هست کلید در گنج حکیم (نظامی گنجوی)

دانشکده مهندسی برق و رباتیک

پایان نامه کارشناسی ارشد مهندسی مخابرات سیستم

تشخیص گوینده بر اساس ویژگی‌های استخراجی ترکیبی در محیط با چند گوینده

نگارنده: میترا جهانیان

استاد راهنما:

دکتر حسین مروی

استاد مشاور:

دکتر سید مسعود میر رضایی

شهریور ۱۳۹۵

پروردگارا:

به‌پاس تعبیر عظیم و انسانی‌شان از کلمه ایثار و ازخودگذشتگان

به‌پاس عاطفه سرشار و گرمای امیدبخش وجودشان که در این سردترین روزگاران بهترین پشتیبان است

به‌پاس قلب‌های بزرگشان که فریادرس است و سرگردانی و ترس در پناهشان به شجاعت می‌گراید

و به‌پاس محبت‌های بی‌دریغشان که هرگز فروکش نمی‌کند

این مجموعه را به پدر و مادر عزیزم ، برادران دوست‌داشتنی‌ام و دوستان عزیزم تقدیم می‌کنم

تقدیم به

معلمان فرهیخته‌ای که صادقانه و عاشقانه تلاش می‌کنند تا نقّالی دانسته‌ها را به نقّادی اندیشه‌ها تبدیل سازند

و دانش‌آموزانی که روح خلاق گریزپایشان، تکرار را نمی‌خواهد

تقدیر و تشکر:

با سپاس فراوان از لطف خدای مهربان.

با تشکر از دو استاد بزرگوارم که شایسته‌ی هر نوع سپاس، تجلیل و تکریم‌اند؛

جناب آقای دکتر حسین مروی؛ استاد راهنمای ارجمند که با ایجاد عشق به نوشتن، صبورانه، با ارائه‌ی رهنمودها، انتقادها و پیشنهادهایشان، در تمامی مراحل اجرای پایان‌نامه مرا حمایت و تشویق نمودند و جناب آقای دکتر سید مسعود میر رضایی؛ استاد مشاور محترم که با نظرهای اصلاحیِ ارزنده‌ی خود، ضمن دلگرمی بنده، موجب تکمیل این اثر شدند.

همچنین از جناب دکتر معروضی و دکتر سلیمانی به جهت زحمت فراوان در داوری این اثر و از استادان محترمی که در طول دوران تحصیلی‌ام ، جهت آموزش و ارتقای علمی بنده، زحمت‌کشیده‌اند سپاسگزارم.

همچنین از جناب آقای دکترگرایلو مسئول محترم آزمایشگاه پردازش سیگنال و دوستان عزیز و گرامی آقایان حسینی و غفاریان فر و خانم‌ها آزاده اصغری و یلدا عابدینی و مریم صادقی کمال تشکر رادارم و از مسئولین محترم دانشکده آقایان یونسیان و رضایی و خانم جعفری صمیمانه متشکرم.باشد که خداوند عمر باعزت و روحی متعالی نصیب تک‌تک این عزیزان نماید.

میترا جهانیان

شهریورماه سنه‌ی یک هزار و سیصد و نودوپنج خورشیدی

تعهدنامه

اینجانب **میترا جهانیان** دانشجوی دوره کارشناسی ارشد رشته **مهندسی برق مخابرات سیستم** دانشکده **برق و رباتیک** دانشگاه شاهرود نویسنده پایان‌نامه **تشخیص گوینده بر اساس ویژگی‌های استخراجی ترکیبی در محیط با چند گوینده** تحت راهنمایی **دکتر حسین مروی** و مشاوره و راهنمایی دکتر **مسعود میر رضایی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط این‌جانب انجام‌شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه 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 شاهرود می‌باشد و مقالات مستخرج بانام « دانشگاه شاهرود » و یا « Shahrood University» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده‌شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده‌شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

•

امضای دانشجو

مالکیت نتایج و حق نشر

کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته‌شده است) متعلق به دانشگاه شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.

چکیده:

امروزه در دنیای ارتباطات پردازش داده با سرعت و دقت بالا از اهمیت بسزایی برخوردار است. پردازش دادگان گفتاری نیز به دلیل کاربرد وسیع در تمامی جهات زندگی بشر سهم بسزایی را ایفا می‌کند. نمایه سازی گوینده یعنی تشخیص دهیم چه کسی چه زمانی صحبت می‌کند. هدفِ طراحی سیستم‌های نمایه سازی گوینده این است که تغییر در گوینده در فایل صوتی تشخیص داده شود و گفتار هر گوینده به‌درستی برچسب‌گذاری و دسته‌بندی شود. این فرآیند امروزه با نام Speaker Diarization شناخته‌شده است. در این پایان‌نامه، سیستمی طراحی کردیم که با استفاده از ویژگی‌های آکوستیکی MFCC[1] و مشتقات مرتبه اول و دوم آن به همراه ویژگی‌های انرژی و نرخ عبور از صفر ویژگی‌های گفتار و غیر گفتار را استخراج کند سپس با استفاده از فریم‌های قطعاً سکوت و قطعاً آهنگ به مدل‌سازی این دو می‌پردازد و در یک فرایند دوبخشی شامل حذف سکوت و حذف آهنگ به جداسازی گفتار از غیر گفتار و آهنگ موجود در فایل صوتی می‌پردازد. در ادامه با استفاده از بردارهای i در فضای برداریِ ویژگی‌ها به کاهش بُعد می‌پردازیم. در ادامه برای تشخیص تغییر در گوینده معیار فاصله را به‌کاربرده و با خوشه‌بندی توسط برنامه‌ریزی خطی عدد صحیح (ILP[2]) گفتار هر گوینده را جدا و برچسب‌گذاری و خوشه‌بندی می‌کنیم. به بهینه‌سازی پارامترها پرداختیم. پایگاه داده‌ی مورداستفاده AMI corpus می‌باشد. نتایج خوبی در خطای بازشناسی گوینده (DER) گزارش داده شد.

کلید واژگان: نمایه سازی گوینده، پردازش گفتار، برنامه عدد صحیح خطی، بردارهای i

[1] Mel Frequency Cepstral Coefficient
[2] Integer Linear Programming

فهرست مطالب

فهرست اشکال:

فهرست جدول‌ها:

فهرست اختصارات

MFCC	mel frequency cepsteral coefficient	ضرایب کپسترال فرکانسی در مقیاس مل
LPC	linear prediction coeficient	ضریب پیشگویی خطی
LFCC	linear frequnecy cepstral coeficient	ضریب کپسترال فرکانسی خطی
ZCR	zero crossing rate	نرخ عبور از صفر
CLR	cross likelihood ratio	نرخ احتمال متقابل
NCLR	normalized cross likelihood ratio	نرخ احتمال متقابل نرمال
BIC	bayesian inference criterion	معیار استنتاج بیزین
GMM	gusian mixture model	مدل مخلوط گوسی
ASR	activity speech recognition	تشخیص فعالیت گفتار
UBM	universal background model	مدل پس‌زمینه کلی
HAC	hierarchical agglomerative clustering	خوشه‌بندی سلسله مراتبی سنتی
ILP	integer linear program	برنامه خطی عدد صحیح
HMM	hidden marcov model	مدل مخفی مارکوف
DER	diarization error rate	نرخ خطای بازشناسی
MSR	missed speech rate	نرخ گفتار ازدست‌رفته
FASR	false alarm speech rate	نرخ هشدار اشتباه گفتار
WCCN	within class covariance normalization	کوواریانس بین کلاسی نرمال
SAD	Speech activity detection	تشخیص فعالیت گفتار

فصل ۱ : مقدمه

۱-۱- اهمیت گفتار

در دنیای مدرن امروز به‌جرئت می‌توان گفت متداول‌ترین راه ارتباطی جوامع بشری با یکدیگر استفاده از گفتار و صحبت کردن است. اهمیت گفتار در تمامی جنبه‌های زندگی انسان بروز کرده است. در تمام ارتباط‌ها اعم از مکالمه متداول بین افراد، سخنرانی‌های عمومی، کلاس‌های آموزشی، سخنرانی‌های دیپلماتیک، گفتگوهای جنگی و یا حتی مکالمات سری نیز گفتار به کار می‌رود. لذا بررسی این ابعاد از راه ارتباطی ضرورت بسیار زیادی دارد. در میان داده‌های جمع‌آوری‌شده توسط بشر گفتار سهم بسزایی را به خود اختصاص می‌دهد. مؤسسات و ارگان‌ها و حتی خود ما به جمع‌آوری این داده‌ها می‌پردازیم. کارایی این اصوات ذخیره‌شده بسیار متنوع است و کاربردهای متفاوت و اهمیت‌های مختلفی دارند. ازاین‌رو بررسی و آنالیز و واکاوی این داده‌ها خود امری جدا در حوزه‌ی علوم را شکل داده است. جستجو و بازیابی و استخراج اطلاعات دلخواه از این حجم اطلاعات نیازمند سیستم‌های کامپیوتری و نرم‌افزارهای کامپیوتری دارد. ساختاربندی و آرشیو این اطلاعات مختلف پیکربندی خاصی را دارد . گزارشات تلویزیونی و رادیوئی و مکالمات تلفنی بخش اعظم اطلاعات گفتاری را تشکیل می‌دهند. الگوهای گفتار را می‌توان به دو صورت دسته‌بندی کرد: دسته‌ی اول بر اساس اندام‌های صوتی و اندازه و شکل گلو و دهان و ویژگی‌های تارهای صوتی دسته‌بندی کرد و دسته‌ی دوم بر اساس الگوهای رفتاری همانند تحصیلات و موقعیت اجتماعی و سبک سخت گفتن دسته‌بندی می‌شوند. در اغلب قالب‌های امروزی دسته‌ی اول رایج‌تر است. تحقیقات وسیعی در این زمینه در سال‌های اخیر صورت گرفته است و نتایج بسیار خوبی حاصل‌شده است. کاربردهای متنوع این پردازش‌ها همچون تشخیص مجرم و جدا کردن صحبت‌های مهم یک شاهد و تشخیص گفتارهای متنوع از هم بر کسی پوشیده نیست.

۱-۲- پردازش گفتار

هدف از سیستم نهایی در این حوزه پاسخ به این سؤال است که چه کسی در چه زمانی صحبت می‌کند؟ این حوزه خود دارای بخش‌های مختلفی ازجمله: قطعه‌بندی گوینده[۱]، تشخیص گوینده[۲]، رونویسی قوی[۳] و اندیس گذاری قوی[۴] است. از چنین سیستم‌هایی برای جابه‌جایی راحت در داده‌های صوتی در فایل‌های صوتی که متعلق به چند گوینده است، استفاده می‌شود. هدف نهایی در این کاربردها، پیاده‌سازی روش‌هایی برای افراز فایل صوتی به نواحی است که در آن‌ها گوینده‌ای خاص صحبت می‌کند.

با افزایش تعداد مدارک متنی موجود در اینترنت، نیاز به فن‌هایی نظیر فهرست‌نگاری متن به‌منظور تسهیل دسترسی و جستجو در این مدارک افزایش پیدا کرد. نظیر همین نیاز نیز با افزایش تعداد مدارک صوتی نظیر سخنرانی‌ها و مصاحبه‌ها و گردهمایی‌ها و ... ایجاد شد. به‌طور مشخص دسترسی به مدارک صوتی بسیار سخت‌تر از دسترسی به متن است و گوش دادن به فایل صوتی ضبط‌شده بیشتر از خواندن متن زمان‌بر است و فهرست‌نگاری دستی مدارک صوتی در مقایسه با فهرست‌نگاری متن، مشکل است. راه‌حل پیشنهادی جهت رفع این مشکل فهرست‌نگاری خودکار مدارک صوتی[۵] است.

۱-۳- تشخیص گوینده

اولین بار سیستم‌های تشخیص گوینده توسط کمپانی NIST در سال ۱۹۹۹ ارائه شد[۱] . در سال ۲۰۰۱، پلکان و سیدهارون به همراه گروهشان با استفاده از کم کردن اثر نویز روی سیگنال،

---

[1] Speaker Segmentation
[2] Speaker Recognition
[3] Robust Transcription
[4] Robust indexing
[5] Automatic Audio Indexing

بهبودهایی در نتایج سیستم ایجاد کردند. در سال ۲۰۰۷، بولیان و کنی با به‌کارگیری بردارهای ویژگی دیگری و ادغام روش‌های قبلی و استفاده از مدل‌های گوسی در سیستم، نتایج متفاوتی به دست آوردند[۲] .درهمان سال با استفاده از ویژگی‌های سیگنال‌های صوتی مانند فرکانس پیچ، انرژی، فرکانس‌های ماکزیمم سیگنال و سه ویژگی دیگر توسط یاماتا و ماتسوناگا بهبود قابل‌توجهی در بخش-بندی گوینده حاصل شد. در سال‌های بعد با انجام روش‌های دیگر بر روی قسمت‌های مختلف آن تا به امروز سیستم تشخیص گوینده در حال تکمیل بوده‌اند.[۳]

هدف از این پایان‌نامه، طراحی و پیاده‌سازی سیستمی است که بتواند در یک فایل صوتی که شامل گفتار چندین گوینده است، تغییر در گوینده را تشخیص و جدا کند و تاحدامکان، گفتار هر گوینده را دسته‌بندی نماید. این سیستم می‌تواند شامل دو بخش اساسی به‌صورت زیر باشد:

- بخش‌بندی گوینده

- خوشه‌بندی گوینده

در قسمت بخش‌بندی[۱]، تقسیم سیگنال گفتاری به قسمت‌هایی که تنها شامل گفتار یک گوینده هستند، صورت می‌گیرد. در مرحله‌ی خوشه‌بندی[۲]، شناسایی و دسته‌بندی بخش‌های گفتاری مرتبط با بازشناسی یا فهرست‌نگاری[۳] گفتار در محیطی که چندین گوینده ممکن است در آن اقدام به سخن گفتن کنند، مانند یک جلسه یا یک کنفرانس و اخبار و نظایر آن، کاربرد دارد. این کار نه‌تنها می‌تواند به سیستم‌های بازشناسی گفتار پیشرفته جهت بهبود نتایج بازشناسی گروهی کمک نماید بلکه در شناسایی و متن نگاری مکالمه‌ها نیز به آن‌ها کمک می‌کند.

شکل (۱-۱) نحوه کار این سیستم را به‌خوبی نشان می‌دهد.

---

[1] Segmentation
[2] Clustering
[3] Indexing

شکل ۱-۱: نمایش بخش‌بندی گویندگان روی گفتار ورودی

در شکل (۱-۱) فایل صوتی موردبررسی یک صوت ضبط‌شده تک کاناله است که شامل چندین منبع صوتی می‌باشد. هر منبع توسط یک یا چند منبع صوتی مشابه یا متفاوت ناشی شده است و می‌تواند شامل گفتار گوینده، موسیقی، انواع نویز، مکث و... باشند. ورودی سیستم بازشناسی فایلی مشابه به شکل (۱-۱)را در بردارد . نوع و جزییات منابع صوتی موجود در فایل به ویژگی کاربردی آن فایل بستگی دارد.

به‌طورکلی سیستم‌های جداسازی گفتار در حوزه‌های مختلف پیش روی خود چالش‌های متفاوتی ازجمله کیفیت ضبط‌صوت، پهنای میکروفون، نویز، میزان و نوع منابع غیر گفتاری، تعداد گویندگان، سبک و ساختار گفتار، طول مدت گفتار و ترتیب گویندگان روبرو هستند. بنابرکاربرد هر حوزه، هرکدام مشکلات و مسائل مربوط به خود را دارا می‌باشند. در همه‌ی این سیستم‌ها تلاش بر بدست-آوردن نتایج قابل‌قبول و مناسب می‌باشد.[۴]

۱-۴- ساختار پایان‌نامه

پایان‌نامه پیش رو از ۵ فصل تشکیل‌شده است که به‌صورت ذیل شکل‌گرفته است:

فصل اول مقدمه‌ای بر پردازش سیگنال و صوت می‌باشد.در این فصل به بیان آنچه در گفتار می‌گذرد و اهمیت بررسی و تحلیل داده‌های صوتی می‌پردازیم.به نیازمندی این حوزه از پردازش

اطلاعات پرداخته و چرایی آن را موردبررسی قرار می‌دهیم.

در فصل دوم به ارائه‌ی پیشینه‌ی مبحث پیش رو می‌پردازیم. بیان آغاز ایده‌ی بازشناسی گوینده و آنچه در طی سالیان متمادی تحقیق و پژوهش دانشمندان و محققان و دانشجویان راه علم به وجود آمده است می‌پردازیم.چه ایده‌هایی ترکیب‌شده‌اند و چه چیزهایی دستخوش تغییر قرارگرفته‌اند.چه کسانی در این زمینه کارکرده‌اند و پیشگامان این عرصه چه کسانی بوده‌اند.

در ادامه به معرفی الگوریتم‌های در دسترس و روش انجام‌شده می‌پردازیم.به تعریف کار خود و الگوریتم مورداستفاده در این پایان‌نامه پرداخته و آزمایش‌های انجام‌شده در این متن را معرفی و بررسی می‌کنیم.مزایا و معایب طرح معرفی‌شده را بیان می‌کنیم .

در فصل پایانی نتیجه‌گیری‌های این آزمایش‌ها و آنچه به‌دست‌آمده و مقایسه‌های به‌دست‌آمده را ذکر کرده و به ارائه پیشنهادهای قابل انجام می‌پردازیم.

فصل ۲ :

بیان مسئله و پیشینه تحقیق

۲-۱- پیشینه بازشناسی گفتار

پردازش گفتار به‌عنوان یکی از زیرشاخه‌های پردازش سیگنال، به‌سرعت در حال گسترش است. تکنیک‌های پیچیده و نوآوری‌های روزافزون این دانش، همگی در راستای دستیابی به این آرزو هستند که امکان بیابیم مفاهیم در قالب ابزارهای ریاضی فراهم گردد. در این متن، به بیان خلاصه‌ای از انواع روش‌های بازشناسی گوینده می‌پردازیم.

هدف بلندمدت سیستم‌های بازشناسی خودکار گفتار، طراحی ماشینی است که سیگنال صوتی مربوط به یک جمله بیان‌شده را به دنباله‌ای از کلمات نوشته‌شده تبدیل نماید. سیستم‌های بازشناسی خودکار گفتار اطلاعات متنوعی از منابع دانش گوناگون را در جهت دستیابی به جمله بیان‌شده از روی سیگنال صوتی دریافت شده، به کار می‌گیرند. اما مشکلات متعددی در بازشناسی گفتار پیوسته بدون قید وجود دارد که عبارت‌اند از:

۱ .تحت تأثیر قرار گرفتن کیفیت سیگنال صوتی به‌وسیله نویز محیط و تابع انتقال سیستم انتقال مانند میکروفون، تلفن و...

۲ .عدم وضوح مرز مابین کلمات و واج‌ها در سیگنال صوتی.

۳ .تنوع وسیع سرعت بیان.

۴ .دقت ناکافی در بیان کلمات و به‌خصوص انتهای آن‌ها در گفتار محاوره‌ای نسبت به گفتار مجزا.

۵ .تأثیر تنوعات متعدد گوینده ازجمله جنسیت، شرایط فیزیولوژیک و روانی بر گفتار.

۶ .به‌کارگیری محدودیت‌های معنائی- نحوی زبان برای گفتار زبان طبیعی به روشی مشابه ارتباط انسان با انسان در سیستم بازشناسی.

