دانشکده برق و رباتیک

گروه الکترونیک

پایان نامه کارشناسی ارشد

تشخیص گوینده در محیط شامل چند گوینده با استفاده از ماشین بردار پشتیبان

دانشجو: مرضیه لشکربلوکی

استاد راهنما:

دکتر حسین مروی

استاد مشاور:

دکتر حسین صامتی

تیر ماه 1390


بسم الله الرحمن الرحيم

تقدیم به:

خانواده عزیزم که همواره نقش اساسی در موفقیت هایم داشته اند.

دوستان عزیزی که همواره همراه و مشوقم بوده اند.

تقدیر و تشکر

سپاس و تشکر از زحمات و راهنمایی های علمی استاد گرانقدر جناب آقای دکتر حسین مروی (استاد راهنمای اینجانب) و جناب آقای دکتر حسین صامتی (استاد مشاور اینجانب). برای ایشان همواره آرزوی سلامتی و موفقیت روز افزون دارم.

چکیده:

شناسایی گوینده یکی از مباحث مطرح در بحث پردازش گفتار می باشد. شناسایی گوینده عبارت است از فرآیندی که طی آن با استفاده از سیگنال صحبت تشخیص دهیم چه کسی چه موقع واقعا صحبت می کند. هدف طراحی سیستمی است که بتواند تغییر در گوینده را مشخص نماید و گفتار هرگوینده را برای سیستم برچسب گذاری نماید. یعنی مشخص نماید که کدام گوینده، در چه بازه هایی صحبت کرده است. امروزه این عمل با یک عنوان جدید که هر دو فرآیند جداسازی و برچسب گذاری را در بر می گیرد بنام Speaker Diarization مشهور گشته است. هدف از بخش بندی تقسیم سیگنال گفتاری به بخش هایی است که تنها شامل گفتار یک گوینده هستند و هدف از خوشه بندی نیز شناسایی بخش های گفتاری مربوط به یک گوینده و اختصاص یک برچسب واحد به آنهاست.

هدف از انجام این پایان نامه طراحی و پیاده سازی یک سیستم بخش بندی و خوشه بندی گوینده با استفاده از الگوریتم های جدید و همچنین بهبود نتایج این الگوریتم ها برای این موضوع می باشد. این سیستم باید بطور صحیح نقاط تغییر گوینده را بدون دانستن اطلاعات قبلی از گوینده تشخیص داده و در نهایت تمام قسمت های صوتی مربوط به یک گوینده را در یک خوشه قرار می دهد.

در این پایان نامه، سیستم تشخیص گوینده، از سه مرحله اصلی تشکیل شده است. درمرحله اول قسمت-های غیر گفتاری، از بخش های گفتاری فایل صوتی حذف می شوند، تا دقت و سرعت عملیات سیستم در مراحل بعدی افزایش پیدا کند. سپس فایل گفتاری به بخش هایی همگن که در آن فقط گفتار یک گوینده وجود دارد، تقسیم می شود. در مرحله سوم با استفاده از خوشه بندی مناسب، بخش های گفتاری مرحله قبل، که متعلق به یک گوینده هستند، در یک خوشه جای می گیرند. جهت پیاده سازی سیستم از چهار نوع بردار ویژگی TDC, root-MFCC, MFCC و root-TDC و سه نوع پایگاه داده استفاده شده است و دقت مرحله بخش بندی 80٪ بوده است و دقت مرحله خوشه بندی نیز 59٪ با استفاده از ماشین بردار پشتیبان بدست آمده است.

کلمات کلیدی:

بخش بندی آماری گوینده

بخش بندی گویندگان

تشخیص بخش های صوتی

خوشه بندی گویندگان

فهرست مطالب

فصل اول: معرفی سیستم های تشخیص گوینده

1-1 –مقدمه..... 2

1-2-مراحل مختلف کاری سیستم های تشخیص گوینده..... 6

1-2-1- قطعه بند آکوستیکی..... 7

1-2-2-تشخیص گفتار از غیر گفتار..... 8

1-2-3-تشخیص جنسیت گوینده..... 9

1-2-4-تشخیص تغییر گوینده..... 9

1-3-روش های بخش بندی و خوشه بندی گویندگان..... 10

1-3-1-روش‌های بر اساس فاصله..... 10

1-3-2-روش‌های بر اساس مدل..... 11

1-3-3-روش‌های هیبرید یا ترکیبی..... 11

1-4-خوشه بندی نمودن..... 11

1-5- خلاصه..... 12

فصل دوم: تشخیص گفتار از نواحی غیرگفتاری

2-1-مقدمه..... 14

2-2-ساختار قسمت تشخیص گفتار از غیر گفتار..... 16

2-2-1-پیش پردازش..... 16

2-2-2-استخراج ویژگی..... 17

2-2-2-1-انرژی..... 18

2-2-2-2-نرخ عبور از صفر..... 19

2-2-2-3- استخراج ویژگی به کمک ضرایب کپسترال فرکانسی در مقیاس مل..... 19

2-2-2-4- ضرایب LPC..... 23

2-2-2-5- آنتروپی..... 24

2-2-2-6- اندازه متناوب بودن..... 26

2-2-2-7- اطلاعات زیر باند..... 28

2-2-2-8- سایر پارامترها..... 28

2-2-3- محاسبه آستانه.....29
2-2-4- تصمیمات VAD.....29
2-2-4-1- تصمیم گیری مبتنی بر مدل مخفی مارکوف.....30
2-2-4-2- تصمیم گیری مبتنی بر شبکه های عصبی.....31
2-2-5- تصحیح نتایج VAD.....33
2-3- بلوک دیاگرام چند VAD استاندارد.....33
2-3-1-استانداردETSI AMR 33
2-3-2- الگوریتم GSM.....34
2-4-خلاصه.....35

فصل سوم: آشکارسازی تغییر گوینده

3-1-مقدمه.....37
3-2-بخش بندی گوینده.....38
3-2-1-بخش بندی بر اساس فاصله.....38
3-2-2-بخش بندی بر اساس مدل.....40
3-2-3-بخش بندی هیبرید.....40
3-3-مقایسه روش‌های بخش بندی.....40
3-4-روش‌های متداول آشکارسازی گوینده.....41
3-4-1- معیار اطلاعات بیزین(BIC) 41
3-4-1-2- بخش بندی با استفاده از مدل آماری گوینده.....42
3-4-2- ترکیب آماره T^2 و BIC.....45
3-4-2-1- سرعت و بهره بیشتر در بخش بندی T^2-*BIC*.....۴۷
3-4-3- فاصله نرخ درستنمایی عمومی(GLR).....۴۹
3-4-4-فاصله KL2.....49
3-4-5- آشکارسازی تغییر گوینده با استفاده از DSD.....51
3-4-6-BIC متقاطع((Cross-BIC (XBIC).....52
3-4-7-درستنمایی مدل مخلوط گوسی (GMM-L).....53
3-5-خلاصه.....53

فصل چهارم: روش‌های دسته‌بندی

4-1-مقدمه.....55

4-2-اجزا سیستم خوشه بندی.....56

4-3-روش های خوشه بندی.....57

4-3-1-روش های خوشه بندی سلسله مراتبی.....58

4-3-1-1-تکنیک‌های خوشه‌بندی بالارونده.....59

4-3-1-2-تکنیک‌های خوشه‌بندی پایین رونده.....60

4-3-2-روش های خوشه بندی افرازی.....61

4-4- روش های خوشه بندی متداول در سیستم های خوشه بندی گوینده.....61

4-5- دسته‌بندی کننده ماشین های بردار پشتیبان.....63

4-5-1- دسته‌بندی کننده ماشین بردار پشتیبان خطی.....63

4-5-1-1- دسته‌بندی کلاس‌های جداپذیر.....63

4-5-1-2- دسته‌بندی کلاس‌های جدا ناپذیر.....68

4-5-1-3- دسته‌بندی داده‌های چند کلاسه با ماشین‌های بردار پشتیبان.....71

4-5-2- ماشین‌های بردار پشتیبان غیر خطی.....72

4-6- خلاصه.....74

فصل پنجم: پیاده سازی و مشاهدات سیستم ترکیبی پیشنهادی

5-1-مقدمه.....76

5-2-ساختار سیستم پیاده سازی شده.....77

5-3-پایگاه داده.....80

5-4-استخراج ویژگی.....82

5-5-معیار ارزیابی سیستم های تشخیص گوینده.....84

5-6-نتایج آزمایشات.....88

5-6-1- اثر اعمال VAD بر روی سیگنال گفتار.....88

5-6-2- اثر تغییر طول پنجره VAD بر روی دقت سیستم.....89

5-6-3- اثر تغییر طول پنجره BIC بر روی نتایج بخش بندی.....89

5-6-4-دقت.حاصل.از.بخش.بندی.بر.دو.نوع.از.دادگان با استفاده از MFCC.....93

5-6-5-اثر تغییر .بردار .ویژگی .بر .روی .دقت .مرحله .بخش بندی.....93

5-6-6-مقایسه .نتایج .مرحله .بخش‌بندی .با .بکارگیری .بردارهای .ویژگی متفاوت.....95

5-6-7-اثر جنسیت، گویندگان .بر تشخیص .درست .مرزهای .بخش بندی.....96

5-6-8-دقت مرحله خوشه‌بندی بکارگیری ماشین بردار پشتیبان(SVM) با بردار ویژگی MFCC.....96

5-6-9-دقت مرحله خوشه بندی ماشین بردار پشتیبان با بکارگیری بردار ویژگی root-MFCC.....97

5-6-10- اثر تغییر نوع تابع کرنل ماشین بردار پشتیبان بر روی دقت مرحله خوشه‌بندی.....98

5-7-خلاصه.....98

فصل ششم: جمع بندی و پیشنهادات

6-1-جمع بندی و خلاصه نتایج.....100

6-2-پیشنهادات.....101

منابع.....103

فهرست شکل ها

شکل (1-1): نمایش بخش بندی و خوشه بندی گویندگان روی گفتار ورودی.....4

شکل (1-2): ساختار کلی سیستم های بخش بندی و خوشه بندی گوینده.....6

شکل(2-1): دیاگرام یک VAD ساده.....16

شکل (2-2): نمایش پنجره همینگ 512 نقطه ای در حوزه زمان.....16

شکل (2-3): شمای کلی سیستم استخراج ویژگی 18

شکل (2-4): مراحل استخراج ویژگی با روش MFCC.....20

شکل (2-5): اعمال بانک فیلتر Mel scaled و محاسبه انرژی در هر زیر باند.....22

شکل (2-6): شبکه ای از HMM ها جهت بررسی دنباله احتمالی گفتار و سکوت.....31

شکل (2-7): دیاگرام ساده ای از یک VAD مبتنی برشبکه های عصبی.....32

شکل (2-8) : دیاگرام ساده ای از الگوریتم AMR2.....34

شکل (2-9) : دیاگرام الگوریتم GSM 35

شکل (3-1):پنجره های همسایه.....38

شکل(3-2) : ترکیب گوسین برای یک سیگنال شامل سکوت/گفتار.....39

شکل (3-3) :منحنی ها با اعمال متریک T^2-statistic.....46

شکل (4-1): انواع دسته بندی.....55

شکل (4-2): مراحل خوشه بندی.....56

شکل (4-3): روش های خوشه بندی.....57

شکل (4-4):روشهای خوشه‌بندی بالا و پایین رونده.....58

شکل (4-5): مثال ساده ای از خوشه بندی سلسله مراتبی.....60

شکل (4-6): یک نمونه از مسئله دو کلاسه خطی جداپذیر که نمونه‌ها توسط دو دسته‌بندی کننده خطی جدا شده.....64

شکل (4-7): حاشیه برای جهت 2 بیشتر از حاشیه در جهت 1 است.....65

شکل (4-8): نمونه‌ای از داده‌هایی که به صورت خطی به طور کامل از همدیگر جدا نمی‌شوند.....68

شکل (4-9): نمایش ماشین بردار پشتیبان غیر خطی.....74

شکل (5-1) : بلوک دیاگرام سیستم پیاده سازی شده.....76

شکل (5-2) : انتقال اطلاعات گفتار با استفاده از یک VAD 77

شکل (5-3) : دیاگرام الگوریتم G.729B 79

شکل (5- 4):بلوک دیاگرام بردار ویژگی TDC.....83

شکل (5-5): تشخیص خطا در سیستم های تشخیص گوینده.....87

شکل (5- 6): جداسازی قسمت های گفتاری از غیر گفتار.....88

شکل(5-7): اثر تغییر طول پنجره VADبر روی دقت سیستم.....89

شکل (5- 8): چگونگی قرار دادن یک آستانه و بعد انتخاب نقاط تغییر گوینده را نمایش میدهد.....90

شکل (5- 9): سیگنال گفتاری گوسی مدل شده در مرحله بخش بندی.....90

شکل(5-10): اثر افزایش طول پنجره BIC بر روی نتیجه مرحله بخش بندی برای 8 نفر دادگان فارس دات.....91

شکل(5-11): اثر افزایش طول پنجره BIC بر روی نتیجه مرحله بخش بندی برای 12 نفر دادگان فارس دات.....92

شکل(5- 12): اثر افزایش طول پنجره BIC بر روی نتیجه مرحله بخش بندی برای 18 نفر دادگان فارس دات.....92

شکل(5-13):مقایسه میزان خطای سیستم با تغییر بردار ویژگی مورد استفاده.....95

شکل(5- 14):تاثیر جنسیت بر روی خروجی مرحله بخش بندی سیستم.....96

شکل(5- 15): مقایسه نتایج خطای حاصل از خوشه‌بندی با تغییر نوع تابع کرنل بکارگرفته شده.....98

فهرست جداول

جدول (5-1): مقادیر خطا برای دادگان تهیه شده فارسی آزمایشگاهی.....93

جدول (5-2): مقادیر خطا برای دادگان AMI.....93

جدول (5-3): مقادیر خطا برای تعداد 3 نفر گوینده در دادگان فارس دات.....93

جدول (5-4): مقادیر خطا برای تعداد 5 نفر گوینده در دادگان فارس دات.....94

جدول (5-5): مقادیر خطا برای تعداد 8 نفر گوینده در دادگان فارس دات.....94

جدول (5-6): مقادیر خطا برای تعداد 11 نفر گوینده در دادگان فارس دات.....94

جدول (5-7): مقادیر خطا برای تعداد 14 نفر گوینده در دادگان فارس دات.....94

جدول(5-8): مقادیر خطا برای تعداد 17 نفر گوینده در دادگان فارس دات.....94

جدول(5-9): مقادیر خطا برای تعداد 20 نفر گوینده در دادگان فارس دات.....95

جدول(5-10): خطای حاصل از دسته‌بندی با استفاده از ماشین بردار پشتیبان با بکارگیری MFCC.....97

جدول(5-11): خطای حاصل از دسته‌بندی با استفاده از ماشین بردار پشتیبان با بکارگیری root-MFCC.....97

فصل اول :

معرفی سیستم های تشخیص گوینده

1-1-مقدمه

امروزه داده های چند رسانه ای بخش قابل توجهی از دانش انسان را در بر می گیرند. حجم پرونده های چند رسانه ای آرشیو شده در موسسه های مختلف در سال های اخیر افزایش چشمگیری داشته است. دسترسی و وضوح بالای این پرونده ها می تواند کمک شایانی به افرادی کند که در جستجوی اطلاعات باشند. بنابراین عملیات جستجو و بازیابی اطلاعات در این حجم بالا کاری است که خود احتیاج به سیستم کامپیوتری دارد. و درنتیجه یکی از حوزه های تحقیقاتی که به تازگی مورد توجه قرار گرفته است، مربوط به ساختاربندی پرونده- های چند رسانه ای است. در میان این داده ها، اطلاعات صوتی اهمیت بالاتری دارد. زیرا بخش اعظم آرشیوها حاوی داده های صوتی از گزارش های تلویزیونی، رادیویی و همچنین مکالمات تلفنی می باشد. در سالهای اخیر تحقیقات وسیعی در این حوزه آغاز شده و نتایج قابل قبولی نیز حاصل شده است. از دیگر کاربردهای این حوزه در تشخیص مجرم، جدا کردن صحبت های مهم یک شاهد یا متهم در دادگاه و ... میتوان اشاره نمود.

در کاربرد صوتی، عمده اطلاعات موجود در پرونده ها، صحبت های تعدادی گوینده است و هدف از سیستم نهایی، پاسخ به این سوال است که چه کسی در چه زمانهایی صحبت کرده است؟ بخش های مختلف این حوزه تحقیقاتی به نامهای مختلفی مانند: قطعه بند گوینده ای[1]، تشخیص گوینده[2] ،رونویسی قوی[3]، و اندیس گذاری گوینده ای[4] نامیده شده اند. از چنین سیستم هایی برای جابجایی راحت در داده های صوتی، در فایل های صوتی طولانی (مانند: اخبار و ملاقات ها و جلسات یک شرکت و ...) که متعلق به چند گوینده باشند بهره- برداری می شود. مکالمات و محاسبات رادیویی طولانی از محیط هایی هستند که در آنها چند گوینده حضور داشته و با هم صحبت می کنند. هدف نهایی چنین سیستم هایی، پیاده سازی روش هایی مناسب برای افراز پرونده صوتی به نواحی است که در آنها گوینده ای خاص صحبت کرده باشد. دسترسی راحت به

1.Speaker Segmentation
2.Speaker Diarization
3.Rich Transcription
4.Speaker Indexing

بخش هایی از صحبت یک گوینده توسط این سیستم فراهم می گردد. با داشتن حجم بالایی از داده های صوتی اهمیت این سیستم ها بیشتر می گردد.

با افزایش تعداد مدارک متنی موجود در اینترنت، نیاز به تکنیک هایی نظیر فهرست نگاری متن به منظور تسهیل دسترسی و جستجو در این مدارک افزایش پیدا کرد. نظیر همین نیاز نیز با افزایش تعداد مدارک صوتی نظیر سخنرانی ها، مصاحبه ها و گردهمایی ها و ... ایجاد شد. بطور مشخص دسترسی به مدارک صوتی بسیار سخت تر از دسترسی به متن است و گوش دادن به یک فایل صوتی ضبط شده بیشتر از خواندن متن زمان بر است و فهرست نگاری دستی مدارک صوتی در مقایسه با فهرست نگاری متن، مشکل است. راه حل پیشنهادی جهت رفع این مشکل، فهرست نگاری خودکار مدارک صوتی[5] است.

اولین بار سیستم هایی تشخیص گوینده توسط کمپانیNIST در سال 1999 ارایه شد. در سال2001، پلکان و سیدهارون به همراه گروهشان با استفاده از کم کردن اثر نویز بر روی سیگنال بهبودهایی در نتایج سیستم دادند و جداسازی بهتر گویندگان را باعث شدند. در سال 2005، بولیان و کنی با بکارگیری بردارهای ویژگی دیگر (یا ادغام روش های قبلی) و استفاده از مدل های گوسی در سیستم نتایج متفاوتی بدست آوردند. در سال 2005 توسط یاماشیتا و ماتسوناگا با استفاده از ویژگی های سیگنال صوتی مانند فرکانس پیچ سیگنال، انرژی، فرکانس های ماکزیمم سیگنال، و سه ویژگی دیگر نتایج در قسمت بخش بندی گوینده این سیستم بهبود داده شد.[1] و در سال های بعدی با انجام روش های مختلف برروی قسمت های متفاوت آن تا به امروز این سیستم ها در حال تکمیل شدن و بهتر شدن نتایج بوده اند.

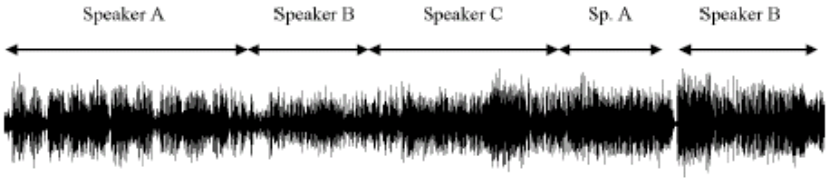
هدف از این پایان نامه، طراحی و پیاده سازی سیستمی است که بتواند در یک فایل صوتی که شامل گفتار چندین گوینده می باشد، تغییر در گوینده را مشخص نماید و تا حد امکان، گفتار هر گوینده را بدون دانستن اطلاعات قبلی از وی، دسته بندی نماید. این سیستم می تواند شامل دو بخش اساسی باشد که عبارتند از:

-بخش بندی گوینده

[1] .Automatic Audio Indexing

-خوشه بندی گوینده

کار قسمت بخش بندی[6]، تقسیم سیگنال گفتاری به سگمنت هایی است که تنها شامل گفتار یک گوینده هستند. در مرحله خوشه بندی[7]، شناسایی و دسته بندی بخش های گفتاری مربوط به یک گوینده و اختصاص یک برچسب واحد به آن انجام می شود. این مطلب در بسیاری از کاربردهای گفتاری که مربوط به بازشناسی یا فهرست نگاری[8] گفتار در محیطی که چندین گوینده ممکن است در آن اقدام به سخن گفتن بنمایند، مانند یک جلسه، کنفرانس، اخبار و نظایر آن کاربرد دارد. این کار نه تنها می تواند به سیستم های بازشناسی گفتار پیشرفته جهت بهبود نتایج بازشناسی گروهی کمک نماید بلکه در شناسایی و متن نگاری مکالمه ها نیز به آنها کمک می نماید. همانطور که قبلا نیز ذکر شد، امکان استفاده از آن در فهرست نگاری صوتی که امکان جستجو در فایل های صوتی را فراهم می نماید نیز ممکن است. شکل (1-1) نحوه کار این سیستم را بخوبی نشان می- دهد.



شکل (1-1): نمایش بخش بندی گویندگان روی گفتار ورودی

فایل صوتی مورد بررسی یک صوت ضبط شده تک کاناله است که شامل چندین منبع صوتی است. این منابع صوتی متفاوتند و می توانند شامل چند گوینده، موسیقی، انواع نویز و ... باشند. نوع و جزییات منابع صوتی موجود در فایل به ویژگی کاربردی آن فایل بستگی دارد.

بطور کلی سیستم های بخش بندی و خوشه بندی گوینده در سه حوزه زیر دارای کاربرد می باشند:

- دادگان اخباری
- جلسات ضبط شده
- مکالمات تلفنی

[6] .Segmentation
[7] .Clustering
[8] .Indexing

همانطور که قبلا نیز اشاره شد این سه حوزه تفاوت هایی مانند کیفیت ضبط صوت (پهنای باند، میکروفون ها و نویز) و میزان و نوع منابع غیرگفتاری، تعداد گویندگان، سبک و ساختار گفتار (طول مدت گفتار، ترتیب گویندگان) دارند و هر حوزه جهت کار بخش بندی و خوشه بندی گوینده، مسائل و مشکلات خاص خود را دارد. البته در سیستم های تشخیص گوینده سعی بر آن است تا برای هر سه حوزه کاری، نتایج قابل قبول و مناسبی حاصل شود.[1]

در سطح پایین تر کار چنین سیستمی دسته بندی داده های صوتی در خوشه هایی است که هر یک متعلق به یک گوینده باشد. در همین جا به راحتی میتوان دید که دو دیدگاه ناظرانه[9] (با سرپرست) و غیر ناظرانه[10] (بدون سرپرست) در این بخش مشاهده می شود. در دیدگاه اول از پیش اطلاعاتی از اینکه چه کسانی در فایل صوتی صحبت می کنند، وجود دارد. ولی در دیدگاه دوم کار سیستم دسته بندی فایل به بازه های زمانی است که در آنها تنها یک گوینده که هویت آن بر ما پوشیده است، صحبت می کند. توجه شود که میتوان از خروجی یک دسته بند غیرناظرانه به عنوان ورودی سیستم های شناسایی[11]، استفاده کرد و به این ترتیب یک سیستم دسته بندی ناظرانه خواهیم داشت. بنابراین کارایی و همچنین زمان اجرای سیستم ناظرانه بدست آمده بهتر است. از سوی دیگر، عملکرد این سیستم ها، به میزان اطلاعات قبلی مجاز نیز بستگی دارد. این اطلاعات قبلی می تواند نمونه گفتار از گویندگان، تعداد گویندگان موجود در فایل صوتی، یا اطلاعاتی از ساختار فایل ضبط شده باشد. ولی در اکثر سیستم های بخش بندی و خوشه بندی گوینده فرض بر نبود هیچگونه اطلاعات قبلی راجع به گویندگان و تعداد آنهاست. در این پروژه نیز با روش های بکار گرفته شده، فرض بر اینست که هیچگونه اطلاعات قبلی از گویندگان، مانند تعداد آنها، هویت آنها و داده آموزشی موجود نمی باشد و بنابراین مدل های گویندگان را نمیتوان از قبل آماده کرد. شکل (1-2) ساختار کلی سیستم های بخش بندی و خوشه بندی گوینده را نشان می دهد.

1.Supervised
2.Unsupervised
3.Identification

شکل (1-2): ساختار کلی سیستم های بخش بندی و خوشه بندی گوینده

چنین سیستمی شامل مراحل کاری مختلفی است و میتوان بخش های ذکر شده در قسمت های بعدی را برای آنها در نظر گرفت.[6-5]

1-2-مراحل مختلف کاری سیستم های بازشناسی گوینده

بطور کلی مراحل مختلف یک سیستم بازشناسی گوینده، بصورت زیر خلاصه می گردد:

1-قطعه بندی آکوستیکی[12]

2-تشخیص گفتار از غیر گفتار[13]

3-تشخیص جنسیت گوینده

4-تشخیص تغییر گوینده

5-جمع زدن گوینده های مشابه

1. Acoustic Segmentation Module
2.Speech Detection

این سیستم دارای بلوک های کاری مستقل از هم می باشد که هر بلوک ورودی خود را از خروجی بلوک قبلی دریافت می کند و ورودی لازم برای بلوک کاری پس از خود را تهیه می کند. در برخی سیستم ها، از بلوک سوم کاری صرف نظر می شود. در ادامه شرح مختصری از بخش های مختلف داده شده است.[4-2]

1-2-1-قطعه بند آکوستیکی

در اولین مرحله، باید جریان داده های صوتی به قطعات همگن آکوستیکی تقسیم شود. برای این امر باید نقاطی که تغییر در خواص آکوستیکی داده های صوتی روی میدهد را، بدست آورد. در واقع این نقاط شکست[14] بعنوان ورودی به بلوک کاری بعدی داده می شود. در بسیاری از کاربردهای چند رسانه ای که داده ها علاوه بر صدا دارای تصویر نیز می باشند، عمل تشخیص نقاط تغییر، هم از روی صدا و هم از روی تصویر امکان پذیراست.[2] بنابراین کارایی چنین سیستم هایی نسبت به داده هایی که تنها شامل صوت یا تصویر هستند، بالاتر خواهد بود.

امروزه روش های کاربردی تعیین نقاط تغییر آکوستیکی، همگی بر پایه ی محاسبه فاصله آماری بین دو قطعه مجاور استوار هستند. تفاوت عمده ی میان آنها معیار فاصله ای است که در آنها بکار می رود. از روش های غیر آماری مورد استفاده میتوان به شبکه عصبی[15] و ماشین بردار پشتیبان[16] اشاره نمود، که در بخش های بعدی توضیح داده خواهند شد.

از دیدگاهی قطعه بندی، یک مساله بهینه سازی[17] است. زیرا هدف نهایی یافتن نقاطی است که در آنها معیار فاصله به ماکزیمم محلی[18] برسد. یکی از پرکاربردترین معیارهایی که امروزه برای تعیین نقاط شکست آکوستیکی بکار می رود، معیار بیزین[19] است. پیش از این، روش های آماری دیگری از سال 1997 ابداع شده بود، که همگی آنها در مقایسه با معیار بیز جواب مناسبی نمی داده اند.[1] ارایه این روش اعتبار روش های دیگر را تا حدودی کمتر نمود.

1.Break Point
2.Artificial Neural Network
3.Support Vector Machine
4.Optimization
5.Local Maximum
1.Bayesian Information Criterion

1-2-2-تشخیص گفتار از غیر گفتار(دسته بندی[20] صوتی)

برای پیاده سازی این سیستم ها، قبل از هر کار دیگری بخش های گفتاری صوت ضبط شده را از بخش های غیر گفتاری آن مانند (سکوت، موسیقی، نویز خیابان، صدای سرفه ، صدای ورق زدن و ...) جدا می نمایند. با حذف بخش های غیرگفتاری میزان بار محاسباتی سیستم کاهش پیدا می کند و سرعت سیستم بیشتر می شود و سپس مراحل بخش بندی و خوشه بندی اجرا می شود. بعد از یافتن نقاط تغییر آکوستیکی، میتوان جریان داده های صوتی را مانند قطعات همگن در نظر گرفت. به عبارت دیگر یک قطعه نباید هم شامل گفتار، هم موسیقی و سکوت با هم باشد. اگر یک قطعه شامل گفتار دو گوینده باشد، باز هم همگن نخواهد بود. بنابراین این بلوک کاری خروجی قطعه بند صوتی را دریافت کرده و از آن قطعاتی را که حاوی داده های صوتی غیرگفتاری اند را حذف می کند. در یک سیستم تشخیص گفتار، معمولا داده های صوتی به 5 کلاس[2] زیر تقسیم می شوند:

1-موسیقی خالص

2-گفتار خالص

3-گفتار همراه با نویز

4-سکوت

5-سکوت همراه با نویز

البته در یک سیستم تشخیص گوینده، تنها احتیاج به تشخیص موارد 2و3 وجود دارد. زیرا هدف سیستم کار با گفتار بوده و هر چیزی غیر از گفتار از جریان داده ی صوتی حذف می شود تا بلوک های کاری پس از این بلوک با تمرکز برروی گفتار عمل نمایند. روشی که برای رسیدن به هدف این سیستم وجود دارد، بیشترین میزان شباهت(ML)[21] مبتنی بر مدل مخلوط گوسی(GMM)[22] می باشد.

2.Classification
[1].Maximum Likelihood.
[2]. Gaussian Mixture Model

1-2-3-تشخیص جنسیت گوینده

این بخش سیستم برای بهبود سرعت اجرای بلوک کاری خوشه بندی داده های گفتاری بکار می رود.[2] به این ترتیب که با برچسب خوردن هریک از قطعات گفتاری به عنوان مرد یا زن، فضای جستجو کاهش می یابد، زیرا لازم نیست که قطعات گفتاری با برچسب جنسی مخالف با یکدیگر مقایسه شوند. روش بکار رفته در این بلوک کاری نیز ML مبتنی بر GMM می باشد.

1-2-4-تشخیص تغییر گوینده

از جهت ترتیب و ترکیب بخش بندی و خوشه بندی نیز روش های موجود پیاده سازی شده در سیستم ها به دو دسته تقسیم می شوند: در روش اول یک روال دو مرحله ای [7-9] اجرا می شود. (همانند ساختار شکل (1-2)) که مرحله اول بخش بندی است. این مرحله مرز سگمنت ها را بر اساس تغییرات آکوستیکی سیگنال مشخص می کند. مرحله دوم خوشه بندی است که سگمنت های متعلق به هر گوینده را در یک خوشه دسته بندی می کند. نقطه ضعف این روش برطرف نشدن خطاهای ناشی از مرحله بخش بندی در پردازش های بعدی سیستم می باشد. و متعاقبا کارآیی مرحله خوشه بندی را نیز کاهش می دهد. در روش دوم بخش بندی و خوشه بندی بصورت توام [23] و تکراری[24] است. این روش کارآیی بیشتری در مقایسه با روش اول دارد. در این روش تکرارهای لازم با استفاده از مدل های پنهان مارکوف (HMM) پیاده سازی شده اند.[10]

1-3-روش های بخش بندی و خوشه بندی گویندگان

تاکنون روش های مختلفی برای بخش بندی و خوشه بندی گویندگان در یک جریان صوتی پیشنهاد شده است. روش های بکارگرفته شده را میتوان در سه دسته طبقه بندی نمود:

1) روش های بر اساس فاصله[25] [7][11]
2) روش های بر اساس مدل[26] [12-14]

---

[23] .Joint
[24] .Iterative
[25] .Distance-based
[26] .Model-based

3) روش های هیبرید یا ترکیبی[27] [15-17]

1-3-1-روش بر اساس فاصله

در این روش بخش بندی گویندگان به دو بخش اصلی تقسیم می شود:

1)آشکارسازی تغییر گوینده[28]: در ابتدا آشکارساز تغییر گوینده جریان صوتی[29] را به سگمنت های کوچکتر که شامل گفتار تنها یک گوینده هستند، تقسیم می کند.

2)خوشه بندی سگمنت ها(بخش ها): ادغام سگمنت های گفتاری متعلق به هر گوینده است. این ادغام با استفاده از یک معیار فاصله که شباهت بین دو سگمنت را اندازه می گیرد انجام می شود. مزیت این روش آن است که به هیچ اطلاعات قبلی نیاز ندارد ولی چون خوشه بندی بر اساس فاصله بین سگمنت های مجزا است وسگمنت های خیلی کوتاه نمی توانند به اندازه کافی مشخصات یک گوینده را توصیف کنند، بنابراین سگمنت های خیلی کوتاه روی دقت این روش تاثیر نامطلوب دارند.[16و17] عیب این روش، متکی بودن بر فاصله است، که موجب می گردد مقاومت و پایداری زیادی نداشته باشد.[15]

1-3-2-روش بر اساس مدل

در روش بخش بندی بر اساس مدل برای هر گوینده موجود در فایل صوتی با استفاده از داده های آموزشی یک مدل آموزش داده می شود. و این کار قبل از بخش بندی انجام می شود و سپس یک بخش بندی با استفاده از مدل های پنهان مارکوف[30] برای یافتن بهترین دنباله زمانی گویندگان[31] انجام می شود. جریان صوتی ورودی با استفاده از این مدل ها توسط انتخاب بیشترین درست نمایی دسته بندی می شود. در این روش، بخش بندی توسط بررسی ماکزیمم درست نمایی کلی[32] انجام می شود. در هر صورت بیشتر روش های بر اساس مدل به اطلاعات قبلی برای آماده سازی مدل های گویندگان نیاز دارند.

