[۵۵] استفاده می‌شد.

یک سری از این ماموریتها که در طول ۱۰ تا ۱۵ سال توسط DARPA، در حوزه بازشناخت گفتار، معرفی شد در زیر آمده است. هر یک از این ماموریتها نسبت به قبلی دارای پیچیدگی و سختی بیشتری بود [۲۸].

Resource Management (RM): این ماموریت برای استفاده در محیط‌های نظامی، برای تحقیق در مورد موقعیت و وضعیت کشتی‌ها در نقاط مختلف کره زمین، بود. تعداد لغات تقریبا ۱۰۰۰ کلمه و بهترین نرخ خطای کلمه‌ایی که بعد از پایان این ماموریت به دست آمد برابر با ۲ درصد بود.

Airline Travel Information System (ATIS): در این ماموریت به کاربر این توانایی را می‌داد که از طریق یک گفتگوی دو طرفه با سیستم، مشابه با چیزی که در حالت واقعی وجود دارد، با توجه به جدول زمانی پروازها در خطوط هوایی، یک برنامه پرواز را تنظیم کند. فرهنگ لغت به کار

---

[50] Discriminative Training
[51] Minimum Classification Error
[52] Kernel-based Method
[53] Support Vector Machines
[54] Natural Language Understanding
[55] Figure of Merit

رفته در این حالت دارای ۲۵۰۰ لغت بود و نرخ خطای کلمه بدست آمده در پایان ماموریت برابر با ۲/۵ درصد بوده است.

North American Business (NAB): در این ماموریت کاربر قادر بود که یک داستان در مورد کسب و کار را که از یکی از روزنامه‌ها مثل وال استریت ژورنال انتخاب شده است بخواند. به عبارت دیگر جمله‌ها به طور مستقیم از روی روزنامه خوانده می‌شد. در این حالت فرهنگ لغت دارای ۶۴۰۰۰ کلمه بود و نرخ خطای کلمه به دست آمده برابر با ۶/۶ درصد بوده است.

Broadcast News (BN): در این ماموریت واژه‌ها و جمله‌ها از شبکه‌های خبری مثل CNBC[۵۶] در زمان پخش خبر انتخاب شده بودند و هدف آن تبدیل صحبت به متن به صورت زنده بود به طوری که متن به دست آمده از سیگنال گفتار در پایین صفحه تلویزیون نمایش داده شود. در این حالت فرهنگ لغت دارای ۲۱۰۰۰۰ کلمه بود و نرخ خطای کلمه‌ی بدست آمده بین ۱۳ تا ۱۷ درصد بوده است.

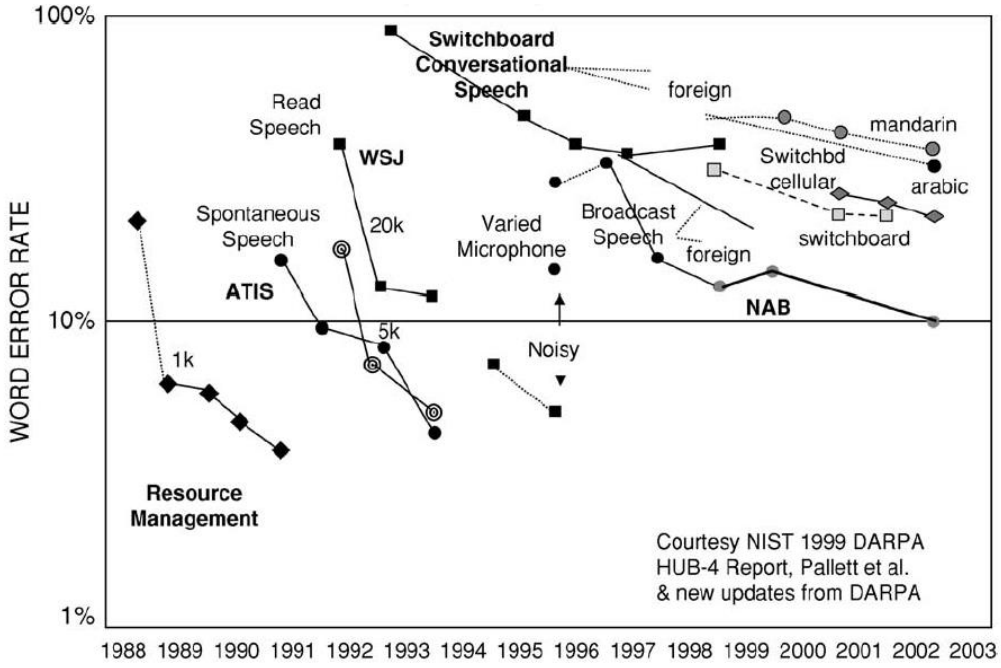
Switchboard (SB): در این ماموریت از گفتار موجود در صفحه سوئیچ مرکز مخابرات استفاده شده بود. به عبارت دیگر دراین حالت صحبت معمولی که در یک ارتباط تلفنی بین دو نفر رد و بدل می‌شود با همان کیفیت و پهنای باند مورد استفاده قرار گرفته بود. اندازه فرهنگ لغت ۴۵۰۰۰ کلمه و درصد خطای کلمه بدست آمده بین ۲۵ تا ۲۹ درصد بوده که نشان دهنده سختی آن است.

Call Home (CH): در این ماموریت از صحبت معمولی که بین دو خانه از طریق شبکه تلفن رد و بدل می‌شود استفاده شده بود. در این حالت دو شخص صحبت کننده معمولا با هم نسبت فامیلی و دوستی دارند و از جمله‌هایی استفاده می‌کنند که اغلب کوتاه هستند و شکل دستوری کاملی ندارند. فرهنگ لغت بکار رفته در این سیستم دارای تقریبا ۲۸۰۰۰ کلمه و نرخ خطای بدست آمده در حدود ۴۰ درصد بوده است.

[56] Consumer News and Business Channel

از ماموریت‌های انجام شده در بالا می‌توان نتیجه گرفت که بازشناخت گفتار محاوره‌ایی که اغلب پای‌بند ساختار و قالب مشخصی نیست بسیار مشکل‌تر از حالتی است که از روی یک متن روخوانی می‌کنیم. همچنین برنامه DARPA نشان داد که افزایش داده‌های گفتار آموزشی، باعث بهتر شدن نرخ خطا می‌شود.

شکل ۱-۶ نمودار مربوط به کارآیی بعضی از سیستم‌های بازشناخت گفتاری که به آنها اشاره شد را در طول زمان نشان می‌دهد. نکته‌ایی که از روی این نمودارها می‌توان به آن پی برد این است که در طول زمان کارآیی تمام ماموریت‌ها با نرخ تقریبا مشترکی بهبود پیدا کرده است.



شکل ۱-۶ ارزیابی بدست آمده توسط DARPA مربوط به بعضی از ماموریت‌های بازشناخت گفتار [۲].

یکی از مهمترین آثار برنامه‌های DARPA ظهور نرم‌افزارهای پیچیده‌ایی بود که به سرعت تبدیل به یکی از ابزارهای اجتناب ناپذیر در تحقیقات پیشرفته، برای ایجاد مفاهیم و الگوریتم‌های جدید، شدند. خوشبختانه بعضی از تولید کنندگان این ابزارهای نرم‌افزاری کد منبع[۵۷] برنامه خود را

[57] Source Code

برای استفاده در محیط‌های تحقیقاتی در اختیار بقیه قرار دادند. مهمترین جعبه‌های ابزار نرم‌افزاری[58] که در زمینه بازشناخت گفتار و درک زبان طبیعی کاربرد دارند عبارتند از:

۱- جعبه ابزار مدل مخفی مارکوف (HTK[59])، مربوط به دانشگاه کمبریج [۲۹].

۲- Sphinx مربوط به دانشگاه کارنگی ملون [۳۰].

۳- جعبه ابزار مدل مخفی مارکوف مربوط به دانشگاه ایالتی میسیسیپی [۳۱].

یکی دیگر از مشکلاتی که در سیستم‌های بازشناخت گفتار وجود داشت و هنوز هم حل نشده است وجود ناهماهنگی بین شرایط آموزش و تست است که باعث کاهش شدید نرخ بازشناخت گفتار در عمل می‌شود. بعضی از عواملی که باعث به وجود آمدن این ناهماهنگی می‌شوند عبارتند از: نویز پس زمینه، تغییرات صدا از یک شخص به شخص دیگر، میکروفن، کانال انتقال، انعکاس صدا در اتاق و غیره. در دهه‌های ۱۹۹۰ و ۲۰۰۰ تکنیک‌های زیادی برای افزایش قدرت سیستم‌های بازشناخت گفتار پیشنهاد شده‌اند که بعضی از آنها عبارتند از: ماکزیمم احتمال برگشت خطی (MLLR[60]) [۳۲]، تجزیه مدل[61] [۳۳]، ترکیب مدل موازی (PMC[62]) [۳۴] و وابستگی بهبود یافته زمان کوتاه (SMC[63]) [۳۵].

یکی دیگر از مباحثی که در دهه ۱۹۹۰ و ۲۰۰۰ برای افزایش نرخ بازشناخت به آن پرداخته شد، بازشناخت گفتار چند وجهی[64] است که در آن هم از اطلاعات تصویر مثل حالت صورت و لبها و هم از اطلاعات صوتی برای بازشناخت گفتار استفاده شده است. استفاده از اطلاعات تصویر زمانی حیاتی می‌شود که پیغام ارسال شده به درستی از طریق صوت قابل انتقال نباشد مثل زمانی که در محیط‌های نویزی قرار گرفته باشیم. تحقیقات نشان داده است که استفاده از اطلاعات تصویر و صوت

---

58 Software Toolkit
59 Hidden Markov Model Toolkit
60 Maximum Likelihood Linear Regression
61 The Model Decomposition
62 Parallel Model Combination
63 Short-time Modified Coherence
64 Multimodal Speech Recognition

با هم نتایج بهتری نسبت به حالتی دارد که فقط از اطلاعات صوت یا تصویر به تنهایی استفاده می‌شود.

۱–۳ بیان مسئله و فصل بندی مطالب

همانطور که در قسمت قبل بیان شد با وجود پیشرفت‌های چشم‌گیری که در زمینهٔ بازشناخت گفتار بدست آمده است هنوز قادر به ساخت سیستمی که بتواند به خوبی انسان و با درصد بسیار بالا جمله‌ها و کلمه‌ها را تشخیص دهد نیستیم. این امر لزوم تحقیقات بیشتر در این زمینه را آشکار می‌کند.

یکی از دلایلی که باعث کاهش کارآیی سیستم‌های بازشناخت گفتار می‌شود وجود ناهماهنگی بین شرایط آموزش و تست است که یکی از مهمترین عواملی که باعث به وجود آمدن این ناهماهنگی می‌شود وجود نویز جمع شونده در محیط آزمایش می‌باشد. در این پایان‌نامه هدف ارائهٔ روشی برای استخراج ویژگی می‌باشد که علاوه بر داشتن کارآیی خوب در محیط بدون نویز نسبت به نویز جمع شونده نیز دارای مقاومت خوبی باشد. برای این منظور از حوزهٔ اتوکرولیشن و خواص تابع اتوکرولیشن استفاده شده است و سعی شده که تعدادی از روش‌های موجود فعلی بهبود داده شود.

در ادامه در فصل دوم در ابتدا با پیچیدگی‌های موجود در سیستم‌های بازشناخت گفتار آشنا می‌شویم. سپس سه بخش کلی موجود در سیستم‌های بازشناخت گفتار و روش‌های متداول برای استخراج ویژگی مثل MFCC و PLP معرفی شده است.

در فصل سوم در ابتدا تئوری اولیهٔ روش آماری مدل مخفی مارکوف و سه مسئلهٔ اصلی موجود در این روش و راه حل آنها آورده شده است. سپس کاربرد این روش آماری و طریقهٔ استفاده از آن در

سیستم‌های بازشناخت گفتار به اختصار توضیح داده شده است. همچنین در این فصل بستهٔ نرم افزاری مدل مخفی مارکوف (HTK) که برای ساخت سیستم‌های بازشناخت گفتار، بر پایهٔ مدل‌های مخفی مارکوف، به طور گسترده مورد استفاده قرار می‌گیرد، معرفی شده است.

در فصل چهارم بعد از معرفی بعضی از خواص تابع اتوکرولیشن، به طور خلاصه به مرور بعضی از کارهای انجام شده برای استخراج ویژگی در حوزهٔ اتوکرولیشن مثل روش‌های رشتهٔ اتوکرولیشن نسبی، مشتق رشتهٔ اتوکرولیشن و فاز اتوکرولیشن پرداخته شده است.

در فصل پنجم روش‌های پیشنهادی مشتق طیف توان در حوزهٔ اتوکرولیشن (DPSA) و مشتق طیف اتوکرولیشن فاز (DPAS) برای استخراج ویژگی در حوزهٔ اتوکرولیشن ارائه شده‌اند. در فصل ششم نتایج آزمایش‌های انجام شده در محیط‌های بدون نویز و نویزی در حضور چهار نویز مختلف آورده شده است. سپس نتایج بدست آمده از روش‌های پیشنهادی اول و دوم به طور جداگانه با بعضی از روش‌های دیگر مقایسه شده است.

فصل دوم

معرفی اجزاء سیستم‌های بازشناخت

گفتار

فصل دوم: معرفی اجزاء سیستم‌های بازشناخت گفتار

۱-۲ ۴-۱ مقدمه

یک سیستم بازشناخت گفتار سیستمی است که به صورت خودکار گفتار را به متن تبدیل می‌کند. به عبارت دیگر به فرآیند نسبت دادن یک کلمه، که در بانک واژگان سیستم بازشناخت وجود دارد، به سیگنال آکوستیکی که توسط دستگاه ضبط شده است بازشناخت گفتار گفته می‌شود.

برای بیشتر مردم گفتار طبیعی‌ترین و مؤثرترین وسیله انتقال اطلاعات است. هدف نهایی تکنولوژی بازشناخت گفتار ساخت ماشینی است که بتواند اطلاعات گفته شده را از طریق سیگنال صوت دریافت کند و متناسب با آنها عمل لازم را انجام دهد. بنابراین مسئله بازشناخت گفتار قسمتی از تحقیق در زمینهٔ ساخت ماشینهای با هوش مصنوعی[۶۵] است که دارای توانایی‌های شنیدن، درک کردن و انجام عمل متناسب با اطلاعات دریافت شده و همچنین صحبت کردن برای تبادل اطلاعات

[65] Artificially Intelligent

هستند. با وجود پیشرفتهای بدست آمده هنوز توانایی ساخت ماشینی که مستقل از محیط و شخص صحبت کننده به خوبی انسان عمل تشخیص را انجام دهد به دست نیامده است.

۱–۵ ۲–۲ ابعاد مختلف سیستم‌های بازشناخت گفتار

سیستم‌های بازشناخت گفتار با توجه به کاربردی که دارند دارای انواع مختلفی هستند که مشخص کننده میزان پیچیدگی آنها می‌باشد. بعضی از این ابعاد که بیشتر مورد توجه قرار دارند عبارتند از [۳۶]:

- بازشناخت کلمات مجزا[۶۶] در مقابل بازشناخت گفتار پیوسته[۶۷]
- اندازه بانک واژگان[۶۸]
- وابسته یا مستقل بودن از گوینده
- محدودیت‌های زبان شناختی[۶۹]
- ابهامات و اشتباهات آکوستیکی[۷۰]
- نویز محیطی

در ادامه هر یک از این ابعاد مورد بررسی قرار گرفته می‌شود.

---

66 Isolated-Word Recognition
67 Continuous-Speech Recognition
68 Vocabulary Size
69 Linguistic Constraints
70 Acoustic Ambiguity and Confusability

۲-۲-۱ بازشناخت کلمات مجزا در مقابل بازشناخت گفتار پیوسته

نحوه صحبت کردن گوینده باید متناسب با نوع سیستم بازشناخت گفتار باشد. معمولا دو نوع سیستم بازشناخت گفتار با توجه به گسسته یا پیوسته بودن گفتار تعریف می‌شود که در ادامه هر یک از آنها به طور مختصر معرفی شده است.

۲-۲-۱-۱ بازشناخت کلمات مجزا

در این حالت سیستم بازشناخت با هر یک از اجزاء مجزا به طور جداگانه آموزش دیده می‌شود. از آنجاییکه در این حالت اجزاء مجزا معمولا کلمه‌ها هستند از لفظ سیستم بازشناخت کلمات مجزا استفاده می‌شود. در این حالت در مرحله بازشناخت فرض بر این است که شخص گوینده وقتی جمله‌ایی را ادا می‌کند حداقل ۲۰۰ میلی ثانیه بین کلمه‌ها مکث می‌کند تا سیستم بازشناخت بتواند به راحتی آنها را از هم جدا کند. تشخیص یک کلمه تنها هم جزء این نوع سیستم بازشناخت گفتار محسوب می‌شود [۳۷].

در این نوع سیستم‌ها برای انجام عمل تشخیص در ابتدا بر روی کلمه گفته شده آنالیز لازم انجام می‌شود و مدل بدست آمده از آن با مدل‌های موجود در بانک واژگان مقایسه می‌شود. این نوع از سیستم‌های بازشناخت گفتار برای کاربردهایی که حالت فرمانی دارند مناسب می‌باشند.

۲-۲-۱-۲ بازشناخت گفتار پیوسته

سیستم‌های بازشناخت گفتار پیوسته دارای پیچیدگی خیلی بیشتری نسبت به نمونه قبلی هستند. این سیستم‌ها باید قادر باشند گفتار پیوسته را که به صورت روان بیان شده است تشخیص دهند. در

این حالت سیستم بازشناخت باید اولا قادر باشد به طریقی نقاط ابتدا و انتهای کلمه‌ها را در سیگنال صوتی تشخیص دهد ثانیا باید بتواند از پس کلماتی که به درستی تلفظ نشده‌اند، یا کلماتی که دارای تلفظ شبیه به هم ولی شکل متفاوت و یا شکل یکسان ولی تلفظ متفاوت، در نقاط مختلف جمله، هستند بر آید.

۲-۲-۲ اندازه بانک واژگان

معمولا اندازه بانک واژگان با دقت سیستم نسبت عکس دارد به طوری که هر چه اندازه بانک واژگان کمتر باشد سرعت و دقت سیستم بازشناخت بیشتر می‌شود. بعضی از تخمین‌های زده شده نشان داده است که میزان دشواری سیستم بازشناخت گفتار به صورت لگاریتمی با اندازه بانک واژگان افزایش پیدا می‌کند [۳۷].

معمولا سیستم‌های بازشناخت گفتار از نظر اندازه بانک واژگان به سه دستهٔ، کوچک، متوسط و بزرگ تقسیم می‌شوند. تعریفهای زیادی برای ملاک این تقسیم‌بندی شده است ولی با یک حساب سر انگشتی می‌توان گفت که سیستم‌های با اندازه بانک واژگان کوچک دارای تعداد لغت بین ۹۹-۱ هستند، اندازه متوسط دارای ۹۹۹-۱۰۰ لغت و بزرگ دارای بیشتر از ۱۰۰۰ لغت می‌باشند. البته تقسیم بندی یک امر نسبی است و ممکن است این تقسیم بندی درست نباشد. به عنوان مثال ممکن است یک سیستم با تعداد ۱۰۰۰ لغت در مقابل یک سیستم با ۲۰۰۰۰۰ لغت کوچک محسوب شود.

سیستم‌های با اندازه بانک واژگان کوچک معمولا در جاهایی کاربرد دارند که احتیاج به یک سیستم مستقل از گوینده برای انجام یک کار مشخص است مثل وارد کردن شماره کارت اعتباری یا تشخیص شماره گفته شده برای شماره‌گیری بوسیله صحبت کردن. سیستم‌های با اندازه بانک واژگان متوسط معمولا در تحقیقات برای ساخت سیستم‌های بازشناخت گفتار پیوسته استفاده می‌شود

(بوسیله دیتابیس‌های موجود). سیستم‌های با اندازه بانک واژگان بزرگ دارای کاربردهای تجاری هستند مثل بازیابی پرونده‌ها در این حالت سیستم‌ها فقط برای یک یا چند کاربر به خصوص ساخته می‌شوند. سیستم‌های با بانک واژگان بزرگ معمولا از نوع سیستم‌های بازشناخت کلمه مجزا هستند. به عبارت دیگر کاربر باید کلمه‌ها را با یک مکث کوتاه بین آنها بیان کند [۳۷].

باید توجه داشت در حالتی که اندازه بانک واژگان افزایش پیدا می‌کند ایجاد یک مدل برای هر کلمه و آموزش آن تقریبا غیر ممکن است در این حالت از مدل مربوط به اجزاء تشکیل دهنده کلمه‌ها مثل واج‌ها[۷۱] یا هجاها[۷۲] استفاده می‌شود. مشکلی که در این حالت در مرحله تشخیص به وجود می‌آید این است که باید عمل جستجو در بین تمام رشته‌های ممکن از این اجزاء انجام شود که کاری بیش از حد وقت گیر است. در این حالت از الگوریتم‌های پیچیده‌ایی برای پایین آوردن تعداد رشته‌های مورد جستجو استفاده می‌شود. مهمترین معیاری که در این الگوریتم‌ها از آن استفاده می‌شود محدودیت‌های زبان شناختی است که بوسیله آنها ترکیبهای بی‌معنی یا از نظر دستوری غلط را حذف می‌کنند [۳۷].

۲-۲-۳ وابسته یا مستقل بودن از گوینده

سیستم‌های بازشناخت گفتار وابسته به گوینده معمولا برای استفاده توسط یک یا حداکثر چند کاربر ساخته می‌شوند. در این سیستم‌ها از صدای همان کاربری که قرار است با سیستم کار کند برای آموزش آن استفاده می‌شود. بر خلاف سیستم وابسته به گوینده در سیستم‌های مستقل از گوینده بعد از انجام عمل آموزش، سیستم باید قادر باشد عمل تشخیص را با صدای هر گوینده‌ایی انجام دهد حتی اگرصدای فرد مورد نظر جزء صدای افرادی که برای آموزش سیستم استفاده شده است نباشد.

---

[71] Phoneme
[72] Syllable

به طور کلی بیشتر الگوریتم‌های بازشناخت گفتار را می‌توان هم در حالت وابسته به گوینده و هم در حالت مستقل از گوینده استفاده کرد و اینکه در نهایت چه سیستمی داشته باشیم بستگی به نوع آموزش سیستم دارد.

معمولا سیستم‌هایی که برای استفاده عموم ساخته می‌شوند باید به صورت مستقل از گوینده باشند مثل سیستم‌های تلفنی که باید توانایی تشخیص صدای افراد مختلفی را داشته باشند. سیستم‌های وابسته به گوینده هم کاربرد زیادی دارند در این سیستم‌ها به دلیل اینکه از صدای کاربر مورد نظر برای آموزش سیستم استفاده شده دارای دقت خیلی خوبی هستند. از نمونه این سیستم‌ها می‌توان به سیستم‌هایی اشاره کرد که برای افراد دارای ناتوانی‌های جسمی ساخته می‌شوند.

۲-۲-۴ محدودیت‌های زبان شناختی

یکی از مشکلاتی که در ساخت سیستم‌های بازشناخت گفتار پیشرفته با آن روبرو هستند چگونگی به کارگیری محدودیت‌های زبان شناختی است. محدودیت‌های زبان شناختی بیان می‌کنند که واحدهای ابتدایی گفتار (آواها، واجها، هجاها و کلمه‌ها) چگونه به هم متصل می‌شوند. به عبارت دیگر محدودیت‌های زبان شناختی مشخص می‌کنند که هر یک از این واحدهای ابتدایی با چه ترتیبی، با چه بافتی و با چه معنی مورد نظری باید به هم متصل شوند. اعمال محدودیت‌های زبان شناختی مسئله‌ایی نیست که بتوان آن را با یک برنامه‌نویسی ساده حل کرد. به عبارت دیگر هر چه محدودیت‌های دستوری را در سیستم بازشناخت زیاد کنیم از آن طرف باعث کاهش آزادی‌های کاربر در بیان بعضی از جملات می‌شویم. بنابراین چالش موجود در مدل کردن زبان این است که باید به نحوی بین محدودیت‌های اعمالی در مرحله تشخیص و آزادی‌های کاربر یک تعادل برقرار شود. این

تعادل باید طوری باشد که محدودیت‌های موجود در سیستم بازشناخت ماکزیمم شود در حالی که احتمال در نظر گرفته نشدن آزادی بیان گوینده مینیمم شود [۳۷].

۲-۲-۵ ابهامات و اشتباهات آکوستیکی

کلمه‌های زیادی وجود دارند که از نظر تلفظ شباهت زیادی با هم دارند وجود این کلمات در بانک واژگان کار را برای سیستم بازشناخت سخت می‌کند و باعث به وجود آمدن ابهامات در مرحله تشخیص می‌شود. به عنوان مثال کلمه‌های "know" و "no" یا کلمه‌های "tow"، "to" و "too" شباهت زیادی با هم دارند. به عبارت دیگر در مرحله آکوستیکی این کلمه‌ها تقریبا غیر قابل تمییز هستند. علاوه بر این کلمه‌ها، بعضی از کلمه‌ها هستند که به خاطر شباهت جزئی که با هم دارند امکان دارد توسط سیستم بازشناخت با هم اشتباه گرفته شوند. به عنوان مثال کلمه‌های "nine" و "one" به خاطر شباهتی که در آخر کلمه با هم دارند می‌توانند با یکدیگر اشتباه گرفته شوند [۳۷].

۲-۲-۶ نویز محیطی

بدون شک سیستم‌های بازشناخت گفتار باید در نهایت در محیط‌هایی استفاده شوند که امکان وجود انواع و اقسام صداهای مزاحم و نویزها وجود دارد. مخصوصا اینکه بعضی از این نویزها ممکن است به سیگنال صوت شخص اصلی شباهت داشته باشند و باعث پایین آمدن دقت سیستم شوند.

یکی از مهمترین چالش‌های موجود در ساخت سیستم‌های بازشناخت گفتار ساخت سیستمی قوی است که بتواند با وجود نویز دقت قابل قبولی داشته باشد. نویز تقریبا در همه جا وجود دارد. نویز ممکن است ناشی از صداهای مربوط به اشخاص دیگر باشد یا اینکه مربوط به صدای تجهیزات موجود

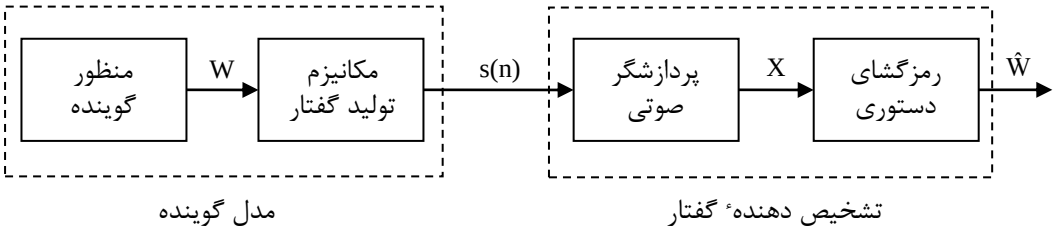
در اتاق مثل دستگاه تهویه، لامپ مهتابی و غیره باشد. بعضی از نویزها هم مختص مکانهای به خصوصی هستند مثل نویز کارخانه و نویز موجود در کابین خلبان هواپیما. همچنین نویز ممکن است به خاطر کانالی باشد که صوت از آن عبور می‌کند تا به سیستم بازشناخت گفتار برسد مثل نویز میکرفن یا نویز موجود در خط تلفن، برای زمانیکه صدا قرار است از طریق خط تلفن عبور داده شود. از نمونه‌های دیگر نویز، نویزهایی هستند که توسط خود شخص ایجاد می‌شوند مثل صدای نفس کشیدن، صدای مربوط به لب‌ها، سرفه، گرفتگی صدا به خاطر سرماخوردگی و غیره [۳۷].

۱-۶ ۲-۳ ساختار سیستم‌های بازشناخت گفتار

همانطور که در قسمت قبل به آن اشاره شد هدف از یک سیستم بازشناخت گفتار اتوماتیک (ASR) این است که مستقل از شخص صحبت کننده و محیط، با دقت و راندمان بالا، سیگنال صحبت را به یک پیام متنی که رونوشت کلمات گفته شده است تبدیل کند. یک مدل ابتدایی از فرآیند تولید گفتار، مربوط به شخصی که قصد صحبت کردن دارد، و تشخیص دهندهٔ گفتار در شکل ۲-۱ نشان داده شده است. در قسمت اول این شکل در ابتدا شخص گوینده تصمیم می‌گیرد که چه چیزی بگوید سپس منظور خود را به وسیله رشته W، که رشته‌ایی از کلمه‌ها یا آواها است، بیان می‌کند. مکانیزم تولید گفتار این رشته W را به سیگنال صوتی S(n)، که شامل تمام کلمه‌های موجود در رشته به اضافه مکث‌های کوتاه اتفاق افتاده و صداهای اضافی است، تبدیل می‌کند. همانطور که در قسمت دوم شکل ۲-۱ نشان داده شده است یک تشخیص دهنده گفتار سعی دارد طوری سیگنال گفتار S(n) را رمزگشایی کند تا بهترین تخمین از رشتهٔ W یعنی $\hat{W}$ بدست آید [۳۸].

همانطور که در شکل ۲-۱ نشان داده شده است اولین مرحله از فرآیند بازشناخت اعمال پردازشگر صوتی است. در این مرحله سیگنال S(n) به رشته‌ایی از بردارهای ویژگی، X، تبدیل می‌شود

که هر ۱۰ میلی ثانیه (یا بیشتر) در طول سیگنال گفتار محاسبه شده است. مرحله دوم از فرآیند بازشناخت استفاده از یک رمزگشای دستوری[۷۳] است تا با توجه به کاربرد مورد نظر تمام جمله‌هایی را که از نظر دستور زبان درست هستند تولید کند و به هر جمله یک امتیاز بدهد (در حقیقت احتمال پسین[۷۴] مربوط به رشته کلمه به شرط داشتن بردار ویژگی). جمله تشخیص داده شده، $\hat{W}$، جمله‌ایی خواهد بود که دارای بیشترین امتیاز است. فرآیند بازشناختی که در بالا به آن اشاره شد با نام تصمیم‌گیری با احتمال پسین ماکزیمم (MAP[۷۵]) خوانده می‌شود که در ابتدا توسط بیز پیشنهاد شده است.



شکل ۲- ۱ مدل کلی از فرآیند تولید و بازشناخت گفتار.

از نظر ریاضی در بازشناخت گفتار هدف پیدا کردن رشته $\hat{W}$ است به طوری که احتمال پسین مربوط به آن رشته به شرط داشتن بردار ویژگی X، ماکزیمم شود. یعنی:

$$\hat{W} = arg_W \max P(W|X) \qquad (۱-۲)$$

با استفاده از قانون بیز داریم:

$$\hat{W} = arg_W \max \frac{P(X|W)P(W)}{P(X)} \qquad (۲-۲)$$

رابطه ۲-۲ نشان می‌دهد که محاسبه احتمال پسین به دو عبارت تجزیه شده است یکی احتمال پیشین مربوط به رشته کلمه W که با P(W) نشان داده می‌شود و دیگری احتمال این را نشان

---

73 Syntactic Decoder
74 Posteriori Probability
75 Maximum A Posteriori Probability

می‌دهد که رشته W بردار X را ایجاد کرده باشد و با P(X|W) نشان داده می‌شود (عبارت مخرج به خاطر اینکه مستقل از رشته W است نادیده گرفته می‌شود). هر یک از احتمالهای P(W) و P(X|W) به ترتیب مربوط به مدل زبان و مدل صوتی می‌شوند و باید بوسیله داده‌های آموزشی در مرحله آموزش به دست آیند تا بتوان از آنها در مرحله بازشناخت استفاده کرد. با توجه به بی اثر بودن مخرج، P(X)، فرآیند بازشناخت به صورت زیر تبدیل می‌شود:

$$\hat{W} = arg_W \max P(X|W)P(W) \qquad (۳-۲)$$

در صورتی که سیگنال گفتار دارای N فریم باشد رشته بردارهای ویژگی (مشاهدات صوتی) به صورت زیر نوشته می‌شود:

$$X = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \qquad (۴-۲)$$

بردار مربوط به هر فریم، $\mathbf{x}_n$، $n = 1,2, \dots, N$ یک بردار ویژگی صوتی است که خواص طیفی و لحظه‌ایی سیگنال گفتار را، در زمان مورد نظر، مشخص می‌کند و به صورت زیر نشان داده می‌شود:

$$\mathbf{x}_n = x_{n1}, x_{n2}, \dots, x_{nD} \qquad (۵-۲)$$

که در این رابطه D تعداد ویژگی‌های صوتی در هر فریم است.

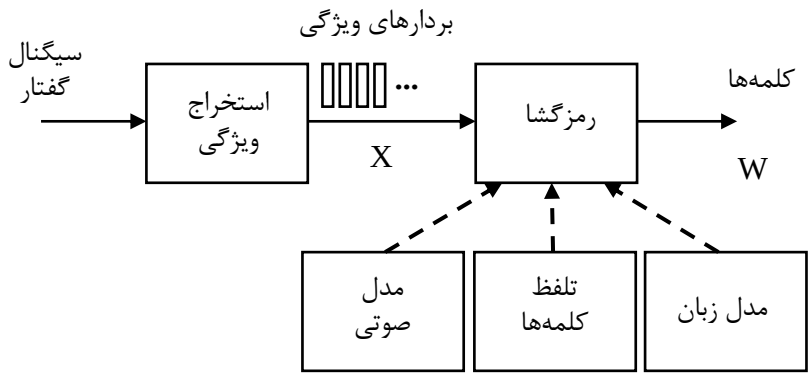
به طور مشابه رشته کلمه‌های رمزگشایی شده نیز به صورت زیر نوشته می‌شود:

$$\hat{W} = w_1 w_2 \dots w_M \qquad (۶-۲)$$

که در این رابطه M تعداد کلمه‌های موجود در رشته رمزگشایی شده است.

به طور کلی اکثر سیستم‌های بازشناخت گفتار امروزی از سه قسمت اصلی تشکیل شده‌اند که عبارتند از: استخراج ویژگی، مدل صوتی و مدل زبان. همانطور که در شکل ۲-۲ نشان داده شده است در ابتدا سیگنال گفتار از طریق میکروفن وارد بخش استخراج ویژگی می‌شود در این بخش پس از

انجام پردازش‌های اولیه بر روی سیگنال گفتار رشته‌ایی از بردارهای ویژگی با اندازه ثابت بدست می‌آید. در مرحله بعد رمزگشا سعی در پیدا کردن رشته‌ایی از کلمه‌ها می‌کند که با بیشترین احتمال رشته X را تولید کرده‌اند. به عبارت دیگر رمزگشا با توجه به معادله ۲-۳ از اطلاعات مدل صوتی و مدل زبان استفاده می‌کند تا بهترین رشته کلمه‌ها را پیدا کند. رمزگشا برای انجام عمل تشخیص احتیاج به ساختار آوایی تمام کلمه‌ها نیز دارد. اطلاعات مربوط به ساختار آوایی کلمه‌ها توسط فایل دیکشنری که حاوی تلفظ تمام کلمه‌های مورد استفاده است، در اختیار رمزگشا قرار می‌گیرد. در ادامه هر یک از قسمت‌های استخراج ویژگی، مدل زبان و مدل صوتی به طور جداگانه بررسی می‌شود.



شکل ۲- ۲ ساختار کلی یک سیستم بازشناخت گفتار اتوماتیک [۶۷].

۲-۳-۱ استخراج ویژگی

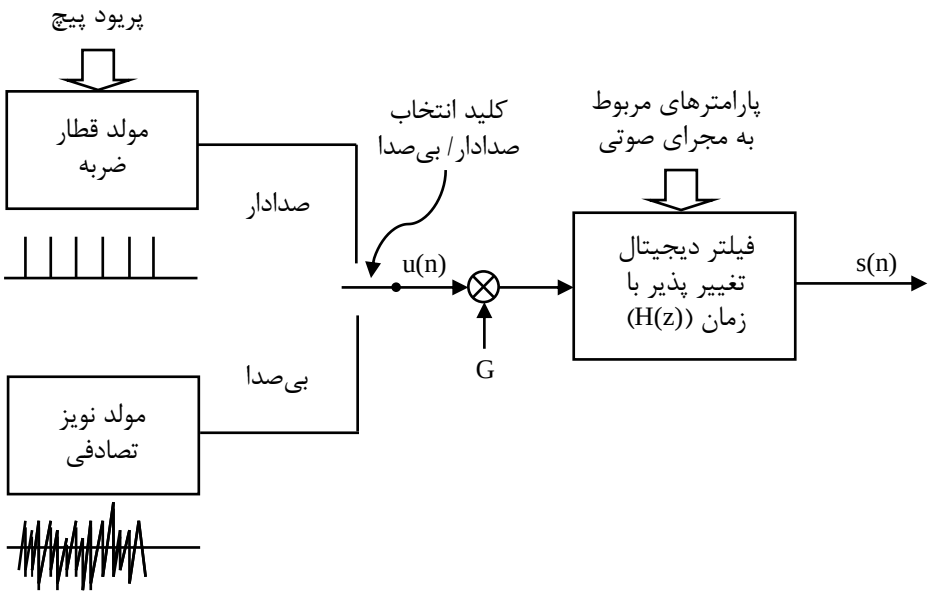
به فرآیند ایجاد بردارهای ویژگی از روی سیگنال گفتار استخراج ویژگی گفته می‌شود. هدف از مرحلهٔ استخراج ویژگی در بازشناخت گفتار تولید یک نمایش فشرده از شکل موج گفتار می‌باشد که دارای بیشترین اطلاعات مربوط به گفتار می‌باشد. این نمایش فشرده باید طوری باشد که بهترین جداسازی را بین صداهای مختلف ایجاد کند. این مرحله را می‌توان مهمترین مرحلهٔ موجود در سیستم‌های بازشناخت گفتار در نظر گرفت که تاثیر بسیار زیادی بر روی قسمت‌های بعدی دارد و دقت سیستم تا حد زیادی وابسته به آن است. اگر ویژگی‌های استخراج شده بتوانند نمایش خوبی از سیگنال داشته

باشند و اطلاعات موجود در سیگنال گفتار را که برای جداسازی بین صداهای مختلف حیاتی هستند حذف نکند، دقت بازشناخت تا حد زیادی بالا خواهد رفت. ویژگی‌های استخراج شده باید تا آنجا که ممکن است نسبت به کانال مخابراتی، نویز و تغییرات ناشی از گوینده‌های مختلف مقاوم باشند.