۲-۲- انواع روش‌های بازشناسی گوینده

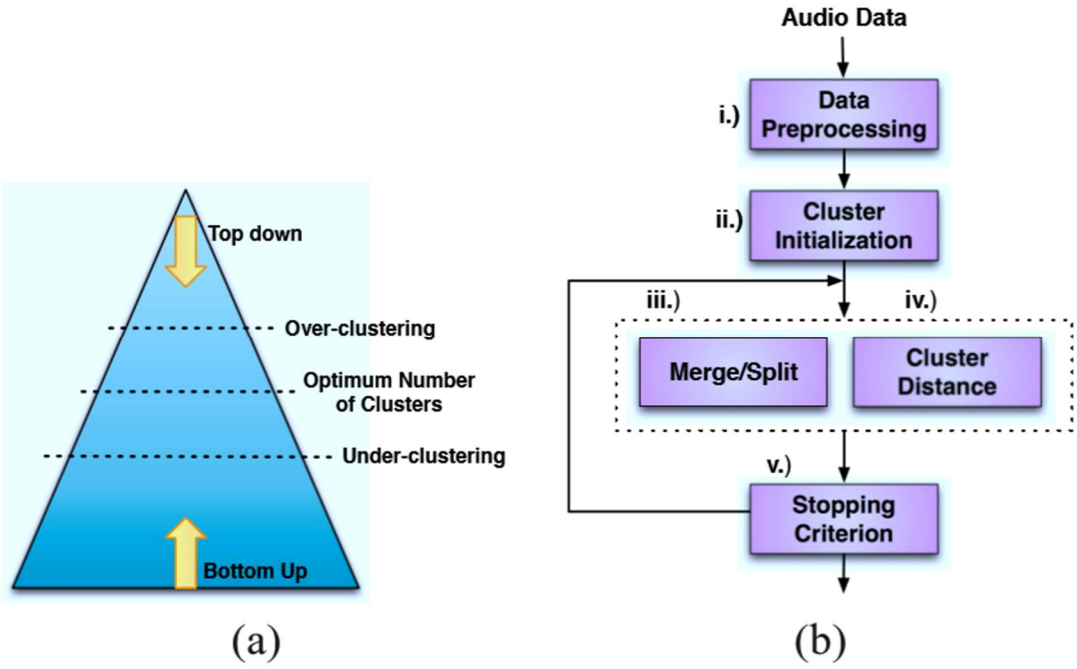
در جهت غلبه بر مشکلات مذکور تاکنون روش‌های متنوعی پیشنهادشده است که ازجمله آن‌ها روش‌های آماری مبتنی بر قانون تصمیم‌گیری بیز، روش‌های مبتنی بر شبکه عصبی و در برخی موارد ترکیب روش‌های آماری و شبکه عصبی است. با بررسی روش‌های فوق می‌توان دریافت که شناسایی کلمه یا واج بدون خطا بدون استفاده از دانش سطوح بالاتر به‌خصوص در بازشناسی گفتار پیوسته با حجم لغت‌نامه بزرگ، امکان‌پذیر نیست. به‌عنوان یک نتیجه، یک سیستم بازشناسی گفتار که با انبوهی از فرض‌ها درباره واج‌ها، کلمات و جملات مواجه است، در حالت ایده آل بایستی محدودیت‌های سطوح بالا را که به‌وسیله واژگان، نحو، معانی و ادراک مشخص می‌شود، در نظر بگیرد. در سیستم‌های مبتنی بر قانون تصمیم‌گیری بیز برخی از این محدودیت‌ها توسط مدل زبانی به سیستم بازشناسی اعمال می‌شود.

نتایج مطالعات و بررسی‌ها نشان داده است که مدل‌های زبانی‌ای که در حالت کلی توالی واحدهای زبانی را مدل می‌کنند، در کاهش خطای بازشناسی نقش عمده‌ای ایفا می‌کنند. در این میان، استفاده از مدل‌های زبانی مبتنی بر شبکه‌های عصبی باوجود قابلیت این شبکه‌ها در یادگیری زنجیره نمادها و نیز به دلیل قابلیت هموارسازی و خاصیت تعمیم دهی آن‌ها بر روش‌های آماری مزیت دارد.

سیستم های نمایه زنی گفتارشامل دو سیستم میشوند :

- سیستم بالا به پایین
- سیستم پایین به بالا

این دو سیستم در مدل سازی اولیه با یکدیگر متفاوت هستند که در شکل زیر نمای کلی این دو سیستم را میبینید :



شکل ۲-۱: سیستم کلی نمایه سازی : الف) نمایه کلی خوشه بندی کلی ب) معماری خوشه بندی نمایه سازی کلی

در شکل بالا در بخش ب روند کلی سیستم را مشاهده میکنید. سیستم بالا به پایین همانطور که دیده میشود در ابتدا خوشه بندی و مدل سازی را با تعداد بیشتری نسبت بهگویندگان احتمالی اغاز میکنیم و در ادامه تا رسیدن به نتایج ایده ال خوشه ها را حذف و ادغام میکنیم .

در سیستم پایین به بالا معمولا با یک خوشه بندی کار را آغاز میکنیم و روند خوشه سازی را تا رسیدن به نتایج ایده ال ادامه میدهیم .

۲-۳- انواع ویژگی‌های مورداستفاده در جداسازی گوینده

در سیستم‌های بازشناسی گفتار نیازمند استخراج ویژگی‌هایی برای جداسازی و شناخت گوینده‌ها هستیم. ویژگی‌های آکوستیکی‌ای که اطلاعات گوینده را در اسپکتوگرام متمایز می‌کند و در مکالمات تلفنی ثابت هستند، مدنظر می‌باشد.در استخراج ویژگی این نکته اهمیت دارد که ویژگی استخراجی به‌گونه‌ای باشد که بتواند تقریباً تمامی اطلاعات را در بربگیرد و همزمان در مقایسه با ویژگی‌های دیگر اجزا دارای مشخصه و تفاوت برای جداسازی و متمایز بودن را داشته باشد.

ضرایب کپسترال مل فرکانسی[1] یا ضرایب پیشگویی خطی[2] اگرچه برای متمایز ساختن گوینده‌ها طراحی نشده‌اند اما به‌طور وسیعی در حوزه شناسایی گوینده و تشخیص گوینده به کار گرفته می‌شوند. از زمانی که مدل‌سازی گویندگان در حوزه بازشناسی گوینده معمول شد، ضرایب کپسترال مل فرکانسی و دیگر ویژگی‌های کپسترال رایج‌ترین ویژگی‌های مورداستفاده می‌باشند.

به‌طورمعمول در طی قسمت‌بندی گویندگان ۱۳ ضرایب کپسترال مل فرکانسی در کنار انرژی زمان کوتاه استفاده می‌شود. درحالی‌که در طول خوشه‌بندی کردن گویندگان استفاده از مشتقات مرتبه بالاتر ضرایب کپسترال مل فرکانسی رایج‌تر است. تعداد ویژگی‌ها و دقت و حجم محاسباتی بر اساس تعداد و نوع ویژگی مورداستفاده افزایش پیدا می‌کند.[۵]

دیگر ویژگی به نام ضرایب کپسترال خطی فرکانسی[3] به‌دست‌آمده با فیلتر بانک خطی به‌جای فیلتر بانک در مقیاس مل[۶] و نیز ضرایب کپسترال پیشگویی خطی[4] نیز مورداستفاده قرارگرفته‌اند[۷] . اما هیچ نتیجه‌ای مبنی بر بهتر بودن کارایی این سیستم‌ها در کارهای انجام‌شده گزارش داده نشده است.

اندازه معمول پنجره تحلیل و آنالیز به‌طور میانگین در اغلب سیستم‌های پردازش ۲۵-۳۰ میلی‌ثانیه با حرکت پنجره‌ی ۱۰ میلی‌ثانیه می‌باشد.حرکت پنجره به این معنا می‌باشد که بین قسمت اول و دوم همپوشانی فاصله زمانی خاصی قرار دارد تا از دست رفتن اطلاعات در مرزهای قسمت‌ها و نیز در خطاهای پنجره‌بندی جلوگیری شود[۷].

برای سیستم شناسایی گفتار ویژگی‌های آکوستیکی که بین گفتار و غیر گفتار را متمایز می‌کنند

---

[1] Mel Frequnecy Cepstral Coeficient (MFCC)
[2] Linear Prediction Coeficient (LPC)
[3] Linear Frequnecy Cepstral Coeficient (LFCC)
[4] Linear Prediction Cepstral Coeficient (LPCC)

مورداستفاده قرار می‌گیرند. ویژگی‌هایی از قبیل انرژی [۷] و نرخ عبور از صفر[۱]، طیف مرکزی[۲]، شار طیفی[۳] [۸] در شناسایی گفتار مورداستفاده قرار می‌گرفتند.

استفاده از حداقل یکی از این بردارهای ویژگی که در بالا ذکر شدند همیشه در اتصال با ویژگی‌های کپسترال دیده‌شده‌اند.این ویژگی‌ها در کنار ویژگی‌های کپسترال در اکثر موارد موجب بهبود نتیجه و بالا رفتن دقت سیستم شناسایی می‌شود اگرچه در برخی کاربردها ترکیب ویژگی‌های مختلف برای اهداف و کاربردهای نادرست دقت را پایین می‌آورد .

جدا از ویژگی‌های آنالیز زمان کوتاه ذکرشده در بالا ویژگی‌های مدولاسیون فرکانسی ۴ هرتز که مشخصه‌های بلندمدت سیگنال آکوستیک را بیان می‌کند نیز موردبررسی قرارگرفته‌اند[۹] .این ویژگی‌ها با چالش‌های زیادی روبرو شده‌اند که درهرحال بعد بالای ویژگی‌ها و هزینه‌های محاسباتی با آن پیوند خورده‌اند. ویژگی‌های ترکیبی بلندمدت محاسبه‌شده در طول پنجره ۵۰۰ میلی‌ثانیه مانند میانه‌ی پیچ[۴]، طیف میانگین بلندمدت[۵]، انحراف فرمنتهای ۵ ام و ۴ ام[۶]، همسازها به نرخ نویز[۷]، پراکندگی فرمنت[۸] و غیره برای کلاس‌بندی سریع مورداستفاده قرار می‌گیرند [۱۰] .این ویژگی‌ها اطلاعات آوایی و دستگاه صوتی منبع را فراهم می‌کنند و شناخت گوینده بهتری نسبت به استفاده از ضرایب کپسترال مل فرکانسی رادارند.

اخیراً، اسلینی و همکارانش ویژگی‌های مشتق شده به‌عنوان فعال‌ساز لایه تنگنای یک شبکه عصبی را استفاده کرده است. شبکه عصبی مصنوعی برای مشخص کردن قسمت ۵۰۰ میلی‌ثانیه‌ای بطوریکه به گوینده‌ای متفاوت یا همسان تعلق دارد آموزش داده می‌شود [۱۱]. در کار دیگر پیشرفت

---

[1] Zero Crossing Rate(ZCR)
[2] Spectral Centroid
[3] Spectral Flux
[4] Median Pitch
[5] Long time average spectrum
[6] Deviation of the 4th and 5th Formants
[7] Harmonics to Noise Ratio
[8] Formant Dispersion

نسبی ۵۰ درصدی برای سیستم تشخیص گفتار زمانی که بیشینه فعالیت دوبعدی با یک شبکه عصبی عمیق و ۱۳ ضرایب کپسترال مل فرکانسی متصل شده‌اند گزارش داده‌شده است[۱۲] .

یک فضای ویژگی جالب دیگر که در سال ۲۰۱۰ منتشر شد، خطای بازشناسی را اندکی قربانی کرد تا به یک سرعت ۱۰ برابری با استفاده از ویژگی‌های با مقدار دودویی برای اجرای دسته‌بندی برسد[۱۳]. در این کار ویژگی‌های آکوستیکی ضرایب کپسترال مل فرکانسی با استفاده از احتمال گرفته‌شده از مدل مخلوط گوسی به یک فضای ویژگی دودویی انتقال داده شد.

۲-۴- بخش‌بندی

در بخش‌بندی گفتار هدف ساخت قطعه‌های متصل‌به‌هم و همگن از فایل صوتی می‌باشد که با قسمت همسایه تفاوت دارد. به این کار همچنین تشخیص تغییر آکوستیک[1] نیز گفته می‌شود. در ادامه به دو روش بخش‌بندی که بیشتر در بازشناسی گفتار مورداستفاده قرار می‌گیرند اشاره می‌کنیم:

۲-۴-۱- بخش‌بندی متریک[2]

یکی از رایج‌ترین روش‌های بخش‌بندی کردن صوت امروزه بخش‌بندی متریک است. این روش در بخش‌بندی آهنگ بسیار مشهور است. در بخش‌بندی متریک یک‌فاصله متریک در ابتدا بین دو بخش صوت تعریف می‌شود که شباهت آن‌ها را مشخص می‌کند. سپس یک راهکار تشخیص تغییر با استفاده از این متریک اجرا می‌شود. در مقایسه با روش بعدی (بر پایه مدل) این روش ویژگی‌های بسیاری همچون عدم نیازمندی به اطلاعات قبلی از داده‌ها را داراست.

برای بخش‌بندی آهنگ، فاصله به‌صورت مستقیم بین ویژگی‌ها محاسبه می‌شود. اگرچه در پردازش صوت ویژگی‌های استفاده‌شده(به‌صورت کلی ویژگی‌های کپسترال) برای محاسبات فاصله

[1] Acoustic Change Detection
[2] Metric Based Segmentation

برای مقایسه شباهت گوینده به دلیل تغییر کامل آن‌ها با تلفن‌ها مناسب نمی‌باشند.

برای جمع‌آوری اطلاعات گوینده از بخش‌های بلندتر، فرض می‌شود که ویژگی‌های هر بخش از یک توزیع احتمال می‌آیند. مقایسه فاصله بین این توزیع‌های احتمال با استفاده از اندازه‌گیری‌های شباهت آماری همانند دیورژانس KL[1]، نرخ شباهت متقابل[2]، معیار استنتاج بیزین[3] و غیره انجام می‌شود. معمول‌ترین روش توزیع احتمال برای مدل‌سازی قطعات بردارهای ویژگی حین بخش‌بندی گوینده توزیع چند متغیره کوواریانس کامل گوسی است.

۲-۴-۱-۱- معیار استنباط بیزین

معیار استنباط بیزین یک معیار انتخاب مدل است. این معیار آماری مدل‌های موجود را برای شرح دادن داده‌ها مقایسه می‌کند. هدف در حین این انتخاب، محاسبه‌ی هر نوع عدم تناسب است. برای یک مجموعه از بردارهای X، BIC همانند زیر تعریف می‌شود.

$$\Delta BIC = R - \lambda P \qquad (۱-۲)$$

عبارت اول احتمال داده‌های مدل را به دست می‌دهد که شامل توزیع مدل‌های قسمت‌هایی است که می‌خواهیم بین آن دو این معیار را بررسی کنیم.این بخش شامل دو عبارت احتمال و تابع توزیع جدا برای هر قسمت به همراه توزیع و احتمالی جدا برای قسمت شامل اتصال دو قطعه به یکدیگر و تشکیل یک قطعه واحد می‌باشد.به‌طور خلاصه عبارت اول را می‌توان به‌صورت زیر تعریف کرد.

$$R = \frac{N_X}{2}\log(|\Sigma_X|) - \frac{N_{X_1}}{2}\log(|\Sigma_{X_1}|) - \frac{N_{X_2}}{2}\log(|\Sigma_{X_2}|) \qquad (۲-۲)$$

[1] Kullback–Leibler divergence
[2] Cross Likelihood Ratio
[3] Bayesian Inference Criterion(BIC)

عبارت دوم شامل پارامتر تجربی λ است که به‌صورت تجربی به دست می‌آید و به‌طورمعمول عددی بین ۰.۵ تا ۲ می‌باشد که باید بهینه شود. تابع P اندازه و بعد داده‌ی آموزش مدل را مشخص می‌کند. عبارت دوم پیچیدگی مدل نام دارد و به‌صورت زیر می‌توان نشان داد.

$$P = \frac{1}{2}(\mathrm{p} + \frac{1}{2}\mathrm{p}(\mathrm{p}+1)) \qquad (۲-۳)$$

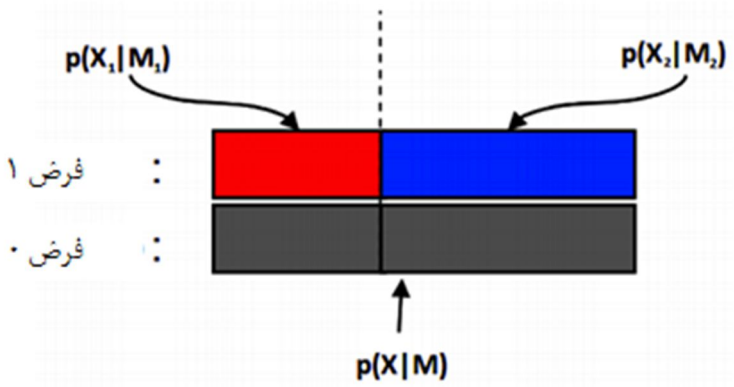
معیار بیزین برای تشخیص تشابه دو مجموعه از بردار ویژگی‌ها با توزیع یکسان یا غیر همسان بکار می‌رود. برای اندازه‌گیری میزان شباهت بین بلوک X_1 و X_2 فرض‌های زیر باید مقایسه شوند:

H_0: بردار ویژگی از X_1 و X_2 توزیع همسان دارند

H_1 : بردار ویژگیX_1 و X_2 از دو توزیع متفاوت می‌آیند.

مدل برایH_0 فرض اول M نام دارد که با داده‌ی X_1 - X و X_2 متصل شده به هم آموزش‌دیده می‌شود. مدل برای H_1 فرض دوم M1 و M_2 به ترتیب برای X_1 و X_2نام‌گذاری می‌شود.

$$\Delta BIC = BIC(M) - BIC(\mathrm{M}_1) - \mathrm{BIC}(\mathrm{M}_2) \qquad (۲-۴)$$



شکل ۲-۲: BIC برای تشخیص تغییر در گوینده

در شکل(۲-۱) فرض‌های ممکن شرح داده‌شده است .دو قطعه‌ای که مورد آزمایش هستند تا تصمیم بگیریم به یک گوینده تعلق دارند یا خیر با دورنگ و بخش با اندازه متفاوت نشان داده‌شده‌اند.تفاوت و تشابه توزیع هر دو بخش با تفاوت و تشابه رنگ در شکل مدل‌سازی شده است.یک مقدار مثبت برای ΔBIC عدم تشابه بین دو بلوک را مشخص می‌کند و این نشان‌دهنده‌ی تغییر در گوینده از بلوک X_1 به بلوک X_2 می‌باشد.

آقای چن و همکارانش یک سیستم بدون نظارت کامل را ساختند. روش آن‌ها بارها و بارها در تعداد زیادی سیستم‌های تشخیص تغییر گوینده و محیط مورداستفاده قرار گرفت[۱۴،۸].

روش آن‌ها در[۱۵] و [۱۶]با استفاده از روش BIC با کاهش محاسبات ناشی از کاهش دقت اندک به‌سرعت اجرای بالاتری دست یافتند.

روش‌های بخش‌بندی متریک گوینده در دو استراتژی اجرا می‌شوند: یک روش پنجره‌بندی ثابت و یک پنجره‌بندی گسترشی. در مورد اول، یک پنجره با اندازه ثابت داریم که برای تشخیص تغییر مرکزیت از آن استفاده می‌شود [۱۶]. اگر بردارهای ویژگی با جداسازی توزیع‌ها در یک سمت نقطه‌ی وسط بهتر مدل شود که باعث فاصله‌ی بیشتر بین توزیع‌ها می‌شود، نقطه‌ی وسط به‌عنوان نقطه‌ی تغییر انتخاب می‌شود. اندازه‌ی پنجره‌ی بخش‌بندی به‌طورمعمول ۵ ثانیه است و دو بخش ۲.۵ ثانیه‌ای برای شباهت مورداستفاده قرار می‌گیرند. با پنجره با سایز بزرگ‌تر دو بخش بهتر مدل می‌شوند. اگرچه ممکن است با این کار احتمال از دست دادن تغییر در گوینده باشد. چون ممکن است تغییر در گوینده‌ی بیشتر از یکی در پنجره تحت نظر ما وجود داشته باشد و به همین دلیل نتوان آن را تشخیص داد.

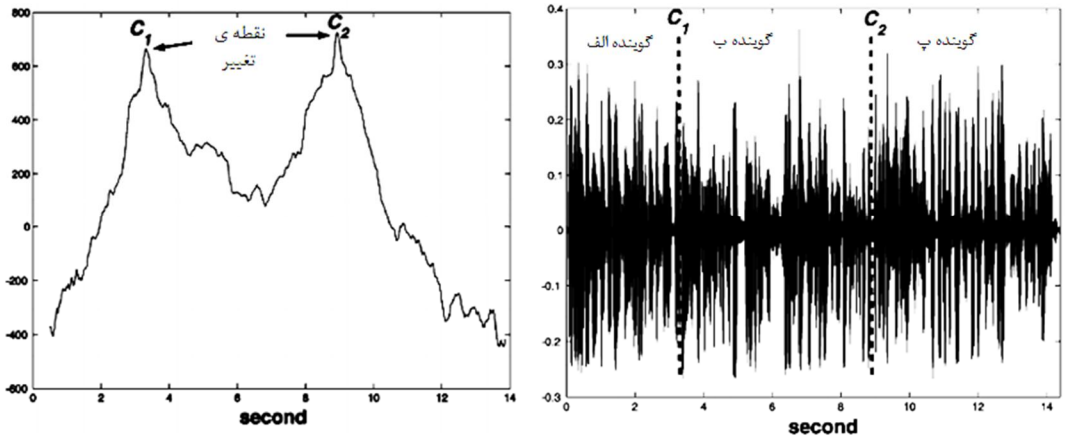
۲-۴-۲- بخش‌بندی بر اساس مدل

روش‌های بخش‌بندی گوینده بر اساس مدل برای هر کلاس بخش‌بندی یک مدل مخلوط

گوسی[1] آموزش داده می‌شود. این مدل مخلوط گوسی‌ها به‌عنوان یک تابع توزیع چگالی در مدل مخفی مارکوف استفاده می‌شود. در مدل مخفی مارکوف هر حالت به هر حالت دیگر با احتمال انتقال حالت خاصی متصل است. الگوریتم دیکد کردن ویتربی با استفاده از این مدل مخفی مارکوف یک بخش‌بندی صوت ضبط‌شده را اجرا می‌کند.

بزرگ‌ترین معایب این نوع بخش‌بندی این است که مدل مخلوط گوسی‌ها نیازمند این هستند که از قبل شناخته‌شده باشند و این بدین معنی است که باید با برخی داده‌های خارجی آموزش‌دیده شوند.

