دانشکده : برق و رباتیک

گروه : الکترونیک

پایان نامه دوره ی کارشناسی ارشد مهندسی برق – الکترونیک

کاربرد روش استخراج ویژگی RootMel جهت تخمین سن افراد با استفاده از سیگنال گفتار

عاطفه دهقانیان

استاد راهنما :

دکتر حسین مروی

استاد مشاور :

دکتر علی سلیمانی

شهریور ۱۳۹۱


بسم الله الرحمن الرحيم

تقدیم به پدر ومادرم

که هستی و همه ی وجودم از آن هاست

و همسرم

به پاس قدرانی و سپاسگذاری از زحمات و

دلسوزی هایشان

تشکر و قدردانی

در ابتدا از استاد راهنمای خود، جناب آقای دکتر حسین مروی به خاطر راهنمایی و حمایتی که از من در طی انجام این تحقیق کرده اند، نهایت تشکر را می نمایم.

و نیز از تمامی دوستانم در دانشگاه صنعتی شاهرود، که با حمایت و کمک بی دریغ خود در انجام این تحقیق، مرا همراهی کردند؛ تشکر می نمایم.

از پدر و مادر و همسرم به خاطر کمک ها و دلسوزی هایشان و به خاطر حمایت های معنوی که در طی انجام این تحقیق از من داشته اند، تشکر ویژه می نمایم.

چکیده :

تخمین سن بر اساس ویژگی های گفتار انسان، یک موضوع قابل توجه در سیستم های شناسایی گفتار اتوماتیک می باشد. مطالعاتی در زمینه ی تخمین سن گوینده صورت گرفته است ولی نیاز به کار های نوین بیشتری، خصوصا برای گوینده های فارسی زبان، می باشد. در تخمین سن، مانند سایر سیستم های پردازش گفتار، با دو چالش مهم مواجه هستیم: یافتن یک روش مناسب برای استخراج ویژگی و انتخاب یک روش قابل اطمینان برای کلاسه بندی.

هدف اصلی از این تحقیق استفاده از ویژگی Root Mel Ferequency Cepstral Coefficients جهت بخش استخراج ویژگی در یک سیستم تخمین سن و یافتن بهترین مقدار برای داشتن درصد خطای کمتر می باشد؛ و همچنین مقایسه ی عملکرد این ویژگی با سایر ویژگی های متداول نظیر MFCC معمول، PLP و LPC نیز مورد بررسی قرار گرفته است .

برای استخراج ویژگی، کل سیگنال صوت را به کوچک ترین جزء آن، یعنی واج ها، تفکیک می کنیم و عملیات استخراج ویژگی و رده بندی را بر سیگنال مربوط به واج ها انجام می دهیم.

از کلاسه بند به روش های تفکیک پذیری خطی و فواصل Mahalanobis استفاده شده است.

نتیجه آزمایشات انجام شده بر پایگاه داده ی FARSDAT به خطای کمترین ۲۸.۶۹ ٪ بازای ریشه ی ۰.۰۰۶ در استفاده از روش RootMFCC و تاثیر مثبت تفکیک سیگنال صوتی به واج های تشکیل دهنده ی آن، در کاهش خطا می باشد.

کلمات کلیدی : تخمین سن، گفتار، واج، استخراج ویژگی RootMel، FARSDAT

فهرست مطالب :

فهرست جدول ها..... د

فهرست شکل ها..... و

فهرست اصطلاحات ..... ی

۱- فصل اول : مقدمه..... ۱

۱-۱ خلاصه ای از چگونگی تغییر صدای انسان با افزایش سن..... ۴

۱-۱-۱ مکانیزم تولید گفتار با گذشت سن..... ۵

۱-۱-۲ سیستم تنفسی..... ۶

۱-۱-۳ حنجره..... ۶

۱-۱-۴ سیستم فرا حنجره ای..... ۶

۱-۱-۵ بالارفتن سن در مردها و زن ها..... ۷

۱-۲ تحلیل صوتی سن..... ۷

۱-۳ درک انسان از سن..... ۹

۱-۴ روش های تکنولوژی گفتار..... ۱۰

۱-۵ ساختار پایان نامه..... ۱۱

۲- فصل دوم : تحقیقات صورت گرفته در زمینه ی تخمین سن اتوماتیک توسط گفتار..... ۱۲

۳- فصل سوم : مروری بر انواع روش های استخراج ویژگی..... ۴۰

۳-۱ مقدمه..... ۴۱

۳-۲ استخراج ویژگی در برابر کلاسه بندی..... ۴۱

۳-۳ مراحل استخراج ویژگی از سیگنال گفتار..... ۴۲

۳-۳-۱ شکل دهی طیفی..... ۴۲

۳-۳-۲ تحلیل طیفی..... ۴۳

۳-۳-۳ تبدیل ویژگی..... ۴۵

۳-۴ دسته بندی تکنیک های استخراج ویژگی گفتار.....۴۵

۳-۵ روش های رایج استخراج ویژگی..... ۴۶

۳-۵-۱ ضرایب Linear Prediction Cepstral..... ۴۷

۳-۵-۲ ضرایب Perceptual Linear Prediction.....۴۸

۳-۵-۳ ضرایب Linear Frequency Cepstral..... ۵۱

۳-۵-۴ ضرایب Mel Frequency Cepstral.....۵۲

۳-۶ کاهش ویژگی :.....۵۳

۳-۶-۱ LDA(linear discriminant analysis).....۵۷

۳-۶-۲ تحلیل مؤلفه‌های اصلی (PCA).....۵۹

۳-۶-۳ مقایسه ی تصویری LDA , PCA.....۶۳

۳-۷ روش های تعمیم ویژگی MFCC برای کاربردهای مختلف.....۶۴

۳-۷-۱ روش استخراج ویژگی از ضرایب قسمت بالایی خود همبستگی

(higher – lag autocorrelation).....۶۴

۳-۷-۲ ضرایب MFCC دو طیفی.....۸۶

۳-۷-۳ غیر حساس کردن MEL- Cepstrum نسبت به اجزای طیفی

نادرست (بدلی) (RootMFCC)، برای شناسایی گفتار پایدار.....۹۱

۳-۷-۴ Autocorrelation MFCC.....۹۶

۳-۷-۵ Relative MFCC برای شناسایی گفتار تلفنی پایدار.....۱۰۰

۴ - فصل چهارم : روش پیشنهادی برای تخمین سن افراد.....۱۰۶

۴-۱ مقدمه..... ۱۰۷

۴-۲ معرفی پایگاه داده.....۱۰۸

۴-۳ معرفی الگوریتم پیشنهادی..... ۱۱۰
۴-۴ پیاده سازی الگوریتم پیشنهادی.....۱۱۱
۴-۴-۱ بکارگیری روش استخراج ویژگی MFCC.....۱۱۱
۴-۴-۲ بکارگیری روش استخراج ویژگی LPC.....۱۲۰
۴-۴-۳ بکارگیری روش استخراج ویژگی RootMel.....۱۲۱
۴-۴-۴ بکارگیری روش استخراج ویژگی PLP..... ۱۲۳
۴-۴-۵ اعمال روش PCA برای کاهش ابعاد ویژگی.....۱۲۴
۴-۴-۶ اعمال روش LDA برای کاهش ابعاد ویژگی..... ۱۲۵
۴-۴-۷ بررسی تاثیر تفکیک داده ها بر اساس جنسیت بر نتایج آزمایشات...۱۲۶
۴-۵ نتایج نهایی بدست آمده..... ۱۲۶
۴-۶ بررسی تاثیر تعداد گروه های سنی بر نتایج بدست آمده..... ۱۲۹
۴-۷ مقایسه نتایج آزمایش با رده بند های Linear , mahalanobis.....۱۳۰
۴-۸ ارزیابی نتایج حاصل ازالگوریتم های مختلف.....۱۳۱
۴-۹ نتیجه گیری کلی ۱۳۲
فهرست منابع..... ۱۳۵

فهرست جدول ها:

جدول ۱-۱: اطلاعاتی که سیگنال گفتار حمل می کند (با توجه به نظر Fujisaki -2004)..... ۱۰

جدول ۲-۱: نتایج بدست آمده از آزمایشات صورت گرفته در زمینه ی تخمین سن در [۹]..... ۱۳

جدول ۲-۲ : انواع ویژگی های استفاده شده در [۱۰]..... ۱۶

جدول ۲-۳ : انواع توابع kernel استفاده شده در SVM ۱۸

جدول ۲- ۴: انواع توابع فاصله مورد استفاده در K-NN ۱۸

جدول ۲-۵ : ماتریس ترکیبی بر پایگاه داده ی DES.[۱۰] ۱۹

جدول ۲-۶ : ماتریس ترکیبی بر پایگاه داده ی ELSDSR در[۱۰]..... ۱۹

جدول ۲-۷ : ماتریس ترکیبی نسبی برای سیستم SVM در [۱۱] ۲۱

جدول ۲-۸ : تعداد داده های آموزشی و آزمایش برای ۳ گروه سنی در [۱۲]..... ۲۲

جدول ۲-۹ : تعداد داده های آموزشی و آزمایشی برای ۱۳ گروه سنی در [۱۲]..... ۲۳

جدول ۲-۱۰: نتیجه ی آزمایشات برای تقسیم بندی افراد به ۳ گروه سنی در [۱۲]..... ۲۳

جدول ۲-۱۱: نتایج کمترین نرخ خطا برای مرتبه های مختلف PLP در [۱۳]..... ۲۵

جدول ۲-۱۲ : نرخ خطا ی بدست آمده برای مقادیر مختلف گاما، طول قاب و تعداد ضرایب MFCC در [۱۳]..... ۲۶