تاکنون روش‌های مختلفی جهت استخراج ویژگی ارائه شده است که اکثر آنها بر اساس پردازش در حوزهٔ کپسترال بدست آمده‌اند. در ادامه در ابتدا پردازش کپسترال توضیح داده می‌شود. سپس به طور خلاصه متداولترین روش‌های استخراج ویژگی مثل MFCC و PLP بررسی می‌شود.

۲-۳-۱-۱ پردازش کپسترال[۷۶]

بلوک دیاگرام مربوط به مدل تولید گفتار در شکل ۲-۳ نشان داده شده است. در این مدل که سعی در به نمایش گذاشتن فرآیند تولید گفتار بر اساس خاصیت سیگنال خروجی دارد از سه جزء اصلی تشکیل شده است که عبارتند از: دو منبع تحریک متفاوت که به وسیله تارهای صوتی تولید می‌شوند و حفره‌های مربوط به عبور صدا که کانال صوتی خوانده می‌شود.



شکل ۲- ۳ بلوک دیاگرام مربوط به مدل ساده تولید گفتار [۴۲].

[76] Cepstral Processing

فرآیند تولید گفتار بوسیله مدل نشان داده شده در شکل ۲-۳ به این صورت است که با توجه به اینکه آوای مورد نظر صدادار یا بی‌صدا باشد کلید موجود در قسمت اول به سمت مولد پالس یا مولد نویز تصادفی جابجا می‌شود. فرکانس اصلی صدا یا همان پریود پیچ[۷۷] با توجه به نوع آوای مورد نظر تغییر می‌کند. تابع انتقال مربوط به مجرای صوتی به وسیله یک فیلتر تغییرپذیر با زمان خطی، H(z)، مدل می‌شود. فرکانسهای تشدید مربوط به مجرای صوتی یا همان فرمنتها[۷۸] تابعی از شکل مجرای صوتی هستند و به آهستگی با زمان تغییر می‌کنند. به منظور مدل کردن این فرکانسها پارامترهای مربوط به فیلتر مجرای صوتی نیز به آهستگی با زمان تغییر می‌کنند. سیگنال خروجی، s(n)، به وسیله کانولوشن منبع، Gu(n)، با پاسخ ضربه فیلتر، h(n)، مدل می‌شود [۳۹].

برای استخراج ویژگی معمولا از پاسخ فرکانسی مربوط به مجرای صوتی استفاده می‌شود و اطلاعات مربوط به سیگنال تحریک که برای اصوات صدادار متناوب و برای اصوات بی‌صدا نویز مانند است نادیده گرفته می‌شوند. دلیل این امر این است که پاسخ فرکانسی مربوط به مجرای صوتی بهترین جداسازی بین صداهای مختلف مربوط به گفتار را ایجاد می‌کند [۳۹].

برای جداسازی سیگنال تحریک و پاسخ فرکانسی مجرای صوتی از پردازش کپسترال استفاده می‌شود. پردازش کپسترال یکی از انواع تبدیل‌های هوموморفیک[۷۹] است که قادر است اطلاعات مربوط به منبع را از فیلتر جدا کند. با توجه به مدل نشان داده شده در شکل ۲- ۳ برای تولید گفتار داریم:

$$s(n) = Gu(n) \otimes h(n) \qquad (۲-۷)$$

که u(n) سیگنال تحریک، h(n) پاسخ فرکانسی مجرای صوتی، s(n) سیگنال گفتار و G بهره می‌باشد. با اعمال تبدیل فوریه به رابطه بالا داریم:

$$S(f) = GU(f) \cdot H(f) \qquad (۲-۸)$$

---

[77] Pitch Period
[78] Formant
[79] Homomorphic Transformation

اگر از هر دو طرف لگاریتم مختلط گرفته شود داریم:

$$\log\big(S(f)\big) = \log\big(GU(f) \cdot H(f)\big)$$

$$= \log\big(GU(f)\big) + \log(H(f)) \qquad (۹ -۲)$$

بنابراین با توجه به رابطه بالا، در حوزه لگاریتم، اطلاعات مربوط به سیگنال تحریک و مجرای صوتی بر روی هم افتاده‌اند و با استفاده از تکنیک‌های موجود در پردازش سیگنال قابل جدا شدن هستند [۴۰].

با گرفتن تبدیل فوریه معکوس از رابطه ۲- ۹ کپستروم مختلط مربوط به سیگنال s(n) بدست می‌آید. در حوزه کپسترال ضرایب قرار گرفته در مرتبه پایین‌تر مربوط به مجرای صوتی می‌باشد در حالی که ضرایب مربوط به تحریک در مرتبه‌های بالاتری قرار گرفته‌اند. بنابراین با حذف ضرایب مرتبه بالا و نگه داشتن ۱۲ یا ۱۳ ضریب اول اطلاعات مربوط به مجرای صوتی از سیگنال تحریک جدا می‌شود. به این عمل در اصطلاح لیفترینگ[۸۰] گفته می‌شود.

۲-۳-۱-۲ ضرایب کپسترال پیشگویی خطی (LPCC[۸۱])

یکی از قویترین تکنیک‌های آنالیز صوت آنالیز پیشگویی خطی است [۴۱]. این روش به یکی از روشهای پرکاربرد در محاسبه پارامترهای مربوط به سیگنال صوت مثل دوره تناوب اصلی و فرکانس‌های فرمنت تبدیل شده است و کاربردهای زیادی در زمینه‌های مختلف پردازش صوت پیدا کرده است. آنالیز پیشگویی خطی به خاطر توانایی آن در محاسبه نسبتا دقیق پارامترهای مربوط به صوت و همچنین سرعت بالای آن اهمیت زیادی در پردازش صوت دارد. ایدهٔ ابتدایی آنالیز پیشگویی

---

[80] Liftering

[81] Linear Prediction Cepstral Coefficients

خطی این است که یک نمونه از سیگنال گفتار را می‌توان به وسیله ترکیبی از نمونه‌های قبل تخمین زد [۴۲].

فیلتر نشان داده شده در شکل ۲-۳ را می‌توان با یک فیلتر تمام قطب با درجه p به صورت زیر به خوبی تخمین زد.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1-\sum_{k=1}^{p} a_k z^{-k}} . \qquad (۲- ۱۰)$$

با توجه به سیستم نشان داده شده در شکل ۲-۳ سیگنال گفتار s(n) با استفاده از معادله تفاضلی ساده زیر به سیگنال تحریک u(n) مربوط می‌شود.

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (۲- ۱۱)$$

یک پیشگوی خطی با ضرایب پیشگویی α_k به صورت سیستمی با خروجی زیر تعریف می‌شود:

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) \qquad (۲- ۱۲)$$

در این رابطه هر نمونه از سیگنال به وسیلهٔ P نمونهٔ قبل تخمین زده می‌شود. در این حالت خطای پیشگویی، e(n)، ناشی از این تخمین به صورت زیر تعریف می‌شود:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \qquad (۲- ۱۳)$$

با توجه به رابطه بالا دیده می‌شود که خطای پیشگویی خروجی سیستمی با تابع انتقال زیر می‌باشد:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k \, z^{-k} \qquad (۲- ۱۴)$$

با مقایسه رابطه‌های ۲- ۱۱ و ۲- ۱۳ دیده می‌شود که اگر سیگنال گفتار دقیقا مدل مربوط به رابطه ۲- ۱۱ را برآورده کند، و اگر $\alpha_k = a_k$ باشد، آنگاه $e(n) = Gu(n)$ است. بنابراین فیلتر خطای پیشگویی، A(z)، فیلتر معکوس سیستم داده شده در رابطه ۲- ۱۰ خواهد بود، یعنی،

$$H(z) = \frac{G}{A(z)} \qquad (۱۵ -۲)$$

مشکل اساسی در آنالیز پیشگویی خطی تعیین مجموعه ضرایب پیشگویی، $\{\alpha_k\}$ به طور مستقیم از روی سیگنال گفتار است به طوری که یک تخمین خوب از خواص طیفی سیگنال گفتار از طریق استفاده از رابطه ۲- ۱۵ بدست آید. به خاطر خاصیت تغییر پذیری با زمان سیگنال گفتار ضرایب پیشگویی باید از بخش‌های کوتاه از سیگنال گفتار تخمین زده شوند. رویکرد اولیه در حل این مسئله پیدا کردن مجموعه‌ایی از ضرایب پیشگویی است که میانگین مربع خطای پیشگویی را در طول بخش کوچکی از شکل موج گفتار مینیمم کند[۴۳]. پارامترهای بدست آمده از این روش به عنوان پارامترهای تابع سیستم، $H(z)$، در مدل تولید گفتار، فرض می‌شوند [۴۲].

بعد از بدست آمدن ضرایب پیشگویی خطی ضرایب کپسترال پیشگویی خطی به کمک رابطه ۲- ۱۶ محاسبه می‌شوند. همانطور که از رابطه ۲- ۱۶ دیده می‌شود با تعداد محدودی ضریب پیشگویی خطی ضرایب LPCC بدست آمده نامحدود می‌باشد. به طور تجربی نشان داده شده است که معمولا ۱۲ تا ۲۰ ضریب اول برای بدست آوردن نتایج خوب در بازشناخت گفتار کافی است [۴۴].

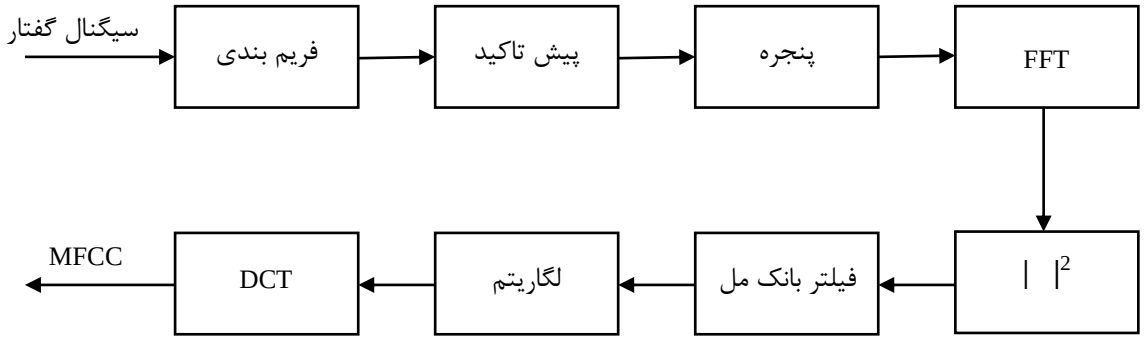
$$c(n) = \begin{cases} 0 & n < 0 \\ \ln G & n = 0 \\ a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & 0 < n \leq p \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & n > p \end{cases} \qquad (۱۶ -۲)$$

۲-۳-۱-۳ ضرایب کپسترال فرکانس مل (MFCC[۸۲])

ضرایب کپسترال فرکانس مل در ابتدا در سال ۱۹۸۰ توسط Davis و Mermelstein به عنوان ویژگی برای استفاده در بازشناخت گفتار پیشنهاد شد [۴۵] و امروزه به عنوان پرکاربردترین روش استخراج

[82] Mel Frequency Cepstral Coefficients

ویژگی در سیستم‌های بازشناخت گفتار مورد استفاده قرار می‌گیرد. روش MFCC، برای مدل کردن مکانیزم شنوایی موجود در انسان، مقیاس مل را به طیف توان سیگنال گفتار اعمال می‌کند [۴۶]. بلوک دیاگرام مربوط به محاسبه ضرایب کپسترال فرکانس مل در شکل ۲- ۴ نشان داده شده است.



شکل ۲- ۴ بلوک دیاگرام مربوط به استخراج ضرایب کپسترال فرکانس مل.

در ادامه به طور مختصر مراحل مختلف این روش استخراج ویژگی مورد بررسی قرار گرفته است.

فریم‌بندی[۸۳]:

سیگنال گفتار یک سیگنال غیر ایستا است و مشخصه‌های آن در طول زمان تغییر می‌کند. برای بدست آوردن ویژگی‌های معتبر از سیگنال گفتار باید آن را به فریم‌های ۲۰ تا ۳۰ میلی ثانیه تقسیم کرد به طوری که هر فریم با هم ۱۰ تا ۱۵ میلی ثانیه هم پوشانی داشته باشد.

پیش تاکید[۸۴]:

برای کاهش اثر حنجره و لبها در مدل مربوط به مجرای صوتی از یک فیلتر دیجیتال مرتبه اول با پارامتر پیش تاکید α (رابطه ۲- ۱۷) استفاده می‌شود. مقدار α معمولا بین ۰/۹ و ۱ انتخاب می‌شود.

$$H(z) = 1 - \alpha z^{-1} \qquad (۱۷-۲)$$

---

83 Frame Blocking
84 Pre-emphasis

پنجره گذاری[۸۵]:

برای از بین بردن ناپیوستگی موجود در مرز بین فریم‌ها، هر فریم در یک تابع پنجره ضرب می‌شود. متداولترین پنجره‌ایی که برای این منظور استفاده می‌شود پنجره همینگ است که رابطه آن به صورت زیر می‌باشد.

$$W(n) = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \qquad 0 \le n \le N-1 \qquad (۱۸-۲)$$

که در این رابطه N طول هر فریم می‌باشد.

آنالیز طیفی:

در این مرحله از تبدیل فوریهٔ گسستهٔ زمان کوتاه برای بردن هر فریم گفتار به حوزه فرکانس استفاده می‌شود. در این مرحله اطلاعات مربوط به فاز به خاطر اینکه از نظر شنوایی حاوی اطلاعات مفیدی نیستند، حذف می‌شوند.

فیلتر بانک مقیاس مل:

فیلتر بانک، در ساده‌ترین شکل خود، مجموعه‌ایی از فیلترهای میانگذر با فرکانس‌های متفاوت است که محدودهٔ مورد نظر از طیف سیگنال را می‌پوشاند. با روشهای متفاوتی می‌توان فرکانس مرکزی این فیلترها را انتخاب کرد. معمولا این فرکانس‌ها طوری انتخاب می‌شوند که حساسیت موجود در گوش انسان، نسبت به حوزه‌های مختلف فرکانسی، مدل شود. یکی از تخمین‌هایی که برای این منظور استفاده می‌شود مقیاس مل است [۴۷].

مل واحد ارزیابی صدای درک شده است. در آزمایشی که دانشمندها برای محاسبهٔ مقیاس مل انجام دادند، در ابتدا فرکانس ۱۰۰۰ هرتز بطور اختیاری برابر ۱۰۰۰ مل قرار داده شده است. سپس از

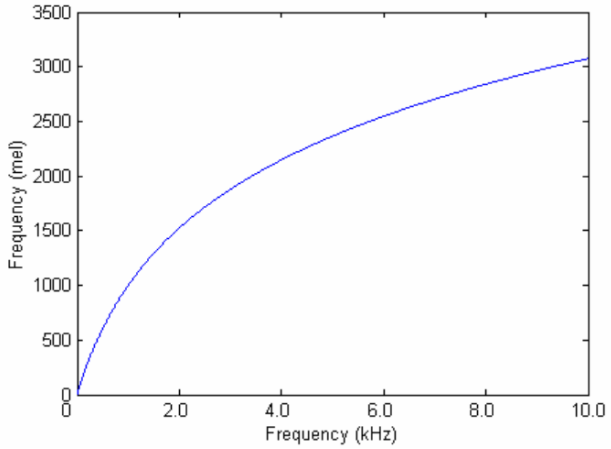
---

[85] Windowing

شنوندگان خواسته شده که فرکانس فیزیکی را تا آنجا افزایش یا کاهش دهند که صدای شنیده شده، ۰/۱، ۰/۵، ۲ و ۱۰ برابر صدای مرجع باشد. سپس فرکانس مربوط به این صداها به ترتیب برابر با ۱۰۰، ۵۰۰، ۲۰۰۰ و ۱۰۰۰۰ مل قرار داده شده‌اند. به این ترتیب نگاشتی بین مقیاس فرکانس واقعی (Hz) و مقیاس فرکانس درک شده (Mel) بدست آمده است. این نگاشت برای فرکانس‌های کمتر از یک کیلو هرتز خطی و برای فرکانس‌های بیشتر از آن لگاریتمی می‌باشد [۴۸]. فرکانس مل به صورت زیر تعریف می‌شود.

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (۲-۱۹)$$

که در این رابطه f فرکانس واقعی و بر حسب هرتز می‌باشد. نمودار شکل ۲- ۵ مقیاس مل را بر حسب هرتز نشان می‌دهد.



شکل ۲- ۵ مقیاس مل [۴۶].

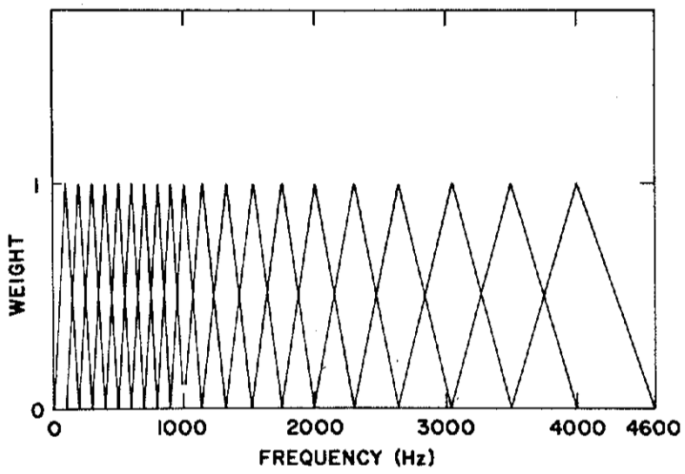
سیستم شنوایی انسان نمی‌تواند بین فرکانس‌هایی که به هم نزدیک هستند تمییز قائل شود. هر چه فرکانس بیشتر می‌شود این باندهای بحرانی[۸۶] نیز بزرگتر می‌شوند. به عبارت دیگر در هر باند بحرانی شنونده نمی‌تواند تفاوتی بین فرکانس مرکزی و فرکانس‌های اطراف آن تشخیص دهد. در

[86] Critical Band

صورتی که از مقیاس مل استفاده شده باشد اندازهٔ باندهای بحرانی با رابطه زیر قابل تخمین زدن است.

$$\mathrm{BW_{critical}} = 25 + 75\left[1 + 1.4\left(\frac{\mathrm{f_{mel}}}{1000}\right)^2\right]^{0.69}. \qquad (۲۰ -۲)$$

یک نمونه از فیلترهایی که برای بدست آوردن ضرایب کپسترال فرکانس مل استفاده می‌شوند در شکل ۲- ۶ نشان داده شده است.



شکل ۲- ۶ فیلتر بانک مقیاس مل [۴۵].

تعداد فیلترهای موجود در فیلتر بانک مقیاس مل برای کاربردهای مختلف فرق می‌کند ولی معمولا تعداد آنها را بین ۱۹ تا ۲۴ انتخاب می‌کنند.

تبدیل کسینوسی گسسته (DCT[۸۷]):

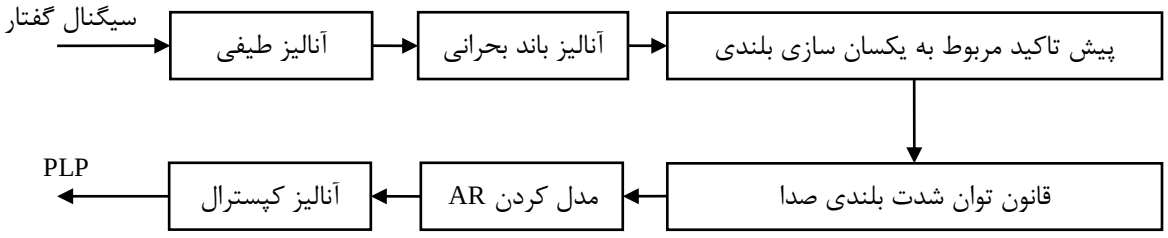
بعداز اعمال فیلتر بانک به طیف سیگنال، لگاریتم انرژی‌های فیلتر بانک محاسبه می‌شود. در نهایت ضرایب کپسترال مقیاس مل با اعمال DCT به لگاریتم انرژی‌های فیلتر بانک بدست می‌آید. ضرایب ۱ تا ۱۲ چون حاوی اطلاعات مربوط به مجرای صوتی هستند نگه داشته می‌شوند و بقیه ضرایب حذف

[87] Discrete Cosine Transform

می‌شوند [۳۹]. معمولا لگاریتم انرژی هر فریم نیز به عنوان ضریب صفرم به ضرایب MFCC مربوط به هر فریم اضافه می‌شود.

۲-۳-۱-۴ پیشگویی خطی مبتنی بر درک انسان (PLP[۸۸])

روش پیشگویی خطی برای بدست آوردن ضرایب کپسترال در قسمت‌های قبل بررسی شد یکی از معایب LPCC این است که مکانیزم شنوایی انسان را در محاسبهٔ ویژگی‌ها منظور نمی‌کند. به عبارت دیگر پیشگویی خطی در تمام فرکانس‌ها سیگنال گفتار را به یک صورت تخمین می‌زند و این مطابق با سیستم شنوایی انسان نیست. برای هماهنگ کردن روش پیشگویی خطی با سیستم شنوایی انسان در ابتدا Hermansky در سال ۱۹۸۹ روش پیشگویی خطی مبتنی بر درک انسان را پیشنهاد کرد [۴۹]. Hermansky برای هماهنگی ویژگی‌ها با سیستم شنوایی انسان آنالیز طیفی را طوری انجام داد که بعضی از نواحی حساستر از بقیهٔ قسمت‌ها شوند. برای این منظور از مقیاسی شبیه مقیاس مل با نام مقیاس بارک[۸۹] استفاده کرد. بلوک دیاگرام نشان داده شده در شکل ۲-۷ مراحل مختلف روش PLP را نشان می‌دهد.



شکل ۲- ۷ بلوک دیاگرام مربوط به بدست آوردن ضرایب PLP [۴۹].

---

88 Perceptual Linear Prediction
89 Bark scale

آنالیز طیفی:

بعد از فریم بندی و اعمال پنجره، طیف توان زمان کوتاه مربوط به هر فریم با استفاده از تبدیل فوریه سریع محاسبه می‌شود.

آنالیز باند بحرانی:

بعد از محاسبهٔ طیف توان محور فرکانس با توجه به رابطهٔ زیر به فرکانس بارک نگاشت داده می‌شود.

$$\Omega(\omega) = 6\ln\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right\}, \qquad (۲۱-۲)$$

که در این رابطه ω فرکانس زاویه‌ایی بر حسب rad/s است. سپس طیف توان نگاشت داده شده با نمودار باند بحرانی پوشاننده[۹۰]، $\Psi(\Omega)$، کانوالو می‌شود. این مرحله تقریبا شبیه به آنالیز کپسترال فرکانس مل است که در قسمت قبل بیان شد. رابطه مربوط به نمودار باند بحرانی به صورت زیر داده شده است [۴۹].

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega \le -0.5, \\ 1 & -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & 0.5 < \Omega < 2.5, \\ 0 & \Omega > 2.5. \end{cases} \qquad (۲۲-۲)$$

پیش تاکید مربوط به یکسان سازی بلندی صدا[۹۱]:

در این مرحله نمودار مربوط به یکسان سازی بلندی صدا به رشته بدست آمده از مرحله قبل اعمال می‌شود. تابع استفاده شده در این مرحله تخمینی است از حساسیت غیر یکنواخت شنوایی انسان در فرکانس‌های مختلف و حساسیت شنوایی انسان را در سطح ۴۰ دسی بل شبیه سازی می‌کند [۴۹].

---

90 Critical Band Masking Curve
91 Equal-loudness Preemphasis

قانون توان شدت بلندی صدا[92]:

آخرین مرحله قبل از مدل کردن AR اعمال ریشه سوم به طیف است تا قانون توان مربوط به سیستم شنوایی انسان تخمین زده شود. به عبارت دیگر در گوش انسان میزان احساس بلندی صدا با ریشه سوم انرژی آن متناسب است.

مدل کردن AR[93]:

در این مرحله برای بدست آوردن ضرایب کپسترال حاصل از پیشگویی خطی مبتنی بر درک انسان در ابتدا تبدیل فوریه معکوس اعمال می‌شود تا رشته اتوکرولیشن بدست آید. سپس به وسیله معادله‌های Yule-Walker (یکی از روشهای موجود در محاسبه ضرایب پیشگویی خطی) فیلتر تمام قطب مدل می‌شود و ضرایب AR (همان ضرایب LP) بدست می‌آید.

آنالیز کپسترال:

بعد از بدست آمدن ضرایب AR، به کمک رابطه ۲- ۱۶ ضرایب کپسترال مانند حالت LPCC بدست می‌آید.

آنالیز باند بحرانی انجام شده در روش PLP شباهت زیادی به آنالیز بانک فیلتر مل استفاده شده در روش MFCC دارد. بعضی از آزمایشهای انجام شده نشان داده‌اند که روش MFCC نسبت به PLP دارای نتایج بهتری در بازشناخت گفتار است [۵۰] ولی روش PLP نسبت به تغییرات تعداد ضرایب و تعداد فیلترهای مورد استفاده در محاسبه ضرایب دارای نتایج پایدارتری نسبت به MFCC می‌باشد [۵۱]. علاوه بر این روش PLP دارای مقاومت بیشتری نسبت به نویز نیز می‌باشد.

---

92 Intensity-loudness Power Law
93 Autoregressive Modeling

۲-۳-۱-۵ ضرایب مشتق

معمولا برای بالا بردن دقت سیستم از مشتق نسبت به زمان ضرایب بدست آمده نیز به همراه ضرایب اصلی برای بازشناخت گفتار استفاده می‌شود. در حقیقت ضرایب کپسترال اطلاعات استاتیک سیگنال گفتار را مدل می‌کنند و به حالتهای گویش و تغییرات آن حساس می‌باشند. در حالیکه مشتقات ضرایب کپسترال حاوی اطلاعات دینامیک و اطلاعات انتقال میان حالتهای مختلف گویش هستند. تلفیق ضرایب کپسترال و مشتقات آن می‌تواند ویژگی‌های بهتری از سیگنال گفتار را بیان نماید [۴۸].

رابطه ۲- ۲۳ تخمینی است از مشتق اول نسبت به زمان ضریب n ام. برای محاسبهٔ مشتق دوم رابطه ۲- ۲۳ به نتایج مربوط به مشتق اول اعمال می‌شود. معمولا هر چه درجه مشتق بیشتر می‌شود اندازهٔ پنجرهٔ آنالیز نیز کاهش پیدا می‌کند [۳۹].

$$\partial c_t(n) = \frac{\sum_{k=-K}^{K} k c_{t+k}(n)}{\sum_{k=-K}^{K} k^2} \qquad K > 0 \qquad (۲۳-۲)$$

۲-۳-۲ مدل زبان

مدل زبان احتمال یک رشته کلمه، که یک جملهٔ درست از نظر زبان شناختی ایجاد می‌کند، را نشان می‌دهد. به عبارت دیگر هدف مدل زبان فراهم کردن مکانیزمی برای تخمین احتمال بعضی از کلمه‌ها، w_k، در جمله است در صورتی که کلمه‌های قبل از آن، $W_1^{k-1} = w_1 \dots w_{k-1}$، داده شده باشد. یکی از روش‌های پرکاربرد و موثر برای این کار استفاده از مدل N-gram است به طوری که کلمهٔ w_k فقط به N-1 کلمهٔ قبل از خود بستگی داشته باشد (رابطه ۲- ۲۴) [۵۲]. مقدار N را معمولاً بین ۱ تا ۴ انتخاب می‌کنند.

$$P(W) = P(w_1, w_2, \dots, w_M) = \prod_{m=1}^{M} P(w_m | w_{m-1}, w_{m-2}, \dots, w_{m-N+1}) \qquad (۲۴-۲)$$

احتمال مربوط به N-gram به طور مستقیم از روی داده‌های متنی قابل محاسبه است و نیازی به تعریف قواعد زبان شناختی، به طوری که در دستور زبان است، نیست [۵۲]. در عمل مدل N-gram بوسیله یک شمارش‌گر تکرار قابل محاسبه است. اگر فرض کنیم یک داده متنی بزرگ برای آموزش داشته باشیم که شامل تمام کلمه‌های موجود در فرهنگ لغت باشد. با فرض اینکه N برابر با ۳ باشد احتمال مربوط به یک رشته سه کلمه‌ایی[۹۴] (احتمال اینکه کلمهٔ w_i بعد از دو کلمهٔ (w_{i-1}, w_{i-2}) قرار گیرد) به صورت زیر قابل محاسبه است:

$$P(w_i|w_{i-1}, w_{i-2}) = \frac{C(w_{i-2},w_{i-1},w_i)}{C(w_{i-2},w_{i-1})} \qquad (۲۵-۲)$$

که $C(w_{i-2}, w_{i-1}, w_i)$ تعداد تکرار سه کلمهٔ (w_{i-2}, w_{i-1}, w_i) و $C(w_{i-2}, w_{i-1})$ تعداد تکرار دو کلمهٔ[۹۵] (w_{i-2}, w_{i-1}) در متن آموزشی است [۳۸].

مشکلی که در روش بالا برای محاسبهٔ N-gram ممکن است به وجود آید این است که بعضی از رشته کلمه‌ها ممکن است در متن آموزشی موجود نباشند و احتمال مربوط به آنها به اشتباه صفر قرار داده شود. مخصوصا زمانی که تعداد لغات موجود در فرهنگ لغت خیلی زیاد باشد این مشکل بیشتر خودش را نشان می‌دهد. در این حالت رشته کلمه‌هایی که احتمال مربوط به آنها صفر قرار داده شده است در سیستم بازشناخت گفتار بی اعتبار می‌شوند [۳۸]. یکی از روش‌هایی که برای حل این مشکل پیشنهاد شده است استفاده از الگوریتم هموارساز[۹۶] [۵۳] است. بعنوان مثال برای حالت N=3 احتمال مربوطه، با استفاده از این الگوریتم، به صورت زیر بدست می‌آید.

$$\widehat{P}(w_i|w_{i-1}, w_{i-2}) = p_3 \frac{C(w_{i-2},w_{i-1},w_i)}{C(w_{i-2},w_{i-1})} + p_2 \frac{C(w_{i-1},w_i)}{C(w_{i-1})} + p_1 \frac{C(w_i)}{\sum_i C(w_i)} \qquad (۲۶-۲)$$

که در این رابطه مجموع احتمالهای هموارسازی p_1، p_2 و p_3 باید برابر با یک باشد و $\sum_i C(w_i)$ اندازهٔ متن آموزشی است.

---

94 Trigram
95 Bigram
96 Smoothing Algorithm

۲-۳-۳ مدل صوتی

نقش مدل صوتی در سیستم‌های بازشناخت گفتار گرفتن رشتهٔ تولید شده توسط مرحلهٔ استخراج ویژگی، O، و معین کردن اینکه با چه احتمالی این رشته توسط رشته کلمهٔ فرضی W تولید شده است، P(O|W) . به عبارت دیگر در این مرحله تغییرات آماری ویژگی‌های مربوط به هر یک از اجزاء صدا (آواها یا کلمه‌ها) مدل می‌شود.

یکی از خصوصیات سیگنال گفتار دینامیک بودن آن است به طوری که حتی در اجزاء کوچکی مثل آواها لحن گفتار به تدریج تغییر می‌کند. به طوری که ابتدا و انتهای آوا به ترتیب تحت تاثیر آوای قبلی و بعدی تغییر می‌کند و وسط آوا تقریبا ثابت است. بنابراین اطلاعات لحظه‌ایی که به وسیلهٔ بردارهای ویژگی از سیگنال گفتار استخراج می‌شوند نقش مهمی در فرآیند بازشناخت گفتار بازی می‌کنند [۵۴].

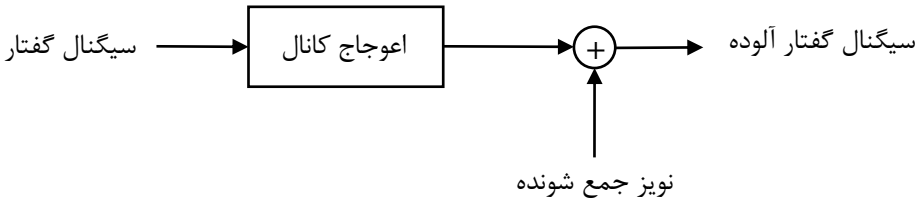
برای در نظر گرفتن این تغییرات موجود در سیگنال گفتار از روش‌های آماری استفاده می‌شود. یکی از استانداردترین روش‌های آماری که در اکثر سیستم‌های بازشناخت امروزی استفاده می‌شود مدل مخفی مارکوف (HMM) می‌باشد. برای اولین بار Baker و Jelinek در دههٔ ۱۹۷۰ از مدل مخفی مارکوف در سیستم‌های بازشناخت گفتار استفاده کردند [۳۹]. از این روش آماری برای مدل کردن تغییرات طیفی مربوط به هر یک از اجزاء صدا مثل آواها استفاده می‌شود و پارامترهای مربوط به هر یک از این مدل‌ها به وسیلهٔ داده‌های آموزشی بدست می‌آید [۳۸].

در فصل بعد مدل مخفی مارکوف به طور کاملتری مورد بررسی قرار می‌گیرد.

۱-۷ ۲-۴ بازشناخت گفتار مقاوم

معمولا دقت سیستم‌های بازشناخت گفتار در عمل به شدت کاهش پیدا می‌کند. علت اصلی این کاهش دقت وجود اختلاف بین محیطی که در آن سیستم آموزش می‌بیند و محیطی است که سیستم در آن مورد استفاده قرار می‌گیرد. دو علت اصلی که باعث به وجود آمدن این اختلاف می‌شوند عبارتند از نویز پس زمینه و اعوجاج کانال انتقال [۵۵].

نویز پس زمینه که معمولا حالت جمع شونده با سیگنال گفتار دارد می‌تواند ناشی از عوامل مختلفی مثل پنکهٔ مربوط به سیستم تهویه، به هم خوردن در، صدای مربوط به صحبت کردن دیگران و غیره باشد. در حالی که اعوجاج ناشی از کانال انتقال می‌تواند ناشی از پاسخ فرکانسی میکرفن مورد استفاده، انعکاس صدا، خط تلفن و غیره باشد [۴۴]. شکل ۲-۸ مدل مربوط به اثر نویز محیطی و کانال انتقال را بر روی سیگنال گفتار نشان می‌دهد.



شکل ۲- ۸ مدل مربوط به اثر نویز جمع شونده و اعوجاج کانال بر سیگنال گفتار.

با توجه به شکل ۲-۸ سیگنال گفتار آلوده شده به نویز به صورت زیر بدست می‌آید:

$$x(t) = s(t) * h(t) + d(t) \qquad (۲۷-۲)$$

که در این رابطه $d(t)$ نویز جمع شونده، $h(t)$ پاسخ ضربهٔ مربوط به کانال انتقال، $s(t)$ سیگنال بدون نویز و $x(t)$ سیگنال نویزی است. معمولا نویز جمع شونده در رابطهٔ بالا به صورت ناهمبسته با سیگنال گفتار در نظر گرفته می‌شود [۵۵].

یکی از هدف‌های اصلی سیستم‌های بازشناخت گفتار مقاوم، بهبود دقت سیستم بازشناخت در چنین محیط‌های نامطلوبی است. روش‌های زیادی برای این منظور ارائه شده است که به طور کلی می‌توان آنها را به سه دستهٔ زیر تقسیم کرد:

۱. استخراج ویژگی‌های مقاوم.
۲. بهبود سیگنال گفتار.
۳. مقاوم سازی مدل بازشناخت.

در ادامه به طور مختصر با هر یک از این روش‌ها آشنا می‌شویم.

۲-۴-۱ استخراج ویژگی‌های مقاوم

در این دسته از روشها ویژگی‌هایی از سیگنال گفتار استخراج می‌شوند که ذاتاً نسبت به نویز مقاوم باشند و نویز تأثیر چندانی بر آنها نداشته باشد. از جملهٔ این روش‌ها می‌توان به RASTA فیلترینگ[۹۷] [۵۶]، نرمالیزه‌سازی میانگین ضرایب کپسترال (CMN[۹۸]) [۵۷]، اتوکرولیشن یک طرفه LPC (OSALPC[۹۹]) [۵۸] و رشتهٔ اتوکرولیشن نسبی (RAS[۱۰۰]) [۵۹] اشاره کرد.

۲-۴-۲ بهبود سیگنال گفتار

روش‌های بر پایهٔ بهبود سیگنال گفتار در بازشناخت گفتار مقاوم همانطور که از نامش هم پیدا است روش‌هایی هستند که سعی دارند در ابتدا سیگنال بدون نویز را از روی سیگنال نویزی بدست آورند و

---

97 RelAtive SpecTrAl Filtering
98 Cepstral Mean Normalization
99 One-Sided Autocorrelation LPC
100 Relative Autocorrelation Sequence

سپس ویژگی‌ها را از سیگنال بدون نویز بدست آمده استخراج کنند. در این نوع روشها یک اطلاعات اولیه از سیگنال گفتار و نویزی که آن را آلوده کرده است مورد نیاز است. از جملهٔ روش‌هایی که در این دسته قرار می‌گیرند می‌توان به تفریق طیفی (SS[۱۰۱]) [۶۰] و [۶۱] اشاره کرد.

۲-۴-۳ مقاوم سازی مدل بازشناخت

در این گروه از روش‌ها سعی در بهبود مدل صوتی بکار رفته در سیستم بازشناخت گفتار است. به عنوان مثال اگر از مدل مخفی مارکوف برای مدل کردن صوتی استفاده شده باشد این نوع روشها سعی دارند که به طریقی مدل صوتی را با شرایط نویزی جدید وفق دهند. به عبارت دیگر این نوع روش‌ها سعی دارند اختلاف موجود بین مدل‌های ایجاد شده در مرحلهٔ آموزش و سیگنال گفتار نویزی را بر طرف کنند. از جملهٔ این روش‌ها می‌توان به ترکیب چند مدل موازی (PMC) [۳۴] و اندازه پروژکشن وزن‌دار شده (WPM[۱۰۲]) [۶۲] اشاره کرد.

---

101 Spectral Subtraction
102 Weighted Projection Measure

فصل سوم

مدل مخفی مارکوف و کاربرد آن در

بازشناخت گفتار

فصل سوم: مدل مخفی مارکوف و کاربرد آن در بازشناخت گفتار

۱-۸ ۳-۱ مقدمه

قبل از تبدیل شدن مدل‌های مخفی مارکوف به عنوان یکی از روش‌های آماری مطرح در سیستم‌های بازشناخت گفتار از روش پیچش زمانی پویا (DTW[۱۰۳]) برای بازشناخت کلمه‌ها استفاده می‌شد. ولی روش DTW دارای مشکلاتی بود که باعث می‌شد از آن فقط در سیستم‌های با فرهنگ لغت کم (تقریبا ۱۰۰ کلمه برای حالت مستقل از گوینده و ۵۰۰ کلمه برای حالت وابسته به گوینده) استفاده شود. دلیل آن هم این بود که در سیستم‌های با تعداد لغات زیاد فضای مورد نیاز برای ذخیرهٔ الگوها و همچنین هزینهٔ جستجو به شدت افزایش پیدا می‌کرد. برای حل این مشکل پیشنهاد شد که از اجزاء تشکیل دهندهٔ کلمه‌ها مثل آواها به عنوان الگو استفاده و ذخیره شود. در این حالت نیز به علت تغییراتی که در مرزهای مربوط به آواها به علت آواهای قبلی و بعدی و همچنین تغییرات ناشی از

[103] Dynamic Time Warping

شرایط مختلف صحبت کردن و تغییرات موجود بین گوینده‌های مختلف وجود داشت، باعث ناکارآمد شدن روش پیچش زمانی پویا می‌شد [۳۷].