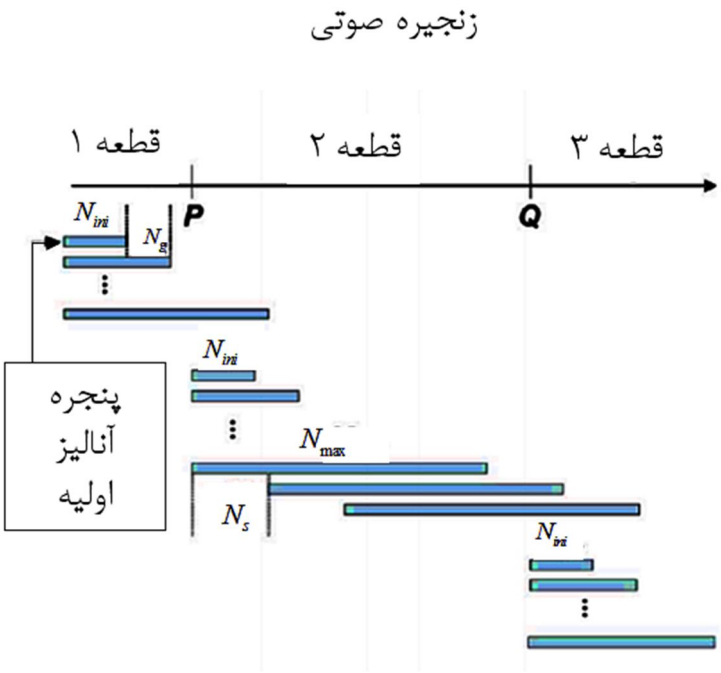
در شکل (۲-۲) مشاهده می‌کنید که برای یافتن تغییر در گوینده بین بخش‌های پشت سر هم فاصله‌ی میانه‌ی دو پنجره محاسبه می‌شود و برحسب زمان در طول پنجره رسم می‌شود.زمانی که فاصله به بیشترین حد خود می‌رسد محل تغییر در گوینده و محل بیشترین تفاوت در دو پنجره است.این خود نشان‌دهنده‌ی تغییر در ویژگی‌ها و یا همان تغییر در گوینده است.ابدیم معنی که این دو قسمت مجاور متعلق به دو گوینده‌ی متفاوت می‌باشند.



شکل ۲-۳: پنجره‌ی متغیر برای تشخیص تغییر. فاصله‌ی بین نیمه‌های پنجره‌ها محاسبه و برحسب زمان نمایش داده‌شده است. قله و اوج در فاصله نمایانگر تغییر است.

[1] GMM : Gusian Mixture Models

در شکل (۲-۳) روش جستجوی پنجره‌ای و جاروب کل قطعه نشان داده‌شده است .روش کار به این صورت است که در ابتدا پنجره‌ای با کمترین طول و طول اولیه در نظر گرفته و به جستجو برای تغییر ویژگی‌های آکوستیکی می‌پردازیم.سپس طول پنجره را اندک‌اندک در هر مرحله افزایش می‌دهیم تا زمانی که نقطه‌ی تغییر در گوینده را بیابیم.سپس جستجو را از سر گرفته و از نقطه‌ی تغییر پنجره گذاری را مجدد آغاز می‌کنیم تا به نقطه‌ی تغییر بعدی برسیم.دقت داشته باشید برای طول پنجره ماکسیمم خاصی در نظر می‌گیریم که اگر تا به این حد نقطه‌ی تغییری پیدا نشد از مرحله‌ی بعدازآن پنجره‌ی بعدی را به میزان خاصی جلوتر شروع می‌کنیم .این بدین معنی است که به میزان مشخصی از اطلاعات بررسی‌شده از نقطه‌ی تغییر قبلی را که بررسی کردیم از لیست بررسی مجدد در جستجوی در حال انجام حذف می‌کنیم و به جستجو در بین ادامه‌ی بخش‌ها می‌پردازیم.



شکل ۲-۴ : جستجوی افزایشی برای تشخیص تغییر گوینده. هر جستجو برای تنها یک تغییر است. بعد از هر یافتن نقطه تغییری جستجو از سر گرفته می‌شود.[۱۶]

در زمان بخش‌بندی و دسته‌بندی کردن پی‌درپی، خروجی مجموعه‌ای از بخش‌های گوینده را به‌عنوان داده‌ی آموزش برای مدل مخلوط گوسی‌های اجرای بعدی بکار می‌برد. این کار باعث تعریف و

نوسازی بخش‌بندی گوینده می‌شود. یک پیش دسته‌بندی به‌طورمعمول برای به دست آوردن اطلاعات دسته‌های اولیه بخش‌های گوینده [۱۰] اجرا می‌شود. هر گروه اطلاعات مشابه گوینده را نشان می‌دهد.

اغلب روش‌های بر پایه مدل برای بخش‌بندی به‌عنوان یک پس پردازش برای دستیابی به بخش‌بندی اصلاح‌شده به کار می‌رود[۱۷]. برخی از این تکنیک‌های بر پایه مدل در بخش‌بندی در تشخیص اتوماتیک گفتار مشهور هستند [۱۸] که در این کاربرد تغییر آکوستیکی در بین گفتار و غیر گفتار جستجو می‌شود.

در برخی سیستم‌های تشخیص گفتار تلفنی [۱۹] یک پیش بخش‌بندی صوت ضبط‌شده بر پایه جنسیت و پهنای باند با استفاده از مدل مخلوط گوسی آموزش داده‌شده برای هر ۴ کلاس (۲ پهنای باند* ۲ جنسیت) استفاده می‌شود.

۲-۵- تشخیص فعالیت گفتار[1]

یافتن بخش‌های همسان گفتار در یک فایل صوتی و جدا کردن آن‌ها از دیگر انواع اصوات تشخیص فعالیت گفتار نامیده می‌شود. در سیستم‌های پردازش گفتار سودمند است که بخش‌هایی از سیستم که بخش‌های شامل گفتار هستند بجای تمام فایل صوتی بررسی شود .

این امر در بهبود حجم محاسباتی و زمان پردازش و استفاده‌ی بهینه از منابع کمک شایانی می‌کند. جدا از مزیت‌های محاسباتی، در نبود سیستم تشخیص فعالیت گفتار در سیستم‌های پردازش گفتار خطاهایی وارد می‌شود. سیستم تشخیص فعالیت گفتار یک‌رویه‌ی پایه‌ای در تقریباًهمه‌ی

[1] SAD:Speech Activity Detection

بخش‌های پردازش و تشخیص و بهبود صوت می‌باشد[۱۸].

در بازشناسی گوینده، خطای متریک خود اهمیت وجود سیستم تشخیص فعالیت گفتار را نمایان می‌کند. چون هردو خطای گفتار ازدست‌رفته و هشدار خطای گفتار در خطای کلی بازشناسی لحاظ شده‌اند. از آن بیشتر با داده‌ای محدود از گوینده از بخش کوچکی از گفتار، حضور داده‌های غیر گفتاری که مدل‌های گوینده را تحت تأثیر و خطا قرار داده‌اند باعث تغییر در کارایی عملکرد سیستم بازشناسی می‌شوند.

سیستم تشخیص فعالیت گفتار اغلب خوشه‌بندی بر اساس فریم را انجام می‌دهد. مدل‌های آماری روی یک فضای ویژگی مناسب برای جداسازی گفتار و غیر گفتار آموزش می‌بینند و تخمین زده می‌شوند. در بسیاری از موارد، مدل‌های مخلوط گوسی مدل‌های آماری هستند که استفاده می‌شوند و فضای ویژگی در اغلب موارد ویژگی‌های کپسترال هستند. در برخی کارها استفاده از ویژگی‌های آکوستیکی مانند انرژی [۷] و نرخ عبور از صفر [۲۰]و شار طیفی [۸] نشان داده‌شده است.

۲-۶- خوشه‌بندی کردن

خوشه‌بندی کردن مشکل معمول در آنالیز داده‌های آماری است. در بسیاری از شاخه‌های علمی من‌جمله داده‌های اکتشافی معدن تا تشخیص جوامع در شبکه‌های اجتماعی مورد کاربرد است. خوشه‌بندی کردن فرایند گروه‌بندی کردن مجموعه‌ای از اهداف -مانند اهداف در هر گروه- که دسته نام دارند است که بیشترین شباهت را نسبت به بقیه‌ی اهداف در دیگر گروه‌ها نسبت به یکدیگر دارند. هدف‌ها باید در فضای برداری یا مدل‌های آماری نشان داده شوند. شباهت گفته‌شده در بالا یک‌فاصله‌ی تعریف‌شده بین اهداف است که توسط کاربر تعریف می‌شود.

واژه شباهت به این دلیل استفاده می‌شود که اندازه‌گیری‌های تعریف‌شده نیازمند هیچ‌گونه از معیارهایی چون غیر منفی بودن و تقارن و نامساوی مثلثی ندارد. واژه شباهت و فاصله در اینجا

به‌صورت معادل بکار می‌روند بدین منظور که فاصله‌ی کمتر نشان‌دهنده‌ی شباهت بیشتر است و برعکس.

فرایند دسته‌بندی کردن عموماً انتقالی ثابت است و جایگاه هر هدف در فضای خودش نسبت به فضای دسته‌بندی خودشان مرتبط‌تر هستند. در مسئله شناسایی گوینده، هدف دسته‌بندی بخش‌های سیگنال صوت بر اساس گوینده‌ی فعال در هر بخش است. هر دسته باید به‌صورت ایده آل یک گوینده را نشان دهد.

به‌جرئت می‌توان گفت ابعاد زمانی و فرکانسی اسپکتوگرام یک تک بخش، بزرگ است و درنتیجه مقایسه‌ی بین اسپکتوگرام دو بخش ازلحاظ محاسباتی به دلیل بسیار زیاد بودن حجم داده‌های محاسباتی غیرقابل انجام است و نمی‌توان با تکیه‌بر مقایسه‌ی دو اسپکتوگرام به نتایج دلخواه دست‌یابیم. درنتیجه نیازمند این هستیم که هر بخش به فضای با بعد کمتر برده شود تا بتوان شباهت اطلاعات گویندگانشان موردبررسی قرار بگیرد. هر بخش باید دارای یک بردار نشانه یا مدل آماری باشد که اطلاعات گوینده را دربرداشته باشد.

۲-۶-۱- مدل‌های گوینده برای خوشه‌بندی کردن

ازآنجایی‌که بازشناسی گوینده نیازمند دربرگیری اطلاعات گوینده از بخش‌های صوتی می‌باشد مدل‌های گوینده معمولاً در شناسایی گوینده و کاربردهای مرتبط بکار گرفته می‌شود.

۲-۶-۱-۱- مدل‌های گوسی مخلوط

مدل‌های گوسی مخلوط معمولاً از ویژگی‌های کپسترال برای مدل‌سازی گوینده‌ها استفاده می‌روند. یک مدل گوسی مخلوط ابزار معروفی برای مدل‌سازی داده‌های چند مدل و پردازش به‌صورت زیر است:

$$p(x) = \sum_{i=1}^{N} \omega_i N(\mu_i, \Sigma_i, \mathrm{x}) \qquad s.t. \sum_{i=1}^{N} \omega_i = 1 \qquad (۵-۲)$$

ازآنجایی‌که طول مدت بخش می‌تواند کوچک باشد تعداد بردارهای ویژگی قابل‌دسترس از هر بخش در برخی موارد برای تخمین و ساخت یک مدل گوسی کامل ناکافی است. برای حل این مشکل مدل‌های پس‌زمینه‌ای همگانی[1] از پیش آموزش‌یافته برای ساخت مدل گوینده هر بخش تنظیم‌شده است [۲۱] مدل‌های پس‌زمینه‌ای همگانی یک مدل جامع برای داده‌های ترکیب‌شده از چند گوینده که اطلاعات آن‌ها را دربردارد می‌باشد. برای مدل مخلوط گوسی از ویژگی‌های کپسترال اندازه‌گیری‌های شباهت آماری مختلفی همانند دیورژانس KL متقارن[2] و نسبت احتمال متقابل و غیره به‌کاربرده شده‌اند [۲۲].دیورژانس KL یک اندازه‌گیری تئوری اطلاعات بدین‌صورت است که دو توزیع احتمال چقدر از یکدیگر متفاوت هستند. نرخ شباهت متقابل دو مقدار $P(\mathrm{X}_1|M_2)$ و $P(\mathrm{X}_2|M_1)$ که توابع گوسی مدل‌سازی شده روی ویژگی‌های استخراجی می‌باشند، را مقایسه می‌کند.

$$CLR(\mathrm{X}_1, \mathrm{X}_2) = \log \frac{P(\mathrm{X}_1, \mathrm{M}_1)}{P(\mathrm{X}_1, \mathrm{M}_2)} + \log \frac{P(\mathrm{X}_2, \mathrm{M}_2)}{P(\mathrm{X}_2, \mathrm{M}_1)} \qquad (۶-۲)$$

$$NCLR(\mathrm{X}_1, \mathrm{X}_2) = \frac{1}{|\mathrm{X}_1|} \log \frac{P(\mathrm{X}_1, \mathrm{M}_1)}{P(\mathrm{X}_1, \mathrm{M}_2)} + \frac{1}{|\mathrm{X}_2|} \log \frac{P(\mathrm{X}_2, \mathrm{M}_2)}{P(\mathrm{X}_2, \mathrm{M}_1)} \qquad (۷-۲)$$

$\mathrm{M_i}$ مدل تخمین زده‌شده روی X_i است. همان‌طور که می‌بینند اگر بردارهای ویژگی $\mathrm{X_1}$ و $\mathrm{X_2}$ متعلق به یک گوینده باشند، مدل $\mathrm{X_1}$ به‌خوبی بر مدل $\mathrm{X_2}$ منطبق است و درنتیجه شباهت متقابل افزایش‌یافته و درنتیجه فاصله کاهش می‌یابد.همان‌طور که در (۲-۷) ملاحظه می‌کنیم عبارت فاصله‌ی CLR به‌صورت نرمال شده خود به‌عنوان معیار فاصله‌ای مجزا تعریف و مورد کاربرد است.

---

[1] Universal Background Model(UBM)

[2] Kullback–Leibler Divergence

[3] Normalized Cross Likelihood Ratio

به دلیل تغییرپذیری بالای ماتریس‌های کوواریانس و وزن‌های ترکیبی نسبت به گفته‌های [۲۱]، شاخص قابل‌اطمینانی برای مشخص کردن اطلاعات گوینده وجود ندارد. درنتیجه به‌جای محاسبه‌ی امتیازات شباهت بالا، متوسط‌های یک مدل مخلوط گوسی را با به دست آوردن یک بردار در فضای برداری بعد بالا به هم پیوند داده می‌شوند. اندازه‌های فاصله همانند فاصله‌ی کسینوسی و فاصله‌ی مهالانوبیس [۲۳]در این فضا موردبررسی قرار گرفتند. برای مقایسه‌ی بین دو ابر بردار، مدل‌های مخلوط گوسی نیازمند این هستند که با یک مدل پس‌زمینه‌ای همگانی وفق داده شوند تا اطمینان پیدا کنیم که بردارهای میانگین متناظر مدل مخلوط گوسی بین بخش‌ها مقایسه شده است. الگوریتم تطبیق در منبع [۲۱]شرح داده‌شده است.

۲-۶-۱-۲- نمایش بردار i

مفهوم بردارهای i برای اولین بار در بحث اثبات گوینده به‌عنوان یک استخراج ویژگی از مدل مخلوط گوسی معرفی شد تا بُعد پارامترهای مدل مخلوط گوسی را کاهش دهند.سیستم استخراج بردار i، سیستمی است که دنباله‌ی بردارهای (به‌طورمعمول ضرایب کپسترال) به‌دست‌آمده از دنباله صوتی را به یک بردار با طول ثابت نگاشت می‌کند به‌طوری‌که مدل‌های مخلوط گوسی با بُعد های مشخص که مدل‌های کلی پس‌زمینه‌ای خوانده می‌شوند برای جمع‌آوری داده‌های الگوریتم باوم-ولچ از دنباله‌ی صوتی مورداستفاده قرار گیرند. مدل پس‌زمینه‌ای همگانی اندازه‌ی آن‌ها در برخی مدل مخلوط گوسی، از مرتبه‌ی ۵۱۲ و ۱۰۲۴ و یا حتی ۲۰۴۸ می‌باشند و اندازه‌ی ابر بردارها[1] برای انجام محاسبات خیلی بزرگ هستند. در عوض، آنالیز برای کاهش بعد ابر بردارها به یک بردار معرف با بعد چند صدتایی می‌رسد. فرض بر این شده است که این زیر فضا، -که زیر فضای تغییر[2] نام دارد-شامل اطلاعات طیفی گوینده و پس‌زمینه می‌باشد.

[1] Supervector
[2] Total Variation Subspace

$$m = M + Tx \qquad (۸-۲)$$

که m ابر بردار متوسط وفق داده‌شده‌ی گفتاری است که X بردارِ i آن بخش می‌باشد . M ابر بردار میانگین مدل‌های پس‌زمینه‌ای همگانی است و T یک ماتریس مرتبه پایین بلند است که زیر فضای تغییر کلی را نشان می‌دهد. این زیر فضا نیازمند این است که روی یک مجموعه داده آموزش داده شود. اگرچه ابر بردارها معمولاً ابعاد ده هزارتایی دارند، این نمایش تمامی ابر بردارها را مجبور می‌کند در یک زیر فضا محدود شوند. بعد این زیر فضا چند صدتا می‌باشد.

استخراج بردار i نیازمند داده‌های آموزش برچسب‌گذاری شده گوینده‌ها با موارد مشابه گوینده می‌باشد. موارد موجود شامل تنوع در نویز زمینه و توازن آوایی می‌باشند. الگوریتم آموزش برای زیر فضای تغییر کلی [۲۳] و استخراج i بردار از الگوریتم باوم-ولچ[1] در ابزار تشخیص MSR اجراشده است[۲۴].

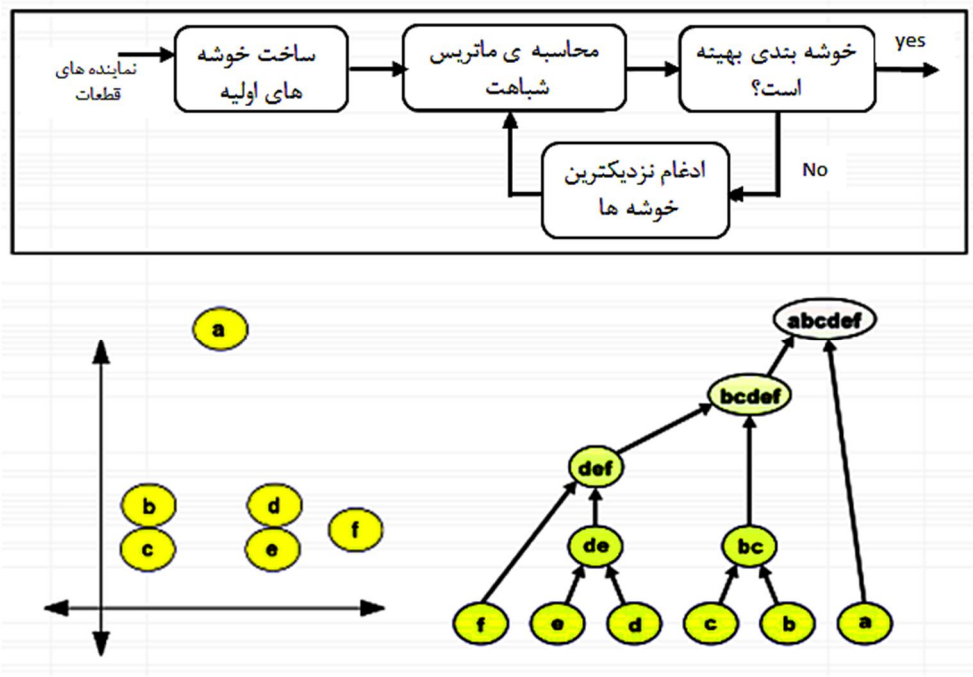
۲-۶-۲- الگوریتم خوشه‌بندی

با داشتن ماتریس شباهت بین مدل مخلوط گوسی مدل‌ها و بردارهای‌i، باهدف بهترین مجموعه دسته‌ها با حداقل واریانس درون دسته‌ای و حداکثر واریانس برون دسته‌ای به دنبال انواع روش‌های دسته‌بندی هستیم. ما دو نوع الگوریتم دسته‌بندی را موردبررسی قراردادیم:۱-دسته‌بندی متراکم طبقه‌ای[2] و ۲- دسته‌بندی برنامه خطی صحیح[3].

هردو مورد بالا روش‌های حل متفاوتی دارند و نیز معیارهای متفاوتی برای رسیدن به بهترین دسته‌بندی رادارند.در ادامه به معرفی هرکدام به‌صورت جداگانه می‌پردازیم تا بتوان در مورد هر روش فهم و ادراک متوسطی به دست آورد.