جدول ۲- ۱۳ : صحت پیش بینی سن و جنس بازای استفاده از رده بندهای متفاوت در [۱۴] ۲۸

جدول ۲-۱۴ : MAE(Mean Absolute Error)، آزمایشات با مجموعه ویژگی های متفاوت در [۸]..... ۳۲

جدول ۲-۱۵ : تعداد گروه های ویژگی مختلف. که بر اساس تحلیل MAXR انتخاب شده اند [۸]..... ۳۲

جدول ۲-۱۶: رده بندی تخمین گروه سنی بر اساس PNN برای [۷]..... ۳۳

جدول ۲- ۱۷ : رده بندی تخمین گروه سنی بر اساس GMM برای [۷] ۳۳

جدول ۲-۱۸ : تعداد گوینده ها در دسته های مختلف سن – جنس در [۱۵]..... ۳۶

جدول ۲- ۱۹ : صحت شناسایی گروه سنی بر حسب درصد در [۱۵]..... ۳۷

جدول ۲- ۲۰ : ماتریس ترکیبی نسبی برای روش مورد نظر در شناسایی ۶ گروه سن – جنسیت در[۱۵]..... ۳۷

جدول ۲- ۲۲ : ماتریس ترکیبی برای ۷ دسته ی سن – جنسیت در [۱۶]..... ۳۸

جدول ۳- ۱ : عملکرد تخمین فرمنت با استفاده از دو روش ذکر شده ، بر یک قاب ۳۲ میلی ثانیه ای از گفتار ساختگی نویزی در [۲۳]..... ۸۳

جدول ۳-۲ : نرخ خطای کلمه برای نویز کارخانه و f16 در [۲۵] ۹۵

جدول ۳- ۳ : نرخ شناسایی گوینده (٪) برای گفتار پاک برای مقایسه MFCC , A-MFCC در [۲۶].....۹۹

جدول ۳-۴: نرخ شناسایی گوینده (٪) برای گفتار آموزشی خراب شده با نویز F16 و کارخانه و سفید برای مقایسه MFCC , A-MFCC در [۲۶]..... ۹۹

جدول ۳- ۵ : اندازه های مختلف SNR برای پایگاه داده در [۲۷]..... ۱۰۵

جدول ۳-۶ : نرخ خطای کلمه با به کارگیری انواع روش های جبران کانال در [۲۷]..... ۱۰۵

جدول ۴-۱ : نتایج بدست آمده برای حالت اولMFCC (خطا ی آزمایش بر حسب درصد).....۱۱۵

جدول ۴-۲: نتایج بدست آمده برای حالت دوم a , b(خطای آزمایش بر حسب درصد).....۱۱۶

جدول ۴-۳ : نتایج بدست آمده برای حالت دوم c (خطای آزمایش بر حسب درصد).....۱۱۶

جدول ۴- ۴ : نتایج بدست امده برای حالت سوم (خطای آزمایش بر حسب درصد)..... ۱۱۸

جدول ۴- ۵ : نتایج بدست آمده برای حالت چهارم (خطای ازمایش بر حسب درصد).....۱۱۹

جدول ۴- ۶ : نتایج بدست آمده برای حالت اول LPC (خطا ی آزمایشی بر حسب درصد).....۱۲۰

جدول ۴- ۷ : نتایج بدست آمده برای حالت دوم LPC (خطای آزمایشی بر حسب درصد).....۱۲۰

جدول ۴-۸ : درصد خطای آزمایش برای روش استخراج ویژگی RootMFCC بازای گامای مختلف.....۱۲۲

جدول ۴-۹ : نتایج بدست آمده برای قسمت ۴-۴-۴ PLP(خطای آزمایش بر حسب درصد).....۱۲۳

جدول ۴-۱۰ :نتایج حاصل از اعمال تابع PCA بر ماتریس ویژگی های بدست آمده..... ۱۲۵

جدول ۴- ۱۱ : نتایج مربوط به اعمال تابع LDA بر روش ذکر شده در قسمت ۴-۴-۱ ،حالت (ب a).....۱۲۶

جدول ۴-۱۲ : نتایج حاصل از تفکیک داده ها بر اساس تفکیک جنسیت.....۱۲۶

جدول ۴-۱۳: مقایسه عملکردهای سیستم های مختلف تخمین سن با استفاده از گفتار.....۱۳۳

فهرست شکل ها:

شکل ۱-۱ : مکانیزم تولید گفتار صدا..... ۵

شکل۲-۱ : نتیجه ی رده بندی سن برای پایگاه داده ی DES در[۱۰] ۱۷

شکل۲-۲ : نتیجه ی رده بندی سن برای پایگاه داده ی ELSDSR در [۱۰]..... ۱۷

شکل ۲-۳: مدل کردن هر گوینده با یک GMM در [۱۱]..... ۲۰

شکل ۲-۴ : نمایشی از بانک فیلتر در مقیاس Mel [۱۲]..... ۲۲

شکل ۲-۵ : نمودار نرخ خطا بر حسب پارامتر گاما در RBF kernel ، با استفاده از PLP با ۳ مقدار متفاوت از مرتبه (به عنوان بردار ویژگی) در[۱۳]..... ۲۴

شکل ۲-۶ : نمودار نرخ خطا بر حسب پارامتر گاما در RBF kernel ، با به کارگیری مرتبه ی ثابت MFCC ، با طول قاب های زمانی مختلف در [۱۳] ۲۵

شکل ۲-۷ : نمودار نرخ خطا بر حسب پارامتر گاما در RBF kernel ، با به کارگیری مرتبه های مختلف MFCC ، با طول قاب زمانی برابر 25ms در [۱۳] ۲۶

شکل ۲-۸ : سلسله مراتب رده بندی در [۱۴]..... ۲۸

شکل ۲-۹ : مدل منحنی pitch در یک قسمت صدادار در [۸]..... ۳۱

شکل ۲-۱۰ : بلوک دیاگرامی از تخمین زننده ی گروه سنی در [۷] ۳۳

شکل ۲- ۱۱: بلوک دیاگرام روش مورد نظر در فاز آموزش اولیه برای آموزش WSNMF(weighted supervised non-negative matrix factorization) در [۱۵]..... ۳۵

شکل ۲- ۱۲ : بلوک دیاگرام روش مورد نظر در فاز آموزش ثانویه برای آموزش GRNN(general regression neural network) در [۱۵]..... ۳۶

شکل ۲- ۱۳ : بلوک دیاگرام روش مورد نظر در فاز آزمایش در [۱۵]..... ۳۶

شکل ۲- ۱۴ نمایی کلی از سیستم و چگونگی ترکیب ۵ روش مختلف و به کارگیری آنها در [۱۶]..... ۳۹

شکل ۳- ۱ : بلوک دیاگرام نشان دهنده ی ۳ گام اصلی در استخراج ویژگی برای شناسایی گفتار..... ۴۲

شکل ۳ -۲ : دسته بندی الگوریتم های تحلیل طیفی..... ۴۳

شکل ۳-۳ : مقایسه ی مقیاس Bark , Mel ۴۴

شکل ۳ – ۴ : گام های محاسبه ی LPCC ۴۷

شکل ۳ -۵ : گام های محاسبه یPLPCC ۴۸

شکل ۳- ۷: بلوک دیاگرام نشان دهنده ی گام های محاسبه ی MFCC ۵۲
شکل ۳ -۸ : نمایشی از تاثیر کاهش ویژگی در ابعاد یک بعدی و دو بعدی..... ۵۴
شکل ۳- ۹ : مسئله ی داده های ناکافی مشابه مسائل مطرح شده در curve fitting می باشد..... ۵۵
شکل ۳- ۱۰ : نمایشی از عملکرد روش LDA..... ۵۷
شکل ۳ -۱۱ : نمایشی از عملکرد PCA..... ۵۹
شکل ۳- ۱۲: تحلیل مقادیر ویژه ماتریس کواریانس..... ۶۲
شکل ۳-۱۳: مقایسه عملکرد PCA , LDA و نمایش میزان تفکیک نمونه های دو کلاس..... ۶۳
شکل ۳-۱۴ : مقایسه عملکرد PCA , LDA و نمایش میزان تفکیک نمونه های دو کلاس..... ۶۳
شکل ۳-۱۵ : طیف توان و تابع خودهمبستگی برای یک قاب ms ۳۲ از گفتار 'ey' voiced در [۲۳]..... ۶۸
شکل ۳-۱۶ : طیف توان و تابع خودهمبستگی برای یک قاب ۳۲ms از گفتار 's' unvoiced در [۲۳]..... ۶۹
شکل ۳- ۱۷ : تحلیل زمان کوتاه سیگنال نویز سفید اتفاقی مصنوعی با استفاده از قاب های ۳۲ms..... ۷۰
شکل ۳- ۱۸ : تحلیل زمان کوتاه سیگنال نویز chrip مصنوعی با استفاده از قاب های ۳۲ms..... ۷۱
شکل ۳- ۱۹ : تحلیل زمان کوتاه سیگنال نویز ضربه ای مصنوعی با استفاده از قاب های ۳۲ms..... ۷۲
شکل ۳-۲۰ : تحلیل سیگنال نویز اتومبیل با استفاده از قاب های ۳۲ میلی ثانیه ای..... ۷۳
شکل ۳- ۲۱ : نمایش روش HASE (با استفاده از رشته ی خود همبستگی higher – lag ، پنجره گذاری شده با hamming) بر یک قاب ۳۲ میلی ثانیه ای از گفتار voiced یک خانم. 'ey' ۷۵
شکل ۳-۲۲ : تابع پنجره و طیف توان آن برای a) پنجره ی hamming و b) پنجره kaiser (α=۱۱.۳) و c) پنجره ی DDR hamming ۷۶
شکل ۳-۲۳ : نمایش روش HASE (با استفاده از رشته ی خود همبستگی higher – lag ، پنجره گذاری شده با kaiser) بر یک قاب ۳۲ میلی ثانیه ای از گفتار voiced یک خانم. 'ey'..... ۷۸
شکل ۳-۲۴ : نمایش روش HASE (با استفاده از رشته ی خود همبستگی higher – lag ، پنجره گذاری شده با DDR hamming) بر یک قاب ۳۲ میلی ثانیه ای از گفتار صدادار یک خانم. 'ey' ۷۹
شکل ۳-۲۵ : مقایسه روش های تخمین طیفی با به کارگیری یک قاب ms ۳۲ از گفتار صدادار ساختگی پاک. ۸۰
شکل ۳- ۲۶ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز سفید اتفاقی ساختگی با 10 dB SNR تخریب شده است ۸۱