---

[27] .Hybrid
[28].Speaker Change Detection
[29].Audio Stream
[30].Hidden Markov Models(HMM)
[31].Best Time-aligned Speaker Sequence
[32].Global Maximum Likelihood Framework

1-3-3-روش هیبرید یا ترکیبی

این روش ها جدید و متنوع هستند و هنوز برای حصول نتایج بهتر بر روی آنها کار می شود. این روش ترکیبی از دو روش بر اساس فاصله و بر اساس مدل می باشد. نتایج حاصل از این روش در سیستم های پردازش گوینده بطور قابل ملاحظه ای بهتر از سایر روش های دیگر است.[15][17] برای نمونه در این روش یک الگوریتم بخش بندی بر اساس فاصله، تنها برای ساخت یک مجموعه اولیه مدل های گویندگان بکار می رود. سپس با شروع از این مدل ها، بخش بندی بر اساس مدل انجام می شود و با ترکیب خوشه بندی بر اساس فاصله و بر اساس مدل دقت خوشه بندی افزایش می یابد.

1-4-خوشه بندی نمودن

خوشه بندی کردن بخشی از علم دسته بندی غیر ناظرانه ی داده های آماری است. یعنی هیچ اطلاعی از نوع، مدل و حتی تعدادخوشه ها در اختیار نداریم و به صورت کورکورانه داده هایی را که با هم شباهت دارند در یک خوشه فرضی دسته بندی می کنیم. بیشتر از روش توده کردن سلسله مراتبی[33] که روشی آماری برای جمع کردن داده هایی است که به هم شباهت دارند، استفاده می شود. نکته مهم یافتن معیار شباهت است. به عبارت دیگر، اصول کلی فرآیند خوشه بندی کردن در تمام کاربردها یکسان است و فقط معیار شباهت برای هر کاربرد متفاوت است. قبلا دیدیم که معیار BIC بهترین معیار برای اندازه گیری شباهت دو قطعه است، بنابراین برای خوشه بندی نمودن نیز از BIC استفاده می شود. بدیهی است که با تغییر دادن معیار شباهت به الگوریتم های دیگری می رسیم که لزوما جواب یکسانی به ما نمی دهند.

1-5-خلاصه

در این فصل سیستم های تشخیص گوینده و مراحل مختلف سیستم معرفی شدند. سه مرحله اصلی کار سیستم توضیح داده شد. که این مراحل عبارتند از: مرحله اول، شامل جداکردن سکوت از سیگنال اصلی است و مرحله دوم، بخش بندی سیگنال خروجی حاصل از مرحله اول به سگمنت های همگن است. مرحله

4.Agglomerative Hirearchical

سوم نیز خوشه- بندی سگمنت های حاصل از مرحله دوم می باشد. انواع روش های بخش بندی و خوشه بندی توضیح داده شد. مزیت ها و معایب آنها گفته شد.

فصل دوم:

تشخیص گفتار از نواحی غیر گفتاری

2-1-مقدمه

سیگنال گفتار از دو ناحیه سکوت و غیرسکوت تشکیل می شود. آشکارسازی گفتار در حضور وقایع آکوستیکی غیرگفتاری و نویزهای زمینه، تشخیص گفتار از غیر گفتار نامیده می شود و با نام های VAD[34]، PDT[35] و یا EDT[36] بکار می رود. این عمل یکی از اجزاء مهم در برخی از کاربردهای پردازش گفتار نظیر تشخیص گفتار[37]، فشرده سازی اطلاعات گفتار[38]، تخمین و حذف نویزها[39]، سیستم بهسازی گفتار[40] و ... می باشد. برای بازشناسی گفتار (بازشناسی گوینده[41]) ،لازم است که گفتار آغشته به نویزهای محیطی مختلف را مورد پردازش قراردهیم. برای این منظور باید نقاط ابتدایی و انتهایی گفتار مشخص شود. این کار سبب می شود تا فرآیند بازشناسی گفتار تنها در آن بخش ها اعمال شود.[18] عدم استفاده از VAD کار تشخیص کلمات را پیچیده تر نموده میزان خطای بازشناسی گفتار را افزایش می دهد. در سیستم مخابرات و یا در سیستم فشرده سازی اطلاعات گفتار، میتوان با اختصاص دادن بیت های کمتر به نواحی سکوت به مقدار قابل توجهی در پهنای باند و یا فضای اختصاص داده شده صرفه جویی نمود، زیرا مشاهده شده که ضریب فعالیت گفتار (VAF[42]) یک گوینده بطور معمول بین 44-36٪ می باشد. این بدان معناست که 64-56٪ اطلاعات گفتار، شامل مکث ها و سکوت می باشد.

نویزهای موجود در گفتار (نویزهای زمینه) به دو گروه نویزهای ضربه ای و نویزهای غیرضربه ای یا تداوم دار تقسیم می شوند. نویزهای ضربه ای (مانند: صدای پف دهان، ته سرفه برای صاف کردن حنجره، کلیک و ضربه میکروفن و ...) توسط طول دوره آن نسبت به سیگنال گفتار قابل تشخیص می‌باشند. نویزهای تداوم دار

34. Voice Activity Detection
35. Pause Detection
36. End Point Detection
37. Speech Recognition
38. Speech Compression
39. Noise Estimation And Cancellation
40. Speech Enhancement
41. Speaker Recognition
42. Voice Activity Factor

(مانند: همهمه، صدای موتور ماشین، صدای فن، صدای کولر و ...) معمولا ماقبل و بعد از سیگنال گفتار وجود دارند. در VAD ها بیشتر این نوع نویزها مورد بررسی قرار می گیرند.

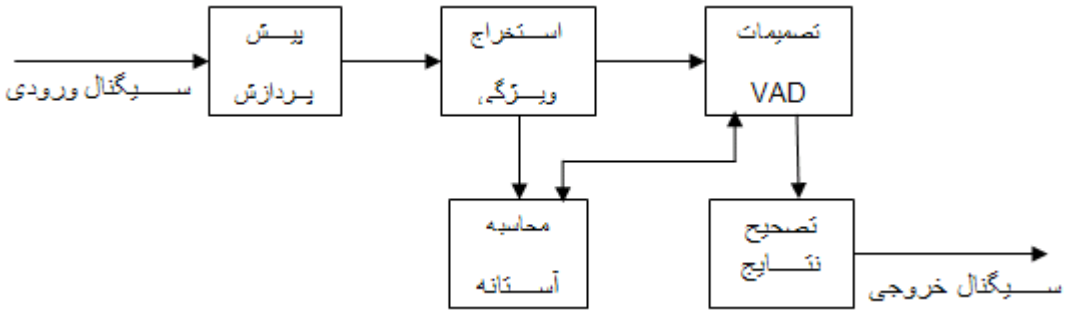
در تقسیم بندی دیگر، نویزها را میتوان به دو گروه سفید[43]و یا رنگی[44] تقسیم نمود. نویز سفید دارای پوش طیف تقریبا هموار و یکنواخت می باشد. ولی نویز رنگی دارای قطب ها و صفرهایی در پوش طیف است.

نویزها را میتوان از لحاظ ثابت بودن و یا متغیربودن مشخصه های آماری به دو دسته ایستان[45] و غیرایستان[46] تقسیم نمود. در واقعیت، نمیتوان ادعا نمود که یک نویز واقعا ایستان است. ایستان بودن نویز در دنیای واقعی فرضی است که میتوان آن را در دوره های زمانی کوتاه (مثلا فاصله بین رخداد سکوت) با درصد خطای اندکی پذیرفت. از لحاظ زمان پاسخگویی، VADها را میتوان به دو گروه دسته ای و بلادرنگ تقسیم نمود. در VADهای مبتنی بر پردازش دسته ای فرض بر این است که اطلاعات تمام فریم های گفتار و مقداری مناسب از فریم های سکوت ماقبل گفتار در اختیار می باشد. در این روش پس از ضبط سیگنال گفتار همراه با سکوت ماقبل، سیستم به فرآیند پردازش و تشخیص می پردازد. این روش ها معمولا در تصدیق و تعیین هویت گوینده و سیستم های شماره گیر صوتی و دستورات صوتی مورد استفاده قرار می گیرد. چون تمام اطلاعات گفتار یکجا در اختیار می باشد انتظار این است که این روش دارای دقت و عملکرد بالاتری نسبت به روش های بلادرنگ باشد. در روش های بلادرنگ تصمیم VAD و بازشناسی، همزمان با پایان گفتار با یک تاخیر کم، انجام می پذیرد. بنابراین منطقی است که این روش ها از نظر محاسباتی سریعتر باشند. ویژگی ها و مشخصه های بارز و مورد بررسی در VADها شامل اطمینان پذیری، مقاوم بودن در مقابل نویز، دقت، قابلیت تطبیق با شرایط جدید، سهولت، سادگی و ... می باشند.

1.White Noise
[44].Colored Noise
[45].Stationary
[46].Non Stationary

2-2- ساختار قسمت تشخیص گفتار از غیرگفتار

بطور معمول برای یک VAD میتوان دیاگرام شکل (2-1) را در نظر گرفت.

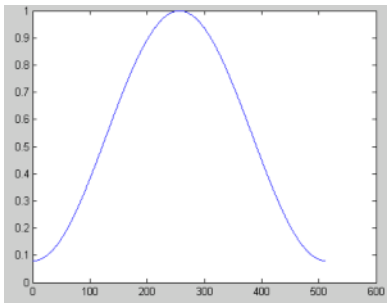


شکل(2-1): نمودار یک VAD ساده[80]

همانگونه که در این شکل دیده می شود تشخیص نواحی سکوت در سیگنال گفتار شامل مراحل زیر است.

2-2-1-پیش پردازش

در این مرحله اعمال پنجره، فیلتر کردن اطلاعات و ... انجام می شود. برای بررسی سیگنال، ابتدا سیگنال قاب- بندی[47] می شود. در اکثر الگوریتم های پردازش گفتار برای کمترکردن تاثیر لبه ها در طیف، از پنجره های همینگ یا هنینگ بجای پنجره مستطیلی و بطور هم پوشان استفاده می شود. شکل (2-2) نمایش یک پنجره همینگ در حوزه زمان می باشد.



شکل (2-2): نمایش پنجره همینگ 512 نقطه ای در حوزه زمان

در تحلیل سیگنال های صوتی عموما از تحلیل های طیفی بهره گرفته می شود. خصوصیت جالب سیگنال صوتی که در کار با آنها بسیار کارآمد است، نیمه ساکن[48] بودن آنها در حوزه فرکانس است. به این معنی که

[47] .Frame
1.Quasi - Stationary

در بازه های زمانی چند 10 میلی ثانیه ای رفتار سیگنال از نظر تحلیل فرکانسی ثابت می ماند. رفتار فازی اهمیتی ندارد، زیرا اطلاعات مهم برای گوش انسان، همان اطلاعات دامنه ی طیفی سیگنال است.[79] بنابراین این سیگنال ها در فواصل زمانی که تغییرات آکوستیکی محسوس ندارند، دارای ویژگی های منحصر به خود هستند. که با استخراج این ویژگی ها میتوان الگوی رفتاری سیگنال را مدل کرد. برای این امر پنجره ای با طول چند ده میلی ثانیه در نظر گرفته و با لغزاندن پنجره روی کل سیگنال، ویژگی ها در هر پنجره استخراج می شود. پنجره ها با همپوشانی هستند.[79]

2-2-2-استخراج ویژگی[49]

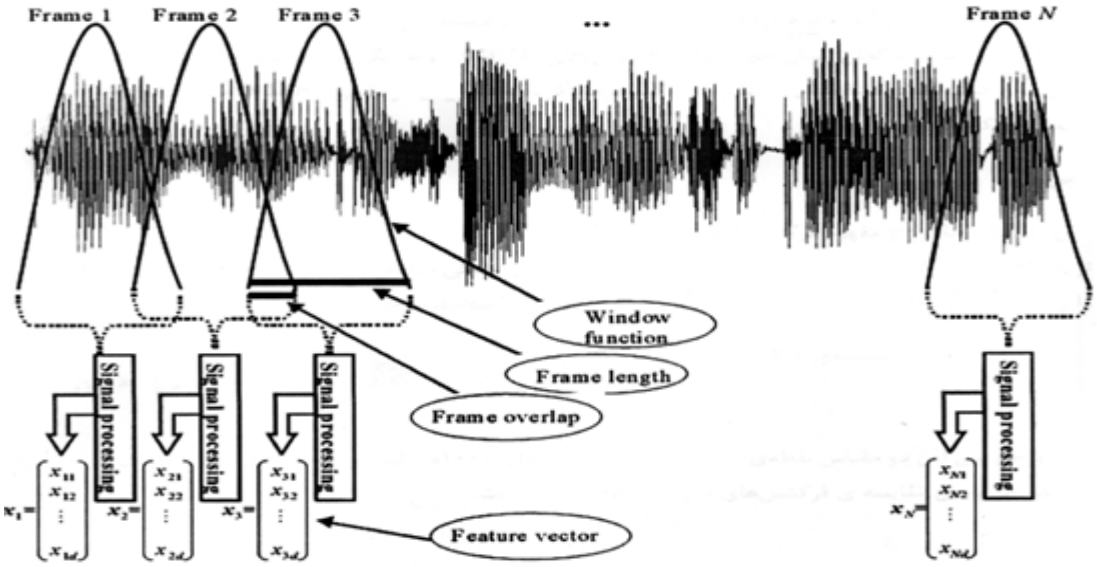
یکی از نکات مهم در پیاده سازی VADها، انتخاب ویژگی و یا ویژگی هایی است که بتوان به کمک آنها به تمایز دو ناحیه سکوت وگفتار پرداخت. در این مرحله پارامترهای مورد نیاز از فریم مربوطه استخراج می شوند. عموما پارامترهایی انتخاب می شوند که فاکتور خوبی برای تمایز[50] نواحی سکوت و غیر سکوت از هم باشند. از یک دیدگاه میتوان ویژگی های سیگنال گفتار را به دو دسته زیر تقسیم نمود:

1-ویژگی هایی که با مفاهیم سطح بالایی مانند: گویش(لهجه)، بستر سخن، شیوه ی صحبت کردن فرد خاص و مواردی مانند شرایط احساسی گوینده سرو کار دارند.

2-ویژگی هایی که با مفاهیم سطح پایینی مانند: فرکانس گام[51] (فرکانس پایه ای که تحت آن ارتعاش تارهای صوتی انجام می شود.) شدت صوت، فرکانس های مشتق[52]، خود همبستگی[53] طیفی سرو کار دارند. در حالت کلی این ویژگی ها با تحلیل طیف سیگنال در بازه های کوتاه زمانی بدست می آیند.[79] شمای کلی سیستم استخراج ویژگی در شکل (2-3) نشان داده شده است.

---

[49].Feature Extraction
[50].Threshold
4. Pitch
5. Formant
6.Autocorrelation



شکل (2-3): شمای کلی سیستم استخراج ویژگی [79]

در ادامه تعدادی از این ویژگی ها شرح داده شده اند.

2-2-2-1-انرژی

انرژی فریم یکی از ساده ترین و قدیمی ترین پارامترهایی است که به تنهایی و یا در کنار پارامترهای دیگر، مورد استفاده قرار گرفته است.[18-21] این پارامتر در SNR[54]های پایین، بدلیل بالا بودن انرژی نویز نسبت به انرژی سیگنال در نواحی رخداد اصوات با انرژی پایین، به تنهایی عملکرد بالایی ندارد. در [18] با بررسی طولانی مدت انرژی فریم جاری و محاسبه SNR فریم، عملیات تشخیص انجام شده است. در [61] با محاسبه $P = \frac{X-\mu}{\delta}$، که X: انرژی فریم، μ: میانگین انرژی نویز و δ: واریانس نویز می باشند و استفاده از یک ماشین حالت محدود در جهت تصحیح خروجی VAD، ماهیت فریم ها مشخص شده است. در [22] در ابتدا با یک روش VE[55]مناسب، سیگنال گفتار از نویز جداسازی می شود و سپس با کمک یک الگوریتم VAD مبتنی بر انرژی عملیات جداسازی انجام می شود.

2-2-2-2-نرخ عبور از صفر[56]

این پارامتر از طریق رابطه (2-1) محاسبه می گردد.

54.Signal to Noise Ratio
55.Voice Extraction
56.Zero Crossing Rate

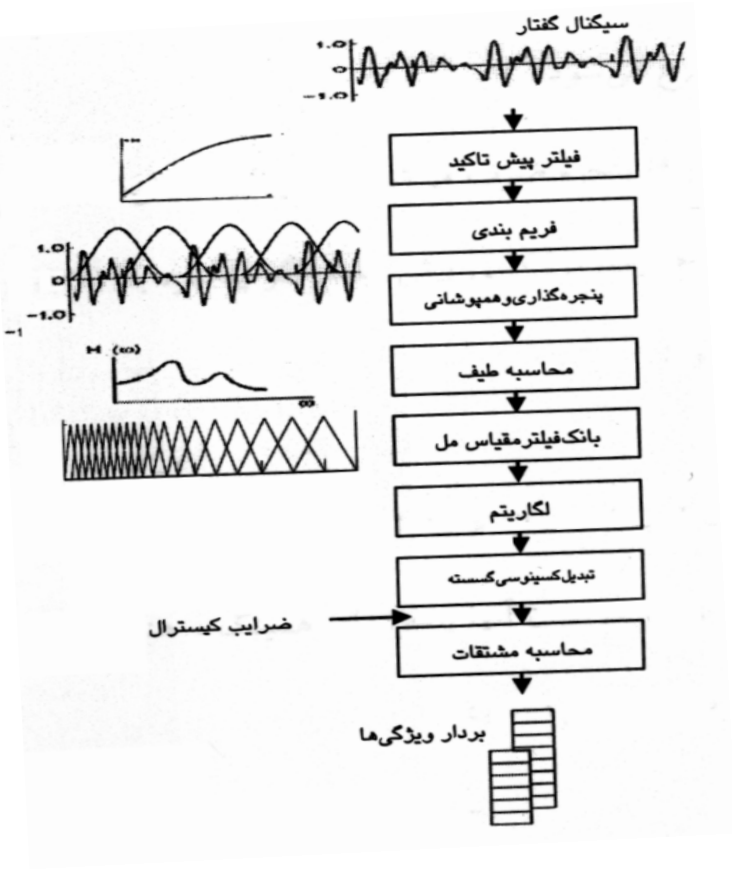
$$ZCR = \sum_{n=1}^{k-1} [1 - sgn(x(n+1)).sgn(x(n))]/2 \quad (1\text{-}2)$$

که در آن k اندازه پنجره و sgn() تابع علامت معمولی می باشد. استفاده از ZCR در شرایطی که انرژی پایین باشد، بسیار کمک کننده خواهد بود. بطور معمول این پارامتر در سیگنال گفتار، در یک رنج بخصوصی می باشد و در فریم های حاوی نویز این عدد تصادفی و اتفاقی خواهد بود. در اکثر الگوریتمهای VAD از پارامترهای انرژی و ZCR در کنار یکدیگر به همراه پارامترهای دیگر استفاده می شود.[21-27]

2-2-2-3-استخراج ویژگی به کمک ضرایب کپسترال[57] فرکانسی در مقیاس مل[58](MFCC)

منظور از ضرایب کپسترال، ضرایب کپسترال حقیقی می باشد. مطالعه روی نحوه شنیدن انسان و مدل گوش، نشان می دهد که درک انسان از محتوای فرکانسی، از یک مقیاس خطی پیروی نمی کند. برای شبیه سازی رفتار غیرخطی گوش در مقابل فرکانس ها از فیلتربانک استفاده می کنیم. در شکل (2_4) مراحل استخراج این ضرایب نشان داده شده است. در این سیستم ورودی سیگنال گفتار و خروجی بردارهای ویژگی (بردار ضرایب متناسب با آن گفتار) است. در ادامه توضیح مختصری از مراحل انجام کار داده می شود.[78]

[57].Cepsteral Coefficient
[58].Mel-Frequency Cepstral Coefficient



شکل(2-4): مراحل استخراج ویژگی با روش MFCC

1)فیلتر پیش تاکید: فیلتری بالاگذر که روی کل سیگنال اعمال می شود، تا اثرات طیفی نامطلوب مانند تغییرات ناگهانی موجود در سیگنال که در اثر نویزهای لحظه ای شدید به وجود می آید را حذف نماید و باعث یکنواخت شدن سیگنال می گردد. رابطه این فیلتر در حوزه زمان و در حوزه فرکانس بصورت روابط (2-4) و (2-5) می باشد.

$$h(n) = S(n) - \propto S(n) \qquad (4\text{-}2)$$

$$H(z) = 1 - \frac{\propto}{z} \qquad (5\text{-}2)$$

$\propto$: ضریب پیش تاکید است (معمولا $0.9 \leq \propto \leq 1$) در کارهای پردازش گفتار نزدیک به1($\propto = 0.97$) انتخاب می شود.

2)قاب بندی، پنجره گذاری و همپوشانی: ابتدا سیگنال را به قطعه های کوچکتر که آنرا قاب می نامند، تقسیم و ویژگی ها از هر قاب استخراج می شود. هر فریم یک بردار ویژگی را نتیجه می دهد. معمولا طول

هر قاب بین 10 تا 50 میلی ثانیه است و قاب ها با هم همپوشانی دارند. میزان همپوشانی بین آنها متفاوت (معمولا 25 تا 75 درصد طول قاب) انتخاب می شود. اگر طول قاب ها کوچکتر انتخاب شود، تعداد بردارهای ویژگی بیشتر و حجم محاسبات بالاتر می رود. و با افزایش طول قاب، تعداد بردارهای ویژگی و حجم محاسبات کمتر می شود ولی فرض ایستان بودن سیگنال در طول قاب خدشه دار می شود و اطلاعات کمتری از سیگنال استخراج می شود. قاب های بدست آمده، در یک پنجره که با w(n) نشان داده می شود، ضرب می شود. تا اثر ناپیوستگی سیگنال در ابتدا و انتهای هر قاب کم شود و تداخلی بین قاب ها در حوزه فرکانسی پیش نیاید. از انواع پنجره، مستطیلی، همینگ، هنینگ، ... وجود دارند. همینگ و هنینگ متداول تر هستند. که با رابطه های زیر نشان داده می شوند.[78]

$$Hamming: W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \leq n \leq N-1 \qquad (6\text{-}2)$$

$$Hanning: W(n) = 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \qquad 0 \leq n \leq N-1 \qquad (7\text{-}2)$$

اعمال پنجره به سیگنال مطابق رابطه زیر خواهد بود.

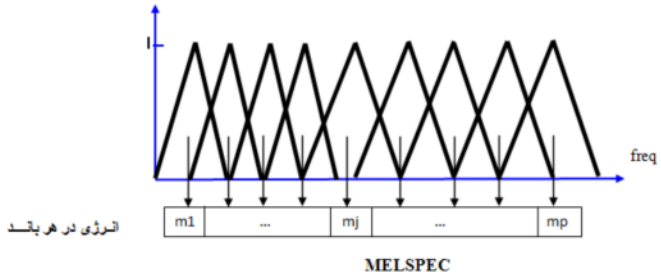
$$\bar{X}(n) = X(n).W(n) \qquad 0 \leq n \leq N-1 \qquad (8\text{-}2)$$

3)محاسبه طیف و بانک فیلتر در مقیاس مل: برای داشتن محاسبات راحت تر و سریعتر، با استفاده از تبدیل فوریه، سیگنال گفتار به حوزه فرکانسی برده می شود. تخمین طیف با استفاده از تبدیل فوریه سریع[59] انجام می- شود. چون گوش انسان در درک فرکانس های صوتی، با آن فرکانس رابطه خطی ندارد، ایده اعمال یک تبدیل غیرخطی به اسم مقیاس مل[60] روی طیف گفتار انجام می شود تا حساسیت گوش انسان را نسبت به حوزه های مختلف فرکانس مدل کند، یعنی مقیاس مل بیان می کند که گوش انسان به اطلاعات حوزه پایین ارزش بیشتری می دهد. به این ترتیب که برای فرکانس های کمتر از 1KHZ خطی است و برای فرکانس های بالاتر لگاریتمی عمل می کند. مقیاس مل با رابطه زیر تعریف می شود:

$$\hat{F} = mel(F) = 2596 .\log(1 + {}^{F}/_{700}) \qquad (9\text{-}2)$$

[59] .Fast Fourier Transform(FFT)
[60] .Mel scale

در این رابطه فرکانس Fبه $\hat{F}$ تبدیل می شود. سپس تعدادی فیلتر میان گذر هم اندازه با همپوشانی های یکسان روی طیف اعمال می شود و انرژی هر فیلتر را به عنوان یک ویژگی محاسبه می کنند.[63] شکل (2-5) این فیلتر را نمایش می دهد. انجام این کار قاب اولیه سیگنال را از تعداد چند صد تایی نمونه ها، به 20 تا 30 ویژگی (معمولا برابر با تعداد فیلترها N_f) کاهش می دهد. ویژگی هایی که با حذف اطلاعات اضافی، بسیاری از اطلاعات مفید سیگنال را نیز دارد.



شکل (2-5): اعمال بانک فیلتر مقیاس مل و محاسبه انرژی در هر زیر باند[16]

4)اعمال لگاریتم و تبدیل کسینوسی گسسته: با استفاده از رابطه (2-10) به منظور تعدیل دامنه ویژگی ها و بهبود ویژگی ها تبدیل غیر خطی زیر اعمال می شود.

$$E(i) = \sum_{k=0}^{M/2} (\log|H(k,m)|) \cdot F_i\left(\frac{2k\pi}{M}\right) \qquad (10\text{-}2)$$

F_i فیلتر iام و E(i) لگاریتم انرژی در باند iام می باشد.

مرحله نهایی در MFCC استفاده از تبدیل کسینوسی گسسته[61] روی ویژگی های بدست آمده جهت برگرداندن ویژگی ها به حوزه زمان و تقریب معکوس FFT است که با رابطه (2-11) انجام می شود. مزیت عمده این کار کاهش تعداد ویژگی ها (N_c) نسبت به تعداد فیلترها (N_f) است. ($N_c \leq N_f$) بعلاوه انجام این کار مستقل کردن ویژگی های بدست آمده و غیر همبسته کردن آنها را بدنبال خواهد داشت که منجر به قطری شدن ماتریس کواریانس ویژگی ها می شود.

$$C_i = \sum_{j=1}^{N_f} E(j) \cdot \cos\left(\frac{i\pi}{N_f}\left(j - \frac{1}{2}\right)\right) \qquad 1 \leq i \leq N_c \qquad (11\text{-}2)$$

[61] .Discrete Cosine Transform(DCT)

خروجی این تبدیل را ضرایب کپستروم[62] می نامند. این ضرایب غیر همبسته ترند و مولفه های پایین آن نشان دهنده اطلاعات مهم تر است. در مقابل مولفه های بالاتر دارای میزان اطلاعات کمتری در بازشناسی گفتار است تنها اطلاعات جزیی طیف و فرکانس را دارد که حذف آنها حتی میتواند در بهبود دقت سیستم موثر باشد.

5)محاسبه مشتقات ضرایب کپسترال: ضرایب استخراج شده از هر فریم فقط شامل اطلاعات استاتیک فریم است و این باعث می شود تا اثر فریم های مجاور در نظر گرفته نشود و بدلیل غیرایستان بودن سیگنال گفتار لازم است که بردار ویژگی هر فریم تغییرات ویژگی طیفی سیگنال را بازگو نماید.[63و62] بردار ویژگی هر فریم شامل مشتقات زمانی ضرایب استخراجی نیز می باشد. مشتقات زمانی هر فریم را میتوان با استفاده از دو روش بنام رگرسیون خطی[62] و روش تفاضل[63] که نسبت به روش قبلی ساده تر میباشد بدست آورد.[78]

2-2-2-4-ضرایب LPC[63]

ضرایب LPC از دیگر پارامترهایی است که میتوان از داخل هر فریم گفتار آنها را استخراج نمود. ضرایب LPCبرای فریم iام بصورت برداری $a_i =(a_i(0),a_i(1),\ldots,a_i(p))$ نشان داده می شود. (p درجه LPC میباشد.) این ضرایب بیشتر در کد نمودن اطلاعات با توجه به چگونگی ایجاد صوت در انسان به کمک یک فیلتر تمام قطب مورد استفاده قرار می گیرد. ایده استفاده از این بردار ویژگی در VAD ها بدین گونه است که سیگنال گفتار را به سه دسته صدادار[64]، بی صدا و سکوت (ناحیه نویزی فاقد گفتار) تقسیم می کنند. سپس بطور میانگین برای هر دسته یک سری ضرایب LPC از درجه P استخراج می شود. حال با ورود یک فریم مشکوک، پس از بدست آوردن ضرایب LPC فریم مزبور، فاصله این بردار از تک تک بردارهای متناظر با سه دسته فوق را بدست می- آورند. چگونگی بدست آوردن فاصله مزبور از راه های مختلف امکان پذیر می باشد. در [48] از محاسبه فاصله به روش Itakura استفاده شده است. در نهایت با بررسی فاصله بردار فریم

[62].Cepstrum
2.Linear Prediction Coding
[64].Voiced

مزبور با هر دسته، نزدیکترین دسته انتخاب می شود. در [41] از جمع مربع های ضرایب LPC، بعنوان یک ویژگی استفاده شده است.

2-2-2-5-آنتروپی[65]

استفاده از پارامتر انرژی در SNRهای پایین به دلیل بالا بودن انرژی نویز و پایین بودن انرژی برخی واج ها، کار VAD را دچار مشکل می کند. نواحی مربوط به گفتار در سیگنال گفتار، منظم تر[66] از نواحی مربوط به نویز می-باشد. برای بررسی نظم در مجموعه ها طبق قانون شانون[67] توسط رابطه (2-12) که برای بدست آوردن تعداد بیت های مورد نیاز به ازاء هر سمبل است، میتوان عمل نمود.

$$H(S) = -\sum_{i=1}^{N} P(s(i)).\log_2(p(s(i)) \qquad (2\text{-}12)$$

که در آن S=[s(1),s(2), … ,s(N)] و N تعداد سمبل ها، P(s(i)) احتمال رخداد سمبل s(i) می باشد. همانند رابطه (2-12)، به کمک رابطه (2-13) میتوان نظم را در طیف فرکانسی سیگنال بررسی نمود.

$$H(|Y(\omega,t)|^2) = -\sum_{\omega=1}^{\Omega} P(|Y(\omega,t)|^2).\log P(|Y(\omega,t)|^2) \qquad (2\text{-}13)$$

$$P(|Y(\omega,t)|^2) = \frac{|Y(\omega,t)|^2}{\sum_{\omega=1}^{\Omega}|Y(\omega,t)|^2} \qquad (2\text{-}14)$$

احتمال رخداد باند فرکانسی ω در فریم t ام می باشد. بیشترین مقدار بی نظمی در صورت رخداد نویزی کاملا تصادفی (نویز سفید) و در صورت رخداد یک تن خالص معادل صفر خواهد بود. میتوان با درنظر گرفتن قاب های ابتدایی بعنوان سکوت، مقدار آستانه ای برای بی نظمی بدست آورد و در نهایت با بدست آوردن این پارامتر در قاب جاری و مقایسه آن با مقدار آستانه، تشخیص را انجام داد.[31] این روش در صورت وجود نویز رنگی و سایر نویزهای منظم، کارایی ندارد. در صورت وجود این نوع نویزها، قبل از محاسبه آنتروپی، عملیات سفیدسازی نویز انجام می شود. برای انجام این کار توسط رابطه (2-15) سیگنال از یک فیلتر سفیدکننده عبور داده می شود.