---

[1] Baum-Welch Statistics
[2] Hierarchical Agglomerative Clustering (HAC)
[3] Integer Linear Program (ILP) clustering (https://en.wikipedia.org/wiki/Integer_programming)



شکل ۲-۵ : روش خوشه‌بندی متراکم سلسله مراتبی

۲-۶-۲-۱- خوشه‌بندی متراکم سلسله مراتبی

این الگوریتم یک الگوریتم حریصانه[1] است بدین معنی که یک انتخاب بهینه‌ی محلی را در هر مرحله با امید یافتن بهینه‌ی کلی اجرا می‌کند. در یک فرایند تکراری، دودسته که بیشترین شباهت رادارند ادغام می‌شوند. تعداد دسته‌ها در هر مرحله یکی کاهش می‌یابد. این فرایند تکراری تا زمانی که تنها یک دسته باقی بماند ادامه پیدا می‌کند. در زمان ادغام دودسته، داده‌ی بخش مرتبط با ۲ دسته باهم یکی مخلوط می‌شوند و یک مدل تک گوینده روی آن‌ها محاسبه می‌شود. فاصله‌ی هر دسته‌ی دیگر با این دسته‌ی جدید مجدداً محاسبه می‌شود و ماتریس شباهت برای هر پله به‌روزرسانی می‌شود.همان‌طور که در شکل (۲-۴) مشاهده می‌کنید در هر مرحله با یافتن بهینه‌ی محلی و ادغام کردن داده‌ها تا جایی که به یک دسته برسند مراحل ادامه پیدا می‌کند و در آخر نتیجه یک دسته‌ی

[1] Greedy Algorithm: (https://en.wikipedia.org/wiki/Greedy_algorithm)

یکتای بهینه است.در زیر مراحل تشریح داده‌شده است:

مرحله ۰: محاسبه ماتریس شباهت (X_i, X_j)

مرحله ۱: یافتن i^* و j^* به‌طوری‌که $i^* \neq j^* \& (X_{i^*}, X_{j^*}) = \min_{i,j}(X_i, X_j)$

مرحله ۲: مرحله ادغام:جایگزین کردن X_{i^*} و X_{j^*} را با یک هدف $X_{K^*}.K^* = \min(i^*, j^*)$

مرحله ۳: به‌روزرسانی کردن ماتریس شباهت (X_i, X_{k^*})

مرحله ۴: رفتن به مرحله اول اگر تعداد دسته‌ها بیشتر از یک

مرحله ۵: محاسبه بهترین مجموعه از دسته‌ها با استفاده از معیار بهینه

مجموعه بهینه از دسته‌ها بر اساس معیار بهینه انتخاب می‌شود. یک معیار بهینه معیاری است که مجموعه‌ای را انتخاب کند که حداقل فاصله برون دسته‌ای بزرگ‌تر از یک حد آستانه است. معیار دیگر توسط[۲۵] معرفی‌شده است که در آن در طول دسته‌ها در هر مرحله‌ی اجرا، مجموعه دسته‌ای که هیستوگرام فاصله‌های درون دسته‌ای و فاصله‌های برون دسته‌ای در بیشترین فاصله از هم قرار دارند انتخاب می‌شود.

$$\arg\max_k \quad \min_{i \neq j}(X_i^{(k)}, X_j^{(k)}) \geq \theta \qquad (۹-۲)$$

که $(X_i^{(k)}, X_j^{(k)})$ ماتریس شباهت در اجرای k^{th} است.

$$\arg\max \quad \frac{|m_{\text{inter}} - m_{\text{int}ra}|}{\sqrt{\frac{\sigma_{\text{inter}}^2}{n_{\text{inter}}} + \frac{\sigma_{\text{intra}}^2}{n_{\text{intra}}}}} \qquad (۱۰-۲)$$

که m_{inter} و σ_{inter} و n_{inter} میانگین و انحراف استاندارد و تعداد عناصر در فواصل برون دسته‌ای و به‌طور مشابه فواصل درون دسته‌ای هستند.

سیستم ICSI[1] [۱۰] یک دسته‌بندی اولیه بخش‌های یک‌ثانیه‌ای با استفاده از ویژگی‌های بلندمدت اجرا می‌کند. سپس این بخش‌بندی در یک مدل مخلوط گوسی-مدل پنهان مارکوف[2] با بردارهای ویژگی ضرایب کپسترال فرکانسی در مقیاس مل (MFCC) دوباره تعریف می‌شوند. هر حالت مدل پنهان مارکوف یک گوینده را نشان می‌دهد و توسط یک مدل مخلوط گوسی مدل شده است. در هر اجرا، تعداد دسته‌ها با ادغام حالت‌های مدل مخلوط گوسی در یک خوشه‌بندی متراکم سلسله مراتبی با استفاده از فاصله‌ی BIC کم می‌شود. سیستم منبع[۷] ضرایب کپسترال پیشگویی خطی[3] برای ۳۰ دسته از صوت یکنواخت را استفاده می‌کند. سپس به‌صورت تکراری مدل پنهان مارکوف-مدل مخلوط گوسی را برای خوشه‌بندی متراکم سلسله مراتبی که از فاصله‌ی CLR[4] استفاده می‌کند اجرا می‌کند. سیستم منبع [۶] همان پیکربندی را استفاده می‌کند ولی به‌جای استفاده از خوشه‌بندی متراکم سلسله مراتبی یک روش بالا به پایین را بکار می‌گیرد. سیستم‌های زیاد دیگری روش خوشه‌بندی متراکم سلسله مراتبی را برای دسته‌بندی با فواصل مختلف بکار گرفته‌اند.[۱۳، ۲۶]

۲-۶-۲-۲- خوشه‌بندی بر اساس ILP

میگنر و همکارانش یک آنالیز نسبی روی هردو روش بازشناسی گوینده که شامل روش پله به پله و روش سیستم یکپارچه است انجام داده‌اند[۲۷].درروش پله‌ای بازشناسی از طریق زیرمجموعه‌های تشخیص فعالیت گفتار، بخش‌بندی، دسته‌بندی انجام می‌شود. درروش همگانی و یکپارچه اطلاعات الگوریتم دسته‌بندی برای بخش‌بندی مجدد گوینده استفاده می‌شود و دسته‌بندی مرتباً تکرار می‌شود.در ادامه‌ی تلاش‌های این تیم و درنتیجه‌ی تلاش‌های روویر و میگنر در سال ۲۰۱۲ [۲۸] ،یک روش بهینه‌سازی کلی برای دسته‌بندی گوینده با استفاده از یک برنامه‌ی خطی

[1] International Computer Science Institute
[2] HMM:hidden marcov model
[3] linear prediction cepstral coeficient (LPCC)
[4] Cross Likelihood Ratio

صحیح[1] اجرا شد. دسته‌بندی به‌عنوان یک مسئله‌ی بهینه‌سازی ترکیبی روی یک گراف کامل (هر گره به گره ی دیگر متصل است) مطرح می‌شود. بخش‌های گوینده به‌عنوان گره‌های گراف در نظر گرفته می‌شود و ماتریس تلاقی ماتریس شباهت است. برنامه‌ی خطی صحیح برای یافتن دسته‌ی بهینه یک نوع دیگری از مسئله‌ی K مرکزی است[2]. به‌طور ساده‌تر مسئله‌ی K مرکزی برای انتخاب K شهر برای ساختن مرکزها است. فاصله‌ی بین یک شهر و نزدیک‌ترین مرکز آن حداقل می‌شود. روش ILPبرای بازشناسی گوینده‌ی بدون ناظر زمانی که تعداد گویندگان نامعلوم است بکار می‌رود.

با معرفی ILP در شاخه‌ی بازشناسی گوینده در حال حاضر سیستم‌های بازشناسی با استفاده از مدل مخلوط گوسی-مدل پنهان مارکوف یک مرحله پس پردازش دیکد کردن ویتربی را بکار می‌برند. در سال‌های اخیر میگنر و همکاران[۲۹] روش بالا را برای کم کردن زوائد برای درست کردن سریع دسته‌ها بکار گرفته‌اند. بدین ترتیب در سال‌های اخیر، روش‌های بر پایه ILP برخلاف روش‌های قدیمی بر پایه‌ی خوشه‌بندی متراکم سلسله مراتبی به دلیل کارکرد بهترشان در سرعت و دقت موردتوجه همگانی قرارگرفته‌اند. جعبه‌ابزار LUIM یک خطای DER برابر با ۱۷.۱۹ درصد را با استفاده از مدل مخلوط گوسی بر پایه دسته‌بندی بر اساس فاصله‌ی CLR گزارش داده است و در مقابل یک خطای DER برابر با ۱۵.۴۶ درصد را با استفاده از بردار i بر پایه دسته‌بندی ILP نشان داده است.[۲۹]

۲-۷- روش‌های ارزیابی:

در بحث ارزیابی نتایج سیستم بازشناسی گوینده معیارهای ارزیابی متفاوتی وجود دارد.در سیستم‌های داده‌ای مختلف در دنیا بنا به کاربرد ، معیار ارزیابی متفاوت است .

[1] Integer Linear Program (ILP)
[2] https://en.wikipedia.org/wiki/Integer_programming

۲-۷-۱- معیار DER

میزان خطای بازشناسی[1] درصد زمانی را مشخص می‌کند که گوینده‌ها به‌اشتباه برچسب‌گذاری شده‌اند[۳۰]. خروجی سیستم با یک قسمت که به‌صورت دستی برچسب‌گذاری گوینده شده است ، مقایسه می‌شود. فرمول زیرسیستم ارزیابی بازشناسی گوینده NIST 2005,2006,2007,2009 را مشخص می‌کند[۳۱]. در این ارزیابی سیستم مورد ارزیابی باید بازشناسی گوینده را بر روی مکالمات ضبط‌شده‌ی گروهی انجام دهد. در این فرمول، برای هر قسمت S با طول زمانی dur(s)، N_{ref} تعداد گوینده‌های مشخص‌شده در منبع اصلی و N_{hyp} تعداد گویندگان فرض شده توسط سیستم ما را به ترتیب نشان می‌دهد.$N_{correct}$ تعداد گوینده‌ها در قسمت s که تطابق درستی بین گوینده‌های منبع و سیستم ما را دارد مشخص می‌کند.

$$DER = \frac{\sum_{s=1}^{S} dur(s).(\max(N_{ref}, N_{hyp}) - N_{correct})}{\sum_{s=1}^{S} dur(s).N_{ref}} \qquad (۱۱-۲)$$

زمانی که همپوشانی نداشته باشیم ، فرمول DER ساده‌سازی می‌شود.

جدول ۲-۱: ساده‌سازی محاسبه‌ی خطا

(NS) زمان کلی غیر گفتار	(s)زمان کلی گفتار
غیر گفتار درست برچسب‌گذاری شده (C2) +	گفتار به‌درستی برچسب‌گذاری شده (C1) +
زمان هشدار اشتباه گفتار(T2)	زمان گفتار ازدست‌رفته (T1) +
	زمان گفتار نادرست برچسب‌گذاری شده (T3)

[1]DER:Diarization Error Rate

همان‌طور که در جدول(۲-۲) ملاحظه می‌کنید زمان‌های به‌دست‌آمده در جداسازی گفتار و غیر گفتار به‌صورت آمده در جدول می‌توانند جداسازی شوند.بدین‌صورت که زمان‌هایی را به‌عنوان زمان گفتار تلقی می‌کنیم که شامل گفتار درست تشخیص داده‌شده(C1) و گفتار تشخیص داده نشده یا همان گفتار ازدست‌رفته (T1) و در آخر شامل زمان گفتاری است که به نادرست غیر گفتار برچسب‌گذاری شده است (T3) می‌باشد.همچنین برای زمان غیر گفتار به‌صورت جمع زمان‌های غیر گفتار درست تشخیص داده‌شده (C2) به همراه زمانی است که به‌غلط گفتار شناسایی‌شده است(T2) می‌باشد.

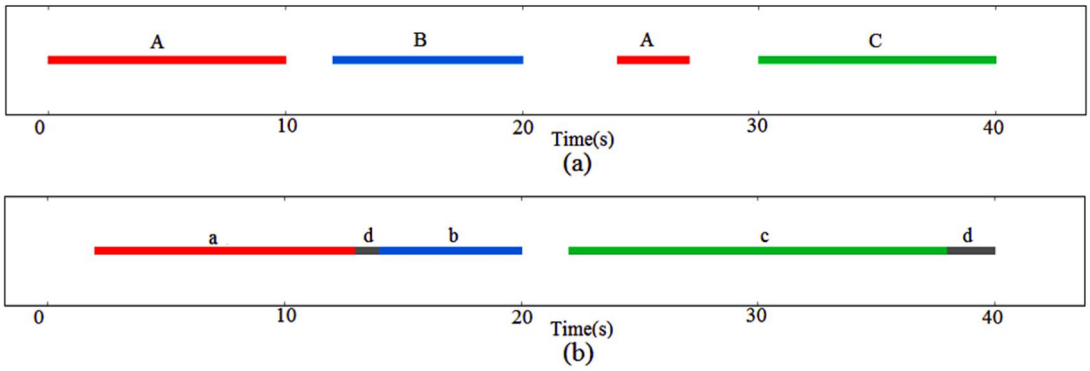
DER می‌تواند به‌صورت سیستماتیک به دو نوع تقسیم شود. فرض کنید صوتی با S ثانیه گفتار و NS ثانیه غیر گفتار به‌صورت دستی مشخص‌شده باشد. قسمت‌های غیر گفتار شامل سکوت ،مکث گوینده و موزیک و نویز می‌باشد. دودسته به‌طور کامل و جزئی به‌صورت شکل زیر می‌توانند دسته‌بندی شوند.

زمان گفتار ازدست‌رفته [۱] زمانی است که الگوریتم به‌غلط قسمتی که در اصل گفتار است را نا گفتار تشخیص می‌دهد. از طرف دیگر زمان هشدار اشتباه گفتار[۲] رمانی است که الگوریتم به‌غلط قسمتی که غیر گفتار می‌باشد را به‌عنوان گفتار تشخیص دهد. این دو خطا طی تشخیص فعالیت گفتار که بخش پیشین در تقریباً همه‌ی سیستم‌های تشخیص گوینده است رخ می‌دهد. تعداد خطاهای نرخ گفتار ازدست‌رفته[۳] و نرخ تشخیص اشتباه گفتار به ترتیب به‌صورت E1=T1*100/S و E2=T2*100/S مشخص می‌شوند.

خطای E3=T3*100/S به‌صورت مجزا توسط هردو بخش قسمت‌بندی گوینده و دسته‌بندی گوینده تولید می‌شود و اغلب بانام خطای گوینده[۴] نام‌برده می‌شود. تغییر گوینده اگر در حین قسمت‌بندی از

[1] Missed Speech Time
[2] False Alarm Speech Time
[3] MSR:Missed Speech Rate
[4] Speaker Error(spk_err)

دست برود باعث دسته‌بندی و کلاس‌بندی شدن نادرست قسمت کوتاه‌تر می‌شود. از طرف دیگر اگر سیستم صوت را به تعداد زیادی قسمت که با زیرنویس مشخص‌شده است تقسیم شود بیش‌ازحد بخش‌بندی[1] نامیده می‌شود. شانسی برای درست کردن این خطا حین دسته‌بندی با یکی کردن قسمت‌های همسایه وجود دارد. هرچند اگر قسمت موردنظر کوچک باشد احتمال خطا حین استخراج اطلاعات گوینده از آن بخش بالا می‌رود. قسمت به‌اشتباه بخش‌بندی شده دلیل اصلی E3 است. در آخر خطای کلی به‌صورت جمع تمامی خطاهای گفته‌شده در بالا محاسبه می‌شود.



شکل ۲-۶ : مثالی از محاسبه‌ی DER : الف)منبع ب) سیستم ما

محاسبه‌ی spk_err برای اجرای یک تطابق بین برچسب‌های الگوریتم و برچسب‌های منبع لازم و ضروری است. برای این منظور بهترین جفت کلاس را بر اساس بیشینه همپوشانی بین دودسته برچسب انجام می‌رود. مثالی در این زمینه در جدول زیر آورده شده است.در جدول ۲-۲ در بخش برچسب‌گذاری منبع همان‌طور که در شکل (۲-۶) مشاهده می‌کنید صوت ورودی ما به بخش‌های مختلف که هرکدام دارای گفتار مربوط به یک گوینده خاص است تقسیم‌شده است.در ستون کناری بخش‌بندی و تخصیص گوینده انجام‌شده است و در کنار داده‌های داشته از منبع برای مقایسه قرارگرفته است .در شرح داده‌های منبع می‌بینیم که صوت در درست دارای سه گوینده می‌باشد که بر اساس قسمت‌های برحسب ثانیه داده‌شده‌اند به‌طور مثال گوینده‌ی اول ثانیه‌ی اول تا دهم صحبت

[1] Oversegmentation

کرده است و از ثانیه‌ی ۱۰ تا ثانیه‌ی ۱۲ شامل غیر گفتار است و از ثانیه‌ی ۱۲ تا ثانیه‌ی ۲۰ گوینـده‌ی دوم صحبت کرده است و به همین طریـق در ادامـه .در بخـش‌بنـدی و مشـخص‌سـازی دسـتی دیـده می‌شود که ثانیه‌ی ۲ تا ۱۳ گوینده‌ی اول تشخیص داده‌شده است و از ثانیه‌ی ۱۳ تا ۱۴ را غیـر گفتـار تشخیص داده است که دیده می‌شود تطابق درستی با فرضیات منبع ندارد و این امـر باعـث بـه وجـود آمدن خطاهای گفته‌شده می‌باشند. در ادامه‌ی تشریح جدول ۲-۲ می‌توان دید کـه در انتهـای صـوت نیز گفتار گوینده سوم را به‌اشتباه غیر گفتار دسته‌بندی کرده است.

جدول ۲-۲: مثالی از بخش‌بندی برای محاسبه‌ی DER

فرضیه	برچسب‌گذاری منبع
قسمت (ثانیه۲ام الی ثانیه ۱۳ام) : (a)(گوینده الف)	قسمت (ثانیه۰ام الی ثانیه ۱۰ام) =گوینده (الف)
قسمت (ثانیه۱۳ام الی ثانیه ۱۴ام) : (d)(غیر گفتار)	قسمت (ثانیه۱۲ام الی ثانیه ۲۰ام) =گوینده (ب)
قسمت (ثانیه۱۴ام الی ثانیه ۲۰ام) : (b)(گوینده دوم)	قسمت (ثانیه۲۴م الی ثانیه ۲۷ام) =گوینده (الف)
قسمت (ثانیه۲۲ام الی ثانیه ۳۸ام) : (c)(گوینده سوم)	قسمت (ثانیه۳۰م الی ثانیه ۴۰ام) =گوینده (ج)
قسمت (ثانیه۳۸ام الی ثانیه ۴۰ام) : (d)(غیر گفتار)	

۲-۸- پایگاه داده‌های مورداستفاده

۲-۸-۱- AMI curpus

اولین پایگاه داده مورداستفاده در این پایان‌نامه AMI curpus [۳۲] مـی‌باشـد کـه شـامل ۱۰۰ ساعت گفتگوی ضبط‌شـده بـین گوینـدگان مختلـف مـی‌باشـد. مکالمـات ضـبط‌شـده بـا اسـتفاده از میکروفون‌های نزدیک و دور و نیز دوربین‌های کنفرانسی و تخته‌های نمایش مجـزا اسـت. جلسـات در زبان انگلیسی با استفاده از سه اتاق‌های مختلف با خواص مختلف صوتی ثبت‌شـده و شـامل سـخنرانان عمدتاً غیربومی است. مدت‌زمان این مکالمات و جنس گویندگان مختلف می‌باشد.مشخصـات تـک‌تـک

ثانیه‌های مکالمات و ویژگی‌های گویندگان در دسترس می‌باشد .این پایگاه داده توسط گروه تحقیقاتی دانشکده انفورماتیک[1] دانشگاه ادینبرگ در اسکاتلند [2]تدوین و منتشرشده است و کارایی‌های مختلفی برای آن ذکرشده است که در این پایان‌نامه برای بازشناسی گوینده مورداستفاده قرار می‌گیرد.

۲-۸-۲ TIMIT

دومین پایگاه داده مورداستفاده در این پایان‌نامه دیتابیس TIMIT [۳۳] است .دادگان TIMIT یک بانک اطلاعاتی از گفتار پیوسته انگلیسی است که با همکاری شرکت TI[3] و دانشگاه MIT[4] منتشرشده است. این دادگان حاوی ۶۳۰۰ جمله است که توسط ۶۳۰ گوینده و با ۸ لهجه معمول آمریکای شمالی بیان‌شده‌اند. ۷۰٪ گویندگان مرد و ۳۰٪ آن‌ها زن هستند. هر گوینده ۱۰ جمله را ادا کرده است که ۲ جمله از این ۱۰ جمله توسط سایر گویندگان نیز ادا شده است.از این پایگاه داده در آموزش داده‌ها و مدل‌ها و در دیگر مواقع موردنیاز بهره گرفته‌شده است.