شکل ۳- ۲۷ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز chrip ساختگی با 10 dB SNR تخریب شده است..... ۸۲
شکل ۳-۲۸ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز ضربه ای اتفاقی ساختگی با 10 dB SNR تخریب شده است..... ۸۲
شکل ۳-۲۹ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز واقعی اتومبیل با 10 dB SNR تخریب شده است..... ۸۳
شکل ۳-۳۰ : طیف نگاره ی گفتار واقعی (عبارت 'MAL_19Z96Z8 A' از پایگاه داده ی Aurora)..... ۸۴
شکل ۳-۳۱ : بلوک دیاگرام الگوریتم استخراج ویژگی در AMFCC..... ۸۵
شکل ۳- ۳۲ : بلوک دیاگرام الگوریتم استخراج ویژگی در MFCC ۸۵
شکل ۳- ۳۳ : بلوک دیاگرام مراحل استخراج ویژگی در روش Mfcc دوطیفی در [۲۴]..... ۸۷
شکل ۳-۳۴ : نسبت تشخیص صحیح بر حسب SNR برای روش طیفی و روش دو طیفی در زمانی که نویز سفید گوسی اضافه شود..... ۸۸
شکل ۳-۳۵ : هیستوگرام مربوط به نویز های babble ، اتومبیل و کارخانه ۸۸
۳-۳۶ :نسبت تشخیص صحیح بر حسب SNR برای روش طیفی (خط توپر) و روش دو طیفی (خط چین) در زمانی که نویز babble اضافه شود..... ۹۰
۳-۳۷ نسبت تشخیص صحیح بر حسب SNR برای روش طیفی (خط توپر) و روش دو طیفی (خط چین) در زمانی که نویز اتومبیل اضافه شود..... ۹۰
۳-۳۸ نسبت تشخیص صحیح بر حسب SNR برای روش طیفی (خط توپر) و روش دو طیفی (خط چین) در زمانی که نویز کارخانه اضافه شود..... ۹۰
شکل ۳-۳۹ : انرژی های بانک log mel-filter از یک گفتار پاک و نویزی (مشوش) در [۲۵] ۹۲
شکل ۳-۴۰ : مربع انرژی های بانک log mel-filter برای گفتار پاک و نویزی..... ۹۳
در شکل ۳-۴۱ :اندازه ی تفاوت بین اولین ۱۳ ضریب DCT برای دو نمونه ی logMelFBS و اندازه ی تفاوت بین اولین ۱۳ ضریب DCT برای توان دوم logMelFBS..... ۹۳
شکل۳- ۴۲ محاسبه ی MFCC در [۲۶]..... ۹۸
شکل ۳- ۴۳ : بلوک دیاگرام پردازشگر A-MFCC در [۲۶]..... ۹۸
شکل ۳- ۴۴ : نمایش اعوجاج سیگنال گفتاردر [۲۷]..... ۱۰۲

شکل ۴-۱ : معرفی اجمالی الگوریتم پیشنهادی..... ۱۱۰
شکل ۴- ۲ : ۹ تکرار مربوط به واج "ای" که توسط شخص شماره ۱۰۰ در پایگاه داده ، ادا شده است..... ۱۱۳
شکل ۴- ۳ : بلوک دیاگرام مربوط به روش حالت اول..... ۱۱۴
شکل ۴-۴ : بلوک دیاگرام مربوط به حالت دوم ۱۱۷
شکل ۴- ۵ : بلوک دیاگرام مربوط به روش حالت سوم ۱۱۸
شکل ۴-۶ : بلوک دیاگرام مربوط به حالت چهارم ۱۱۹
شکل ۴- ۷ : بلوک دیاگرام مربوط به محاسبه ی ضرایب ویژگی RootMFCC ۱۲۱
شکل ۴-۸ : نمودار تغییرات خطای کل بر حسب تغییرات گاما ۱۲۳
شکل ۴- ۹ : نمودار مقایسه نتایج درصد خطای کل برای روش های مختلف..... ۱۲۷
شکل ۴- ۱۰ درصد خطا برای هر ۳گروه سنی مختلف (گروه hو گروه g و گروه w) برای همه ی روش های گفته شده ۱۲۸
شکل ۴- ۱۱ : درصد خطای کل برای تقسیم داده ها به ۲ گروه سنی ، ۳ گروه سنی و ۵ گروه سنی ۱۲۹
شکل ۴- ۱۲ : مقایسه درصد خطا در هر گروه سنی h,g,w ، برای ۴ روش نشان داده شده در شکل، با استفاده از روش mahalanobis distance ۱۳۰
شکل ۴- ۱۳ : مقایسه در صد خطای کل بدست آمده برای رده بند , linear و روش Mahalanobis برای ۴ روش استخراج ویژگی مذکور ۱۳۱

فهرست علائم و اختصارات:

Pitch : فرکانس صدا

Octave : فاصله ی زمانی(مدت) بین یک pitch و pitch دیگری با فرکانس دو برابر و یا نصف

Jitter : حداکثر مقدار انحراف از فرکانس پایه (F0)

Shimmer : تغییرات محلی توان

Formant :قله های موجود در طیف صوت (تجمع انرژی صوت در یک فرکانس خاص از شکل موج صوت)

linguistic information :

اطلاعات نمادین که با مجموعه ای از نمادهای مجزا و قواعدی برای ترکیب آن ها، تعریف می شوند.

paralinguistic information :

اطلاعاتی که از نوشته ها قابل استنتایج نمی باشند ولی توسط گوینده برای اصلاح و یا تکمیل اطلاعات نمادین، اضافه می شوند.

non-linguistic information :

در رابطه با سن، جنس، ویژگی شخصی و حالات فیزیکی و احساسی گوینده و... می باشد. اگرچه گوینده برای بیان احساسی خاص ، روش صحبت کردن خود را کنترل می کند، این ویژگی ها قابل کنترل نیستند.

ANN : artificial NEURAL NETWORK

HMM : hidden markove modle

GMM : Gussian mixture modle

KNN:K- nearest neighber

SVM: support vector machine

SVR:support vector regression

PNN : PROBABILISTIC NEURAL NETWORK

DFT : discrite fuorier transform

KLT : Karhunen-Loeve transform

ICA : INDEPENDENT COMPONENT ANALYSIS

EM : EXPECTATION MAXIMIZATION

MAP : MAXIMUM A POSTERIORI TRAINING ALGORITHM

RBF(Radial basic function)

RASTA: Relative Spectral

MAXR : MAX-Relevance (feature selection method)

فصل اول

مقدمه

در تولید گفتار، انواع متعددی از اطلاعات نیز به طور موازی تولید می شوند، هم اطلاعات زبان شناسی و هم تعداد زیادی از صفات شخصی گوینده. که اطلاعات اخیر شامل نشانه هایی است که به ویژگی های فیزیکی دستگاه صوتی گوینده مربوط می شود. بعلاوه تعدادی از ویژگی های غیر فیزیولوژیکی گوینده نیز تاثیر خود را بر سیگنال صوتی می گذارند. مانند اطلاعاتی درباره ی حالت احساسی گوینده، ویژگی های مذهبی و اجتماعی.

تعیین کیفی دقیق این اطلاعات کار دشواری است زیرا اکثر این اطلاعات با مجموعه نشانه های یکسانی علامت گذاری می شوند. مثلا فرکانس اصلی هم به رابطه ی بین محتوی علمی _معنایی عبارت و آهنگ تلفظ بستگی دارد و هم به بسیاری فاکتورهای دیگر مانند جنس و گروه سنی و حالت احساسی گوینده .

امروزه شاهد گسترش استفاده ی افراد مختلف جامعه از کودک و خردسال تا میان سال و کهنسال و همچنین قشرهای گوناگون جامعه کودکان و دانش آموزان، دانشجویان و محققان دانشگاهی، صنعتی و کارکنان ادارات شرکتی، استفاده های خانگی تا مدیریت های شرکتی و کارخانجات بزرگ صنعتی و کشاورزی از کامپیوتر هستیم.

این واقعیات به این معنی است که برای ارتباط سریع تر و آسان تر افراد از کامپیوتر، سیستم های مکالمه ای- گفتاری باید گسترش یابند تا بتوانند با کاربر ارتباط برقرار کرده و راهبرد های مکالمه ی صمیمانه با همه ی نسل ها را بهبود بخشد. اگرچه این کار با به کارگیری یک شیوه ی واحد و عمومی برای همه ی نسل ها امکان پذیر است، ولی کنترل دینامیک، انعطاف پذیر و دقیق بر ارتباط با کاربر و طرح های مکالمه ای، را می توان با تخمین اتوماتیک سن افراد و گوینده ها، محقق ساخت.