---

[65] .Entropy
[66] .Organized
[67] .Shannon

$$|\bar{Y}(\omega,t)| = \frac{|Y(\omega,t)|}{\frac{1}{T}\sum_{i=1}^{T}|Y(\omega,t)|} \qquad (15\text{-}2)$$

در [32] نیز بهبودهایی بر روش های معمول مبتنی بر آنتروپی انجام شده است. مثلا برای بررسی آنتروپی باند فرکانسی 250-6000 هرتز، برای محاسبه، در نظر گرفته شده است و یا برای کم کردن اثر نویزهایی که در همه نقاط فرکانسی مولفه دارند (نویز سفید) و یا نویزهایی که در برخی زیرباندها به شدت مولفه دارند از رابطه (2-16) برای بهبود استفاده شده است.

$$P(|Y(\omega,t)|^2 = 0 \quad if \quad P(|Y(\omega,t)|^2 < \delta_2 \quad or \quad P(|Y(\omega,t)|^2 > \delta_1 \qquad (16\text{-}2)$$

در [32] محاسبه آنتروپی از روی وزن گذاری بر روی باندها انجام می شود. در همین مرجع، در نهایت بر روی اعداد آنتروپی بدست آمده، بر روی فریم های متوالی مجاور، فیلتر میانه ای اعمال شده است. در [33] با همین ایده آنتروپی پارامتر دیگری بنام LEC[68] در جهت تشخیص از روی رابطه (2-17) تعریف شده است.

$$LEC(n) = \frac{E - (E_1 + E_2)}{|E + E_1 + E_2|} \qquad (17\text{-}2)$$

که در آن E_2,E_1,E به ترتیب آنتروپی فریم جاری، آنتروپی فریم بعدی و آنتروپی فریم قبلی می باشد، نشان داده شده در صورتیکه LEC>0 فریم غیرایستان می باشد. در حقیقت فریم غیرایستان نشان دهنده گذار و فریم ایستان نشان دهنده حالت پایداری می باشد. میتوان با بررسی حالت LEC در سیگنال و حداقل طول دوره گفتار و سکوت، نواحی ایستان گفتار را تشخیص داد.

2-2-2-6- اندازه متناوب بودن[69]

این پارامتر یکی از پارامترهای رایج دیگری است که در VAD ها مورد استفاده قرار می گیرد. میتوان سیگنال گفتار را از لحاظ چگونگی ایجاد به سه دسته صدادار، بی صدا و سکوت تقسیم نمود. واج های صدادار به دلیل تحریک شدن و ارتعاش تارهای صوتی، دارای ماهیت تناوبی می باشند. از این نکته در

1.Local Entropy Criterion
2.Periodicity

تشخیص نواحی صدادار در سیگنال گفتار استفاده می شود. اما استفاده از این پارامتر در VAD کار تشخیص را در نواحی رخداد واج- های بی صدا کمی سخت می کند. نحوه محاسبه این پارامتر بطور خلاصه در زیر آمده است.[34] فرض می- کنیم که $s(i) = s_o(i) + n(i)$ و در آن $s_o(i)$، قسمت متناوب و $n(i)$ قسمت نامتناوب سیگنال $s(i)$ باشد بطوری که $s_o(i)= s_o(i+kp_o)$ و p_o تناوب $s_o(i)$ در طول قاب باشد. فرض می کنیم که از روی $s(i)$ یک تخمین از p_o یعنی $\hat{P}_0$ را بدست بیاوریم. برای این کار $\hat{S}_0(i,\hat{P}_0)$ و یا $\hat{S}_0(i)$ را که تخمینی از $s_o(i)$ است را بصورت رابطه (2-18) می نویسیم.

$$\hat{S}_o(i) = \sum_{h=0}^{k_o} \frac{S(i+h\hat{P}_0)}{k_o} \qquad 1 \le i \le \hat{P}_0 \quad , \quad P_{min} \le \hat{P}_0 \le P_{max} \tag{18-2}$$

که در آن P_{min} و P_{max} حداقل و حداکثر تعداد نمونه ها در تناوب گام، $k_o = [(\frac{N-i}{p_o})]$، تعداد تناوب های $\hat{S}_0(i)$ در قاب جاری باشد. هدف در اینجا بدست آوردن $\hat{P}_0$ است که با آن خطای مابین $\hat{S}_0(i)$ و $S(i)$ حداقل شود. برای محاسبه پارامتری که بتواند ما را در این کار کمک کند، فریدمن از رابطه (2-19) ، $R_1(\hat{P}_0)$ را محاسبه کرده است.

$$R_1(\hat{P}_0) = \frac{I_o(\hat{P}_0) - I_1(\hat{P}_0)}{\sum_{i=1}^{N} S^2(i) - I_1(\hat{P}_0)} \tag{19-2}$$

که $I_1(\hat{P}_0)$ و $I_o(\hat{P}_0)$از رابطه های (2-20) و (2-21) بدست می آیند.

$$I_1(\hat{P}_0) = \sum_{i=1}^{p} \sum_{h=0}^{k_o} \frac{S(i+h\hat{p}_0)^2}{k_0} \tag{20-2}$$

$$I_o(\hat{P}_0) = \sum_{i=1}^{N} \hat{S}_o^2(i) = \sum_{i=1}^{\hat{P}_0} \frac{\left[\sum_{h=0}^{k_o} S(i+h\hat{P}_0)\right]^2}{k_0} \tag{21-2}$$

برای هر قاب $R_1(\hat{P}_0)$ به ازاء $P_{min} \le P_0 \le P_{max}$ محاسبه می شود. ماکزیمم مقدار $R_1(\hat{P}_0)$ در حقیقت میزان قابلیت متناوب بودن قاب جاری می باشد. در حقیقت LSPE[70]، محاسبه مقدار قابلیت متناوب بودن در سیگنال می باشد. مقداری که برای یک قاب حاوی نویز سفید از این طریق بدست می آید حدود 0.5

[1].Least Square Periodicity Estimator

است. به همین دلیل مقدار بدست آمده را از 0.5 کم می کنند و در صورت منفی شدن، صفر را برای این مقدار اتخاذ می نمایند. بعد از هموار کردن مقدار بدست آمده در طول چند قاب (فیلتر میانه) و مقایسه تک تک مقادیر با یک مقدار آستانه، ماهیت قاب مشخص می شود. یکی از نکات مهم در محاسبه و بدست آوردن این پارامتر و بررسی آن، روش پیش پردازشی است که در [34] به آن اشاره شده است. در این فاز (پیش پردازش)، تن ها و برخی فرکانس های تداخلی بدلیل امکان ایجاد مشکل، حذف می گردند. این کار با بررسی فرکانس های هارمونیک گام و بررسی مولفه های فرکانسی در طول چند قاب قابل بررسی می باشد. این پارامتر به تنهایی در [34] و در جهت کمک به پارامترهای دیگر در پیاده سازی در [25]،[36و35] مورد استفاده قرار گرفته است. از پارامترهای دیگر وابسته به میزان متناوب بودن، اندازه پارامتر اختلاف پریود گام است که در [30] به آن اشاره شده و از طریق رابطه (2-22) قابل محاسبه می باشد.

$$P = \max\left[\frac{log\sum_{i=0}^{N-1}\left(S(t)S(t-\tau)\right)}{log\sum_{i=0}^{N-1}S(t)S(t)}\right] \qquad (2\text{-}22)$$

در این رابطه r بین 20 و 160 می باشد. مساله تناوبی بودن قاب گفتار، در [27] نیز به گونه ای در بررسی قابلیت اطمینان عملکرد VAD مورد استفاده قرار گرفته است.

2-2-2-7- اطلاعات زیر باند

یکی دیگر از روش هایی که در VADها مورد استفاده قرار می گیرد، بررسی انرژی فریم در زیرباندهای مختلف می باشد. برخی VADها مانند [42] به سادگی، با محاسبه انرژی در هر زیرباند و سپس محاسبه SNR در زیر باند مربوطه و در نهایت محاسبه SNR فریم و مقایسه آن با مقدار آستانه، ماهیت فریم را مشخص می کنند. در برخی VADها نظیر [43] از اطلاعات زیرباند برای تعیین صداداربودن، بی صدا بودن و یا سکوت بودن قاب به طریق زیر استفاده می کنند. برای تشخیص اینکه فریم موردنظر صدادار است، ابتدا سیگنال را از یک دسته فیلتر میان گذر گوسی، با مرکز فرکانس هایی که مضاربی از فرکانس گام می باشد، عبور می دهد. انرژی حاصله (E_f) محاسبه و با کمک انرژی کل (E)، عدد ρ از رابطه $\rho= E / (E-E_f)$ محاسبه می شود. بالا بودن ρ نشان دهنده صدا دار بودن سیگنال می باشد. در این مرجع برای شناسایی سیگنال بی صدا هم از رابطه $\gamma=E_h / E_l$ که E_h، انرژی در باند بالا و E_l، انرژی در باند پایین می باشد استفاده می شود.

در نهایت به کمک γ وρ ، VAD کارش را انجام می دهد. یکی از روش های تصمیم گیری در VADها استفاده از آزمایش مربع کای[71]می باشد.[81] این روش نیز بر اساس اطلاعات زیر باند قاب می باشد. از این آزمایش در دو قسمت، یکی در تخمین طیف نویز در جهت بهسازی از روش EM[72] که نوعی تفریق طیفی است و دیگری در قسمت تشخیص استفاده می شود. برای انجام عمل آزمایش، ابتدا با کمک بردار مشاهدات (فریم جاری) و بردار مقادیر موردانتظار (طیف تخمینی نویز) عدد مربع شای بدست می آید، عدد بدست آمده با مقدار آستانه مقایسه و عملیات تشخیص انجام می شود.

2-2-2- 8-سایر پارامترها

بردار واریانس خودهمبستگی[73] (AVV) پارامتر دیگری است که در [24] مورد توجه قرار گرفته است. در این روش قاب به چند زیر قاب از لحاظ زمانی تقسیم می شود. انرژی در هر زیر قاب E(i,j) (فریمiام، زیر فریم jام)، محاسبه می شود. سپس واریانس دنباله انرژی در زیر فریم ها محاسبه می شود. از روی مقایسه این مقدار با مقداری مربوط به نویز، ماهیت قاب مشخص می شود. در [36] نیز، از محدوده رخداد فرکانس گام، برای تشخیص استفاده شده است. به خاصیت ایستان بودن فرکانس گام طول چند قاب در [44] و خاصیت ایستان بودن طیف سیگنال گفتار در طول چند قاب، بهنگام رخداد یک واج صدادار نیز در [35] اشاره شده است. در [45] نیز با فرض ایستان بودن طیف نویز در طول دوره سکوت و انتخاب بردار ویژگی بر اساس انرژی قاب در زیرباندها و اعمال یک تابع تفاضلی بر روی بردار ویژگی قاب جاری با قاب های قبلی، نواحی سکوت تشخیص داده شده است. دو پارامتر دیگر مورد استفاده LTSE[74] وLTSD[75] می باشند. **که** ایده اساسی استفاده از این پارامترها، تغییرات زمانی اندازه طیف سیگنال می باشد.[38و37]

2-2-3- محاسبه آستانه

پس از استخراج پارامتر و یا پارامترهای دلخواه از داخل هر قاب، این پارامترها را با پارامتر متناظرش در قاب نویزی مقایسه می نماییم. سطح آستانه ای برای پارامتر مزبور در شرایط نویزی با توجه به اطلاعاتی که از

[71] .CHI-Square
[72].Ephraim and Malah
[73].Autocorrelation Vector Variance
[74] .Long Term Spectral Estimation
[75] . Long Term Spectral Divergence

قاب- های ابتدایی بدست می آوریم، در نظر گرفته می شود. این کار با فرض اینکه قاب های ابتدایی(مثلا 10 قاب) حاوی سکوت می باشند و با فرض اینکه نویز زمینه ایستان است، امکان پذیر می باشد. میتوان در طول انجام فعالیت VAD، در صورتیکه VAD قاب را سکوت تشخیص داد، به بهنگام سازی این مقادیر پرداخت.

2-2-4-تصمیمات VAD

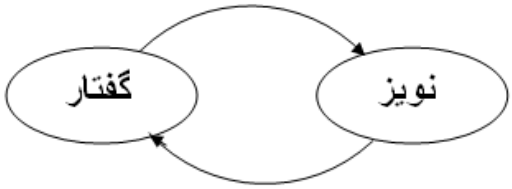
پس از استخراج ویژگی و در نظر گرفتن یک بردار ویژگی میانگینی برای نویز، میتوان از راه محاسبه فاصله بین این دو بردار، به بررسی تفاوت این بردارها از هم پرداخت. بدلیل اینکه نویز واقعا ایستان نیست، میتوان یک دامنه ای برای تغییرات میزان فاصله در قاب های سکوت درنظر گرفت. در اینجا با در نظر گرفتن مقدار آستانه- ای برای تفاوت ها، میتوان ماهیت قاب را از روی تفاوت بردار ویژگی قاب با بردار ویژگی نویز مشخص نمود. این عمل با رابطه ای ساده قابل انجام است. ولی با استفاده از روش های دیگر مانند اعمال مدل های فازی، شبکه عصبی، مدل های مارکوف مخفی، ... بر روی بردار ویژگی استخراج شده نیزمیتوان این عمل را انجام داد. در ادامه توضیح مختصری از مدل های معرفی شده داده شده است.

2-2-4-1- تصمیم گیری مبتنی بر مدل مخفی مارکوف

در این مدل، با در نظر گرفتن یک سری حالت، احتمال گذار حالات، یک بردار ویژگی و توابع توزیع عناصر بردار در هر حالت، ... اعمال الگوریتم جستجوی ویتربی[76]بر روی مشاهدات، بهترین دنباله حالت برای رخداد این مشاهدات، بدست می آید. در [46] از ویژگی های ZCR، انرژی، تابع متوسط اندازه تفاضلات AMDF[77] و اعمال آنها بر روی مدل مارکوف مخفی تک حالته، دو حالته و یک شبکه رایج حالت های مختلف بررسی شده است. در این روش میتوان برای نویزهای مختلف مدل مارکوف مناسبی را در نظر گرفت و سپس به کمک الگوریتم ویتربی میتوان نواحی سکوت و حتی نویز را تشخیص داد. در [47] با در نظر گرفتن ویژگی اختلاف لگاریتم انرژی فریم و لگاریتم انرژی نویز و ویژگی دلتای انرژی ، به عنوان بردار ویژگی و با در نظر گرفتن دو مدل مارکوف متناظر با گفتار و سکوت، نواحی سکوت تشخیص داده شده است. در این مرجع

[76] .Vterbi Algorithm
[77].Average Magnitude Difference Function

برای مدل گفتار، یک مدل مخفی مارکوف چهار حالته و برای نویز یک مدل مخفی مارکوف سه حالته در نظر گرفته شده است. سپس الگوریتم جستجوی ویتربی، برای یافتن بهترین دنباله از حالات و در نهایت بهترین و محتمل ترین دنباله از سکوت و گفتار، بر روی شبکه ای مانند شکل (2-6) اعمال شده است. در [48] با در نظر گرفتن بردار ویژگی ای متشکل از 12 ضریب MFCC و توان سیگنال و یک مدل مارکوف 7 حالته متشکل از حالات سکوت قبل از گفتار، گفتار (5 حالت)، سکوت بعد گفتار به بازشناسی نواحی غیرسکوت (2 تا 6 حالت) پرداخته شده است. با استفاده از روش مبتنی بر موجک جهت بهسازی گفتار و استفاده از مدل مارکوف 3 حالته (سکوت قبل از گفتار، گفتار، سکوت بعد از گفتار) نواحی سکوت تشخیص داده شده است.



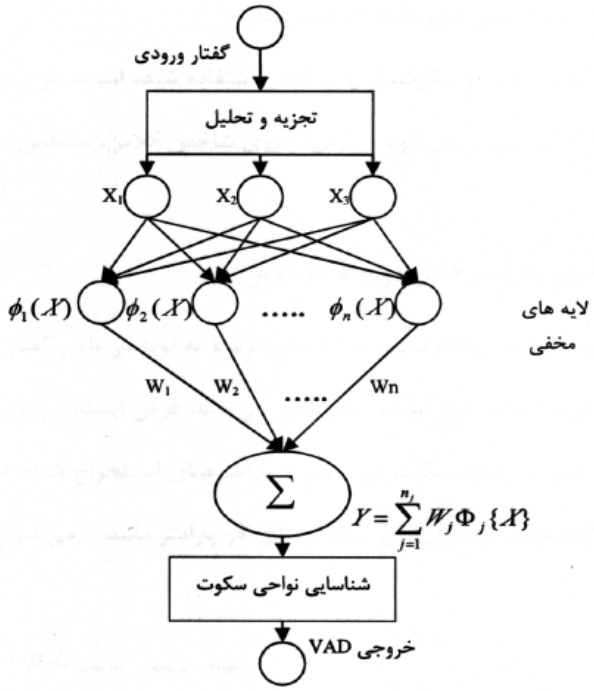
شکل(2-6): شبکه ای از مدل مخفی مارکوف جهت بررسی دنباله احتمالی گفتار و سکوت [47]

در [49] از یک مدل مارکوف دو حالته(گفتار و سکوت) و در نظر گرفتن احتمال گذار متغیر و بردار ویژگی مشابه با بردار ویژگی مورد استفاده در استاندارد G.729B، نواحی سکوت تشخیص داده شده است. مقدار احتمال گذار از حالت گفتار به حالت سکوت از زمان ورود به حالت گفتار، به مرور زمان کمتر می شود. در [50] سیگنال گفتار از لحاظ چگونگی بیان به 14 دسته تقسیم و با در نظر گرفتن 14 مدل و یک مدل برای نویز، یعنی با 15 مدل سه حالته عملیات تشخیص انجام شده است.

2-2-4-2-تصمیم گیری مبتنی بر شبکه های عصبی

در این روش بعد از فاز استخراج ویژگی، برای تصمیم گیری از یک شبکه عصبی استفاده می شود. شبکه های عصبی دارای ورودی، خروجی و لایه های مخفی میانی[78]می باشند. شکل (2-7) دیاگرام ساده ای از این روش می باشد.

[78] .Hidden Layers



شکل (2-7): نمودار ساده ای از یک VAD مبتنی برشبکه های عصبی[30]

در [30] در لایه های میانی مقدار φ(X) ویژگی X از رابطه (2-23) محاسبه می شود.

$$\emptyset(x) = \exp\left(\frac{(x-c)^2}{p^2}\right) \qquad (2\text{-}23)$$

که در آن C مرکز و مقدار میانگین و P دامنه تغییرات ویژگی مذکور می باشد. خروجی Y که یک تابع خطی از مقادیر لایه های آخرین سطح در لایه های پنهان است از رابطه (2-24) محاسبه می شود.

$$Y = \sum_{i=1}^{n_i} W_i \emptyset_i(x) \qquad (2\text{-}24)$$

که n تعداد واحدهای محاسباتی می باشد. پارامترهای استفاده شده در [30] شامل انرژی، مجموع مربع های ضرایب LPC و پارامترهای وابسته به فرکانس گام می باشد. در [54] از یک شبکه 3 لایه ای با 400 گره مخفی و استفاده از پارامترهای انرژی، اعوجاج طیفی (نسبت انرژی باند بالا به انرژی باند پایین) و میزان صدادار بودن قاب، استفاده شده است. در [55] بجای در نظر گرفتن دو کلاس گفتار و سکوت، از چند کلاس استفاده شده است. در این مرجع با در نظر گرفتن ضرایب MFCC تعلق قاب به هر کلاس بررسی می شود و سپس از روی شاخص کلاس، مشخص می شود قاب مورد نظر گفتار و یا سکوت می باشد.

2-2-5-تصحیح نتایج VAD

در این مرحله با بررسی میانگین طول دوره گفتار و میانگین طول دوره سکوت از تشخیص غیر صحیح قاب حاوی سکوت در بین قاب های حاوی گفتار مجاور و تشخیص غیرصحیح قاب های گفتار بین قاب های حاوی سکوت مجاور جلوگیری بعمل می آید. در این مرحله میتوان گذر سکوت به گفتار و گفتار به سکوت را با دقت بیشتری انجام داد.

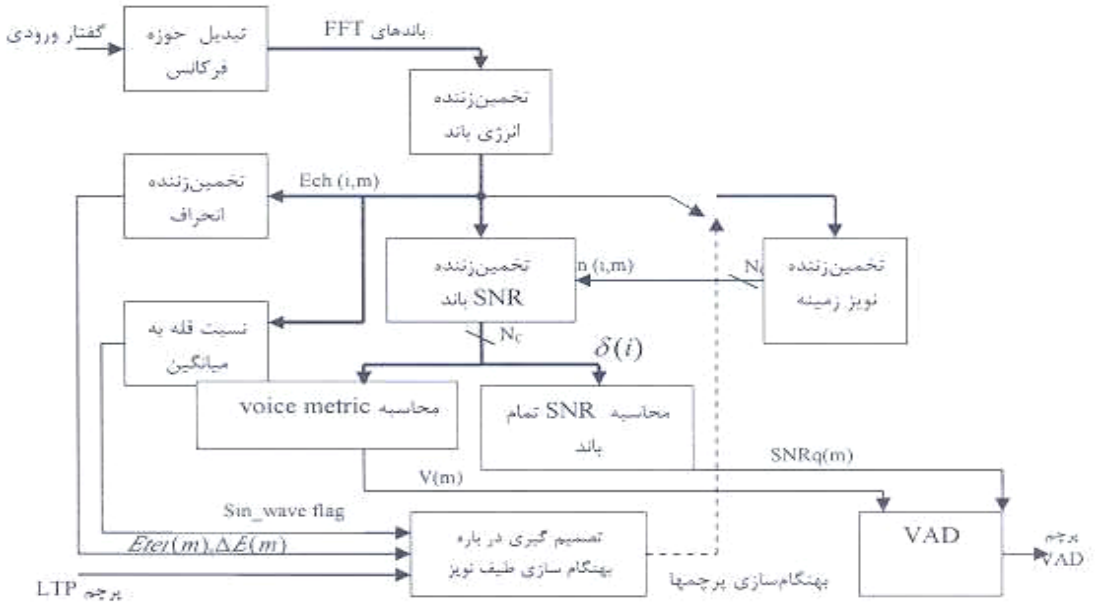
2-3-بلوک دیاگرام چند VAD استاندارد

سه VADاستانداردی که در کاربردهای کدگذاری و انتقال اطلاعات گفتار مورد استفاده قرار می گیرند عبارتند از: AMR[79], GSM[80] , G.729B. از G.729B در این پایان نامه استفاده شده است، که درفصل پنجم توضیح داده شده است. در زیر بلوک دیاگرامهای دو موردآخر آورده شده است.

2-3-1-استاندارد ETSI[81] AMR

موسسه استانداردسازی اروپا (ETSI)، دو نسخه از استاندارد AMR جهت کدینگ ارایه نموده است. اساس کار این دو نسخه بر اساس اطلاعات زیر باند می باشد و بلوک دیاگرام آنها در شکل (2-8) نشان داده شده است.

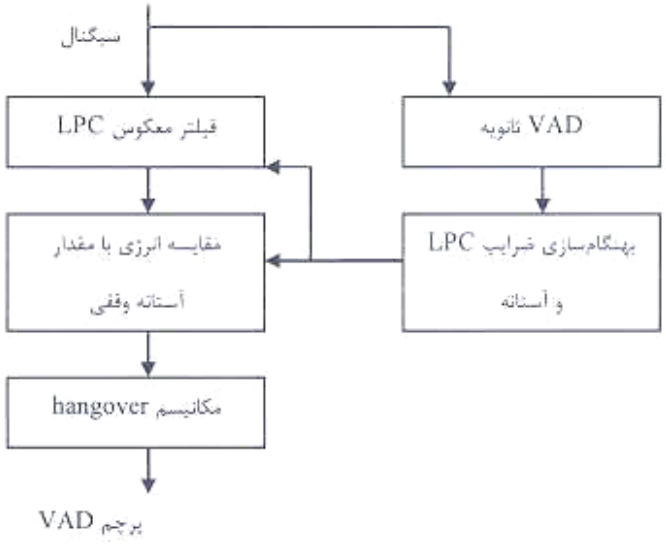
[79] .Adaptive Multi Rate
[80] .Global System for Mobile Communicat
[81] .European Telecommunication Standard Institute



شکل (2-8) : نمودار ساده ای از الگوریتم AMR2 [57]و[58]

2-3-2- الگوریتم GSM

در شکل (2-9) دیاگرام ساده ای از این VAD نشان داده شده است.[59] سیگنال ورودی در ابتدا توسط یک فیلتر معکوس که از ضرایب LPC نویز بدست آمده، فیلتر می شود. سپس انرژی سیگنال فیلتر شده، با مقدار آستانه ای مقایسه می شود. در انتها با کمک یک مکانیزم پیشامدگی داشتن تصمیم نهایی اتخاذ می شود. در صورتی که VAD ثانویه، رخداد سکوت را اعلام نماید، مقادیر ضرایب LPC مربوط به نویز، بهنگام می شوند. VAD ثانویه در صورت رخداد طیف متناوب و یا طیف غیرایستان وجود گفتار را اعلام می کند.



شکل (2-9) : نمودار الگوریتم GSM [59]

2-4-خلاصه

در سیستم های تشخیص گفتار، مرحله اول کار سیستم که یکی از مهمترین مراحل پردازش سیگنال گفتارنیز می باشد، تشخیص قسمت های گفتاری از دیگر قسمت ها می باشد، که با حذف قسمت های اضافی سرعت و دقت سیستم در مراحل بعدی افزایش می یابد. بنابراین در این فصل روش عملکرد سیستم توضیح داده شد. بردارهای ویژگی معرفی شدند. نحوه استخراج بردارهای ویژگی، که این بردارها بجای سیگنال اصلی به دلیل ثابت بودن و تغییر نکردن، در مراحل بعدی کار سیستم مورد استفاده قرار می گیرند، معرفی شدند. ضرایب MFCC و نحوه استخراج آنها که در سیستم پیشنهادی مورد استفاده اند بطور کامل شرح داده شده اند. انواع مدل های سیستم های تشخیص گفتار از غیر گفتار توضیح داده شدند.

فصل سوم:

آشکار سازی تغییر گوینده

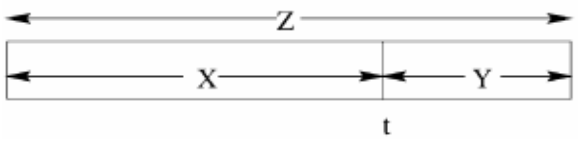
3-1-مقدمه

در یک فایل صوتی که شامل صحبت چندین گوینده می باشد، یکی از مهمترین موضوعات مشخص نمودن بازه زمانی بخش های گفتاری مربوط به هر گوینده می باشد. بنابراین باید مشخص نماییم که هر نفر در چه بازه ای صحبت نموده است. در سال های اخیر این مسئله موضوع پژوهش در بسیاری موارد بوده است. از جمله کاربردهای این بخش استفاده در سیستم های ردیابی گوینده[82] و افزایش دقت سیستم های بازشناسی گفتار و گوینده و فهرست نگاری اصوات ضبط شده و ... را میتوان نام برد. هدف یافتن نقاط تغییر گوینده در فایل صوتی می باشد. فایل صوتی به بخش های کوچکتری که در هر بخش(سگمنت) تنها گفتار یک گوینده

[82].Speaker Tracking

وجود دارد تقسیم می شود. این مرحله لازمه اصلی در این سیستم ها می باشد. بنابراین باید از الگوریتم های مناسبی استفاده نماییم تا بهترین نتایج بدست آیند. در سیستم های بخش بندی گفتار تکنیکی که اولین بار مورد استفاده قرار گرفت بوسیله چن و گوپالاکریشنان در سال 1998 بکار گرفته شد.[۱]

از اولین کارها در این زمینه میتوان به تحقیقات و سیستم های پیاده سازی شده که توسط گیش و دیگران[66] انجام شده است، اشاره نمود. در این روش پارامترهای سیگنال های گفتاری در ابتدا برحسب بردارهای ویژگی تعیین می شوند و سپس فاصله بین دو سگمنت همسایه بطور پی در پی برای آشکارسازی تغییر گوینده محاسبه می شود.[15] دو پنجره با طول نسبتا کم مانند شکل(3-1) در نظر گرفته می شوند و محتویات این دو پنجره بردارهای ویژگی استخراج شده از روی سیگنال صوتی هستند. این دو پنجره در طول سیگنال صوتی حرکت می کنند و شباهت محتویات آنها با هر قسمت از سیگنال با استفاده از یک تابع فاصله محاسبه می شود. مقایسه مقدار بهینه محلی این تابع فاصله، با مقدار آستانه تعیین می کند که آیا مرز این دو پنجره، t نقطه تغییر گوینده هست یا نه؟



شکل(3-1):پنجره های همسایه

الگوریتم های مختلف آشکارسازی گوینده در نوع تابع فاصله (ناهمانندی)[83] مورد استفاده، اندازه دو پنجره، افزایش زمانی شیفت دو پنجره، راه های آستانه گذاری و ارزیابی نتایج حاصل از مقادیر فاصله محلی فرق دارند. پارامترهای آستانه نیز معمولا تجربی محاسبه می شوند و نسبت به شرایط محیطی و آکوستیکی متفاوت مقاوم نیستند. الگوریتم های مختلفی در این بخش مورد استفاده قرار می گیرند.[80]

3-2-بخش بندی گوینده

کل روش های بخش بندی گوینده درسه گروه خلاصه می شوند که عبارتند از:

1)بخش بندی بر اساس فاصله(متریک)

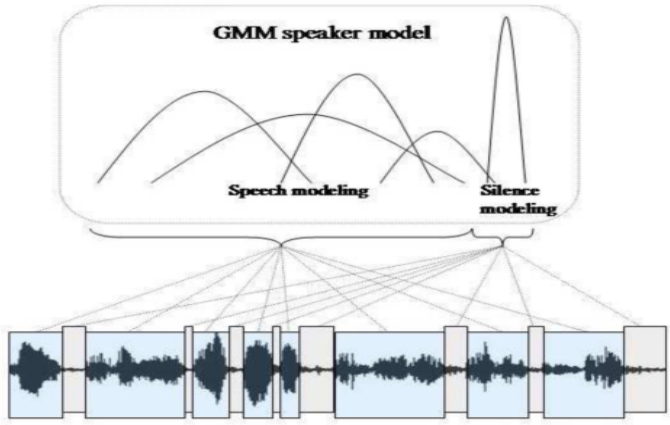
[1].Dissimilarity Function

2)بخش بندی بر اساس مدل

3)بخش بندی هیبرید

3-2-1-بخش بندی بر اساس فاصله

روش های متریک شاید یکی از تکنیک هایی است که تا بحال بیشترین استفاده را داشته اند. و بر اساس فاصله بین دو بخش تصمیم می گیریم که آیا دو بخش مختلف، مربوط به یک گوینده می باشد یا هر بخش به گوینده متفاوتی تعلق دارد. روش کار به این صورت است که برای دو بخش سیگنال بردارهای آکوستیکی X_i , X_j با تعداد نمونه های N_i , N_j و میانگین و واریانس های μ_i و σ_i و μ_j و σ_j که می توانند یک سیگنال گوسی یا یک مدل مخلوط گوسی باشند، در نظر گرفته می شود. به عبارت دیگر اگر دو بخش را بصورت پیوسته در نظربگیریم، میتوان بردار آکوستیکی X و میانگین و واریانس μ وσ مدل گوسی $M(\mu,\sigma)$ را برای آنها در نظر بگیریم. این روش تابحال بیشترین کاربرد را داشته است. بدلیل استفاده از مدل مخلوط گوسی در قسمت های مختلف، توضیح مختصری از این مدل ارائه شده است. در این مدل داده ها توسط منحنی های گوسی که در شکل (3-2) نشان داده شده اند، توصیف می شوند.



شکل(3-2) : ترکیب مدل های گوسین برای یک سیگنال شامل سکوت/گفتار[1]

برای یک ترکیب با K مولفه که از نمودارهای گوسی متفاوت تشکیل شده اند، داریم:

$$P(O) = \sum_{K=1}^{K} w_k p_k(O) = \sum_{K=1}^{K} w_k N(O;\mu_k,\Sigma_k) \qquad (27\text{-}3)$$

P مجموع k توزیع گوسی که هر توزیع دارای وزنی می‌باشد را مشخص می‌کند.

μ_k, Σ_k میانگین و کواریانس ماتریس می باشند. و مجموع وزن ها 1 می باشد. $(\sum W_i = 1)$ و داریم:

$$P_K(O) \sim N(O; \mu_K, \Sigma_K) \quad (28\text{-}3)$$

P_k توزیع نرمال گوسی o را نشان می دهد. اغلب مواقع کواریانس قطری گوسی مانند فرمول زیر مورد استفاده قرار می گیرد:[1]

$$\Sigma_k = \mathrm{diag}(\sigma_{k,1}^2, \ldots, \sigma_{k,N}^2) \quad (29\text{-}3)$$

$$\log P_K(O) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{d}\log \sigma_{k,t}^2 - \frac{1}{2}\sum_{t=1}^{d}\frac{(o_t - \mu_{k,t})^2}{\sigma_{k,t}^2} \quad (30\text{-}3)$$

d ابعاد O می باشد.

3-2-2-بخش بندی بر اساس مدل

این بخش بندی از دو مرحله تشکیل می شود. این مراحل عبارتند از:

1)مرحله آموزش: داده های آموزشی اولیه به سیستم انتقال می یابد و از آن مدل های اولیه که در نهایت توسط سیستم مورد استفاده قرار می گیرد، استخراج می شود.

2)مرحله آزمایش یا تشخیص گفتار : بازشناسی گفتار جدید با توجه به مدل های بدست آمده از مرحله قبل، هدف اصلی مرحله آزمایش یا بازشناسی است. هر سیستم با توجه به الگوریتم کاری خود می تواند گفتار جدید را با مدل های اولیه ذخیره شده در مرحله آموزش مقایسه و در نهایت نتیجه را اعلام کند.