۲-۹- خلاصه و جمع‌بندی فصل

در این فصل برخی مفهوم‌های پس‌زمینه‌ای مورداستفاده در بخش‌بندی و دسته‌بندی و تشخیص فعالیت گفتار موردبررسی قرار گرفت. مابقی کارهای بازشناسی گوینده همانند تعداد گویندگان نامشخص است. بسیاری از روش‌های نشانه‌گذاری‌های صوت برای مثال تشخیص حضور آهنگ[۳۴] و کمک به یافتن ساختار برنامه‌ی رادیوتلویزیونی[۳۵] درگذشته موردبررسی قرار گرفتند. سیستم‌های معمول دیگر شامل موارد مختلفی من‌جمله ابزارهایی جون SHOUT [۳۶] و INRIA [۳۷] و IDIAO [۳۸] و DIARTK [۳۹] می‌باشند.

[1] http://www.ed.ac.uk/informatics
[2] http://www.ed.ac.uk/
[3] Texas Instruments
[4] Massachusetts Institute of Technology

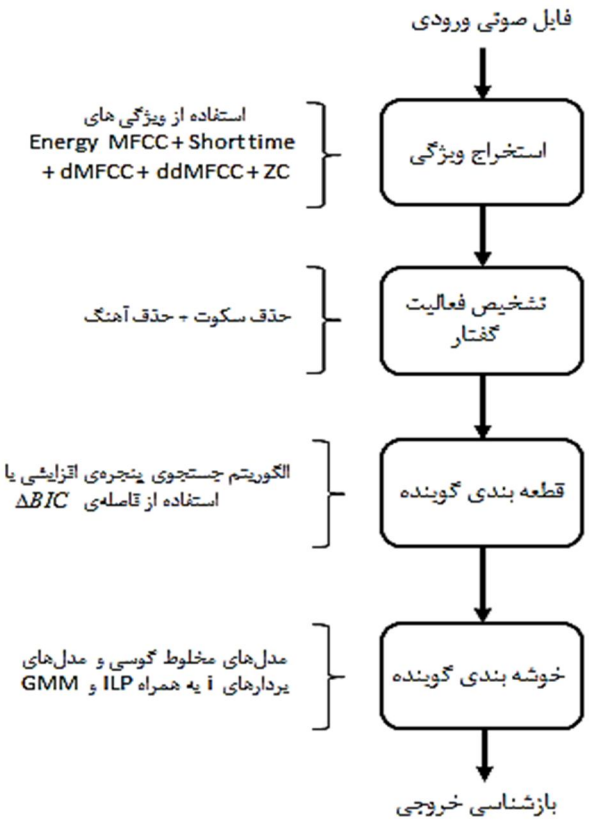
فصل ۳ :

روش پیشنهادی در بازشناسی گوینده

۳-۱- مقدمه

در این بخش به معرفی سیستم مورداستفاده در این پایان‌نامه می‌پردازیم.یک سیستم کامل در نرم‌افزار متلب برای بازشناسی گوینده فایل‌های صوتی اجراشده است. این سیستم تست و ارزیابی‌شده است. برای ارزیابی معیارهای توضیح داده‌شده در بخش قبل مورداستفاده قرارگرفته‌اند. در این بخش توضیح کاملی از این سیستم از استخراج ویژگی تا ارزیابی خطاها را موردبررسی قرار می‌دهیم. انتخاب پارامترها و مدل‌ها در هر بخش توضیح داده‌شده است.

با پیدایش الگوریتم‌های دسته‌بندی بهتر با مدل بردارهای i، تمرکز روی اجرای جدای از هم بخش‌بندی گوینده و دسته‌بندی متمرکز شد. اگرچه استفاده از مدل مخلوط گوسی-مدل مخفی مارکوف همچنان معروف و مشهور است[۱۶]. شکل (۱-۴) بلوک دیاگرام سیستم پیشنهادی که روش پله به پله را نشان می‌دهد رسم شده است.



شکل ۳-۱: بلوک دیاگرام سیستم پیشنهادی

۳-۲- استخراج ویژگی

ویژگی ضرایب فرکانسی در مقیاس مل[1] مکرراً در بازشناسی برای تمام زیرمجموعه‌های تشخیص فعالیت گفتار و بخش‌بندی و دسته‌بندی مورداستفاده قرار می‌گیرد. سیستم پیشنهادی ۱۹ ویژگی ضرایب کپسترال فرکانسی در مقیاس مل برای تمام پردازش‌ها بکار می‌گیرد.این استفاده به دلیل دربرگیری تمامی اطلاعات فریم‌های گفتار و غیر گفتاراست تا بتوانیم در ادامه در مدل‌سازی قسمت‌های مختلف به نتیجه‌ی بهتری دست پیدا کنیم. علاوه بر این ویژگی‌ها، انرژی زمان کوتاه و نرخ عبور از صفر و مشتق‌های مرتبه اول و دوم آن‌ها در بخش تشخیص فعالیت گفتار بکار می‌روند. بخش‌بندی گوینده تنها ۱۹ ویژگی ضرایب کپسترال فرکانسی در مقیاس مل و انرژی زمان کوتاه را

[1] MFCC : Mel Frequency Cepstral Coeficient

استفاده می‌کند.در بخش دسته‌بندی مشتق‌های مرتبه اول و دوم آن‌ها نیز بکار می‌رود. اندازه فریم‌ها برای آنالیز ۳۰ میلی‌ثانیه با همپوشانی ۲۰ میلی‌ثانیه است.[۲۳]

۳-۳- تشخیص فعالیت گفتار[1]

تشخیص فعالیت عبارت است از گفتار جداسازی قسمت‌های گفتار از غیر گفتار در فایل صوتی ضبط‌شده است. در حین انجام این جداسازی هدف مدنظر ما حداقل کردن خطاهای به وجود آمده در این بخش است. این خطاها شامل نرخ گفتار ازدست‌رفته[2] و نرخ هشدار خطای گفتار[3] می‌باشند. درصد گفتاری که به‌اشتباه در غیر گفتار توسط این بخش دسته‌بندی می‌شود نرخ گفتار ازدست‌رفته نام دارد و درصد غیر گفتاری که اشتباهاً گفتار دسته‌بندی‌شده است را نرخ هشدار خطای گفتار می‌نامند. عملاً این دو معیارهای ارزیابی بخش تشخیص فعالیت گفتار می‌باشند. معمولاً ۱الی ۳ درصد خطای گفتار ازدست‌رفته و ۲الی۴ درصد خطای هشدار خطای گفتار خطاهای معمول امروزی در سیستم‌های تشخیص فعالیت گفتار است. در بازشناسی گفتار، هشدار خطای گفتار باعث خرابی در مدل‌های گوینده طی دسته‌بندی و بخش‌بندی می‌شود و فرایند دسته‌بندی را تحت تأثیر قرار می‌دهد.

یک روش معمول در سیستم‌های تشخیص فعالیت گفتار سعی بر دست بندی کردن همه‌ی انواع صداهای موجود در فایل صوتی است. اگر داده‌های بازشناسی شده از قبل معلوم باشند و ویژگی‌های خاصی مثل شاخص‌های صوتی برای مشخص کردن هر بخش داشته باشند، امکان آموزش مدل‌های آماری برای این شاخص‌ها درداده‌ها و دسته‌بندی مستقیم می‌باشد[۳۴].متأسفانه هنگام بهبود سیستم‌های کلی این‌چنین ویژگی‌هایی همانند شاخص‌ها و ویژگی‌های صوتی برای جداسازی داده‌های صوتی ورودی نداریم. در این بخش مشکل تخمین مدل‌های غیر گفتار برای تشخیص فعالیت گفتار بدون اطاعات قبلی در مورد داده‌ها بررسی‌شده است.

---

[1] Speech Activity Detection
[2] Missed Speech Rate (MSR)
[3] False Alarm Speech Rate (FASR)

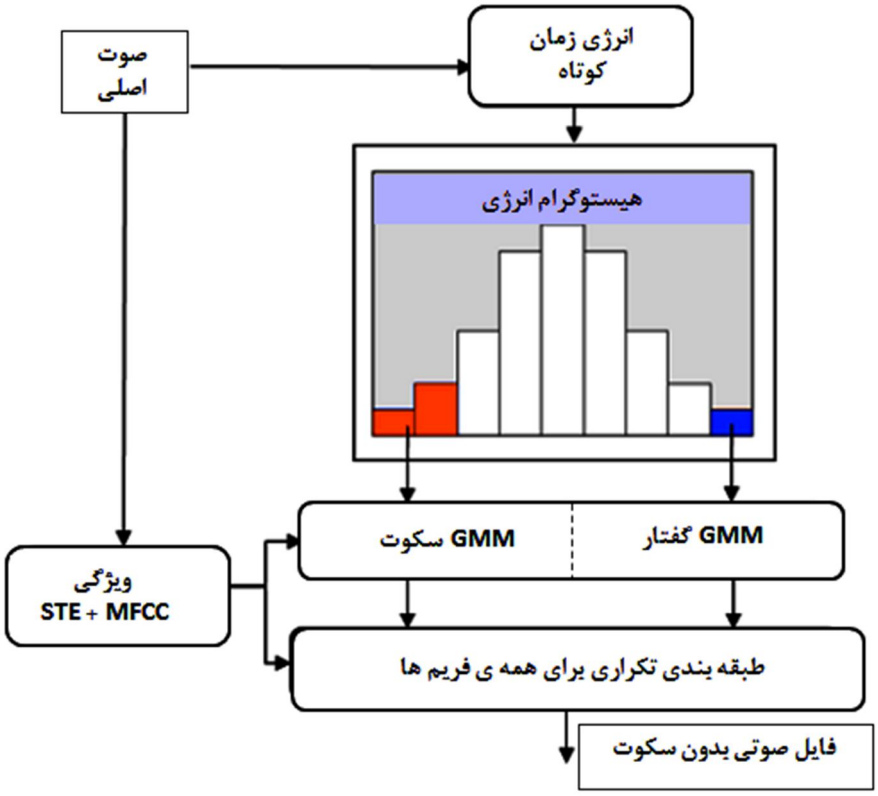
صداها علاوه بر گفتار معمولی که ما در صوت می‌بینیم انواع صداهای تولیدی انسان همانند ملچ‌ملوچ کردن و خندیدن و دست زدن و فوت کردن و یا صداهای ابزاری مانند آهنگ و جیرینگ جیرینگ و غیره نیز وجود دارند. قسمت‌های سکوت و مکث نیز سهم مهمی از بیشتر پایگاه‌های داده صوتی را دربرمیگیرند. این صداها با همدیگر مجموعه‌ی غیر گفتار را تشکیل می‌دهند. روش‌های بر پایه مدل در تشخیص فعالیت گفتار معمول هستند. مدل‌های آماری برای گفتار و غیر گفتار با داده‌های خارجی آموزش‌دیده شده‌اند. اگرچه ایراد این سیستم‌ها اتکای آن‌ها بر شرایط آکوستیکی خارج از داده‌ی نمونه است. سیستم‌های هیبرید[۱۰،۶] از یک خوشه بند آموزش‌دیده شده با داده‌های خارجی استفاده می‌کنند تا یک بخش‌بندی اولیه‌ی متکی بر خود بسازند. سپس مدل‌های گفتار و غیر گفتار مرتبط با بخش‌بندی تعریف می‌شوند تا بتوانند تغییر آکوستیکی در غیر گفتار را مدل‌سازی کنند.

تشخیص‌دهنده‌ی گفتار در این سیستم دسته‌بندی بر اساس مدل است. همچنین مستقل از داده‌های آموزش برای مدل‌سازی گفتار و غیر گفتار است. این نوع سیستم بر پایه مدل تشخیص فعالیت گفتار بر اساس سیستم ارزیابی NIST RT2009می‌باشد [۷]. در سیستم ما تشخیص فعالیت گفتار در دو مرحله‌ی مجزا صورت می‌گیرد. اولین مرحله شامل جداسازی سکوت از کل فایل صوتی با استفاده از انرژی توسط دسته‌بندی تکراری است. در مرحله دوم آهنگ و دیگر اصوات غیر گفتار تشخیص داده می‌شوند. برای حذف آهنگ، صوت بدون سکوت به سیستم تشخیص گری که در بالا گفته شد داده می‌شود. فریم‌هایی از صوت که دارای آهنگ با اطمینان بالا هستند برای آموزش مدل‌های آهنگ که مکرراً تکرار می‌شود بکار می‌روند. در هر دو مرحله، تنها بخش‌های با طول ۱ ثانیه و یا بیشتر به‌عنوان غیر گفتار برچسب می‌خورند به این دلیل که هر از چند گاهی از غیر گفتار به گفتار منتقل نشویم. این قید و محدودیت در منبع [۳۸] و [۳۶] با استفاده از مدل مخلوط گوسی-مدل مخفی مارکوف بکار گرفته‌شده است.

۳-۳-۱- **حذف سکوت**

حذف سکوت درروش اجرایی توسط ۱۹ ضرب کپسترال در مقیاس مل به همراه انرژی زمان کوتاه و مشتقات اول و دوم آن‌ها صورت گرفته است. بخش‌بندی خود راه‌انداز یک مقدار مشخص را در فریم به هر دودسته‌ی سکوت و گفتار اختصاص می‌دهد. مدل سکوت با استفاده از یک گوسی با اندازه‌ی ۴ روی فضای ویژگی ۶۰ بعدی آموزش داده می‌شود. مدل گفتار با همان اندازه از فریم‌های مطمئن گفتار آموزش داده می‌شود.

در یک مرحله‌ی تکراری دسته‌بندی کردن، هر فریم به دودسته سکوت و گفتار دسته‌بندی می‌شود. فریم‌های گفتار با اطمینان بالا و سکوت از این مرحله برای آموزش مدل‌های گفتار و سکوت برای اجرای بعدی بکار می‌روند. همان‌طور که تعداد اجراها بالا می‌رود تعداد گوسی‌های ۶۰ بعدی مورداستفاده برای مدل کردن مدل مخلوط گوسی‌های گفتار و سکوت بالا می‌رود. این روند تا مقدار حداکثری ادامه دارد .همان‌طور که در شکل (۳-۲) مشاهده می‌کنید از فایل صوتی اصلی فریم‌های باانرژی حداکثر و حداقل جداگانه برای تشخیص گفتار و سکوت از هم بکار می‌روند و با استفاده از آن‌ها مدل‌های گوسی برای سکوت و گفتار تولیدشده و توسط این مدل‌ها و دسته‌بندی آن‌ها فریم‌های سکوت از فایل اصلی جدا می‌شود و فایل صوتی بدون سکوت در آخر منتج می‌شود.



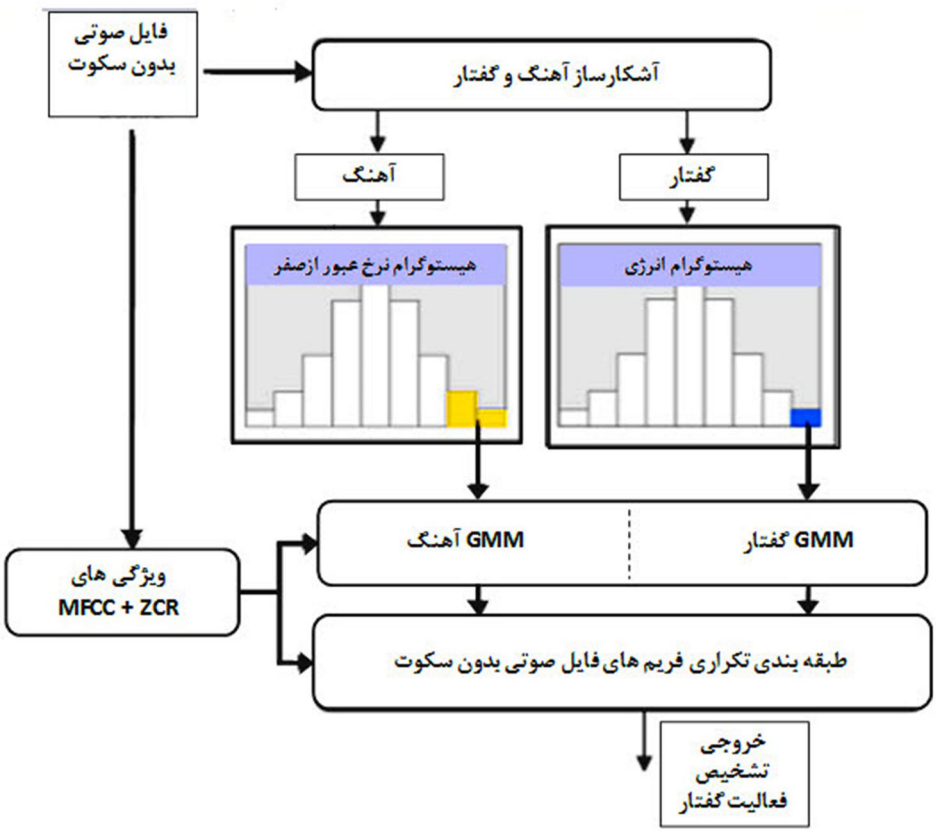
شکل ۳-۲ : حذف سکوت با استفاده از خوشه‌بندی تکراری

۳-۳-۲- حذف موزیک

در سال ۲۰۰۵ در [۲۰] یک مدل سازگار بر پایه دسته‌بند آهنگ و گفتار که درصد دسته‌بندی ۹۵ درصد را گزارش داده معرفی شد. نویسنده فایل صوتی را به قطعه‌های ۱ ثانیه‌ای بخش‌بندی کرده است و سپس ۵۰ بردار ویژگی را در طول پنجره‌های ۲۰ میلی‌ثانیه استخراج کرده است. این بردارهای ویژگی دوبعدی بودند-بعد اول انرژی کوتاه‌مدت و بعد دوم نرخ عبور از صفر سیگنال پنجره‌بندی شده- هیستوگرام انرژی کوتاه‌مدت و نرخ عبور از صفر برای قطعات ۱ ثانیه‌ای محاسبه شدند و با هیستوگرام مدل‌های موزیک و آهنگی که از یک دیتابیس بزرگ موزیک و گفتار به‌دست‌آمده‌اند مقایسه می‌شوند. این هیستوگرام‌های ایده آل با توزیع X^2مدل شدند. قطعه‌ی موردبررسی بعد از مقایسه‌ی بین هیستوگرام قطعه‌ی یک‌ثانیه‌ای و مدل‌های X^2 به‌عنوان آهنگ یا گفتار برچسب‌گذاری می‌شود.

این شاخص گفتار و آهنگ [۲۰] زمانی که گفتار و آهنگ باهم هستند شکست می‌خورد. در

اغلب داده‌های گفتاری برای جداسازی بخش‌های مختلف گفتاری آهنگ‌های پس‌زمینه خاصی به آن اضافه می‌شود که این باعث پیدایش خطا و از دست رفتن اطلاعات اصلی ما می‌شود. ازآنجاکه ما خروجی دسته‌بند را به‌عنوان یک بخش‌بندی قوی در نظر می‌گیریم، تخمین‌های اولیه مدل‌ها برای آهنگ و گفتار بر اساس فریم‌های با اطمینان بالا از هر دو کلاس آموزش می‌بینند. یک دسته‌بندی مشابه با سیستم حذف سکوت برای جداسازی دسته‌های موزیک و گفتار و حذف بخش‌های موزیک انجام شد(شکل (۳-۳)). ویژگی‌های مورداستفاده ۱۹ ضریب ضرایب کپسترال فرکانسی در مقیاس مل به همراه نرخ عبور از صفر و مشتقات اول و دوم آن‌ها است. انرژی زمان کوتاه در این فرایند استفاده‌نشده است. بعد از نادیده گرفتن انرژی زمان کوتاه گفتار با آهنگ پس‌زمینه که در دسته‌بند موزیک قرارگرفته بود جداسازی شد.



شکل ۳-۳ : حذف موسیقی با استفاده از یک آشکارساز موسیقی گفتار

۳-۳-۳- اندازه‌گیری‌های تشخیص فعالیت گفتار

در طول حذف سکوت، هیستوگرام انرژی فریم‌ها برای مرتب‌سازی فریم‌ها بر اساس انرژی آن‌ها استفاده می‌شود. فریم‌های در ۲۰ درصد پایین هیستوگرام به‌عنوان فریم‌های قطعاً سکوت و فریم‌های در ۱۰ درصد بالایی نمودار به‌عنوان فریم‌های گفتار با درصد قاطعیت بالا در نظر گرفته می‌شوند. در هر اجرا و مرحله، این فریم‌ها برای آموزش مدل مخلوط گوسی[1] بکار می‌روند. برای حذف آهنگ، هدف ما جمع‌آوری فریم‌هایی است که گفتار همراه با آهنگ پس‌زمینه هستند اما به‌عنوان غیر گفتار دسته‌بندی‌شده‌اند. بدین منظور تنها ۴۰ درصد بالای فریم‌های نرخ عبور از صفر از هیستوگرام نرخ عبور از صفر را به‌عنوان فریم‌های خالص آهنگ برای آموزش مدل‌های آهنگ بکار گرفتیم.