بیشتر سیستم های گفتاری – مکالمه ای رایج تنها دنباله ای از کلمات را از صدای گوینده استخراج می کنند. این کار به طور وسیعی از اطلاعات مفید دیگری که می توان از گفتار بدست آورد، چشم پوشی می کند. انسان می تواند یک سری صفات مهم را درباره ی گوینده، مانند جنسیت، سن، لهجه، نژاد، احساسات، سطح تحصیلات و حتی قد و وزن را استنباط کند. این نوع ویژگی های صوت می تواند برای طبیعی ساختن رابط های گفتاری انسان - ماشین، به طور بهینه ای مورد استفاده قرار گیرند. مثلا یک سیستم مشتری - محور با دانستن اینکه یک شخص مسن از قسمت جنوبی کشور که ناراحت و افسرده است، می تواند با ساختن یک دنباله ی مکالمه ای مناسب به لهجه ی جنوبی، مایه ی آرامش کاربر شود.

در ارتباطات گفتاری انسان - با - انسان، مخصوصا در ارتباطات شخص – با – شخص، انتظار می رود که گوینده، شیوه ی سخن گفتن خود را با توجه به خصوصیات و پاسخ های شنونده، تغییر دهد. در ارتباطات بر پایه ی گفتار، به ویژه در اولین ملاقات برای تنظیم رفتار مناسب، یک فاکتور مهم برای هر کسی سن می باشد. گفتار مردان و زنان نه تنها حامل بار معنایی عبارات گفته شده است بلکه شامل ویژگی هایی است که اطلاعاتی غیر زبانی، وابسته به گوینده را تامین می کنند مانند جنس، سن گوینده، حالت احساسی گوینده و ...با استخراج این ویژگی ها برای هر فرد، می توانیم نحوه ی صحبت کردن خود را با طرف مقابل تنظیم کنیم.

یک کاربرد رایج تخمین سن، هر روزه در مکالمات تلفنی صورت می گیرد. هم چنین بعضی شرکت ها نیاز به سیستم تخمین سن اتوماتیک دارند تا مثلا بتوانند برای گروه های سنی مختلف از مشتریان خود، موسیقی های مناسب پخش کنند.

علاوه بر سیستم های مکالمه ای، سن به عنوان یک پارامتر کلاسه بندی مهم در بسیاری دیگر از کاربردها، در نظر گرفته می شود. مثلا اجازه دادن به گروه سنی خاصی برای دسترسی به حقوقی ویژه و یا لحاظ کردن قیمت های متفاوتی در خرید کالاها برای گروه های سنی مختلف.

نماینده های اجرای قوانین برای تشخیص هویت یک فرد، تکنیک های بیومتریک مختلفی را مد نظر قرار می دهند. مشخصه های مختلف بیومتریک، می توانند برای هویت شناسی قانونی مورد استفاده قرار گیرند. مانند الگوهای اثر انگشت، مشخصه های چهره، ترکیب هندسی دست، تغییرات امضا ء و الگوهای صدا . انتخاب یک روش مناسب به اطلاعات موجود و نیز به درجه اطمینان آن روش در یک کاربرد مشخص، بستگی دارد. در بعضی از جُرم ها، مدرک در دسترس ممکن است به شکل مکالمات ضبط شده باشد. الگوهای گفتاری می توانند دارای اطلاعات مهمی برای ماموران قانون باشد. برای مثال، نمونه ی گفتاری یک شخص می تواند اطلاعاتی درباره ی سن، جنس، لهجه، حالات فیزیولوژیکی باشد ویا حتی اینکه آن شخص عضو یک گروه اجتماعی یا مذهبی خاصی باشد. در نتیجه، از گفتار می توان برای شناسایی گوینده که در موارد زیادی مانند آدم دزدی، تماس های تهدید آمیز و ... مورد نیاز است، استفاده کرد.

۱-۱: خلاصه ای از چگونگی تغییر صدای انسان با افزایش سن

هر موجود زنده ای گذر عمر و افزایش سن را تجربه می کند. این یک مکانیزم پیچیده است که از جنبه های مختلفی بر یک شخص تاثیر می گذارد. در نتیجه مورد بررسی قرار دادن مفهوم تغییر سن در اکثر علوم طبیعی و قوانین بشری، گریز ناپذیر است .

بالارفتن سن در شیوه ی سخن گفتن ما تغییراتی ایجاد می کند. صدای ما و الگوهای گفتاری ما از کودکی تا پیری تغییر می کند. اگرچه اکثر تغییرات در کودکی و بلوغ اتفاق می افتد، تغییرات وابسته به سن را می توان از دوران بزرگسالی تا پیری نیز مشاهده کرد. درنتیجه، سن ما در گفتار ما منعکس می شود. بدین ترتیب سن گوینده با به کارگیری روش های متعددی، عمدتا تحلیل صوتی و آزمایشات ادراکی، مورد مطالعه قرار می گیرد.

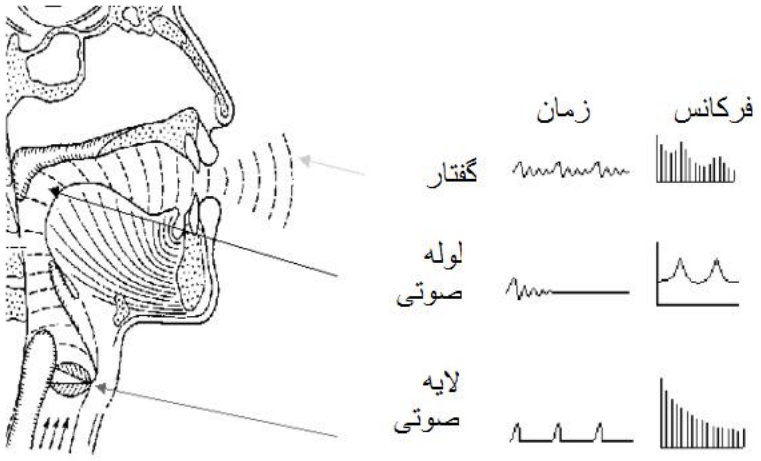
تغییرات وابسته به سن در گفتار بزرگسالان به طور گسترده از ۱۹۶۰ مورد مطالعه قرار گرفته است. اکثر مطالعات جنبه های صوتی و ادراکی را در نظر گرفته اند، اگرچه بعضی روش های تکنولوژی گفتار نیز دنبال شده است. بدلیل پیچیدگی فرایند افزایش سن، در بیشتر تحقیقات نیاز است تا تغییرات وابسته به سن در گفتار به خوبی درک شود. [۱]

با گذشت سن مقداری تغییرات فیزیکی و هورمونی در مرد ها و زن ها رخ می دهد که می تواند بر صدای آنها تاثیر گذارد. برای مثال یک موجی ازهورمون های فعال سازی در زمان بلوغ بوجود می آید که بر انتقال صدا از کودکی به یک فرد بالغ تاثیر می گذارد .

در زمان بلوغ، بر اثرافزایش هورمون تستوسترون، صدای یک مرد با تقریب ۱ octave در pitch ، کاهش می یابد. در خانم ها نیز در زمان بلوغ بر اثر هورمون های استروژن و پروژسترون، صدای آنها کاهش می یابد ولی با درجه ی کمتر تقریبا یک چهارم octave در pitch .

با بالا رفتن سن در مردها، مقدار هورمون تستوسترون کاهش می یابد و میانگین حجم صدا و تغییرپذیری در حجم صدای آنها افزایش می یابد. با بالا رفتن سن در خانم ها، کاهش فعالیت هورمون های جنسی، مخصوصا در دوران یائسگی، بر لایه های صوتی خانمها و عملکرد حنجره تاثیر می گذارد. صدای خانم ها عمیق تر می شود، فرکانس آوائیِ حداکثر کاهش می یابد و بازه ی صوتی گسترده می شود که باعث ادا شدن pitch های پایین تر می شود. تغییرپذیری pitch و jitter بر اثر مرور سن افزایش می یابد. [۲]

۱-۱-۱ مکانیزم تولید گفتار با گذشت سن :



شکل ۱-۱ : تولید گفتار صدا . فشار هوای شش ها تولید ارتعاش لایه های صوتی می کند ، که یک سیگنال منبع صدا ی ضربه ای شبه – متناوب را نتیجه می دهد . سیگنال منبع صدا باعث تحریک لوله ی صوتی ، که مانند یک بدنه ی رزونانسی عمل می کند که فرکانس های معینی را تقویت و یا تضعیف می کند و باعث تولید گفتار می شود .[۳]

مکانیزم تولید صدا در انسان را می توان به ۳ بخش تقسیم کرد: شش ها، لایه های صوتی و لوله ی صوتی. در شکل ۱-۱ مکانیزم صوتی انسان نشان داده شده است. فشار هوای شش ها باعث جریان یافتن هوا از بین حنجره، که فضای بین لایه های صوتی است، می شود . لایه های صوتی دو توده از گوشت، رباط و ماهیچه هستند که بین جلو و عقب حنجره کشیده شده و بسته به رانش و ربایش لایه های صوتی، در حالت های مختلف ارتعاشی (صوت های صدا دار) قرار می گیرند و یا اینکه اصلا ارتعاشی صورت نمی گیرد (صوت های بی صدا).

برای صوت های صدادار، لایه های صوتی به صورت شبه تناوبی باز و بسته می شوند و در نتیجه جریان هوای حنجره ای را به جریان ضربه هایی تبدیل می کند که سیگنال منبع صدا نامیده می شود. سپس سیگنال منبع صدا از لوله ی صوتی، که از حنجره آغاز و به لب ها ختم می شود، عبور می کند. لوله ی صوتی مانند یک بدنه با رزونانس ها (فرکانس های فرمنت) و ضد رزونانس ها (صفر ها) عمل می کند. لوله ی صوتی به عنوان فیلتر صوتی که طیف صوت را شکل می دهد، انجام وظیفه می کند. صداهای متنوع صوت با تنظیم شکل لوله ی صوتی و هم چنین سیگنال منبع صدا، تولید می شوند.