در دهه‌های ۱۹۷۰ و ۱۹۸۰ محققان برای غلبه بر مشکل تغییرپذیری موجود در سیگنال گفتار، مخصوصا در سیستم‌های با تعداد لغت زیاد، روی به روش‌های آماری آوردند. دو روش آماری متفاوتی که در این دوره برای استفاده در سیستم‌های بازشناخت گفتار مورد استفاده قرار گرفتند عبارت بودند از: مدل مخفی مارکوف و شبکه‌های عصبی مصنوعی (ANN) [۳۷].

تاریخچه مدل مخفی مارکوف برمی‌گردد به قبل از استفادهٔ آن در پردازش گفتار. تئوری اولیهٔ مدل مخفی مارکوف را Baum و همکارانش در اواخر دههٔ ۱۹۶۰ معرفی کردند و ارائهٔ آن برای استفاده در بازشناخت گفتار نتیجهٔ کارهای مستقل Baker در دانشگاه کارنگی ملون و Jelinek و همکارانش در IBM بود [۳۷].

مدل مخفی مارکوف علاوه بر بازشناخت گفتار [۱۹] در بسیاری از زمینه‌های دیگر مثل تشخیص پزشکی [۶۳]، تشخیص کاراکتر [۶۴] و غیره نیز کاربردهایی پیدا کرده است. در این فصل علاوه بر معرفی اجمالی مدل مخفی مارکوف مشهورترین بسته نرم افزاری موجود در این زمینه با نام HTK نیز معرفی می‌شود.

۱-۹ ۳-۲ مدل مخفی مارکوف

به طور کلی می‌توان گفت مدلهای مخفی مارکوف بیشتر در مسائلی می‌تواند قدرت خود را به خوبی نشان دهد که احتیاج به گرفتن رشته‌ایی از تصمیم‌ها است. در چنین فرآیندهایی ممکن است حالت سیستم در زمان t به طور مستقیم تحت تاثیر حالت زمانهای قبل باشد [۶۵]. مدل مخفی مارکوف در

واقع یک ماشین حالت محدود آماری[۱۰۴] است، که جهت مدل‌سازی یک پدیده مثلاً سخن گفته شده بکار می‌رود. این سخن می‌تواند لغت، حرف، واج و غیره باشد [۶۶].

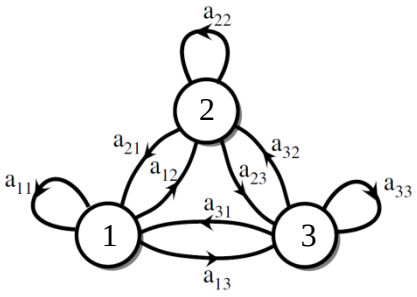
در ادامه به بررسی مدل مخفی مارکوف و سه مساله موجود در آن پرداخته می‌شود و سپس طریقهٔ استفاده از آن در سیستم‌های بازشناخت گفتار به طور مختصر بیان می‌شود.

۳-۲-۱ مدل‌های مارکوف مرتبهٔ اول

در صورتی که رشته‌ایی از حالتها به صورت پشت سر هم در نظر گرفته شود و حالت در هر زمان t با s_t نمایش داده شود آنگاه یک رشته حالت با طول T به صورت زیر نمایش داده می‌شود.

$$s_1^T = \{s_1, s_2, \dots, s_T\} \qquad (۱-۳)$$

به عنوان مثال شکل ۳- ۱ یک نمونه رشتهٔ مارکوف[۱۰۵] با سه حالت ۱، ۲ و ۳ را نشان می‌دهد. در این شکل احتمال انتقال از حالت i به حالت j با a_{ij} نشان داده شده است.



شکل ۳- ۱ یک نمونه رشتهٔ مارکوف با ۳ حالت [۶۵].

با افزایش زمان t حالت در سیستم نشان داده شده در شکل ۳-۱ با احتمال مربوطه تغییر می‌کند. همانطور که در شکل قابل مشاهده است یک حالت می‌تواند به خودش نیز برگردد.

---

104 Stochastic Finite State Automaton
105 Markov Chain

اگر فرض کنیم که $\mathbf{X} = X_1, X_2, \ldots X_n$ یک رشته از متغیرهای تصادفی باشد (هر کدام از X_iها معادل با یک حالت است). با استفاده از قانون بیز احتمال مربوط به رشته **X** به صورت زیر بدست می‌آید.

$$P(X_1, X_2, \ldots, X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_1^{i-1}) \qquad (۲-۳)$$

که در این رابطه $X_1^{i-1} = X_1, X_2, \ldots X_{i-1}$ است. همانطور که از رابطهٔ ۳-۲ نیز قابل مشاهده است انتقال به حالتی معین در زمان t به تمام حالتهای قبلی رخ داده بستگی دارد. به وضعیتی که انتقال حالت در زمان t فقط به n حالت قبلی خود بستگی داشته باشد فرآیند مارکوف درجهٔ n گفته می‌شود. در یک حالت خاص که معمولا در سیستم‌های بازشناخت گفتار استفاده می‌شود حالت در هر زمان فقط به حالت زمان قبل بستگی دارد و با نام مدل مارکوف مرتبهٔ اول شناخته می‌شود [۶۶]. به عبارت دیگر برقرار بودن رابطهٔ زیر نمایان‌گر مدل مارکوف مرتبهٔ اول است.

$$P(X_i \mid X_1^{i-1}) = P(X_i \mid X_{i-1}) \qquad (۳-۳)$$

در نتیجه با توجه به این شرط رابطهٔ ۳- ۲ به صورت زیر تبدیل می‌شود.

$$P(X_1, X_2, \ldots, X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1}) \qquad (۴-۳)$$

با استفاده از فرض بالا می‌توان از رشته مارکوف برای مدل کردن رویدادهای تغییر ناپذیر با زمان یا ایستا استفاده کرد، به طوری که احتمال متغیر تصادفی در هر لحظه فقط به لحظهٔ قبل بستگی داشته باشد.

$$P(X_i = s \mid X_{i-1} = s') = P(s \mid s') \qquad (۵-۳)$$

مدل مارکوفی که در بالا معرفی شد مدل مارکوف مشاهده پذیر نیز نامیده می‌شود زیرا خروجی فرآیند در هر زمان t مجموعه‌ایی از حالتها است، که هر حالت مربوط می‌شود به رخداد قابل مشاهدهٔ X_i . به عبارت دیگر یک تناظر یک به یک بین رشته رخداد قابل مشاهدهٔ **X** و رشتهٔ حالت مدل مارکوف وجود دارد [۴۴].

۳-۲-۲ مفهوم مدل مخفی مارکوف مرتبهٔ اول

در این بخش مفهوم مدل‌های مارکوف طوری گسترش می‌یابد تا حالتی که در آن مشاهده‌ها توابع احتمالی از حالت‌ها باشند را در بر بگیرد. این مدل جدید با نام مدل مخفی مارکوف شناخته می‌شود که به آن می‌توان به صورت یک فرآیند آماری دو لایه نگاه کرد که فرآیند آماری زیرین (رشتهٔ حالتها) به طور مستقیم قابل مشاهده نمی‌باشد. فرآیند آماری زیرین فقط می‌تواند به صورت احتمالی با فرآیند آماری قابل مشاهده که رشته‌ایی از ویژگی‌های قابل مشاهده را ایجاد می‌کند مربوط باشد [۴۴].

برای روشن‌تر شدن مفهوم مدل مخفی مارکوف می‌توان مدل کاسه و گلوله [۱۹] را مطرح کرد:

اگر فرض کنیم که تعداد N ظرف بزرگ در اتاق وجود دارد و درون هر کدام از این ظرفها تعداد زیادی گلوله با رنگ‌های مختلف وجود دارد، که هر رنگ در هر کاسه دارای احتمال مربوط به خود است، فرآیند بدست آوردن رشته قابل مشاهده را می‌توان به صورت زیر بیان کرد. فردی کاسه‌ایی را به طور تصادفی و بدون مشاهدهٔ آن انتخاب می‌کند و یک گلوله از آن برمی‌دارد و رنگ آن را یادداشت می‌کند سپس گلوله را به همان کاسه برمی‌گرداند. در صورتی که این فرآیند چندین بار تکرار شود یک رشتهٔ قابل مشاهده از رنگ‌ها به وجود می‌آید که به عنوان خروجی HMM در نظر گرفته می‌شود. در این مثال هر کاسه معادل با یک حالت بود که مخفی است و هر رنگ نیز در هر حالت دارای احتمال مربوط به خودش است [۱۹].

۳-۲-۳ پارامترهای لازم برای تعریف یک مدل مخفی مارکوف گسسته

یک مدل مخفی مارکوف با پارامترهای زیر تعریف می‌شود:

- $O=\{o_1,o_2,...,o_M\}$: رشتهٔ قابل مشاهدهٔ خروجی که مربوط می‌شود به خروجی فیزیکی سیستمی که می‌خواهد به وسیلهٔ HMM مدل شود. در مثال کاسه و گلوله رشتهٔ خروجی رشته رنگ‌های بدست آمده است.
- $\Omega=\{1,2,...,N\}$: مجموعه‌ایی از حالت‌ها که فضای حالت را به وجود می‌آورد. که در اینجا s_t نمایانگر حالت در زمان t است. در مثال کاسه و گلوله هر یک از کاسه‌ها معادل با یک حالت در مدل مخفی مارکوف هستند.
- $\mathbf{A}=\{a_{ij}\}$: ماتریس انتقال به طوری که a_{ij} احتمال انتقال از حالت i به حالت j است و به صورت زیر تعریف می‌شود.

$$a_{ij}=P(s_t=j\mid s_{t-1}=i) \qquad (۶-۳)$$

- $\mathbf{B}=\{b_i(k)\}$: ماتریس احتمال خروجی قابل مشاهدهٔ گسسته به طوری که $b_i(k)$ احتمال داشتن خروجی o_k است وقتی که سیستم به حالت i وارد شده باشد. به عبارت دیگر اگر $\mathbf{X}=X_1,X_2,...,X_t,...$ رشتهٔ خروجی مشاهده شده از مدل مخفی مارکوف باشد و $S=s_1,s_2,...,s_t,...$ رشته حالت‌های مخفی باشد آنگاه $b_i(k)$ به صورت زیر تعریف می‌شود:

$$b_i(k)=P(X_t=o_k\mid s_t=i) \qquad (۷-۳)$$

- $\pi=\{\pi_i\}$: توزیع اولیهٔ حالتها. در مدل کاسه و گلوله توزیع اولیهٔ حالتها مربوط می‌شود به اینکه در بار اول با چه احتمالی کدام کاسه را انتخاب کنیم.

$$\pi_i=P(s_0=i)\quad 1\le i\le N \qquad (۸-۳)$$

با توجه به اینکه تمام پارامترهای a_{ij}، $b_i(k)$ و π_i احتمال می‌باشند باید محدودیت‌های زیر را برآورده سازند:

$$a_{ij} \geq 0, \quad b_i(k) \geq 0, \quad \pi_i \geq 0 \quad \forall \quad all \quad i, j, k \qquad (۹-۳)$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad (۱۰-۳)$$

$$\sum_{k=1}^{M} b_i(k) = 1 \qquad (۱۱-۳)$$

$$\sum_{i=1}^{N} \pi_i = 1 \qquad (۱۲-۳)$$

به طور کلی برای تعیین کامل یک مدل مخفی مارکوف باید پارامترهای N و M که به ترتیب نمایانگر تعداد کل حالتها و تعداد عناصر قابل مشاهده است و همچنین سه ماتریس احتمال **A**، **B** و **π** مشخص شوند. برای راحتی از نماد زیر برای نشان دادن مجموعهٔ کامل پارامترهای یک مدل مخفی مارکوف استفاده می‌شود.

$$\Phi = (\mathbf{A}, \mathbf{B}, \pi) \qquad (۱۳-۳)$$

۳-۲-۴ سه مسألهٔ اصلی در مدل‌های مخفی مارکوف

قبل از اینکه بتوان مدل‌های مخفی مارکوف را به مسائل واقعی اعمال کرد سه مسألهٔ اولیه باید حل شود. این سه مسأله عبارتند از:

۱. **ارزیابی**[۱۰۶]: در این حالت مدل مخفی مارکوف با پارامترهای Φ و رشتهٔ مشاهده شدهٔ $\mathbf{X}=(X_1,X_2,...,X_T)$ داده شده است و هدف پیدا کردن احتمال مربوط به تولید شدن این رشته توسط HMM با پارامترهای داده شده است. یعنی: $P(\mathbf{X}|\Phi)$.

۲. **رمزگشایی یا آشکارسازی**[۱۰۷]: در این حالت مدل مخفی مارکوف با پارامترهای Φ و رشتهٔ مشاهده شدهٔ $\mathbf{X}=(X_1,X_2,...,X_T)$ داده شده است و هدف پیدا کردن رشته حالتهایی از مدل مخفی مارکوف داده شده است که با بیشترین احتمال رشتهٔ مشاهده شدهٔ داده شده را تولید کرده باشند.

۳. **یادگیری یا آموزش**[۱۰۸]: در این حالت مدل مخفی مارکوف با پارامترهای فرضی و اولیهٔ Φ و مجموعه‌ایی از مشاهده‌ها داده شده است و هدف پیدا کردن پارامترهای جدید $\hat{\Phi}$ برای مدل مخفی مارکوف است به طوری که احتمال توأم $\prod_{\mathbf{X}} P(\mathbf{X}|\Phi)$ را ماکزیمم کند.

با حل شدن مسألهٔ ارزیابی می‌توان گفت که یک مدل مخفی مارکوف داده شده تا چه اندازه با رشتهٔ مشاهده شده مطابقت دارد. بنابراین از مدل مخفی مارکوف می‌توان برای تشخیص الگو استفاده کرد. به عبارت دیگر با محاسبه احتمال $P(\mathbf{X}|\Phi)$ می‌توان با استفاده از قانون بیز احتمال پسین $P(\Phi|\mathbf{X})$ را بدست آورد و مدل مخفی مارکوفی که بیشترین احتمال پسین را داشته باشد به عنوان مدلی که با بیشترین احتمال رشتهٔ مشاهده شده را تولید کرده است، انتخاب می‌شود. با حل شدن مسألهٔ رمزگشایی می‌توان بهترین رشته از حالتهای موجود در مدل مخفی مارکوف را که با بیشترین احتمال رشتهٔ مشاهده شده را تولید کرده‌اند، بدست آورد. به عبارت دیگر می‌توان رشتهٔ حالتهای مخفی را آشکار کرد. بالاخره با حل شدن مسألهٔ نه چندان سادهٔ یادگیری می‌توان به صورت

---

106 Evaluation
107 Decoding
108 Learning

اتوماتیک پارامترهای مربوط به مدل مخفی مارکوف را از روی مجموعه‌ایی از داده‌های آموزشی بدست آورد [۴۴].

۳-۲-۴-۱ حل مسألهٔ ارزیابی در مدل مخفی مارکوف

در مسألهٔ ارزیابی هدف پیدا کردن احتمال $P(\mathbf{X}|\Phi)$ است در صورتی که رشتهٔ مشاهده شده، $\mathbf{X}=(X_1,X_2,...,X_T)$، و پارامترهای HMM مورد نظر، Φ، داده شده باشد. برای این منظور مستقیم‌ترین راه، محاسبه مجموع احتمال‌های مربوط به تمام رشته حالتهای ممکن است. به عبارت دیگر برای محاسبهٔ این احتمال باید در ابتدا تمام رشته حالتهای ممکن که طول آنها T است پیدا شود سپس باید معین شود که هر یک از این رشته حالتها با چه احتمالی رشتهٔ مشاهده شده را تولید کرده‌اند. در نهایت احتمال مورد نظر مجموع تمام این احتمال‌ها است [۴۴]. این روش مستقیم را به صورت زیر می‌توان نوشت:

$$P(\mathbf{X}|\Phi)=\sum_{all\ \mathbf{S}} P(\mathbf{S}|\Phi)P(\mathbf{X}|\mathbf{S},\Phi) \qquad (۱۴-۳)$$

که در این رابطه $\mathbf{S}$ نمایانگر تمام مسیرهای موجود برای تولید رشتهٔ مشاهده شده است. به عنوان نمونه برای رشتهٔ حالت $\mathbf{S}=(s_1,s_2,...,s_T)$، جملهٔ اول موجود در طرف دوم رابطه بالا با اعمال فرض اول مارکوف به صورت زیر محاسبه می‌شود:

$$P(\mathbf{S}|\Phi)=P(s_1|\Phi)\prod_{t=2}^{T}P(s_t|s_{t-1},\Phi)=\pi_{s_1}a_{s_1s_2}..a_{s_{T-1}s_T}=a_{s_0s_1}a_{s_1s_2}..a_{s_{T-1}s_T} \qquad (۱۵-۳)$$

جملهٔ دوم نیز به صورت زیر قابل محاسبه است:

$$\begin{aligned}P(\mathbf{X}|\mathbf{S},\Phi)&=P(X_1^T|S_1^T,\Phi)=\prod_{t=1}^{T}P(X_t|s_t,\Phi)\\&=b_{s_1}(X_1)b_{s_2}(X_2)..b_{s_T}(X_T)\end{aligned} \qquad (۱۶-۳)$$

با قرار دادن دو رابطهٔ ۳-۱۵ و ۳-۱۶ در رابطهٔ ۳-۱۴ داریم:

$$P(\mathbf{X}|\Phi) = \sum_{all\ \mathbf{S}} P(\mathbf{S}|\Phi)P(\mathbf{X}|\mathbf{S},\Phi)$$
$$= \sum_{all\ \mathbf{S}} a_{s_0 s_1} b_{s_1}(X_1) a_{s_1 s_2} b_{s_2}(X_2)..a_{s_{T-1} s_T} b_{s_T}(X_T) \qquad (۱۷-۳)$$

برای محاسبهٔ رابطهٔ فوق نیاز به $(2\mathrm{T}-1)\mathrm{N}^{\mathrm{T}}$ ضرب و $\mathrm{N}^{\mathrm{T}}-1$ جمع است. عملا با این حجم محاسبات راه حل مستقیم فوق تقریبا غیر ممکن است. برای کاهش حجم محاسبات از الگوریتم پیش‌رو[۱۰۹] استفاده می‌شود. برای استفاده از این الگوریتم پارامتر $\alpha_t(i)$ به صورت زیر تعریف می‌شود.

$$\alpha_t(i) = P(X_1^t, s_t = i|\Phi) \qquad (۱۸-۳)$$

به عبارت دیگر منظور از $\alpha_t(i)$ احتمال بودن مدل مخفی مارکوف در حالت i در زمان t و تولید کردن رشتهٔ مشاهده شدهٔ $X_1, X_2, ...X_t$ تا این زمان است.

$\alpha_t(i)$ها به وسیله دو رابطهٔ زیر بدست می‌آیند:

$$\alpha_1(i) = \pi_i b_i(X_1) \quad 1 \le i \le N \qquad (۱۹-۳)$$

$$\alpha_t(j) = [\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}] b_j(X_t) \quad 2 \le t \le T; \quad 1 \le j \le N \qquad (۲۰-۳)$$

در نهایت احتمال مورد نظر به صورت زیر بدست می‌آید:

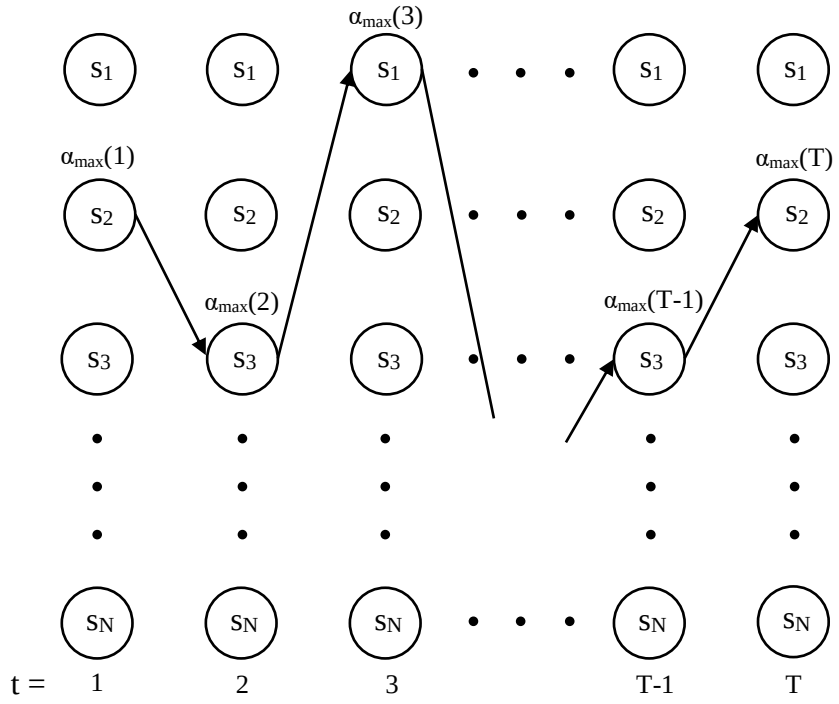
$$P(\mathbf{X}|\Phi) = \sum_{i=1}^{N} \alpha_T(i) \qquad (۲۱-۳)$$

۳-۲-۴-۲ حل مسألهٔ رمزگشایی در مدل مخفی مارکوف

[109] Forward Algorithm

الگوریتم پیش‌رو معرفی شده در قسمت قبل فقط احتمال تولید شدن رشتهٔ مشاهده شده را توسط مدل مخفی مارکوف معین می‌کرد و این کار را با جمع کردن احتمال مربوط به تمام مسیرهای ممکن انجام می‌داد. در این قسمت هدف پیدا کردن مسیری است که با بیشترین احتمال رشتهٔ مشاهده شده را تولید کرده است. به عبارت دیگر هدف پیدا کردن رشته حالت مخفی $\mathbf{S} = (s_1, s_2, ..., s_T)$ است به طوری که احتمال $P(\mathbf{S}, \mathbf{X} | \Phi)$ ماکزیمم شود. برای این منظور از الگوریتم ویتربی[۱۱۰] استفاده می‌شود.

به الگوریتم ویتربی می‌توان به عنوان الگوریتم پیش‌رو اصلاح شده نیز نگاه کرد. به عبارت دیگر، همانطور که در شکل ۳-۲ نشان داده شده است، الگوریتم ویتربی در هر بار محاسبهٔ $\alpha_t(i)$ها، $\alpha_t(i)$ با بیشترین احتمال را انتخاب و در حافظه ذخیره می‌کند. به این ترتیب در انتها رشتهٔ ذخیره شده همان رشتهٔ حالتهای مخفی است.



شکل ۳- ۲ فرآیند پیدا کردن بهترین رشتهٔ حالت به کمک الگوریتم ویتربی [۴۴].

[110] Viterbi Algorithm

۳-۲-۴-۳ حل مسألهٔ یادگیری در مدل مخفی مارکوف

اگر پارامترهای مدل‌های مخفی مارکوف موجود نباشد نمی‌توان از آنها در تشخیص الگو استفاده کرد. بنابراین باید به طریقی این پارامترها را بدست آورد. بهترین گزینه برای بدست آوردن این پارامترها استفاده از داده‌های آموزشی است. بنابراین در مسئلهٔ یادگیری در مدل مخفی مارکوف هدف پیدا کردن پارامترهای مدل مخفی مارکوف است به طوری که رشتهٔ مشاهده شده را با دقت توصیف کند. این مسأله به وسیلهٔ الگوریتم بازگشتی بام ولچ[۱۱۱] که با نام پیش‌رو-پس‌رو نیز شناخته می‌شود، قابل حل است [۴۴].

الگوریتم بام ولچ برای محاسبهٔ پارامترهای مدل مخفی مارکوف از احتمال پیش‌رو و احتمال پس‌رو استفاده می‌کند. احتمال پس‌رو به صورت زیر تعریف می‌شود:

$$\beta_t(i) = P(X_{t+1}^T \mid s_t = i, \Phi) \qquad (۲۲-۳)$$

در این رابطه $\beta_t(i)$ احتمال تولید شدن رشتهٔ قابل مشاهدهٔ X_{t+1}^T (از زمان t+1 تا آخر) برای مدل مخفی مارکوف داده شده است به شرطی که مدل مخفی مارکوف در لحظهٔ t در حالت i بوده باشد. $\beta_t(i)$ها به صورت زیر قابل محاسبه هستند:

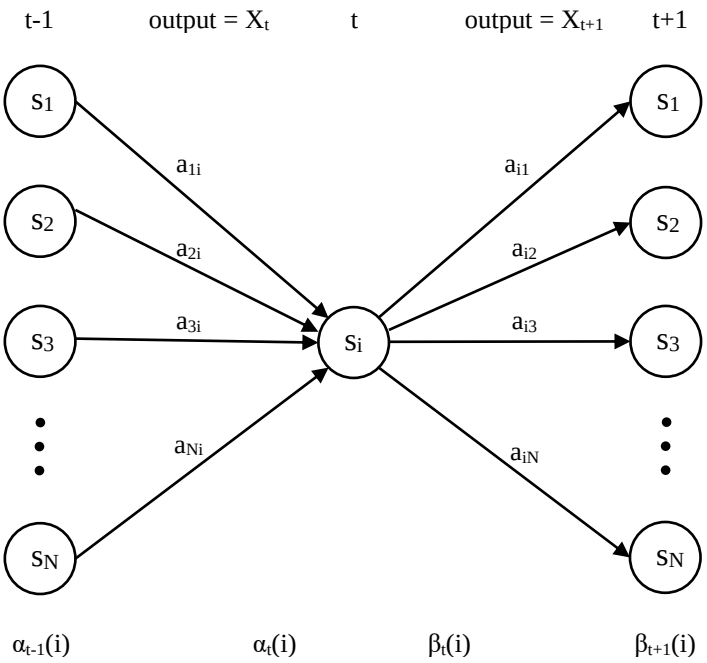
$$\beta_T(i) = 1/N \quad 1 \le i \le N \qquad (۲۳-۳)$$

$$\beta_t(i) = [\sum_{j=1}^{N} a_{ij} b_j(X_{t+1}) \beta_{t+1}(j)] \quad t = T-1 \ldots 1; \quad 1 \le i \le N \qquad (۲۴-۳)$$

به عبارت دیگر همانطور که در شکل ۳-۳ نشان داده شده است $\alpha_t(i)$ها به صورت بازگشتی از چپ به راست ولی $\beta_t(i)$ها به صورت بازگشتی از راست به چپ محاسبه می‌شوند [۴۴].

---

[111] Baum-Welch Algorithm



شکل ۳- ۳ رابطهٔ بین α و β در الگوریتم بام ولچ [۴۴].

احتمال انتقال از حالت i به حالت j در زمان t به شرطی که پارامترهای مدل مخفی مارکوف و رشتهٔ قابل مشاهدهٔ خروجی داده شده باشد با $\gamma_t(i,j)$ نشان داده می‌شود و به وسیله رابطهٔ زیر بدست می‌آید.

$$\begin{aligned}\gamma_t(i,j) &= P(s_{t-1}=i, s_t=j \mid X_1^T, \Phi) \\ &= \frac{P(s_{t-1}=i, s_t=j, X_1^T \mid \Phi)}{P(X_1^T \mid \Phi)} \\ &= \frac{\alpha_{t-1}(i)a_{ij}b_j(X_t)\beta_t(j)}{\sum_{k=1}^{N}\alpha_T(k)}\end{aligned} \qquad (۲۵ -۳)$$

الگوریتم بازگشتی بام ولچ در هر بار تکرار با ماکزیمم کردن احتمال $P(\mathbf{X}|\Phi)$، به وسیلهٔ الگوریتم EM[۱۱۲]، پارامترهای جدید، $\hat{\Phi}$، را از روی پارامترهای قبلی، Φ، بدست می‌آورد.

پارامترهای بهینه یافته $\hat{a}_{ij}$ از روی پارامترهای قبلی، a_{ij}، به وسیلهٔ رابطهٔ زیر بدست می‌آید.

[112] Expectation Maximization Algorithm

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{k=1}^{N} \gamma_t(i,k)} \qquad (۲۶-۳)$$

این رابطه در حقیقت نسبت تعداد انتقال‌های مورد انتظار بین حالت i و حالت j به تمام انتقال‌های مورد انتظار از حالت i می‌باشد. به طور مشابه پارامترهای بهینه یافتهٔ $\hat{b}_j(k)$ از روی پارامترهای $b_j(k)$ با استفاده از رابطهٔ زیر بدست می‌آیند [۴۴].

$$\hat{b}_j(k) = \frac{\sum_{t \in X_t = o_k} \sum_{j} \gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{j} \gamma_t(i,j)} \qquad (۲۷-۳)$$

۳-۲-۵ انواع مدل مخفی مارکوف از لحاظ ساختار

مدل مخفی مارکوفی که در قسمت‌های قبل در نظر گرفته شد، مدل مخفی مارکوف از نوع ارگودیک[۱۱۳] بود که از هر حالت به حالت دیگر امکان ارتباط مستقیم وجود داشت (شکل ۳-۴-الف). در این نوع مدل مخفی مارکوف تمام ضرایب a_{ij} بزرگتر از صفر هستند.

نوع دیگری از مدل مخفی مارکوف که در بعضی از کاربردها مخصوصا در بازشناخت گفتار مورد استفاده قرار می‌گیرد، مدل نشان داده شده در شکل ۳-۴-ب است که با نام مدل چپ به راست[۱۱۴] شناخته می‌شود. در این مدل مخفی مارکوف هر حالت می‌تواند به خود یا به حالت‌های بعدی برود و امکان برگشت به حالت‌های قبل وجود ندارد. در تمام مدل‌های مخفی مارکوف چپ به راست ضرایب انتقال باید دارای خاصیت زیر باشند.

---

113 Ergodic Model
114 Left-Right Model

$$a_{ij} = 0, \qquad j < i \qquad (۲۸ -۳)$$

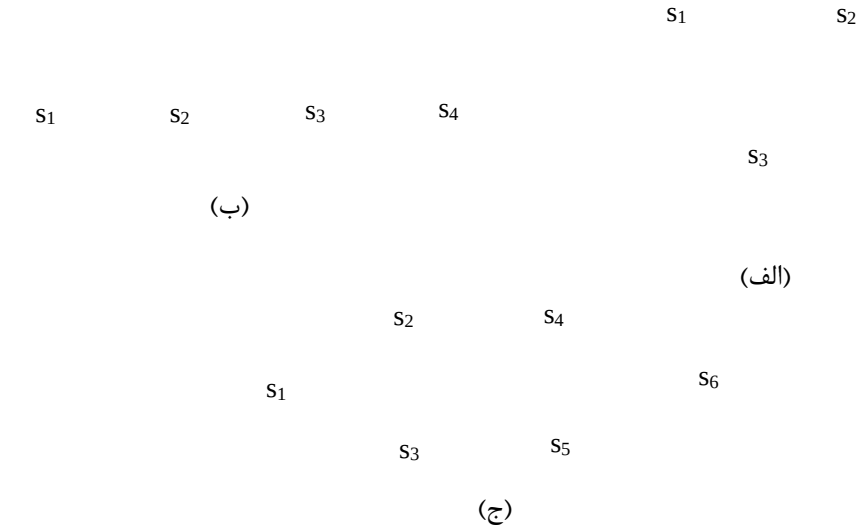
علاوه بر این احتمال اولیه نیز باید دارای خاصیت زیر باشد:

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \qquad (۲۹ -۳)$$

گاهی اوقات تعداد پرشها در مدل چپ به راست محدود می‌شود. یعنی:

$$a_{ij} = 0, \qquad j > i + \Delta \qquad (۳۰ -۳)$$

به عنوان مثال در مدل مخفی مارکوف نشان داده شده در شکل ۳-۴-ب مقدار Δ برابر با ۲ است. ساختارهای متفاوت دیگری هم وجود دارند به عنوان مثال شکل ۳-۴-ج ترکیب موازی دو مدل مخفی مارکوف چپ به راست را نشان می‌دهد.



شکل ۳- ۴ نمایش سه نوع HMM (الف) ارگودیک. (ب) چپ به راست. (ج) ترکیب موازی دو مدل چپ به راست.

۳-۲-۶ مدل‌های مخفی مارکوف با چگالی‌های مشاهده‌ایی پیوسته[۱۱۵]

مدل‌های مخفی مارکوفی که تاکنون مورد بحث قرار گرفتند دارای رشتهٔ قابل مشاهدهٔ گسسته بودند که از یک مجموعهٔ محدود انتخاب شده بود. همچنین احتمال‌های مربوط به آنها نیز به صورت گسسته بود. در اغلب مسائل عملی با سیگنال‌های پیوسته سروکار داریم و مدل مخفی مارکوف پیوسته دارای کارآیی بیشتری است [۱۹].

در مدل‌های مخفی مارکوف پیوسته خروجی قابل مشاهده از یک مجموعهٔ محدود نیست بلکه از یک فضای پیوسته می‌آید. تفاوت اصلی بین مدل مخفی مارکوف پیوسته و گسسته مربوط می‌شود به تابع احتمال خروجی آنها.

برای انتخاب تابع چگالی احتمال خروجی $b_j(\mathbf{x})$، بهترین انتخاب مخلوطی از گوسی‌ها است. چون آنها می‌توانند تقریبا هر تابع چگالی احتمال پیوسته‌ایی را به خوبی تخمین بزنند [۴۴]. در صورتی که از M تابع چگالی گوسی استفاده شود، تابع چگالی احتمال خروجی به صورت زیر نمایش داده می‌شود:

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk} N(\mathbf{x}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) = \sum_{k=1}^{M} c_{jk}\, b_{jk}(\mathbf{x}) \qquad (۳۱-۳)$$

که در این رابطه $N(\mathbf{x}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ یا $b_{jk}(\mathbf{x})$ نمایانگر تابع چگالی گوسی واحد با بردار میانگین $\boldsymbol{\mu}_{jk}$ و ماتریس کواریانس $\boldsymbol{\Sigma}_{jk}$ مربوط به حالت j ام مدل مخفی مارکوف می‌باشد. همچنین در این رابطه M تعداد اجزاء مخلوط گوسی را مشخص می‌کند و c_{jk} وزن داده شده به تابع k ام است که باید شرط زیر را برآورده کند:

$$\sum_{k=1}^{M} c_{jk} = 1 \qquad (۳۲-۳)$$

---

[115] Continuous Hidden Markov Model

فرآیند بهینه سازی ضرایب مربوط به تابع‌های مخلوط گوسی شبیه به حالت گسسته می‌باشد. فرمول‌های بازگشتی مربوط به بهینه‌سازی ضرایب به صورت زیر می‌باشند [۴۴]:

$$\hat{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\mathbf{x}_t}{\sum_{t=1}^{T} \xi_t(j,k)} \qquad (۳۳-۳)$$

$$\hat{\boldsymbol{\Sigma}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\,(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_{jk})\,(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_{jk})^{t}}{\sum_{t=1}^{T} \xi_t(j,k)} \qquad (۳۴-۳)$$

$$\hat{c}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)}{\sum_{t=1}^{T} \sum_{k} \xi_t(j,k)} \qquad (۳۵-۳)$$

که در این روابط $\xi_t(j,k)$ به صورت زیر محاسبه می‌شود:

$$\xi_t(j,k) = \frac{\sum_i \alpha_{t-1}(i)\, a_{ij}\, c_{jk}\, b_{jk}(\mathbf{x}_t)\, \beta_t(j)}{\sum_{i=1}^{N} \alpha_T(i)}. \qquad (۳۶-۳)$$

۳-۲-۷ اصول بازشناخت گفتار بوسیلهٔ مدل‌های مخفی مارکوف

اجزاء اصلی یک سیستم بازشناخت گفتار در فصل قبل معرفی شدند. همانطور که در شکل ۲-۲ نشان داده شد در ابتدا سیگنال صحبت ورودی وارد قسمت استخراج ویژگی می‌شود و تبدیل به رشته‌ایی از بردارهای ویژگی با اندازهٔ یکسان می‌شود. سپس قسمت رمزگشا یا آشکارساز سعی در پیدا کردن رشته‌ایی از کلمه‌ها می‌کند که با بیشترین احتمال رشتهٔ بردارهای ویژگی را تولید کرده باشند [۶۷].

همانطور که قبلا نیز گفته شد رمزگشا برای پیدا کردن رشته کلمهٔ گفته شده از مدل صوتی، مدل زبان و دیکشنری که شامل تلفظ تمام کلمه‌ها است استفاده می‌کند. قسمت دیکشنری زمانی کاربرد دارد که به جای تولید مدل صوتی مربوط به هر یک از کلمه‌ها، مدل صوتی مربوط به هر یک از اجزاء تشکیل دهندهٔ کلمه‌ها مثل آواها ایجاد شده باشد. در این حالت اجزاء تشکیل دهندهٔ هر کلمه به کمک قسمت دیکشنری بدست می‌آید.

برای آشنایی با اساس بازشناخت گفتار به کمک مدل مخفی مارکوف، در ابتدا بازشناخت کلمات جدا از هم معرفی شده است.