۳-۴- ارزیابی تشخیص فعالیت گفتار

در این بخش نتایج ۳ آزمایش مجزا نشان داده‌شده است. اثر اندازه‌ی مدل‌های مخلوط گوسی در مراحل اجزای SADو وضعیت ماتریس کوواریانس مدل‌های مخلوط گوسی و اثر اتصال این دو سیستم به هم دیگر موردبررسی قرار گرفت. نرخ گفتار ازدست‌رفته[2] درصد خطای زمانی است که گفتار به‌صورت غیر گفتار دسته‌بندی شود. نرخ هشدار اشتباه گفتار[3] خطای زمانی است که صوت غیر گفتاری به‌اشتباه گفتار تلقی شود. خطای SAD شامل جمع آن دو خطاست.

۳-۴-۱- اندازه‌ی مدل‌های مخلوط گوسی‌ها حین جداسازی سکوت

پس از بخش‌بندی، مدل‌ها برای فریم‌های محدود گفتار و غیر گفتار آموزش می‌بینند که

[1] GMM:Guasian Mixture Model
[2] Missed Speech rate (MSR)
[3] False Alarm Speech Rate (FASR)

مدل‌های مخلوط گوسی با اندازه‌ی ۴ هستند. به میزانی که داده‌های موجود در دسته‌ها افزایش می‌یابد، اندازه‌ی مدل‌های مخلوط گوسی استفاده‌شده برای مدل‌سازی گفتار و سکوت نیز افزایش می‌یابد. در هر مرحله‌ی اجرا، اندازه‌ها دو برابر می‌شوند تا به یک حد بیشینه‌ای برسند. در این آزمایش‌ها اندازه‌ی مناسبی را که برای هردو بخش بهتر است انتخاب کرده‌ایم. در جدول (۳-۱) درصد خطای SAD را برای اندازه‌های مختلف مدل‌های مخلوط گوسی نشان داده‌شده است. این اندازه‌های بهینه برای گفتار ۳۲ مدل مخلوط گوسی و برای غیر گفتار ۱۶ مدل مخلوط گوسی به دست آمدند.

جدول ۳-۱: خطای SAD برای اندازه‌های مختلف مدل‌های مخلوط گوسی برای مدل‌های گفتار و سکوت(برحسب درصد./)

گفتار					
سکوت	اندازه مدل‌های مخلوط گوسی	۸	۱۶	۳۲	۶۴
	۴	۱۳.۴۲	۹.۸۲	۵.۹۴	-
	۸	۱۰.۶۲	۵.۷۹	۵.۵۴	۵.۸۹
	۱۶	۶.۴۴	۵.۷۴	۵.۲	۵.۴۹
	۳۲	۱۳.۵	۶.۵۲	۵.۹۶	-

۳-۴-۲- ماتریس کوواریانس مدل‌های مخلوط گوسی در دسته‌بندی تکراری حذف سکوت و حذف آهنگ

در این بخش، مدل‌های مخلوط گوسی گفتار، سکوت و آهنگ با استفاده از ۳۲ و ۱۶ و ۳۲ گوسی در آخرین مرحله از دسته‌بندی مدل شدند. یک مدل مخلوط گوسی کوواریانس کامل برای آموزش بهتر کلاس گفتار می‌تواند مفید باشد. این باعث MSR پایین می‌شود هرچند برای دسته‌های سکوت و آهنگ تعداد فریم‌ها محدود شدند اما یک مدل مخلوط گوسی کوواریانس کامل مناسب است. هرچند خطای FASR به‌طور مشخص افزایش می‌یابد.

جدول ۳-۲ :خطای برای SADحالت ماتریس کوواریانس مدل‌های مخلوط گوسی (برحسب درصد./)

حالت ماتریس / مرحله انجام	کامل	قطری
حذف سکوت	۵.۴۹	۵.۲
حذف آهنگ	۸.۱۳	۷.۶۹

۳-۴-۳- اثر متوالی کردن حذف سکوت و حذف آهنگ

استفاده از سیستم حذف موزیک یک نرخ خطای غلط زیادی را می‌دهد. متوالی کردن دو سیستم بهبود خوبی را در مقایسه با استفاده مجزای آن‌ها می‌دهد. همان‌طور که انتظار می‌رفت، MSR افزایش می‌باید ولی FASR به دلیل دسته‌بندی فریم‌های موزیک در غیر گفتار کاهش می‌یابد.

جدول ۳-۳: متوالی کردن سیستم حذف آهنگ و حذف سکوت برای SAD(برحسب درصد./)

	MSR	FASR	جمع دو خطا
حذف سکوت	۱.۴۹	۳.۷۱	۵.۲۰
حذف آهنگ	۱.۵۹	۶.۱۱	۷.۶۹
سیستم متوالی	۲.۳	۲.۷۱	۵.۰۱

۳-۴-۴- نرخ عبور از صفر برای آموزش مدل‌های آهنگ

استفاده از فریم‌هایی که به‌طور یقین آهنگ هستند که از مقادیر نرخ عبور از صفر به‌دست‌آمده‌اند، برای آموزش مدل‌های آهنگ و نادیده گرفتن فریم‌های با نرخ عبور از صفر پایین باعث کاهش نرخ خطای از دست رفتن گفتار می‌شود. این افزایش به علت برگرداندن فریم‌های گفتار

از فریم‌های گفتار مخلوط با آهنگ پس‌زمینه است. افزایش FASR به دلیل گفتاری است که از فریم‌های با صدای جیرینگ و ...به‌دست‌آمده است.

جدول ۳-۴: تعریف مجدد مدل‌های آهنگ با استفاده از فریم‌های با ZCR بالا(برحسب درصد./)

	MSR	FASR	جمع هر دو
فریم‌های آهنگ با ZCR بالا	۲.۳	۲.۷۱	۵.۰۱
همه فریم‌های آهنگ	۳.۰۸	۲.۳۲	۵.۴۱

۳-۵- بخش‌بندی گوینده

الگوریتم بخش‌بندی گوینده مورداستفاده در این پایان‌نامه یک الگوریتم جستجوی پنجره‌ی افزایشی [۱۴] با استفاده از فاصله‌ی ΔBIC که در ادامه معرفی می‌شود، است. در ابتدا، جستجو برای یافتن یک تغییر گوینده صورت می‌گیرد. در هر تغییر پیداشده، جستجو از فریم بعدی شروع می‌شود. پنجره‌ی اولیه با مقدار ۵ ثانیه مقداردهی اولیه می‌شود و یک مقدار ΔBIC برای هر فریم از پنجره محاسبه می‌شود. اگر حداکثر این آرایه از حد آستانه‌ی θ، فراتر رود، تغییری در حد آستانه تشخیص داده می‌شود. اگر حداکثری در طول پنجره‌ی جستجو یافت نشود، اندازه پنجره به ۲ ثانیه تغییر می‌کند و روند دوباره تکرار می‌شود. دقت شود تنها فریم‌های گفتار به‌دست‌آمده از بخش تشخیص فعالیت گفتار مورد پردازش قرار می‌گیرد. بعد از یافتن نقاط تغییر در فریم‌ها، مکان آن‌ها در فایل اولیه و اصلی مشخص می‌شنود و به‌عنوان نقاط تغییر شناخته می‌شنود.

در دو سیستم قبل [۳۸،۳۶] بخش‌بندی در دو بخش صورت می‌گیرد: مرحله اول ΔBIC تشخیص نقاط تغییر همان‌طور که گفته شد با آستانه‌ی صفر اجرا می‌شود سپس ادغام بخش‌های متوالی برای آن‌هایی که امتیاز ΔBICشان مثبت است صورت می‌گیرد. دومرحله‌ای بودن به علت

بیش‌ازحد بخش‌بندی کردن بر اساس حد آستانه‌ی ΔBIC صفر می‌باشد. برای جلوگیری از این امر، تنها بیشینه‌ای که از یک آستانه‌ی(θ) صفر بزرگ‌تر است انتخاب می‌شود. این امر به‌مراتب بیش‌ازحد دسته‌بندی کردن را کاهش می‌دهد:

$$\Delta BIC(\mathrm{x}_i) = \mathrm{N}\log|\Sigma| - \mathrm{N}_1\log|\Sigma_1| - \mathrm{N}_2\log|\Sigma_2| - \frac{\lambda}{2}(\mathrm{d} + \frac{1}{2}\mathrm{d}(\mathrm{d}+1))\log\mathrm{N} \qquad (۱-۳)$$

در (۳-۲) پارامتر Σ ماتریس کوواریانس داده‌ها و d بعد فضای بررسی ما و N تعداد داده‌های بردارهای ویژگی را در هر مدل گوسی نشان می‌دهد. برای بخش‌بندی گوینده، ۱۹ ضریب کپسترال فرکانسی در مقیاس مل باانرژی زمان کوتاه استفاده شدند. اغلب سیستم‌های بخش‌بندی گوینده[۳۸،۳۶،۶] مشتقات ویژگی‌های کپسترال را هنگام بخش‌بندی استفاده نمی‌کنند.

۳-۶- دسته‌بندی گوینده

بعد از تشخیص نقاط تغییر در گوینده با استفاده از بخش‌بندی گوینده زیر بخش دسته‌بندی گوینده باهدف جمع‌آوری بخش‌هایی که گوینده یکسان دارند اجرا می‌شود. بدین منظور بخش‌های فراهم‌شده با مدل‌های گوینده نشان داده می‌شوند. شباهت دوبه‌دویی بین همه‌ی مدل‌های گوینده به دست می‌آید و الگوریتم دسته‌بندی برای گروه‌بندی کردن آن‌ها بکار می‌رود.

۳-۶-۱- انتخاب مدل گوینده

سیستم پیشنهادی دو مدل گوینده را بررسی می‌کند که به‌طور وسیعی در حیطه‌ی تشخیص گوینده و بازشناسی گوینده موردمطالعه قرار می‌گیرد: ۱- مدل‌های مخلوط گوسی ۲- مدل‌های بردارهای i. مدل مخلوط گوسی (۳-۵) یک مدل احتمالی روی فضای ویژگی است. ویژگی‌های استفاده‌شده انرژی زمان کوتاه و ۱۹ ویژگی ضریب ضرایب کپسترال فرکانسی در مقیاس مل و مشتقات اول و دوم آن‌ها در یک فضای ویژگی ۶۰ بعدی است. شباهت بین مدل مخلوط گوسی ، بر پایه‌ی مدلِ بیشترین شباهت متقابل هر بخش از داده با داده‌ی دیگر است. مدل مخلوط گوسی برای

هر بخش روی بردارهای ویژگی بخش با استفاده از الگوریتم بیشینه امید[1] آموزش داده می‌شود تا یک مدل مخلوط گوسی با ماتریس کوواریانس قطری با اندازه ۳۲ به دست آید. در حین محاسبه‌ی سیستم با استفاده از مدل‌های گوینده مدل مخلوط گوسی، فاصله‌های CLR[2]و NCLR[3]((۳-۶)،(۳-۷)) با استفاده از الگوریتم‌های دسته‌بندی HAC[4] وILP[5] محاسبه می‌شوند.

۳-۶-۲- استخراج بردار i

برای به دست آوردن بردار i ، ابتدا یک مدل زمینه‌ی کلی روی‌داده‌های آموزش، آموزش داده می‌شوند. مدل پس‌زمینه کلی[6] یک مدل مخلوط گوسی با تعداد بسیار زیادی از گوسی‌ها است بنابراین می‌تواند تمامی تنوع‌های ممکن در گفتار را در فضای ویژگی جمع‌آوری کند. در سیستم پیشنهادی با استفاده از دیتابیس TIMITمدل پس‌زمینه کلی آموزش داده می‌شوند. مدل پس‌زمینه کلی یک مدل مخلوط گوسی با ماتریس کوواریانس قطری با اندازه ۵۱۲ است. آموزش مدل پس‌زمینه کلی تنها یک‌بار انجام می‌شود. متوسط مدل پس‌زمینه کلی و مدل مخلوط گوسی بخش وفق یافته برای به دست آوردن یک ابر بردار با اندازه ۳۰۷۲۰ به هم متصل می‌شوند.

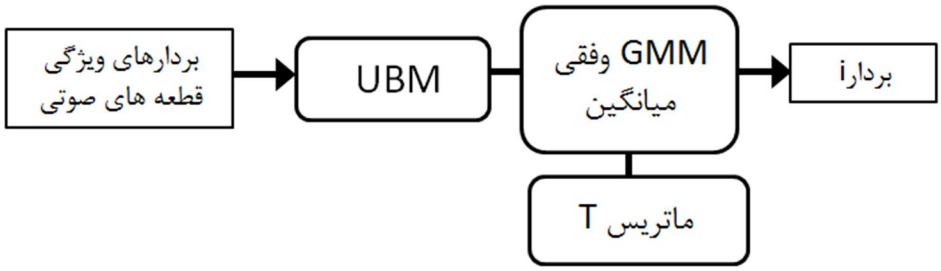
فضای تغییر کلی یک زیر فضا از ابر فضای مدل مخلوط گوسی است که تمام گوینده‌ها و اطلاعات مرتبط با آن‌ها را دربردارد. ماتریس T ماتریس مرتبه پایینی است که ستون‌هایش کل زیرمجموعه‌ی تغییر را در برمی‌گیرد. برای سیستم پیشنهادی، ماتریس T، با استفاده از همان گویندگانی که برای آموزش مدل پس‌زمینه کلی استفاده شدند آموزش می‌بینند. فرایند آموزش ماتریس T نیز یک فرایند یک مرحله‌ای است. بردار i هر بخش تصویر ابر بردار مدل مخلوط گوسی بر روی زیر فضای تغییر کلی است.

---

[1] Expectation-Maximization Algorithm
[2] Cross Likelihood Ratio
[3] Normalized Cross Likelihood Ratio
[4] Hierarchical Agglomerative Clustering
[5] Integer Linear Program
[6] Universal Background Model (UBM)

$$m = M + Tx \qquad (۲-۳)$$

M ابر بردار مدل پس‌زمینه کلی است و m ابر بردار مدل مخلوط گوسی متوسط وفقی هر بخش است. پس برای هر بخش استخراج بردار i که x است شامل دو مرحله است: تطبیق مدل پس‌زمینه کلی برای به دست آوردن ابر بردار مدل مخلوط گوسی[۱] مرتبط با آن و استخراج فاکتورهای بردارهای ویژه‌ی تغییر کلی است. الگوریتم آموزش ماتریس T از گویندگان برچسب خورده داده‌های آموزش در [۲۳] شرح داده‌شده است. سیستم پیشنهادی ابزار تشخیص MSR را برای آموزش مدل پس‌زمینه کلی و آموزش استخراج بردار i استفاده می‌کند.[۲۴]



شکل ۳-۴ : استخراج بردارهای i

حین ارزیابی سیستم با بردارهای i، بعد زیر فضاها و انتخاب فواصل بین بردارهای i امتحان می‌شوند. دو فاصله‌ی متریک برای اندازه‌گیری شباهت بین بردارهای i بکار می‌رود.: شباهت متریک کسینوسی[۲] (۳-۳) و فاصله متریک مهالانوبییس[۳] (۳-۴) که W ماتریس کوواریانس بین دسته‌ای اندازه‌گیری شده از n بردار آموزش i از S گوینده شرح داده‌شده در بخش‌های قبلی است. فاصله‌ی مهالانوبیس ازاین‌پس کوواریانس نرمال بین دسته‌ای[۴] نامیده می‌شود. در (۳-۵) محاسبه‌ی WCCN برای فاصله‌ی مهالانوبیس، بردارهای $\overline{w^s}$ متوسط n_s بردار i از s گوینده است.

[۱] ابر بردار برداری است که بعد از اتصال همه‌ی بردارهای میانگین مدل‌های مخلوط گوسی به دست می‌آید

[2] Cosine Similarity Metric

[3] Mahalanobis Distance Metric

[4] within class covariance normalization

$$D(\mathrm{x},\mathrm{y}) = 1 - \frac{x^T y}{\|x\|.\|y\|} \qquad (۳-۳)$$

$$D(\mathrm{x},\mathrm{y}) = (\mathrm{x}-\mathrm{y})^T W^{-1} (\mathrm{x}-\mathrm{y}) \qquad (۳-۴)$$

$$W = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s} (\mathrm{w}_i^s - \overline{w^s})(\mathrm{w}_i^s - \overline{w^s})^T \qquad (۳-۵)$$

۳-۶-۳- انتخاب الگوریتم خوشه‌بندی

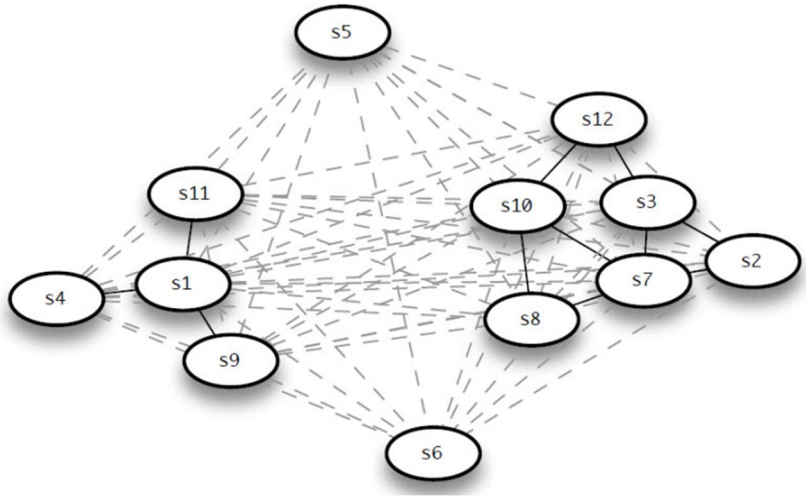
سیستم پیشنهادی با دو روش خوشه‌بندی کار می‌کند: الگوریتم خوشه‌بندی متراکم سلسله مراتبی سنتی[۱] و یک الگوریتم خوشه‌بندی گرافیکی به نام الگوریتم ILP[۲] که اخیراً در بحث بازشناسی گوینده در سال ۲۰۱۳ معرفی شد [۳۸] .

HAC[۳] -۴-۶-۳

در سیستم HAC، در هر مرحله اجرا مدل‌های گوینده‌ی خیلی مشابه برای ادغام انتخاب می‌شوند. در مرحله‌ی ادغام، یک مدل گوینده جدید با استفاده از داده‌های تمام بخش‌های هردو گوینده مشابه ادغام‌شده تولید می‌شود. ماتریس شباهت با شباهت مدل گوینده‌ی جدید از مدل‌های دیگر به‌روزرسانی می‌شود. در هر مرحله تعداد خوشه‌ها یکی کاهش می‌یابد. فرایند تا زمانی که تنها یک خوشه باقی بماند ادامه پیدا می‌کند. مجموعه بهینه‌ی خوشه‌ها از میان خروجی‌های هر مرحله بر اساس یک معیار بهینگی انتخاب می‌شوند. خوشه‌بندی سلسله مراتبی متراکم می‌تواند با استفاده از هریک از مدل‌های گوینده کار کند. در هر مرحله از خوشه‌بندی سلسله مراتبی متراکم، ادغام نیازمند

[1] Traditional Hierarchical Agglomerative Clustering Algorithm
[2] Integer Linear Program
[3] Hierarchical Agglomerative Clustering

محاسبات اضافی است چون مدل‌ها دوباره باید آموزش ببینند ماتریس فاصله باید طبق مدل گوینـده‌ی ادغام‌شده به‌روز شود.



شکل ۳-۵ : خوشه‌بندی ILPروی یک گراف کامل از مدل‌های گوینده[۲۹]

همان‌طور که در شکل ۳-۵ دیده می‌شود خوشه بند ILP بر روی گراف مـدل‌هـای گوینـده کـار می‌کند و با استفاده از معیارهای بهینه‌ی خوشه‌بندی به تصمیم‌گیـری بـرای انتخـاب خوشـه‌ی بهینـه می‌پردازد.

دو معیار خوشه‌بندی بهینه وجود دارد که برای این خوشه‌بندی اجراشده‌اند.اولین معیار آسـتانه‌ی فاصله است که در آن ، مرحله‌ای از اجرا که در آن حداقل شباهت دوبه‌دویی بزرگ‌تـر از یـک آسـتانه باشد،به‌عنوان دسته‌ی بهینه از خوشه‌ها شناخته می‌شود. (۲-۹)معیار بعدی معیار Ts خوشه‌ی بهینه‌ی معرفی‌شده در [۲۵] است ، مجموعه خوشه‌ها با حداقل شباهت بـرون دسـته‌ای و بیشـترین شـباهت درون دسته‌ای با استفاده از (۲-۱۰)به‌عنوان دسته‌ی بهینه انتخاب می‌شود.

۳-۶-۵- خوشه‌بندی ILP

در خوشه‌بندی ILP، مسئله‌ی K مرکزی برای به دست آوردن مجموعه دسته‌ها اصلاح می‌شـود. مسئله‌ی اصلی K مرکزی اصلی تشخیص K شهر از N شهر برای تشکیل مراکـز بـه‌طـوری اسـت کـه

بیشترین فاصله‌ی بین یک شهر و نزدیک‌ترین مرکز به آن حداقل شود. در مورد فرمول سازی ILP، N قسمت مشابه N شهر و K تا از قسمت‌ها به‌عنوان بهترین قسمت‌ها برای داده‌های K گوینده انتخاب شوند. بااین‌حال در مورد بازشناسی، تعداد گوینده‌ها (K) نامعلوم است؛ بنابراین مسئله خارج از مسئله‌ی بهینه‌سازی می‌شود.