گفتار براثر ارتعاش تارهای صوتی هنگامی که نفس از شش ها بیرون داده می شوند، تولید می شود. تنوع صداهای گفتار عمدتا به دلیل تنوع در جرم و طول تارهای صوتی و بندهای موجود در لوله ی صوتی می باشد.

مردها دارای تارهای صوتی به طول ۱۷-۲۴ میلی متر و متوسط فرکانس اصلی 125 Hzهستند، درحالی که خانم ها دارای تارهای صوتی به طول ۱۳-۱۷ میلی متر و متوسط فرکانس اصلی برابر 200 Hz می باشند. [۳]

۱-۱-۲ سیستم تنفسی:

سیستم تنفسی بعد از بلوغ به اندازه ی کامل خود می رسد ولی از بزرگسالی تا پیری، باز هم در حال تغییر می باشد. این تغییرات شامل کاهش توانایی ششها(عمدتا به دلیل از دست دادن قابلیت ارتجاعی در بافت شش ها)، سفت شدن قفسه ی سینه و ضعیف شدن ماهیچه های تنفسی، می باشد.

۱-۱-۳حنجره

تغییرات وابسته به سن زیادی در حنجره پس از اینکه به اندازه ی کامل خود در سن بلوغ رسید، روی می دهد که عمدتا بر فرکانس اصلی و کیفیت صدا تاثیر می گذارد. استخوانی شدن غضروف ها بعدا اتفاق می افتد که در خانم ها در دهه ی چهارم زندگی و در مردها در سومین دهه ی زندگی شان می باشد. در حالی که فرایند آهکی شدن ممکن است بعد از استخوانی شدن روی دهد (هم در مرد ها و هم در خانم ها) تضعیف ماهیچه های حنجره نیز اتفاق می افتد.

تغییرات ساختار پیچیده ی لایه های صوتی با افزایش سن، قابل توجه می باشد. گذشته از انحطاط و تضعیف کلی، لایه های صوتی در مردها کوتاه می شوند (بعد از سن ۷۰ سالگی). هم چنین، در خانم ها لایه ی محافظ خارجی نازک بر بافت لایه ها، بر اثر مرور زمان به صورت تصاعدی ضخیم می شوند، مخصوصا بعد از ۷۰ سالگی، در حالی که در مردها تا ۷۰ سالگی این فرایند ضخیم شدن ادامه دارد و سپس مجددا نازک می شود.

غدد مخاطی ترشحات خود را کاهش می دهند، که باعث می شود لایه های صوتی کمتر مرطوب باشند، به ویژه در مردها. بعضی از اعصاب حنجره ای از بین می روند و هم چنین تغییراتی در خون رسانی به ماهیچه های حنجره صورت میگیرد.

۱-۱-۴ سیستم فرا حنجره

اسکلت جمجمه به طور پیوسته حدود ۳-۵٪ از جوانی تا پیری رشد میکند. تضعیف ماهیچه ای در ماهیچه های صورت، جویدنی و حلقی روی می دهد. افتادگی حنجره در گردن، طول لوله ی صوتی را افزایش می دهد.

تغییرات مخرب گسترده ای در مفصل گیجگاه روی می دهد که شامل کاهش تدریجی در اندازه و کاهش منابع خونی می باشد. در حفره ی دهانی، لا یه های مخاطی نازک تر می شوند و قابلیت ارتجاعی خود را از دست می دهند که اکثرا بعد از ۷۰ سالگی روی می دهد، و سطح مخاطی خشن تر می شود.

تغییرات گلو (حلق) و کام (سقف دهان) شامل نازک شدن مخاط ها، تضعیف ماهیچه ها و کاهش احساس می باشد. سطح زبان نازک تر و شکاف دار می شود در حالی که ماهیچه های زبان ضعیف می شود و نفوذ چربی مانندی ایجاد می شود، که از دهه های دوم و سوم زندگی آغاز می شود. [۱]

۱-۱-۵ بالارفتن سن در مردها و زن ها

یک تفاوت آشکار این است که تغییرات ظاهری در سن بلوغ در مرد ها روی می دهد؛ دیگر اینکه در زن ها این تغییرات در دوران یائسگی پدیدار می شود.

به طور کلی، با درنظر گرفتن (۱) اندازه ی تغییر در ساختار حنجره (۲) کنترل ظریف حرکات عضله ی ربایشی و رانشی حنجره ای (۳) حرکات زبان (۴) نرخ گفتار، تغییرات وابسته به سن در بزرگسالان، معمولا در مردها بیشتر از زن ها است.

پوسته های مخاطی حنجره، در خانم ها نسبت به مردها حساس تر است و در نتیجه خانم ها نسبت به تغییرات وابسته به سن آسیب پذیر ترند.

از طرف دیگر، مردها و خانم ها در تنفس گفتاری دارای تغییرات وابسته به سن یکسانی هستند.[۱]

۱-۲ : تحلیل صوتی سن

تعداد زیادی از تحقیقات ویژگی های صوتی متعددی از سیگنال گفتار که وابسته به سن گوینده است را بیان می کنند مانند: F0 و ثبات F0 ، رزونانس، کیفیت صدا، طول مدت سیگنال ، سطح فشار صدا (sound pressure level (SPL ، توزیع انرژی طیفی.

در تخمین سن، صحت عملکرد سیستم اولا به میزان دقت مورد نظر بستگی دارد، سپس به طول مدت و نوع نمونه های گفتار (مصوت بلند، گفتار متنی) بستگی دارد. هر چه اطلاعات صوتی موجود در نمونه ی گفتار کمتر باشد، کار مشکل تر می شود.

مطالعات قبلی نشان داده اند که با افزایش سن در خانم ها، F0 کاهش می یابد.

تحقیقات قابل مقایسه ای درباره ی گوینده های مرد نیز صورت گرفته است که قطعی نیست :

بعضی مواقع گزارش شده است که F0 بدون تغییر می ماند، در بعضی موارد کاهش می یابد و در بعضی مطالعات افزایش F0 گزارش شده است. این الگوی متفاوت در نتایج، به دلیل این است که در بازه ی سنی ۳۰-۵۰ سال در مردها F0 کاهش می یابد و سپس در سنین بالا افزایش می یابد.

کاهش F0 در خانم ها و افزایش آن در مردها در اواخر عمر، به دلیل تاثیر تغییرات هورمونی بر ارتعاش لایه ی صوتی می باشد. در حالی که کاهش F0 در هر دو جنسیت به دلیل افزایش ضخامت و جرم لایه ی صوتی نیز تعبیر شده است.

یکی از دلایلی که مطالعات درباره ی تاثیر سن بر پارامترهای صوتی، نتایج متناقض و ناسازگاری ارائه می دهند، این است که آنها معمولا بر گروه های سنی متفاوتی تحقیق انجام می دهند. در نتیجه مورفولوژی های متفاوت لوله ی صوتی که تاثیر زیادی بر F0 و فرمنت ها دارد، ممکن است تغییرات وابسته به سن را مبهم نشان دهد.

به طور مشابه، تفاوت های اجتماعی – واج شناسی بین گوینده های جوان تر و مسن در یک جامعه ی یکسان ممکن است تفاوت های فیزیولوژیکی اساسی بین آنها نشان دهد.

مطالعات طولانی مدت، کاهش F0 را بر اثر افزایش سن نشان داده اند و هم چنین کاهش میانگین فرمنت ها نیز مشخص شده است (ولی تغییرات هر فرمنت به طور جداگانه را نمی توان تعیین کرد.)[۴]

در یک تحقیق، برای یافتن چگونگی تغییر F0 و فرکانس های فرمنت با تغییر سن، از ۱۰ مصوت واحد برای ۴۹۰ نفر در بازه ی سنی ۵-۵۰ سال استفاده شده است. این مطالعه نشان می دهد که کودکان دارایF0 و فرکانس فرمنت های بالاتری هستند. و نسبت به بزرگسالان تغییرپذیری طیفی و زمانی بیشتری دارند. این یافته ها به دلیل تفاوت های آناتومی لوله ی صوتی و تفاوت های احتمالی در توانایی شمرده سخن گفتن می باشد.[۵]

مطالعه ای بر گفتار به زبان سوئدی نشان می دهد که پارامترهای عروضی، فرکانس اصلی و طول مدت برای تشخیص سن به تنهایی کافی نیستند. افزایش فضای پارامتر ها با افزودن پارامترهای کیفیت صدا، توزیع انرژی طیفی، نسبت هارمونیک ها به نویز و دقت شمرده سخن گفتن نتایج شناسایی را بهبود می بخشد.[۶]

نتایج نشان می دهند که ویژگی های صوتی طیف، اطلاعاتی درباره ی سن گوینده دارد. مطالعات قبلی که بر گفتار افراد مسن صورت گرفته است نشان می دهد که طیف توان در باندهای فرکانسی بالا برای گفتار افراد مسن در مقایسه با افراد غیر مسن، کاهش می یابد.[۷]

Linville تعدادی از ویژگی های صوتی که معمولا شنونده ها برای گوینده های مسن در نظر می گیرند را توضیح می دهد. مانند بهم آمیختگی در pitch برای صدای مردها و خانم های بزرگسال، افزایش درشتی، کشیدگی، لرزش صدا و نفس نفس زدن، کاهش بلندی صدا، سرعت پایین صحبت کردن و مکث های طولانی تر .[۸]

بعلاوه، .Rastatter et al (1997) نشان داده است که در گوینده های مسن، تمرکز بیشتری در فضای مصوت ها وجود دارد، اگرچه این تاثیر در مردها بیش از خانم ها مشهود است.