۳-۲-۷-۱ بازشناخت کلمه‌های جدا از هم

اگر هر کلمهٔ گفته شده به صورت رشته‌ایی از بردارهای ویژگی یا بردارهای مشاهده شده، **O**، به صورت زیر تعریف شود:

$$\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T \qquad (۳۷-۳)$$

که $\mathbf{o}_t$ بردار صوتی مشاهده شده در زمان t می‌باشد، مسألهٔ بازشناخت کلمات جدا از هم را می‌توان به صورت زیر در نظر گرفت:

$$\arg\max_i \{P(\omega_i \mid \mathbf{O})\} \qquad (۳۸-۳)$$

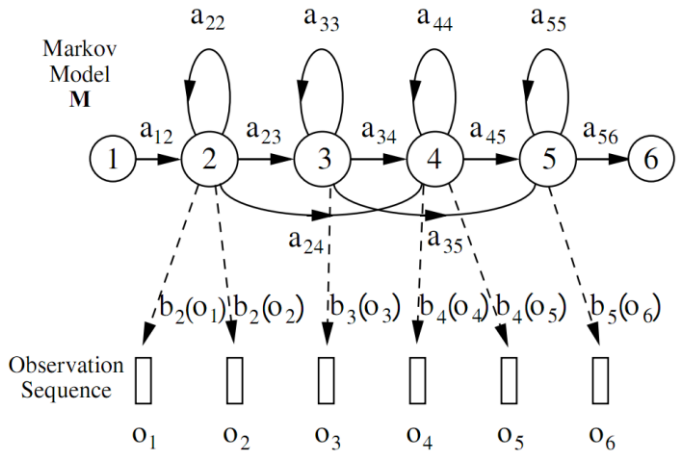
که ω_i i امین کلمهٔ موجود در فهرست لغات است. احتمال بالا را نمی‌توان به طور مستقیم حساب کرد در نتیجه به کمک قانون بیز داریم:

$$P(\omega_i \mid \mathbf{O}) = \frac{P(\mathbf{O} \mid \omega_i) P(\omega_i)}{P(\mathbf{O})} \qquad (۳۹-۳)$$

بنابراین برای مجموعهٔ اطلاعات پیشین داده شده ($P(\omega_i)$)، محتمل‌ترین کلمهٔ گفته شده فقط به احتمال $P(\mathbf{O} \mid \omega_i)$ بستگی دارد. در صورتی که ابعاد رشتهٔ مشاهده شدهٔ **O** داده شده باشد، محاسبهٔ مستقیم احتمال توأم شرطی $P(\mathbf{o}_1, \mathbf{o}_2, \ldots \mid \omega_i)$ از روی کلمه‌های نمونهٔ داده شده عملی نیست. اما اگر از یک مدل پارامتری مثل مدل مخفی مارکوف برای تولید کلمه استفاده شود، آنگاه

تخمین از روی داده‌ها امکان‌پذیر است زیرا تخمین چگالی‌های مشاهده شده مربوط به کلاس شرطی $P(\mathbf{O}\,|\,\omega_i)$ تبدیل به یک مسألهٔ سادهٔ تخمین پارامترهای مدل مخفی مارکوف می‌شود [۲۹].

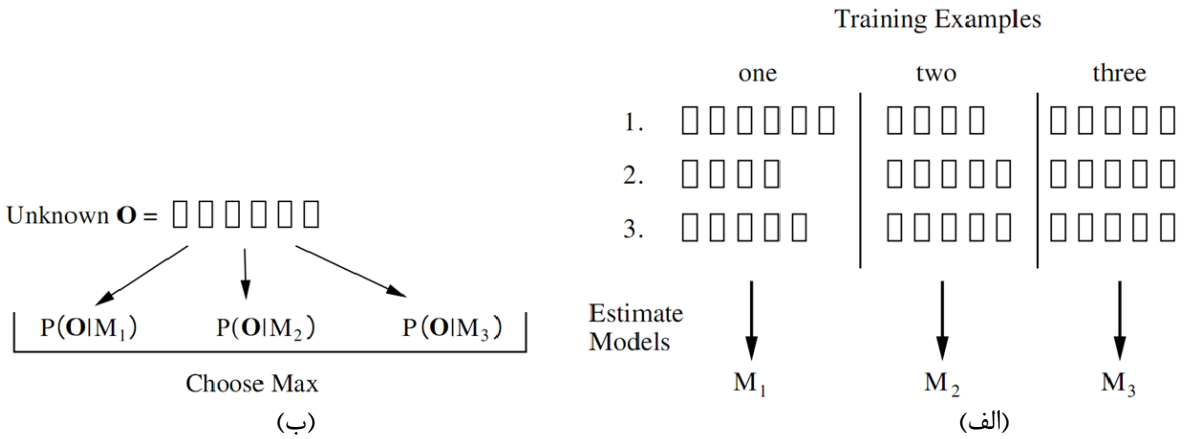
همانطور که در شکل ۳-۵ نشان داده شده است در بازشناخت کلمه‌ها بر پایهٔ مدل مخفی مارکوف، فرض می‌شود که رشته بردارهای صوتی مشاهده شده مربوط به هر کلمه به وسیلهٔ مدل مخفی مارکوف تولید شده است.



شکل ۳- ۵ یک نمونه مدل مخفی مارکوف برای یک کلمه [۲۹].

در مدل مخفی مارکوف نشان داده شده در شکل ۳-۵ هر بار که یک واحد به زمان افزوده می‌شود حالت مربوط به آن تغییر می‌کند و با توجه به تابع خروجی مربوط به آن حالت یک بردار صوتی قابل مشاهده تولید می‌شود. به عنوان مثال اگر مدل مخفی مارکوف در زمان t وارد حالت j شود، بردار صوتی $\mathbf{o}_t$ از روی چگالی احتمال $b_j(\mathbf{o}_t)$ تولید می‌شود. علاوه بر این احتمال انتقال از حالت i به حالت j نیز به صورت گسسته و با ضرایب a_{ij} داده می‌شود. همانطور که در شکل ۳-۵ نشان داده شده است حالت‌های اول و آخر در مدل مخفی مارکوف نشان داده شده دارای خروجی قابل مشاهده نیستند، علت این امر در بخش بعد توضیح داده خواهد شد[۲۹].

در صورتی که برای هر کلمه یک مدل مخفی مارکوف ایجاد شود، آنگاه به کمک الگوریتم پیش‌رو بیان شده در بخش ۳-۲-۴ می‌توان احتمال $P(\mathbf{O}\,|\,\omega_i)$ مربوط به تمام کلمه‌ها را حساب کرد و کلمه‌ایی که دارای بیشترین احتمال باشد به عنوان کلمهٔ مورد نظر تشخیص داده می‌شود. البته تمام اینها در صورتی است که پارامترهای مدل مخفی مارکوف مربوط به هر کلمه یعنی $\{a_{ij}\}$ و $\{b_j(\mathbf{o}_t)\}$ از قبل بدست آمده باشد. همانطور که در فصل قبل گفته شد مدل مخفی مارکوف قادر است به کمک الگوریتم بام ولچ و تعدادی دادهٔ آموزشی پارامترهای خود را به صورت کارآمدی تخمین بزند و این یکی از مزیت‌های مهم مدل مخفی مارکوف می‌باشد. شکل ۳-۶ به طور خلاصه عملیات بازشناخت کلمه‌های جدا از هم به کمک مدل مخفی مارکوف را نشان می‌دهد [۲۹].



شکل ۳- ۶ بازشناخت کلمه‌های جدا از هم به کمک مدل مخفی مارکوف [۲۹]. (الف) مرحلهٔ آموزش. (ب) مرحلهٔ بازشناخت.

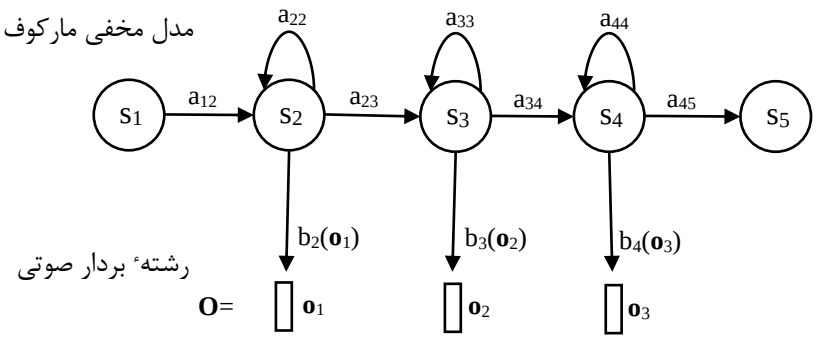
همانطور که در شکل ۳-۶-الف نشان داده شده است در ابتدا مدل مخفی مارکوف مربوط به هر کلمهٔ موجود در فرهنگ لغت (در اینجا فقط سه کلمه وجود دارد) به کمک فایل‌های گفتاری مربوط به همان کلمه آموزش داده می‌شود. سپس در مرحلهٔ بازشناخت همانطور که در شکل ۳-۶-ب نشان داده شده است، برای تشخیص کلمهٔ نامعلوم ورودی، احتمال اینکه هر مدل بردار مشاهده شدهٔ مربوط به کلمهٔ نامعلوم ورودی را تولید کرده باشد حساب می‌شود و کلمهٔ نامعلوم به مدلی تعلق می‌گیرد که احتمال بیشتری داشته باشد [۲۹].

۳-۲-۷-۲ بازشناخت گفتار پیوسته

با کمی دقت می‌توان متوجه شد که حالت تعمیم یافتهٔ بازشناخت گفتار پیوسته فقط شامل به هم پیوستن مدل‌های مخفی مارکوف در یک رشته است. هر مدل در طول این رشته به طور مستقیم مربوط می‌شود به نماد اصلی فرض شده در سیگنال گفتار. این نمادها هم می‌توانند کلمه‌ها باشند که در این حالت بازشناخت گفتار متصل خوانده می‌شود و هم می‌تواند اجزاء تشکیل دهندهٔ کلمه‌ها مثل آواها باشند که در این حالت بازشناخت گفتار پیوسته خوانده می‌شود. دلیل قرار دادن حالت‌های ابتدا و انتهای بدون خروجی قابل مشاهده در مدل مخفی مارکوف در قسمت قبل این بود که بتوان در این مرحله مدل‌ها را به یکدیگر متصل کرد[۲۹].

۳-۲-۷-۲-۱ مدل‌های صوتی بر پایهٔ مدل مخفی مارکوف

همانطور که در بالا به آن اشاره شد در بازشناخت گفتار پیوسته در ابتدا کلمه‌های گفته شده موجود در رشتهٔ W به کمک قسمت دیکشنری به رشته‌ایی از آواها تبدیل می‌شوند. سپس هر آوا به وسیلهٔ یک مدل مخفی مارکوف مدل می‌شود. مدل مخفی مارکوفی که در بازشناخت گفتار پیوسته برای مدل کردن آواها استفاده می‌شود معمولا دارای ۵ حالت است که فقط ۳ حالت آن دارای خروجی قابل مشاهده است. حالتهای ابتدا و انتها به خاطر اتصال مدل‌های مختلف به یکدیگر دارای خروجی نیستند (شکل ۳-۷).



شکل ۳- ۷ یک نمونه مدل مخفی مارکوف مربوط به یک آوا.

اگر رشتهٔ آواهای ایجاد شده از کلمه‌های گفته شده با Q نشان داده شود، احتمال صوتی مورد نیاز برای بازشناخت با رابطهٔ زیر بدست می‌آید [۶۸].

$$p(\mathbf{O}|Q) = \sum_S p(S, \mathbf{O}|Q) , \qquad (۴۰ -۳)$$

که در این رابطه $S = s(0), s(1), \dots, s(T)$ رشته‌ایی از حالتها در طول مدل مخفی مارکوف مرکب ایجاد شده است و

$$p(S, \mathbf{O}|Q) = a_{s(0),s(1)} \prod_{t=1}^{T} b_{s(t)}(\mathbf{o}_t) a_{s(t),s(t+1)} . \qquad (۴۱ -۳)$$

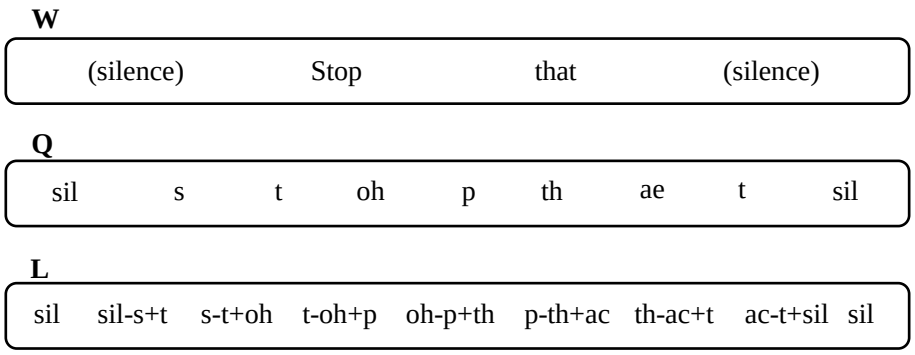
مهمترین مشکلی که در این حالت وجود دارد این است که با تبدیل هر کلمه به رشته‌ایی از آواها که نسبت به متن مستقل هستند امکان در نظر گرفتن و ذخیرهٔ تغییرات موجود در متن ناشی از تأثیر آواها بر هم وجود ندارد. برای حل این مشکل به جای اینکه مدل مخفی مارکوف را برای هر آوا ایجاد کنند، مدل مخفی مارکوف را برای تمام رشته‌های سه تایی از آواها[۱۱۶] ایجاد می‌کنند. در این حالت تغییرات ناشی از آواهای قبل و بعد از هر آوا نیز مدل خواهد شد. به عنوان مثال همانطور که در شکل ۳-۸ برای جملهٔ stop that نشان داده شده است در ابتدا کلمه‌های مورد نظر توسط دیکشنری به ساختار آوایی تبدیل می‌شوند سپس هر یک از آواها با آوای قبل و بعد از خودشان ترکیب می‌شوند و مدل مخفی مارکوف برای هر یک از اجزاء این رشته جدید ایجاد می‌شود. در این شکل علامت x-q+y نمایانگر یک رشته سه آوایی است که در آن آوای q به ترتیب بعد و قبل از دو آوای x و y قرار گرفته است.

مشکلی که در این حالت ممکن است به وجود آید تعداد زیاد مدل‌ها و نبودن دادهٔ آموزشی کافی برای آموزش آنها است. چون به عنوان مثال اگر به تعداد N آوای پایه وجود داشته باشد تعداد N^3 رشتهٔ سه آوایی قابل بدست آمدن است. برای حل این مشکل پارامترهای بعضی از حالتهای مشابه

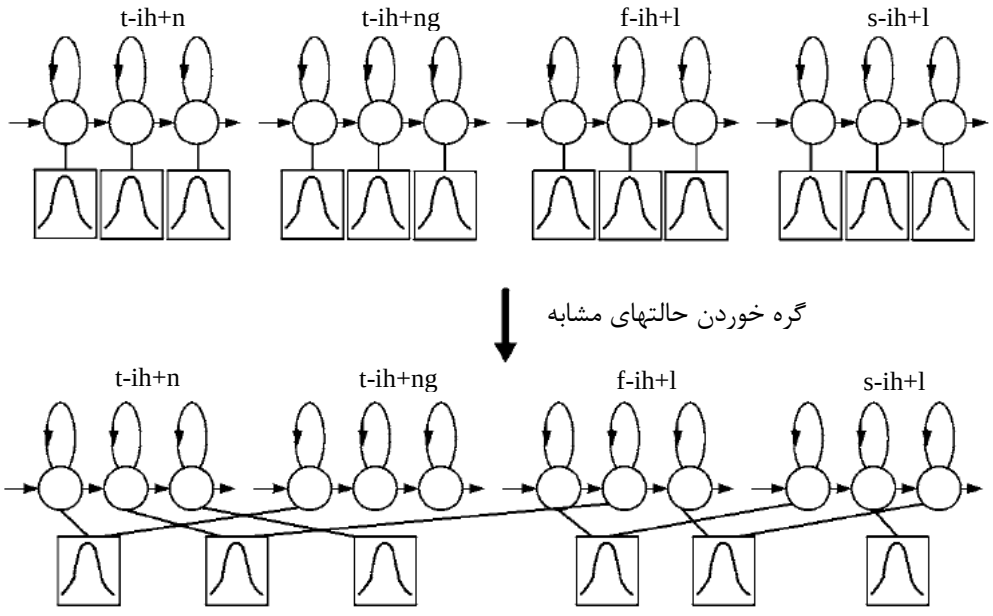
---

[116] Triphone

مربوط به مدل‌های مخفی مارکوف با هم به اشتراک گذاشته می‌شوند. به عبارت دیگر بعضی از مدل‌ها به هم گره می‌خورند [۶۸]. در شکل ۳-۹ یک نمونه از فرآیند گره خوردن حالت‌های مشابه نشان داده شده است. باید توجه داشت که اینکه چه حالتهایی باید به هم گره بخورند بستگی به آواهای موجود در رشته و میزان شباهت آنها دارد. برای این منظور از درخت تصمیم گیری استفاده می‌شود [۶۹].



شکل ۳- ۸ بدست آوردن رشته‌های سه آوایی از روی کلمه‌ها [۶۸].



شکل ۳- ۹ ایجاد پیوند بین حالت‌ها در مدل‌های مربوط به آواهای مشابه [۶۸].

۳-۲-۷-۲-۲ آموزش ادغام شده[۱۱۷]

داده‌های آموزشی مورد نیاز برای آموزش یک سیستم بازشناخت گفتار پیوسته باید به صورت پیوسته باشد. در حالت کلی مرزهای مربوط به اجزاء تشکیل دهندهٔ کلمه‌ها در داده‌های آموزشی پیوسته مشخص نیست. اگر تعداد داده‌های آموزشی کم باشد شاید بتوان با دست عمل مرزبندی را انجام داد ولی در حالت کلی به تعداد زیادی دادهٔ آموزشی نیاز است و انجام این عمل امکان پذیر نیست. برای حل این مشکل از روش به اصطلاح آموزش ادغام شده استفاده می‌شود.

آموزش ادغام شده هم مانند حالت گسسته از الگوریتم بام ولچ استفاده می‌کند ولی با این تفاوت که به جای آموزش هر مدل به صورت جداگانه تمام مدل‌ها به صورت موازی با هم آموزش می‌یابند. مراحل مربوط به فرآیند آموزش ادغام شده به صورت زیر می‌باشد [۲۹]:

۱. در ابتدا به انباره‌های مربوط به پارامترهای مدل‌های مخفی مارکوف مقدار صفر اختصاص داده می‌شود.
۲. جملهٔ مربوط به آموزش به همراه ساختار متنی آن دریافت می‌شود.
۳. در این مرحله مدل‌های مخفی مارکوف مربوط به هر یک از آواها، با توجه به تلفظ مربوط به کلمه‌ها و ساختار متنی جملهٔ آموزشی، به یکدیگر متصل می‌شوند تا یک مدل مخفی مارکوف مرکب ایجاد شود.
۴. احتمال‌های پیش‌رو و پس‌رو مربوط به مدل مخفی مارکوف مرکب محاسبه می‌شود.
۵. بوسیلهٔ احتمال‌های پیش‌رو و پس‌رو پارامترهای مربوط به مدل‌ها محاسبه و موجودی انباره‌ها با این پارامترهای جدید جایگزین می‌شود.
۶. برای دریافت جمله آموزشی بعدی دوباره از مرحلهٔ ۲ شروع می‌شود. این فرآیند تا جایی ادامه پیدا می‌کند که پردازش بر روی تمام جمله‌های آموزشی انجام شود.

---

[117] Embedded Training

۷. از روی پارامترهای موجود در انباره پارامترهای مربوط به تمام مدل‌های مخفی مارکوف بدست می‌آید.

این مراحل را می‌توان هر چند بار که لازم است انجام داد تا همگرایی مورد نظر بدست آید. باید توجه داشت که هر چند در این روش احتیاجی به مرز هر کدام از آواها در فایل صوتی نیست ولی به ساختار آوایی مربوط به تمام کلمه‌ها احتیاج است [۲۹].

۳-۳-۱۰-۱ بستهٔ نرم افزاری مدل مخفی مارکوف HTK

به طور کلی HTK یک بستهٔ نرم افزاری برای ایجاد مدل‌های مخفی مارکوف است. هدف اصلی ساخت HTK تحقیق در زمینهٔ بازشناخت گفتار بوده است ولی از آن در کاربردهای دیگری مثل تشخیص کاراکتر و ترتیب دهی DNA[۱۱۸] نیز استفاده شده است. امروزه HTK کاربرد گسترده‌ایی در سیستم‌های بازشناخت گفتار اتوماتیک پیدا کرده است.

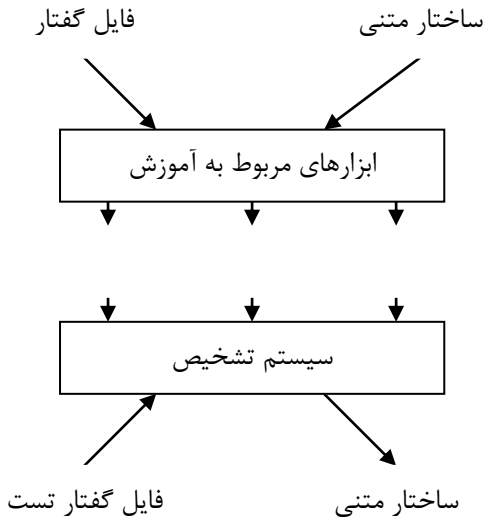
اولین نگارش HTK در سال ۱۹۸۹ توسط Steve Young در دانشگاه کمبریج ایجاد شد. از آن زمان تاکنون تغییرات زیادی در این بستهٔ نرم افزاری به وجود آمده و قابلیت‌های زیادی به آن افزوده شده و بسیاری از اشکالات موجود در آن برطرف شده است. HTK شامل یک مجموعه از کتابخانهٔ ماژولها و ابزارها می‌باشد که کدهای آن به زبان C موجود می‌باشد. از این ابزارها برای آموزش و تست مدل‌های مخفی مارکوف و آنالیز نتایج آنها استفاده می‌شود. همچنین بعضی از ابزارهای مورد نیاز برای پردازش صوت نیز در طول زمان به آن اضافه شده است.

همانطور که در شکل ۳-۱۰ نشان داده شده است به طور کلی دو مرحلهٔ اصلی پردازش در یک سیستم بازشناخت گفتار وجود دارد و ابزارهای مورد نیاز برای هر یک از این دو مرحله در HTK قرار

---

118 DNA Sequencing

داده شده است. در ابتدا ابزارهای مربوط به آموزش موجود در HTK با استفاده از فایل‌های گفتار به همراه فایل‌های متنی مربوط به آنها، پارامترهای مربوط به مدل‌های مخفی مارکوف را تخمین می‌زنند. سپس در مرحلهٔ بازشناخت، فایل گفتار ورودی به وسیلهٔ ابزارهای مربوط به بازشناخت موجود در HTK به فایل نوشتاری تبدیل می‌شود. از HTK می‌توان هم در بازشناخت گفتار پیوسته و هم گسسته استفاده کرد. HTK شامل تعدادی از ابزارها است که قادر به انجام بیشتر عملیاتهای لازم در بازشناخت گفتار مثل استخراج ویژگی، آموزش معمولی و ادغام شده به وسیله الگوریتم بام ولچ، رمزگشایی به وسیلهٔ الگوریتم ویتربی و همچنین آنالیز نتایج هستند.

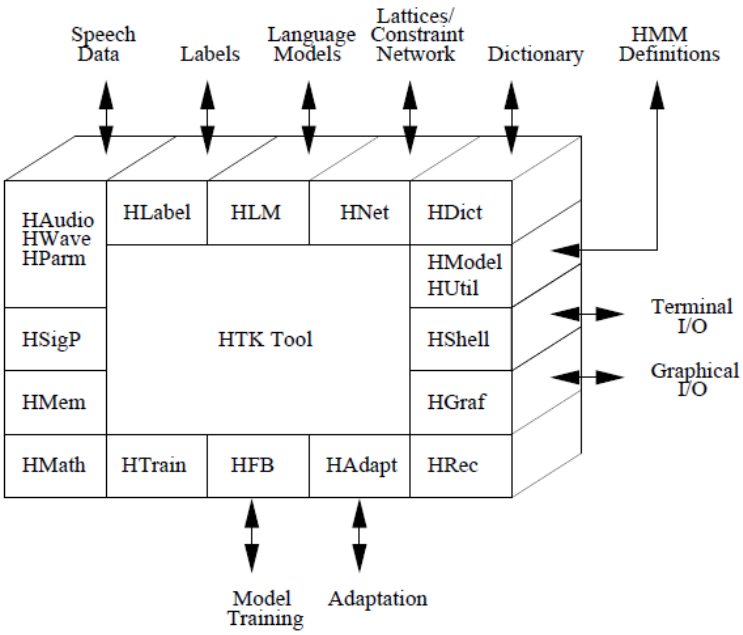


شکل ۳- ۱۰ دو مرحلهٔ اصلی موجود در HTK [۲۹].

۳-۳-۱ ساختار کلی HTK

شکل ۳-۱۱ ساختار نرم افزاری HTK و رابط‌های ورودی و خروجی آن را نشان می‌دهد. همانطور که در این شکل قابل مشاهده است ورودی و خروجی و ارتباط با سیستم عامل توسط ماژول کتابخانه‌ایی HShell کنترل می‌شود و تمام تنظیمات مربوط به حافظه توسط HMem کنترل می‌شود. عملیات‌های ریاضی به وسیلهٔ HMath پشتیبانی می‌شود و عملیات‌های مربوط به پردازش سیگنال

مورد نیاز برای آنالیز صوت درون HSigP قرار داده شده است. هر نوع فایل مورد نیاز برای HTK ماژول رابط مربوط به خود را دارد. HLabel رابط مورد نیاز برای فایلهای برچسب را فراهم می‌کند، HLM برای فایلهای مدل زبان، HNet برای فایلهای شبکه، HDict برای دیکشنریها و HModel برای تعریف مدل‌های مخفی مارکوف است [۲۹].



شکل ۳- ۱۱ ساختار نرم افزاری HTK [۲۹].

تمام ورودی و خروجی‌های در سطح شکل موج از طریق HWave و در سطح پارامتری از طریق HParm انجام می‌شود. ورودی مستقیم صوتی از طریق HAudio پشتیبانی می‌شود و رابط گرافیکی ساده توسط HGraf فراهم می‌شود. HTrain و HFB از ابزارهای مختلف موجود در HTK مربوط به آموزش مدلهای مخفی مارکوف پشتیبانی می‌کنند. HAdapt از ابزارهای مختلف مربوط به تطبیق دادن موجود در HTK پشتیبانی می‌کند. و در نهایت HRec شامل تابع‌های پردازش اصلی مربوط به بازشناخت است [۲۹].

ابزارهای HTK طوری طراحی شده‌اند که باید آنها را از طریق خط فرمان اجرا کرد. هر ابزار دارای یک سری آرگومانهای اختیاری و یک سری آرگومانهای اجباری است، آرگومانهای اختیاری با علامت منفی قبل از آنها مشخص می‌شوند. مقدار هر آرگومان بلافاصله بعد از آن قرار می‌گیرد. به عنوان مثال فرمان زیر باعث اجرای ابزار فرضی HFoo می‌شود [۲۹].

```
HFoo -T 1 -f 34.3 -a -s myfile file1 file2
```

در این فرمان آرگومانهای T، f، a، و s اختیاری می‌باشند. به عنوان مثال مقدار T عدد ۱ است و مقدار s myfile می‌باشد.

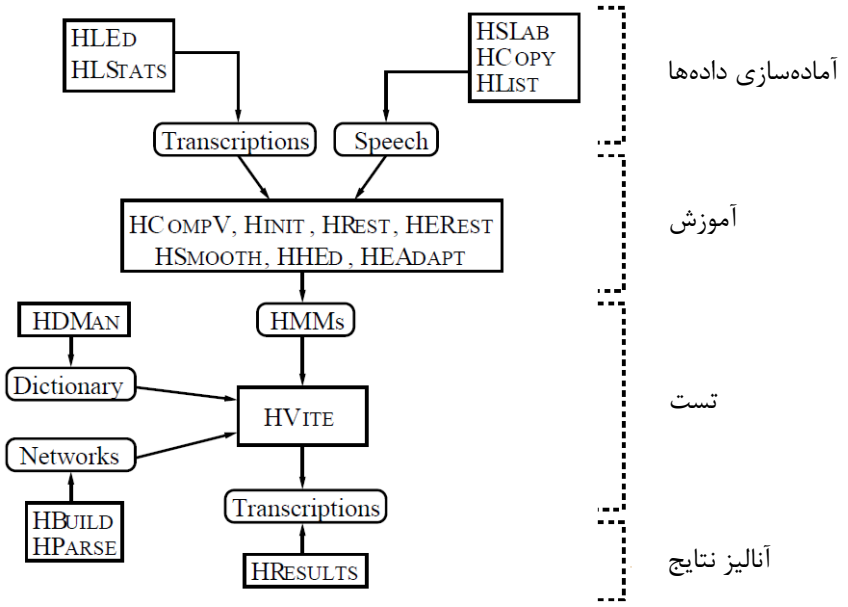
علاوه بر آرگومانهای اختیاری که در خط فرمان قرار داده می‌شوند برای کنترل عملکرد ابزار مورد نظر می‌توان پارامترهای مورد نظر را در فایل پیکربندی[۱۱۹] قرار داد. نام این فایل بلافاصله بعد از آرگومان C در خط فرمان قرار داده می‌شود. به عنوان نمونه در مثال زیر پارامترهای مربوط به پیکربندی در فایل config قرار داده شده‌اند.

```
HFoo -C config -f 34.3 -a -s myfile file1 file2
```

به طور کلی ۴ فاز کلی در HTK برای ساخت یک سیستم بازشناخت گفتار وجود دارد. این ۴ فاز عبارتند از: آماده‌سازی داده‌ها[۱۲۰]، آموزش، تست و آنالیز نتایج. در ادامه هر یک از این قسمت‌ها و بعضی از ابزارهای مورد استفاده در هر قسمت به طور خلاصه شرح داده می‌شود. لیست ابزارهای موجود در HTK برای ساخت یک سیستم بازشناخت گفتار در شکل ۳-۱۲ نشان داده شده است. در مرجع [۲۹] وظیفهٔ هر یک از این ابزارها به طور کامل توضیح داده شده است.

---

119 Configuration file
120 Data Preparation



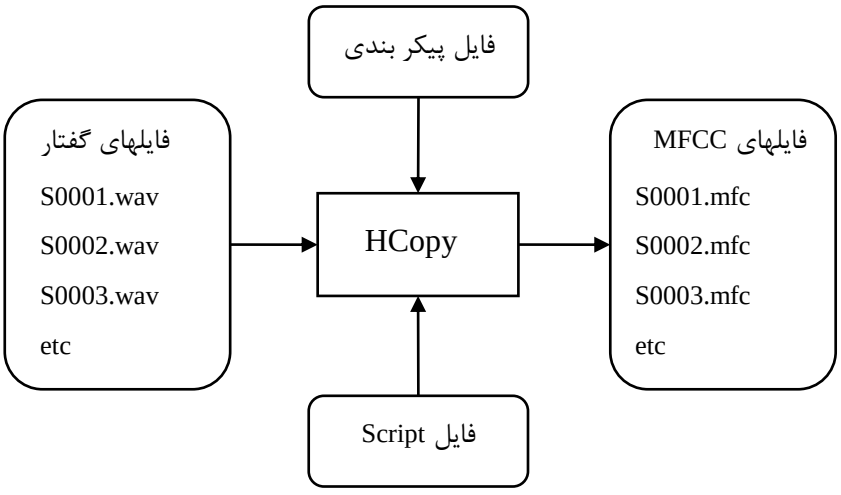
شکل ۳- ۱۲ ابزارهای مختلف موجود در HTK در مراحل مختلف پردازش [۲۹].

۳-۳-۱-۱ ابزارهای مربوط به آماده‌سازی داده‌ها

برای بدست آوردن مدل صوتی مناسب برای هر یک از اجزاء تشکیل دهندهٔ سیگنال گفتار احتیاج به تعداد زیادی فایلهای صوتی و ساختار متنی آنها برای آموزش است. فایلهای صوتی را می‌توان از دیتابیس‌های موجود بدست آورد. در صورتی که دیتابیس در دسترس نباشد به کمک ابزار HSlab و یک میکرفن می‌توان اقدام به تهیهٔ فایل‌های صوتی کرد. در صورت نیاز به کمک این ابزار می‌توان به صورت دستی سیگنال گفتار را نیز برچسب زد [۲۹].

مرحلهٔ بعد در آماده‌سازی داده‌ها استخراج ویژگی از فایل‌های صوتی است. برای این منظور ابزار HCopy در HTK در نظر گرفته شده است. این ابزار امکان استخراج ویژگی به وسیلهٔ چند روش معروف مثل MFCC و PLP را در اختیار ما قرار می‌دهد. با تغییر پارامترهای موجود در فایل پیکربندی این ابزار می‌توان ویژگی‌های مورد نظر را از فایل‌های گفتار استخراج کرد. بلوک دیاگرام مربوط به ورودی‌ها و خروجی ابزار HCopy در شکل ۳-۱۳ نشان داده شده است. همانطور که در این

شکل قابل مشاهده است ابزار HCopy احتیاج به سه ورودی دارد که عبارتند از فایل‌های گفتار، فایل پیکربندی و فایل script. فایل پیکربندی اطلاعات مورد نیاز در مورد نوع روش استخراج ویژگی و پارامترهای به کار رفته در آن را فراهم می‌کند و فایل script حاوی آدرس محل فایل‌های گفتار و آدرس محلی است که ویژگی‌های استخراج شده باید در آنجا ذخیره شوند. بعد از استخراج ویژگی می‌توان به کمک ابزار HList ویژگی‌های استخراج شده را مشاهده و از صحت آنها اطمینان حاصل کرد [۲۹].



شکل ۳- ۱۳ بلوک دیاگرام مربوط به ورودی‌ها و خروجی ابزار HCopy [۲۹].

برای آموزش مدل‌های مخفی مارکوف احتیاج به ساختار کلمه‌ایی و آوایی فایل‌های گفتار، یا به اصطلاح فایل‌های برچسب، نیز است. برای این منظور ابزار HLEd در نظر گرفته شده است که از آن می‌توان برای ویرایش فایل‌های برچسب استفاده کرد.

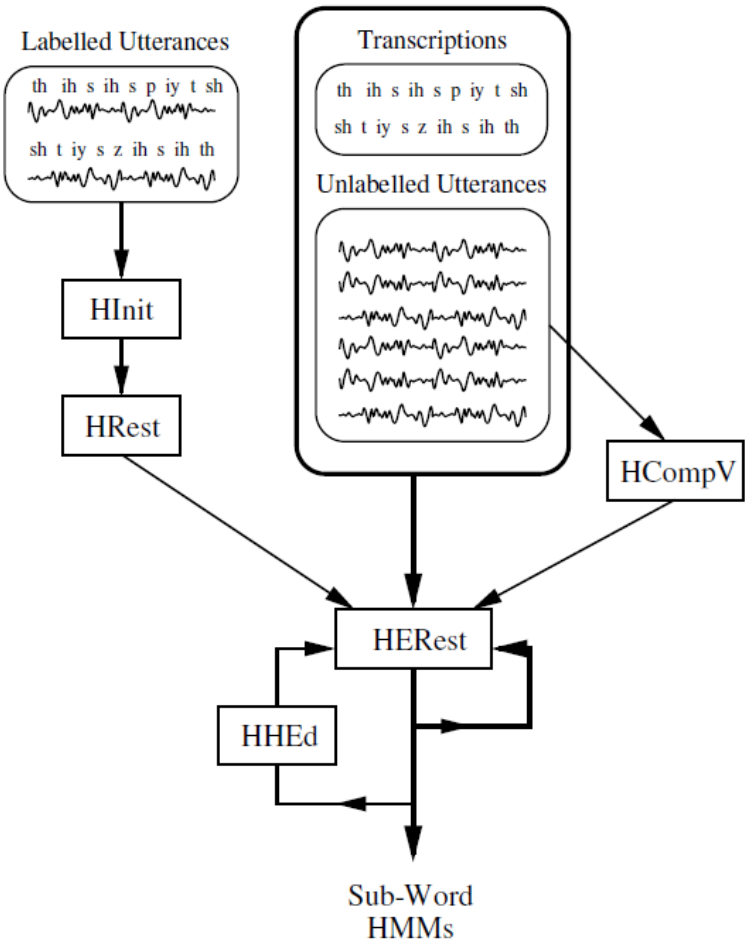
۳-۳-۱-۲ ابزارهای مربوط به آموزش

در این قسمت باید در ابتدا ساختار اولیه مدل‌های مخفی مارکوف به صورت متنی با یکی از ابزارهای ویرایش متن مثل WordPad ایجاد شود و پارامترهای مربوط به مدل‌ها با یک مقدار اولیه مقدار دهی

شود. این مدل اولیه خصوصیات و توپولوژی کلی مدل‌های مخفی مارکوف را مشخص می‌کند. مقدارهای اولیه به کار رفته در این مدل اولیه مهم نیست چون در مراحل بعد مقدارهای اصلی به کمک داده‌های آموزشی بدست می‌آیند و جایگزین این مقدارها می‌شوند. معمولا در مرحلهٔ اولیه مقدار تمام پارامترها را یکسان در نظر می‌گیرند به غیر از پارامترهای مربوط به ماتریس انتقال که دارای فرم به خصوصی است (جمع هر سطر باید برابر با یک شود به غیر از سطر آخر که همگی صفر هستند) [۲۹].

مراحل مختلف مربوط به فرآیند آموزش در شکل ۳-۱۴ نشان داده شده است. همانطور که در این شکل دیده می‌شود دو حالت برای آموزش مدل‌های مخفی مارکوف وجود دارد. حالت اول زمانی است که داده‌های آموزشی مرزبندی شده باشند و محل آواها در جمله مشخص شده باشد. در این حالت در ابتدا به کمک ابزار HInit یک مقدار اولیهٔ جدید برای مدل‌های مخفی مارکوف بدست می‌آید. سپس به کمک ابزار HRest مدل‌های مخفی مارکوف به صورت جداگانه آموزش داده می‌شوند. سپس ابزار HERest مرحلهٔ اصلی آموزش را انجام می‌دهد. ابزار HERest از روش آموزش ادغام شده که در بخش‌های قبل به آن اشاره شد استفاده می‌کند. در حالت کلی که داده‌های آموزشی مرزبندی شده در دسترس نیست در ابتدا بوسیلهٔ ابزار HCompV مقدار اولیهٔ جدید برای مدل‌های مخفی مارکوف بدست می‌آید. سپس از ابزار HERest برای آموزش مدل‌ها استفاده می‌شود [۲۹].

ابزار HHEd برای ویرایش مدل‌های مخفی مارکوف ایجاد شده است. به عنوان مثال از این ابزار می‌توان برای تبدیل مدل‌های تک آوایی به سه آوایی استفاده کرد و یا اینکه تعداد تابع‌های گوسی به کار رفته در تابع خروجی را افزایش داد. باید توجه داشت که بعد از ویرایش مدل‌های مخفی مارکوف باید دوباره مدل‌های مخفی مارکوف جدید را به وسیلهٔ ابزار HERest آموزش داد [۲۹].



شکل ۳- ۱۴ مراحل مختلف مربوط به آموزش مدل‌های مخفی مارکوف [۲۹].