3-2-3-بخش بندی هیبرید

این روش ترکیبی از دو تکنیک بر اساس مدل و بر اساس فاصله است. در این روش یک الگوریتم بخش بندی بر اساس فاصله، تنها برای ساخت یک مجموعه اولیه مدل های گویندگان بکار می رود، سپس با شروع از این مدل ها بخش بندی بر اساس مدل انجام می شود و با ترکیب خوشه بندی بر اساس فاصله و بر اساس مدل دقت خوشه بندی افزایش می یابد.[1]

3-3-مقایسه روش های بخش بندی

مزیت روش های بر اساس فاصله در این آن است که به هیچ اطلاعات قبلی نیاز ندارد ولی چون بخش بندی براساس فاصله بین سگمنت های مجزاست و سگمنت های خیلی کوتاه، نمی توانند به قدر کافی مشخصات یک گوینده را توصیف کنند، بنابراین سگمنت های خیلی کوتاه روی دقت این روش تاثیر نامطلوب دارند. بنابراین این روش مقاومت وپایداری زیادی ندارد. سیستم های بخش بندی بر اساس مدل به اطلاعات قبلی برای آماده سازی مدل های گویندگان نیاز دارند. که این نقطه ضعف این سیستم ها می باشد. سیستم های هیبرید چون از هر دو روش بالا استفاده می نماید، بطور قابل ملاحظه ای نتایج بهتری از دو روش دیگر دارند.[1]

3-4-روش های متداول آشکارسازی گوینده

متداول ترین معیارهایی که برای آشکارسازی تغییر گوینده بکار می روند عبارتند از:

- معیار اطلاعات بیزین
- نرخ درست نمایی عمومی[84]
- فاصله کالبک لیبلر[85]
- فاصله دیورژانس اشکال[86]
- BIC متقاطع[87]
-درستنمایی مدل مخلوط گوسی[88]

در ادامه هر یک از معیارها را توضیح می دهیم.[80]

3-4-1-معیار اطلاعات بیزین(BIC)

معیار بیزین بیشترین مورد استفاده در بخش بندی و کلاسه بندی را در بین روش های متریک به خود اختصاص داده است. در ضمن روش ساده و کارآمدی می باشد. معیار بیزین یک معیار مشخص نمودن پیچیدگی مدل مورد استفاده با توجه به تعداد پارامترهای آزاد بکار گرفته شده در مدل می باشد.[1] این

[84]. Generalize Likelihood Ratio (GLR)
[85]. Kullback-Leibler distance (KL or KL2)
[86]. Divergence Shape Distance(DSD)
[87].Cross-BIC (XBIC)
[88].GMM Likelihood Measure(GMM-L)

مدل در سال 1971 توسط اسچوارز و 1978 باز هم توسط اسچوارز بعنوان یک مدل معیار مورد استفاده قرار گرفت.[۶۹] و بطور گسترده ای در مقالات آماری بکار می رود. مسئله انتخاب مدل، انتخاب یک مدل از میان مجموعه مدل- های کاندید $M = M_1 , M_2 , ... , M_n$ است. که این مدل، مجموعه داده مشخص $D = D_1 , D_2 , ... , D_N$ را نمایش می دهد. مقدار BIC از رابطه (۳-۱) بدست می آید.

$$BIC(M_i) = \log P(D_1, D_2, ..., D_N | M_i) - \frac{\lambda}{2} d_i \log N \qquad (۱-۳)$$

$P(D_1, D_2, ..., D_N | M_i)$ درستنمایی ماکزیمم مدل M_i و d_i تعداد پارامترهای مستقل مدل است. و λ فاکتور جریمه است که برای جبران مواردی با مقدار N کوچک مورد استفاده قرار می گیرد. (چون BIC تمایل به انتخاب یک مدل ساده دارد، در زمانی که نمونه انتخابی کوچک می باشد، برای تنظیم λ کوچکتر، احتمال انتخاب مدل پیچیده تر افزایش می یابد. در تئوری مقدار λ باید ۱ قرار داده شود، ولی در عمل λ یک پارامتر قابل تنظیم است. در رابطه فوق، مقدار $\frac{\lambda}{2} d_i \log N$، از مقدار درستنمایی، بخاطر پیچیدگی کم می شود. بر اساس معیار BIC، برای مقادیر بقدر کافی بزرگ N، بهترین مدل برای نمایش داده، مدلی با مقدار BIC ماکزیمم است. امکان انتخاب مدل صحیح توسط BIC با بزرگتر شدن سایز نمونه ها ($N \to \infty$) به ۱ نزدیک می شود.

۳-۴-۱-۲-بخش بندی با استفاده از مدل آماری گوینده BIC

فرض کنید $X = x_i \in R^d$, $i = 1,2,...,N$ دنباله ای از بردارهای ویژگی کپسترال بر اساس فریم استخراج شده از روی جریان صوتی باشد. (معمولا از بردارهای ویژگی MFCC استفاده می شود. البته معیار BIC هیچ فرضی درباره ی روش استخراج ویژگی ندارد. بنابراین این روش قابل تعمیم به مواردی است که از روش های دیگر استخراج ویژگی استفاده می کنند.) که در آن حداکثر یک مرز سگمنت وجود دارد. مسئله تعیین اینکه آیا یک تغییر گوینده (مرز سگمنت) در فریم $b \in (1, N)$ وجود دارد یا نه، می تواند به یک مسئله انتخاب مدل تبدیل شود. دو مدل تعریف شده عبارتند از:

۱) مدل M_1 فرض می کند که همه نمونه های X مستقل هستند و بطور یکسان توسط یک فرآیند گوسین چند متغیره[89] توزیع می شود.

$$M_1 : \quad X = x_1, x_2, \dots, x_N \sim N(\mu, \Sigma)$$

۲) مدل M_2 فرض می کند که X توسط دو فرآیند گوسین چند متغیره ایجاد شده است.

$$M_1 : \quad X = x_1, x_2, \dots, x_N \sim N(\mu, \Sigma)$$

$$M_2 : \quad x_1, x_2, \dots, x_b \sim N(\mu_1, \Sigma_1)$$

$$x_{b+1}, x_{b+2}, \dots, x_N \sim N(\mu_2, \Sigma_2)$$

اگر $BIC(M_1) - BIC(M_2) > 0$ باشد، داده یکنواخت بوده و نقطه شکست (تغییر گوینده) نداریم. اگر این مقدار منفی باشد، نقطه شکست در این بازه وجود دارد.

میتوان نشان داد که با فرض یک توزیع نرمال $N(\mu, \sum)$، درستنمایی داده های $x_1, x_2, \dots, x_N$ وقتی ماکزیمم است که $\mu = \hat{\mu}$, $\Sigma = \hat{\Sigma}$ شود و داشته باشیم:

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (۲-۳)$$

$$\hat{\Sigma} = \frac{1}{N}\sum (x_i - \bar{x})(x_i - \bar{x})' \qquad (۳-۳)$$

طبق رابطه (۳-۱) مقادیر BIC این دو مدل بصورت زیر محاسبه می شود:

$$\widehat{BIC}(M_1) = -\frac{d}{2}N\log 2\pi - \frac{N}{2}\log|\hat{\Sigma}| - \frac{N}{2}\log|\hat{\Sigma}| - \frac{N}{2} - \frac{\lambda}{2}\big(d + d(d+1)\big)\log N \qquad (۴-۳)$$

$$\widehat{BIC}(M_2) = -\frac{d}{2}N\log 2\pi - \frac{b}{2}\log|\hat{\Sigma}_1| - \frac{N-b}{2}\log|\hat{\Sigma}_2| - \frac{N}{2}\lambda\big(d + d(d+1)\big)\log N \qquad (۵-۳)$$

$\hat{\Sigma}$و$\hat{\Sigma}_1$و$\hat{\Sigma}_2$ برآوردهای کواریانس با ماکزیمم درستنمایی از روی داده نظیر هستند. D بعد ویژگی کپسترال است. همچنین تفاضل BIC می تواند با استفاده از روابط بالا بعنوان تابعی از نقطه شکست b طبق رابطه (۳-۶) محاسبه شود.

[1].Multivariate Gaussian Process

$$\Delta BIC(b) = \overline{BIC(M_2)} - \overline{BIC(M_1)} =$$
$$\frac{1}{2}\left(N\log|\widehat{\Sigma}| - b\log|\widehat{\Sigma_1}| - (N-b)\log|\widehat{\Sigma_2}|\right) - \frac{\lambda}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N \qquad (۶-۳)$$

براساس معیار *BIC* بخش بندی جریان صوتی به دو بخش در فریم b وقتی صحیح است که $\Delta BIC(b) > 0$ باشد. مقدار مثبت بدین معناست که مدل M_2 سیگنال را بهتر توصیف می کند و نقطه شکست b وجود دارد. نقطه بخش بندی نهایی می تواند از طریق برآوردگر درستنمایی بیشینه[90] *(MLE)* بصورت زیر بدست آید:

$$\hat{b} = \arg\max \Delta BIC(b) \qquad 1 < b < N\ , \quad \Delta BIC(b) > 0 \qquad (۷-۳)$$

باید توجه داشت که BIC تنها برای بدست آوردن حداکثر یک نقطه تغییر آکوستیکی در داده های صوتی کاربرد دارد. تنظیم پارامترهای λ , N برای رسیدن به سیستمی مناسب بسیار اهمیت دارد. (در واقع خروجی سیستم قطعه بند صوتی بسیار حساس به تنظیم این دو پارامتر است.) بنابراین لازم است از الگوریتم هایی برای بدست آوردن نقاط شکست بیشتر استفاده نماییم. بنابراین برای صوتی که شامل چندین مرز بخش بندی است، یک الگوریتم آشکارسازی ترتیبی در [۸]پیشنهاد شده است. در [۸] یک پنجره متحرک کل جریان صوتی را جاروب می کند. پنجره از ابتدای جریان با عرض ۱ ثانیه آغاز می شود و داخل هر پنجره تست *BIC* به ترتیب برای هر نقطه $1 < b < N$ انجام می شود تا تعیین کند که آیا در این فاصله یک مرز سگمنت وجود دارد یا نه؟ اگر مرزی یافت نشود، پنجره 1 ثانیه به سمت جلو سیگنال جابجا می شود و اگر مرزی یافت شود، از محل این مرز پنجره جدیدی آغاز می شود.[۸۰]

۳-۴-۲-ترکیب آماره T^2 و BIC

بخش بندی بر اساس معیار *BIC* پیچیدگی زیادی دارد. اگرچه میتوان سرعت را با جستجو روی یک شبکه (در هر ۳۰ فریم) بهبود داد، ولی بار محاسباتی زیادی دارد. چون در یک پنجره برای هر نقطه شکست *b*، دترمینان- های دو ماتریس کواریانس کامل[91] تعیین می شوند و میانگین و کواریانس این توزیع ها باید

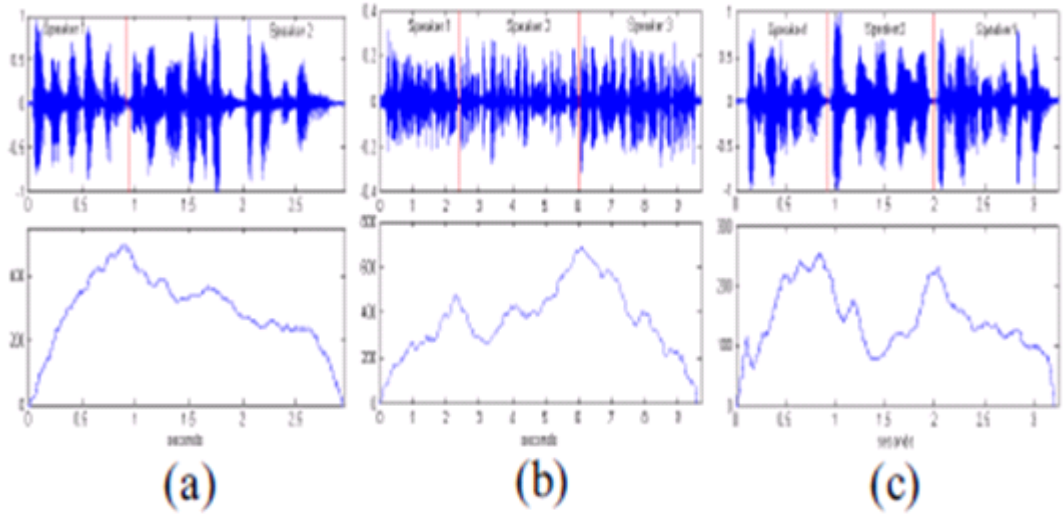
[1].Maximum Likelihood Estimator
[1]. Full Covariance Matrice

برآورده شوند، میزان خطا بالا خواهد بود. این مشکلات منجر به پیشنهاد روشی سریعتر و مفیدتر برای آشکارسازی مرز از طریق آماره T^2 توسط هانسن و زیو گردید.[۱۱]

آماره T^2 هتلینگ، یک آماره چند متغیره از توزیع معروف t است.[۷۰] یکی از کاربردهای آماره T^2 هتلینگ، اندازه گیری فاصله بین میانگین دو نمونه در زمانی است که ماتریس کواریانس آن دو نمونه یکسان ولی نامعلوم فرض می شود. در اینجا فرض می کنیم: دو نمونه در یک جریان گفتاری داریم. اولی شامل فریم های $[1,b]$ و دومی شامل فریم های $[b+1,N]$ است. آماره T^2 بصورت زیر تعریف می شود.

$$T^2 = \frac{b(N-b)}{N}(\mu_1 - \mu_2)\sum^{-1}(\mu_1 - \mu_2) \qquad (۸-۳)$$

μ_1 و μ_2 بترتیب میانگین های دو نمونه هستند و $\sum$ ماتریس کواریانس مشترک است. مقدار کوچکتر T^2 نشان دهنده شباهت بیشتر توزیع های دو نمونه است. بنابراین، هنگامی که T^2 به ماکزیمم خود می رسد، نقطه تغییر گوینده است که بین نمونه های کاملا متفاوت و نامشابه رخ می دهد. این مطلب در شکل (۳-۳) نشان داده شده است.



شکل(3-3): منحنی ها با اعمال متریک T^2-statistic رسم شده اند. شکل a مربوط به دو بخش متفاوت می باشد. شکل b سه بخش را نشان می دهد. شکل c سه بخش که به دو گوینده متعلقند را نمایش می دهد.[82]

الگوریتم جدید پیشنهادی [۶۸] تلفیقی از دو روش بالاست. بدین ترتیب که در داخل یک پنجره آشکارسازی، آماره T^2 در هر نقطه محاسبه می شود. نقطه با مقدار پیک بعنوان کاندید یک تغییر گوینده

انتخاب می شود. سپس معیار *BIC* تنها در این نقطه، در کنار پنجره آشکارسازی امتحان می شود. اگر تغییر گوینده توسط تست *BIC* تایید شود، یک پنجره آشکارسازی جدید از این نقطه شروع می شود که به جستجوی نقطه بعدی تغییرگوینده می پردازد. در غیر اینصورت، پنجره آشکارسازی برای بزرگتر کردن بازه جستجو، بزرگ می شود. این روش ترکیبی ساده مزایای فراوانی دارد که عبارتند از:

۱)با پیش انتخاب نقاط شکست ممکن از طریق آماره T^2، از محاسبه دو ماتریس کواریانس کامل در سایر نقاط جلوگیری می کنیم و بنابراین در مقایسه با *BIC* برای هر پنجره لغزان، تعداد ضرب ها $(N+2).d^2$ مرتبه کاهش می یابد. بنابراین این روش با استفاده از یک روش آشکارسازی ترتیبی موثرتر است و مزایای *BIC* مانند مستقل از آستانه بودن و یک پایه ریاضی ثابت را نیز هنوز دارد.

۲)هنگامی که اندازه نمونه کوچک است یا نقطه شکست نزدیک مرز پنجره است، روش BIC نتایج مطلوبی ندارد. چون وجود یک نقطه شکست درون یک پنجره با اندازه کوچک بدلیل ناکافی بودن داده ها باعث انحراف آماره مرتبه دوم می شود و تصمیم گیری در مورد نقاط شکست درست نخواهد بود. از طرفی ارزیابی آماره T^2 تنها به آماره های مرتبه اول نیاز دارد و نسبت به مواردی با اندازه نمونه کوچک، مقاوم است. در نتیجه پیش انتخاب نقاط شکست از طریق آماره T^2 از رخ دادن اشتباه در تعیین مرزها[92] در بخش بندی *BIC* جلوگیری می- نماید.[۸۰]

۳-۴-۲-۱-سرعت و بهره بیشتر در بخش بندی T^2-BIC

برای افزایش سرعت و بازدهی در این روش، افزایش اندازه پنجره متغیر و آزمایش پرش قاب نیز پیشنهاد شده است. میزان و خلوص داده موجود در هر پنجره، برای گرفتن تصمیمات آماری قابل اطمینان اهمیت فراوانی دارد. در الگوریتم بخش بندی ترتیبی، پهنای پنجره جاری اثر مهمی در پیش انتخاب نقاط شکست کاندید از طریق آماره T^2 و تصمیم گیری *BIC* بعدی دارد. اگر پنجره از نظر مدت زمان، خیلی پهن باشد و بیشتر از یک نقطه تغییر را شامل شود، فرض انتخاب مدل صحیح نیست و اگر پنجره خیلی کوتاه انتخاب شود، کمبود داده باعث می شود که برآورد گوسین ضعیف باشد و منجر به تصمیم نادرست در بخش بندی

[1]. Miss Locations

می شود. ضمنا این خطاها آماره های گوسین بعدی را نیز آلوده می کنند و روی آشکارسازی مرز سگمنت بعدی تاثیر می گذارند. بنابراین در این روش از یک پنجره پویا[93] استفاده می نماییم.[۸۰] یک پنجره با پهنای $W_0 = 200$ در ابتدا مورد استفاده قرار می گیرد. اگر هیچ نقطه شکستی در پنجره قبلی W_{i-1} نباشد، پهنای پنجره فعلی W_i بصورت زیر تنظیم می شود:

$$W_i = W_{i-1} + \Delta W_i \qquad (۹-۳)$$

$$\Delta W_i = \begin{cases} 100 & if\ W_{i-1} < 500 \\ \Delta W_{i-1} + 50 & else \end{cases} \qquad (۱۰-۳)$$

علاوه براین، پهنای پنجره فعلی W_i با موقعیت پیک (قله) آماره T^2 پنجره قبلی نیز کنترل می شود. اگر این پیک نزدیک مرز انتهایی پنجره در محدوده یک آستانه در پنجره قبلی ظاهر شود، $\Delta W_i = 50$ قرار می گیرد. با استفاده از این پنجره با افزایش قابل تنظیم، بهتر میتوان نقاط شکست سگمنت های کوچکتر را یافت و زمانی که هیچ نقطه شکستی در داده نباشد، جریان را با نرخ سریعتری جستجو نمود. دومین بهبود در کارآیی این الگوریتم از آزمایش پرش قاب حاصل می شود. نکته جالب توجه این است که نیازی نیست که همه قاب های داخل پنجره، بعنوان یک مرز در نظر گرفته شوند (بویژه زمانی که پنجره فعلی بزرگ باشد) برای مثال قاب های داده نزدیک به پنجره بویژه زمانی که پنجره فعلی بزرگ باشد، می توانند در تست T^2 در نظر گرفته نشوند، چون نمیتوان برآورد گوسین مقاومی را با این داده محدود بدست آورد. بعلاوه در پنجره های بزرگ (بزرگتر از ۱۰۰۰ قاب) احتمال اینکه یک شکست در بخش آغازین پنجره فعلی روی دهد، خیلی کم است. چون بعید به نظر می- رسد که شکستی از آزمایش بخش بندی قبلی در پنجره قبلی بجای مانده باشد. بنابراین در مورد این قاب ها آزمایش آماره T^2 را انجام نمی دهیم. بهسازی دیگری نیز می تواند با محاسبه پویای ماتریس کواریانس پنجره کامل $\sum$ بدست آید. که این ماتریس بوسیله آماره T^2 و تست BIC مورد استفاده قرار می گیرد. فرض کنید که در چند پریود زمانی، شکستی آشکار نشده باشد و در نتیجه طول پنجره همچنان زیاد می شود. میتوان کواریانس پنجره فعلی را با ترکیب آماره های داده پنجره قبلی و گسترش جدید محاسبه کرد. به این روش، از محاسبات تکراری کواریانس، روی داده های همپوش

[2]. Dynamic

بین پنجره های پی در پی جلوگیری می شود. اگر پهنای پنجره فعلی $W_i = W_{i1} + \Delta W_i$ باشد، ماتریس کواریانس پنجره فعلی بصورت زیر بدست می آید.

$$\Sigma_i(k,l) = \frac{1}{W_i}\{W_{i-1}(\Sigma_{i-1}(k,l) + \mu_{i-1}(k)\mu_{i-1}(l)) + \Delta W_i(\Sigma_{\Delta i}(k,l) + \mu_{\Delta i}(k)\mu_{\Delta i}(l))\} \quad (۳-$$

$$- \mu_i(k)\mu_i(l) \qquad 0 \le k,l \le d \quad (۱۱$$

μ_{i-1} و Σ_{i-1} و $\mu_{\Delta i}$ و $\Sigma_{\Delta i}$ به ترتیب ماتریس های میانگین و کواریانس پنجره قبلی و داده اضافه شده جدید هستند و μ_i پنجره فعلی کامل بصورت زیر بدست می آید.

$$\mu_i(k) = \frac{1}{W_i}\left(\mu_{i-1}(k).W_{i-1} + \mu_{\Delta i}(k).\Delta W_i\right) \quad (۱۲-۳)$$

۳-۴-۳-فاصله نرخ درستنمایی عمومی(GLR)

این روش اولین بار در سالهای 1976 توسط ویلسکی و جونز و در 1982 توسط برندت و اپل مورد استفاده قرار گرفت و روشی متریک مبتنی بر احتمال است. در آشکارسازی تغییر گوینده، دو زیر سگمنت همسایه را با استفاده از فاصله GLR نیز میتوان مقایسه نمود. اگر X_1 , X_2 به ترتیب مجموعه بردارهای ویژگی دو سگمنت همسایه باشند و $L(X_j , \lambda_j)$ و $L(X_i , \lambda_i)$ درستنمایی X_i , X_j باشند و λ_j, λ_i پارامترهای مدلی باشند که درستنمایی را ماکزیمم می کند و همینطور X ناشی از اتحاد دو سگمنت $X = X_i \cup X_j$ با هم و $L(X,\lambda)$ ماکزیمم درستنمایی برآورد شده برای X باشد، داریم:

$$GLR = \frac{L(X,\lambda)}{L(X_i,\lambda_i)L(X_j,\lambda_j)} \quad (۱۷-۳)$$

$$D(i,j) = -\log(GLR(i,j)) \quad (۱۸-۳)$$

با استفاده از اعمال آستانه به این فاصله میتوان نقاط تغییر نهفته را آشکار نمود. اگر دو سگمنت مورد بررسی i , j متعلق به یک گوینده باشند، مقدار فاصله GLR بیشتر از ۱ می شود و اگر دو سگمنت به یک گوینده متعلق نباشند، مقدار فاصله GLR به سمت صفر نزدیک می شود. معمولا مقدار GLR فاصله را

بطریق تجربی به گونه- ای مشخص می کنند که بتوان نقاط تغییر موجود در سیگنال صوتی را کاملا آشکار نمود. برای محاسبه این معیار معمولا مدل های گوسین مورد استفاده قرار می گیرند و λ شامل میانگین و واریانس مدل گوسین می- باشد که از روی داده های سگمنت بدست می آید.[۶۵]

3-4-4-فاصله KL2

روش فاصله ی متقارن KL2 مانند روش BIC از روش هایی است که هم برای قطعه بندی آکوستیکی کاربرد دارد و هم برای خوشه بندی مورد استفاده قرار می گیرد. روش KL در سال ۱۹۹۷ برای اولین بار و همچنین در سال ۲۰۰۰ توسط هانگ و ونگ و لی مورد استفاده قرار گرفت. و نتایج قابل قبولی برای دو توزیع رندم X,Y بدست می دهد. فاصله KL (اغلب دیورژانس نامیده می شود) از فرمول (3-19) محاسبه می شود:

$$KL(X;Y) = E_X(\log\frac{P_X}{P_Y}) \qquad (19\text{-}3)$$

وقتی دو بخش مورد بررسی دارای توزیع گوسی باشند میتوان میزان فاصله KL را از رابطه (3-20) محاسبه نمود:

$$KL(X,Y) = \frac{1}{2}tr[(\Sigma_X - \Sigma_Y)(\Sigma_Y^{-1} - \Sigma_X^{-1})] + \frac{1}{2}tr[(\Sigma_Y^{-1} - \Sigma_X^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T] \qquad (20\text{-}3)$$

معیار KL2 خود از معیار فاصله KL مشتق می شود. معیار فاصله ی KL در پیدا کردن فاصله ی دو متغیر تصادفی کاربرد دارد. به این ترتیب که برای بدست آوردن فاصله ی دو متغیر تصادفی A,B کافی است B را با کدبندی بهینه برای A، کد کنیم. در این صورت نرخ بیت اضافی (کمتری) که برای B بدست می آید، همان فاصله ی دو متغیر خواهد بود. هر چه این فاصله ی KL بزرگتر باشد، به معنی این است که توابع توزیع چگالی دو متغیر تصادفی از هم فاصله ی بیشتری دارند. بنابراین می بینیم که روابط بالا خاصیت تقارنی ندارند و نمیتوان آن را به عنوان معیار فاصله در نظرگرفت. با کمی بازبینی در این رابطه میتوان رابطه متقارن KL2 را مطرح نمود. برای توزیع های گوسین چند متغیره، فرمول بالا را میتوان به صورت رابطه زیر تغییر داده و نام فاصله KL2 برای آن انتخاب گردید.

$$KL2(X;Y) = KL(X;Y) + KL(Y;X) \qquad (21\text{-}3)$$

اگر دو زیر سگمنت همسایه توسط توزیع های گوسین چند متغیره[94] $N(\mu_i, \Sigma_i), N(\mu_j, \Sigma_j)$ مدل شوند، فاصله KL2 بین دو زیر سگمنت همسایه بصورت زیر تعریف و محاسبه می شود:[68]

$$KL2_{i,j} = \frac{1}{2}(\mu_i - \mu_j)'(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + \frac{1}{2} tr(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I) \qquad (22\text{-}3)$$

این معیار فاصله، مثبت و متقارن[95] است. هنگامی که فاصله KL2 در مرز دو سگمنت به مقدار ماکزیمم محلی خود می رسد، این محل، نقطه تغییر گوینده و شروع بخش جدید را مشخص می کند. این معیار فاصله بطور معمول بصورت پیوسته بین پنجره های همسایه در طول جریان صوتی محاسبه و منحنی فواصل محاسبه شده رسم می شود. برای جلوگیری از ایجاد نوسانات[96] در این منحنی، فواصل بدست آمده را با استفاده از یک فیلتر پایین گذر، ملایم سازی می نمایند. در نهایت، پیک های محلی منحنی بعنوان نقاط کاندید تغییر گوینده در نظر گرفته می شوند. تعیین نهایی نقاط تغییر گوینده از روی نقاط کاندید شده، کار مشکلی می باشد، چون به آستانه ای برای مقایسه نیاز دارد که این آستانه معمولا از روی داده های آموزشی تعیین می شود و به همین دلیل نمی تواند برای همه داده های آزمایشی مقاوم و پایدار باشد.

3-4-5-آشکارسازی تغییر گوینده با استفاده از DSD

این روش نیز مانند روش های قبلی با استفاده از معیار فاصله کار میکند و روشی بسیار شبیه به روش گیش است و در سال 1991 توسط گیش مطرح گردید. این روش به راحتی تحت تاثیر شرایط محیطی قرار می گیرد. در این روش برای آشکارسازی تغییرات گوینده، دو زیر سگمنت همسایه از بردارهای ویژگی MFCC، روی جریان گفتاری حرکت داده می شوند. شباهت بین محتویات این دو زیر سگمنت با استفاده از یک تابع فاصله دیورژانس محاسبه می شود.[15] ناهمانندی D بین دو زیر سگمنت همسایه بصورت فاصله تعیین شده توسط کواریانس دو زیر سگمنت با رابطه زیر تعریف می شود:

$$D = \frac{1}{2} tr\left[(\Sigma_i - \Sigma_j)(\Sigma_j^1 - \Sigma_i^1)\right] \qquad (23\text{-}3)$$

1. Multivariate Gussian Distribution

2.Symmetric

1.Fluctuation

Σ_i ،Σ_j ماتریس های کواریانس برآورد شده از زیر سگمنت های i,j هستند. زمانی یک تغییر گوینده بالقوه[97] بین زیر سگمنت های i و i+1 وجود دارد که شرایط زیر برآورده شود:

$$D(i,i+1) > D(i+1,i+2)$$
$$D(i,i+1) > D(i-1,i) \qquad (24\text{-}3)$$
$$D(i,i+1) > T_i$$

D(i , j) فاصله بین زیر سگمنت i , j می باشد و T_i یک مقدار آستانه می باشد. دو شرط اول وجود پیک محلی را تضمین می کنند و شرط سوم از آشکارسازی پیک های کوچک جلوگیری می کند. نتایج حاصل از معیار DSD نتایج قابل قبولی می باشند. در معیار DSD، تنظیم آستانه کار سختی می باشد و اگر آستانه خیلی کوچک باشد، آشکارسازی نادرستی انجام می گیرد. همچنین مقدار آستانه تحت تاثیر شرایط محیطی مختلف نیز قرار می گیرد، بعنوان مثال اگر گفتار در یک محیط نویزی باشد، فاصله بین زیر سگمنت های گفتاری افزایش می یابد و بنابراین برای یک محیط نویزی مقدار آستانه باید افزایش یابد. به کمک رابطه زیر میتوان آستانه مناسب را با استفاده از N فاصله پیاپی بدست آورد:[67]

$$T_i = a \frac{1}{N} \sum_{i=0}^{N} D(i-n-1, i-n) \qquad (25\text{-}3)$$

N تعداد فواصل قبلی بکار رفته برای پیش بینی آستانه و a یک ضریب تقویت کننده با مقادیر $1 \leq a \leq 1.5$ است. نتیجه گیری در این سیستم بدین صورت است که اگر فاصله بین دو سگمنت گفتاری بزرگتر از یک آستانه باشد، این دو سگمنت گفتاری متعلق به دو گوینده مختلف می باشند.[16]

3-4-6- BIC متقاطع

این معیار فاصله در سال 2004 توسط هرناندو و آنگرا و در سال 2005 توسط آنگرا مورد استفاده قرار گرفت. این معیار فاصله با استفاده از BIC، میزان فاصله میان دو بخش مجاور را با استفاده از معیار درستنمایی[98] اندازه- گیری می نماید، رابطه زیر چگونگی محاسبه فاصله را نمایش می دهد.[۱]

2.Potential Speaker Change

[98].Cross-likelihood

$$XBIC(x_1 ; x_2) = L(x_1 , M_2(\mu_2, \sigma_2)) + L(x_2, M_1(\mu_1, \sigma_1)) \quad (۲۶-۳)$$

3-4-7-درستنمایی مدل مخلوط گوسی (GMM-L)

برای اندازه گیری درستنمایی مدل مخلوط گوسی بصورت زیر عمل می نماییم: اگر دو زیر سگمنت همسایه A وB توسط توزیع های گوسین چند متغیره $N(\mu_A , \Sigma_A)$ و $N(\mu_B , \Sigma_B)$ مدل شوند ، فاصله GMM-L بین این دو زیر سگمنت بصورت زیر محاسبه می شود.[17]

$$(GMM - L) = \frac{1}{N_B} \log L\left(B , N(\mu_A , \Sigma_A)\right) \quad (31-3)$$

اگر مقدار فاصله دو زیر سگمنت کم باشد، احتمال اینکه نقطه تغییر گوینده بین A و B باشد، بیشتر است. [1]

3-5-خلاصه

سیگنال گفتاری شامل گفتار چند گوینده، باید به بخش هایی که فقط شامل گفتار یک گوینده است، تقسیم شود. برای این مهم، سیگنال بخش بندی می شود. در این فصل، سه روش بخش بندی بر اساس فاصله، بخش بندی بر اساس مدل، بخش بندی ترکیبی جهت انجام تقسیم بندی سیگنال گفتار توضیح داده شد. بخش بندی بر اساس فاصله از پرکاربردترین این روش ها می باشد. سپس مزیت ها و معایب روش ها با هم مقایسه شد. پرکاربردترین روش بخش بندی، بخش بندی بر اساس فاصله است که انواع آن معرفی شد. معیار فاصله بیزین (BIC) که در این پایان نامه استفاده شده است به همراه سایر روش ها بطور کامل شرح داده شد.