بعضی از مطالعات که به مقایسه ی گوینده ها در گروه های سنی مختلف پرداخته اند ،تاثیر سن را بر فرمنت ها نشان داده اند و سازگارترین تاثیر کاهش F1 براثر افزایش سن بوده است. کاهش فرمنت ها بدلیل طولانی تر شدن لوله ی صوتی می باشد که بر اثر افتادگی حنجره و افتادگی رشته های نای و نایژه ها و شش ها، رشد اسکلت صورت ایجاد می شود.

به هر حال، نتایج متناقضی برای مکان نسبی حنجره نسبت به مهره های گردن بدست آمده است: در حالی که Zemlin(1998) یک نزول عمودی حنجره در طی طول عمر انسان را که به طور میانگین از مهره ی ۶ ام در سن ۲۰ سالگی به مهره ی ۷ ام در سن ۸۰ سالگی می رسد را اعلام می کنند، Flu ¨gel و Rohen (1991) نتوانستند افتادگی حنجره را در بزرگ سالی بیابند.

ودر حالی که Xue and Hao's (2003) با استفاده از تکنیک انعکاس صوتی (acoustic reflection) هیچ تفاوتی در طول لوله ی صوتی در افراد جوان (۱۸-۳۰ سال) و افراد مسن (۶۲- ۷۹ سال) دریافت نکرده است، افزایشی را هم در حجم لوله ی صوتی و هم در طول و حجم حفره ی دهانی افراد مسن گزارش کرده اند. [۴]

۱- ۳ درک انسان از سن

بعضی نشانه های ادراکی عبارتند از Pitch ، سرعت صحبت کردن، بلندی صدا، کیفیت صوت.

بعلاوه فاکتورهای دیگری نیز بر درک سن تاثیر می گذارند:

۱- گوینده(جنسیت، شرایط فیزیولوژیکی و زبان گفتاری)

۲- شنونده (سن، فرهنگ و انگیزه)

۳- نمونه ی گفتار (نوع گفتار – گفتار خواندنی و یا گفتار فوری و طول گفتار)

۴- هدف (دسته بندی گوینده ها به دو گروه سنی و یا بیشتر و یا تخمین دقیقی از سن)

اگرچه اکثر مطالعات pitch و سرعت صحبت کردن را مهم ترین نشانه ی ادراکی برای سن گوینده معرفی می کنند، بعضی دیگر کیفیت طیفی را نیز با اهمیت می دانند.

تخمین سن افراد نسبت به تشخیص جنس آنها مشکل تر است زیرا در جنسیت با دو کلاس جدا از هم روبرو هستیم در حالی که در تخمین سن با تغییرات غیر خطی صدا در طول عمر مواجه ایم؛ و تعریف دقیقی از یک مرز بین دو گروه سنی مجاور مشکل می باشد.

مطالعات قبلی اثبات می کنند که جنس و سن یک گوینده ی ناشناس را می توان تنها با گوش کردن به صدای ضبط شده ی او، حتی با کیفیت پایین، مثلا از روی خط تلفن انالوگ، تخمین زد.

در طول افزایش سن، تغییرات صدا پیوسته است، درنتیجه تخمین دقیق سن حتی با شنونده های انسان، امکان پذیر نیست. مطالعات نشان می دهند که دقت تخمین سن فاعلی (subjective) که توسط افراد صورت می گیرد، به فاکتور های متعددی بستگی دارد. استفاده از جملات بلند به جای کلمات جدا ازهم، دقت تخمین را افزایش می دهد.[۱]

۱-۴ روش های تکنولوژی گفتار

روش های شناسایی سن اتوماتیک اکثرا از ویژگی های cepstral و آشفتگی (perturbation) و ترکیب آنها با تکنیک های آموزش ماشین مانند HMM , ANN استفاده می کنند.

چرا ترکیب تکنولوژی گفتار و آوا شناسی؟ با ایجاد پلی بین قوانین آواشناسی و تکنولوژی گفتار، جنبه های متعددی در زمینه ی سن گوینده، چه جنبه های آواشناسی و چه تکنولوژی گفتار، مورد بررسی قرار می گیرد؛ مانند درک انسان، قطعه بندی اتوماتیک و استخراج ویژگی از یک مجموعه گفتار بزرگ، شناسایی اتوماتیک مقادیر ویژه ی گوینده، و ترکیب گفتار (ساختن گفتار).

سن گوینده از جنبه ی تئوری اطلاعات نیز مورد برسی قرار می گیرد. اطلاعاتی که سیگنال گفتار حمل می کند (با توجه به نظر Fujisaki -2004) در جدول ۱-۱ آورده شده است. [۱]

گسسته / پیوسته	مثال	دسته
گسسته (نمادین) که با گوینده کنترل می شود	لغوی (کلمه ، لهجه ، ..)	linguistic
	ساختاری (ساختار عبارت ، ...)	
	Paramatic	
گسسته و پیوسته که می تواند توسط گوینده کنترل شود	قصدی (نصیحت ، تشویق ، ....)	paralinguistic
	رفتاری(ادب ، ...)	
	سبک ادبی- سلیسی و روانی(تند ، ارام ،....)	
گسسته و/یا پیوسته که معمولا نمی توانند کنترل شوند	فیزیکی(سن ، جنس ، ...)	non- linguistic
	احساسی (لذت ، تاسف ، ...)	
	طرز فکر ویژه	

جدول ۱-۱

۱-۵ ساختار پایان نامه :

بعد از ذکر مقدمه ی پیشین و نگاهی اجمالی به ضرورت تخمین سن و هم چنین چگونگی تغییر گفتار انسان با تغییر سن او، در این تحقیق سعی بر معرفی روشی برای تخمین سن افراد با استفاده از سیگنال صوت هستیم .

در فصل دوم، اشاره ای به مطالعات و تحقیقات انجام شده در زمینه ی تخمین سن اتوماتیک با بکارگیری صوت شده است تا با انواع روش ها آشنایی مختصری بدست آید.

در فصل سوم، بدلیل اهمیت بخش استخراج ویژگی در یک سیستم تخمین سن با استفاده از گفتار، به معرفی روش های استخراج ویژگی متداول در پردازش صوت می پردازیم.

و درنهایت در فصل چهارم، بعد از معرفی روش پیشنهادی برای تخمین سن با استفاده از گفتار و بررسی نتایج بدست آمده، به مقایسه ی نتایج حاصل از روش های سابق با نتیجه ی روش پیشنهادی می پردازیم.

فصل دوم :

تحقیقات صورت گرفته در زمینه ی

تخمین سن اتوماتیک توسط سیگنال گفتار

در این فصل به تحقیقاتی در زمینه ی تخمین سن افراد با استفاده از سیگنال صوت و گفتار آنها، به طور خلاصه اشاره ای می کنیم تا هم با روش های قبلی آشنا شویم و هم اینکه نتایجی که تاکنون در این زمینه بدست آمده است را بررسی نماییم.

در طی تحقیقی [۹] به بررسی چگونگی دریافت اطلاعاتی نظیر جنس و سن و لهجه و احساسات گوینده با استفاده از سیگنال صوت او پرداخته شده است. در این مطالعه با اطلاعاتی که توسط سیستم گسترش یافته ی مشتری – محور، سیستم AT&T 's "How may I help you" (HMIHY) جمع آوری شده است، کار می شود. و از پایگاه داده ی DARPA switch board برای تشخیص سن استفاده شده است.

افراد به ۳ گروه سنی اصلی تقسیم شده اند :

جوان (کمتر از ۲۵ سال) بزرگسال (۲۵- ۵۰ سال) مسن (بالای ۵۰ سال)

مجمو عه ی داده های آموزش شامل ۲۰۰ تماس تلفنی است .(۶۲۷ عبارت)

از ویژگی های صوتی و عروضی که بیشتر به اینکه گوینده چگونه صحبت می کند، تمرکز می کند تا اینکه آنها دقیقا چه می گویند، برای تشخیص خصیصه های گوینده استفاده می شود.

از رده بند هایی بر پایه ی speech/ non-speech HMM که بر ویژگی های cepstral اعمال می شود، برای تشخیص مشخصات گوینده ها استفاده می شود.

قسمت speech هر عبارت با HMM های یک حالته مدل شده است. قسمت non-speech عبارات نیز با یک مدل نویز یک حالته ی کلی، مدل شده اند. همه ی قسمت های عبارات (speech , non- speech) با مدل های مخلوطی گوسی مدل شده اند (با تعداد اجزای مخلوطی متغیر) .

ویژگی های MFCC استاندارد که معمولا در سیستم های شناسایی گفتار اتوماتیک (ASR) به کار برده می شوند ، بدون هیچ نرمالیزه کردنی، به عنوان ورودی در نظر گرفته می شوند.

درقسمت دوم آزمایش ها، اطلاعات pitch هم به بردار ویژگی MFCC اضافه می شود. که نتایج بدست آمده از این آزمایشات در جدول ۲-۱ آمده است .

ویژگی ها	Cepstral	Cepstral + pitch
دقت رده بندی در این آزمایشات (تخمین سن)	68.4	70.2

جدول ۲-۱

در سال ۲۰۰۹ مقایسه ای بین رده بندی سن و جنس یک فرد با استفاده از سیگنال گفتار در مرجع [۱۰] انجام گرفته است. در این تحقیق از پایگاه داده ی (Danish Emotional Speech)DES و ELSDSR(English Language Speech Database for Speaker Recognition) استفاده شده است.