۳-۳-۱-۳ ابزارهای مربوط به بازشناخت

اصلی‌ترین ابزاری که در مرحلهٔ بازشناخت استفاده می‌شود ابزار HVite است. این ابزار برای بازشناخت احتیاج به فایل دیکشنری، مدل‌های مخفی مارکوف آموزش دیده، فایل‌های تست، توصیف شبکه یا در صورت وجود مدل زبان دارد. این ابزار قادر است عملیات بازشناخت را هم بر روی فایلهای صوتی ذخیره شده و هم بر روی فایل‌هایی که به صورت زنده از طریق میکرفن بدست می‌آیند انجام دهد [۲۹].

۳-۳-۱-۴ ابزار آنالیز

برای بررسی کارآیی سیستم بازشناخت ساخته شده بوسیلهٔ HTK ابزاری به نام HResults ایجاد شده است که نتایج خروجی بدست آمده از ابزار HVite را با مقدار واقعی آنها مقایسه و نرخ بازشناخت را در اختیار ما قرار می‌دهد.

۳-۳-۲ انواع فایل‌های مورد استفاده در HTK

به غیر از فایل‌های صوتی که برای آموزش و تست وجود دارد HTK کاملا با فایل‌های متنی سروکار دارد. فایل‌های متنی مورد استفاده در HTK دارای پسوند مخصوص به خود هستند ولی به خاطر ساختار متنی که دارند می‌توان آنها را در محیط‌های ویرایش متن مثل Notepad یا WordPad باز و ویرایش کرد [۷۰]. توضیحات کاملی در مورد فایل‌های مورد استفاده در HTK در مرجع [۲۹] آمده است ولی در اینجا به طور خلاصه به بعضی از آنها اشاره می‌شود.

۳-۳-۲-۱ فایل‌های برچسب

فایل‌های برچسب محتوای متنی فایل‌های گفتار مورد استفاده در مرحلهٔ آموزش و تست را در خود نگه می‌دارند. فایل‌های برچسب می‌توانند در سطح کلمه یا آوا باشند. به عنوان مثال در بازشناخت گفتار پیوسته که احتیاج به ساختار آوایی جمله‌ها است فایل‌های برچسب با ساختار آوایی مورد نیاز است.

در فایل‌های برچسب با ساختار آوایی یا کلمه‌ایی هر آوا یا کلمه باید در یک خط قرار بگیرد به عنوان مثال فایل برچسب مربوط به ساختار آوایی جملهٔ زیر با نام */sample1* بعد از آن قرار داده

شده است. در این ساختار آوایی sil نشان دهندهٔ مکث بلند اتفاق افتاده در ابتدا و انتهای جمله است. در بازشناخت گفتار پیوسته یک مدل مخفی مارکوف نیز برای sil ایجاد می‌شود.

*/sample1 DIAL ONE.

"*/sample1.lab"

sil

d

ay

l

w

ah

n

sil

ابزار HLed در HTK برای ویرایش فایل‌های برچسب ایجاد شده است. به کمک این ابزار می‌توان بعضی از برچسب‌ها را حذف، اضافه یا جابجا کرد.

۳-۳-۲-۲ فایل دیکشنری

برای ایجاد فایل برچسب ایجاد شده در قسمت قبل احتیاج به تلفظ یا ساختار آوایی تمام کلمه‌های مورد استفاده در مرحلهٔ آموزش و تست است. برای این منظور ساختار آوایی تمام کلمه‌ها در فایل دیکشنری قرار می‌گیرد. باید توجه داشت که در فایل دیکشنری باید تمام کلمه‌ها بر حسب حروف الفبا مرتب شده باشند و هر کلمه به همراه ساختار آوایی آن در یک خط قرار گرفته باشد. قسمتی از یک فایل دیکشنری به صورت زیر است.

ABALON ae b ax l aa n sp

ABDOMINALS ae b d aa m ih n ax l z sp

ABOLISH ax b aa l ih sh sp

sp نشان دهندهٔ مکث کوتاه بین دو کلمه است. در بازشناخت گفتار پیوسته علاوه بر ایجاد مدل مخفی مارکوف برای هر آوا به کمک مدل ایجاد شده برای sil یک مدل مخفی مارکوف نیز برای sp ایجاد می‌شود.

ابزار HDMan برای کار با فایل‌های دیکشنری در HTK ایجاد شده است. به کمک این ابزار می‌توان چند فایل دیکشنری را یکی کرد و یا از روی یک فایل دیکشنری با ساختار تک آوایی می‌توان فایل دیکشنری با ساختار سه آوایی را بدست آورد.

۳-۳-۲-۳ فایل‌های مربوط به تعریف مدل‌های مخفی مارکوف

هدف اصلی HTK ایجاد مجموعه‌ایی از مدل‌های مخفی مارکوف است. هر یک از این مدل‌ها باید طوری تعریف شوند که ساختار آنها، مقدار پارامترهای ماتریس انتقال و پارامترهای مربوط به تابع توزیع خروجی به خوبی معین شوند [۲۹].

برای تعریف مدل‌های مخفی مارکوف، HTK از یک فرمت مخصوص به خود استفاده می‌کند. قبل از اینکه بتوان از ابزارهای مربوط به آموزش و تست استفاده کرد باید تمام مدل‌ها، به صورت دستی و به کمک برنامه‌های ویرایش متن مثل WordPad، با این زبان مخصوص تعریف شوند و پارامترها با مقدارهای اولیه مقدار دهی شوند [۷۰].

۳-۳-۲-۴ فایل‌های پیکربندی

فایل‌های پیکربندی حاوی پارامترهایی هستند که به کمک آنها می‌توان کار ابزارهای مختلف را کنترل کرد. به عنوان مثال در هنگام استفاده از ابزار HCopy که برای استخراج ویژگی از داده‌های صوتی استفاده می‌شود به کمک فایل پیکربندی می‌توان فرمت داده‌های ورودی، روش مورد نظر برای استخراج ویژگی، طول پنجرهٔ مورد نظر برای استخراج ویژگی، تعداد فیلترها در بانک فیلتر و غیره را مشخص کرد.

فصل چهارم

مروری بر کارهای انجام شده در حوزهٔ اتوکرولیشن برای استخراج ویژگی

فصل چهارم: مروری بر کارهای انجام شده در حوزهٔ اتوکرولیشن برای استخراج ویژگی

۱-۴۱۱-۱ مقدمه

همانطور که قبلاً گفته شد یکی از مهمترین مراحل موجود در سیستم‌های بازشناخت گفتار مرحلهٔ استخراج ویژگی است. سیگنال گفتار زمانی که وارد سیستم بازشناخت می‌شود حاوی اطلاعات متفاوتی از منابع گوناگون است که خیلی از این اطلاعات بی ارزش است و حاوی اطلاعات مهمی برای بازشناخت گفتار نیستند. هدف مرحلهٔ استخراج ویژگی که اولین مرحلهٔ حیاتی در سیستم‌های بازشناخت گفتار است تبدیل سیگنال گفتار به یک سری از ویژگی‌های صوتی است به طوری که این ویژگی‌ها تمام اطلاعات مهم مورد نیاز برای بازشناخت گفتار را در خود داشته باشند.

یک مجموعه ویژگی خوب برای بازشناخت گفتار باید دارای سه معیار جداپذیری، مقاومت و حجم محاسباتی کم باشد. معیار جداپذیری نشان دهندهٔ آن است که مجموعهٔ ویژگی مورد نظر باید

این قابلیت را داشته باشد که بتواند به خوبی اجزاء مختلف صوتی مثل آواها را از هم جدا کند. معیار مقاومت مبین آن است که مجموعهٔ ویژگی استخراج شده باید تا آنجا که ممکن است نسبت به تغییرات ناشی از نویز جمع شونده و اعوجاج مربوط به کانال انتقال مقاوم باشد. از آنجایی که بیشتر سیستم‌های بازشناخت گفتار باید به صورت زنده مورد استفاده قرار گیرند، الگوریتم‌هایی که برای استخراج ویژگی‌ها در آنها مورد استفاده قرار می‌گیرد باید تا آنجا که امکان دارد بهینه باشند به عبارت دیگر پیچیدگی محاسباتی آنها کم باشد [۷۱].

تحقیقات بسیاری در زمینهٔ استخراج ویژگی انجام شده و روش‌های متفاوتی پیشنهاد شده است که از میان آنها روش MFCC [۴۵] به عنوان یک روش استاندارد شناخته شده است و در اکثر سیستم‌های بازشناخت گفتار امروزی از آن استفاده می‌شود.

مشکلی که روش MFCC و بسیاری از روش‌های دیگر استخراج ویژگی با آن روبرو هستند عدم مقاومت آنها نسبت به نویز است. به عبارت دیگر در عمل نویز جمع شونده و اعوجاج ناشی از کانال انتقال باعث آلوده شدن سیگنال گفتار و کاهش شدید دقت بازشناخت می‌شوند. همانطور که قبلاً در فصل دوم اشاره گردید روش‌هایی که برای حل این مشکل ارائه شده است شامل سه دستهٔ کلی روش‌های استخراج ویژگی مقاوم، روش‌های بهبود گفتار و روش‌های بر پایهٔ مقاوم سازی مدل بازشناخت می‌باشند. روش‌هایی که در این فصل و در فصل بعد مورد بررسی قرار می‌گیرند جزء دستهٔ روش‌های استخراج ویژگی مقاوم هستند.

اخیراً بعضی از روش‌های موفقیت آمیز در بازشناخت گفتار مقاوم از حوزهٔ اتوکرولیشن برای بدست آوردن ویژگی‌های مقاوم استفاده کرده‌اند. از جملهٔ این روش‌ها می‌توان به وابستگی بهبود یافته زمان کوتاه (SMC) [۳۵]، اتوکرولیشن یک طرفه LPC (OSALPC) [۵۸، ۷۲]، رشتهٔ اتوکرولیشن

نسبی (RAS) [۵۹، ۷۳]، مشتق رشتهٔ اتوکرولیشن (DAS[۱۲۱]) [۷۴]، اتوکرولیشن MFCC (AMFCC[۱۲۲]) [۷۵، ۷۶] و اتوکرولیشن فاز (PAC[۱۲۳]) [۷۷] اشاره کرد.

در این فصل در ابتدا بعضی از خواص تابع اتوکرولیشن، که آن را به حوزه‌ایی مناسب برای استخراج ویژگی تبدیل کرده است، معرفی می‌شود. سپس به مرور کلی چند روش که از حوزهٔ اتوکرولیشن برای استخراج ویژگی مقاوم استفاده کرده‌اند پرداخته می‌شود.

۱-۱۲ ۴-۲ بعضی از خواص تابع اتوکرولیشن

تابع اتوکرولیشن مربوط به یک سیگنال نمایانگر میزان شباهت سیگنال و نمونه‌های انتقال یافتهٔ آن در طول زمان است. از این تعریف می‌توان نتیجه گرفت که تابع اتوکرولیشن مربوط به سیگنال‌های متناوب با همان فرکانس، متناوب هستند [۷۸].

مطلب مهمی که در مورد تابع اتوکرولیشن وجود دارد توانایی آن در تخمین طیف توان سیگنال است به عبارت دیگر تابع اتوکرولیشن یک شکل موج حاوی اطلاعات یکسانی نسبت به طیف توان آن شکل موج است. هر چند این اطلاعات به صورت تابعی از زمان است نه فرکانس [۷۸]. برای بدست آوردن طیف توان یک سیگنال کافی است از تابع اتوکرولیشن آن تبدیل فوریه گرفته شود [۷۹].

دو خاصیت مهم تابع اتوکرولیشن عبارتند از:

۱. خاصیت جمع‌پذیری: در صورتی که دو سیگنال ناهمبسته باشند تابع اتوکرولیشن مجموع آنها برابر است با مجموع تابع اتوکرولیشن هر یک از آنها.

---

121 Differentiated Autocorrelation Sequence
122 Autocorrelation Mel Frequency Cepstral Coefficient
123 Phase AutoCorrelation

۲. خاصیت مقاومت: تابع اتوکرولیشن مربوط به سیگنال نویز تصادفی سفید در هر جایی به غیر از تأخیر زمانی صفر، صفر است. علاوه بر این تابع اتوکرولیشن مربوط به سیگنال‌های نویز پهن‌باند در تأخیرهای زمانی کوچک و صفر دارای مقدارهای بزرگ و در تأخیرهای زمانی بزرگ دارای مقدارهای کوچک و نزدیک صفر هستند.

با توجه به این دو خاصیت، سیگنال نویز جمع شونده تأثیر کمی بر روی ضرایب تابع اتوکرولیشن مربوط به تأخیرهای زمانی بزرگ می‌گذارد [۷۶].

علاوه بر این دو خاصیت یک خاصیت مهم دیگر حوزهٔ اتوکرولیشن خاصیت حفظ قطب آن است. به عبارت دیگر خواص طیفی سیگنال گفتار بعد از انتقال آن به حوزهٔ اتوکرولیشن از بین نمی‌رود [۷۴].

۱–۱۳ ۴–۳ روش رشتهٔ خود همبستگی نسبی (RAS)

RAS یکی از روش‌های استخراج ویژگی مقاوم است که برای کم کردن اثر نویز جمع شونده معرفی شده است. روش RAS این کار را با اعمال یک فیلتر بالا گذر در حوزهٔ اتوکرولیشن انجام می‌دهد. به عبارت دیگر ایدهٔ اصلی روش RAS برای از بین بردن نویز موجود در سیگنال، فیلتر کردن خط سیر زمانی[۱۲۴] مربوط به رشتهٔ اتوکرولیشن یکطرفهٔ سیگنال گفتار است. سپس ضرایب کپسترال از رشتهٔ جدید بدست آمده استخراج می‌شوند.

---

124 Temporal Trajectories

۴-۳-۱ فیلتر کردن رشتهٔ اتوکرولیشن

در صورتی که سیگنال گفتار نویزی با $x(m,n)$ نشان داده شود، داریم:

$$x(m,n) = s(m,n) + d(m,n) \quad 0 \le m \le M-1,\ \ 0 \le n \le N-1, \qquad (۱-۴)$$

که در این رابطه M تعداد فریم‌ها، N طول هر فریم، $s(m,n)$ سیگنال گفتار بدون نویز و $d(m,n)$ نویز جمع شونده است.

در صورتی که نویز و سیگنال ناهمبسته باشند با توجه به خاصیت اول تابع اتوکرولیشن داریم:

$$r_{xx}(m,i) = r_{ss}(m,i) + r_{dd}(m,i) \quad 0 \le m \le M-1,\ \ 0 \le i \le N-1, \qquad (۲-۴)$$

که در این رابطه $r_{xx}(m,i)$، $r_{ss}(m,i)$ و $r_{dd}(m,i)$ به ترتیب رشتهٔ اتوکرولیشن یک طرفهٔ مربوط به سیگنال نویزی، سیگنال بدون نویز و نویز جمع شونده هستند. در رابطه ۴-۲ برای بدست آوردن رشته اتوکرولیشن از رابطهٔ بایاس نشده، به صورت زیر، استفاده شده است:

$$r_{xx}(m,i) = \frac{1}{N-i} \sum_{j=0}^{N-1-i} x(m,j)\, x(m,j+i), \quad 0 \le i \le N-1. \qquad (۳-۴)$$

در صورتی که سیگنال نویز ایستا باشد می‌توان فرض کرد که رشتهٔ اتوکرولیشن مربوط به نویز در تمام فریم‌های سیگنال گفتار یکسان است و در نتیجه $r_{dd}(m,i)$ فقط به شاخص i مربوط به اتوکرولیشن بستگی دارد، یعنی:

$$r_{xx}(m,i) = r_{ss}(m,i) + r_{dd}(i) \quad 0 \le m \le M-1,\ \ 0 \le i \le N-1, \qquad (۴-۴)$$

با مشتق‌گیری از دو طرف رابطهٔ بالا نسبت به شاخص فریم، m برای تمام i ها داریم:

$$\frac{\partial r_{xx}(m,i)}{\partial m} = \frac{\partial r_{ss}(m,i)}{\partial m} \quad 0 \le m \le M-1,\ \ 0 \le i \le N-1. \qquad (۵-۴)$$

رشتهٔ $\{\partial r_{xx}(m,i)/\partial m\}_{i=0}^{N-1}$ معرف رشتهٔ RAS مربوط به فریم m ام سیگنال نویزی است. با توجه به رابطهٔ ۴-۵ می‌توان دید که رشتهٔ RAS سیگنال نویزی برابر است با رشتهٔ RAS سیگنال بدون نویز، بنابراین اثر نویز در آن حذف شده است [۷۳].

برای محاسبهٔ مشتق موجود در رابطهٔ ۴-۵ می‌توان از تخمین زیر استفاده کرد

$$\frac{\partial r_{xx}(m,i)}{\partial m} \cong \frac{1}{T_L}\sum_{t=-L}^{L} t r_{xx}(m+t,i) \qquad (۶-۴)$$
$$0 \le m \le M-1, \quad 0 \le i \le N-1,$$

که در این رابطه

$$T_L = \sum_{t=-L}^{L} t^2 \qquad (۷-۴)$$

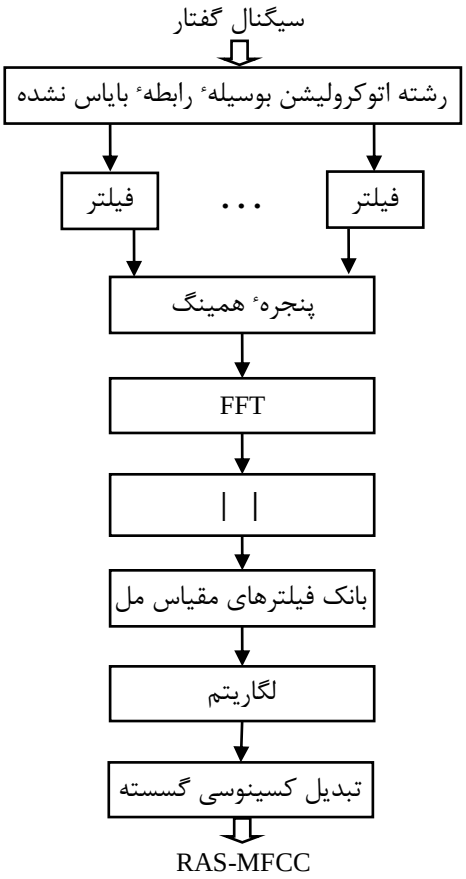
عملیات انجام شده در رابطهٔ ۴-۶ معادل است با فرآیند فیلتر کردن به وسیلهٔ یک فیلتر FIR بالاگذر با تابع انتقال زیر:

$$H(z) = \frac{1}{T_L}\sum_{t=-L}^{L} t z^t . \qquad (۸-۴)$$

۴-۳-۲ محاسبهٔ ضرایب کپسترال از رشتهٔ RAS

بلوک دیاگرام مربوط به محاسبهٔ ضرایب کپسترال فرکانس مل از رشتهٔ RAS (RAS-MFCC) در شکل ۴-۱ نشان داده شده است.

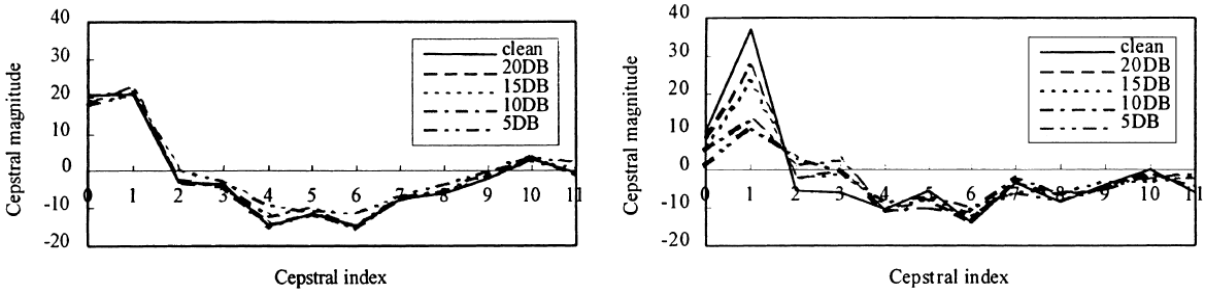
ضرایب MFCC که از رشتهٔ RAS بدست می‌آیند دارای مقاومت بیشتری نسبت به نویز هستند چون همانطور که در رابطهٔ ۴-۵ نشان داده شد اثر نویز جمع شونده در رشتهٔ RAS حذف شده است. شکل ۴-۲-الف و ۴-۲-ب به ترتیب میزان مقاومت روش‌های MFCC و RAS-MFCC را در مقابل نویز سفید جمع شونده با توان‌های مختلف نشان می‌دهد [۷۳].



شکل ۴- ۱ بلوک دیاگرام مربوط به محاسبهٔ ضرایب RAS-MFCC [۷۳].

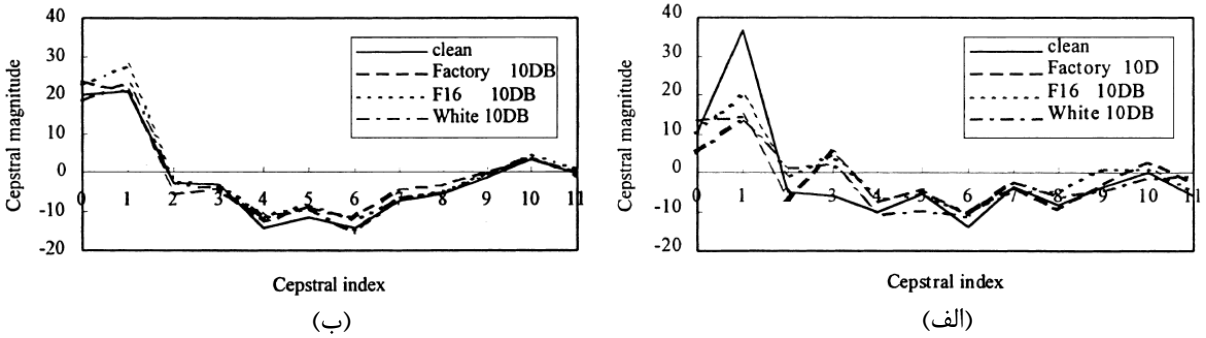
همانطور که در شکل ۴-۲-الف قابل مشاهده است دامنهٔ ضرایب کپسترال بدست آمده با روش MFCC با کاهش میزان سیگنال به نویز به شدت تحت تأثیر نویز سفید قرار می‌گیرد. این تأثیر پذیری در ضرایب کپسترال با اندیس پایین بیشتر نمایان است. این در حالی است که با توجه به شکل ۴-۲-ب ضرایب کپسترال بدست آمده با روش RAS-MFCC دارای مقاومت نسبتاً خوبی در مقابل نویز سفید هستند و در مقایسه با ضرایب کپسترال بدست آمده با روش MFCC، با افزایش میزان سیگنال به نویز، تغییر چندانی در دامنهٔ آنها به وجود نمی‌آید.

برای بررسی میزان مقاومت ضرایب کپسترال بدست آمده با روش RAS-MFCC در مقابل نویز رنگی و مقایسهٔ آن با روش MFCC شکل ۴-۳ اثر نویز رنگی (کارخانه و F16) را بر روی این دو روش نشان می‌دهد [۷۳].



شکل ۴- ۲ مقایسه دو روش MFCC و RAS-MFCC در حضور نویز سفید [۷۳]. (الف) اثر نویز سفید جمع شونده در MFCC؛ (ب) اثر نویز سفید جمع شونده در RAS-MFCC.

همانطور که در شکل ۴-۳-الف نشان داده شده است هم ضرایب کپسترال با اندیس پایین و هم ضرایب کپسترال با اندیس بالا مربوط به روش MFCC تحت تأثیر نویز رنگی جمع شونده تغییر می‌کنند. این در حالی است که شکل ۴-۳-ب مقاومت نسبتاً خوب روش RAS-MFCC را در مقابل نویز رنگی نشان می‌دهد.



شکل ۴- ۳ مقایسه دو روش MFCC و RAS-MFCC در حضور نویز رنگی [۷۳]. (الف) اثر نویز رنگی جمع شونده در MFCC؛ (ب) اثر نویز رنگی جمع شونده در RAS-MFCC.

۱۴–۱ ۴–۴ روش مشتق رشتهٔ اتوکرولیشن (DAS)

یکی دیگر از روش‌هایی که اخیراً برای استخراج ویژگی‌های مقاوم در مقابل نویز جمع شونده در مقالات معرفی شده است روش DAS [۷۴] می‌باشد. در این روش از دو روش استخراج ویژگی مقاوم با نام RAS (که در قسمت قبل توضیح داده شد) و مشتق طیف توان (DPS) [۷۱] استفاده شده است.

همانطور که در روش RAS (در بخش قبل) نشان داده شد فیلتر کردن رشتهٔ اتوکرولیشن باعث حذف اثر نویز جمع شونده می‌شود. همچنین در روش مشتق طیف توان (DPS) نشان داده شده است که، برای اصوات صدادار آلوده شده با نویز، میزان سیگنال به نویز مربوط به مشتق طیف توان بیشتر از سیگنال به نویز مربوط به طیف توان معمولی است و برای اصوات بی‌صدا و سکوت این دو مقدار تقریباً یکسان هستند [۷۱]. روش DAS از مزایای موجود در این دو روش برای بدست آوردن ویژگی‌های مقاوم برای استفاده در بازشناخت گفتار استفاده کرده است.

محاسبهٔ ضرایب DAS شامل سه قسمت کلی می‌باشد که در ادامه هر یک از آنها توضیح داده شده است.

۱–۴–۴ بدست آوردن رشتهٔ RAS

همانطور که در قسمت قبل نیز نشان داده شد برای بدست آوردن رشتهٔ RAS در ابتدا با استفاده از رابطهٔ بایاس شده ۴-۳ رشتهٔ اتوکرولیشن یک طرفه مربوط به سیگنال نویزی بدست آورده می‌شود. با فرض ایستا بودن سیگنال نویز می‌توان رشتهٔ اتوکرولیشن مربوط به آن را در تمام فریم‌های سیگنال

گفتار یکسان فرض کرد و با حذف شاخص مربوط به فریم در سیگنال نویز جمع شونده رابطهٔ ۴-۴ بدست می‌آید.

در نهایت رشتهٔ RAS با مشتق گیری از دو طرف رابطهٔ ۴-۴ بدست می‌آید. برای محاسبهٔ این مشتق از رابطه‌های ۴-۶ و ۴-۷ استفاده شده است.

۴-۴-۲ محاسبهٔ مشتق طیف توان (DPS)

با مشتق‌گیری از طیف توان علاوه بر اینکه پیک‌های موجود در طیف توان که دارای اطلاعات مهمی در بازشناخت گفتار هستند باقی می‌مانند، قسمت هموار طیف توان که معمولاً مربوط به نویز می‌باشد نیز تقریباً برابر با صفر می‌شود [۷۱].

با فرض ناهمبسته بودن سیگنال گفتار و نویز با اعمال تبدیل فوریهٔ گسستهٔ زمان کوتاه به رشتهٔ RAS، طیف توان به صورت زیر بدست می‌آید:

$$X(\omega) = S(\omega) + D(\omega), \qquad (۹-۴)$$

که در این رابطه $X(\omega)$، $S(\omega)$ و $D(\omega)$ به ترتیب طیف توان مربوط به سیگنال گفتار نویزی، سیگنال گفتار بدون نویز و نویز جمع شونده هستند و ω فرکانس زاویه‌ایی است. در صورتی که از دو طرف رابطهٔ بالا نسبت به ω مشتق گرفته شود، داریم:

$$\mathrm{Diff}_X(\omega) = \frac{dX(\omega)}{d\omega} = \frac{dS(\omega)}{d\omega} + \frac{dD(\omega)}{d\omega} = \mathrm{Diff}_S(\omega) + \mathrm{Diff}_D(\omega), \qquad (۱۰-۴)$$

که در این رابطه $\mathrm{Diff}_S(\omega)$ و $\mathrm{Diff}_D(\omega)$ به ترتیب مشتق طیف توان مربوط به سیگنال گفتار بدون نویز و نویز است. تعریف مربوط به مشتق طیف توان داده شده در رابطهٔ ۴-۱۰ مربوط به حوزهٔ فرکانس پیوسته است. معادل گسستهٔ آن به وسیلهٔ رابطهٔ تفاضلی زیر تخمین زده می‌شود:

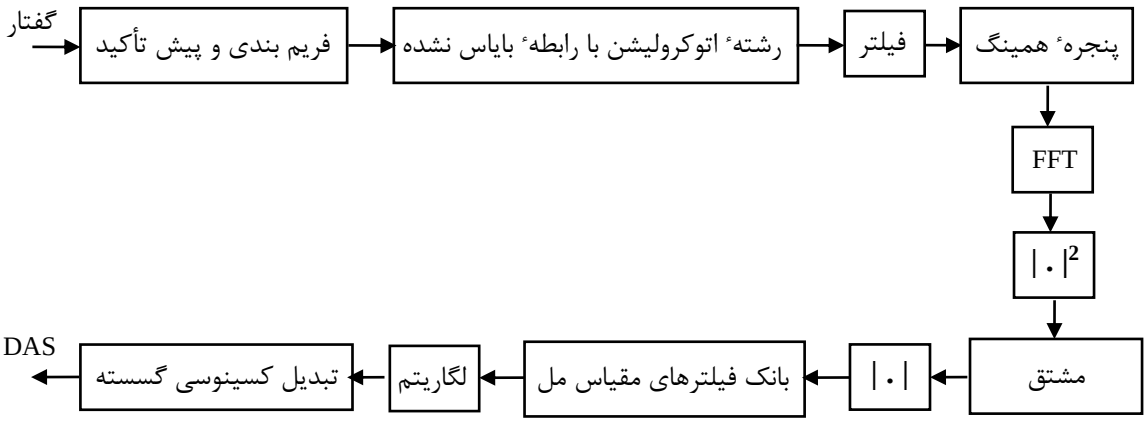
$$\begin{aligned} \mathrm{Diff_X(k)} &= \mathrm{Diff_S(k)} + \mathrm{Diff_D(k)} \approx \sum_{l=-Q}^{P} b_l\, X(k+l) \\ &\approx \sum_{l=-Q}^{P} b_l\, [S(k+l) + D(k+l)]\,, \end{aligned} \qquad (۱۱-۴)$$

که در این رابطه Q و P درجه‌های رابطهٔ تفاضلی، b_l ها ضرایب وزن دهی و $0 \le k \le K$، که K طول FFT است.

۴-۴-۳ بدست آوردن ضرایب کپسترال

بلوک دیاگرام مربوط به محاسبه ضرایب کپسترال با روش DAS در شکل ۴-۴ نشان داده شده است.

همانطور که از روی شکل مشاهده می‌شود بعد از بدست آوردن رشتهٔ RAS و عبور دادن آن از پنجرهٔ همینگ، طیف توان با محاسبهٔ اندازهٔ تبدیل فوریهٔ این رشته بدست می‌آید. سپس فیلتر بانک مقیاس مل به اندازهٔ مشتق این طیف توان اعمال می‌شود. سپس لگاریتم انرژی‌های فیلتر بانک محاسبه می‌شود و در نهایت ویژگی‌های DAS با اعمال تبدیل کسینوسی گسسته بدست می‌آیند.



شکل ۴- ۴ بلوک دیاگرام مربوط به روش DAS برای استخراج ویژگی‌های مقاوم [۷۴].

۱۵–۱ ۵–۴ روش اتوکرولیشن MFCC (AMFCC)

همانطورکه قبلاً گفته شد یکی از پرکاربردترین روش‌های استخراج ویژگی که در بازشناخت گفتار مورد استفاده قرار می‌گیرد روش MFCC است. این روش در صورتی که شرایط آموزش و تست یکسان باشند دارای نتایج نسبتاً خوبی است. ولی وجود نویز به شدت بر روی کارآیی این روش استخراج ویژگی تأثیر می‌گذارد و نرخ بازشناخت گفتار را پایین می‌آورد. یکی از دلایل این امر تأثیر پذیری زیاد طیف توان، مورد استفاده در این روش، از نویز محیط است.

همانطور که در فصل ۲ اشاره شد اولین مرحله در روش MFCC، برای استخراج ویژگی، تخمین طیف توان است که هر چه این طیف دقیق‌تر باشد و فرکانس‌های فرمنت را بهتر در اختیار ما قرار دهد مطلوب‌تر است. در روش MFCC برای تخمین طیف توان از روش پریودگرام[۱۲۵] استفاده می‌شود. از روی طیف توان بدست آمده از این روش فرکانس مربوط به ۳ فرمنت اولیه به خوبی قابل بدست آمدن است ولی متأسفانه در شرایطی که سیگنال با نویز جمع می‌شود طیف توان تخمین زده شده با این روش دچار تغییراتی می‌شود که از کارآیی ویژگی‌های استخراج شده با روش MFCC می‌کاهد [۷۶].

یکی از ایده‌هایی که در مقالات برای حل مشکل MFCC بیان شده است استفاده از تابع اتوکرولیشن و مزایای مهم آن می‌باشد. همانطور که قبلاً ذکر شد (بخش ۴-۲). تابع اتوکرولیشن علاوه بر داشتن قابلیت تخمین طیف توان دارای دو خاصیت مهم دیگر که عبارتند از: خاصیت جمع شوندگی و خاصیت مقاومت، نیز است. در روش AMFCC از این دو خاصیت تابع اتوکرولیشن به همراه قابلیت تخمین طیف توان برای استخراج ویژگی‌های مقاوم استفاده شده است. به عبارت دیگر در روش AMFCC به جای استفاده از روش پریودگرام برای تخمین طیف توان از ضرایب اتوکرولیشن

---

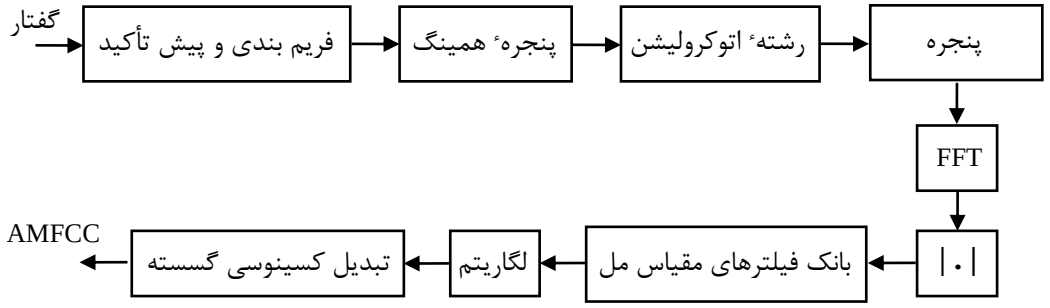
125 Periodogram

با تأخیر زمانی بالا استفاده شده است. در ادامه دلیل استفاده از ضرایب اتوکرولیشن با تأخیر زمانی بالا در این روش به طور مختصر توضیح داده شده است.

با بررسی‌های انجام شده بر روی تابع اتوکرولیشن سیگنال گفتار این نتیجه به دست آمده است که از ضرایب تابع اتوکرولیشن با تأخیرهای زمانی بالا نیز می‌توان برای تخمین طیف توان استفاده کرد. در حالی که قبلاً تصور می‌شد فقط ضرایب اتوکرولیشن با تأخیر زمانی کم حاوی اطلاعات مربوط به طیف توان هستند و ضرایب با تأخیر زمانی بالا بیشتر مربوط به اطلاعات فرکانس پیچ هستند[۸۰].

مزیت استفاده از این ضرایب اتوکرولیشن با تأخیر زمانی بالا، برای محاسبهٔ طیف توان، این است که تأثیر نویز در آنها حداقل است (با توجه به خواص تابع اتوکرولیشن). بنابراین در روش AMFCC در ابتدا ضرایب اتوکرولیشن با تأخیر زمانی کم (کمتر از ۲ میلی ثانیه) حذف می‌شوند و سپس یک پنجرهٔ مناسب به باقیماندهٔ ضرایب اعمال می‌شود. از این رشتهٔ اتوکرولیشن جدید برای بدست آوردن طیف توان و استخراج ویژگی استفاده شده است. در روش AMFCC این روش تخمین طیف توان با نام تخمین طیف توان از روی ضرایب اتوکرولیشن با تأخیر زمانی بالا (HASE[۱۲۶]) نامیده شده است [۷۶].

بلوک دیاگرام مربوط به روش AMFCC برای استخراج ویژگی در شکل ۴-۵ نشان داده شده است.



شکل ۴- ۵ بلوک دیاگرام مربوط به روش AMFCC برای استخراج ویژگی [۷۶].

[126] Higher-lag Autocorrelation Spectrum Estimation

همانطور که در شکل ۴-۵ نشان داده شده است پنجرهٔ همینگ به سیگنال گفتار فریم بندی شده و پیش تأکید شده اعمال می‌شود. سپس رشتهٔ اتوکرولیشن یک طرفه با استفاده از رابطهٔ بایاس شده بدست می‌آید. رابطهٔ بایاس شده برای محاسبهٔ رشتهٔ اتوکرولیشن یکطرفه مربوط به هر فریم در زیر آورده شده است:

$$r_{xx}(i) = \frac{1}{N} \sum_{j=0}^{N-1-i} x(j)\, x(j+i), \quad 0 \le i \le N-1. \qquad (۱۲-۴)$$

که در این رابطه x(j) یک فریم از سیگنال گفتار نویزی است که از پنجرهٔ همینگ عبور کرده است و N طول هر فریم است.

بعد از محاسبهٔ رشتهٔ اتوکرولیشن یکطرفه مربوط به هر فریم، ضرایب اتوکرولیشن با تأخیر زمانی کمتر از ۲ میلی ثانیه حذف می‌شوند و یک پنجرهٔ مناسب به باقیماندهٔ ضرایب اعمال می‌شود (روش بدست آوردن این پنجره به طور کامل در مرجع [۷۶] آورده شده است). سپس از اندازهٔ تبدیل فوریهٔ رشتهٔ اتوکرولیشن بدست آمده به عنوان تخمینی از طیف توان استفاده می‌شود و ضرایب کپستروم مانند روش MFCC از آن استخراج می‌شوند.

۱۶-۱ ۴-۶ روش اتوکرولیشن فاز (PAC)

در بیشتر روش‌های استخراج ویژگی از طیف توان برای بدست آوردن ویژگی‌های سیگنال گفتار استفاده می‌شود. طیف توان مورد نیاز در این روش‌ها معمولاً از طریق محاسبهٔ تبدیل فوریهٔ ضرایب اتوکرولیشن سیگنال گفتار بدست می‌آید. متأسفانه این روش‌ها حساسیت زیادی نسبت به نویز دارند و در حضور نویز کارآیی خود را از دست می‌دهند. یکی از دلایل این امر حساسیت ضرایب اتوکرولیشن،

که برای بدست آوردن طیف توان مورد استفاده قرار گرفته، نسبت به نویز است. یکی دیگر از روش‌هایی که برای بهبود این مشکل مطرح شده است، روش PAC [۷۷] است.