مجموعه‌ی دودویی از متغیرهای تصمیم X_{ii} را در نظر بگیرید.

$X_{ii}=1$ مشخص می‌کند که خوشه‌ی i ام یک خوشه نماینده است.

$X_{ij}=1$ مشخص می‌کند که دسته‌ی i ام به خوشه نماینده‌ی j ام تعلق دارد و تخصیص داده‌شده است. (حتماً لازم است که $X_{jj}=1$ باشد)

توجه کنید که $X_{ji}=1$ و $Xij=1$ معانی متفاوتی دارند اگرچه هردوی آن‌ها نشان می‌دهند که بردارهای i برای بخش‌های i و j به یک خوشه تعلق دارند. اکنون، مسئله‌ی بهینه‌سازی زیر را در نظر بگیرید.

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} X_{ii} + \frac{1}{\delta}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}X_{ij} \\ s.t \quad & \sum X_{ij} = 1 \qquad \forall j \\ & X_{ij} \le X_{ii} \qquad \forall j \\ & d_{ij}X_{ij} \le \delta \qquad \forall i,j \\ & X_{ij} \le \{0,1\} \qquad \forall i,j \end{aligned} \qquad (۶-۳)$$

تابع هدف مدنظر برای بهینه‌سازی شامل دو قسمت است، اول N تعداد نمایندگان خوشه‌ها (تعداد گویندگان) و دوم δ پراکندگی کلی همه‌ی خوشه‌هاست. قید اول تأکید می‌کند که یک قسمت فقط به یک خوشه تعلق می‌گیرد. قید دوم بر این نکته اشاره دارد که مرکز خوشه به همان خوشه تعلق داشته باشد. قید سوم جلوی تخصیص یک بردار به خوشه را وقتی‌که از یک آستانه‌ی خاصی دورتر از مرکز خوشه است، می‌گیرد.

در نظر داشته باشید که الگوریتم خوشه‌بندی برنامه عدد صحیح خطی[1] نیازمند هیچ اطلاعات درباره‌ی اهدافی که قصد خوشه‌بندی داریم ندارد و تنها وابسته به ماتریس شباهت است. برنامه‌ی عددی نیازمند این است که به یک ILP تک‌بعدی تبدیل شود.

تنها معایب استفاده از الگوریتم خوشه‌بندی برنامه عدد صحیح خطی این است که مدل‌های گوینده در فرایند به دست آوردن مدل‌های گوینده در خوشه‌بندی سلسله مراتبی متراکم[2] مرتباً تعریف نمی‌شوند؛ بنابراین اگر اندازه‌ی بخش‌ها کوچک باشند بردارهای i ای که به‌عنوان رهبر بردارها انتخاب می‌شوند ممکن است نتوانند تمامی اطلاعات گوینده را در بربگیرند. این الگوریتم خوشه‌بندی می‌تواند همراه با دو مدل گوینده‌ها بکار گرفته شود- مدل‌های مدل مخلوط گوسی و بردارهای i- این به این دلیل است که تنها ماتریس شباهت برای استخراج مجموعه‌ی بهینه‌ی خوشه‌ها لازم است.

[1] Integer Linear Program (ILP)
[2] HAC

فصل ۴ : نتایج

۴-۱- مقدمه

در این بخش به بررسی و تحلیل و دلایل نتایج به‌دست‌آمده می‌پردازیم و در انتها پیشنهادات و کارهایی پیش رو را بیان می‌کنیم .

۴-۲- پارامترهای بخش‌بندی

منظور از پارامترهای بخش‌بندی دو پارامتر θ و λ در (۴-۱) هستند که برای کم کردن خطاها هرکدام را طوری انتخاب می‌کنیم که بهینه شوند.

اثر هشدار اشتباه تغییر گوینده این است که تعداد بخش‌بندی‌ها را بالا می‌برد و درنتیجه اندازه‌ی بخش‌ها پایین می‌آید.برای بررسی اثرات ناشی از هشدار اشتباه بهترین الگوریتم برای خوشه‌بندی (خوشه‌بندی ILPبا بردارهای i) مورداستفاده قرار گرفت .مشاهده‌شده است که مقادیر پایین θ و λ بخش‌بندی بیش‌ازحد را به وجود میاورند.به میزانی که θ را افزایش دهیم میانگین طول بخش‌ها نیز افزایش می‌یابد. درنتیجه‌ی این افزایش مدل‌سازی بهتری از گوینده‌ها را نتیجه می‌دهد و باعث کم کردن خطای DER می‌شود.

جدول ۴-۱ : مقادیر DER با بهترین الگوریتم‌های خوشه‌بندی(برحسب ٪)

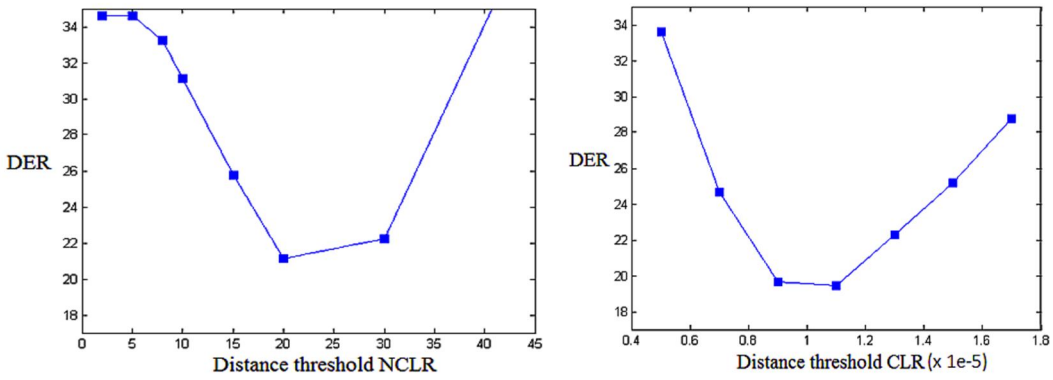
مقادیر پارامترها		λ	
		۱	۱۰
θ	۰	۳۱.۵۲	۳۳.۴۱
	۱۰۰۰	۲۳.۶۷	۱۶.۵۴
	۲۰۰۰	۱۲.۳۵	۱۶.۵۹

همان‌طور که در جدول (۴-۱) مشاهده می‌شود با ثابت نگه‌داشتن λ و افزایش پارامتر دیگر مقدار خطا در هر دو مقدار کاهش پیدا می‌کند.در مورد خود پارامتر λ با ثابت نگه‌داشتن θ و افزایش پارامتر دیگر مقادیر خطا افزایش می‌یابد پس درنتیجه‌ی جدول بالا ترکیب مقادیر $\theta = 2000$ و $\lambda = 1$ اتخاذ می‌شوند.

۴-۳- ارزیابی خوشه‌بندی گوینده‌ها

در این بخش ارزیابی خوشه‌بندی بر اساس خروجی‌های بخش‌های بخش‌بندی و تشخیص فعالیت گفتار را بیان می‌کنیم. در بخش‌های قبل خوشه‌بندی سلسله مراتبی سنتی را با استفاده از مدل‌های مخلوط گوسی و مدل‌های بردار i نشان دادیم. در ادامه الگوریتم خوشه‌بندی برنامه عدد صحیح خطی و مدل‌های مدل مخلوط گوسی و مدل‌های بردار i را فراهم می‌کنیم.

در این قسمت مقادیر DER نرخ خطای بازشناسی کلی هستند که میانگین DER های مجزای هر بخش که بر اساس مدت هر بخش وزن بندی شده است را نشان می‌دهد.در شکل (۴-۱) مشاهده می‌کنید با افزایش حد آستانه‌ی فاصله میزان خطای به وجود آمده در سیستم افزایش پیدا می‌کند و این به دلیل دسته‌بندی داده‌ها به‌صورت اشتباه در خوشه‌ی نادرست است .چون حد آستانه‌ی فاصله را زیاد کردیم پس داده‌های درون یک خوشه فاصله‌های زیادتری از هم پیدا می‌کنند و پراکندگی داده‌ها در یک خوشه افزایش پیدا می‌کند.این باعث می‌شود که این امکان که داده‌های خوشه‌های دیگر در این خوشه نمایان شوند افزایش پیدا می‌کند.لذا یک حد وسط برای آستانه‌ی فاصله لازم و ضروری است .



شکل ۴-۱ :: HAC با حد آستانه‌ی فاصله برای مدل‌های مخلوط گوسی

۴-۳-۱- آزمایشات خوشه‌بندی سلسله مراتبی متراکم با مدل‌های گوینده‌ی مخلوط گوسی

خوشه‌بندی سلسله مراتبی روی یک ماتریس شباهت بین دسته‌ها اجرا شد و نیازمندیم معیار توقف که بتواند مجموعه بهینه‌ی خوشه‌ها را مشخص کند دارد. دو معیار توقف بررسی شدند- معیار آستانه‌ی فاصله[1] و- معیار توقف بهینه T_S[2]

۴-۳-۲- معیار آستانه‌ی فاصله

در طول خوشه‌بندی سلسله مراتبی همان‌طور که تعداد اجراها زیاد می‌شود، داده‌های بدون خوشه کم می‌شوند تا زمانی که حداقل DER خوشه‌بندی به دست آید. تعداد اجراها مجموعه خوشه‌های بهینه را دنبال می‌کنند. این امکان نیز وجود داردکه بیش‌ازحد خوشه‌بندی هم رخ دهد.

۴-۳-۲-۱- معیار بهینگی Ts

با استفاده از معیار بهینگی [۲۵] که توسط (۲-۱۰) به دست می‌آید، خوشه‌ای با دورترین هیستوگرام برای فواصل درون خوشه‌ای و برون خوشه‌ای انتخاب می‌شود.در جدول۴-۲ نتایج نشان داده‌شده است. با استفاده از فاصله‌ی NCLR یک DER ۲۲.۱۵ درصدی به دست آمد درصورتی‌که با

[1] distance threshold criterion
[2] Ts optimal stopping criterion

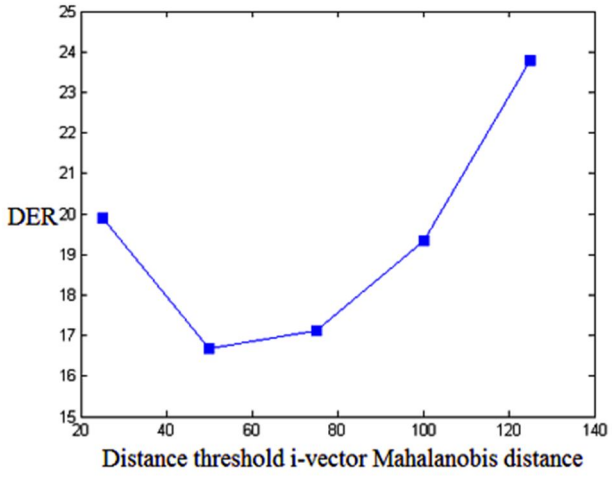
CLR یک DER ۱۹.۸۳ درصدی به دست آمد.

جدول ۴-۲:نتایج خطای بازشناسی بر اساس فواصل مختلف

فاصله مورداستفاده	(٪)DER
CLR	۱۹.۸۳
NCLR	۲۲.۱۵

۴-۳-۳- HAC با مدل‌های گوینده‌ی بردار i

با استفاده از خوشه‌بندی سلسله مراتبی با بردارهای i، بردارهای i جدیدی برای هر بخش به‌دست‌آمده در مرحله‌ی ادغام خوشه استخراج می‌شود.همان‌طور که در شکل (۴-۲) مشخص است کمترین میزان خطا در کار با فاصله‌ی مهالانوبیس به دست آمد که مقدار ۱۶.۶۹ درصد را به خود اختصاص می‌دهد.



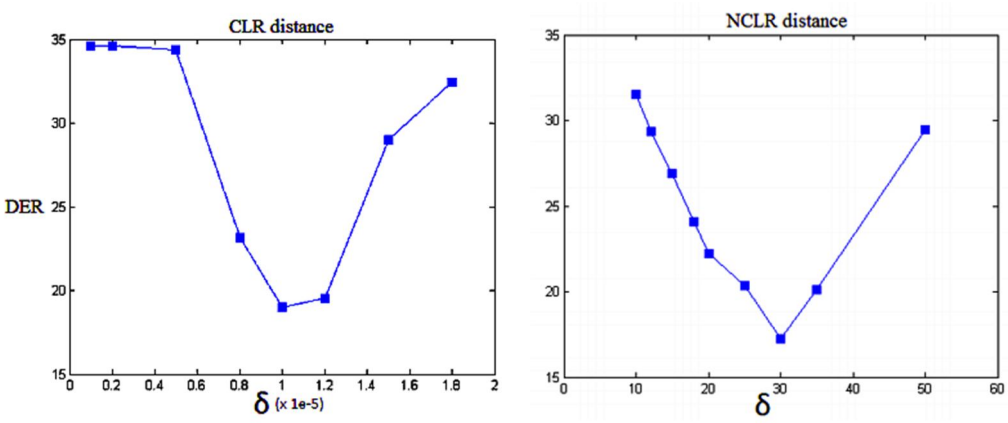
شکل ۴-۲ :HAC با آستانه‌ی فاصله برای مدل‌های مخلوط گوسی گوینده

۴-۳-۴- ILP با آزمایشات با مدل‌های گوینده‌ی مدل مخلوط گوسی

خوشه‌بندی برنامه عدد صحیح خطی با استفاده از فواصل CLRو NCLR برای ساخت ماتریس

فاصله اجرا شد.با توجه به شکل (۴-۳) بهترین نتیجه به‌صورت ۱۹.۰۳ درصد بود درصورتی‌که با استفاده از فاصله‌ی NCLR بهترین نتیجه ۱۷.۲۷ درصد بود. محور x آستانه‌ی حاضر در قیود مسئله‌ی بهینه‌سازی برنامه عدد صحیح خطی را نشان می‌دهد (۴-۶).NCLR نمایش بهتری از فاصله نسبت به CLR است هرچند برای استفاده در خوشه‌بندی سلسله مراتبی متراکم مناسب نیست. این بدین دلیل است که هرقدر اندازه‌ی خوشه‌ی ادغام‌شده افزایش یابد اندازه‌ی قسمت‌ها نقش بسزایی در کاهش فاصله‌ی NCLR دارد.

از طرف دیگر فرمول سازی برنامه‌ی خطی عددی یک مسیر همبسته برای دستیابی به خوشه‌بندی بهینه را پیشنهاد می‌دهد. برای اثبات این، فرمول سازی برنامه عدد صحیح خطی برای ماتریس‌های شباهت NCLR و CLR تولیدشده توسط مدل‌های گوینده‌ی مدل مخلوط گوسی اجرا شد .در این پایان‌نامه برنامه عدد صحیح خطی تنها با استفاده از بردارهای i محاسبه شد.

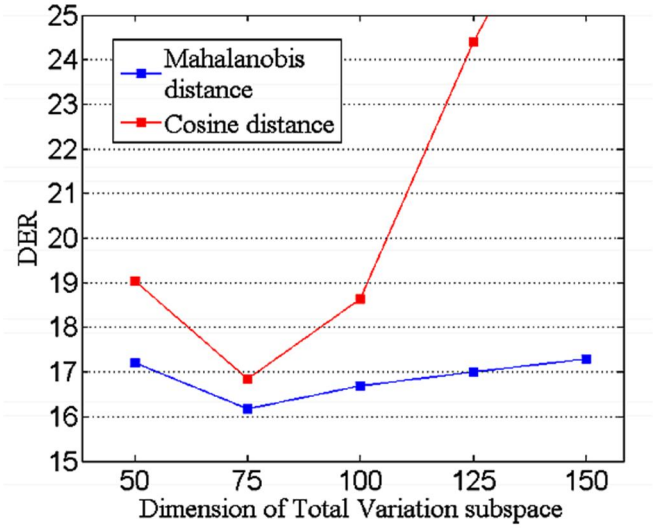


شکل ۴-۳: ILP با آستانه فاصله برای مدل‌های مخلوط گوسی گوینده

۴-۳-۵- خوشه‌بندی برنامه عدد صحیح خطی با مدل‌های گوینده‌ی بردار i

خوشه‌بندی برنامه عدد صحیح خطی با استفاده از بردارهای i آموزش‌دیده اجراشده‌اند. با بررسی آزمایش‌های انجام‌شده به بهترین بُعد فضای تغییر کلی و بهترین انتخاب فاصله دست می‌یابیم.مقادیر به‌دست‌آمده در فصول قبلی موردبحث و بررسی قرار گرفت . فاصله‌ی مهالانوبیس یک روش جبران

پس‌زمینه‌ای را پیشنهاد می‌دهد که شباهت بین بخش‌های با گوینده مشابه و پس‌زمینه متفاوت را بهبود می‌دهد.



شکل ۴-۴ : کارکرد خوشه یابی برنامه عدد صحیح با مدل‌های گوینده بردار i با ابعاد متفاوت از زیر فضای تغییر.

۴-۳-۵-۱- مدل مخلوط گوسی در مقایسه با بردارهای i

مدل‌سازی گوینده‌ها بر اساس مدل مخلوط گوسی یک نمایش بعد بالا از قسمت‌ها را می‌دهد و نیز اطلاعات پس‌زمینه را به‌خوبی در برمی‌گیرد. شباهت در پس‌زمینه می‌تواند سبب شباهت بین بخش‌های گویندگان مختلف شود. طرح جبران پس‌زمینه نیازمند به‌کارگیری در فضای ویژگی است. از طرف دیگر بردارهای i به جبران پس‌زمینه از طریق WCCN[1] امکان می‌دهد. موضوع دیگر در مورد مدل‌های گوینده با مدل مخلوط گوسی، زمان جبران بالای آن‌ها برای شباهت قسمت‌ها به دلیل عبارت‌های احتمال متقابل در (۲-۶) و (۲-۷) است.در جدول (۴-۳) خطای کلی در انواع ترکیب‌های روش‌های خوشه‌بندی و روش‌های مدل‌سازی را نشان می‌دهد.مشاهده می‌شود که در ترکیب خوشه‌بندی ILP با روش مدل‌سازی بر اساس بردارهای i به کمترین میزان خطا دست‌یافته‌ایم.

[1] Within Class Covariance Normalization

جدول ۴-۳:نتایج خطای DER از دو مدل گوینده و دو الگوریتم خوشه‌بندی

	HACخوشه‌بندی	ILPخوشه‌بندی
مدل مخلوط گوسی	٪۱۹.۴۵	٪۱۷.۲۷
بردار i-	٪۱۷.۱۱	٪۱۶.۱۸

۴-۳-۵-۲- HAC در مقابل ILP

خوشه‌بندی سلسله مراتبی باوجودآنکه یک الگوریتم افزایشی برای خوشه‌بندی است به‌عنوان یک تقریب خوب کار می‌کند. هرچند اگر طی یک مرحله در خوشه‌بندی، یک ادغام اشتباه رخ دهد، به‌طور مشخص کارایی مراحل بعدی را تحت تأثیر قرار می‌دهد. بنابراین یک تخمین مجدد از خوشه برای هر مرحله لازم است. بنابراین خوشه‌بندی سلسله مراتبی بیشتر از خوشه‌بندی برنامه عدد صحیح خطی هزینه دارد. خوشه‌بندی برنامه عدد صحیح خطی جستجوی دقیق‌تری نسبت به خوشه‌بندی سلسله مراتبی توسط بررسی کردن تمامی ترکیبات خوشه‌های ممکن انجام می‌دهد.

در این فصل به بیان نتایج آزمایشات و بررسی‌های به‌دست‌آمده بر اساس الگوریتم‌ها و روش‌های ارائه‌شده در فصل‌های قبل پرداختیم و به تحلیل و ذکر معایب و کمبودها و برتری‌های این نتایج پرداختیم.در ادامه آنچه اهمیت دارد تکمیل بحث با نتیجه‌گیری کلی از تمامی بحث‌های گفته‌شده است.در انتها ذکر راهکارهای پیش رو می‌تواند در ادامه‌ی روند این راه مؤثر باشد.