* پایگاه داده ی DES

شامل عبارت هایی است که توسط ۲ هنرمند حرفه ای مرد و ۲ هنرمند حرفه ای زن که در ۵ احساس متفاوت (خشم، شادی، عادی، ناراحت، تعجب) بیان شده اند، می باشد. یک هنرمند مرد و یک هنرمند زن، هر دو ۵۲ سال دارند. هنرمند مرد دیگر ۳۸ ساله و هنرمند زن دیگر ۳۴ ساله است. سن آستانه ی مورد نظر ۴۵ در نظر گرفته شده است. در کل ۱۱۶۰ عبارت شامل ۳۶۰ عبارت از کلمه ها و جمله ها و ۸۰۰ عبارت از پاراگراف ها وجود دارد که ۵۸۰ عبارت توسط مردها و ۵۸۰ عبارت توسط زن ها بیان شده است.

* پایگاه داده ی ELSDSR

عبارات موجود در این پایگاه داده توسط یک فرد ایسلندی، یک فرد کانادایی و ۲۱ نفر دانمارکی ادا شده اند. ۱۰ نفر از گوینده ها مرد و ۱۳ نفر گوینده ها زن هستند. ۶ زن و ۳ مرد بالای ۳۳ سال سن دارند. برای داده های آموزش، ۱۶۱ عبارت ضبط شده(۷ جمله از کل ۹ جمله ای که ۲۳ نفر موجود در پایگاه داده ادا کرده اند.) و برای داده های آزمایش۴۶ عبارت(۲ جمله از کل ۹ جمله ای که۲۳ نفر موجود در پایگاه داده ادا کرده اند) در نظر گرفته شده است.

- استفاده از رده بند Bayes با به کارگیری SFFS(Sequential floating forward selection) برای انتخاب ویژگی
- استفاده از رده بند های GMM , k-NN , SVM , PNN ، برای تعیین بهترین رده بند برای دسته بندی گروه سنی و جنسیت افراد از روی سیگنال صوت آنها.
- تقسیم بندی افراد تنها دو گروه سنی (تصمیم گیری دودویی) سن های بیشتر و یا کمتر از یک سن تعیین شده، و یا سن آستانه
- ویژگی های آماری که در این تحقیق به کار برده شده شامل :

۱- ویژگی های فرمنت (۱-۱۶) ۲- ویژگی های Pitch (۱۷- ۵۱)

۳- ویژگی های انرژی (۵۲-۸۵) ۴- ویژگی های طیفی (فرکانسی)(۸۶-۱۱۳)

که اعداد داخل پرانتز نشان دهنده ی شماره ی ویژگی در بردار ویژگی نهایی می باشد.

انواع ویژگی های به کاربرده شده به صورت جزئی در جدول ۲-۲ نام برده شده اند.

Formant features	
1-4	Mean value of first,second, third and fourth formant
5-8	Maximum value of first,second, third and fourth formant
9-12	Minimum value of first,second, third and fourth formant
13-16	Variance of first,second, third and fourth formant
Pitch features	
17-21	Maximum, minimum, mean, median, inter-quartile range of pitch values
22	Pitch existence in the utterance expressed in percentage (0- 100 %)
23-26	Maximum, mean, median, interquartile range of durations for the plateaux at minima
27-29	Mean, median, interquartile range of pitch values for the plateaux at minima
30-34	Maximum, mean, median, interquartile range, upper limit (90%) of durations for the plateaux at maxima
35-37	Mean, median, interquartile range of the pitch values within the plateaux at maxima
38-41	Maximum, mean, median, interquartile range of durations of the rising slopes of pitch contours
42-44	Mean, median, interquartile range of the pitch values within the rising slopes of pitch contours
45-48	Maximum, mean, median, interquartile range of durations of the falling slopes of pitch contours
49-51	Mean, median, interquartile range of the pitch values within the falling slopes of pitch contours
Energy features	
52-56	Maximum, mean, median, interquartile range of energy value
57-60	Maximum, mean, median, interquartile range of durations for the plateaux at minima

61-63	mean, median, interquartile range of energy value for the plateaux at minima
64-68	Maximum, mean, median, interquartile range , upper limit(90%) of durations for the plateaux at maxima
69-71	mean, median, interquartile range of energy value for the plateaux at maxima
72-75	Maximum, mean, median, interquartile range of durations of the rising slopes of energy contours
76-78	Mean, median, interquartile range of the energy values within the rising slopes of energy contours
79-82	Maximum, mean, median, interquartile range of durations of the falling slopes of energy contours
83-85	Mean, median, interquartile range of the energy values within the falling slopes of energy contours
Spectral features	
86- 93	Energy below 250, 600, 1000, 1500, 2100, 2800, 3500, 3950 Hz
94-100	Energy in the frequency bands 250-600, 600-1000, 1000-1500,1500-2100 , 2100-2800,2800-3500,3500-3950 Hz
101-106	Energy in the frequency bands 250-1000, 600-1500,1000-2100,1500-2800,2100-3500,2800-3950 Hz
107- 111	Energy in the frequency bands 250-1500, 600-2100,1000-2800,1500-3500,2100-3950 Hz
112-113	Energy ratio between the frequency bands (3950 – 2100) and (2100-0) and between the frequency bands (2100- 1000) and (1000-0) Hz

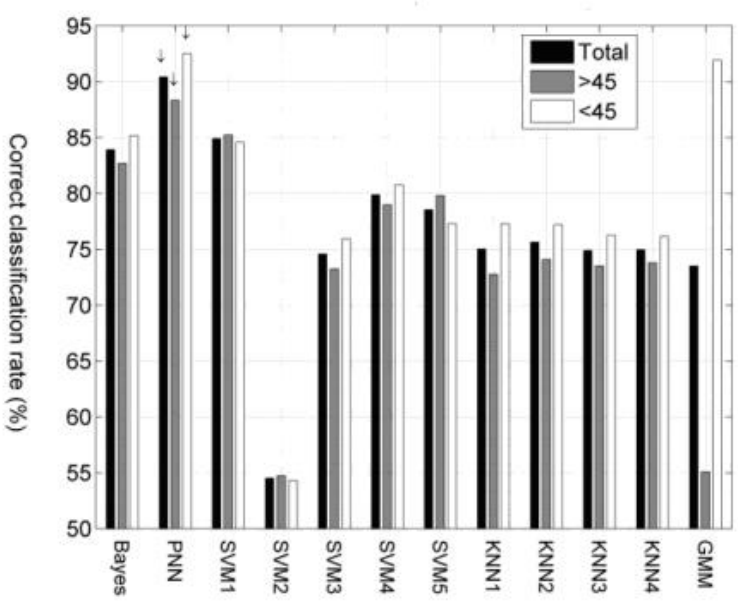
جدول ۲-۲ : انواع ویژگی های استفاده شده در تحقیق [۱۰]

همه ی این ویژگی ها را نمی توان ازهمه ی عبارت ها استخراج کرد. مثلا بعضی عبارات به دلیل اینکه بی صدا (unvoived) هستند، کلا pitch ندارند. به این ترتیب یک سری ویژگی ها حذف می شوند که شماره ی آنها طبق جدول بالاعبارتند از : 8,23-29, 33,34,41,48,57-63,67,75,82,105

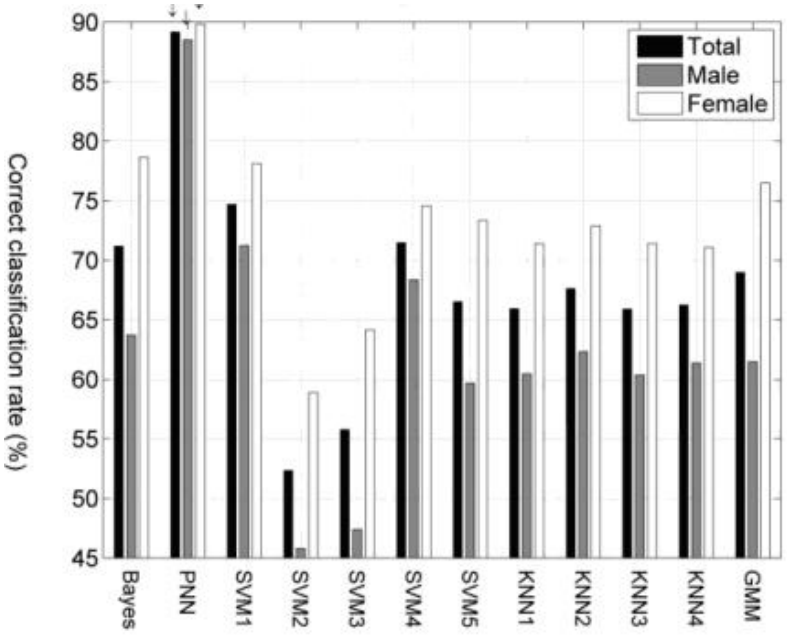
و درنهایت تعداد ویژگی برابر ۹۰ انتخاب می شوند.

۸۰ ٪ داده ها به عنوان داده های آموزش و ۲۰ ٪ داده های آزمایش انتخاب می شوند.

نتایج بدست آمده در شکل های ۲-۱ و ۲-۲ و جدول های ۲-۵ و ۲-۶ آورده شده است.