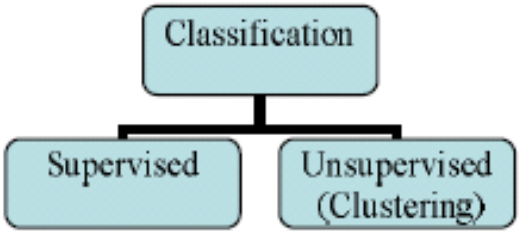
فصل چهارم:

روش های خوشه بندی

4-1-مقدمه

خوشه بندی، سازمان دهی مجموعه ای از الگوها[99] بر اساس شباهت در خوشه ها است. به نحوی که الگوهای داخل یک خوشه شبیه به هم بوده و دارای بیشترین تفاوت با الگوهای خوشه های دیگر باشند. بطور کلی، فرآیند خوشه بندی بصورت یک دسته بندی بدون سرپرست، تعریف می شود که هیچ اطلاع قبلی در مورد کلاس ها و یا تعداد آنها موجود نیست. از طرفی باید به تفاوت میان دسته بندی و خوشه بندی توجه داشت. شکل(4-1) در دسته بندی با سرپرست، مجموعه ای از الگوهای برچسب دار (دسته بندی شده) موجود است و مسئله، برچسب زدن به داده جدید است. این الگوهای برچسب دار، کلاس ها را توصیف می کنند که این کلاس ها برای برچسب زنی الگوی جدید بکار می روند. ولی در خوشه بندی که نوعی دسته بندی بدون سرپرست است، هیچ اطلاع قبلی راجع به کلاس ها یا تعداد آنها موجود نیست و مسئله دسته بندی مجموعه ای از الگوهای بدون برچسب در خوشه های ممکن است و به هر خوشه نیز یک برچسب اختصاص می یابد. ولی این برچسب ها از خود داده ها ناشی می شوند.

1.Pattern



شکل (4-1): انواع خوشه بندی

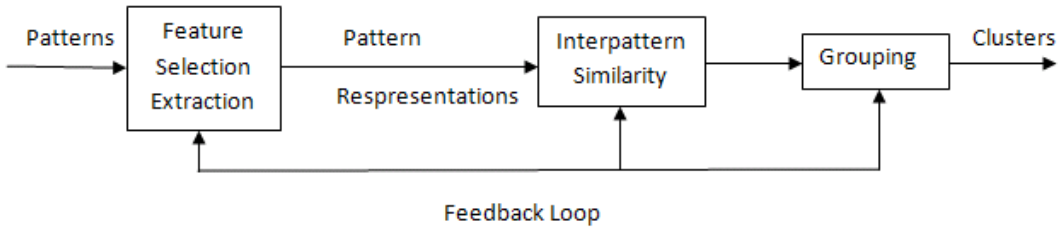
در این پایان نامه، هدف از خوشه بندی گوینده، شناسایی و دسته بندی سگمنت های گفتاری مربوط به یک گوینده و اختصاص یک برچسب واحد به آنهاست و در نهایت نتیجه فرآیند خوشه بندی، یک خوشه در ازای هر گوینده خواهد بود.

4-2-اجزا سیستم خوشه بندی

اجزا این سیستم عبارتند از:

1) نمایش الگو: این بخش اختیاری است و می تواند شامل استخراج ویژگی ها و یا انتخاب ویژگی باشد.
2) تعیین معیار نزدیکی الگو متناسب با حوزه داده.
3) گروه بندی یا خوشه بندی[76]

که یک نمونه ترتیب قرار گرفتن این مراحل در شکل (4-2) همراه با مسیر فیدبک نمایش داده شده است. با وجود این مسیر فیدبک، خروجی مرحله دسته بندی می تواند مراحل استخراج ویژگی و محاسبه شباهت را تحت تاثیر قرار دهد.



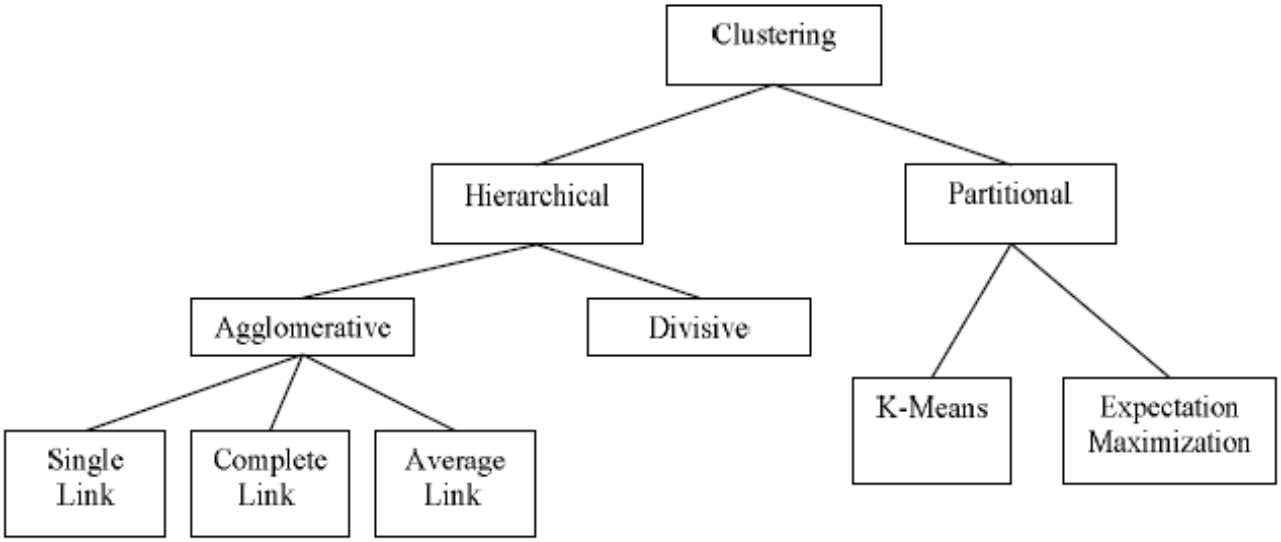
شکل(4-2): مراحل خوشه بندی

واحد نمایش الگو به تعداد کلاس ها و تعداد الگوهای موجود و تعداد و نوع و مقیاس[100] ویژگی های موجود برای الگوریتم خوشه بندی اشاره می کند. در این گام، انتخاب ویژگی، فرآیند تشخیص کارآمدترین زیر مجموعه از ویژگی های اولیه برای استفاده در خوشه بندی است و منظور از استخراج ویژگی نیز استفاده از یک یا چند تابع تبدیل برای تبدیل ویژگی های ورودی به ویژگی های مناسب است. یک یا هر دوی این تکنیک ها را میتوان برای رسیدن به مجموعه مناسب ویژگی ها بکار برد تا این مجموعه در خوشه بندی استفاده شود. نزدیکی دو الگو نیز معمولا توسط یک تابع فاصله بین دو الگو اندازه گیری می شود. برای اندازه گیری فاصله نیز معیارهای متفاوتی پیشنهاد شده است، از جمله این معیارها به نرخ لگاریتم درستنمایی[101]، ΔBIC و GLR و KL2 اشاره نمود، که در قسمتهای قبلی مفصلا توضیح داده شدند. از چندین روش برای انجام مرحله خوشه بندی استفاده می شود که در ادامه توضیح داده می شوند.

4-3-روش های خوشه بندی

از جمله کاربردهای خوشه بندی به بازیابی اطلاعات، پردازش تصاویر، شناسایی الگو، سنجش از راه دور، داده کاوی و ... میتوان اشاره نمود. برای پیاده سازی الگوریتم های مختلفی بکار گرفته شده اند. این الگوریتم ها را در دو دسته سلسله مراتبی و افرازی [102] تقسیم بندی می نمایند. شکل (4-3). روش های سلسله مراتبی شامل دو دسته کلی تجمعی و تقسیمی می باشند. از روش های خوشه بندی افرازی میتوان الگوریتم K-Means و EM[103] را نام برد.

[1].Scale
[2]. Log-Likelihood ratio
1.Partitional
[2]. Expectation Maximization



شکل (4-3): روش های خوشه بندی

4-3-1-روش های خوشه بندی سلسله مراتبی[104]

این روش خوشه بندی به دو صورت زیر انجام می شود:

- روش های خوشه بندی بالارونده
- روش های خوشه بندی پایین رونده

در سیستم هایی که بصورت برون خط کار می کنند و در جاهایی که بخش های گفتار مشخص هستند، عمل ادغام[105] تا جایی ادامه می یابد تا به تعداد بهینه از گویندگان دست یابیم. روش های خوشه بندی سلسله مراتبی بیشتر بر اساس تئوری گراف استوار است. ساختار این روش را بصورت گرافیکی میتوان با نمودار درختی[106] نمایش داد. روش غالب خوشه بندی مورد استفاده در سیستم ها، خوشه بندی تجمعی سلسله مراتبی[107] است که از یک معیار توقف برای پایان خوشه بندی استفاده می کند. مراحل به شرح زیر است:

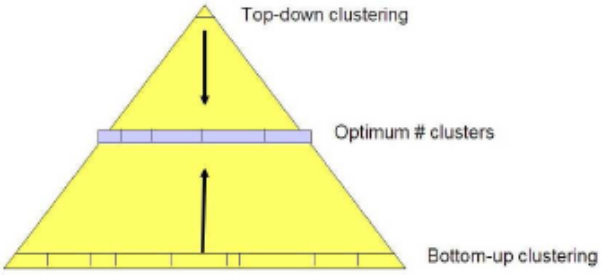
1) در ابتدا هر شی را به عنوان یک خوشه در نظر می گیریم.(گره های درخت خوشه بندی)
2) محاسبه فواصل بین هر دو خوشه

[104].Hierarchical Clustering Techniques
[105].Merge
[106].Dendrogram
4.Hiererchical Agglomerative Clustering

3) ادغام نزدیکترین زوج دو خوشه

4) به هنگام کردن فواصل خوشه های باقیمانده با خوشه جدید

5) تکرار مراحل 2 تا 4 تا زمان ارضای شرط توقف



شکل(4-4):روش های خوشه بندی بالا و پایین رونده[1]

4-3-1-1-روش های خوشه بندی بالارونده[108] :با توجه به مطالب فوق، در این روش هر داده را بطور مستقل به عنوان یک خوشه در نظر می گیریم و با تعریف معیار شباهت، دو یا چند خوشه، خود تشکیل یک خوشه بزرگتر می دهند و این روند ادامه می یابد تا شرط پایانی الگوریتم که معمولا تعداد خوشه ها است، تحقق پذیرد. بعبارت دیگر با تعداد زیادی از بخش ها کار خود را شروع می نماید و با استفاده از تکنیک های ادغام، به یک اندازه بهینه از کلاس ها دست می یابد.[1] این تکنیک سالهای زیادی در دسته بندی آماری برای دسته- بندی مورد استفاده قرار گرفته است. معمولا یک ماتریس فاصله بین همه خوشه ها تعریف می شود و هر دو خوشه نزدیک به هم ادغام می شوند تا کار خاتمه یابد. برای توقف می توانیم از فرمول زیر استفاده نماییم:

$$W_{jin} = \left| \sum_{k=1}^{K} N_k \Sigma_k \right| \sqrt{k} \qquad (1\text{-}4)$$

K : تعداد خوشه های موجود است. N_K تعداد بخش های آکوستیکی و Σ_K کواریانس ماتریس خوشه k.

[108].Bottom-up Clustering Techniques

در سال 2006 توسط راجی فاصله میان دو مدل مبتنی بر مدل مخلوط گوسی از طریق محاسبه به وسیله فاصله KL مطرح شد. دو مدل M_1 , M_2 هر کدام با K_1 , K_2 ترکیب گوسی و وزن های گوسی W_{jin} مفروضند.

فاصله دو مدل M_1 , M_2 از فرمول زیر محاسبه می شود:

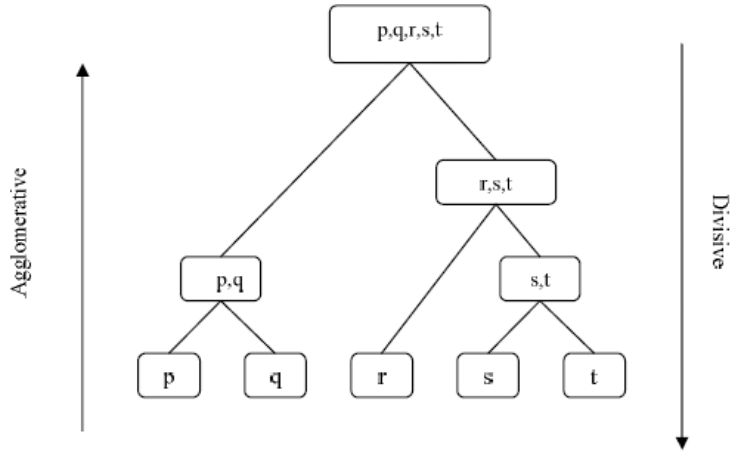
$$d(M_1,M_2) = \sum_{i=1}^{K_1} W_1(i) min_{j=1}^{K_2} KL(N_1(i),N_2(j)) \quad (2\text{-}4)$$

حال اگر از KL2 برای محاسبه فاصله استفاده نماییم داریم:[1]

$$D(M_1,M_2) = \sqrt{\sum_{m=1}^{M}\sum_{d=1}^{D} W_m \frac{(\mu_1(m,d)-\mu_2(m,d))^2}{\sigma_{m,d}^2}} \quad (3\text{-}4)$$

4-3-1-2-روش های خوشه بندی پایین رونده[109]: روش سلسله مراتبی تقسیمی برخلاف روش تجمعی، همه اشیا را یک خوشه در نظر می گیرد و بطور تکراری هر خوشه را به خوشه های کوچکتر تا زمان ارضای شرط پایانی، تقسیم می کند. بعبارت دیگر معمولا با تعداد کلاس های کمی شروع به کار می کند و با تقسیم به بخش های کوچکتر تا رسیدن به مقدار بهینه تقسیمات را ادامه می دهد.

این الگوریتم در ابتدا با تقسیم شدن به قسمت های کوچکتر، سرانجام به تعداد بهینه خوشه تقسیم و در نهایت متوقف می شود. شکل (4-5) این روش را بخوبی نشان می دهد:



[109].Top-down Clustering Techniques

شکل (4-5): مثال ساده ای از خوشه بندی سلسله مراتبی[80]

در الگوریتم های خوشه بندی سلسله مراتبی، میزان مشابهت بین دو خوشه که هر کدام از آنها فقط حاوی یک کلاس است، برابر با میزان مشابهت بین دو کلاس خواهد بود. در غیر این صورت مشابهت بین دو خوشه به سه روش زیر محاسبه می شود:[71]

1) **پیوند تکی**[110]: مشابهت بین دو خوشه، برابر با میزان مشابهت زوجی از اعضای دو خوشه خواهد بود که نسبت به مشابهت سایر زوج های ممکن بین دو خوشه، کمترین مقدار را داشته باشد.

$$sim(P, Q) = min\, sim(p, q) \qquad p \in P\,, q \in Q \tag{4-4}$$

2) **پیوند کامل**[111]: مشابهت بین دو خوشه برابر با میزان مشابهت زوجی از اعضای دو خوشه خواهد بود که نسبت به مشابهت سایر زوج های ممکن بین دو خوشه، بیشترین مقدار را داشته باشد.

$$sim(P, Q) = \max sim(p, q) \qquad p \in P\,, q \in Q \tag{5-4}$$

3) **پیوند میانگین گروهی**[112]: مشابهت بین دو خوشه برابر با میانگین میزان مشابهت همه زوج های ممکن بین اعضای دو خوشه خواهد بود.

$$sim(P, Q) = \frac{1}{(|P||Q|)} \sum_{p \in P\,, q \in Q} sim(p, q) \tag{6-4}$$

4-3-2-روش های خوشه بندی افرازی

روش های افرازی، داده ها را بر اساس معیار تشابه به تعداد خاص خوشه ها تقسیم بندی می کنند. در این روش تمام تکنیک ها بر این فرض استوار است که هر داده تنها متعلق به یک خوشه می باشد. روش های افرازی در مواردی که مجموعه داده بزرگی در اختیار است و تشکیل نمودار درختی از نظر محاسباتی مقرون به صرفه نیست، از نظر کاربرد مزیت دارند. معروفترین الگوریتم در این گروه الگوریتم K-means می باشد که داده ها را به k خوشه مستقل تقسیم بندی می کند. از دیگر الگوریتم های این دسته، میتوان به

2. Single Link
1. Complete Link
2. Group Average Link

الگوریتم های خوشه- بندی فازی[113] و روش های مبتنی بر شبکه های عصبی (ANN) ، مثل SOM[114] را نام برد.[76]

4-4- روش های خوشه بندی متداول در سیستم های خوشه بندی گوینده

در این سیستم ها هدف از خوشه بندی گوینده، شناسایی و دسته بندی سگمنت های گفتاری مربوط به یک گوینده و اختصاص یک برچسب واحد به آنهاست و در نهایت نتیجه فرآیند خوشه بندی، یک خوشه در ازای هر گوینده خواهد بود. در سیستم های بخش بندی و خوشه بندی گوینده، روش خوشه بندی تجمعی سلسله مراتبی بیشتر از سایر روش ها دارای کاربرد می باشد که از یک معیار توقف بر اساس BIC استفاده می کند. در خوشه بندی بر اساس BIC عموما فاصله بین خوشه ها با برآورد نرخ درست نمایی عمومی مشخص می شود و بررسی می شود که آیا این زوج خوشه با دو توزیع گوسین مجزا و یا یک توزیع گوسین بهتر توصیف می شوند. اگر خوشه ها ادغام شوند، داده هر دو خوشه برای برآورد یک توزیع گوسین ترکیب می شود. مراحل ادغام وقتی پایان می یابد که حداقل فاصله، از یک آستانه مشخص(صفر) بیشتر شود. در مواردی نیز که تعداد گویندگان از قبل مشخص بوده است، از روش K-means برای خوشه بندی سگمنت ها استفاده شده است.[72] سیستم های خوشه بندی موجود اساسا از نظر تابع فاصله، نحوه ادغام خوشه ها و معیار توقف متفاوت هستند. همچنین از روش های خوشه بندی دیگری نظیر روش تقسیمی[115][73] و تلفیق بخش بندی و خوشه بندی [10][74] نیز با موفقیت استفاده شده است. برای اندازه گیری فاصله بین خوشه ها نیز معیارهای متفاوتی پیشنهاد شده است. از جمله این معیارها میتوان به نرخ لگاریتم درستنمایی ΔBIC، GLR، KL2 اشاره کرد که قبلا مفصلا توضیح داده شده اند. باید توجه داشت که انتخاب روش مناسب خوشه بندی دارای اهمیت زیادی می باشد و علاوه بر آن معیار توقف مورد استفاده نیز نقش حیاتی در تعیین میزان کارآیی دارد و تنظیم این معیار به نحوه استفاده از نتیجه خوشه بندی بستگی دارد. چون اگر خوشه بندی کمتر از حد[116] لازم صورت گرفته باشد، داده های گفتاری یک گوینده در چند خوشه

3.Fuzzy

4.Self-organizing map

1.Divisive

2.Under-clustering

پراکنده می شود و اگر خوشه بندی بیش از حد[117] باشد، خوشه های نهایی خالص نخواهند بود و شامل گفتار چند گوینده خواهند بود. که هر دو مورد برای انجام عمل فهرست نگاری مطلوب نیستند، اما ممکن است در کاربردهای دیگر مورد استفاده قرار گیرند. بعنوان مثال هرگاه مجموع گفتار چند گوینده مشابه را می خواهیم، خوشه بندی بیش از حد می تواند مفید باشد و زمانی که نتیجه خوشه بندی برای تطبیق گوینده در آماده سازی مدل های بازشناسی گفتار بکار می رود و گوینده در چند محیط آکوستیکی متفاوت سخن گفته باشد، خوشه بندی کمتر میتواند مفید باشد. معمولا برای معیار توقف از تغییر درستنمایی کلی بعد از خوشه بندی استفاده می کنند. متداول ترین معیار توقف برای این منظور BIC می باشد. [8] [75] در این پایان نامه برای خوشه بندی از ماشین بردار پشتیبان (SVM) که امروزه یکی از پرکاربردترین دسته بندی کننده ها می باشد، استفاده گردیده است. در ادامه به توضیح مراحل کاری آن می پردازیم.

4-5- دسته‌بندی کننده ماشین های بردار پشتیبان

ماشين‌هاي بردار پشتيبان يك تكنيك دسته‌بندي و رگرسيون است كه توسط وپنيک[118] و گروهش در آزمايشگاه AT&T Bell پيشنهاد شده است و در حال حاضر در بسياري از زمينه‌ها مثل تشخيص چهره، تشخيص صوت، بازشناسي ديجيتالي هويت با استفاده از دست‌خط وغيره استفاده مي‌شود. این دسته‌بندی کننده یک دسته‌بندی کننده خطی است که می‌توان با اعمال برخی تغییرات از آن به عنوان دسته‌بندی کننده غیر خطی نیز بهره‌جست. توانایی این دسته‌بندی کننده در یافتن فوق صفحه بهینه برای دسته‌بندی می‌باشد. در ادامه به بررسی این دسته‌بندی کننده و توانایی آن در دسته بندی داده‌های خطی در دو حالت جداپذیر کامل و جدا ناپذیر کامل می‌پردازیم و سپس نحوه استفاده از این دسته‌بندی کننده را برای داده‌های غیرخطی شرح می‌دهیم.

4-5-1- دسته‌بندی کننده ماشین بردار پشتیبان خطی

4-5-1-1- دسته‌بندی کلاس‌های جداپذیر

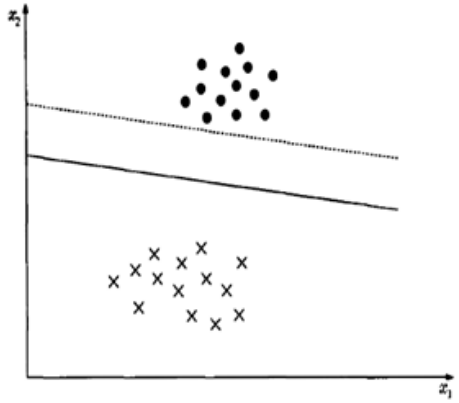
در این بخش هدف طراحی یک دسته‌بندی کننده خطی به منظور جداسازی کلاس‌های جداپذیر از هم می‌باشد. در ابتدا یک دسته‌بندی کننده خطی به منظور جداسازی داده‌های دو کلاس جداپذیر از هم

3.Over-clustering

[1].Vapnik

معرفی می‌گردد و سپس این دسته‌بندی کننده را به منظور جدا سازی داده‌های کلی و جداناپذیر گسترش می‌دهیم.[28]

فرض کنید $x_i \ , \ i = 1,2,\ldots,N$ بردارهای ویژگی برای داده‌های آموزشی[119] X باشد. این داده‌ها به دو کلاس w_1 و w_2، که به صورت خطی از هم جداپذیرند، تعلق دارند. هدف بدست آوردن فوق صفحه $g(x) = \mathbf{w}^{\mathrm{T}} x + w_0 = 0$ به نحوی است که تمامی داده‌های آموزشی را صحیح دسته‌بندی کند. به صورت کلی این فوق صفحه یکتا نبوده و می‌توان مقادیر مختلفی را برای $\mathbf{w}$ و w_0 بدست آورد. شکل (6-4) دو نمونه از فوق صفحه‌هایی را که می‌توان به منظور دسته‌بندی صحیح نقاط داده شده در نظرگرفت را نشان می‌دهد.



شکل(4-6): یک نمونه از مسئله دو کلاسه خطی جداپذیر که نمونه‌ها توسط دو دسته‌بندی کننده خطی جدا شده.

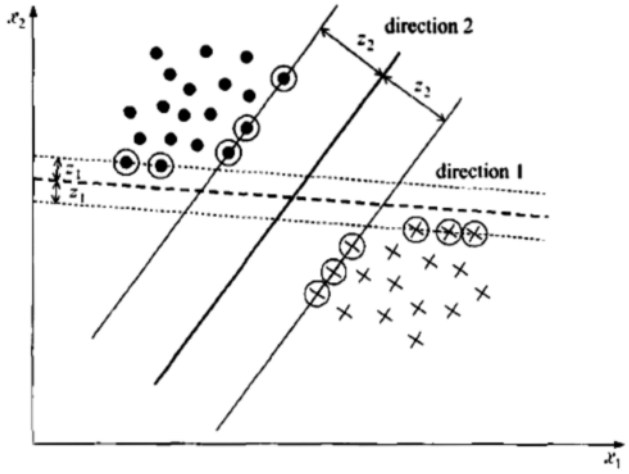
هر دو فوق صفحه نشان داده شده در شکل (6-4) عمل جداسازی را به درستی انجام می‌دهند، اما واضح است که خط متصل نسبت به خط منقطع دسته‌بندی کننده مناسب‌تری است. زیرا این فوق صفحه فضای بیشتری را در دو طرف خود ایجاد می‌کند. این بدان معنی است که این دسته‌بندی کننده در زمان دسته‌بندی نمونه‌های آزمایشی نتایج بهتری را از خود نشان می‌دهد. این نتیجه امری بسیار مهم در طراحی دسته‌بندی کننده‌هاست و آنرا عمومیت عملکرد[120] دسته‌بندی کننده می‌نامند. این مسئله به توانایی دسته‌بندی کننده در دسته‌بندی رضایت‌بخش داده‌های خارجی برمی‌گردد.

بنابراین در دسته‌بندی کننده ماشین‌های بردار پشتیبان قصد بر این است که علاوه بر دسته‌بندی صحیح داده‌های آموزشی، فوق صفحه‌ای را بدست آوریم که حاشیه[121] بین دو کلاس را بیشینه کند. منظور از حاشیه بین دو کلاس فاصله‌ای است که فوق‌صفحه بین دو کلاس باقی می‌گذارد.

---

[2].Training Set
[1] .Generalization Performance
[2].Margin

هر فوق صفحه بوسیله جهت و مکان دقیقش در فضا مشخص می‌گردد که جهت فوق‌صفحه را $\mathbf{w}$ و مکان دقیق آنرا در فضا W_0 مشخص می‌کند. به دلیل اینکه هیچ برتری بین دو کلاس وجود ندارد، لازم است فاصله فوق صفحه از نزدیک‌ترین نقاط هر دو کلاس w_1 و w_2 به یک اندازه باشد. بنابراین لازم است فوق صفحه‌ای را بیابیم که بیشترین حاشیه ممکن را ایجاد کند و به این منظور به دنبال جهت بهینه و مکان دقیق فوق‌صفحه هستیم. [28] شکل (4-7) این امر را به وضوح روشن می‌سازد.



شکل(4-7): حاشیه برای جهت 2 بیشتر از حاشیه در جهت 1 است.

فوق‌صفحه‌ای که در شکل (4-7) با خط ضخیم‌تر نشان داده شده است فوق‌صفحه مورد نظر جهت جداسازی دو کلاس می‌باشد. زیرا علاوه بر جداسازی دو کلاس، حاشیه بین فوق‌صفحه و دو کلاس را هم بیشینه کرده است.

فاصله یک نقطه از یک فوق‌صفحه برابر است با:

$$z = \frac{|g(x)|}{\|\mathbf{w}\|} \qquad (4\text{-}7)$$

$\mathbf{w}$ و w_0 را طوری مقیاس‌دهی می‌کنیم که اندازه $g(x)$ در نزدیک‌ترین نقاط w_1 برابر 1 و در نزدیک‌ترین نقاط w_2 برابر 1- شود. حاشیه بین دو کلاس برابر خواهد بود با:

$$\frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \qquad (4\text{-}8)$$

به این ترتیب خواهیم داشت:

$$\begin{aligned}&\mathbf{w}^T x + w_0 \geq 1 \qquad \forall \mathrm{x} \in \mathrm{w}_1\\&\mathbf{w}^T x + w_0 \leq -1 \qquad \forall x \in w_2\end{aligned} \tag{9-4}$$

حال به جایی رسیده‌ایم که مسئله را می‌توانیم به صورت ریاضی حل کنیم. برای هر x_i یک برچسب کلاس به صورت y_i ، $y_i = 1$ اگر $x_i \in w_1$ و $y_i = -1$ اگر $x_i \in w_2$، در نظر می‌گیریم. برای بدست آوردن فوق‌صفحه دلخواه می‌بایستی $\mathbf{w}$ و w_0 بر اساس شرایط زیر بدست آوریم:

$$\begin{aligned}&\text{minimize} \quad \mathrm{J}(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|^2\\&\text{subject to} \quad \mathrm{y_i}\left(\mathbf{w}^T x_i + w_0\right) \geq 1, \quad i = 1,2,\ldots,N\end{aligned} \tag{10-4}$$

واضح است که کمینه کردن مقدار نرم[122] باعث می‌گردد که حاشیه فوق‌صفحه با کلاس‌ها بیشینه گردد. این شرایط حل یک معادله غیرخطی براساس یک سری محدودیت‌های خطی می‌باشد.

شرایط Karush-Kuhn-Tucher (KKT) به منظور کمینه کردن معادله (4-8) برابر خواهد بود با:

$$\frac{\partial}{\partial \mathbf{w}} L\left(\mathbf{w}, w_0, \lambda\right) = \mathbf{0} \tag{11-4}$$

$$\frac{\partial}{\partial w_0} L\left(\mathbf{w}, w_0, \lambda\right) = 0 \tag{12-4}$$

$$\lambda_i \geq 0, \quad i = 1,2,\ldots,N \tag{13-4}$$

$$\lambda_i \left[y_i\left(\mathbf{w}^T x_i + w_0\right) - 1\right] = 0, \quad i = 1,2,\ldots,N \tag{14-4}$$

در معادلات بالا λ بردار ضرایب لاگرانژ و $L\left(\mathbf{w}, w_0, \lambda\right)$ تابع لاگرانژ است که به صورت زیر تعریف می‌گردد:

$$L\left(\mathbf{w}, w_0, \lambda\right) = \frac{1}{2}\mathbf{w}^T \mathbf{w} - \sum_{i=1}^{N} \lambda_i \left[y_i\left(\mathbf{w}^T x_i + w_0\right) - 1\right] \tag{15-4}$$

با ترکیب رابطه (4-15) با (4-11) و (4-12) خواهیم داشت:

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{16-4}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{17-4}$$

1. Norm

باید توجه داشت که ضرایب لاگرانژ، λ_iها، می‌توانند صفر و یا مثبت باشند. بنابراین براساس معادله(4-16) پاسخ بدست آمده برای بردار پارامترهای $\mathbf{w}$ برابر است با ترکیب خطی از $N_s \le N$ بردار ویژگی که ضریب لاگرانژ آنها مخالف صفر است.

$$\mathbf{w} = \sum_{i=1}^{N_s} \lambda_i y_i x_i \qquad \mathrm{N_s} \le N \tag{18-4}$$

این بردارها که ضرایب لاگرانژ غیرصفر دارند و برای بدست آوردن فوق‌صفحه بهینه استفاده می‌گردند را بردارهای پشتیبان[123] می‌نامند. بر اساس دسته محدودیت‌های موجود در رابطه (4-14) برای ضرایب لاگرانژ غیرصفر، بردارهای پشتیبان در دو فوق‌صفحه زیر قرار می‌گیرند:

$$\mathbf{w}^T x + w_0 = \pm 1 \tag{19-4}$$

بردارهای پشتیبان بردارهایی هستند که نزدیک‌ترین فاصله را از دسته‌بندی‌کننده خطی را دارا می‌باشند. باند جداسازی فاصله بین دو فوق‌صفحه معرفی شده توسط معادله (4-19) می‌باشد. از بین بردارهای ویژگی، بردارهایی که ضریب لاگرانژ مربوط به آنها صفر می‌باشد، $\lambda_i = 0$، می‌توانند داخل و یا خارج باند جداسازی کلاس‌ها و یا حتی بر روی خود دو فوق صفحه محدود کننده باند قرار گیرند.

به منظور حل معادلات مربوطه و بدست آوردن فوق صفحه بهینه، که اثبات می‌شود که این فوق صفحه یکتا می‌باشد، روش‌های گوناگونی مورد استفاده قرار می‌گیرد. این معادلات با استفاده از خاصیت مرسوم به دوگانی لاگرانژ قابل حل هستند و نمایش هم‌ارز دوگان Wolfe آنها برابر خواهد بود با:

$$\text{maximize} \quad L(\mathbf{w}, w_0, \lambda) \tag{20-4}$$

$$\text{Subject to} \quad \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{21-4}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \qquad \lambda_i \ge 0 \tag{22-4}$$

محدودیت‌های بدست آمده در رابطه (4-22) با صفر قرار دادن گرادیان لاگرانژ نسبت به $\mathbf{w}$ و w_0 بدست آمده‌اند. با قرار دادن روابط (4-22) و (4-21) در رابطه (4-20) و انجام محاسبات ریاضی خواهیم داشت:

---

[1]. Support Vectors

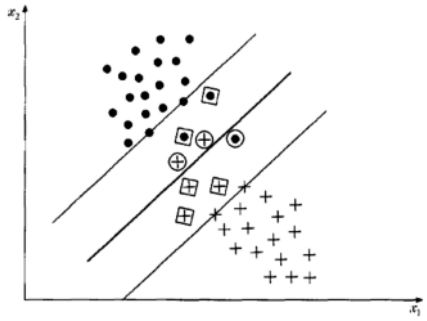
$$\max_{\lambda}\left(\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\lambda_i\lambda_j y_i y_j x_i^T x_j\right) \tag{23-4}$$

$$\text{Subject to} \quad \sum_{i=1}^{N}\lambda_i y_i = 0 \tag{24-4}$$

$$\lambda \geq 0$$

4-5-1-2- دسته‌بندی کلاس‌های جدا ناپذیر

در این حالت کلاس‌های موجود با استفاده از دسته‌بندی کننده خطی به صورت کامل از همدیگر جدا نمی‌شوند. همان‌گونه که در شکل (4-8) مشخص می‌باشد دو کلاس در این حالت جداپذیر نمی‌باشند و نمی‌توان با یک دسته‌بندی کننده خطی آن‌ها را کاملا از هم جدا کرد. هر گونه تلاشی به منظور جداسازی داده‌ها با یک فوق‌صفحه به نتیجه نمی‌رسد مگر آنکه یک سری از داده‌ها در داخل باند جداسازی قرار گیرد. در این شرایط حاشیه فاصله بین دو فوق‌صفحه با معادله $\mathbf{w}^T x + w_0 = \pm 1$ تعریف می‌گردد.