۴-۴- مقایسه نتایج به‌دست‌آمده

با توجه به گستردگی روش‌های خوشه‌بندی و الگوریتم‌های دسته‌بندی و نیز انواع مختلف

ویژگی‌های مورداستفاده در بحث‌های پردازش گفتار و بازشناسی گوینده و نیز با در دست داشتن انواع پایگاه داده‌های مختلف به زبان‌های متنوع بشری این حوزه از علوم گستردگی بسیاری دارد.این گستردگی حیطه‌ی مقایسه و بررسی نتایج با دیگر کارهای انجام‌شده را محصور کرده و این مقایسات را بسته می‌کند.در منبع [۶۸] نویسنده با ترکیب ویژگی‌های فضایی با ویژگی‌های طیفی بر روی پایگاه داده‌ی AMIبه همراه دیگر پایگاه‌های داده با بررسی گفتگوهای مختلف نتایج خوبی روی این پایگاه‌های داده به دست آوردند.آن‌ها با استفاده از این ویژگی‌ها و روش خوشه‌بندی HAC و با مدل‌سازی توسط HMM و GMM بر اساس فاصله‌ی BIC نتایج مطلوبی را در بازشناسی گوینده‌ها به دست آوردند.در منبع [۶۹] نویسنده با استفاده از ویژگی‌های زمانی بلندمدت برای تخمین حالات بعدی سیستم و با استفاده از ویژگی‌های آکوستیکی استخراج‌شده از فایل ورودی همانند MFCC و LP با مدل‌های GMM درزمینه‌ی بازشناسی گوینده بر روی پایگاه داده‌هایی چون AMI و NIST و ICSI نتایج درخور توجهی به دست آوردند.در جدول (۴-۴)در کنار روش پیشنهادی ما این دو روش نشان داده‌شده است.ذکر این نکته مهم است که درروش ما فرآیند و مدت‌زمان به دست آوردن نتیجه به دلیل برون‌خطی بودن تحلیل داده‌های در دست ما اهمیت چندانی نداشته است و ذکر نشده است هرچند در کاربردهای مختلف اهمیت زمانی مسئله و قابل انجام بودن آن ممکن است اهمیت خاصی داشته باشد.

جدول ۴-۴ : مقایسه با روش‌های دیگر

	روش پیشنهادی	منبع [۶۸]	منبع [۶۹]
درصد خطای DER	٪۱۶.۱۸	٪۱۸.۷	٪۱۲.۳

همان‌طور که در جدول(۴-۴) مشاهده می‌کنید در منبع [۶۸] به دلیل استفاده از ویژگی‌های زمانی بلندمدت ممکن است برخی تغییر در گوینده‌ها در این فواصل زمانی از دست برود و این خود باعث بالا رفتن میزان خطا می‌شود.در منبع [۶۹] به دلیل دربرداشتن ویژگی‌های آکوستیکی زمان کوتاه و

نیز ویژگی‌های بلندمدت کلی توانسته در مدل‌های گوینده به‌خوبی تمـامی ویژگـی‌هـا را در بربگیـرد و مدل‌سازی قابل‌توجهی را اراده دهد که نتایج خوبی را به دست داده اسـت.در مـدل پیشـنهادی مـا بـا به‌کارگیری ویژگی‌هایی نظیر MFCC ویژگی‌های آکوستیکی پوشش داده‌شده است و نیز با استفاده از بردارi برای مدل‌سازی توانسته‌ایم کلیت صوت در دست را دربربگیریم.

در آخر قابل‌ذکر است که تحلیل‌های ما فهـم شخصـی و بـه‌تنهـایی از دانسـته‌هـای خودمـان اسـت و نمی‌توان گفت تحلیلی کلی و همه‌جانبه درزمینه‌ی بازشناسی گوینده است زیرا در هـر کـاربردی و در هر پایگاه داده‌ای ویژگی‌ها و نتایج با دید و رویکرد خاصی باید بررسی شوند.

فصل ۵

نتیجه‌گیری و پیشنهادات

۵-۱- نتیجه‌گیری

هدف از این پایان‌نامه مطالعه‌ی روش‌های به‌روز در بحث بازشناسی گوینده و بهبود سیستم مربوطه بود. سیستم پیشنهادی با استفاده از نرخ خطای بازشناسی متریک ارزیابی شد. سیستم سه بخش اساسی دارد: تشخیص فعالیت گفتار، تشخیص تغییر گوینده بر اساس ΔBIC و بلوک خوشه‌بندی به‌روزشده .

هدف کلی سیستم تشخیص فعالیت گفتار توانایی حذف سکوت‌های غیر گفتاری همانند موزیک از فایل صوتی است. بخش خوشه‌بندی گوینده به مدل‌های جدید گوینده اجازه می‌دهد تا قسمت‌ها را با بردارهای i نمایش دهند که این راه گشای کارهای بعدی در زمینه بازشناسی سریع می‌تواند باشد.

سیستم قادر به اجرای تشخیص فعالیت گفتار بدون نیاز به داده‌های آموزش خارجی برای مدل‌های گفتار و غیر گفتار است. انرژی فریم و نرخ عبور از صفر به‌عنوان ویژگی‌های قوی برای ساخت مدل‌های آهنگ و سکوت بکار گرفته شدند. یک تشخیص گر رقابتی گفتار با یک سیستم SAD دومرحله‌ای به دست آمد: شاخص سکوت و به دنبال آن شاخص آهنگ. نتایج با سیستم جدید بر پایه‌ی مدل مخلوط گوسی که با استفاده از داده‌ی خارجی آموزش داده‌شده بود مقایسه شد.

مدل‌های گوینده‌ی بردار i که اکنون جدیدترین روش در بازشناسی گوینده هستند یک نمایش بُعد پایین از اطلاعات گوینده در مقایسه با مدل‌های مدل مخلوط گوسی سنتی ارائه داده است. همچنین آن‌ها یک مزیت محاسباتی ارائه داده‌اند زیرا محاسبه‌ی فاصله‌ی بین بردارهای i سریع‌تر از محاسبه‌ی شباهت بر اساس احتمال متقابل در مدل‌های مدل مخلوط گوسی است؛ بنابراین برای سیستم‌های بازشناسی زمان واقعی بردارهای i به نظر جذاب می‌آیند.

در این پایان‌نامه همان‌طور که در [۲۸] به دست آمد، خوشه‌بندی گوینده با استفاده از یک‌رویه‌ی بهینه‌سازی کلی برای به دست آوردن مجموعه خوشه‌های بهینه، نسبت به روش بهینه‌سازی افزایشیِ

الگوریتم خوشه‌بندی متراکم سلسله مراتبی عملکرد بهتری را داراست. روش خوشه‌بندی سلسله مراتبی متراکم ازنظر محاسباتی روش هزینه بری است و یک ادغام اشتباه طی دسته‌بندی باعث اثراتی در مراحل بعدی می‌شود. از طرف دیگر فرمولاسیون خوشه‌بندی برنامه‌ریزی خطی عددی یک مسیر همبسته برای خوشه‌بندی بهینه پیشنهاد می‌دهد. یک روش گرافیکی خوشه‌بندی برای مسئله‌ی رایج K مرکزی در بهینه‌سازی ترکیبی وفق داده شد. برای دستیابی به کارایی بهتر ILP در مقایسه با خوشه‌بندی سلسله مراتبی متراکم، فرمول سازی خوشه‌بندی برنامه عدد صحیح خطی برای ماتریس‌های شباهت NCLR و CLR تولیدشده توسط مدل‌های گوینده‌ی مدل مخلوط گوسی به دست آمد و اجرا شد و ۳.۲۷ درصد بهبود در خطاها را در مقایسه با بهترین خطای الگوریتم خوشه‌بندی مدل مخلوط گوسی -HAC حاصل شد. در متن، خوشه‌بندی برنامه عدد صحیح خطی تنها با استفاده از بردارهای i بررسی‌شده‌اند.

۵-۲- پیشنهادات

مسیر پیشرفت علم در تمامی حوزه‌ها همیشه بدون وقفه ادامه داشته است و محققان و پژوهشگران همواره در این راه نقش‌های اساسی و بسزایی را ایفا می‌کنند. در بحث بازشناسی گوینده در هر بخش و در هر الگوریتم و روش طرح‌های مبتکرانه و روش‌های جدید تا به امروز تحول و کارکردهای بهینه‌ی زیادی وارد کرده‌اند . در ادامه‌ی کار انجام‌شده در این پایان‌نامه می‌توان به ذکر اندکی از هزار راه‌کار موجود به‌صورت زیر پرداخت :

- اصلاح خروجی بازشناسی توسط عبور از یک رمزگشا ویتربی میتواندموثر باشد.

- بازشناسی متقابل عمل خوشه‌بندی گوینده‌ها در بین اسناد متفاوت است که برای تشخیص قسمت‌های با گوینده‌های یکسان مؤثر است. تلاش‌های حال حاضر درزمینه‌ی تحقیقات بازشناسی در راستای حل این مشکل است.

- بهبود در برنامه عدد صحیح خطی به‌طور مؤثر با کاهش بخش‌های زائد برنامه عدد صحیح خطی به اصلی اجرای سریع‌تری را دارد هرچند برنامه‌ی متلب[1] حل چنین مسائل بهینه‌سازی پیچیده‌ای در بعد بالا را پشتیبانی نمی‌کند. حل‌کننده‌هایی چونGUROBI[2] حل برنامه‌های خطی عددی را پشتیبانی می‌کنند.

-مشاهده شد که طی خوشه‌بندی گوینده، بخش‌هایی که دارای آهنگ پس‌زمینه بودند قادر به نمایش شباهت با بخش‌هایی که پس‌زمینه‌ی مشخص و تمیزی داشتند با استفاده از مدل‌های مدل مخلوط گوسی نبودند. حتی پس از استفاده از روش‌های جبران تغییر پس‌زمینه روی مدل‌های بردار i مسئله و مشکل همچنان پابرجاست.جداسازی گفتار و آهنگ پیش از پارامتر سازی باید موردتوجه قرار بگیرد تا آهنگ پس‌زمینه‌ی گفتار گوینده خنثی شود.

[1] Matlab
[2] http://www.gurobi.com/

فهرست منابع و مراجع

[۱] Martin, A., & Przybocki, M. (2000). The NIST 1999 speaker recognition evaluation—an overview. *Digital signal processing*, *10*(1), 1-18.

[۲] Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1435-1447

[۳] لشکر بلوکی م ،(۱۳۹۰)، پایان‌نامه کارشناسی ارشد:"تشخیص گوینده در محیط شامل چند گوینده با استفاده از ماشین بردار پشتیبان"، دانشکده برق و رباتیک، دانشگاه صنعتی شاهرود.

[۴] L.Docio, C.Garcia, ”Speaker Segmentation, detection and tracking in multi-speaker long audio recordings”, Third COST275 Workshop Bimetrics on the internet. 2005

[۵] Meignier, S., & Merlin, T. (2010, March). LIUM SpkDiarization: an open source toolkit for diarization. In *CMU SPUD Workshop* (Vol. 2010).

[۶] Bozonnet, S., Evans, N. W., & Fredouille, C. (2010, March). The LIA-EURECOM RT'09 speaker diarization system: enhancements in speaker modelling and cluster purification. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4958-4961). IEEE.

[۷] Nguyen, T., Sun, H., Zhao, S., Khine, S. Z. K., Tran, H. D., Ma, T. L. N ... & Li, H. (2009, May). The IIR-NTU speaker diarization systems for RT 2009. In *RT’09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA* (Vol. 14, pp. 17-40).

[۸] Arlindo Veiga, Carla Lopes, and Fernando Perdig~ao. Speaker diarization using gaussian mixture turns and segment matching. Proc. FALA, 2010

[۹] Maganti, H. K., Motlicek, P., & Gatica-Perez, D. (2007, April). Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-*

ICASSP'07 (Vol. 4, pp. IV-1037). IEEE.

[۱۰] Friedland, G., Janin, A., Imseng, D., Anguera, X., Gottlieb, L., Huijbregts, M., ... & Vinyals, O. (2012). The ICSI RT-09 speaker diarization system.*IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 371-381

[۱۱] Yella, S. H., Stolcke, A., & Slaney, M. (2014, December). Artificial neural network features for speaker diarization. In *Spoken Language Technology Workshop (SLT), 2014 IEEE* (pp. 402-406). IEEE

[۱۲] Ryant, N., Liberman, M., & Yuan, J. (2013, August). Speech activity detection on youtube using deep neural networks. In *INTERSPEECH* (pp. 728-731)

[۱۳] Anguera, X., & Bonastre, J. F. (2011, May). Fast speaker diarization based on binary keys. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4428-4431). IEEE.

[۱۴] Chen, S., & Gopalakrishnan, P. (1998, February). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop* (Vol. 8, pp. 127-132)

[۱۵] Anguera, X., & Hernando, J. (2005). Xbic: Real-time cross probabilities measure for speaker segmentation. *Univ. California Berkeley, ICSIBerkeley Tech. Rep*

[۱۶] Cheng, S. S., Wang, H. M., & Fu, H. C. (2010). BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. *IEEE transactions on audio, speech, and language processing*,*18*(1), 141-157

[۱۷] Moraru, D., Meignier, S., Fredouille, C., Besacier, L., & Bonastre, J. F. (2004, May). The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In*Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (Vol. 1, pp. I-373). IEEE

[١٨] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356-370.

[١٩] Reynolds, D. A., & Torres-Carrasquillo, P. (2004). *The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations*. MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB

[٢٠] Panagiotakis, C., & Tziritas, G. (2005). A speech/music discriminator based on RMS and zero-crossings. *IEEE Transactions on multimedia*, *7*(1), 155-166.

[٢١] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1), 19-41

[٢٢] Jensen, J. H., Ellis, D. P., Christensen, M. G., & Jensen, S. H. (2007, October). Evaluation of distance measures between Gaussian mixture models of MFCCs. In *International Conference on Music Information Retrieval* (pp. 107-108)

[٢٣] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798.

[٢٤] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," Speech and Language Processing Technical Committee Newsletter, IEEE, November 2013

[٢٥] Nguyen, Trung Hieu, Engsiong Chng, and Haizhou Li. "T-test distance and clustering criterion for speaker diarization." *INTERSPEECH*. 2008

[٢٦] Vijayasenan, D., Valente, F., & Bourlard, H. (2007, December). Agglomerative information bottleneck for speaker diarization of meetings data. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on* (pp. 250-255). IEEE

[۲۷] Meignier, S., Moraru, D., Fredouille, C., Bonastre, J. F., & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language*, *20*(2), 303-330.

[۲۸] Rouvier, M., & Meignier, S. (2012, June). A global optimization framework for speaker diarization. In *Odyssey* (pp. 146-150)

[۲۹] Dupuy, G., Meignier, S., Deléglise, P., & Esteve, Y. (2014). Recent improvements on ilp-based clustering for broadcast news speaker diarization. In *Proceedings of Odyssey*.

[۳۰] Anguera Miró, X. (2006). ROBUST SPEAKER DIARIZATION FOR MEETINGS

[۳۱] Sun, H., Ma, B., Khine, S. Z. K., & Li, H. (2010, March). Speaker diarization system for RT07 and RT09 meeting room audio. In *ICASSP* (pp. 4982-4985).

[۳۲] http://groups.inf.ed.ac.uk/ami/corpus

[۳۳] https://catalog.ldc.upenn.edu/LDC93S1

[۳۴] Saunders, J. (1996, May). Real-time discrimination of broadcast speech/music. In *icassp* (Vol. 96, pp. 993-996)

[۳۵] Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI signal processing systems for signal, image and video technology*, *20*(1-2), 61-79

[۳۶] Huijbregts, M. A. H. (2008). Segmentation, diarization and speech transcription: surprise data unraveled.

[۳۷] Dalal, N. INRIA Object Detection and Localization Toolkit (2008). *Software available at http://pascal. Inrialpes. Fr/soft/olt*

[۳۸] Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., & Meignier, S. (2013). *An*

open-source state-of-the-art toolbox for broadcast news diarization (No. EPFL-REPORT-192605). Idiap

[۳۹] Vijayasenan, D., & Valente, F. (2012). DiarTk: An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings. In *INTERSPEECH* (pp. 2170-2173)

[۴۰] Xavier, A. M. (2006). *Robust speaker diarization for meetings* (Doctoral dissertation, PhD Thesis, Speech Processing Group Department of Signal Theory and Communications Universitat Politecnica de Catalunya Barcelona (Espagnol)).

[۴۱] Luque, J., Anguera, X., Temko, A., & Hernando, J. (2008). Speaker diarization for conference room: The UPC RT07s evaluation system. In*Multimodal Technologies for Perception of Humans* (pp. 543-553). Springer Berlin Heidelberg.

[۴۲] Friedland, G., Gottlieb, L., & Janin, A. (2009, October). Joke-o-mat: browsing sitcoms punchline by punchline. In *Proceedings of the 17th ACM international conference on Multimedia* (pp. 1115-1116). ACM.

[۴۳] Delgado, H., Fredouille, C., & Serrano, J. (2014). Towards a complete binary key system for the speaker diarization task. In *INTERSPEECH* (pp. 572-576).

[۴۴] Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1557-1565.

[۴۵] Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., & Joly, P. (2012, June). A presentation of the REPERE challenge. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on* (pp. 1-6). IEEE.

[۴۶] Chan, W. N., Lee, T., Zheng, N., & Ouyang, H. (2006, May). Use of vocal source features in speaker segmentation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 1, pp. I-I). IEEE.

[۴۷] Siegler, M. A., Jain, U., Raj, B., & Stern, R. M. (1997, February). Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA speech recognition workshop* (Vol. 1997).

[۴۸] Bredin, H., & Poignant, J. (2013). Integer linear programming for speaker diarization and cross-modal identification in tv broadcast. In *the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH*.

[۴۹] Zhu, X., Barras, C., Meignier, S., & Gauvain, J. L. (2005, September). Combining speaker identification and BIC for speaker diarization. In*INTERSPEECH* (Vol. 5, pp. 2441-2444).

[۵۰] Favre, B., Damnati, G., Bechet, F., Bendris, M., Charlet, D., Auguste, R., ... & Fredouille, C. (2013, August). PERCOLI: A Person Identification System for the 2013 REPERE Challenge. In *SLAM@ INTERSPEECH* (pp. 55-60).

[۵۱] Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. A., & Glass, J. R. (2011, August). Exploiting Intra-Conversation Variability for Speaker Diarization. In *INTERSPEECH* (Vol. 11, pp. 945-948).

[۵۲] Kenny, P., Reynolds, D., & Castaldo, F. (2010). Diarization of telephone conversations using factor analysis. *IEEE Journal of Selected Topics in Signal Processing*, *4*(6), 1059-1070.

[۵۳] Silovsky, J., & Prazak, J. (2012, March). Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4193-4196). IEEE.

[۵۴] Johnson, S. E., & Woodland, P. C. (2000). A method for direct audio search with applications to indexing and retrieval. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on* (Vol. 3, pp.

1427-1430). IEEE.

[۵۵] Kubala, D. L. F. (1999). Fast speaker change detection for broadcast news transcription and indexing.

[۵۶] Wu, T. Y., Lu, L., Chen, K., & Zhang, H. (2003, January). Universal Background Models for Real-time Speaker Change Detection. In *MMM* (pp. 135-149).

[۵۷] Povey, D., Chu, S. M., & Varadarajan, B. (2008, March). Universal background model based speech recognition. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4561-4564). IEEE.

[۵۸] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1), 19-41.

[۵۹] Majetniak, A., & Tan, Z. H. (2011). Speaker recognition using universal background model on YOHO database.

[۶۰] Shum, S. (2011). *Unsupervised methods for speaker diarization* (Doctoral dissertation, Massachusetts Institute of Technology).

[۶۱] Senoussaoui, M., Kenny, P., Dehak, N., & Dumouchel, P. (2010, June). An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech. In *Odyssey* (p. 6).

[۶۲] Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., ... & Schwarz, P. (2010, March). Subspace Gaussian mixture models for speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4330-4333). IEEE.

[۶۳] Shum, S. H., Dehak, N., Dehak, R., & Glass, J. R. (2013). Unsupervised methods for speaker diarization: An integrated and iterative approach. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(10), 2015-2028.

[۶۴] Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings* (pp. 1-7).

[۶۵] Geiger, J. T., Wallhoff, F., & Rigoll, G. (2010, September). GMM-UBM based open-set online speaker diarization. In *INTERSPEECH* (pp. 2330-2333).

[۶۶] Siohan, O., & Bacchiani, M. (2013). Ivector-based acoustic data selection. In*INTERSPEECH* (pp. 657-661).

[۶۷] Lei, H. (2011). Joint Factor Analysis (JFA) and i-vector Tutorial. *ICSI. Web. 02 Oct*.

[۶۸] Zelenak, M., Segura, C., Luque, J., & Hernando, J. (2012). Simultaneous speech *detection* with spatial features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 436-446.

[۶۹] Yella, S. H., & Bourlard, H. (2014). Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1688-1700.

Abstract:

Nowadays, in the world of data processing with high speed and accuracy is very important. Processing speech Data because of the widespread use will contribute in all aspects of human life. Speaker diarization is to recognize who speaks when. The goal is to design a speaker recognition system that identify the speaker change in the audio files and correctly labeled and cluster each speaker's speech. This process is known speaker diarization named today. In this context, the aim is design a system that uses acoustic features MFCC and its first and second order along with energy and zero-crossing rate. Then model silence and music using frame that are absolutely non-speech or music and in tow-steps -silence removal and music removal- seprate speech from audio file. Therefore by using i-vectors in Feature space, reduce dimentional of system. We use integer linear programming (ILP) to label and cluster speaker and modify parameters in this models. Designed system, was examined on AMI corpus. We achieved good result and show in case of DER[1] error.

Keywords: speaker diarization, integer linear programming, i-vector, speech processing

[1] Diarization Error Rate

Shahrood University of Technology

Faculty of Electrical and Robotics Engineering

M.Sc. Thesis in Communication Systems Engineering

Speaker diarization based on mixed feature extraction in multi -speaker environment

By: Mitra Jahanian

Supervisor:

Dr. Hossein Marvi

Advisor:

Dr. Seyyed Masoud Mirrezaei

September 2016