در روش معمولی برای بدست آوردن ضرایب اتوکرولیشن از ضرب نقطه‌ایی بین بردارهای یک سیگنال که با یک زمان مشخص نسبت به هم اختلاف دارند استفاده می‌شود. در روش PAC به جای استفاده از ضرب نقطه‌ایی برای بدست آوردن میزان شباهت بین دو سیگنال از زاویهٔ بین آن دو سیگنال به عنوان معیاری از شباهت استفاده شده است. دلیل این کار هم این بوده است که زاویه نسبت به ضرب نقطه‌ایی حساسیت کمتری نسبت به نویز دارد [۷۷].

در صورتی که هر فریم از سیگنال گفتار به صورت زیر نشان داده شود:

$$s_t(n) = \{s(Kt + 0), s(Kt + 1), \dots, s(Kt + N - 1)\}, \qquad (۱۳-۴)$$

که در این رابطه N طول هر فریم و K میزان انتقال هر فریم است. بردارهای ویژگی از طیف توان بدست آمده از هر یک از این فریم‌ها استخراج می‌شوند. در صورتی که فرض کنیم هر فریم $s_t(n)$ قسمتی از یک سیگنال متناوب $\tilde{s}_t(n)$ به صورت زیر باشد:

$$\tilde{s}_t(n) = \sum_{k=-\infty}^{+\infty} s_t\,(n + kN), \qquad (۱۴-۴)$$

آنگاه مربع اندازهٔ طیف سیگنال برابر است با اندازهٔ تبدیل فوریهٔ گسستهٔ تابع اتوکرولیشن مربوط به سیگنال متناوب $\tilde{s}_t(n)$ در طولی به اندازه طول هر فریم [۷۷، ۷۹]. برای بدست آوردن ضرایب اتوکرولیشن از رابطهٔ زیر استفاده می‌شود:

$$R(k) = \sum_{n=0}^{N-1} \tilde{s}_t(n)\, \tilde{s}_t(n + k)\,, \quad k = 0,1, \dots, N - 1. \qquad (۱۵-۴)$$

اگر از دید دیگری به رابطهٔ ۴-۱۵ نگاه شود متوجه می‌شویم که $R(k)$ تخمینی از میزان شباهت بین نمونه‌های قرار گرفته در فاصلهٔ k از هم، را در اختیار ما قرار می‌دهد که از طریق ضرب نقطه‌ایی بین دو بردار در فضای N بعدی، به صورت زیر، قابل بدست آمدن است. اگر،

$$\begin{aligned} \mathbf{x}_0 &= \{\tilde{s}_t(0), \tilde{s}_t(1), \dots, \tilde{s}_t(N-1)\} \\ \mathbf{x}_k &= \{\tilde{s}_t(k), \dots, \tilde{s}_t(N-1), \tilde{s}_t(0), \dots, \tilde{s}_t(k-1)\} \end{aligned} \qquad (۱۶-۴)$$

آنگاه $R(k)$ به صورت زیر بدست می‌آید:

$$R(k) = \mathbf{x}_0^T \mathbf{x}_k \,. \qquad (۱۷-۴)$$

اندازهٔ دو بردار داده شده در رابطهٔ ۴-۱۶ یکسان است اگر این اندازه با $\|\mathbf{x}\|$ نشان داده شود با توجه به رابطهٔ ۴-۱۷ زاویهٔ بین دو بردار در فضای N بعدی به صورت زیر محاسبه می‌شود:

$$P(k) = \theta_k = \cos^{-1}\left(\frac{R(k)}{\|\mathbf{x}\|^2}\right) \qquad (۱۸-۴)$$

در روش PAC از زاویهٔ محاسبه شده در رابطهٔ بالا به عنوان تخمین جدیدی از میزان شباهت بین دو بردار استفاده شده است.

در صورتی که تبدیل فوریهٔ گسسته به زاویه‌های بدست آمده از رابطهٔ ۴-۱۸ اعمال شود نمونهٔ معادل طیف توان معمولی بدست می‌آید که با نام طیف PAC شناخته می‌شود. در نهایت ضرایب کپسترال از این طیف توان جدید، که دارای مقاومت بیشتری نسبت به نویز است، بدست می‌آیند.

فصل پنجم

روش‌های پیشنهادی

فصل پنجم: روش‌های پیشنهادی

۱–۱۷ ۱–۵ مقدمه

تأثیر نویز بر سیگنال گفتار و کاهش دقت بازشناخت گفتار یکی از مهمترین مسائل موجود در سیستم‌های بازشناخت گفتار است که امروزه سعی می‌شود به گونه‌ایی آن را حل کرد. برای این منظور روش‌های متفاوتی ارائه شده است که هر کدام دارای معایب و مزایای خاص خود می‌باشد و برای کاربرد و نویز به خصوصی مناسب می‌باشند. در بعضی از این روش‌ها مثل PAC، ویژگی‌های استخراج شده در محیط‌های نویزی دارای کارآیی خوبی هستند ولی در شرایط بدون نویز دارای نرخ بازشناخت کمتری نسبت به روش‌های استاندارد استخراج ویژگی مثل MFCC هستند.

در این فصل، دو روش در زمینه استخراج ویژگی مقاوم برای بازشناخت گفتار پیشنهاد شده است. در این الگوریتم‌ها از روش MFCC به عنوان روش استاندارد موجود در سیستم‌های بازشناخت گفتار استفاده شده است. بیشترین اختلاف الگوریتم‌های پیشنهادی با روش MFCC در مرحلهٔ تخمین طیف توان می‌باشد. در روش‌های پیشنهادی از مزایای حوزهٔ اتوکرولیشن برای بدست آوردن

ویژگی‌های مقاوم در مقابل نویز مخصوصاً نویز جمع شونده استفاده شده است. در ادامه روش‌های پیشنهادی تحت عناوین مشتق طیف توان در حوزهٔ اتوکرولیشن (DPSA[۱۲۷]) [۸۱] و مشتق طیف اتوکرولیشن فاز (DPAS[۱۲۸]) [۸۲] مطرح می‌شوند.

۱–۱۸ ۵–۲ روش پیشنهادی اول: مشتق طیف توان در حوزهٔ اتوکرولیشن (DPSA)

همانطور که قبلاً نیز گفته شد سیستم‌های بازشناخت گفتار معمولاً در شرایطی مورد استفاده قرار می‌گیرند که با شرایط محل آموزش آنها متفاوت است. این تفاوت موجود، که بیشتر ناشی از نویز محیط در محل آزمایش و اعوجاج ناشی از کانال است، باعث کاهش نرخ بازشناخت گفتار می‌شود. از سه دستهٔ کلی از روش‌هایی که برای مقابله با این مشکل وجود دارد روش پیشنهادی در این بخش جزء دسته اول یعنی استخراج ویژگی مقاوم قرار می‌گیرد و برای مقابله با نویز جمع شونده ارائه شده است.

در روش پیشنهادی در این بخش برای بهبود روش MFCC و مقاوم سازی آن در برابر نویز جمع شونده از روش‌های AMFCC و DPS که در فصل قبل به آنها اشاره شد استفاده شده است. در این روش از مشتق طیف توان بدست آمده از ضرایب اتوکرولیشن با تأخیر زمانی بالا برای استخراج ویژگی استفاده شده است.

در ادامه مراحل مختلف مربوط به این روش پیشنهادی توضیح داده شده است.

---

127 Differential Power Spectrum in Autocorrelation domain

128 Differentiated Phase Autocorrelation Spectrum

۵-۲-۱ محاسبهٔ تابع اتوکرولیشن سیگنال نویزی

در صورتی که یک فریم از سیگنال نویزی با $x(n)$ نشان داده شود، داریم

$$x(n) = s(n) + d(n), \qquad n = 0,1, \dots, N-1 \qquad (۱-۵)$$

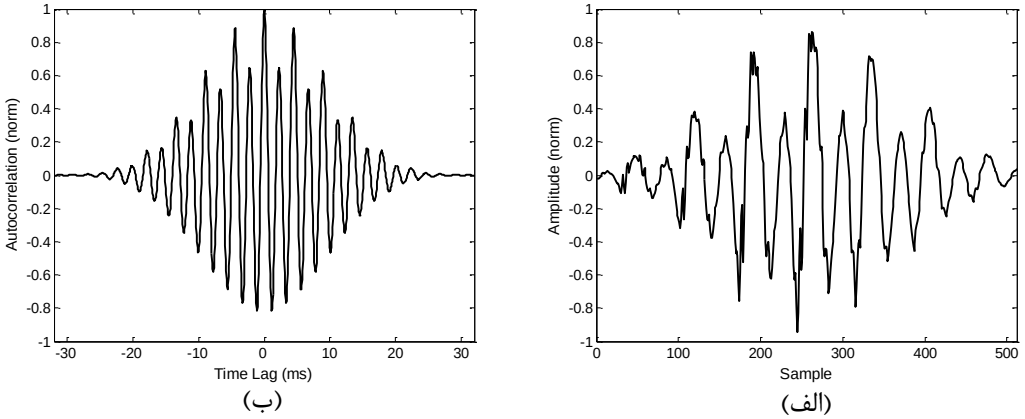
که در این رابطه $s(n)$ سیگنال بدون نویز و $d(n)$ نویز جمع شونده است. با فرض ناهمبسته بودن دو سیگنال $s(n)$ و $d(n)$ ، با اعمال تابع اتوکرولیشن به دو طرف رابطهٔ بالا، داریم

$$r_{xx}(i) = r_{ss}(i) + r_{dd}(i), \qquad i = 0,1, \dots, N-1 \qquad (۲-۵)$$

که در این رابطه $r_{xx}(i)$، $r_{ss}(i)$ و $r_{dd}(i)$ به ترتیب ضرایب اتوکرولیشن مربوط به سیگنال گفتار نویزی، سیگنال گفتار بدون نویز و سیگنال نویز جمع شونده هستند. برای بدست آوردن ضرایب اتوکرولیشن از رابطهٔ بایاس شدهٔ ۵-۳ استفاده شده است. به عبارت دیگر توابع موجود در رابطهٔ ۵-۲ ضرایب اتوکرولیشن یک طرفهٔ مربوط به سیگنال‌های موجود در رابطهٔ ۵-۱ هستند.

$$r_{xx}(i) = \frac{1}{N}\sum_{n=0}^{N-i-1} x(n)\, x(n+i), \qquad i = 0,1, \dots, N-1. \qquad (۳-۵)$$

یک فریم از آوای 'iy' و تابع اتوکرولیشن دو طرفهٔ مربوط به آن، به ترتیب در شکل‌های ۵-۱-الف و ۵-۱-ب نشان داده شده است.



(الف) (ب)

شکل ۵- ۱ (الف) یک فریم از آوای /iy/ (ب) تابع اتوکرولیشن دو طرفهٔ مربوط به آن.

۵-۲-۲ محاسبهٔ طیف توان از روی ضرایب اتوکرولیشن با تأخیر بالا

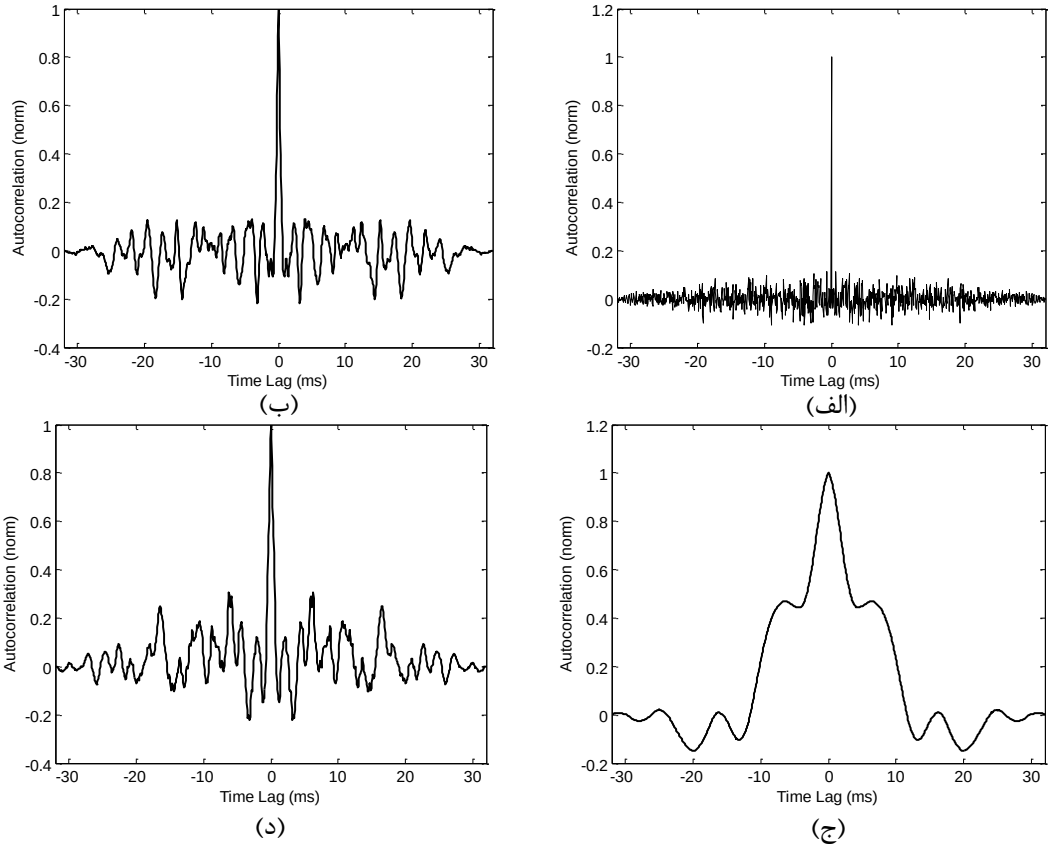
در روش MFCC اولین مرحله برای استخراج ویژگی، محاسبهٔ طیف توان می‌باشد. در روش معمول برای بدست آوردن طیف توان ابتدا تبدیل فوریهٔ زمان کوتاه یک فریم سیگنال گفتار بدست می‌آید، سپس از روش پریودوگرام طیف توان سیگنال بدست می‌آید [۷۶]. روش دیگری که برای محاسبهٔ طیف توان استفاده می‌شود، محاسبهٔ تبدیل فوریهٔ ضرایب اتوکرولیشن زمان کوتاه سیگنال گفتار است [۷۹].

با توجه به خواص تابع اتوکرولیشن، ضرایب اتوکرولیشن مربوط به سیگنال‌های نویز، بیشتر در تأخیرهای زمانی کم متمرکز شده‌اند و ضرایب با تأخیر زمانی بالا دارای مقدارهای کوچک و یا صفر هستند [۳۵، ۸۳]. برای مشاهدهٔ درستی این مطلب، تابع اتوکرولیشن زمان کوتاه مربوط به یک فریم ۳۲ میلی ثانیه‌ایی از سیگنال‌های نویز سفید، نویزکارخانه، نویز ماشین و نویز همهمه به ترتیب در شکل‌های ۵-۲-الف تا ۵-۲-د نشان داده شده است.

همانطور که از روی شکل ۵-۲ قابل مشاهده است بیشتر تمرکز ضرایب تابع اتوکرولیشن مربوط به نویزهای مختلف در تأخیرهای زمانی کم است و ضرایب تابع اتوکرولیشن مربوط به تأخیرهای بالا دارای مقدارهای کم و نزدیک به صفر هستند. بنابراین با توجه به این خاصیت تابع اتوکرولیشن می‌توان با حذف ضرایب اتوکرولیشن با تأخیر زمانی پایین اثر نویز را در طیف توان سیگنال گفتار کاهش داد. ولی سئوالی که در اینجا مطرح می‌شود این است که آیا می‌توان از روی ضرایب اتوکرولیشن با تأخیر زمانی بالا طیف توان را بدست آورد. در ادامه به بررسی این مطلب پرداخته شده است.

مدل متداولی که برای تولید صدا در انسان مورد استفاده قرار می‌گیرد، سیگنال صدا را به عنوان خروجی یک سیستم خطی تغییر پذیر با زمان، که برای اصوات بی‌صدا با نویز سفید و برای

اصوات صدادار با قطار ضربه متناوب تحریک می‌شود، در نظر می‌گیرد (شکل ۲-۳). در بازشناخت گفتار هدف بدست آوردن اندازه پاسخ این سیستم است.



شکل ۵- ۲ رشتهٔ اتوکرولیشن مربوط به یک فریم ۳۲ میلی ثانیه‌ایی از چهار نمونه نویز مختلف. (الف) رشتهٔ اتوکرولیشن مربوط به نویز سفید (ب) رشتهٔ اتوکرولیشن مربوط به نویز کارخانه (ج) رشتهٔ اتوکرولیشن مربوط به نویز ماشین (د) رشتهٔ اتوکرولیشن مربوط به نویز همهمه.

با بررسی ضرایب اتوکرولیشن مربوط به یک آوای صدادار و با جدا کردن اطلاعات مربوط به سیستم و سیگنال تحریک، می‌توان نتیجه گرفت که اطلاعات مربوط به سیستم که برای بازشناخت گفتار ضروری می‌باشند فقط در ضرایب اتوکرولیشن با تأخیر زمانی کم قرار دارند و اطلاعات مربوط به سیگنال تحریک به صورت متناوب در طول رشتهٔ اتوکرولیشن قرار گرفته‌اند. ولی با توجه به اینکه تابع اتوکرولیشن مربوط به یک سیگنال گفتار برابر است با کانولوشن تابع اتوکرولیشن مربوط به اطلاعات سیستم و تابع اتوکرولیشن مربوط به سیگنال تحریک و همچنین با توجه به متناوب بودن

تابع اتوکرولیشن مربوط به سیگنال تحریک، می‌توان نتیجه گرفت که اطلاعات مربوط به سیستم نیز به صورت متناوب در ضرایب با تأخیرهای بالای تابع اتوکرولیشن سیگنال گفتار تکرار می‌شوند. بنابراین برای بدست آوردن طیف توان می‌توان ضرایب با تأخیر زمانی پایین سیگنال گفتار را حذف و فقط از ضرایب با تأخیر زمانی بالا استفاده کرد [۷۶، ۸۰].

با بررسی ضرایب اتوکرولیشن مربوط به یک آوای بی‌صدا نیز می‌توان متوجه شد که اطلاعات مربوط به سیستم فقط در ضرایب با تأخیر زمانی پایین و اطلاعات مربوط به سیگنال تحریک به صورت غیر متناوب و غیر صفر در تمام طول رشتهٔ اتوکرولیشن قرار گرفته‌اند. در نتیجه با انجام کانولوشن بین این دو، اطلاعات مربوط به سیستم نیز در ضرایب اتوکرولیشن با تأخیر زمانی بالای سیگنال گفتار قرار می‌گیرند. بنابراین در مورد اصوات بی‌صدا نیز می‌توان طیف توان را فقط از ضراب اتوکرولیشن با تأخیر زمانی بالا بدست آورد [۷۶، ۸۰].

این روش تخمین طیف توان، تخمین طیف از روی ضرایب اتوکرولیشن با تأخیر زمانی بالا (HASE) نامیده می‌شود [۷۶]. در این بخش برای تخمین طیف با روش HASE از روشی مشابه به روش پیشنهاد شده در روش AMFCC [۷۶] استفاده شده است. برای این منظور بعد از محاسبهٔ ضرایب اتوکرولیشن یک طرفه مربوط به سیگنال گفتار نویزی، ضرایب با تأخیر زمانی کمتر از ۲ میلی ثانیه حذف می‌شوند. سپس برای بهبود بیشتر مقاومت طیف توان تخمین زده شده، یک پنجرهٔ مناسب به باقیماندهٔ ضرایب اعمال می‌شود. اثر این پنجره تضعیف بیشتر ضرایب با تأخیر زمانی پایین و همچنین ضرایب با تأخیر زمانی خیلی زیاد است [۷۶].

برای اینکه ضرایب تابع اتوکرولیشن بتوانند به خوبی طیف توان را تخمین بزنند گسترهٔ دینامیکی[۱۲۹] پنجره‌ایی که برای ضرب کردن در ضرایب با تأخیر زمانی بالا مورد استفاده قرار می‌گیرد باید دو برابر گسترهٔ دینامیکی پنجره‌ایی باشد که در مرحلهٔ پیش پردازش به هر فریم اعمال شده

---

129 Dynamic Range

است [۷۶]. این به این دلیل است که گسترهٔ دینامیکی طیف توان مربوط به رشتهٔ اتوکرولیشن دو برابر گسترهٔ دینامیکی طیف توان سیگنال است [۳۵].

از آنجا که پنجرهٔ مورد استفاده در مرحلهٔ پیش پردازش پنجرهٔ همینگ است و گسترهٔ دینامیکی این پنجره تقریباً ۴۳ دسی بل می‌باشد، پنجرهٔ مورد استفاده در مرحلهٔ تخمین طیف توان باید دارای گسترهٔ دینامیکی ۸۶ دسی بل باشد. برای این منظور از پنجرهٔ طراحی شده در روش AMFCC با نام پنجرهٔ همینگ با گسترهٔ دینامیکی دو برابر (DDR[۱۳۰]) استفاده شده است. روش بدست آوردن پنجرهٔ همینگ DDR با طول M، همانطور که در مرجع [۷۶] آورده شده است، به صورت زیر می‌باشد:

۱. ساخت یک پنجرهٔ همینگ معمولی با طول M.
۲. محاسبهٔ رشتهٔ اتوکرولیشن دو طرفه (بایاس شده) با طول M-1 که مقدار ماکزیمم آن در تأخیر زمانی صفر قرار داده شده است.
۳. برای بدست آوردن پنجرهٔ مورد نظر با طول M یک صفر به انتهای رشتهٔ بدست آمده از مرحلهٔ قبل اضافه می‌شود.

بعد از حذف ضرایب اتوکرولیشن با تأخیر زمانی کمتر از ۲ میلی ثانیه و اعمال پنجرهٔ همینگ DDR به باقی ماندهٔ ضرایب داریم

$$r_{xw}(i) = r_{sw}(i) + r_{dw}(i) \qquad (۴-۵)$$

که در این رابطه $r_{xw}(i)$ و $r_{sw}(i)$ به ترتیب ضرایب اتوکرولیشن یک طرفه مربوط به سیگنال گفتار نویزی و سیگنال گفتار بدون نویز عبور کرده از پنجره است و $r_{dw}(i)$ ضرایب اتوکرولیشن باقی مانده مربوط به نویز است که از پنجره عبور کرده‌اند.

---

130 Double-Dynamic-Range Hamming Window

برای بدست آوردن طیف توان به انتهای رشتهٔ اتوکرولیشن یک طرفه، که بعد از اعمال پنجرهٔ همینگ DDR بدست آمده است، 2N-M صفر اضافه می‌شود. سپس با اعمال تبدیل فوریهٔ گسسته و محاسبهٔ قدر مطلق آن طیف توان مورد نظر بدست می‌آید.

۵-۲-۳ محاسبهٔ مشتق طیف توان

در روش پیشنهادی DPSA ضرایب کپسترال از مشتق طیف توان بدست آمده از ضرایب اتوکرولیشن با تأخیر زمانی بالا بدست می‌آیند. با فرض ناهمبسته بودن نویز و سیگنال گفتار، طیف توان بدست آمده از مرحلهٔ قبل به صورت زیر نمایش داده می‌شود.

$$X(\omega) = S(\omega) + D(\omega), \qquad (۵-۵)$$

که در این رابطه $X(\omega)$، $S(\omega)$ و $D(\omega)$ به ترتیب طیف توان بدست آمده از ضرایب اتوکرولیشن با تأخیر زمانی بالا مربوط به سیگنال گفتار نویزی، سیگنال گفتار بدون نویز و سیگنال نویز می‌باشد و ω فرکانس زاویه‌ایی است. در این حالت مشتق طیف توان به صورت زیر نمایش داده می‌شود.

$$\mathrm{Diff}_x(\omega) = X'(\omega) = \frac{dX(\omega)}{d\omega}, \qquad (۶-۵)$$

که در این رابطه $\frac{d}{d\omega}$ و پرین نمایانگر مشتق نسبت به ω هستند. با توجه به فرض ناهمبستگی سیگنال‌ها و با توجه به رابطهٔ ۵-۵ داریم،

$$\mathrm{Diff}_x(\omega) = \frac{dX(\omega)}{d\omega} = \frac{dS(\omega)}{d\omega} + \frac{dD(\omega)}{d\omega} = \mathrm{Diff}_S(\omega) + \mathrm{Diff}_D(\omega) \qquad (۷-۵)$$

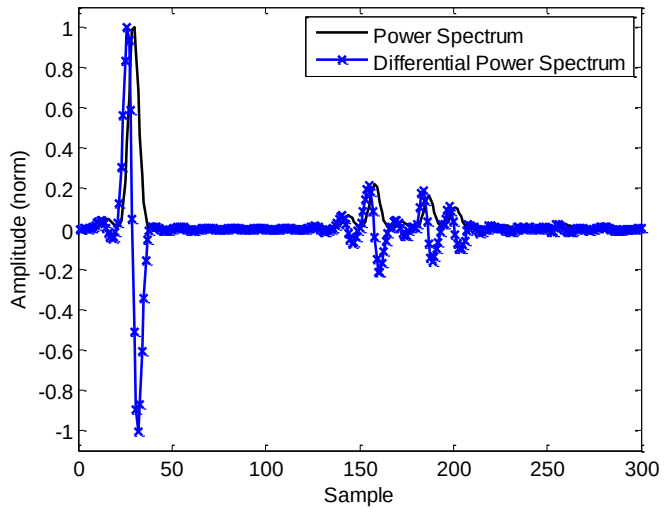
که در این رابطه $\mathrm{Diff}_S(\omega)$ و $\mathrm{Diff}_D(\omega)$ به ترتیب نمایانگر مشتق طیف توان بدست آمده از ضرایب اتوکرولیشن با تأخیر زمانی بالا مربوط به سیگنال بدون نویز و نویز است. تعریف داده شده از مشتق

طیف توان در رابطهٔ ۵-۷ مربوط به حوزهٔ فرکانس پیوسته می‌باشد معادل حوزهٔ گسسته آن بوسیلهٔ رابطهٔ تفاضلی زیر قابل تخمین است:

$$\begin{aligned} \mathrm{Diff_X(k)} &= \mathrm{Diff_S(k)} + \mathrm{Diff_D(k)} \approx \sum_{l=-Q}^{P} b_l\, X(k+l) \\ &\approx \sum_{l=-Q}^{P} b_l\, [S(k+l) + D(k+l)]\,, \end{aligned} \qquad (۵-۸)$$

که در این رابطه Q و P درجه‌های رابطهٔ تفاضلی، b_l ها ضرایب وزن دهی و $0 \le k \le K-1$، که K طول FFT است.

پیک‌های موجود در طیف توان سیگنال گفتار حاوی اطلاعات مهمی برای بازشناخت گفتار می‌باشند. با مشتق‌گیری از طیف توان پیک‌های موجود در طیف توان حذف نمی‌شوند و با عمل مشتق‌گیری هر پیک تبدیل به دو پیک مثبت و منفی می‌شود. یکی دیگر از مزایای مشتق‌گیری این است که قسمت هموار طیف توان که اغلب مربوط به نویز می‌باشد، تبدیل مقدار کوچکی نزدیک به صفر می‌شود [۷۱]. شکل ۵-۳ طیف توان مربوط به سیگنال گفتار به دست آمده از روی ضرایب اتوکرولیشن با تأخیر زمانی بالا مربوط به آوای نشان داده شده در شکل ۵-۱-الف را به همراه مشتق آن نشان می‌دهد.



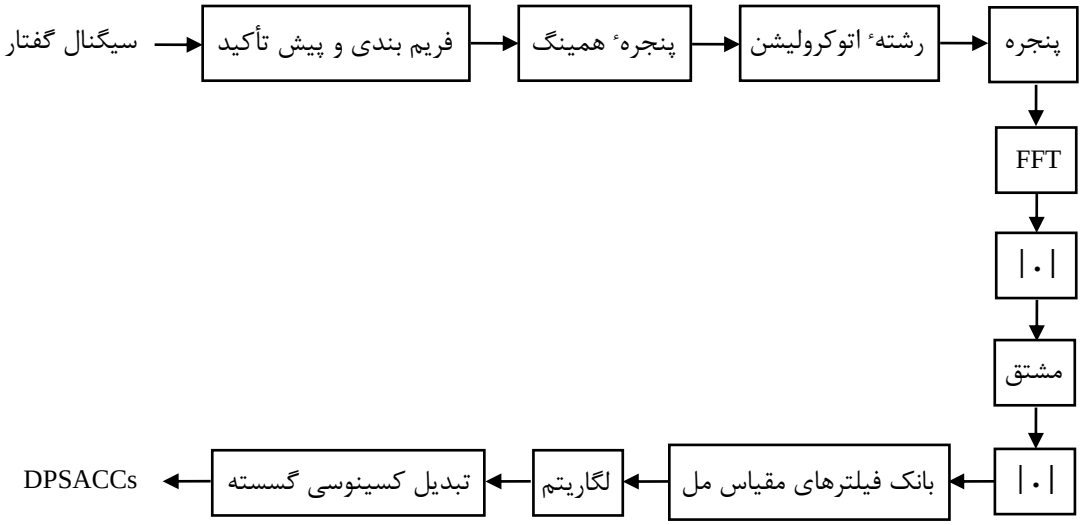
شکل ۵- ۳ طیف توان و مشتق طیف توان مربوط به یک فریم از آوای /iy/. (برای مشاهدهٔ بهتر پیک‌ها فقط ۳۰۰ نمونهٔ اول نشان داده شده است.)

برای بدست آوردن مشتق طیف توان نشان داده شده در شکل ۵-۳ از رابطهٔ زیر استفاده شده است.

$$\mathrm{Diff}_X(k) = X(k) - X(k+1) \qquad (۹-۵)$$

۵-۲-۴ مرحلهٔ نهایی: محاسبهٔ ضرایب کپسترال (DPSACCs)

در این قسمت ضرایب کپسترال فرکانس مل از مشتق طیف توان، بدست آمده در قسمت قبل، استخراج می‌شوند. بلوک دیاگرام کلی مربوط به روش پیشنهادی اول در شکل ۵-۴ نشان داده شده است.



شکل ۵- ۴ بلوک دیاگرام کلی مربوط به روش پیشنهادی اول (DPSACCs).

همانطور که در شکل ۵-۴ قابل مشاهده است در ابتدا سیگنال گفتار به فریم‌های با طول یکسان که با هم همپوشانی دارند تقسیم می‌شوند. سپس فیلتر پیش تأکید با رابطهٔ زیر به هر فریم اعمال می‌شود.

$$H(z) = 1 - 0.97z^{-1} \qquad (۱۰ -۵)$$

بعد از اعمال فیلتر پیش تأکید پنجرهٔ همینگ به هر فریم اعمال می‌شود و ضرایب اتوکرولیشن یکطرفه زمان کوتاه مربوط به هر فریم با استفاده از رابطهٔ ۵-۳ بدست می‌آیند. مانند روش AMFCC ضرایب اتوکرولیشن با تأخیر زمانی کمتر از ۲ میلی ثانیه حذف و پنجرهٔ همینگ DDR به باقیماندهٔ ضرایب اعمال می‌شود.

برای بدست آوردن ضرایب کپسترال فرکانس مل اندازهٔ مشتق طیف بدست آمده از ضرایب اتوکرولیشن با تأخیر بالا به فیلتر بانک مل اعمال می‌شود. در انتها تبدیل کسینوسی گسسته به لگاریتم انرژی‌های فیلتر بانک اعمال می‌شود.

۱۹–۱ ۵–۳ روش پیشنهادی دوم: مشتق طیف اتوکرولیشن فاز (DPAS)

همانطور که قبلاً گفته شد در محیط‌های بدون نویز، روش‌های متداول استخراج ویژگی مثل MFCC دارای کارآیی قابل قبولی هستند ولی در محیط‌های نویزی این روش‌ها کارآیی خود را از دست می‌دهند و دقت سیستم بازشناخت گفتار افت پیدا می‌کند. برای حل این مشکل روش‌های مختلفی ارائه شده است که هر کدام از این روش‌ها مزایا و معایب خاص خود را دارند. مشکلی که در بعضی از روش‌های ارائه شده برای بازشناخت گفتار مقاوم وجود دارد این است که در محیط‌های بدون نویز دارای دقت بازشناختی کمتر از روش‌های معمول مثل MFCC هستند. یکی از این روش‌ها، روش اتوکرولیشن فاز (PAC) [۷۷] است.

روش PAC با استفاده از دو عملیات اضافی، نرمالیزه سازی انرژی و تبدیل کسینوسی معکوس، ویژگی‌های جدیدی را بدست می‌آورد که دارای مقاومت خوبی نسبت به نویز هستند. همانطور که قبلاً

گفته شد در روش PAC به جای استفاده از اندازهٔ تبدیل فوریهٔ گسستهٔ ضرایب اتوکرولیشن، از اندازهٔ تبدیل فوریهٔ زاویه‌های بین بردارها برای بدست آوردن یک طیف توان جدید با نام طیف PAC استفاده شده است. به عبارت دیگر در این روش به جای استفاده از ضرب نقطه‌ایی بین بردارها، برای بدست آوردن ضرایب اتوکرولیشن، از زاویهٔ بین بردارها به عنوان معیاری برای شباهت بین آنها استفاده شده است. این تخمین جدید از میزان شباهت بین دو بردار "اتوکرولیشن فاز" نامیده می‌شود. علت استفاده از زاویه در این روش این است که زاویه نسبت به ضرب نقطه‌ایی تأثیر پذیری کمتری نسبت به نویز جمع شونده دارد [۸۴].

همانطور که گفته شد ایرادی که در روش PAC وجود دارد کارآیی کم آن، در مقایسه با دیگر روش‌های استخراج ویژگی، در محیط بدون نویز است. در این بخش سعی شده است که کارآیی روش PAC در محیط بدون نویز و تا حدودی نویزی بهبود داده شود.

برای این منظور در محاسبهٔ زاویهٔ بین بردارها (رشتهٔ اتوکرولیشن فاز) از ضرایب اتوکرولیشنی استفاده می‌شود که پنجرهٔ همینگ به آنها اعمال شده است. علاوه بر این از اندازهٔ مشتق طیف بدست آمده از زاویه‌ها به جای خود طیف برای استخراج ویژگی استفاده شده است. نتایج آزمایش‌های انجام شده نشان داده است که این دو عملیات باعث افزایش کارآیی روش PAC در محیط بدون نویز و تا حدودی در محیط نویزی می‌شود. در ادامه مراحل مختلف روش پیشنهادی توضیح داده می‌شود.

۵-۳-۱ بدست آوردن رشتهٔ اتوکرولیشن فاز

برای بدست آوردن رشتهٔ اتوکرولیشن فاز در ابتدا سیگنال گفتار به فریم‌های با اندازهٔ مساوی که با هم همپوشانی دارند تقسیم و هر فریم از فیلتر پیش تأکید و پنجرهٔ همینگ عبور داده می‌شود. در صورتی که یک فریم از سیگنال گفتار نویزی با $x(n)$ نمایش داده شود، داریم:

$$x(n) = s(n) + d(n), \qquad n = 0,1, \dots, N-1 \qquad (۱۱-۵)$$

که در این رابطه $x(n)$ و $d(n)$ به ترتیب سیگنال گفتار بدون نویز و نویز و N طول هر فریم است. در روش‌های مرسوم برای استخراج ویژگی مثل MFCC از طیف توان سیگنال گفتار برای استخراج ویژگی استفاده می‌شود. برای بدست آوردن طیف توان سیگنال گفتار معمولاً از مربع اندازهٔ تبدیل فوریهٔ گسستهٔ هر فریم استفاده می‌شود [۷۷]. در صورتی که هر فریم $x(n)$ قسمتی از سیگنال متناوب $\tilde{x}(n)$ ، تعریف شده به صورت زیر، باشد

$$\tilde{x}(n) = \sum_{k=-\infty}^{+\infty} x(n + kN) \qquad (۱۲-۵)$$

آنگاه طیف توان از طریق محاسبهٔ تبدیل فوریهٔ گسستهٔ رشتهٔ اتوکرولیشن، بدست آمده از سیگنال متناوب $\tilde{x}(n)$ در طولی برابر با طول هر فریم، نیز قابل بدست آمدن است [۷۹]. برای بدست آوردن رشتهٔ اتوکرولیشن از رابطهٔ زیر استفاده می‌شود:

$$r_{xx}(i) = \sum_{n=0}^{N-1} \tilde{x}(n)\tilde{x}(n + i), \qquad i = 0,1, \dots, N-1. \qquad (۱۳-۵)$$

۵-۳-۱-۱ اعمال تابع پنجره به رشتهٔ اتوکرولیشن

ضرایب اتوکرولیشن بدست آمده بوسیلهٔ رابطهٔ ۵-۱۳ مربوط به سیگنال گفتار نویزی می‌باشد. بنابراین هر چه اثر نویز در این ضرایب کمتر شود ویژگی‌های بدست آمده در مراحل بعدی کمتر تحت تأثیر نویز قرار خواهند گرفت.

با توجه به خاصیت تابع اتوکرولیشن همانطور که در شکل ۵-۲ نشان داده شده است ضرایب اتوکرولیشن مربوط به سیگنال‌های نویز بیشتر در تأخیرهای زمانی کم متمرکز شده‌اند و ضرایب اتوکرولیشن با تأخیر زمانی بالای آنها دارای مقدار کم نزدیک به صفر است. بنابراین همانطور که در

قسمت‌های قبل نیز گفته شد می‌توان با حذف یا تضعیف ضرایب با تأخیر زمانی پایین، اثر نویز را تا حدودی در ضرایب اتوکرولیشن کاهش داد.

بنابراین در این قسمت با اعمال یک پنجره همینگ به ضرایب اتوکرولیشن، ضرایب اتوکرولیشن با تأخیر زمانی پایین، که بیشتر تحت تأثیر نویز قرار دارند، تضعیف می‌شوند و ضرایب میانی، که حاوی اطلاعات مهمی برای بازشناخت گفتار هستند، وزن دهی می‌شوند. همانطور که در قسمت قبل گفته شد با انجام این کار اطلاعات مفیدی در مورد سیستم موجود در مدل تولید گفتار حذف نمی‌شود چون ضرایب اتوکرولیشن با تأخیر زمانی بالا نیز حاوی اطلاعات مورد نیاز در مورد سیستم هستند.

۵-۳-۱-۲ رشتهٔ اتوکرولیشن فاز

برای استخراج رشتهٔ اتوکرولیشن فاز از روی ضرایب اتوکرولیشن، بدست آمده بعد از اعمال پنجرهٔ همینگ، به صورت زیر عمل می‌شود.