شکل(4-8): نمونه‌ای از داده‌هایی که به صورت خطی به طور کامل از همدیگر جدا نمی‌شوند.

بردار ویژگی نمونه‌های آموزشی به یکی از سه دسته زیر تعلق دارند.[28]

بردارهایی که خارج از باند قرار می‌گیرند و صحیح دسته‌بندی شده‌اند. این بردارها شرایط در نظر گرفته شده در معادله (4-10) را برقرار می‌سازند.

بردارهایی که در داخل باند قرار می‌گیرند و صحیح دسته‌بندی شده‌اند. این بردارها در شکل (4-8) با مربع مشخص گردیده‌اند و در شرایط زیر صدق می‌کنند.

$$0 \leq y_i\left(\mathbf{w}^T x + w_0\right) < 1 \tag{25-4}$$

بردارهایی که به اشتباه دسته‌بندی شده‌اند. این بردارها در شکل (4-8) با دایره مشخص گردیده‌اند و از شریط زیر تبعیت می‌کنند.

$$y_i\left(\mathbf{w}^T x_{\ +}w_0\right)<0 \tag{26-4}$$

تمامی سه حالت فوق را می‌توان با اضافه کردن یک محدودیت جدید و معرفی یک دسته پارامتر جدید در نظر گرفت:

$$y_i\left[\mathbf{w}^T x_{\ +}w_0\right]\geq 1-\xi_i \tag{27-4}$$

گروه اول داده‌ها با در نظر گرفتن $\xi_i = 0$، گروه دوم با در نظر گرفتن $0 < \xi_t \leq 1$ و گروه سوم با در نظر گرفتن $\xi_t > 1$ بدست می‌آیند. پارامتر ξ_i به نام پارامتر Slack شناخته می‌شود. در شرایط کنونی هدف این است که تا حدممکن حاشیه را با در نظر گرفتن تعداد نقاطی است که در آنها $\xi_t > 0$ است، افزایش دهیم. در فرم ریاضی این عمل برابر است با مینیمم کردن تابع هزینه زیر:

$$J\left(\mathbf{w},w_0,\xi\right)=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{N}l\left(\xi_i\right) \tag{28-4}$$

که در آن ξ بردار پارامترهای ξ_i است و

$$l\left(\xi_t\right)=\begin{cases}1 & \xi_i>0\\ 0 & \xi_i=0\end{cases} \tag{29-4}$$

پارامتر C یک عدد ثابت مثبت است که تاثیر نسبی دو قسمت رابطه (28-4) را کنترل می‌کند. بهینه-سازی رابطه (28-4) مشکل است زیرا شامل تابع ناپیوسته $l(\cdot)$ است. بنابراین شرایط به صورت زیر تغییر می‌کند:

$$\text{minimize}\quad J\left(\mathbf{w},w_0,\xi_i\right)=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{N}\xi_i \tag{30-4}$$

$$\text{Subject to}\quad y_i\left[\mathbf{w}^T x_i+w_0\right]\geq 1-\xi_i,\quad i=1,2,\ldots,N \tag{31-4}$$

$$\xi_i\geq 0,\qquad i=1,2,\ldots,N \tag{32-4}$$

محدودیت لاگرانژ برای حل معادلات فوق برابر است با:

$$\begin{aligned}L\left(\mathbf{w},w_0,\xi,\lambda,\mu\right)=&\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\mu_i\xi_i\\ &-\sum\lambda_i\left[y_i\left(\mathbf{w}^T x_i+w_0\right)-1+\xi_i\right]\end{aligned} \tag{33-4}$$

شرایط Karush-Kuhn-Tucker متناظر برابر است با:

$$\frac{\partial L}{\partial \mathbf{w}}=0\ \text{ or }\ \mathbf{w}=\sum_{i=1}^{N}\lambda_i y_i x_i \tag{34-4}$$

$$\frac{\partial L}{\partial w_0} = 0 \text{ or } \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{35-4}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \text{ or } C - \mu_i - \lambda_i = 0, \quad i = 1, 2, \ldots, N \tag{36-4}$$

$$\lambda_i \left[y_i \left(\mathbf{w}^T x_i + w_0 \right) - 1 + \xi_i \right] = 0, \quad i = 1, 2, \ldots, N \tag{37-4}$$

$$\mu_i \xi_i = 0, \quad i = 1, 2, \ldots, N \tag{38-4}$$

$$\mu_i \geq 0, \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, N \tag{39-4}$$

نمایش دوگانی Wolf معادلات فوق برابر خواهد بود با:

$$\begin{aligned} &\text{maximize} \quad L\left(\mathbf{w}, w_0, \lambda, \xi, \mu\right) \\ &\text{Subject to} \quad \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i x_i \\ &\qquad \sum_{i=1}^{N} \lambda_i y_i = 0 \\ &\qquad \mathrm{C} - \mu_i - \lambda_i = 0, \quad i = 1, 2, \ldots, N \\ &\qquad \lambda_i \geq 0, \ \mu_i \geq 0, \quad i = 1, 2, \ldots, N \end{aligned} \tag{40-4}$$

با اعمال محدودیت فوق به معادلات لاگرانژ خواهیم داشت:

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j \right) \tag{41-4}$$

$$\text{Subject to} \quad 0 \leq \lambda_i \leq C, \quad i = 1, 2, \ldots, N \tag{42-4}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{43-4}$$

تنها تفاوت معادلات فوق با معادلات کلاس‌های خطی جداپذیر از هم در محدودیت اعمال شده بر روی ضرایب لاگرانژ است که با مقدار C محدود گشته‌اند. حالت کلاس‌های خطی جداپذیر زمانی که $C \to \infty$ برقرار می‌گردد. همان‌گونه که ملاحظه می‌گردد پارامتر Slack، ξ_i و ضرایب لاگرانژ متناظرشان و μ_i در معادلات وارد نگشته‌اند. حضور این پارامترهای به صورت غیرمستقیم در پارامتر C منعکس گشته است.

4-5-1-3- دسته‌بندی داده‌های چند کلاسه با ماشین‌های بردار پشتیبان

در تمامی قسمت‌های گذشته، به بررسی عمل دسته‌بندی برای دو کلاس مختلف پرداختیم. اگر بخواهیم دسته‌بندی کننده ماشین‌های بردار پشتیبان را برای M کلاس گسترش دهیم کافی است روش ارائه شده در قسمت‌های پیشین را برای M مسئله دو کلاس اعمال کنیم. به این ترتیب مسئله دسته بندی M

کلاس به M دسته‌بندی دو کلاسه تبدیل خواهد شد. برای هر یک از کلاس‌ها، ما به دنبال بدست آوردن تابع تفکیک بهینه $g_i(x),\ i=1,2,\ldots,M$ هستیم. نمونه x متعلق به کلاس w_i است اگر

$$g_i(x) \geq g_j(x), \quad \forall j \neq i \tag{44-4}$$

4-5-2- ماشین‌های بردار پشتیبان غیر خطی

در بخش‌های پیشین ماشین‌های بردار پشتیبان را به صورت یک دسته‌بندی کننده خطی بهینه معرفی کردیم. حال فرض کنید یک نگاشت به صورت

$$x \in \Re^l \rightarrow y \in \Re^k \tag{45-4}$$

از فضای ویژگی ورودی به یک فضای k بعدی داریم که در این فضای جدید کلاس‌ها با یک فوق-صفحه از یکدیگر قابل جداسازی هستند. حال در این فضای k بعدی جدید یک دسته‌بندی کننده ماشین-های بردار پشتیبان طراحی می‌کنیم.[28] در زیر قضیه‌ای را به نام ***Mercer*** معرفی می‌کنیم.

قضیه Mercer:

فرض کنید $x \in \Re^l$ و نگاشت ϕ به صورت زیر در نظر گرفته شده باشد:

$$x \rightarrow \phi(x) \in H \tag{46-4}$$

H یک فضای اقلیدسی[124] است. در این صورت عملگر ضرب داخلی به صورت زیر قابل نمایش است:

$$\sum_r \phi_r(x)\phi_r(z) = k(x,z) \tag{47-4}$$

$\phi_r(x)$ جزء r ام از نگاشت $\phi(x)$ بوده و $k(x,z)$ یک تابع متقارن است که از شرایط پیروی می-کند:

$$\int k(x,z)g(x)g(z)dx\ \mathrm{dz} \geq 0 \tag{48-4}$$

به ازای هر $g(x),\ x \in \Re^1$ که داشته باشیم

$$\int g(x)^2 dx < +\infty \tag{49-4}$$

توابعی که شرایط فوق را داشته باشد کرنل می‌نامیم. به هر حال قضیه ***Mercer*** نحوه بدست آوردن فضای H را شامل نمی‌شود. بنابراین اگر ضرب داخلی مرتبط با فضای مورد نظر را داشته باشیم ابزاری کلی برای بدست آوردن نگاشت $\phi(.)$ را در اختیار نداریم. بعلاوه ما ابزاری برای شناخت

---

[1].Euclidean

ابعاد فضا نیز در اختیار نداریم. به صورت کلی کرنل‌هایی که به منظور نگاشت در شناسایی آماری الگو استفاده می‌گردد عبارتند از:

- **کرنل چند جمله‌ای:**

- $k\left(x,z\right)=\left(x^{T}z+1\right)^{q}, \quad q>0$ (50-4)

- **کرنل تابع پایه شعاعی[125] :**

- $k\left(x,z\right)=\exp\left(-\frac{\left\|x-z\right\|^{2}}{\sigma^{2}}\right)$ (51-4)

- **کرنل تانژانت هیپربولیک :**

- $k\left(x,z\right)=\tanh\left(\beta x^{T}z+\gamma\right)$ (52-4)

از این کرنل‌ها برای نگاشت به فضای بالاتر در دسته‌بندی کننده‌های ماشین‌های بردار پشتیبان استفاده می‌گردد. زمانی که یک کرنل مناسب به منظور نگاشت به فضایی با ابعاد بالاتر استفاده گردید با استفاده از بهینه ساز دوگان Wolfe خواهیم داشت:

$$\max_{\lambda}\left(\sum_{i}\lambda_{i}-\frac{1}{2}\sum_{i,j}\lambda_{i}\lambda_{j}y_{i}y_{j}k\left(x_{i},x_{j}\right)\right) \quad (53\text{-}4)$$

$$\text{Subject to} \quad 0\le\lambda_{i}\le C, \quad i=1,2,\ldots,N \quad (54\text{-}4)$$

$$\sum_{i}\lambda_{i}y_{i}=0 \quad (55\text{-}4)$$

و درنتیجه نمونه x متعلق به کلاس w_1 خواهد بود اگر

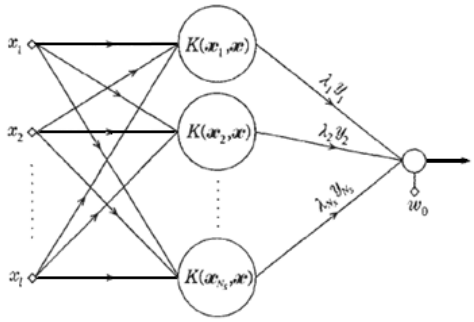
$$g\left(x\right)=\sum_{i=1}^{N}\lambda_{i}y_{i}k\left(x_{i},x\right)+w_{0}>0 \quad (56\text{-}4)$$

و بصورت مشابه نمونه x متعلق به کلاس w_2 خواهد بو د اگر

$$g\left(x\right)=\sum_{i=1}^{N}\lambda_{i}y_{i}k\left(x_{i},x\right)+w_{0}<0 \quad (57\text{-}4)$$

شکل (4-9) نحوه نگاشت به فضای بالاتر را در SVM غیرخطی نشان می‌دهد. تعداد گره‌های موجود در لایه میانی را تعداد بردارهای پشتیبان در دسته‌بندی کننده SVM تعیین می‌کند.[28]

[1].Radial Basis Function(RBF)



شکل(4-9): نمایش ماشین بردار پشتیبان غیر خطی.

4-6-خلاصه

مرحله نهایی کار سیستم های تشخیص گفتار، خوشه بندی خروجی های حاصل از مرحله بخش بندی سیگنال گفتاری می باشد. باید سگمنت های همگن متعلق به یک گوینده در یک خوشه قرار گیرند. در این فصل، انواع روشهای خوشه بندی معرفی گردید. دو تکنیک خوشه بندی بالارونده و پایین رونده که از پرکاربردترین روش ها می باشند، معرفی گردید. الگوریتم های مربوط به این روش ها توضیح داده شد. و روش مورد استفاده در این پایان نامه، جهت خوشه بندی، که استفاده از "ماشین بردار پشتبیبان (SVM)" است، بطور کامل شرح داده شد.

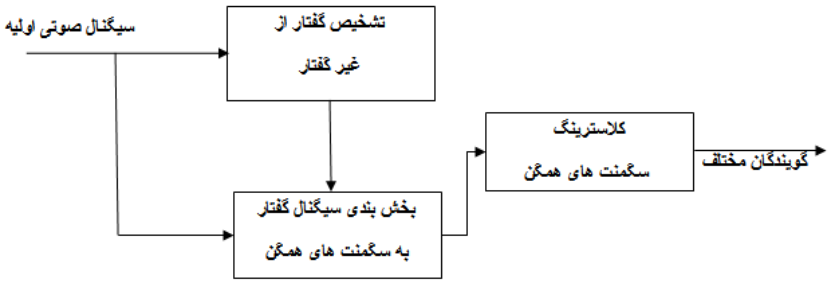
فصل پنجم:

پیاده سازی سیستم

پیشنهادی و نتایج حاصله

5-1-مقدمه

سیستم تشخیص گفتار، سیگنال صوتی را به سگمنت های مجزا و همگن تقسیم می نماید و سپس عملیات خوشه بندی سگمنت های همگن انجام می شود. بطوریکه هر خوشه نشان دهنده اطلاعات یک گوینده به تنهایی می باشد. در این پایان نامه، یک الگوریتم بخش بندی بر اساس فاصله(BIC)، برای بخش بندی گفتار به سگمنت های مجزا بکار گرفته شده است. سپس خوشه بندی نیز بر اساس فاصله با استفاده از ماشین بردارهای پشتیبان انجام گرفته است. بلوک دیاگرام مراحل کاری مختلف در شکل (5-1) نشان داده شده است. در ادامه به توضیح کامل سیستم خواهیم پرداخت.



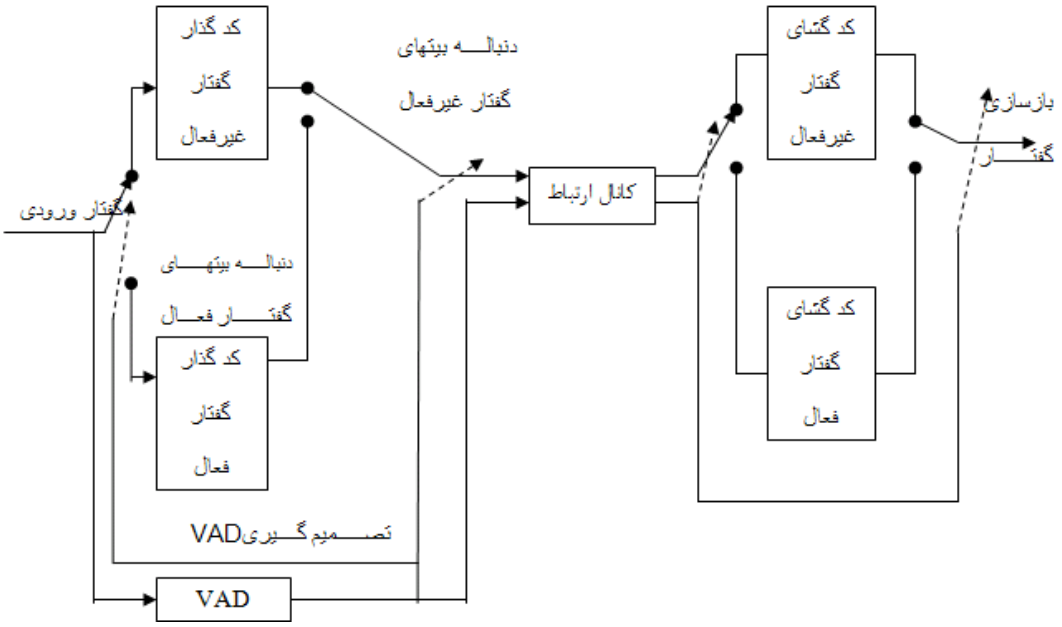
شکل(5-1): نمودار سیستم پیاده سازی شده

ابتدا بخش های سکوت موجود در سیگنال صوتی حذف می شوند. (حذف سکوت) این کار به علت بالاتر رفتن سرعت سیستم انجام می شود و همچنین میزان خطای مراحل بعد را کاهش می دهد. خروجی این مرحله فقط شامل گفتار گویندگان موجود می باشد. (البته برای تخمین میزان تاثیر این مرحله بر روی مراحل بعدی و کل سیستم، آزمایشاتی بدون اعمال این قسمت بر روی دادگان انجام شد، نتایج حاصله دارای خطای بیشتری نسبت به زمانی می باشند، که این مرحله نیز در سیستم پیاده سازی می شود. در نتیجه در ادامه کار، نتایج با اجرای این بخش مورد بررسی قرار گرفتند.) بخش گفتاری باقیمانده به دنباله ای از بردارهای ویژگی تبدیل می شود. و سپس بخش آشکارسازی تغییرات گوینده، خروجی حاصل از بخش قبلی را بعنوان ورودی دریافت می دارد و به آشکارسازی تغییرات می پردازد. (مرزهایی را برای ما مشخص می نماید که در آن مرزها، گوینده به گوینده ای جدید تبدیل شده است) و گفتار را به بخش هایی که در

هر بخش تنها یک گوینده صحبت می نماید، تقسیم می نماید. بر روی بخش ها عمل خوشه بندی انجام می شود و بخش های گفتاری مربوط به هر گوینده در یک خوشه جای می گیرد.

5-2-ساختار سیستم پیاده سازی شده

آشکارسازی سکوت از روش متکی بر انرژی، نرخ عبور از صفر، و ضرایب LSF[126] استفاده می نماید. بدین منظور از الگوریتم G.729B استفاده می نماید. الگوریتم G.729B از الگوریتم های رایج مورد استفاده در رمزنگاری می- باشد. که توسط موسسه ITU-T استاندارد شده است.[60] این VADعملیات رمزنگاری را در دو مد، که در شکل (5-2) نشان داده شده به اجرا می رساند. از رمزگزار G.729 جهت رمزی نمودن داده ها هنگامی که در سیگنال نواحی فعال تشخیص داده شود، استفاده می گردد. در غیر اینصورت از یک رمزگزار نواحی غیرفعال مانند SID[127]جهت انتقال گفتار استفاده می شود.این عمل را انتقال غیرپیوسته[128] می نامند. شکل (5-2) دیاگرام این VAD را نمایش می دهد.



شکل (5-2) : انتقال اطلاعات گفتار با استفاده از یک VAD [60]

در مرحله ابتدایی، پارامترهای انرژی تمام باند، انرژی باند پایین (0-1Khz) ، نرخ عبور از صفر و ضرایبLSF که شامل 11 ضریب بدست آمده از G.729A هستند، استخراج می شود. در قسمت بعدی یک تخمین از

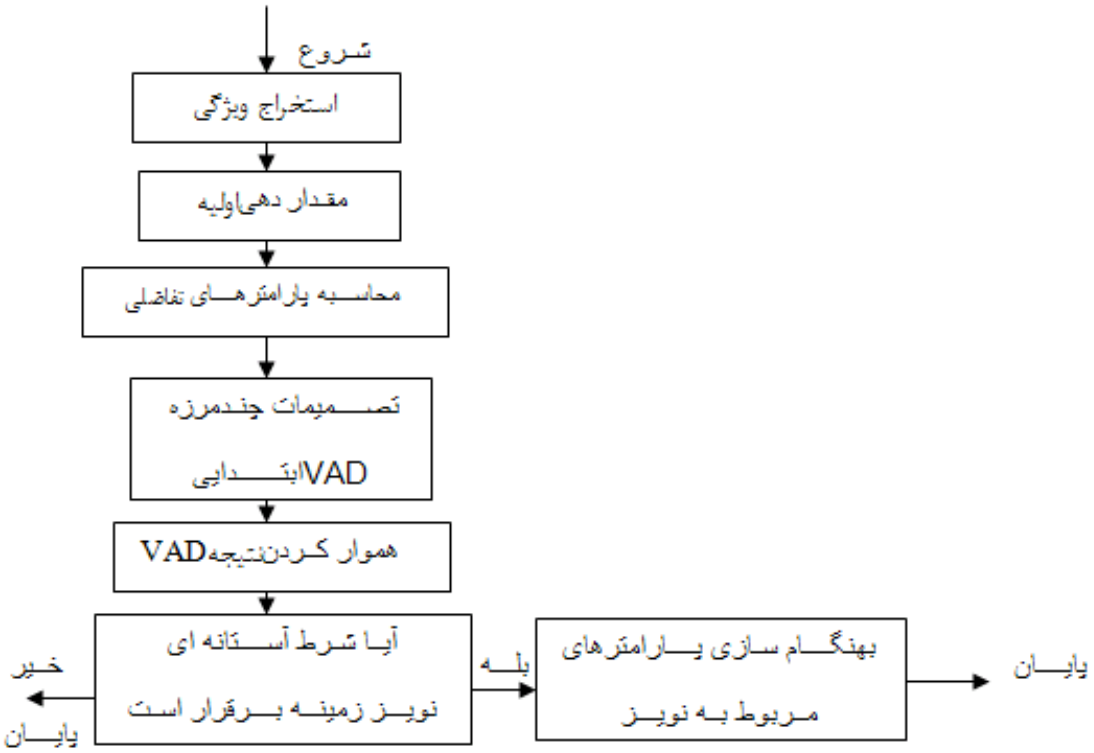
[126] .Line Spectral Frequencies
[127] .Silence Insertion Descriptor
[128] .Discontinues Transmission

طیف نویز با توجه به سکوت بودن فریم های ابتدایی انجام می شود. مرحله بعد، محاسبه مقدار تفاضلی برای هر پارامتر می باشد. در این مرحله تفاضل هر پارامتر با پارامتر متناظر مربوط به طیف نویز که در فاز قبل محاسبه شد، بدست می آید. برای محاسبه تفاضل 11 ضریب LSF، از مجموع مربع تفاضل هر ضریب استفاده می شود. میتوان یک فضای چهار بعدی را فرض کرد که قسمتی از این فضا به ناحیه گفتار(فعال) و قسمت دیگر به عنوان مکمل قسمت فعال، قسمت غیر فعال یا سکوت اختصاص دارد. برای انجام عمل تصمیم گیری از 14مرز(خط) در جهت تشخیص استفاده شده است. بعضی از شرط های مورد استفاده در این بخش، با رابطه (5-1) بیان شده اند، که در آن Δp_i و Δp_k دو پارامتر تفاضلی و a و b ثابت می باشند.[39]

$$if \quad \Delta p_i < a.\Delta p_k + b \qquad then \qquad flag = 1 \qquad i,k = 1,\dots,4; \tag{5-1}$$

اگر چهارده شرط گفته شده، هیچکدام صدق نکنند، مقدار پرچم صفر قرار داده می شود. این عمل را تصمیم گیری VAD چند مرزه[129] می نامیم. در نهایت برای هموار کردن نتیجه خروجی VAD از مقدار انرژی و تصمیم VAD راجع به دو فریم قبل استفاده می شود. برای مثال در قدم اول اگر انرژی فریم از حدی بالاتر بود، فریم گفتار اعلام می شود و در قدم بعدی اگر اختلاف انرزی فریم با دو فریم قبل که گفتار بودند، از حدی کمتر بود، فریم گفتار اعلام می شود. مراحل بالا در چهار مرحله به اجرا در می آیند.[39] همانطور که در شکل (5-3) نشان داده شده، در نهایت در صورت رخداد سکوت درسیگنال مقادیر مربوط به فریم نویزی، بهنگام می شود.

[129] .Multi Boundary Decisions



شکل (5-3) : نمودار الگوریتم G.729B [60]

با حذف سکوت میزان بار محاسباتی سیستم کاهش پیدا می کند و سرعت سیستم بیشتر می شود.[47] بدلیل بالا بودن حجم سیگنال گفتار (شامل اطلاعات غیر مفید می باشد.) و همچنین قابل استفاده نبودن سیگنال نمونه برداری شده بدلیل تغییر حالات افراد در هنگام صحبت کردن و ایجاد حالتهای مختلف در زمان های متفاوت، شکل سیگنال گفتار تغییرات بسیاری دارد و ثابت و یکنواخت نیست. بنابراین بصورت مستقیم از روی سیگنال گفتار نمیتوان تصمیم گیری نمود. در نتیجه از بردارهای ویژگی استفاده می نماییم. این بردارها حالت- های ثابتی دارند و ویژگی ثابت گوینده و صدایش را بیان می دارند. در این مرحله سعی می گردد ویزگی هایی استخراج شود که مناسب ترین اطلاعات مربوط به سیگنال اصلی را بیان دارد و در عین حال حجم پایینی داشته باشد. عملیات فوق بر روی فریم های با طول 10 میلی ثانیه تا 50 میلی ثانیه انجام می گیرد. سپس با استفاده از معیار فاصله BIC (که در فصل های قبلی مفصلا توضیح داده شد.) بخش بندی سیگنال صوتی به سگمنت های مجزا انجام می شود. در این مرحله علاوه بر اینکه سیستم نقاطی را بعنوان نقاط تغییر مشخص می نماید، نقاطی نیز بعنوان نقاط تغییر که زمان تغییرات را نشان می دهد، بصورت دستی و توسط آزمایشگر تعیین شده و در فایل هایی با قالب txt. به برنامه داده شده اند تا

تعیین دقیقتر مکان نقاط تغییر امکان پذیر شود. و اگر فاصله بین نقطه تشخیصی سیستم و نقطه مشخص شده در فایل تهیه شده کمتر از 200 میلی ثانیه باشد، آن نقطه بعنوان نقطه صحیح پذیرفته می شود. در این مرحله اگر نقطه تغییر گوینده، اضافه تعیین شود، امکان تصحیح خطا در مرحله خوشه بندی وجود دارد. ولی اگر نقطه تغییر واقعی گوینده را در این مرحله از دست بدهیم، کارایی مرحله خوشه بندی و در نهایت کل سیستم را خراب می نماییم. بنابراین باید فاکتورها (طول پنجره BIC و ...) را مناسب انتخاب نماییم. این مراحل با استفاده ازکد نویسی در محیط برنامه MATLAB اجرا گردیده اند. در نهایت با استفاده از ماشین بردارهای پشتیبان (SVM) عمل خوشه بندی انجام می شود. برای این مرحله از نرم افزار WEKA (این نرم افزار توسط دانشگاه نیوزلند در سال 2002 به صورت رایگان جهت استفاده عموم عرضه شده است. نمونه های به روز شده نرم افزار نیز توسط این دانشگاه ارائه می‌گردد. یکی از کاربردهای آن دسته بندی نمودن دادگان مورد آزمایش می‌باشد.) استفاده شده است.

5-3-پایگاه داده

جهت بررسی عملکرد سیستم و اعلام نتایج، باید از دادگان گفتاری معتبر استفاده شود. به همین دلیل در این قسمت به معرفی دادگان مورد استفاده می پردازیم. این داده ها به شرح زیر می باشند:

1)دادگان فارس دات: دادگان گفتاری استاندارد برای زبان فارسی یا همان فارس دات، مجموعه‌ای از عبارات و جملات است که توسط گویندگان فارسی زبان از مناطق مختلف کشور بیان شده است. این دادگان در سطح واج (آوا) با دقت میلی‌ثانیه تقطیع و برچسب‌دهی شده و بصورت فایل‌های مجزا ذخیره گردیده است. این دادگان، به عنوان دادگان استاندارد گفتاری زبان فارسی در داخل و خارج کشور شناخته شده است. بعضی از ویژگی ها و قابلیت‌های دادگان فارس‌دات:

1- استخراج پرکاربردترین کلمات زبان فارسی از روزنامه‌ها.

2- طراحی 386 جمله با استفاده از 1000 کلمه شامل کلیه دنباله‌های دوآوایی در زبان فارسی.

3- متوازن بودن 386 جمله مذکور از لحاظ آوایی.

4- انتخاب 304 گوینده بر حسب جنسیت، سن، میزان تحصیلات و لهجه از نقاط مختلف کشور.

5- پوشش 10 لهجه رایج فارسی در کشور (تهرانی، ترکی، اصفهانی، جنوبی، شمالی، خراسانی، بلوچی، کردی، لری و یزدی)

6- تولید 20 جمله در 2 جلسه توسط هر گوینده با کیفیت صوتی بسیار بالا در اتاقک ضد صدا.

این داده ها ، شامل گفتار اشخاص به تنهایی در یک فایل می باشند. برای استفاده در این پروژه این داده ها در کنار هم قرار گرفتند تا هر فایل صوتی شامل گفتار تعدادی گوینده باشد که دارای جنسیت های مختلف می- باشند. در این مجموعه از داده ها، تعداد گویندگان از 3 تا 20 نفر تغییر داده شد و همچنین نسبت تعداد مرد به زن نیز متغیر انتخاب شد تا در حالت های مختلف، نتایج سیستم مورد بررسی قرار گیرد. فرکانس نمونه برداری داده ها 16 کیلوهرتز و داده ها به صورت 16 بیتی می باشند. قالب مجموعه داده ها (.wav) Wave می باشد. (دادگان فارسی دارای فرکانس 22 کیلوهرتز بوده اند و با عمل تغییر نرخ نمونه برداری و جهت مقایسه با نتایج دیگر بدست آمده، فرکانس دادگان به 16 کیلوهرتز تغییر داده شدند.)

2)دادگان جلسه ای AMI: این مجموعه داده شامل 100 ساعت جلسه ضبط شده است. این جلسات در سه اتاق با شرایط آکوستیکی متفاوت و با گوینده های غیربومی ضبط شده اند. فرکانس نمونه برداری داده ها 16 کیلوهرتز و داده ها به صورت 16 بیتی می باشند. قالب مجموعه داده ها (.wav) Wave می باشد. نام فایل های صوتی برای دادگان AMI به طور کلی به صورت زیر است.

حرف اول زبان فایل گفتاری را نشان می‌دهد. (E: انگلیسی)، حرف دوم: S نشان دهنده تک کاناله بودن فایل صوتی ضبط شده می‌باشد. اعداد (2002) سال ضبط فایل صوتی را نشان می‌دهند. حرف چهارم نشان دهنده نام فایل می‌باشد. (a,b,c,d یعنی چهار فایل مختلف در سال 2002 می‌باشد.

ES2002a, ES2002b, ES2002c, ES2002d.

3)دادگان آزمایشگاهی ضبط شده: علاوه بر داده های بالا، یک سری داده هایی که در آزمایشگاه ضبط گردید، نیز مورد استفاده قرار گرفت. این فایل های صوتی ضبط شده به صورت 16 بیتی و تک کاناله در اتاق معمولی ضبط شده اند و دارای فرکانس نمونه برداری 8 کیلوهرتز می باشند. نام گذاری این فایل ها بدین صورت بوده است:

نام ها به صورت انگلیسی، از چپ به راست به شرح زیر می باشند:

-حرف اول نشان دهنده زبان مکالمه است.F نشانگر فارسی، E نشانگر انگلیسی، M نشانگر ترکیب زبان ها است.

-حرف دوم شرایط ضبط است. R: نشانگر اتاق ساکت و 1 میکروفن، M: نشانگر میکروفن دهنی، T: نشانگر مکالمه تلفنی است.

-حرف سوم تعداد گویندگان را نمایش می دهد.

-حروف بعدی هم جنسیت گویندگان را مشخص می کنند. M نشانگر آقا و F نشانگر خانم است.

فرمت مجموعه داده ها (.wav) Wave می باشد.

5-4-استخراج ویژگی ها

در مرحله بخش بندی سعی بر استخراج ویژگی هایی از سیگنال اصلی می باشد که، مناسب ترین اطلاعات سیگنال اصلی را شامل شود و همچنین کمترین حجم ممکن را دارا باشد. در سیستم پیشنهادی ابتدا از 4 نوع بردار ویژگی بهره برده ایم. این بردارهای ویژگی عبارتند از:

1)ضرایب MFCC با طول 13،که از تحلیل فیلتر بانک بدست می آید.

2) ضرایب root MFCC(در MFCC بجای تابع لگاریتم از توابع نمایی استفاده نموده ایم.)