اگر از دید دیگری به رابطهٔ ۵-۱۳ نگاه شود، می‌توان آن را به صورت ضرب نقطه‌ایی بین دو بردار در فضای N بعدی در نظر گرفت. در صورتی که دو بردار $\mathbf{x}_0$ و $\mathbf{x}_i$ در فضای N بعدی به صورت زیر تعریف شوند

$$\begin{aligned}\mathbf{x}_0 &= \{\tilde{x}(0), \tilde{x}(1), \dots, \tilde{x}(N-1)\} \\ \mathbf{x}_i &= \{\tilde{x}(i), \dots, \tilde{x}(N-1), \tilde{x}(0), \dots, \tilde{x}(i-1)\}\end{aligned} \qquad (۱۴-۵)$$

آنگاه رابطهٔ ۵-۱۳ به صورت ضرب نقطه‌ایی دو بردار داده شده در رابطهٔ ۵-۱۴ به صورت زیر نمایش داده می‌شود،

$$r_{xx}(i) = \mathbf{x}_0^T \mathbf{x}_i \ , \qquad (۱۵-۵)$$

و یا به طور معادل، با توجه به برابر بودن اندازهٔ بردارهای داده شده در رابطهٔ ۵-۱۴، داریم

$$r_{xx}(i) = \|\mathbf{x}\|^2 \cos(\theta_i), \qquad (۱۶ -۵)$$

که در این رابطه $\|\mathbf{x}\|^2$ انرژی هر فریم و θ_i زاویهٔ بین بردارهای $\mathbf{x}_0$ و $\mathbf{x}_i$ در فضای N بعدی است.

با توجه به رابطهٔ ۵-۱۶ رشتهٔ اتوکرولیشن فاز، $p_{xx}(i)$، از روی ضرایب اتوکرولیشن، $r_{xx}(i)$، به صورت زیر بدست می‌آیند:

$$p_{xx}(i) = \theta_i = \cos^{-1}\left(\frac{r_{xx}(i)}{\|\mathbf{x}\|^2}\right). \qquad (۱۷ -۵)$$

که در این رابطه $r_{xx}(i)$ ضرایب اتوکرولیشن بدست آمده بعد از اعمال پنجرهٔ همینگ است.

زاویه‌های بدست آمده بوسیله رابطه ۵-۱۷ به عنوان معیاری جدید برای میزان شباهت بین دو بردار، به جای ضرایب اتوکرولیشن بدست آمده از ضرب نقطه‌ایی، مورد استفاده قرار می‌گیرند و رشتهٔ اتوکرولیشن فاز نامیده می‌شوند.

با اعمال تبدیل فوریهٔ گسسته به رابطهٔ ۵-۱۷ معادلی برای طیف توان معمولی سیگنال بدست می‌آید که طیف PAC خوانده می‌شود. از این طیف می‌توان برای استخراج ویژگی استفاده کرد.

۵-۳-۲ مشتق طیف اتوکرولیشن فاز

در این قسمت در ابتدا با اعمال تبدیل فوریهٔ گسسته به رشتهٔ اتوکرولیشن فاز، بدست آمده در قسمت قبل، طیف اتوکرولیشن فاز (طیف PAC) بدست می‌آید. طیف اتوکرولیشن فاز به دست آمده در این قسمت به عنوان معادلی برای طیف توان معمولی سیگنال گفتار برای استخراج ویژگی مورد استفاده قرار می‌گیرد. در ادامه از مشتق این طیف جدید برای استخراج ویژگی استفاده می‌شود.

همانطور که در مرجع [۷۱] نشان داده شده است مشتق گیری از طیف توان سیگنال گفتار نسبت به فرکانس در روش MFCC می‌تواند باعث افزایش کارآیی سیستم بازشناخت گفتار هم در محیط بدون نویز و هم در محیط نویزی شود.

همانطور که قبلاً گفته شد پیک‌های موجود در طیف توان حاوی اطلاعات مهمی برای بازشناخت گفتار هستند. با گرفتن مشتق از طیف توان این پیک‌ها حذف نمی‌شوند بلکه تبدیل به دو پیک مثبت و منفی می‌شوند. بنابراین با مشتق گیری از طیف توان نه تنها اطلاعات مهمی، که برای بازشناخت گفتار حیاتی است، حذف نمی‌شود بلکه قسمت هموار طیف توان که معمولاً مربوط به نویز می‌باشد تبدیل به مقدار کوچکی نزدیک به صفر می‌شود. بنابراین در این مرحله برای افزایش بیشتر کارآیی از مشتق طیف اتوکرولیشن فاز برای استخراج ویژگی استفاده شده است.

اگر از رشتهٔ اتوکرولیشن فاز بدست آمده بوسیلهٔ رابطهٔ ۵-۱۷ تبدیل فوریه گرفته شود، داریم:

$$X(\omega) = FT\{p_{xx}(i)\} \qquad (۱۸-۵)$$

که در این رابطه $X(\omega)$ طیف اتوکرولیشن فاز می‌باشد. با مشتق گیری نسبت به ω مشتق طیف اتوکرولیشن فاز به صورت زیر بدست می‌آید:

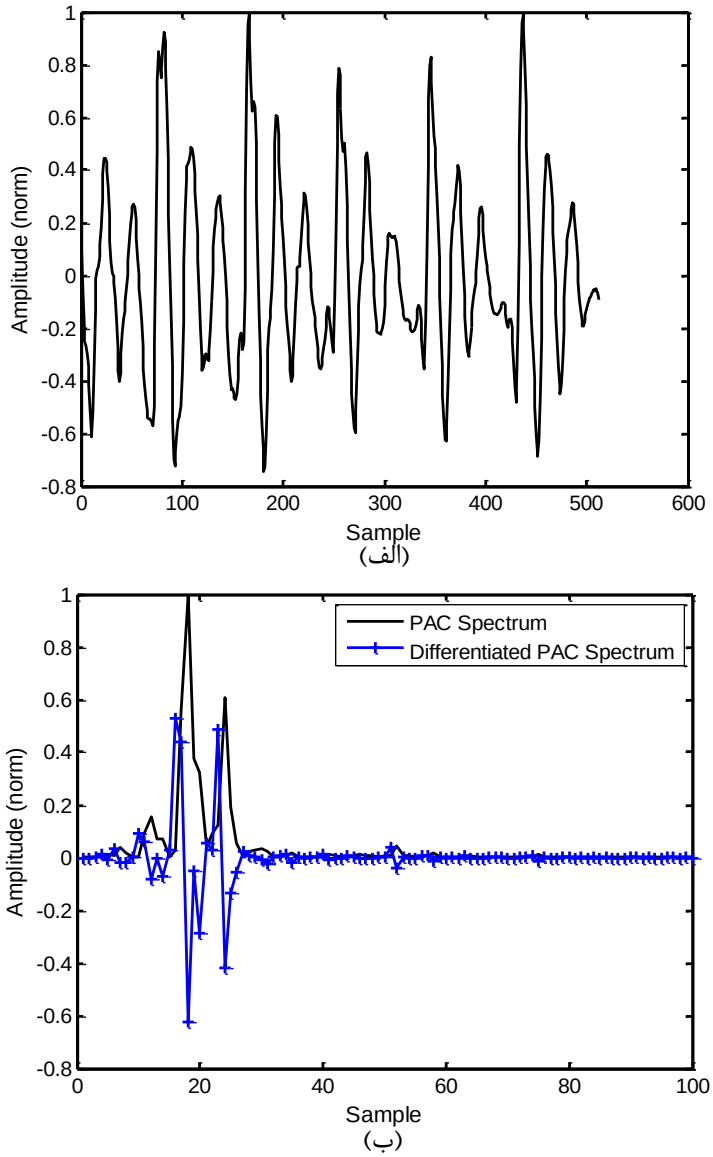
$$Diff_x(\omega) = \frac{dX(\omega)}{d\omega}. \qquad (۱۹-۵)$$

رابطهٔ بالا مربوط به حوزهٔ فرکانس پیوسته می‌باشد. معادل حوزهٔ گسستهٔ آن به صورت زیر تخمین زده می‌شود:

$$Diff_x(k) \approx \sum_{l=-Q}^{P} b_l X(k+l), \qquad 0 \le k \le K-1 \qquad (۲۰-۵)$$

که در این رابطه Q و P درجه‌های رابطهٔ تفاضلی، b_l ها ضرایب وزن دهی و K طول FFT است.

شکل ۵-۵-الف یک فریم از آوای /aa/ را نشان می‌دهد که از دیتابیس TIMIT انتخاب شده است و شکل ۵-۵-ب طیف PAC بدست آمده از آن را به همراه مشتق طیف اتوکرولیشن فاز نشان می‌دهد. برای بدست آوردن طیف PAC نشان داده شده در شکل ۵-۵-ب از ضرایب اتوکرولیشن عبور کرده از پنجرهٔ همینگ استفاده شده است و مشتق آن بوسیلهٔ رابطهٔ ۵-۹ بدست آمده است.

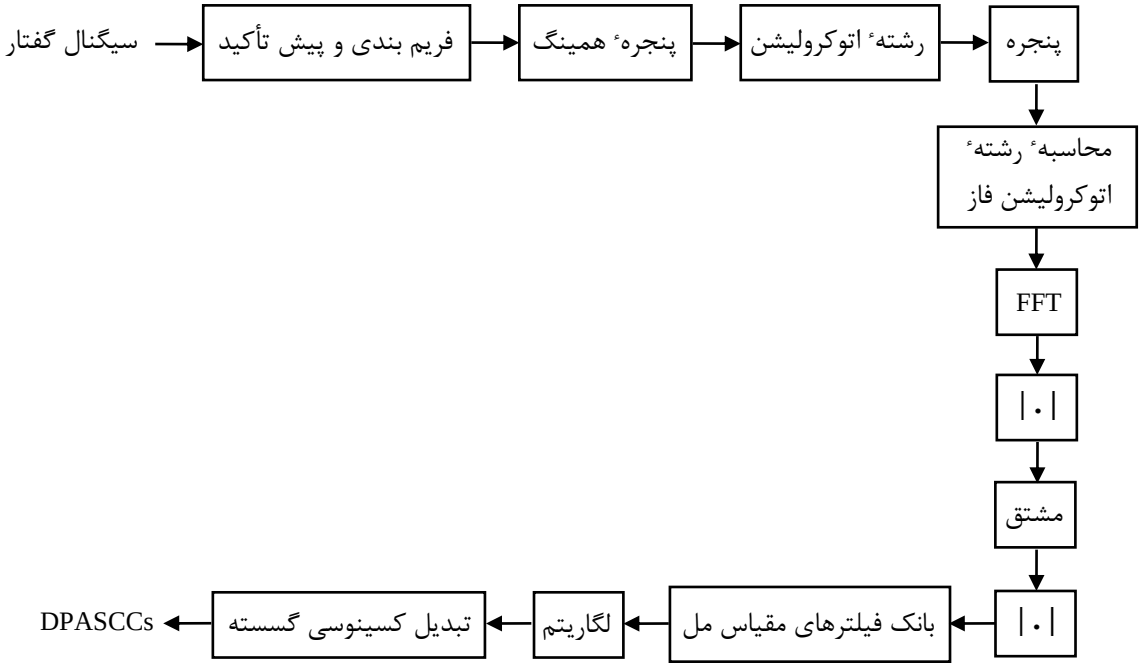


شکل ۵- ۵ طیف اتوکرولیشن فاز و مشتق طیف اتوکرولیشن فاز مربوط به یک فریم از آوای /aa/. (الف) شکل موج یک فریم از آوای /aa/ (ب) طیف اتوکرولیشن فاز و مشتق طیف اتوکرولیشن فاز مربوط به آوای /aa/. (برای نمایش بهتر پیک‌ها فقط ۱۰۰ نمونهٔ اول طیف نشان داده شده است.)

همانطور که در شکل نشان داده شده است بعد از مشتق گیری پیک‌های موجود در طیف اتوکرولیشن فاز در مشتق طیف اتوکرولیشن فاز باقی می‌مانند و هر پیک به دو پیک مثبت و منفی تبدیل می‌شود. از اندازهٔ مشتق طیف اتوکرولیشن فاز (DPAS) می‌توان برای استخراج ویژگی استفاده کرد.

۵-۳-۳ مرحلهٔ نهایی: محاسبهٔ ضرایب کپسترال (DPASCCs)

در این مرحله از خروجی بدست آمده از مرحلهٔ قبل ضرایب کپسترال فرکانس مل بدست می‌آیند و از این ضرایب جدید به عنوان ویژگی‌های مقاوم در برابر نویز، برای بازشناخت گفتار استفاده می‌شود. بلوک دیاگرام کلی مربوط به استخراج ویژگی‌ها توسط روش پیشنهادی دوم در شکل ۵-۶ نشان داده شده است.



شکل ۵- ۶ بلوک دیاگرام کلی مربوط به روش پیشنهادی دوم (DPASCCs).

همانطور که در شکل ۵-۶ نشان داده شده است در ابتدا سیگنال گفتار به فریم‌های با اندازهٔ مساوی که با هم همپوشانی دارند تقسیم می‌شود. بعد از عبور هر فریم از فیلتر پیش تأکید با رابطهٔ ۵-۱۰ پنجرهٔ همینگ به هر فریم اعمال می‌شود. ضرایب اتوکرولیشن با رابطهٔ ۵-۱۳ محاسبه و پنجرهٔ همینگ به ضرایب اعمال می‌شود. سپس رشتهٔ اتوکرولیشن فاز با استفاده از رابطه ۵-۱۷ بدست می‌آید. با اعمال تبدیل فوریهٔ گسسته به رشتهٔ اتوکرولیشن فاز طیف اتوکرولیشن فاز بدست می‌آید.

در ادامه بانک فیلترهای مقیاس مل به اندازهٔ مشتق طیف اتوکرولیشن فاز اعمال می‌شود. و در انتها ضرایب کپسترال مربوط به روش پیشنهادی دوم با اعمال تبدیل کسینوسی گسسته به لگاریتم انرژی‌های فیلتر بانک مل بدست می‌آیند.

فصل ششم

نتایج آزمایش‌های انجام شده

فصل ششم: نتایج آزمایش‌های انجام شده

۶-۱ مقدمه

در این فصل نتایج آزمایش‌های انجام شده برای تست کارآیی روش‌های پیشنهادی آورده شده است. برای این منظور نتایج روش‌های پیشنهادی DPSA و DPAS با روش‌های MFCC، DPS، AMFCC و PAC ، هم برای حالت سیگنال گفتار نویزی و هم بدون نویز، مقایسه شده است. برای بدست آوردن سیگنال گفتار نویزی، سیگنال نویز با توان‌های متفاوت به سیگنال گفتار تمیز اضافه شده است. برای مشخص کردن میزان آلودگی سیگنال گفتار با نویز، از نسبت سیگنال به نویز (SNR[۱۳۱]) استفاده شده است. SNR مربوط به سیگنال گفتار نویزی معرف انرژی سیگنال گفتار تمیز به انرژی سیگنال نویز می‌باشد. در آزمایش‌های انجام شده در این فصل برای آموزش سیستم بازشناخت فقط از سیگنال گفتار تمیز استفاده شده است.

---

131 Signal to Noise Ratio

در ادامه دیتابیس و پارامترهای مورد استفاده برای پیاده‌سازی روش‌ها شرح داده شده است. سپس نتایج آزمایش‌های انجام شده مربوط به روش‌های پیشنهادی با بعضی از روش‌های معمول دیگر مقایسه شده است.

۶-۲ دیتابیس مورد استفاده

دیتابیس مورد استفاده در این پایان نامه دیتابیس TIMIT می‌باشد. TIMIT یکی از دیتابیس‌هایی است که توسط NIST و با حمایت DARPA به صورت CD در آمده است. این دیتابیس حاوی جمله‌های گفتاری پیوسته است که محل آواها و کلمه‌ها در هر جمله مشخص شده است.

این دیتابیس حاوی ۶۳۰۰ جملهٔ گفتاری پیوسته می‌باشد که هر ۱۰ جمله توسط یک گوینده ادا شده است. به عبارت دیگر در این دیتابیس از ۶۳۰ گویندهٔ متفاوت استفاده شده است. گوینده‌های استفاده شده از ۸ ناحیهٔ جغرافیایی آمریکا انتخاب شده‌اند و جمله‌های گفتاری با توجه به این ۸ ناحیه به ۸ قسمت متفاوت تقسیم شده‌اند.

دیتابیس شامل دو قسمت کلی تست و آموزش است که گوینده‌های مورد استفاده برای ایجاد داده‌های تست و آموزش با هم متفاوت می‌باشند. در این پایان‌نامه برای تست و آموزش سیستم بازشناخت از تمام داده‌های مربوط به ناحیهٔ جغرافیایی ۲ استفاده شده است. به عبارت دیگر برای آموزش سیستم بازشناخت از ۷۶۰ جمله (۵۳ مرد و ۲۳ زن) و برای تست از ۲۶۰ جمله (۱۸ مرد و ۸ زن) استفاده شده است.

سیگنال‌های نویز مورد استفاد برای نویزی کردن سیگنال‌های گفتار از مجموعهٔ NATO RSG-10 [۸۵] استخراج شده است. سپس برای بدست آوردن سیگنال گفتار نویزی، سیگنال‌های نویز با توان‌های متفاوت به سیگنال گفتار اصلی اضافه شده است.

۳–۶ پارامترها و کلاسیفایر مورد استفاده

برای استخراج ویژگی در تمام روش‌ها فرکانس نمونه برداری ۱۶ کیلو هرتز، طول هر فریم ۳۲ میلی ثانیه و میزان انتقال هر فریم ۱۰ میلی ثانیه انتخاب شده است. ضریب پیش تأکید مورد استفاده در فیلتر پیش تأکید مقدار ۰/۹۷ قرار داده شده است و در مرحلهٔ اعمال بانک فیلترها نیز از ۲۳ فیلتر در مقیاس مل استفاده شده است.

در تمام روش‌ها علاوه بر ۱۲ ضریب کپسترال استخراج شده، یک ضریب لگاریتم انرژی مربوط به هر فریم عبور کرده از پنجرهٔ همینگ نیز به ضرایب اضافه شده است تا جمعاً ۱۳ ضریب پایه داشته باشیم. مشتق‌های اول و دوم این ۱۳ ضریب نیز به ضرایب پایه اضاف شده است تا یک بردار ویژگی با طول ۳۹ بدست آید. از این بردار به عنوان ویژگی‌های استخراج شده از هر فریم برای بازشناخت گفتار استفاده شده است.

برای ایجاد مدل‌های صوتی از مدل مخفی مارکوف استفاده شده است و برای ایجاد مدل‌های مخفی مارکوف، آموزش و تست آنها از بستهٔ نرم افزاری مدل مخفی مارکوف (HTK) [۲۹] استفاده شده است. برای این منظور در HTK رشته‌های سه آوائی گره خورده به هم با یک تابع گوسی در هر خروجی ایجاد شده است. برای هر مدل مخفی مارکوف از ۵ حالت با ساختار چپ به راست استفاده شده است که فقط ۳ حالت دارای خروجی می‌باشند. برای آموزش مدل‌های ایجاد شده از تمام داده‌های آموزشی مربوط به ناحیهٔ جغرافیایی ۲ (۷۶۰ جمله) و برای تست از تمام جمله‌های مربوط به تست (۲۶۰ جمله) استفاده شده است.

بعد از انجام بازشناخت برای ارزیابی کارآیی سیستم بازشناخت، بوسیلهٔ HTK جمله‌های تشخیص داده شده با جمله‌های صحیح مقایسه می‌شود و با استفاده از رابطهٔ زیر نرخ بازشناخت (نرخ کلمه‌های درست تشخیص داده شده) بدست می‌آید.

$$\text{نرخ بازشناخت} = \frac{N-D-S}{N} \times 100\% \qquad (۱-۶)$$

که در این رابطه N تعداد کل کلمه‌ها، D خطای ناشی از حذف کلمه‌ها و S خطای ناشی از جابجایی بین کلمه‌ها (تشخیص یک کلمه به جای کلمهٔ دیگر) است.

۶-۴ بررسی کارآیی روش پیشنهادی اول (DPSA)

در این قسمت کارآیی روش DPSA با روش‌های MFCC، DPS و AMFCC مقایسه شده است. در ابتدا به صورت تجربی یک رابطهٔ مناسب برای مشتق گیری بدست آمده است. سپس کارآیی ضرایب کپسترال بدست آمده به کمک روش DPSA با ضرایب کپسترال بدست آمده با بعضی از روش‌های دیگر مقایسه شده است.

۶-۴-۱ تنظیم پارامترهای مشتق

همانطور که در فصل قبل گفته شد برای محاسبهٔ مشتق طیف توان از رابطهٔ زیر استفاده می‌شود:

$$\text{Diff}_X(k) \approx \sum_{l=-Q}^{P} b_l X(k+l) \qquad (۲-۶)$$

مسئله‌ایی که در استفاده از این رابطه مطرح می‌شود طریقهٔ انتخاب مرتبهٔ مشتق یعنی پارامترهای P و Q و همچنین وزن‌های b_l است. متأسفانه پیدا کردن یک راه حل بهینه برای حل این مسئله کاری مشکل می‌باشد. بنابراین مانند مرجع [۷۱] از روش تجربی برای انتخاب پارامترها استفاده می‌شود. برای این منظور از سه رابطهٔ زیر استفاده و نتایج آنها با هم مقایسه می‌شود.

$$\mathrm{Diff}_X(k) = X(k) - X(k+1), \qquad (۳-۶)$$

$$\mathrm{Diff}_X(k) = X(k) - X(k+2), \qquad (۴-۶)$$

$$\mathrm{Diff}_X(k) = X(k-2) + X(k-1) - X(k+1) - X(k+2). \qquad (۵-۶)$$

برای این منظور سه روش DPSA1، DPSA2 و DPSA3 که در هم کدام از آنها به ترتیب از رابطه‌های ۶-۳، ۶-۴ و ۶-۵ استفاده شده است، تعریف می‌گردند. نرخ بازشناخت مربوط به ضرایب کپسترال بدست آمده با هر یک از این سه روش در حالت بدون نویز و در حالتی که سیگنال گفتار با SNR های مختلف با نویز سفید جمع شده در جدول ۶-۱ نشان داده شده است.

جدول ۶- ۱ نرخ بازشناخت مربوط به روش پیشنهادی (DPSA) با سه رابطهٔ مشتق متفاوت.

SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB
DPSA1	۷۷/۱۲	۷۶/۰۶	۷۴/۱۳	۷۰/۴۵	۶۱/۵۱	۴۰/۲۰	۱۹/۶۸
DPSA2	۷۷/۷۳	۷۶/۱۵	۷۴/۵۷	۷۰/۶۷	۶۲/۰۸	۳۹/۸۵	۱۹/۷۷
DPSA3	۷۷/۵۱	۷۶/۷۶	۷۵/۰۵	۷۰/۷۶	۶۲/۷۸	۴۰/۵۵	۲۰/۳۰

همانطور که از جدول ۶-۱ قابل مشاهده است رابطهٔ ۶-۵ (DPSA3) در کل دارای نتایج بهتری می‌باشد. بنابراین در تمام پیاده‌سازی‌های انجام شده مربوط به روش پیشنهادی در این بخش از رابطه ۶-۵ استفاده شده است. برای اختصار در نمایش‌های بعدی عدد ۳ در DPSA3 حذف شده است.

۶-۴-۲ مقایسهٔ کارآیی روش DPSA با بعضی از روش‌های دیگر

در این قسمت کارآیی روش DPSA با روش‌های MFCC، DPS و AMFCC در حضور نویزهای مختلف و با نسبت‌های سیگنال به نویز مختلف مقایسه شده است. جدول‌های ۶-۲، ۶-۳، ۶-۴ و ۶-۵

به ترتیب نرخ بازشناخت مربوط به ۴ روش بالا را در حضور نویز سفید، نویز کارخانه، نویز همهمه و نویز هواپیما (F16) نشان می‌دهد.

جدول ۶- ۲ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی در حضور نویز سفید با توان‌های متفاوت.

SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB	0dB
MFCC	۷۶/۹۰	۷۵/۶۷	۷۳/۵۶	۶۹/۴۴	۵۸/۶۱	۳۸/۷۵	۱۸/۴۶	۵/۷۹
DPS	۷۷/۲۰	۷۵/۷۱	۷۳/۹۶	۶۹/۹۳	۶۰/۰۲	۴۰/۰۷	۱۹/۹۰	۷/۴۱
AMFCC	۷۶/۵۹	۷۵/۲۳	۷۴/۴۰	۷۰/۶۳	۶۱/۴۲	۴۰/۶۸	۱۹/۷۷	۶/۴۰
DPSA (روش پیشنهادی ۱)	۷۷/۵۱	۷۶/۷۶	۷۵/۰۵	۷۰/۷۶	۶۲/۷۸	۴۰/۵۵	۲۰/۳۰	۶/۵۳

جدول ۶- ۳ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی در حضور نویز کارخانه با توان‌های متفاوت.

SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB	0dB
MFCC	۷۶/۹۰	۷۶/۲۴	۷۵/۹۸	۷۴/۶۲	۷۰/۴۵	۵۷/۹۱	۳۰/۴۷	۹/۲۹
DPS	۷۷/۲۰	۷۶/۶۳	۷۶/۵۵	۷۴/۹۷	۷۱/۰۲	۶۱/۲۰	۳۶/۴۸	۱۱/۵۳
AMFCC	۷۶/۵۹	۷۶/۶۸	۷۶/۲۸	۷۴/۷۵	۷۲/۲۵	۶۳/۰۰	۳۸/۸۴	۱۱/۸۸
DPSA (روش پیشنهادی ۱)	۷۷/۵۱	۷۷/۴۷	۷۶/۸۵	۷۶/۱۵	۷۲/۶۹	۶۵/۴۵	۴۱/۳۰	۱۵/۲۶

جدول ۶- ۴ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی در حضور نویز همهمه با توان‌های متفاوت.

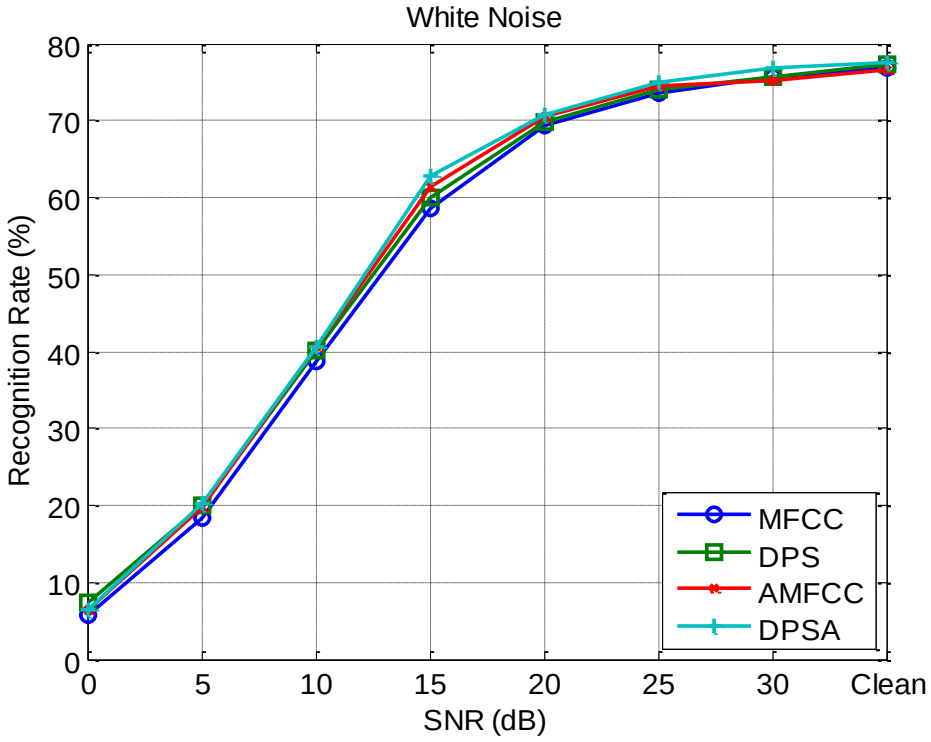
SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB	0dB
MFCC	۷۶/۹۰	۷۶/۸۱	۷۶/۵۰	۷۵/۳۶	۷۲/۷۳	۶۴/۸۴	۴۳/۸۴	۱۷/۰۱
DPS	۷۷/۲۰	۷۷/۲۵	۷۷/۲۹	۷۵/۳۲	۷۱/۹۰	۶۴/۷۵	۴۵/۷۷	۱۷/۹۷
AMFCC	۷۶/۵۹	۷۶/۵۰	۷۵/۶۲	۷۵/۰۵	۷۲/۲۹	۶۶/۴۶	۴۸/۹۷	۲۱/۰۰
DPSA (روش پیشنهادی ۱)	۷۷/۵۱	۷۷/۲۰	۷۶/۲۸	۷۵/۲۳	۷۳/۰۸	۶۶/۹۹	۴۸/۲۷	۲۰/۷۴

جدول ۶- ۵ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی در حضور نویز F16 با توان‌های متفاوت.

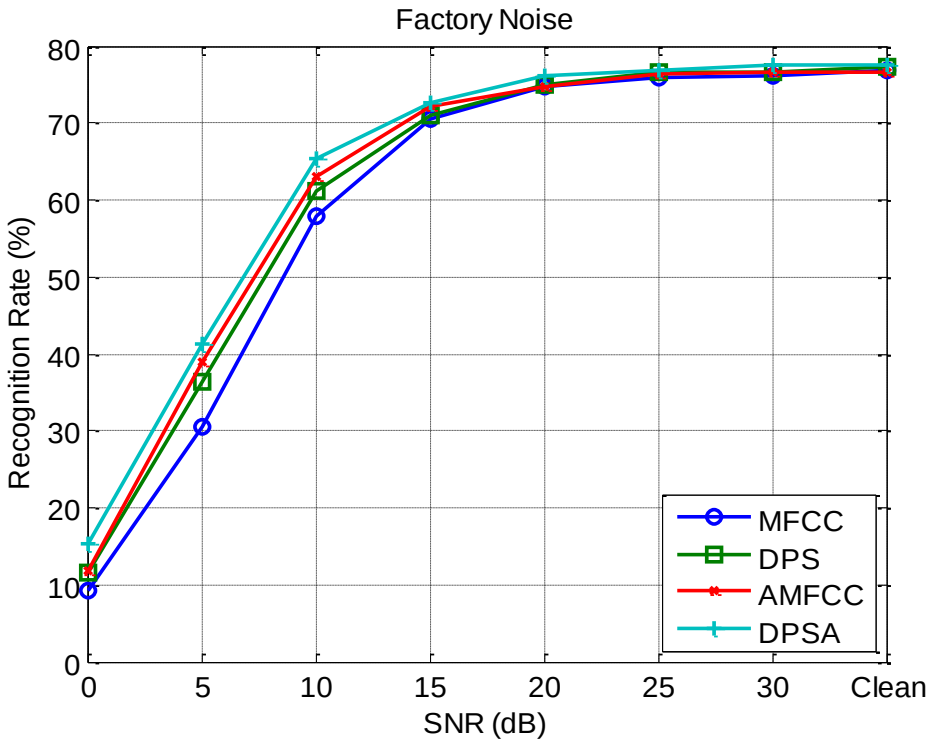
SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB	0dB
MFCC	۷۶/۹۰	۷۶/۲۴	۷۵/۰۵	۷۳/۰۸	۶۷/۹۵	۵۸/۷۵	۳۷/۶۲	۱۳/۴۶
DPS	۷۷/۲۰	۷۶/۱۹	۷۵/۸۴	۷۲/۹۵	۶۷/۹۱	۵۹/۹۷	۳۹/۲۸	۱۴/۱۶
AMFCC	۷۶/۵۹	۷۶/۵۵	۷۵/۸۴	۷۳/۷۰	۶۹/۲۲	۶۰/۵۹	۴۱/۲۱	۱۴/۸۶
DPSA (روش پیشنهادی ۱)	۷۷/۵۱	۷۷/۵۵	۷۶/۵۰	۷۳/۷۸	۷۰/۱۰	۶۱/۷۳	۴۱/۹۱	۱۵/۷۸

با توجه به جدول ۶-۲ در حالتی که سیگنال گفتار با نویز سفید آلوده شده است و در سیگنال به نویزهای مختلف روش پیشنهادی دارای نتایج بهتری نسبت به سه روش دیگر است به جز در سیگنال به نویزهای ۱۰ و صفر دسی بل که به ترتیب روش‌های AMFCC و DPS نتایج بهتری را بدست آورده‌اند. همچنین با بررسی جدول ۶-۴ در حالتی که سیگنال گفتار با نویز همهمه آلوده شده است، می‌توان مشاهده کرد که در نسبت‌های سیگنال به نویز ۱۵، ۱۰ و ۵ دسی بل و همچنین در حالت سیگنال تمیز روش پیشنهادی نسبت به سه روش دیگر دارای نتایج بهتری می‌باشد. در دو جدول ۶-۳ و ۶-۵ نیز که سیگنال گفتار به ترتیب با نویزهای کارخانه و F16 آلوده شده است، در تمام موارد روش پیشنهادی نتایج بهتری را بدست آورده است. برای راحتی مقایسه، نتایج موجود در جدول‌های ۶-۲ تا ۶-۵ به ترتیب در شکل‌های ۶-۱-الف تا ۶-۱-د به صورت نمودار نشان داده شده است.

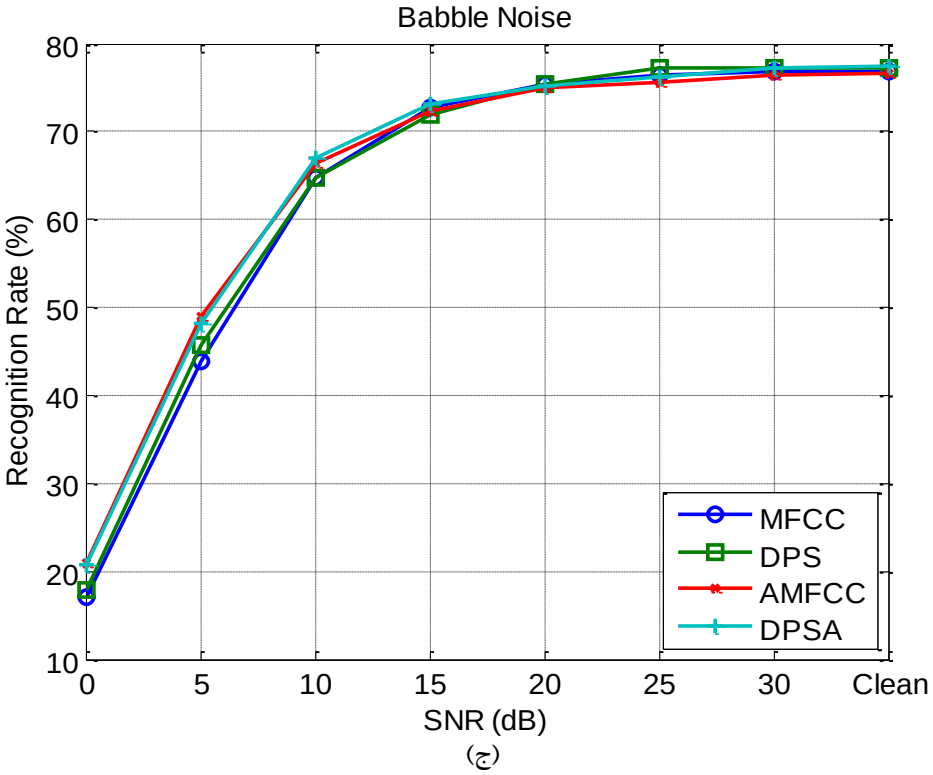
همانطور که از روی جدول‌ها و شکل‌ها قابل مشاهده است روش پیشنهادی اول (DPSA) نه تنها در حالت بدون نویز بلکه در شرایط نویزی نیز دارای نتایج کلی بهتری است.


White Noise
Recognition Rate (%)
SNR (dB)
0
5
10
15
20
25
30
Clean
MFCC
DPS
AMFCC
DPSA

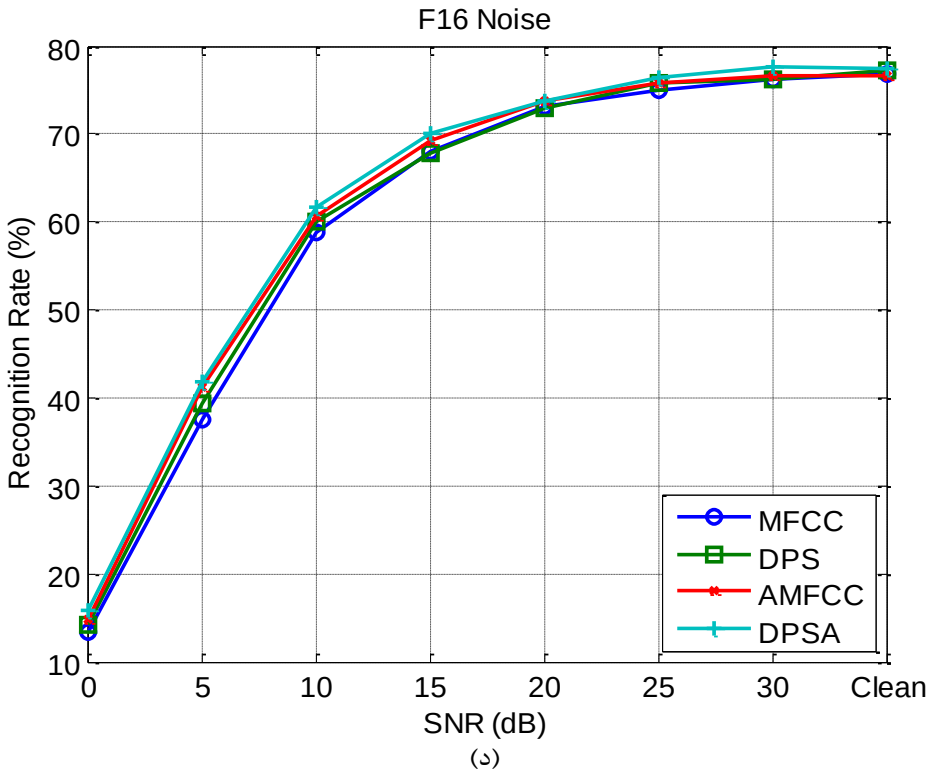
(الف)


Factory Noise
Recognition Rate (%)
SNR (dB)
0
5
10
15
20
25
30
Clean
MFCC
DPS
AMFCC
DPSA

(ب)



(ج)



(د)

شکل ۶- ۱ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی در حضور نویز با توان‌های متفاوت. (الف) نویز سفید (ب) نویز کارخانه (ج) نویز همهمه (د) نویز F16.