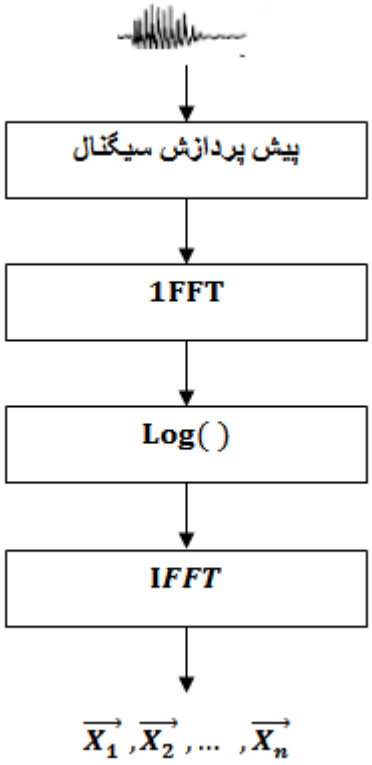
3)ضرایب TDC[130] (ضرایب کپسترال دو بعدی)

4) root TDC

در مورد ضرایب در فصول قبلی کاملا توضیح داده شد، در اینجا به توضیح مختصری راجع به ضرایب کپسترال دو بعدی می پردازیم. برای بدست آوردن این ویژگی ها ابتدا از سیگنال، تبدیل فوریه دو بعدی گرفته می شود، سپس از نتایج حاصل از مرحله قبل لگاریتم گرفته می شود و در نهایت تبدیل فوریه یک بعدی معکوس از نتایج مرحله دوم گرفته می شود تا به بردارهای ویژگی برسیم. در شکل (5- 4) بلوک دیاگرام روش نشان داده شده است. برای بدست آوردن ضرایب root از توابع نمایی، بجای لگاریتم استفاده

[130] .Two-dimensional cepstrum

می نماییم. توان تابع نمایی بکار رفته عددی بین صفر و یک می باشد. مقادیر مختلف در این بازه مورد آزمایش قرار گرفتند و در نهایت مقدار 0.9 مناسب تشخیص داده شد. بدلیل تاثیر نامحسوس تغییر این پارامتر بر روی نتایج بخش بندی، نمودار این تغییرات رسم نشده است.



شکل(5- 4): نمودار بردار ویژگی TDC

نتایج حاصل از هر چهار روش محاسبه و در نهایت با هم مقایسه شده اند. عملیات فوق بر روی فریم های با طول نمونه های متفاوت انجام شده است(از 80 نمونه تا 400 نمونه) و بهترین مقادیر بدست آمده نوشته شده اند.

5-4-معیار ارزیابی[131] سیستم های تشخیص گوینده

برای ارزیابی سیستم از نظیرکردن یک به یک گوینده های مرجع به گوینده هایی که سیستم آنها را تشخیص داده است، استفاده می شود. این موضوع را در نظیر کردن مورد توجه قرار می دهیم که هر گوینده در مرجع باید حداکثر به یک گوینده در خروجی سیستم نظیر شود و هر گوینده در خروجی سیستم باید حداکثر به یک گوینده در مرجع نسبت داده شود. معیار اصلی ارزیابی سیستم، عبارت است از کسری از

[131] .Evaluation Criterion

رشته داده ی صوتی که به درستی به گوینده ای نسبت داده شده است. معیارهای مختلفی در مقالات مختلف برای ارزیابی کارایی الگوریتم های بخش بندی و خوشه بندی مورد استفاده قرار می گیرد. در ادامه به توضیح رایج ترین آنها می-پردازیم.

اگر یک الگوریتم بخش بندی به خوبی تغییرات صحیح گوینده ها را نتیجه دهد، هر بخش تشخیصی تنها شامل گفتار یک گوینده خواهد بود. در آشکارسازی تغییرات گوینده ها، با دو نوع خطا روبرو هستیم:

-خطای درج[132]: زمانی که یک تغییر گوینده آشکار شود ولی در مرجع این تغییر گوینده وجود ندارد.

-خطای حذف[133]: تغییر گوینده وجود دارد، ولی این تغییر آشکار نشده است.

این خطاها تاثیرات متفاوتی با توجه به کاربرد بر روی سیستم دارند. در سیستم هایی که ابتدا بخش بندی و سپس کلاسه بندی انجام می شود، خطاهای درج که باعث بخش بندی بیش از اندازه می شوند از خطاهای حذف کم اهمیت تر هستند، چون در مرحله خوشه بندی با دسته بندی سگمنت های متعلق به هر گوینده، امکان تصحیح خطاهای درج وجود دارد. در حالیکه خطاهای حذف را نمیتوان در این مرحله تصحیح نمود.[40] برای تحلیل خطاها به یک مرجع تغییرات گوینده نیاز داریم، که این مرجع توسط بخش بندی دستی که خیلی هم دقیق نیست، بدست می آید. همانطور که قبلا توضیح داده شد، نقطه تغییر گوینده مرز هر سگمنت می- باشد. محل درست مرز برای یک سگمنت بطور دقیق تعریف نمی شود و اکثرا دو سگمنت بوسیله سکوت کوتاهی از هم جدا می شوند. و هر مرز سگمنت که داخل این محدوده سکوت قرار گیرد، مرز صحیح تلقی می- شود. بنابراین یک محدوده زمانی Δt تعریف می شود که اگر مرز سگمنت فرضی داخل فاصله زمانی $t_0 - \Delta t < t < t_0 + \Delta t$ از مرز مرجع t_0 قرار گیرد، این مرز به عنوان یک مرز صحیح در نظر گرفته می- شود.[64] با آزمایشاتی که انجام شد، مقادیر مناسب، 100تا 200 میلی ثانیه تشخیص داده شد. معیارهای ارزیابی کارایی واحد بخش بندی بصورت زیر تعریف می شوند:

[132] .Insertion errors
[133] .Deletion errors

-دقت[134](PRC) : نسبت تعداد مرزهای صحیح آشکار شده به مجموع تعداد مرزهای آشکار شده توسط واحد بخش بندی. این خطا هنگامی اتفاق می افتد که نقطه تغییر گوینده آشکارشده، صحیح نباشد. (خطای درج)

-فراخوانی [135](RCL): نسبت تعداد مرزهای صحیح آشکارشده به کل تعداد مرزها (تمام نقاط تغییر گوینده در مرجع). این خطا هنگامی اتفاق می افتد که واحد بخش بندی، نقطه تغییر گوینده ای را از دست بدهد. (خطای حذف)

سیستمی که دو معیار دقت و فراخوانی بالایی داشته باشد، مطلوب تر است. برای اینکه هر دو عامل فوق را برای بیان میزان کارایی الگوریتم به راحتی مورد استفاده قرار دهیم، یک معیار هارمونیکی که F نام دارد و ترکیبی از دو عامل فوق با وزن یکسان است، و با رابطه (2-5) تعریف می شود، مورد استفاده قرار می گیرد.

$$F = \frac{2 * PRC * RCL}{PRC + RCL} \qquad (2\text{-}5)$$

با توجه به رابطه (2-5)، میتوان گفت که $0 < F < 1$ تغییر می کند. اگر F=1 باشد، به معنای بخش بندی کاملا درست است و F=0 یعنی بخش بندی کاملا نادرست است.[15] البته در بعضی مقالات به جای موارد فوق از روابط زیر استفاده می نمایند.

$$FR = 1 - RCL \quad , \quad FD = 1 - PRC \qquad (3\text{-}5)$$

این معیارها مقایسه‌ای است بین قسمت‌های تشخیص داده شده و قسمت‌هائی که در دادگان مورد آزمایش موجود بوده است. $\%FD$ و $\%FR$ توسط روابط (4-5) و (5-5) محاسبه می شوند.

$$\%FD = \frac{\text{تعداد نقاطی که در دادگان مرجع نقطه‌ی تغییر نبوده‌اند ولی توسط سیستم به عنوان نقطه‌ی تغییر معرفی شده‌اند}}{\text{کل تعداد نقاطی که توسط سیستم به عنوان نقطه‌ی تغییر معرفی شده‌اند}} \qquad (4\text{-}5)$$

$$\%FR = \frac{\text{تعداد نقاطی که در دادگان مرجع نقطه‌ی تغییر بوده‌اند ولی توسط سیستم تشخیص داده نشده‌اند}}{\text{تعداد نقاطی که توسط سیستم به درستی به عنوان نقطه‌ی تغییر معرفی شده‌اند}} \qquad (5\text{-}5)$$

[134] .Precision
[135] .Recall

برای بررسی سیستم پیاده سازی شده در این پایان نامه نیز از این معیارها استفاده شده است. البته معیار دیگری که در سال های اخیر برای محاسبه خطای سیستم های تشخیص گوینده مورد استفاده "نرخ خطای سیستم تشخیص گوینده " یا DER[136] نامیده می شود.[29] با استفاده از رابطه (5-6) محاسبه می‌شود.

$$\%DER = \frac{SE + MS + I}{SPK} \quad (6\text{-}5)$$

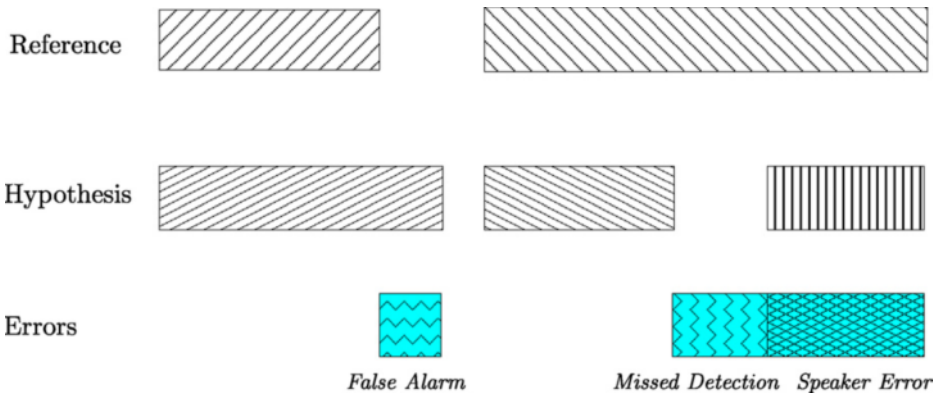
SE[137]: کل زمانی که برای گوینده‌ی اشتباهی اختصاص داده شده است.

MS[138]: کل زمانی که در آن تعداد گویندگان کمتری نسبت به حالت صحیح تشخیص داده شده است.

FA[139] : کل زمانی که در آن تعداد گویندگان بیشتری نسبت به حالت صحیح تشخیص داده شده است.

SPK[140] : جمع کل زمان سخنان هر گوینده که به مرجع اختصاص داده شده است.

به بیان دیگر می‌توانیم بگوئیم که DER نسبت کل زمانی است که خطا اتفاق افتاده به کل زمان عملیات میباشد. این مفهوم در شکل (5-5) نشان داده شده است.



شکل (5-5): تشخیص خطا در سیستم های تشخیص گوینده[29]

کارایی مرحله کلاسترینگ نیز با معیار K سنجیده می شود.[70]

$$K = \sqrt{acp * asp} \quad (7\text{-}5)$$

acp : عبارت است از درستی(خلوص) خوشه متوسط و از رابطه (5-4) بدست می آید.

[136]. Diarization Error Rate (DER)
[137] . Speaker Error time
[138] . Missed Speaker time
[139] . False Alarm Speaker time
[140] . Scored Speaker time

$$acp = \frac{1}{N}\sum_{i=1}^{N_c} p_{i_2} n_{i_2} \quad (8\text{-}5)$$

p_{i_2}مقدار فوق خلوص یک خوشه است و با رابطه (5-5) در زیر تعریف می شود:

$$P_i = \sum_{j=1}^{N_s} \frac{n_{ij}^2}{n_i^2} \quad (9\text{-}5)$$

n_{ij} :مجموع تعداد فریم های صحبت گوینده j در خوشه i

N_s : مجموع تعداد گویندگان و N_c : مجموع تعداد خوشه ها و N: مجموع تعداد قاب ها.

n_i : مجموع تعداد قاب ها در خوشه i و n_j : مجموع تعداد قاب های صحبت شده توسط گوینده j

همچنین P_j (خلوص گوینده) و *asp* از روابط زیر محاسبه می شود:

$$P_j = \sum_{i=1}^{N_c} \frac{n_{ij}^2}{n_j^2} \quad (10\text{-}5)$$

$$asp = \frac{1}{N}\sum_{j=1}^{N_s} p_j \ \ n_j \quad (11\text{-}5)$$

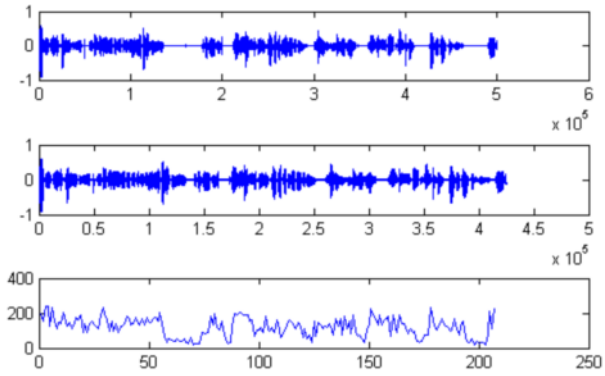
مقدار asp بیان می کند که تا چه اندازه یک گوینده به تنها یک خوشه محدود می شود و مقدار acp نشان می- دهد که به چه میزان یک خوشه به تنها یک گوینده محدود می شود. در ادامه فصل، مقادیر محاسبه شده خطا در جداول ذکر شده اند.

5-6-نتایج آزمایشات

در این بخش، با توجه به توضیحات داده شده قبلی، نتایج حاصل را مشاهده و بررسی می نماییم. نکته قابل توجه این است که، مراحل کار شامل پارامترهایی است که تغییرات آنها روی نتایج تاثیرگذار است. از مهمترین آنها طول پنجره VAD ، طول پنجره BIC می باشد. در ادامه نمودار اثر تغییرات پارامترهای تاثیرگذار روی دقت سیستم آورده شده است. جداول با مقادیر بهینه تنظیم شده اند و در نهایت با بهترین مقادیر به محاسبه خطا در سیستم پرداخته شده است.

5-6-1-اثر اعمال VAD بر روی سیگنال گفتار

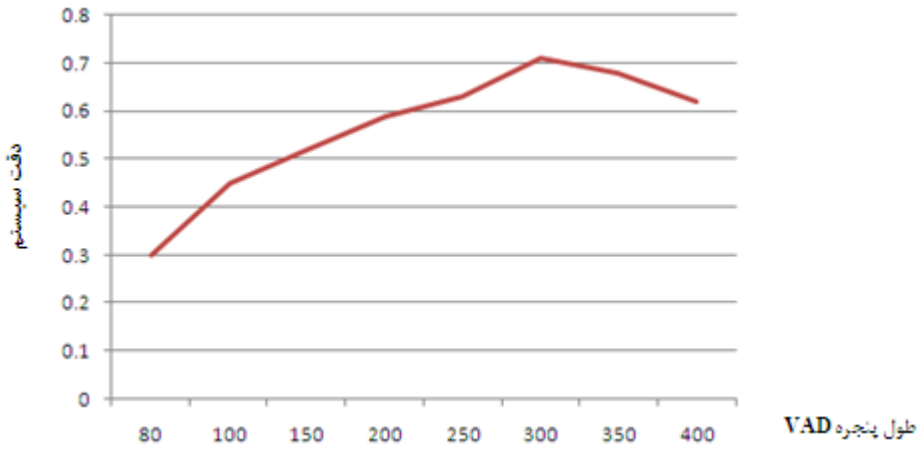
همانطور که قبلا توضیح داده شد، مرحله اول آزمایش، جدا کردن قسمت های گفتاری سیگنال صوتی از قسمت های غیر گفتاری آن می باشد، که در شکل (5-6) برای یک سیگنال گفتاری که شامل گفتار تعدادی گوینده می باشد، نشان داده شده است. شکل اول سیگنال ورودی، شکل دوم سیگنال خروجی مرحله اول است که قسمت های سکوت آن حذف شده اند و بخش های گفتاری باقی مانده است. شکل سوم مدل گوسی متناظر با سیگنال خروجی را نشان می دهد.



شکل(5- 6): جداسازی قسمت های گفتاری از غیر گفتار

5-6-2-اثر تغییر طول پنجره VAD بر روی دقت سیستم

در این مرحله، همانطور که در بالا توضیح داده شد، انتخاب طول پنجره مناسب اهمیت دارد و بر روی دقت سیتم تاثیر گذار است. طبق نتایج مشاهده شده، با افزایش طول پنجره دقت در حال افزایش می باشد. ولی این افزایش تا 43 میلی ثانیه خوب بوده است و بعد از این اندازه، طول های بزرگتر پنجره باعث حذف جزییات شده اند و افزایش خطای سیستم را در پی داشته اند. مینیمم طول پنجره 80 نمونه در نظر گرفته شده است و مقادیر کوچکتر جواب مناسبی ندارند و نهایتا اندازه پنجره تا 450 نمونه بزرگ شده است. در نمودار شکل (5-7) این مطلب نشان داده شده است.

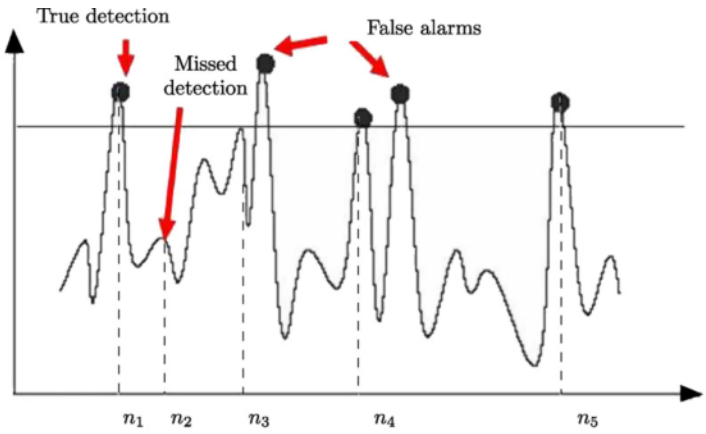


شکل(5-7): اثر تغییر طول پنجره VADبر روی دقت سیستم (محور افقی طول پنجره برحسب نمونه می باشد و محور عمودی میزان دقت سیستم را نشان می دهد.)

همانطور که از شکل مشخص است بهترین طول پنجره انتخابی که کمترین میزان خطا را دارند، از 250 تا 350 نمونه می باشد.

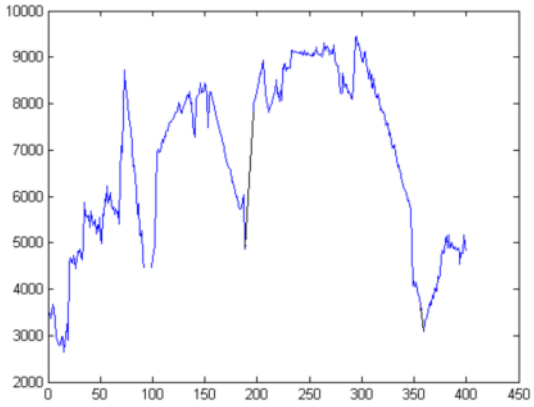
5-6-3-اثر تغییر طول پنجره BIC بر روی نتایج بخش بندی

مرحله دوم پیاده سازی سیستم ، استفاده از BIC و بخش بندی خروجی مرحله قبل به سگمنت های همگن می باشد. این مرحله بسیار مهم و تاثیرگذار بر روی نتایج نهایی سیستم است. همه روزه روش هایی جدید جهت بهبود عملکرد این بخش ارایه می شود. کار اصلی پایان نامه، بکارگیری بردار ویژگی های متفاوت جهت بهبود نتیجه خروجی این بخش می باشد. همانطور که می دانیم سیگنال صوتی به وسیله مدل گوسی شبیه سازی می شود. میزان مقادیر ماکزیمم قله ها در یک پنجره با هم مقایسه می شوند و سپس تصمیم گیری با توجه به این مقادیر انجام می شود. نقاط تغییر مشخص می شوند. در این مرحله نیز بزرگی و یا کوچکی اندازه پنجره هایی که انتخاب می شوند مهم می باشند تا در حد امکان نقاط تغییر، در زمان های مناسب و بطور صحیح تشخیص داده شوند. در شکل (5-8) این موضوع به خوبی نشان داده شده است.



شکل(5-8): چگونگی قرار دادن یک آستانه و نحوه انتخاب نقاط تغییر گوینده.

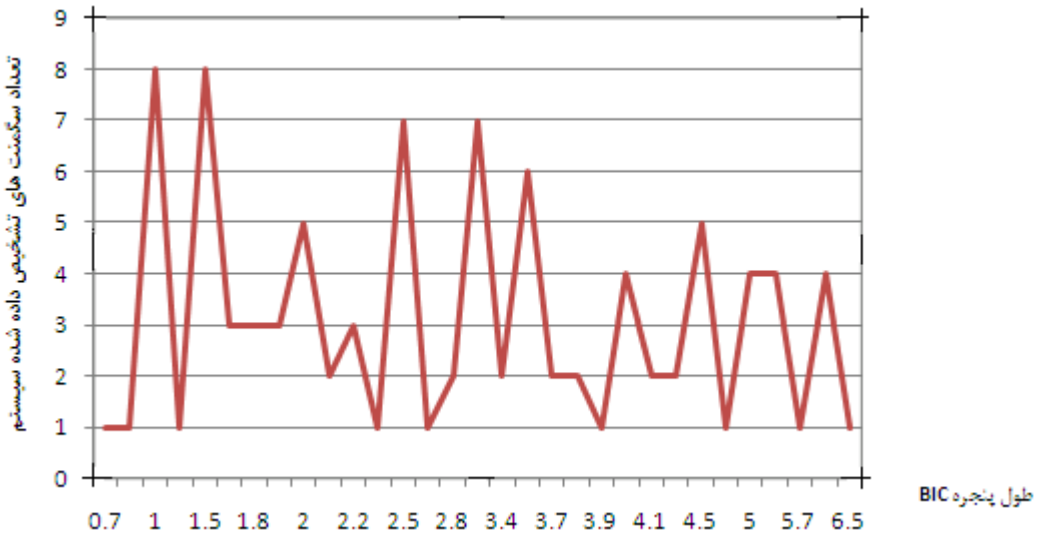
همانطور که شکل نشان می دهد، با توجه به اندازه پنجره انتخابی، بعضی از نقاط به درستی و تعدادی به اشتباه نقطه تغییر گوینده تشخیص داده شده اند. در شکل (5-9) مدل گوسی متناظر با سیگنال گفتار نشان داده شده است.



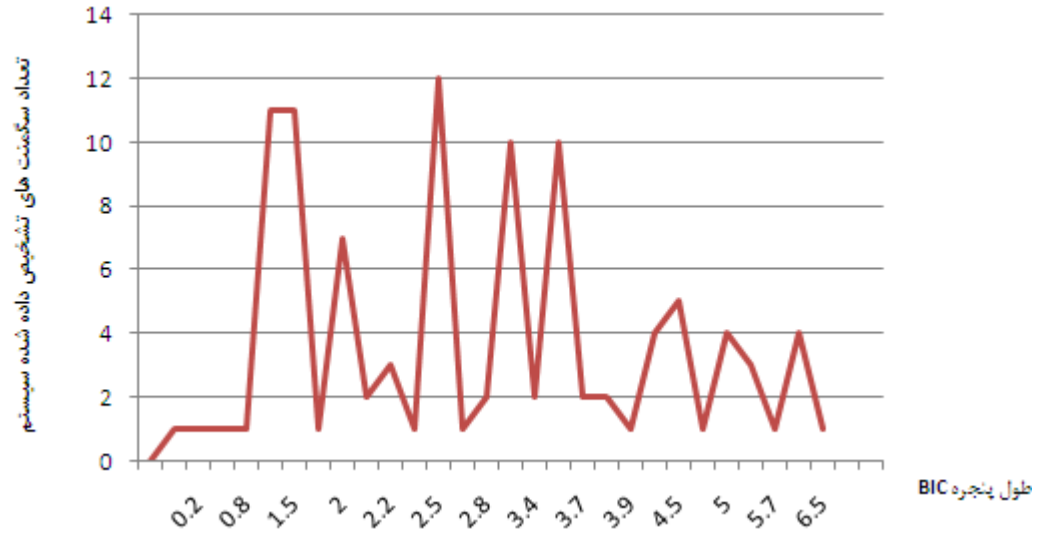
شکل(5- 9): سیگنال گفتاری گوسی مدل شده در مرحله بخش بندی

با توجه به طول پنجره ای که در الگوریتم BIC تعیین می شود، در محدوده پنجره تمام مقادیر سیگنال بررسی می شوند و سپس ماکزیمم مقدار در هر پنجره به عنوان نقطه تغییر گوینده در نظر گرفته می شود.اندازه پنجره BIC بر روی دقت و سرعت سیستم تاثیر گذار است. هر چه اندازه پنجره انتخابی کوچکتر باشد، زمان محاسبه بیشتر خواهد بود و دقت سیستم بالاتر می باشد. در این پایان نامه مقادیر مختلف برای طول پنجره مورد بررسی قرار گرفت. همانطور که در نمودارهای زیر مشخص است، پنجره های با طول 1 تا 2.5 ثانیه به خوبی توانسته اند سگمنت های گفتاری در فایل های با تعداد گویندگان مختلف را جداسازی

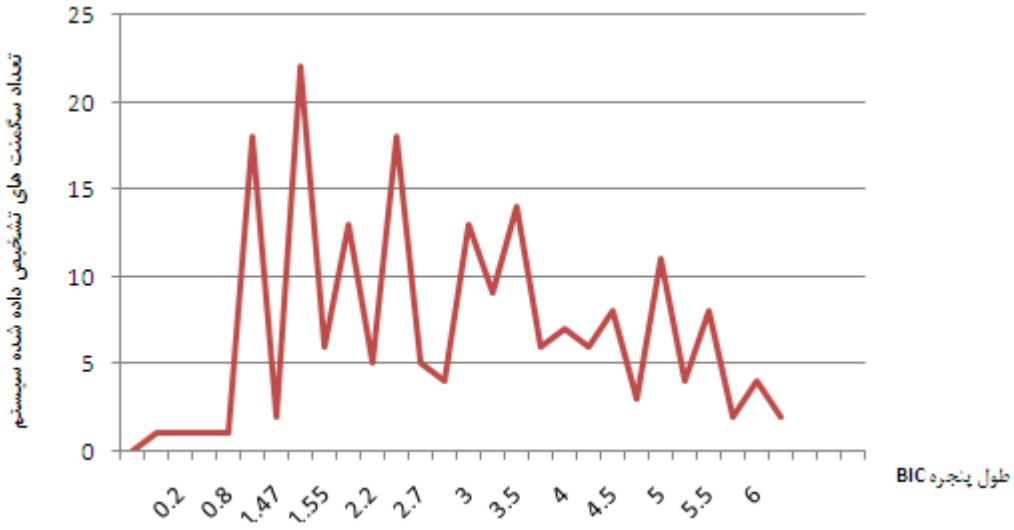
نمایند. در نمودارهای (5-10) و (5-11) و (5-12) اثر تغییر طول پنجره برای چندین فایل مختلف نشان داده شده اند. محور افقی طول پنجره بکار رفته جهت بخش بندی را نشان می دهد. (بر حسب ثانیه) و محور عمودی نشان دهنده تعداد سگمنت های تشخیصی توسط معیار بکار رفته BIC می باشد.



شکل(5-10): اثر افزایش طول پنجره BIC بر روی نتیجه مرحله بخش بندی برای 8 نفر دادگان فارس دات



شکل(5-11): اثر افزایش طول پنجره BIC بر روی نتیجه مرحله بخش بندی برای 12 نفر دادگان فارس دات



شکل(5- 12): اثر افزایش طول پنجره BIC بر روی نتیجه مرحله بخش بندی برای 18 نفر دادگان فارس دات

همانطور که از نمودار های بالا نیز مشخص است، انتخاب طول مناسب پنجره در سیستم های تشخیص گوینده در مرحله بخش بندی اهمیت زیادی دارد. با توجه به ویژگی های فایل گفتاری باید اندازه مناسب را انتخاب نماییم. بزرگتر شدن بیش از اندازه این طول و کوچک شدن بیش از حد، هر دو مطلوب نیستند. با توجه به نمودارهای بالا مقدار 2.5 ثانیه برای پنجره BIC مقدار مناسبی است. مقادیر بزرگتر از 4.5 ثانیه نتایج چندان مطلوبی ندارند. در این مرحله باید توجه داشت که اگر فاصله زمانی بین نقطه تشخیص داده شده سیستم و نقطه مشخص شده در فایل مرجع (این نقاط توسط آزمایشگر بصورت تجربی (با گوش دادن به فایل صوتی) و با استفاده از نرم افزارهای ویرایش فایل های صوتی مشخص شده اند.) کمتر از 200 میلی ثانیه باشد، آن نقطه به عنوان نقطه صحیح تغییر گوینده توسط سیستم مورد پذیرش قرار می گیرد.

5-6-4-دقت حاصل از بخش بندی بر دو نوع از دادگان با استفاده از MFCC

در این قسمت با استفاده از بردار مقیاس مل، دقت مرحله بخش بندی بر روی دادگان فارسی آزمایشگاهی تهیه شده و دادگان AMI اندازه گیری شده است.

جدول(5-1): مقادیر خطا برای دادگان تهیه شده فارسی آزمایشگاهی

دیتاها / خطاها	%FD	FR	RCL	PRC	%F
FR3MMM1	48.82	0	1	0.5118	67.71
FR3MMM2	47.97	0	1	0.5203	68.44
FR4MFMF1	45.69	0	1	0.5431	70.39
FR4MFMF2	46.08	0	1	0.5392	70.06
EM3MFF	48.69	0	1	0.5131	67.82

جدول(5-2): مقادیر خطا برای دادگان AMI

ویژگی ها / خطاها	%FD	FR	RCL	PRC	%F
ES2002a	42.87	0.632	0.994	0.5713	72.55
ES2002b	41.62	0.5618	0.9438	0.5838	73.56
ES2002c	44.073	0.1449	0.9855	0.5592	71.36
ES2002d	38.884	0	1	0.6111	75.87

5-6-5-اثر تغییر بردار ویژگی بر روی دقت مرحله بخش بندی

در ادامه در جداول زیر مقادیر دقت حاصل از مرحله بخش بندی با اعمال بر روی دادگان فارس دات، با اعمال بردارهای ویژگی متفاوت نشان داده شده است.

جدول(5-3): مقادیر خطا برای تعداد 3 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها	%FD	FR	RCL	PRC	%F
MFCC	34.88	0	1	0.6512	78.87
Root-MFCC	34.09	0	1	0.6591	79.45
TDC	35.78	0	1	0.6422	78.21
Root-TDC	35.44	0	1	0.6456	78.46

جدول(5-4): مقادیر خطا برای تعداد 5 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها ا	%FD	FR	RCL	PRC	%F
MFCC	35.00	0.0556	0.9444	0.6500	77.00
Root-MFCC	34.93	0.0556	0.9444	0.6507	77.05
TDC	35.71	0.0605	0.9395	0.6429	76.34
Root-TDC	35.93	0.0683	0.9317	0.6407	74.97

جدول(5-5): مقادیر خطا برای تعداد 8 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها ا	%FD	FR	RCL	PRC	%F
MFCC	36.32	0.0667	0.9333	0.6368	75.70
Root-MFCC	36.25	0.0667	0.9333	0.6375	75.76
TDC	35.89	0.0701	0.9299	0.6368	75.59
Root-TDC	34.97	0.0794	0.9206	0.6503	76.21

جدول(5-6): مقادیر خطا برای تعداد 11 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها	%FD	FR	RCL	PRC	%F
MFCC	35.23	0.1750	0.8250	0.6477	72.57
Root-MFCC	35.56	0.1750	0.8250	0.6444	72.36
TDC	36.08	0.1903	0.8097	0.6392	71.44
Root-TDC	35.74	0.1884	0.8116	0.6426	71.72

جدول(5-7): مقادیر خطا برای تعداد 14 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها	%FD	FR	RCL	PRC	%F
MFCC	35.68	0.0385	0.9615	0.6432	77.08
Root-MFCC	35.48	0.0192	0.9808	0.6452	77.84
TDC	36.56	0.0332	0.9668	0.6344	74.06
Root-TDC	36.79	0.0385	0.9615	0.6321	76.27

جدول(5-8): مقادیر خطا برای تعداد 17 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها	%FD	FR	RCL	PRC	%F
MFCC	36.20	0.0313	0.9688	0.6380	76.93
Root-MFCC	35.63	0.0156	0.9844	0. 6437	77.84
TDC	36.11	0.0289	0.9711	0.6389	77.07
Root-TDC	36.27	0.0306	0.9694	0.6373	76.90

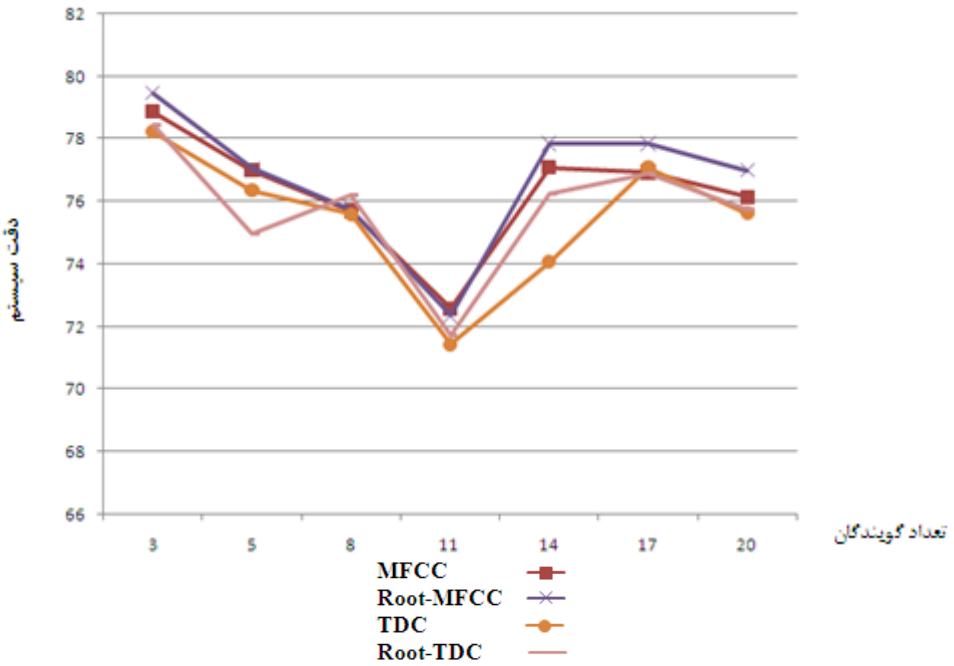
جدول(5-9): مقادیر خطا برای تعداد 20 نفر گوینده در دادگان فارس دات

ویژگی ها / خطاها	%FD	FR	RCL	PRC	%F
MFCC	36.28	0.0541	0.9459	0.6372	76.14
Root-MFCC	35.71	0.0405	0.9595	0.6429	76.99
TDC	36.75	0.0601	0.9399	0.6325	75.61
Root-TDC	36.34	0.0645	0.9355	0.6366	75.76

می بینیم که با توجه به بردار ویژگی بکار رفته میزان خطای سیستم متفاوت خواهد بود.