۶–۵ بررسی کارآیی روش پیشنهادی دوم (DPAS)

در این بخش نیز مانند بخش قبل در ابتدا یک رابطهٔ مناسب برای مشتق گیری از طیف اتوکرولیشن فاز به صورت تجربی بدست آمده است. سپس کارآیی روش پیشنهادی با بعضی از روش‌های دیگر از جمله روش PAC مقایسه شده است.

۶–۵–۱ تنظیم پارامترهای مشتق

همانطور که در بخش قبل گفته شد بدست آوردن یک راه حل بهینه برای محاسبهٔ پارامترهای مشتق موجود در رابطهٔ ۶-۲ مشکل است. بنابراین در این قسمت نیز به صورت تجربی از بین سه رابطهٔ ۶-۳ تا ۶-۵ یکی انتخاب و مورد استفاده قرار گرفته است.

برای این منظور ضرایب کپسترال با سه روش DPAS1، DPAS2 و DPAS3، که در هر کدام از آنها به ترتیب از رابطه‌های ۶-۳، ۶-۴ و ۶-۵ استفاده شده، بدست آمده است. نرخ بازشناخت مربوط به ضرایب کپسترال بدست آمده با هر یک از این سه روش در حالت بدون نویز و در حالتی که سیگنال گفتار با نسبت‌های سیگنال به نویز مختلف با نویز سفید جمع شده است در جدول ۶-۶ آورده شده است.

جدول ۶- ۶ نرخ بازشناخت مربوط به روش پیشنهادی دوم (DPAS) با سه رابطهٔ مشتق متفاوت.

SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB
DPAS1	۷۶/۳۳	۷۶/۵۹	۷۵/۱۹	۷۲/۱۶	۶۶/۲۹	۴۷/۷۰	۲۰/۹۱
DPAS2	۷۷/۳۳	۷۶/۹۰	۷۶/۳۳	۷۳/۱۷	۶۷/۳۰	۴۷/۳۹	۱۹/۵۵
DPAS3	۷۱/۹۴	۷۰/۹۳	۶۹/۰۹	۶۵/۸۵	۵۸/۷۹	۴۲/۶۱	۱۷/۱۰

همانطور که از روی جدول بالا قابل مشاهده است رابطهٔ ۶-۴ (DPAS2) دارای نتایج بهتری نسبت به دو روش دیگر است. بنابراین در تمام پیاده‌سازی‌های انجام شدهٔ بعدی مربوط به این روش پیشنهادی از رابطهٔ ۶-۴ استفاده شده است. برای اختصار در نمایش‌های بعدی عدد ۲ در DPAS2 حذف شده است.

۶-۵-۲ مقایسهٔ کارآیی روش DPAS با بعضی از روش‌های دیگر

در این بخش نرخ بازشناخت گفتار بدست آمده به وسیلهٔ ضرایب کپسترال استخراج شده با روش پیشنهادی دوم با نرخ بازشناخت بدست آمده با روش‌های MFCC، DPS، AMFCC و PAC مقایسه شده است. برای بررسی میزان مقاومت ویژگی‌های جدید در مقابل نویز، نویزهای سفید، کارخانه، همهمه و F16 با توان‌های متفاوت به سیگنال گفتار تمیز اضافه شده است.

جدول‌های ۶-۷ تا ۶-۱۰ نتایج بدست آمده مربوط به نرخ بازشناخت روش پیشنهادی دوم و روش‌های دیگر را به ترتیب در حضور نویزهای سفید، کارخانه، همهمه و F16 نشان می‌دهند. در این جدول‌ها ستون اول نوع روش استخراج ویژگی را مشخص می‌کند و سطر اول میزان نسبت سیگنال به نویز بر حسب دسی بل می‌باشد.

جدول ۶- ۷ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی و روش پیشنهادی دوم (DPAS) در حضور نویز سفید با توان‌های متفاوت.

SNR	Clean	30dB	25dB	20dB	15dB	10dB	5dB	0dB
MFCC	۷۶/۹۰	۷۵/۶۷	۷۳/۵۶	۶۹/۴۴	۵۸/۶۱	۳۸/۷۵	۱۸/۴۶	۵/۷۹
DPS	۷۷/۲۰	۷۵/۷۱	۷۳/۹۶	۶۹/۹۳	۶۰/۰۲	۴۰/۰۷	۱۹/۹۰	۷/۴۱
AMFCC	۷۶/۵۹	۷۵/۲۳	۷۴/۴۰	۷۰/۶۳	۶۱/۴۲	۴۰/۶۸	۱۹/۷۷	۶/۴۰
PAC	۶۷/۸۶	۶۷/۴۷	۶۷/۵۶	۶۶/۵۵	۵۸/۸۸	۲۷/۸۸	۱/۸۰	۰/۲۶
DPAS (روش پیشنهادی ۲)	۷۷/۳۳	۷۶/۹۰	۷۶/۳۳	۷۳/۱۷	۶۷/۳۰	۴۷/۳۹	۱۹/۵۵	۳/۵۵

جدول ۶- ۸ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی و روش پیشنهادی دوم (DPAS) در حضور نویز کارخانه با توان‌های متفاوت.

0dB	5dB	10dB	15dB	20dB	25dB	30dB	Clean	SNR
۹/۲۹	۳۰/۴۷	۵۷/۹۱	۷۰/۴۵	۷۴/۶۲	۷۵/۹۸	۷۶/۲۴	۷۶/۹۰	MFCC
۱۱/۵۳	۳۶/۴۸	۶۱/۲۰	۷۱/۰۲	۷۴/۹۷	۷۶/۵۵	۷۶/۶۳	۷۷/۲۰	DPS
۱۱/۸۸	۳۸/۸۴	۶۳/۰۰	۷۲/۲۵	۷۴/۷۵	۷۶/۲۸	۷۶/۶۸	۷۶/۵۹	AMFCC
۲۶/۶۵	۵۷/۳۰	۶۶/۴۶	۶۶/۹۰	۶۷/۳۴	۶۷/۷۳	۶۷/۷۳	۶۷/۸۶	PAC
۸/۵۹	۳۷/۴۰	۶۷/۶۵	۷۵/۱۴	۷۷/۶۰	۷۸/۱۷	۷۸/۰۴	۷۷/۳۳	DPAS (روش پیشنهادی ۲)

جدول ۶- ۹ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی و روش پیشنهادی دوم (DPAS) در حضور نویز همهمه با توان‌های متفاوت.

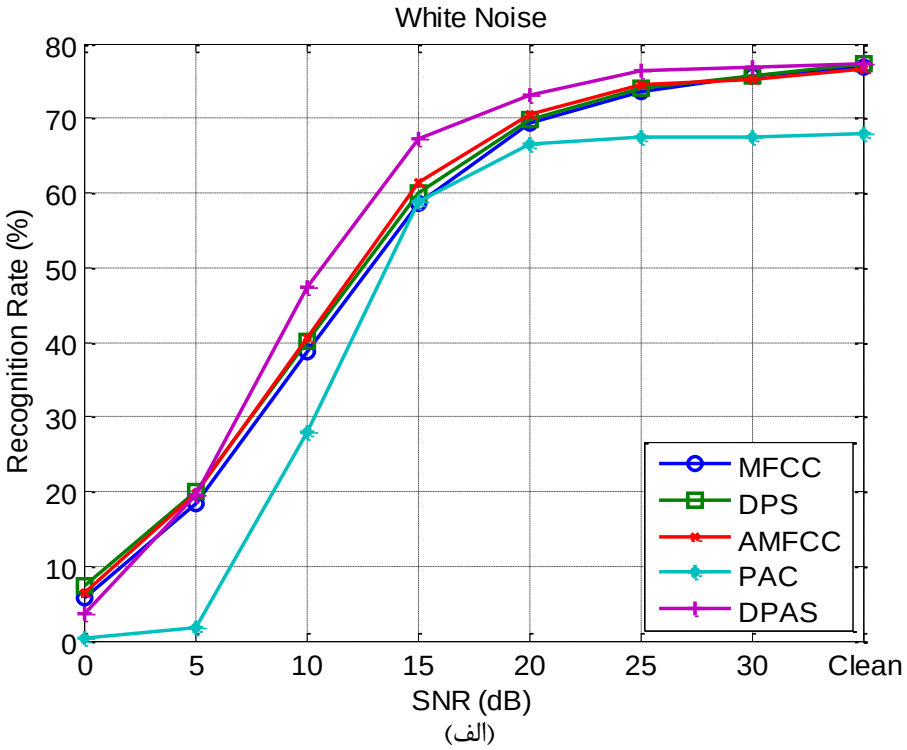
0dB	5dB	10dB	15dB	20dB	25dB	30dB	Clean	SNR
۱۷/۰۱	۴۳/۸۴	۶۴/۸۴	۷۲/۷۳	۷۵/۳۶	۷۶/۵۰	۷۶/۸۱	۷۶/۹۰	MFCC
۱۷/۹۷	۴۵/۷۷	۶۴/۷۵	۷۱/۹۰	۷۵/۳۲	۷۷/۲۹	۷۷/۲۵	۷۷/۲۰	DPS
۲۱/۰۰	۴۸/۹۷	۶۶/۴۶	۷۲/۲۹	۷۵/۰۵	۷۵/۶۲	۷۶/۵۰	۷۶/۵۹	AMFCC
۳۲/۳۱	۶۱/۲۰	۶۷/۷۳	۶۷/۰۸	۶۷/۳۰	۶۷/۵۱	۶۷/۷۸	۶۷/۸۶	PAC
۱۳/۸۵	۴۵/۱۱	۷۰/۳۲	۷۵/۶۲	۷۶/۹۸	۷۸/۲۱	۷۷/۷۷	۷۷/۳۳	DPAS (روش پیشنهادی ۲)

جدول ۶- ۱۰ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی و روش پیشنهادی دوم (DPAS) در حضور نویز F16 با توان‌های متفاوت.

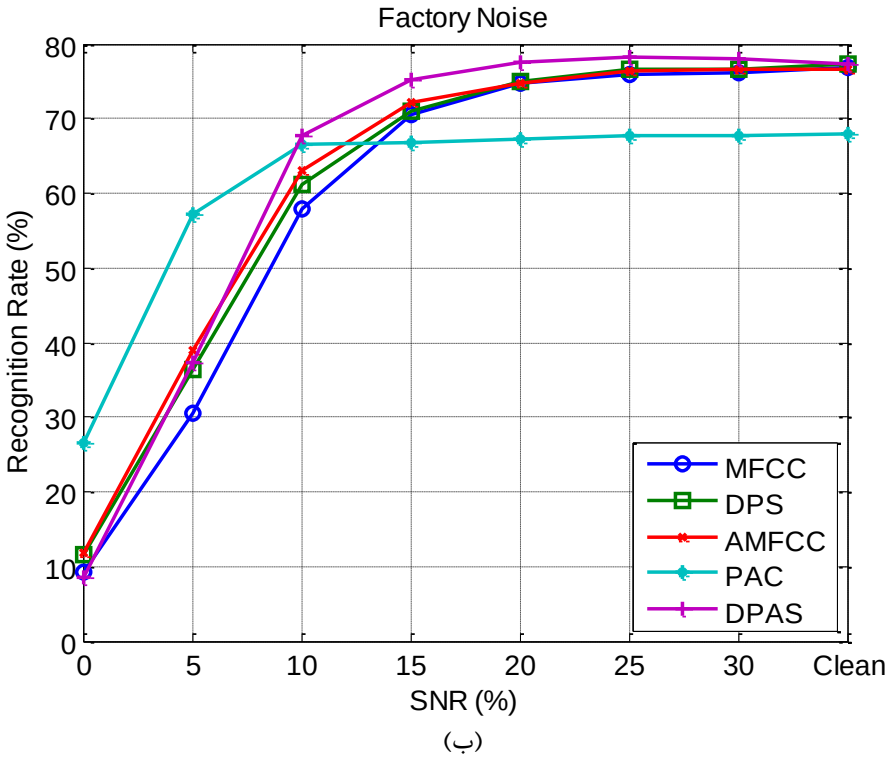
0dB	5dB	10dB	15dB	20dB	25dB	30dB	Clean	SNR
۱۳/۴۶	۳۷/۶۲	۵۸/۷۵	۶۷/۹۵	۷۳/۰۸	۷۵/۰۵	۷۶/۲۴	۷۶/۹۰	MFCC
۱۴/۱۶	۳۹/۲۸	۵۹/۹۷	۶۷/۹۱	۷۲/۹۵	۷۵/۸۴	۷۶/۱۹	۷۷/۲۰	DPS
۱۴/۸۶	۴۱/۲۱	۶۰/۵۹	۶۹/۲۲	۷۳/۷۰	۷۵/۸۴	۷۶/۵۵	۷۶/۵۹	AMFCC
۱۹/۰۷	۵۲/۷۴	۶۵/۰۲	۶۶/۶۴	۶۷/۴۷	۶۷/۶۹	۶۷/۸۶	۶۷/۸۶	PAC
۱۰/۰۰	۴۰/۲۵	۶۴/۱۸	۷۲/۳۴	۷۶/۰۲	۷۷/۳۸	۷۷/۶۹	۷۷/۳۳	DPAS (روش پیشنهادی ۲)

با توجه به جدول‌های ۶-۷ تا ۶-۱۰ مشاهده می‌شود که روش پیشنهادی دوم در تمام حالتها به غیر از نسبت‌های سیگنال به نویز ۵ و صفر دسی بل (و در یک حالت ۱۰ دسی بل) نتایج بهتری را

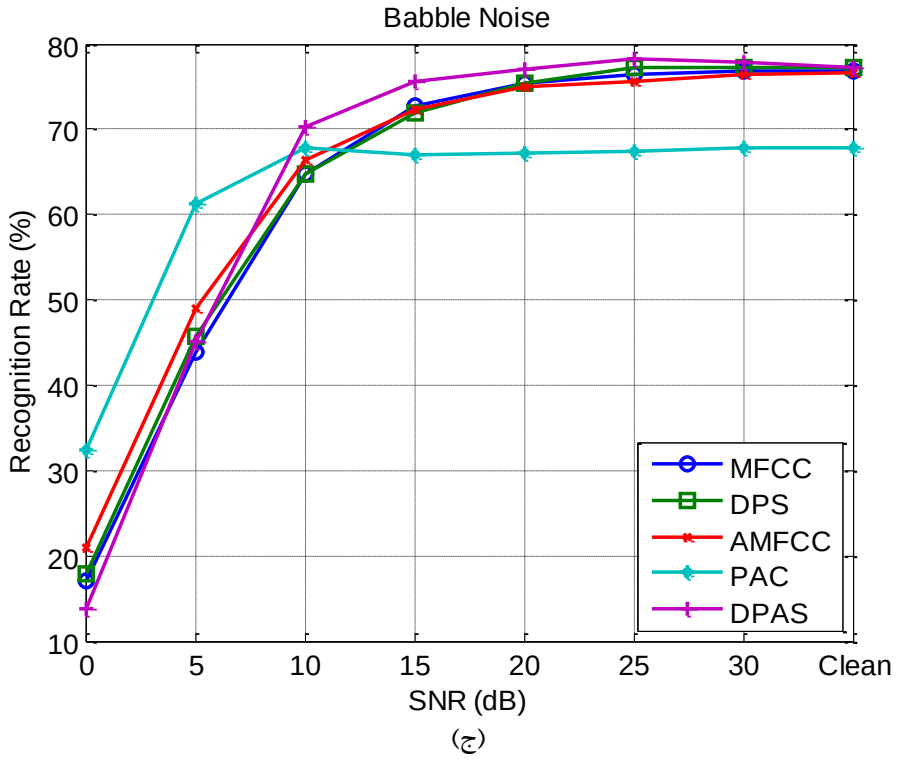
بدست آورده است. برای راحتی مقایسه، نتایج موجود در جدول‌های ۶-۷ تا ۶-۱۰ به ترتیب در شکل‌های ۶-۲-الف تا ۶-۲-ب به صورت نمودار نشان داده شده است.



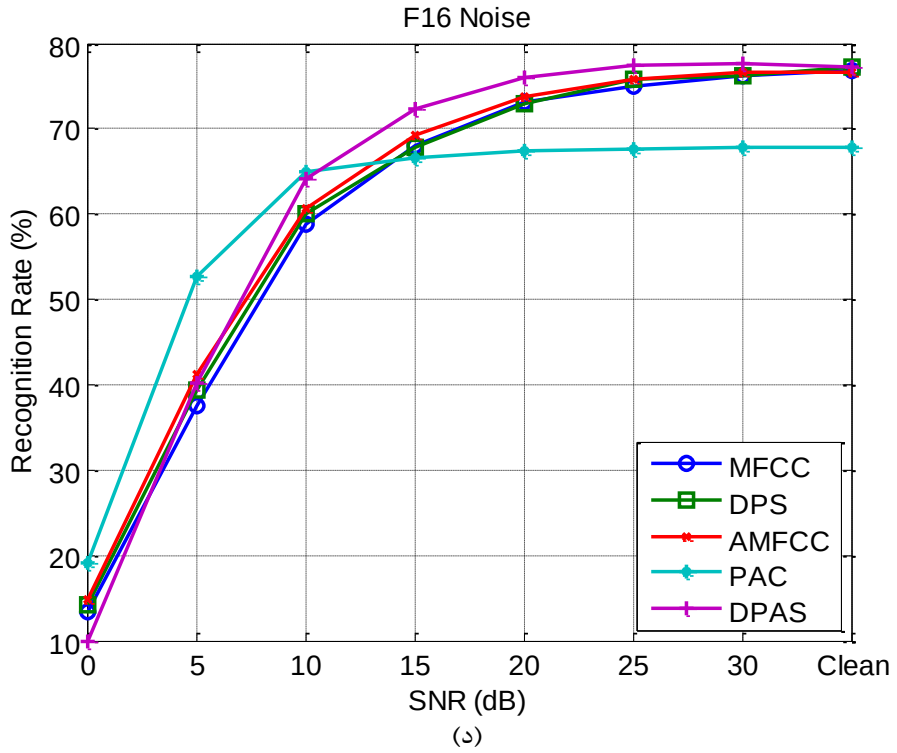
(الف)



(ب)



(ج)



(د)

شکل ۶- ۲ نرخ بازشناخت مربوط به روش‌های مختلف استخراج ویژگی و روش پیشنهادی دوم (DPAS) در حضور نویز با توان‌های متفاوت. (الف) نویز سفید (ب) نویز کارخانه (ج) نویز همهمه (د) نویز F16.

همانطور که از روی شکل‌ها و جدول‌ها ملاحظه می‌شود روش پیشنهادی تقریباً در تمام موارد دارای نتیجهٔ بهتری نسبت به بقیهٔ روش‌ها می‌باشد. همانطورکه از روی شکل‌ها قابل مشاهده است روش پیشنهادی دوم در نویزهای کمتر از ۱۰ دسی بل دارای نرخ بازشناخت کمتری نسبت به روش PAC می‌باشد (به غیر از نویز سفید). علت این امر می‌تواند به خاطر زیاد شدن دامنهٔ نویز و تأثیر عمل مشتق گیری بر روی آن باشد. ولی در بقیهٔ موارد خصوصاً در حالتی که از نویز سفید برای نویزی کردن سیگنال استفاده شده است و در حالت سیگنال بدون نویز نتایج خوبی نسبت به روش PAC و بقیه روش‌ها بدست آمده است.

۶-۶ نتیجه گیری

با توجه به آزمایشات انجام شده و بررسی نرخ بازشناخت بدست آمده از روشهای پیشنهادی اول و دوم و همچنین مقایسهٔ آنها با نرخ بازشناخت مربوط به برخی از روش‌های متداول می‌توان به نتایج زیر رسید.

با مقایسهٔ نرخ بازشناخت روش‌های مختلف استخراج ویژگی ذکر شده در این فصل ملاحظه می‌گردد که در حالت بدون نویز و همچنین در حالتی که سیگنال نویز سفید با توان‌های مختلف به سیگنال گفتار اضافه شده است بهترین نرخ بازشناخت مربوط به روش پیشنهادی دوم می‌باشد. بنابراین اگر بخواهیم از یک سیستم بازشناخت گفتار در محیطی استفاده کنیم که نویز موجود در آن محیط بیشتر به نویز سفید شبیه می‌باشد و یا اینکه هیچ نویزی وجود ندارد بهترین گزینه برای استخراج ویژگی روش پیشنهادی دوم می‌باشد.

با بررسی نرخ بازشناخت مربوط به حالتهایی که سیگنال گفتار با نویزهای کارخانه، همهمه و F16 آلوده شده است نتیجه می‌گردد، در شرایطی که میزان توان سیگنال گفتار به توان سیگنال نویز

بیشتر از ۱۰ دسی بل است بهترین نرخ بازشناخت مربوط به روش پیشنهادی دوم است و در حالتی که SNR کمتر از ۱۰ دسی بل باشد بهترین نرخ بازشناخت مربوط به روش PAC است. لذا در صورتی که سیستم بازشناخت مورد نظر ما جهت استفاده در محیط‌هایی می‌باشد که آلوده به یکی از نویزهای کارخانه، همهمه و F16 است و نسبت توان سیگنال به توان نویز نیز بیشتر از ۱۰ دسی بل می‌باشد بهترین گزینه برای استخراج ویژگی روش پیشنهادی دوم است. در حالیکه اگر نسبت توان سیگنال به توان نویز کمتر از ۱۰ دسی بل باشد بهترین گزینه روش PAC خواهد بود.

۶-۷ پیشنهادها برای کارهای آینده

برای بهبود نتایج انجام تحقیقات بیشتر ضروری می‌باشد. که در ادامه پیشنهادهایی آورده شده است:

- بررسی رابطه‌های بیشتری برای مشتق گیری (با استفاده از رابطهٔ کلی ۶-۲) در روش‌های پیشنهادی اول و دوم.
- بررسی تأثیر پنجره‌های دیگر مخصوصاً در روش پیشنهادی دوم.
- استفاده از روش‌های پیشنهادی برای بدست آوردن ضرایب PLP.
- بررسی اثر حذف ضرایب اتوکرولیشن با تأخیر زمانی کم در روش پیشنهادی دوم.
- بررسی اثر حذف ضرایب اتوکرولیشن با تأخیر زمانی کم و اعمال پنجرهٔ مناسب در روش‌های RAS و DAS.

مراجع

[1]. F. Jelinek (**1997**), "**Statistical Methods for Speech Recognition**", The MIT Press, USA.

[2]. B. H. Juang and L. R. Rabiner (2006), Speech Recognition, Automatic: History, pp 806-819, In: "**Encyclopedia of Language & Linguistics**", Vol. 11, Second Edition, K. Brown , Elsevier.

[3]. H. Perez-Meana (**2007**), "**Advances in Audio and Speech Signal Processing: Technologies and Applications**", The Idea Group Publishing, USA.

[4]. K. H. Davis, R. Biddulph, and S. Balashek (**1952**), "Automatic Recognition of Spoken Digits", **J. Acoust. Soc. Am.**, Vol. 24, No. 6, pp. 637-642.

[5]. S. Furui (**2005**) "50 Years of Progress in Speech and Speaker Recognition Research", **ECTI Trans. Computer Information Technology**, Vol. 1, No. 2, pp. 64-74

[6]. J. W. Forgie and H. Belar (**1959**), "Results Obtained from a Vowel Recognition Computer Program", **J. Acoust. Soc. Am.**, Vol. 31, No. 11, pp. 1480-1489.

[7]. J. Sakai and S. Doshita (1962), "The Phonetic Typewriter", Proc. IFIP Congress, Munich.

[8]. H. Sakoe and S. Chiba (**1978**), "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-26, No. 1, pp. 43-49.

[9]. A. J. Viterbi (**1967**), "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", **IEEE Trans. Information Theory**, Vol. IT-13, No. 2, pp. 260-269.

[10]. F. Itakura (**1975**), "Minimum Prediction Residual Principle Applied to Speech Recognition", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-23, No. 1, pp. 67-72.

[11]. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon (**1979**), "Speaker Independent Recognition of Isolated Words Using Clustering Techniques", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-27, No. 4, pp. 336-349.

[12]. J. K. Baker (**1975**), "The DRAGON System–An Overview", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-23, No. 1, pp. 24-29.

[13]. F. Jelinek, L. R. Bahl, and R. L. Mercer (**1975**), "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech", **IEEE Trans. Information Theory**, Vol. IT-21, No. 3, pp. 250-256.

[14]. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon (1979), "Speaker Independent Recognition of Isolated Words Using Clustering Techniques", ICASSP, Vol. 4, pp. 574-577.

[15]. B. H. Juang, S. E. Levinson and M. M. Sondhi (**1986**), "Maximum Likelihood Estimation for Multivariate Mixture Observation of Markov Chains", **IEEE Trans. Information Theory**, Vol. IT-32, No. 2, pp. 307-309.

[16]. J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman (**1990**), "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-38, No. 11, pp. 1870-1878.

[17]. S. E. Levinson, L. R. Rabiner and M. M. Sondhi (**1983**), "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process", **Bell Syst. Tech. J.**, Vol. 62, No. 4, pp. 1035-1074.

[18]. L. R. Rabiner, S. E. Levinson and M. M. Sondhi (**1983**), "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition", **Bell Syst. Tech. J.**, Vol. 62, No. 4, pp. 1075-1105.

[19]. L. R. Rabiner, (**1989**), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" **Proceedings of the IEEE**, Vol. 77, No. 2, pp. 257-286.

[20]. S. Furui (**1986**), "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-34, No. 1, pp. 52-59.

[21]. K. F. Lee, H. W. Hon and R. Reddy (**1990**), "An Overview of the SPHINX Speech Recognition System", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-38, No. 1, pp. 35-45.

[22]. Y. L. Chow, et. al. (1987), "BYBLOS, the BBN Continuous Speech Recognition System", ICASSP, pp. 89-92.

[23]. M. Weintraub et. al. (1989), "Linguistic Constraints in Hidden Markov Model Based Speech Recognition", ICASSP, pp. 699-702, UK.

[24]. D. B. Paul (1989), "The Lincoln Robust Continuous Speech Recognizer", ICASSP, pp. 449-452, UK.

[25]. V. Zue, J. Glass, M. Phillips and S. Seneff (1989), "The MIT Summit Speech Recognition System, a Progress Report", DARPA Speech and Natural Language Workshop, pp. 179-189.

[26]. C. H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon (**1990**), "Acoustic Modeling for Large Vocabulary Speech Recognition", **Computer Speech & Language**, Vol.4, No. 2, pp. 127-165.

[27]. B. H. Juang, C. H. Lee and Wu Chou (**1997**), "Minimum Classification Error Rate Methods for Speech Recognition", **IEEE Trans. Speech and Audio Proc.**, Vol. 5, No. 3, pp. 257-265.

[28]. L. R. Rabiner and B. H. Juang (2008), Historical Perspective of the Field of ASR/NLU, In: "**Springer Handbook of Speech Processing**", J. Benesty, M. M. Sondhi and Y. Huang, Springer.

[29]. S. Young et. al. (**2006**), "**The HTK Book**", Cambridge University Engineering Department.

[30]. Carnegie Mellon University: http://cmusphinx.soundforge.net/sphinx4/

[31]. J. Piccone: http://www.isip.msstate.edu/projects/speech/ (Mississippi State University)

[32]. C. J. Leggetter and P. C. Woodland (**1995**), "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", **Computer Speech and Language**, Vol. 9, pp. 171-185.

[33]. A. P. Varga and R. K. Moore (1990), "Hidden Markov Model Decomposition of Speech and Noise", ICASSP, pp. 845-848.

[34]. M. J. F. Gales and S. J. Young (**1996**), "Robust Continuous Speech Recognition Using Parallel Model Combination", **IEEE Trans. Speech and Audio Proc.**, Vol. 4, No. 5, pp. 352-359.

[35]. D. Mansour and B. H. Juang (**1989**), "The Short-Time Modified Coherence Representation and Noisy Speech Recognition", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-37, No. 6, pp. 795-804.

[36]. A. Waibel and K. F. Lee (**1990**), "**Readings in Speech Recognition**", Morgan Kaufmann Publishers Inc., USA.

[37]. J. R. Deller, J. H. L. Hansen and J. G. Proakis (**2000**), "**Discrete-Time Processing of Speech Signals**", Wiley-IEEE Press, USA.

[38]. L. R. Rabiner and B. H. Juang (2006), Speech Recognition: Statistical Methods, pp 1-18, In: "**Encyclopedia of Language & Linguistics**", Vol. 12, Second Edition, K. Brown , Elsevier.

[39]. B. J. Shannon (2006), PhD. thesis, "Speech Recognition and Enhancement using Autocorrelation Domain Processing", School of Engineering, Griffith University, Brisbane, Australia.

[40]. J. W. Picone (**1993**), "Signal Modeling Techniques in Speech Recognition", **Proceedings of the IEEE**, Vol. 81, No. 9, pp. 1215-1247.

[41]. J. Makhoul (**1975**), "Linear Prediction: A Tutorial Review", **Proceedings of the IEEE**, Vol. 63, No. 4, pp. 561-580.

[42]. L. R. Rabiner and R. W. Schafer (**1978**), "**Digital Processing of Speech Signals**", Prentice Hall, USA.

[43]. L.R. Rabiner and B. H. Juang (**1993**), "**Fundamentals of Speech Recognition**", Prentice Hall, USA.

[44]. X. Huang, A. Acero and H. W. Hon (**2001**), "**Spoken Language Processing: a Guide to Theory, Algorithm, and System Development**", Prentice Hall, USA.

[45]. S. B. Davis and P. Mermelstein (**1980**), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-28, No. 4, pp. 357-366.

[46]. A. Koc (2002), MS thesis, "Acoustic Feature Analysis for Robust Speech Recognition", Bogazici University, Turkey.

[47]. M. Rosell (2006), “An Introduction to Front-End Processing and Acoustic Features for Automatic Speech Recognition”, Term Paper in Swedish National Graduate School of Language Technology.

[۴۸]. م. رحمانی، (۱۳۸۰)، پایان نامهٔ ارشد، "اعمال روشهای بهبود گفتار به عنوان پیش پردازش، جهت بالا بردن دقت بازشناسی گفتار"، دانشکده کامپیوتر، دانشگاه علم و صنعت ایران.

[49]. H. Hermansky (**1989**), “Perceptual Linear Predictive (PLP) Analysis of Speech”, **J. Acoust. Soc. Am.**, Vol. 87, No. 4, pp. 1738-1752.

[50]. B. Milner (2002), “A Comparison of Front-End Configurations for Robust Speech Recognition”, ICASSP, pp. 797-800.

[51]. J. Psutka, L. Muller and J. V. Psutka (2001), “Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task”, Eurospeech, Scandinavia.

[52]. S. Young (**1996**), “A Review of Large-Vocabulary Continuous-Speech Recognition”, **IEEE Signal Processing Magazine**, pp. 45-57.

[53]. L. R. Bahl, F. Jelinek and R. L. Mercer (**1983**), “A Maximum Likelihood Approach to Continuous Speech Recognition”, **IEEE Trans. Pattern Analysis & Machine Intelligence**, Vol. PAMI-5, No. 2, pp. 179-190.

[54]. D. Yuk (1999), PhD. thesis, “Robust Speech Recognition using Neural Networks and Hidden Markov Models”, New Brunswick, New Jersey.

[55]. M. J. F. Gales (1995), PhD. thesis, “Model-Based Techniques for Noise Robust Speech Recognition”, Gonville and Caius College, Cambridge University.

[56]. H. Hermansky and N. Morgan (**1994**), “RASTA Processing of Speech”, **IEEE Trans. Speech and Audio Proc.**, Vol. 2, No. 4, pp. 578-589.

[57]. C. Kermorvant (1999), “A Comparison of Noise Reduction Techniques for Robust Speech Recognition”, IDIAP RR 99-10.

[58]. J. Hernando and C. Nadeu (**1997**), “Linear Prediction of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition”, **IEEE Trans. Speech and Audio Proc.**, Vol. 4, No. 1, pp. 80-84.

[59]. K. H. Yuo and H. C. Wang (1998), “Robust Features Derived from Temporal Trajectory Filtering for Speech Recognition Under the Corruption of Additive and Convolutional Noises” ICASSP, Vol. 1, pp. 577-580.

[60]. S. F. Boll (**1979**), “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-27, No. 2, pp. 113-120.

[61]. K., Hermus and P. Wambacq (2004), “Assessment of Signal Subspace Based Speech Enhancement for Noise Robust Speech Recognition” ICASSP, pp. 945-948.

[62]. D. Mansour and B. H. Juang (**1989**), “A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition”, **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-37, No. 11, pp. 1659-1671.

[63]. H. Uguz, A. Arslan and I. Turkoglu (**2007**), “A Biomedical System Based on Hidden Markov Model for Diagnosis of the Heart Valve Diseases”, **Pattern Recognition Letters**, Vol. 28, No. 4, pp. 395-404.

[64]. M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar (**2001**) “Handwritten Farsi (Arabic) Word Recognition: A Holistic Approach Using Discrete HMM”, **Pattern Recognition**, Vol. 34, No. 5, pp. 1057-1065.

[65]. R.O. Duda, P. E. Hart and D. G. Stork (**2001**), “**Pattern Classification**”, Second Edition, Wiley-Interscience, Canada.

[۶۶]. ه. کلانتری، (۱۳۷۹)، پایان نامهٔ ارشد، "بهبود دقت بازشناسی واجهای درون طبقه‌ایی در گفتار پیوسته فارسی"، گروه مهندسی برق و کامپیوتر، دانشکده فنی، دانشگاه تهران.

[67]. M. Gales and S. Young (**2008**), “**The Application of Hidden Markov Models in Speech Recognition**”, now Publishers Inc., USA.

[68]. S. Young (2008), HMMs and Related Speech Recognition Technologies, In: “**Springer Handbook of Speech Processing**”, J. Benesty, M. M. Sondhi and Y. Huang, Springer.

[69]. S. J. Young, J. J. Odell and P. C. Woodland (1994), “Tree-Based State Tying for High Accuracy Acoustic Modelling”, In Proc. of Human Language Technology Workshop, pp. 307-312.

[70]. M. A. Comez (2003), MS thesis, “Large Vocabulary Continuous Speech Recognition for Turkish Using HTK”, School of Natural and Applied Sciences, Middle East Technical Univ.

[71]. J. Chen, K. K. Paliwal and S. Nakamura (**2003**), "Cepstrum Derived from Differentiated Power Spectrum for Robust Speech Recognition", **Speech Communication**, Vol. 41, pp. 469-484.

[72]. J. Hernando and C. Nadeu (1994), "Speech Recognition in Noisy Car Environment Based on OSALPC Representation and Robust Similarity Measuring Techniques", ICASSP, Vol. 2, pp. 69-72.

[73]. K. H. Yuo and H. C. Wang (**1999**), "Robust Features for Noisy Speech Recognition Based on Temporal Trajectory Filtering of Short-Time Autocorrelation Sequences", **Speech Communication**, Vol. 28, pp. 13-24.

[74]. G. Farahani, S. M. Ahadi and M. M. Homayounpur (**2007**), "Features Based on Filtering and Spectral Peaks in Autocorrelation Domain for Robust Speech Recognition", **Computer Speech and Language**, Vol. 21, pp. 187-205.

[75]. B. Shannon and K. K. Paliwal (2004), "MFCC Computation from Magnitude Spectrum of Higher Lag Autocorrelation Coefficients for Robust Speech Recognition", Proc. Intern. Conf. Spoken Language Processing, Jeju, South Korea.

[76]. B. Shannon and K. K. Paliwal (**2006**), "Feature Extraction from Higher-Lag Autocorrelation Coefficients for Robust Speech Recognition", **Speech Communication**, Vol. 48, pp. 1458-1485.

[77]. S. Ikbal, H. Misra and H. Bourlard, (2003), "Phase AutoCorrelation (PAC) Derived Robust Speech Features", ICASSP, Vol. 2, pp. 133-136, Hong Kong.

[78]. N. A. Anstey (**1966**), "Correlation Techniques", **Can. J. Explor. Geophys.**, Vol. 2, No. 1, pp. 55-82.

[79]. A. V. Oppenheim, R. W. Schafer and J. R. Buck (**1999**), "**Discrete-Time Signal Processing**", Second Edition, Prentice Hall, USA.

[80]. B. J. Shannon and K. K. Paliwal (2005), "Influence of Autocorrelation Lag Ranges on Robust Speech Recognition", ICASSP, Vol.1, pp. 545-548.

[81]. S. H. Akhlagh and H. Marvi (2009), "Features Derived from Differential Power Spectrum in Autocorrelation Domain for Continuous Speech Recognition", ICSV16, Kracow, Poland.

[82]. S. H. Akhlagh and H. Marvi (2010), "Robust Features Derived from Differentiated Phase Autocorrelation Spectrum for Speech Recognition", NCC 2010, India.

[83]. S. M. Kay (1979), “The Effects of Noise on the Autoregressive Spectral Estimator”, **IEEE Trans. Acoustics, Speech and Signal Proc.**, Vol. ASSP-27, No. 5, pp. 478-485.

[84]. S. Ikbal, H. Hermansky and H. Bourlard (2003), “Nonlinear Spectral Transformations for Robust Speech Recognition”, ASRU, pp. 393.398.

[85]. SPIB (1995), “SPIB Noise Data”, Available from: <http://spib.rice.edu/spib/select _noise.html>.

Abstract

One of the most important issues in the field of speech recognition is the effect of noise on speech signal and degradation of the speech recognition rate. Therefore, robust speech recognition is an active branch of researches. In recent years many researches in this field have been done and different methods are proposed. A group of existing methods in the field of robust speech recognition is the extraction of speech features which are robust against noise. The aim of this thesis is to propose some new methods to extract robust features for speech recognition when the noise is additive to speech signal. In this thesis the autocorrelation domain processing has been used for robust feature extraction. For this purpose, after studying the most important components of speech recognition systems and explaining some of the works that done in the field of robust feature extraction in autocorrelation domain, the proposed methods are presented. The most important methods that used autocorrelation domain for robust feature extraction are RAS, DAS, AMFCC and PAC. In this thesis some new ideas for improving PAC and AMFCC methods based on using differential power spectrum and appropriate windowing has been proposed. The results obtained from the implementation of the proposed methods on the TIMIT database represents improvement in the continuous speech recognition rate compared to the previews methods.

Keywords: Speech recognition; Feature extraction; Autocorrelation; Hidden Markov model toolkit

Shahrood University of Technology
Faculty of Electrical and Robotic Engineering

Feature extraction based on autocorrelation domain processing for speech recognition using HTK

By
Seyed Hamid Akhlagh

Under supervision of
Dr. Hossein Marvi

Consultant
Dr. Omid Reza Ma'rouzi

A Thesis Presented to the Faculty of Electrical and Robotic Engineering
in partial fulfillment of the requirements for the degree of
Master of Science
in
Electronics Engineering

January, 2010