5-6-6-مقایسه نتایج مرحله بخش بندی با بکارگیری بردارهای ویژگی متفاوت

با توجه به مقادیر جداول بالا، در نمودار (5-13) مقادیر دقت های بدست آمده مرحله بخش بندی با ویژگی های متفاوت بکارگرفته شده، نشان داده شده اند.



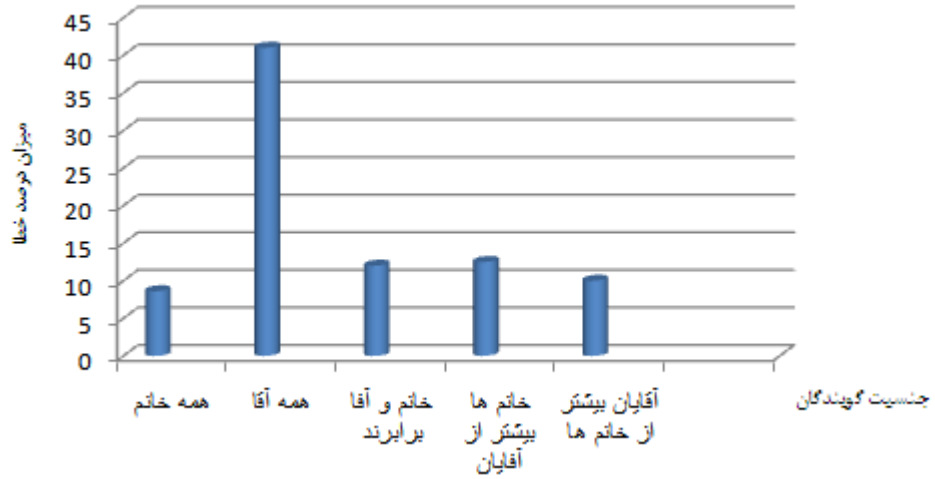
شکل(5-13): مقایسه میزان خطای سیستم با تغییر بردار ویژگی مورد استفاده. (محور افقی تعداد گویندگان و محور عمودی میزان دقت سیستم می باشد.)

نتایج دقت حاصل از بکارگیری چهار بردار ویژگی ، نزدیک به هم بوده اند. میزان درصد دقت های بدست آمده، تقریبا در هر چهار روش نزدیک به هم بوده است. کمترین خطا با استفاده از root-MFFCC و بیشترین خطا متعلق به TDC بوده است. بطور کلی MFCC و root-MFCC بدلیل دارا بودن دقت بالاتر در این سیستم، مناسب تر می باشند.

5-6-7-اثر جنسیت گویندگان بر تشخیص درست مرزهای بخش بندی

در این مرحله عواملی مانند جنسیت گویندگان موجود در فایل صوتی نیز بر روی نتایج حاصل تاثیرگذار بوده اند. بطوریکه اگر سیگنال صوتی فقط شامل گویندگان مرد بوده است، نسبت به حالتی که تعداد گویندگان به صورت دیگری بوده است، میزان خطای مرحله بخش بندی سیستم بالاتر بوده است. و سیستم به خوبی عمل تشخیص مرزها را به انجام نرسانده است. در صورتی که اگر سیگنال فقط شامل گوینده های

خانم بوده است، عمل سگمنت بندی با خطای کمتری انجام شده است و با تعداد گویندگان برابر خانم و آقا این نتایج بهتر بوده است. نمودار (5-14) مطالب فوق را نمایش می دهد.



شکل(5- 14): تاثیر جنسیت بر روی خروجی مرحله بخش بندی سیستم

5-6-8-دقت مرحله خوشه بندی با بکارگیری ماشین بردار پشتیبان(SVM) با بردار ویژگی MFCC

مرحله سوم پیاده سازی در این سیستم، دسته بندی نمودن سگمنت های حاصل از مرحله دوم می باشد. این مرحله با استفاده از ماشین بردار پشتیبان SVM به اجرا در می آید. در این مرحله، آزمایشات زیادی انجام شد. در جدول (5-10) بهترین نتایج بدست آمده، از این مرحله با توجه به دسته دادگان گفتاری، نشان داده شده است. زمان انجام عمل دسته بندی نیز در جدول ذکر شده است. این زمان هر چقدر کمتر باشد، سیستم سریع تری خواهیم داشت، که در برخی کاربردها حایز اهمیت می باشد.

جدول(5-10): دقت مرحله کلاسترینگ با استفاده از ماشین بردار پشتیبان با بکارگیری MFCC

نوع دادگان	دادگان فارسی	AMI	دادگان آزمایشگاهی فارسی
دقت خوشه بندی	%57	50%	%43
زمان خوشه بندی	1.63 S	1.08 S	S 1.12

5-6-9-دقت مرحله خوشه بندی با ماشین بردار پشتیبان با بکارگیری بردار ویژگی root-MFCC

چون بهترین نتایج مرحله بخش بندی با بکارگیری بردار ویژگی root-MFCC به دست آمده است، این مرحله با اعمال این بردار ویژگی نیز اجرا شده است و نتایج در جدول (5-11) نشان داده شده اند.

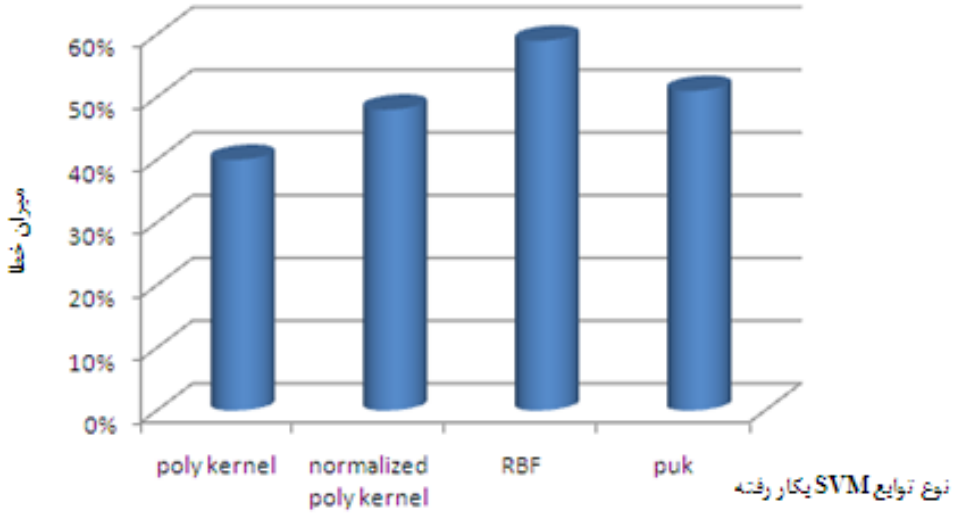
جدول(5-11): دقت مرحله خوشه بندی با استفاده از ماشین بردار پشتیبان با بکارگیری root-MFCC

نوع دادگان	دادگان فارسی	AMI	دادگان آزمایشگاهی فارسی
دقت خوشه بندی	%59	51%	%44
زمان خوشه بندی	1.41 S	0.84 S	S 1.2

میزان خطای سیستم با دادگان مختلف متفاوت است. این موضوع می تواند ناشی از تفاوت پایگاه دادگان، میزان نویز فایل گفتاری، تفاوت شرایط ضبط دادگان، فارسی یا انگلیسی بودن دادگان و ... باشد. در این مرحله داده آموزشی(ترین) بکار رفته در سیستم بسیار مهم می باشد. این موضوع، که تا چه میزان داده های مناسبی انتخاب شده اند و تا چه اندازه این داده ها به داده های آزمایش سیستم نزدیک می باشند، بر روی نتیجه سیستم و میزان دقت اثرگذار است. در این پایان نامه سعی شده است که تا حد امکان دادگان مناسبی تهیه شوند. همانطور که نتایج موجود در جدول نشان می دهند، نتایج حاصل از بکارگیری بردار ویژگی root-MFCC که ایده مورد استفاده در این پایان نامه بوده اند، نسبت به بکارگیری بردار ویژگی MFCC بهتر می باشند.

5-6-7-اثر تغییر نوع تابع کرنل ماشین بردار پشتیبان بر روی دقت مرحله خوشه بندی

همچنین در این مرحله و با استفاده از ماشین بردار پشتیبان، با تغییر تابع کرنل بکار رفته جهت خوشه بندی، نتایج متفاوتی حاصل شده است. در شکل (5-15) این نتایج نشان داده شده اند.



شکل(5- 15): مقایسه نتایج میزان خطای حاصل از خوشه بندی با تغییر نوع تابع کرنل بکارگرفته شده

همانطور که از نمودار نیز مشخص است، تابع کرنل به نام poly kernel نتایج بهتری نسبت به مابقی توابع کرنل دربر دارد و کمترین میزان خطا را نسبت به سایر توابع بکار گرفته شده، داشته است.

5-7-خلاصه

این فصل به پیاده سازی و مشاهدات سیستم پیشنهادی اختصاص داده شد.با توجه به معیارهایی که از آنها میتوان در ارزیابی سیستم های تشخیص گفتار استفاده نمود، و راهکارهایی که در فصول قبلی به آنها اشاره شد، در این فصل به ارزیابی این روش پیشنهادی پرداخته ایم. نتایج حاصل از تغییرات طول پنجره VAD بر روی دقت سیستم و تغییرات طول پنجره در بخش بندی سیگنال گفتاری و تغییر نتایج مرحله بخش بندی با تغییر بردار ویژگی بکارگرفته شده، تاثیر جنسیت گویندگان بر نتیجه مرحله بخش بندی، اثر تغییر تابع بکارگرفته شده در ماشین بردار پشتیبان بر روی نتایج مرحله کلاسترینگ، مورد بررسی قرار گرفتند و نتایج در جداول و نمودارها نشان داده شدند.

فصل ششم:

جمع بندی و پیشنهادات

6-1-جمع بندی و خلاصه نتایج

در این پایان نامه، جهت پیاده سازی سیستم های تشخیص گفتار، سه مرحله اصلی به اجرا درآمدکه این مراحل در تمامی سیستم های اینچنینی مورد استفاده اند و فقط الگوریتم ها و روشهای انتخابی برای هر مرحله در سیستم های مختلف متفاوت می باشد. در این پایان نامه در مرحله اول از یک جداساز گفتار از غیر گفتار استاندارد مخابراتی (G.729B) استفاده گردید. که این جداساز از ویژگی های انرژی، نرخ عبور از صفر و ضرایب LSF برای انجام عمل جداسازی استفاده می نماید و خروجی های گفتاری مناسبی را تولید می نماید. در مرحله بخش بندی که از مراحل پراهمیت در این سیستم ها می باشد، از الگوریتم BIC که الگوریتمی است که بر اساس معیار فاصله بین دو سگمنت مجاور عمل می نماید، استفاده گردید. و همچنین در این پایان نامه، از چهار بردار ویژگی (MFCC,root-MFCC,TDC,root-TDC) در این بخش استفاده گردید. نتایج حاصله، بالاتر بودن دقت مرحله بخش بندی با استفاده از بردار ویژگی root-MFCC را نشان داد. در مرحله سوم نیز از ماشین بردار پشتیبان برای دسته بندی نمودن سگمنت های گفتاری مرحله قبل استفاده گردید. این الگوریتم به صورت مقایسه بین دو خوشه انجام می شود و باید تمام خوشه ها را دو به دو با هم مقایسه نموده، و درصورت نزدیکی بین دو خوشه، آنها را با هم ترکیب می نماید. در نهایت سگمنت های گفتاری مربوط به یک گوینده در یک خوشه قرار می گیرند و هر خوشه به یک گوینده اختصاص می یابد. برای ارزیابی سیستم های اینچنینی، مطلوب این است که میزان خطای سیستم کم باشد و هر چه مقادیر خطا کمتر باشند، سیستم طراحی شده و انتخاب الگوریتم ها در بخش های مختلف مناسب تر بوده اند. که با توجه به بالاتر بودن دقت سیستم با بکارگیری root-MFCC نتایج حاصل در این پایان نامه با استفاده از ساختار سیستم بیان شده، نتایج خوبی می- باشند.

6-2- پیشنهادات

-این سیستم به صورت برون خط عمل می کند، یعنی سیستم از داده هایی که قبلا توسط افراد مختلف ضبط و ذخیره شده اند استفاده می نماید. میتوان برای پروژه های بعدی بر روی روش های برخط که بطور همزمان با پخش و ضبط گفتار، عمل تشخیص گوینده را انجام می دهند، کار نمود. این روش ها با این سیستم کنونی قابل اجرا نیستند.

-این پروژه بر روی داده هایی اجرا می شود که گفتار گوینده های مختلف با همدیگر همپوشانی ندارند. و اگرگفتارها همپوشانی داشته باشند، سیستم دچار اشتباه در تشخیص خواهد شد و خطا بالا می رود. برای کارهای آینده میتوان بر روی دیتاهای اینچنین عمل نمود.

-از الگوریتم و یا ترکیب الگوریتم هایی استفاده نماییم که جنسیت گویندگان بر روی میزان دقت و عملکرد سیستم تاثیرگذار نباشد.

-همانطور که دیده شد، این سیستم در تشخیص گفتارهای با طول کوتاه، خیلی خوب عمل نمی نماید، بنابراین برای رفع این مشکل از الگوریتم ها و روشهای مناسب دیگری باید استفاده نمود.

-در قسمت خوشه بندی که از ماشین بردار پشتیبان استفاده گردید میتوان از روشهای خوشه بندی متفاوتی که دارای دقت و سرعت های بالاتر می باشند ، استفاده نمود.

-بهتر است از الگوریتم هایی استفاده شود که با افزایش یا کاهش تعداد گویندگان، سیستم ضعف کمتری در جداسازی و تشخیص گویندگان از خود نشان دهد.

-نوع بردار ویژگی انتخابی، بر روی نتیجه تاثیر مستقیمی دارد، بررسی بردارهای ویژگی مناسب دیگری برای این منظور می تواند موضوع مناسبی برای کارهای آینده باشد.

-می دانیم که در ضبط گفتارهای گویندگان توسط میکروفون های مختلف، تاخیر در دریافت اطلاعات وجود دارد، بنابراین با استفاده از موقعیت گویندگان و لحاظ تاثیر آن در نتایج سیستم، میتوان سیستم های بهتری بدست آورد.

- حذف نویز همواره یکی از مهمترین مسایل در این سیستم ها بوده است و وجود نویز باعث خطا مخصوصا در مراحلی مانند جداسازی گویندگان می شود، بنابراین اگر سیستم بتواند نویز، موسیقی و ... را از گفتار تشخیص دهد، میتوان سیستم را نسبت به داده های مختلف مقاوم کرد.

منابع

[1].Xavier.Anguera.Mir, Phd Thesis, “Robust Speaker Diarization for meetings”, 2006.

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

[3]. Janes.Zibert, B.Vesnicer, F.Mihelie, ”A System for speaker detection and tracking in audio broadcast news”, IEEE proceeding, pp.51-61, 2008.

[4].A.F.Martin, M.A.Przybocki, “Speaker recognition in a multi-speaker environment”, Euro speech 2001 Scandinavia, Coference on Speech Communication and Technology, 2001.

[5]. R.O.Duda, P.E.Hart, D.G.Stork, “Pattern Classification” ,john wiley and sons , 2nd edition, 2007.

[6]. Christopher M.Bishop, “Pattern Recognition and Machine learning”, pp.738, Springer2006.

[7]. M.A.Siegler,U.Jain,B.Raj, M.Stern, “Automatic Segmentation, Classification and Clustering of Broadcast News Audio”, Proc.DARPA Speech Recognition Workshop, Chantilly, Virginia, pp.97-99, 1997.

[8].S.Chen, P.Gopalakrishnan, “ Speaker , Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion”, Proc .Darpa Broadcast News Transcription Understanding Workshop, Lansdowne, VA, USA, pp . 127-132, 1998.

[9].T.Hain, S.E.Johnson, A.Tuerk, P.C.Woodland, S.J.Young, “Segment generation and clustering in the HTK broadcast news transcription system”, Proc.Darpa Broadcast News Transcription and Understanding Workshop , Landsdowne, pp.133-137, 1998.

[10].J.Amera, C.Wooters, “ A Robust speaker clustering algorithm”, Proc.ASRU(Automatic Speech Recognition Understanding) Workshop, U.S. Virgin Islands, pp.411-416, 2003

[11].B.Zhou, J.H.L.Hansen, “Unsupervised Audio Stream Segmentation and clustering via the Baysian Information Criterion”, Proc. ICSLP, Beijing, China, pp. 714-717, 2000.

[12].K.Sommez, L.Heck, M.Weintraub, “Speaker Tracking and Detection with Multiple Speakers”, Proc. EUROSPEECH , Budapest, Vol. 5, pp. 2219 – 2222, 1999.

[13].P.C.Woodland, T.Hain, S.Johnson, T.Niesler, A.Tuerk, S.B.Young, “ Experiments in Broadcast News Transcription”, Proc.ICASSP, Seattle, Washington, pp.909 ff, 1998.

[14].L.Wilcox, F.Chen, D.Kimber, V.Balasubramanian, “Segmentation of Speech Using Speaker Identification “, Proc. ICASSP, Adeliade, Australia, Vol, pp. 161-164, 1994.

[15].H.Kim, D.Ertelt, T.Sikora, “ Hybrid speaker-based segmentation system using model-level clustering”, Proc. ICASSP, Philadelphia, USA, Vol,pp. 745-748, 2005.

[16].H.Kim, T.Sikora, “Automatic Segmentation of Speakers in Broadcast Audio Material”, Proc. SPIE, Vol. 5307, pp.429-438, 2003.

[17].P.Yu, F.Seide, C.Ma, E.Chang, “ An Improved Model-based Speaker Segmentation System”, Proc. EUROSPEECH, Geneva, Switzerland, pp. 2025-2028, 2003.

[18].D.Valj, B.Kacic, B.Horvat, “Usage of frame dropping and frame attenuation algorithms in automatic speech recognition system”, IEEE proceeding, pp.149-152, 2003.

[19].J.Faneuff, “Spatial, spectral, and perceptual nonlinear noise reduction for hands-free microphones in a car”, Master Thesis Electrical and computer Engineering, 2002.

[20].L.Karray, C.Mokbel, J.Monne, “ Solutions for robust speech\non speech detection in wireless environment”, IEEE proceeding, pp.166-170, 2002.

[21].همایونپور.م، ا.ش.نبوی، "مقایسه و ارزیابی روشهای تشخیص گفتار از سکوت"، کنفرانس بین المللی فن آوری اطلاعات، دی ماه 1382. صفحه 629-639

[22].D.R.Paoletti, G.Erten, “Enhanced silence detection in variable rate coding systems using voice extraction “, proc. 43IEEE Midwest symp, vol.2, PP.592-594, 2000.

[23].A.Benyassine, E.Shlomot, H.Yu Su, E.Yuen, “ Arobust low complexity voice activity detection algoritm for speech communication systems “, IEEE proceeding, pp. 97-98, 1997.

[24].A.Sangwan, M.C.Chiranth, H.S.Jamadagni, R.Sah, R.V.Prasad, V.Gaurav, “ VAD techniques for real-time speech transmission on the Internet”, 5th IEEE Internetional conference on High-speed Networks and Multimedia communications, pp. 46-50, 2002.

[25].S.G.Tanyer, H.Ozer, “Voice activity detection in non-stationary Gaussian noise” proceeding of ICSP,pp. 1620-1623. 1998.

[26].W.Shin, B.Lee, Y.Lee, “Speech/ non-speech classification using multiple features for robudt end point detection”, IEEE ICASSP, pp.876-881, 2000.

[27].B.V.Harsha, “Anoise robust activity detection algorithm”, proc. Of int. symposium of intelligent multimedia video and speech processing, pp. 322-325, 2004.

[28].R.Khemchandani, “Twin Support Vector Machines for Pattern Classification”, IEEE Transactions on pattern analysis and machin intelligence, pp.905-910, 2007.

[29].B.Fergani, M.Davy, A.Houacine, “ Speaker Diarization using one-class support vector machines”, Sience Direct, Speech Communication50, pp.355-365, 2008.

[30].H.I.Kim, S.K.Park, “ Voice activity detection algorithm using radial basis function network”, Electronics Letters, Vol.40, No.22, 2004.

[31].P.Renevey, A.Drygajlo, “Entropy based Voice Activity Detection in very noisy conditions”, Eurospeech’01 , pp.1883-1886 , 2001.

[32].Jia-Lin Shen, Jeih-Weih Hung, Lin-Shan Lee, “Robust entropy-based endpoint detection for speech recognition in noisy environments”, International Conference on Spoken Language, Sydney, Australia, November 30-December4, 1998.

[33].I.Abdullah, S.Montresor, M.Baudry, “Robust speech/non-speech detection in adverse conditions using an entropy based estimator”, IEEE proceeding, pp.757-760, 1977.

[34].R.Tucker, " Voice activity detection using a periodicity measure", IEEE Proceeding-I. VoI. 139, No.4, pp.377-380, 1992.

[35].I.D.Lee, H.P.Stern, S.A.Mahmoud, " A voice activity detection algorithm for communication systems with dynamically varying back ground acoustic noise", IEEE proceeding, pp.1214-1218, 1998.

[36].H.Kobatake, K.Tawa, A.Ishida, "Speech/non-speech discrimination for speech recognition system under real life noise environment ", IEEE proceeding, pp.365-368, 1989.

[37].J.Ramirez, J.C.Segura, C.Benitez, A.De la Torre, A.Rubio, " A new adaptive long-term Spectral Estimation voice activity detector", EUROSPEECH, pp.3041-3044, 2003.

[38].Ramirez et al, "Efficient voice activity detection algorithms using long-term speech information", speech communication, Vol.42, Issues 3-4, pp.271-278, 2004.

[39].F.Beritelli, S.Casale, A,Cavallaro, "A robust voice activity detector for wireless communication using soft computing", IEEE proceeding, pp.1818-1828, 1998.

[40].Q.Jin, K.Laskowski, T.Schultz, A.Waibel, "Speaker Segmentation and Clustering in meetings", ICSLP, JAEJU Island, Korea, pp.945-951, 2004.

[41].J.Rmirez, J.C.Segura, C.Benitez, A.De la Torre, A. Rubio, "An Effective Subband OSF-Based VAD with Noise Reduction for robust speech recognition" IEEE 2005.

[42].J.Wei, L.Du, Z.Yan, H.Zeng, "A new algorithm for voice activity detection ", IEEE proceeding, pp.588-590, 2003.

[43].Vijayachander, Shobha Devi, " A novel algorithm for voice activity detection", IEEE proceeding, pp.222-225, 2005.

[44].M.Jelinek, F.Labonte, "Robust signal/noise discrimination for wideband speech and audio coding", proc.IEEE Workshop on speech Coding, Delevan, Wisconsin, USA,pp.151-153, September 17-20, 2000.

[45].N.R.Garner, P.A.Barrett, D.M.Howard, A.M.Tyrrell, " Robust noise detection for speech detection and enhancement", electronics letters, Vol.33, No.4, pp.270-271, 1997.

[46].M.Orlandi, A.Santarelli, D.Falavigna, "Maximum Likelihood endpoint detection with time-domain features", eurospeech 2003, Geneva, pp.1757-1760.

[47].A.Acero, C.Crespo, C.Del La Torre, J.C.Torrecilla, "Robust HMM-based endpoint detection", Euro speech, pp.1551-1554, 1993.
[48].W.H.Abdullah, "HMM-based techniques for speech segments extraction", science programming, pp.221-239, 2002.

[49].H.Othman, T.Abdulnasr, "Asemi-continuos state transition propability HMM-based voice activity detection ", IEEE proceeding-I. Vol.139, No.4, pp.821-824, 2004.

[50].R.Sarikaya, J.H.L.Hansen, "Robust speech activity detection in the presence of noise", ICSLP, 1998.

[51].F.Beitelli, S.Casale, A.Cavallaro, "Adaptive voice activity detection for wireless communications based on hybrid fuzzy learning", IEEE proceeding, pp.1729-1734, 1998.

[52].A.Cavallaro, F.Beritelli, S.Casale, ”Afuzzy logic based speech detection algorithm for communications in noisy environment”, IEEE proceeding, pp.565-568, 1998.

[53].Y.Tian, J.Wu, Z.Wang, D.Lu,”Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection”, IEEE proceeding, pp.444-447, 2003.

[54].F.Beaufays, D.Boies, M.Weintraub, Q.Zhu, “Using speech/non-speech detection to bias recognition search on noisy data”, IEEE proceeding, pp.424-427, 2003.

[55].S.Grashey, “A new voice activity detection based on self organizing maps” ,Euro Speech, pp.1733-1736. 2003.

[56].A.Sangwan, H.S.Jamadagni, M.C.Chiranth, R.Sah, R.V.Prasad, V.Guarav, “Second and third adaptable threshold for VAD in VoIP”, IEEE proceeding, pp.1693-1696, 2002.

[57].C.Dong, K.Jinming, “ A robust voice activity detector applied for AMR”, proceeding of ICASP, pp.687-692, 2000.

[58].E.Cornu, H.Shikhzadeh, R.L.Brennan, H.R.Abutalebi, E.C.Y.Tam, P.Iles, K.W.Waong, “ETSI AMR2 VAD:Evaluation and ultra low resource implementation”, IEEE proceeding, pp.585-587, 2003.

[59].P.A.Barrette,”Information tone handling in the half rate GSM voice activity detector” , IEEE proceeding, pp.72-76, 1995.

[60].A.Benyassine, E.Shlomot, H.Yusu, ”ITU-T recommendation G.729 Annex B:A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application”, IEEE procedding, pp.64-73, 1997.

[61].J.Shaojun, G.Hitato, Y.Fuliang, ” Anew algorithm for voice activity detection based on wavelet transform”, proc.of int.symposium of intelligent multimedia, video and speech processing, pp.222-225, 2004.

[62].L.Rabiner, B.H.Juang,”Fundamentals of Speech Recognition” Prentice Hall, 1993.

[63].J.R.Deller, J.G.Proakis, J.H.L.Hansen, “Discrete-Time Processing of Speech Signals”, Macmillan Publishing Company, 1993.

[64].T.Kemp, M.Schmidt, M.Westphal, A.Waibel, ”Strategies for automatic segmentation of audio data”, Proc.ICASSP, Istanbul. Turkey, Vol.3, 1423-1426, 2000.

[65].S.Kwon, Sh.N, ”Unsupervised Speaker Indexing Using Generic Models”, IEEE Transactions on Speech and Audio Processing, Vol. 13, no.5, pp. 1004-1013, 2005.

[66].H.Gish, M.H.Siu, R.Rohlicek, “ Segregation of Speakers for Speech Recognition and Speaker Identification”, Proc. ICASSP, Toronto, Canada, Vol.2, pp.873-876, 1991.

[67].L.Lu, H.J.Zhang, “Content Analysis for Audio Classification and Segmentation “, IEEE Transaction on Speech and Audio Processing, Vol. 10, NO. 7, pp. 504-516, 2002.

[68].B.Zhou, J.H.L.Hansen, “Efficient Audio Stream Segmentation via the Combined T^2-Statistic and Bayesian Information Criterion”, IEEE Transsactions on speech and audio processing, Vol. 13, No.4, pp. 467-474, 2005.

[69].G.Schwarz, “Estimating the Dimension of a Model”, The Annals of statistics, Vol. 6, No. 2, pp.462-464, 1978.

[70].J.Ajmera, H.Bourlard, I.Lapidot, I.Mccowan,”Unknown-Multiple speaker clustering using HHM”, Proc.ICSLP,Denver, USA, PP.573-576, 2002.

[71].Laura Docio-Fernandez, Carmen Garcia-Mateo, “ Speaker Segmentation , Detection and Tracking in Multi Speaker Long Audio Recordings”, Third COST275 Workshop “Biometrics on the Internet”, University of Hertfordshire, Hatfield, UK, 2004.

[72].W.H.Tsai, S.S.Cheng, and H.M.Wang, “Speaker Clustering of Speech Utterancves using a voice characteristic reference space”, Proc. ICSLP, Jeju Island, Korea, pp.1237-1241, 2004.

[73].S.E.Tranter, M.J.F.Gales, R.Sinha, S.Umesh, P.C.Woodland, “ The Development of The Cambridge University RT-04 Diarisation System”, RT-04F Workshop, pp.1557-1565, 2004.

[74].C.Barras, X.Zhu, S.Meignier, J.-L.Gauvain, “Improving Speaker Diarization”, proc.RT-04F Workshop (Fall 2004 Rich Transcription Workshop), pp.1498-1503, 2004.

[75].Daniel.Moraru, Mathieu.Ben, Guillaume Gravier, “Experiments on Speaker tracking and segmentation in radio broadcast news”, INTERSPEECH, Lisbon, Portugal, pp.3049-3052, 2005.

[76].A.K.Jain, M.N.Murty and P.J.Flynn,” Data Clustering: A review”, ACM Computing Surveys, Vol. 31, No.pp.264-323, 1999.

[77].Kh.Aghajani, M.S Thesis, “Voice Activity Detection in the Speech Signal With Stationary Noise Based By Wavelet Transform”, sharifuniversity of technology, computer engineering department, 2006.

[78].H.Veisi,M.SThesis,”Model-based methods for noise robust speech recognitionsystems”, sharifuniversity of technology, computer engineering department, 2005.

[79].Y.Seyyedin, M.S Thesis, “Acoustic segmentation”, sharif university of technology, computer engineering department , 2009.

[80].L.Ardakanian, M.S Thesis, “ Speaker Clustering and Segmentation in a Multi-Speaker Environment”, amirkabiruniversity of technology electrical engineering department, 2006.

[81].B.Ahmed,W.Harvey,”A voice activity detector using Chi-Square test” IEEE proceeding, pp.625-628, 2004.

[82].S.Zhang, S.Zhang, B.Xu,”A Two-Level Method for Unsupervised Speaker-based Audio Segmentation”, IEEE, 18th international conference on pattern recognition, pp.1536-1540, 2006.

Abstract

In the every audio signal, it becomes very important answer questions like: “what was said?”, but also “who said it?” as information varies depending on who utters the spoken words. Within the speech technologies, The broad topic of acoustic indexing studies the classification of sounds into different classes/sources. Algorithms used for acoustic indexing worry about the correct classification of the sounds, but not necessarily about the correct separation of them when more than one exist in the same audio segment. These purely classification techniques have sometimes been called audio clustering, which benefit from the broad topic of clustering, well studies in many areas. When multiple sounds appear in the same audio signal one must turn his attention to techniques called as audio diarization to process them.These can include particular speakers, music, background noise sources.
When the possible classes correspond to the different speakers in a recording these techniques
are called speaker diarization. Speaker diarization can be defined in terms of being a subtype of audio diarization, where the speech segments of the signal are broken into the different speakers They aim at answering the question “Who spoke when?” given an audio signal. Algorithms doing speaker diarization need to locate each speaker turn and assign them to the appropriate speaker cluster. The output of the system is a set of segments with a unique ID assigned to each person that intervenes in the recording.
In this project using VAD’s G.729B in once step for seprate voice & unvoice. Then in this system using BIC algorithm for speech’segmentation by using MFCC’s feature, root-MFCC’s feature, TDC & root-TDC feature for second step ,and at last in the system using SVM for clustering.

Keywords:
Speaker Diarization, Voice Activity Detection, Speech Segmentation, Speaker Clustering

Shahrood University of Technology

Electronic Engineering Department

Speaker diarization in a multi-speaker environment using support vector machines

By: Marzieh Lashkarbolouki

Supervisor: Dr.Hossein Marvi

July 2011