شکل۲-۱ : نتیجه ی رده بندی سن برای پایگاه داده ی DES از مرجع [۱۰]



شکل۲-۲ : نتیجه ی رده بندی سن برای پایگاه داده ی ELSDSRاز مرجع [۱۰]

که رده بند SVM را با ۵ نوع تابع kernel متفاوت به کار رفته است:

	نوع تابع kernel
SVM1	Guassian RBF(radial basis function)
SVM2	Multilayer perceptron
SVM3	Quadratic
SVM4	Linear
SVM5	cubic polynomial

جدول ۲-۳

و رده بند K-NN با ۴ تابع فاصله ی (Distance functions) متفاوت به کار رفته است :

	تابع فاصله
KNN1	اقلیدسی -Euclidean
KNN2	مجموع اندازه ی اختلافات -cityblock(sum of absolute differences)
KNN3	Cosine-based (one minus the cosine of the included angle between patterns) بر پایه ی کسینوسی(یک منهای کسینوس زاویه ی بین نمونه ها)
KNN4	correlation-based (one minus the sample correlation between patterns) بر اساس همبستگی (یک منهای همبستگی بین نمونه ها)

جدول ۲- ۴

PNN	Correctly classified response (%)	
Ground Truth ↓	>45	<45
>45	**88.34**	11.66
<45	7.49	**92.51**
Total rate (%)	90.43±1.84	
Bayes-SFFS	Correctly classified response (%)	
Ground Truth ↓	>45	<45
>45	**82.68**	17.32
<45	14.84	**85.16**
Total rate (%)	83.92±1.48	
SVM1	Correctly classified response (%)	
Ground Truth ↓	>45	<45
>45	**85.23**	14.77
<45	15.43	**84.57**
Total rate (%)	84.90±2.05	
SVM4	Correctly classified response (%)	
Ground Truth ↓	>45	<45
>45	**78.98**	21.02
<45	19.24	**80.76**
Total rate (%)	79.87±2.32	

جدول ۲-۵ : ماتریس ترکیبی بر پایگاه داده ی DES برای ۴ تا از بهترین روش ها برای رده بندی سن که از ۲۰ ٪ عبارات برای آزمایش استفاده شده است. که تعداد داده هایی که به درستی دسته بندی شده اند را بر حسب درصد بیان می کند.

PNN	Correctly classified response (%)	
Ground Truth ↓	>33	<33
>33	**88.51**	11.49
<33	10.17	**89.83**
Total rate (%)	89.17±0.93	
Bayes-SFFS	Correctly classified response (%)	
Ground Truth ↓	>33	<33
>33	**63.72**	36.28
<33	21.36	**78.64**
Total rate (%)	71.18±10.55	
SVM1	Correctly classified response (%)	
Ground Truth ↓	>33	<33
>33	**71.23**	28.77
<33	21.88	**78.12**
Total rate (%)	74.68±4.87	
SVM4	Correctly classified response (%)	
Ground Truth ↓	>33	<33
>33	**68.36**	31.64
<33	25.44	**74.56**
Total rate (%)	71.46±4.38	

جدول ۲-۶ : ماتریس ترکیبی بر پایگاه داده ی ELSDSR برای ۴ تا از بهترین روش ها برای رده بندی سن که از ۲۰ ٪ عبارات برای آزمایش استفاده شده است. که تعداد داده هایی که به درستی دسته بندی شده اند را بر حسب درصد بیان می کند.[۱۰]

تخمین سن کودکان پیش دبستانی و ابتدایی نیز در [۱۱] مورد مطالعه قرار گرفته است. در این تحقیق برای هر کودک یک GMM آموزش داده می شود.

به عنوان روش آموزش دهی، روش تطبیق MAP(Max A Posteriori) بکاربرده شده است. MAP ، مدل گوینده را از یک UBM(Universal Background model) استنتاج می کند.

میانگین های هر GMM استخراج میشود و تلفیق می گردند. که فرابردارهای GMM (GMM supervector) خوانده می شوند. این فرابردارها به عنوان ویژگی های فراتر(meta) برای رده بندی با SVM و یا SVR بکار برده می شود.

همه ی کودکان کلمات موجود در آزمایش PLAKSS که شامل ۹۹ کلمه به زبان آلمانی است، را می خوانند.

*پایگاه داده:

۱- یک پایگاه داده شامل ۳۸ کودک از یک پیش دبستانی با میانگین سنی 5.7 ± 0.7

۲- یک پایگاه داده شامل ۱۷۷ کودک از یک دبستان ابتدایی با میانگین سنی 8.5 ± 1.4

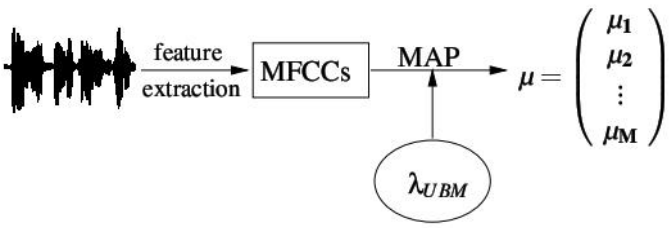
۳- یک پایگاه داده شامل ۱۲۸ کودک از یک دبستان ابتدایی با میانگین سنی 8.6 ± 1.1

که در نهایت ۲۱۲ کودک به عنوان داده های آموزش و ۱۰۰ کودک به عنوان داده های آزمایش با گروه های سنی کمتر از ۷ سال ، ۷ ساله، ۸ ساله، ۹و۱۰ ساله، بیشتر از ۱۰ سال، می باشند.

ابتدا بردارهای ویژگی MFCC از سیگنال صوت استخراج می شوند. این ویژگی ها برای آموزش GMM ها با M چگالی گوسی بکار برده می شوند. هر GMM برای یک کودک است.

میانگین هر GMM تلفیق می شود و به عنوان ویژگی های فراتر(meta) در SVM ها به کار برده می شود. در نتیجه هر کودک با یک فرابردار نمایش داده می شود.

ایده ی اصلی فرابردارهای GMM، مدل کردن هر گوینده با یک GMM متفاوت می باشد که نمایشی از آن در شکل ۲-۳ آورده شده است .



شکل ۲-۳: از مرجع [۱۱]

برای استخراج MFCC ، طول پنجره برابر 16ms و با زمانی شیفت 10ms برای هر پنجره در نظر گرفته می شود. بردار ویژگی نهایی شامل ۲۴ المان است (log energy,1-11 MFCC,first derivatives of MFCC) دو سری آزمایش اولیه صورت گرفته است:

۱- آزمایشاتی برای انتخاب بهترین پارامترها برای تولید مدل گوینده

۲- آزمایشاتی برای تعیین بهترین تابع kernel برای SVM/SVR و ترکیب فرابردارهای GMM .

در آزمایشات اولیه ی (۱)، تاثیر تعداد چگالی های گوسی، الگوریتم آموزش (EM , MAP) و تاثیر شکل های مختلف ماتریس های کوواریانس (کامل و یا قطری) بر نتایج تشخیص سن گوینده بررسی شده است.

بهترین نتایج با GMM هایی که با الگوریتم MAP تنظیم شده اند و دارای ۲۵۶ چگالی گوسی با ماتریس های کوواریانس کامل هستند، بدست آمده است.

در آزمایشات اولیه ی (۲) تعیین مناسب ترین تابع kernel برای SVM/ SVR مهم است. توابع kernel خطی ، نمایی و RBF مد نظر است.

با تغییر مر تبه ی چند جمله ای مربوط به تابع kernel نمایی و عرض تابع kernel شعاعی پایه برای RBF ، بهترین نتیجه با ساده ترین نوع kernel ، یعنی تابع خطی بدست آمده است.

نتایج کلاسه بندی در جدول۲-۷ نشان داده شده است (نتایج بر حسب در صد می باشند.):

	<7	7	8	9+10	>10
<7	**60**	33		7	
7	5	**55**		35	5
8		20	**6**	47	27
9+10			3	**97**	
>10					**100**

جدول ۲-۷ : ماتریس ترکیبی نسبی برای سیستم SVM .

محور y مربوط به کلاس واقعی و محور x مربوط به کلاس شناسایی شده

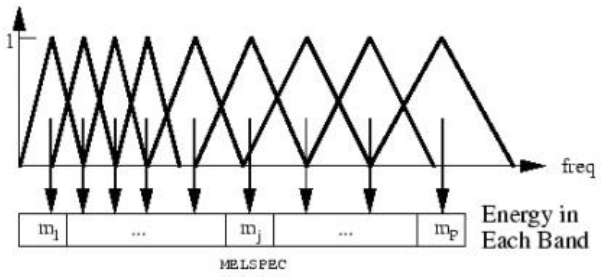
تحقیق دیگری در زمینه ی تخمین سن و جنس از سیگنال گفتار ضبط شده بر اساس GMM در [۱۲] مورد مطالعه قرار گرفته است.

در این تحقیق از پایگاه داده ی Czech Speech Dat (E) استفاده شده است. که شامل صحبت های تلفنی می باشد. ۱۰۵۲ نفر گوینده دارد . ۱۲ عبارت از هر گوینده مورد استفاده قرار گرفته است. مجموعه داده های آموزش ۸۱٪ کل داده ها می باشد که شامل گوینده هایی در بازه ی سنی ۹- ۷۹ ساله است (۱۰۲۰۷ عبارت) مجموعه داده های آزمایش ۱۹ ٪ کل داده ها ست که گوینده ها در بازه ی سنی ۱۲- ۷۵ سال هستند. (۲۳۹۷ عبارت)

از ویژگی های MFCC استفاده شده است (شامل ۳۹ پارامتر) :

c_1	...	c_{12}	E	Δc_1	...	Δc_{12}	ΔE	$\Delta^2 c_1$	...	$\Delta^2 c_{12}$	$\Delta^2 E$

تعداد کانال های بانک فیلتر برابر ۲۴ می باشد.



شکل ۲-۴ : بانک فیلتر در مقیاس Mel از مرجع [۱۲]

*مراحل تشخیص سن

- استخراج ویژگی های MFCC

- تخمین پارامترهای GMM با استفاده از ML(max liklihood)

- ابتدا داده ها در ۳ گروه سنی با بازه ی ۲۵ سال تقسیم می شوند(جوان ، میان سال ، مسن)

	Young	Middle Aged	Old
Range	9..30	31..55	56..79
Train.	4259	3333	969
Test	1125	984	276

جدول ۲-۸ : تعداد داده های آموزشی و آزمایش

- سپس هر فرد به طور دقیق تری در یکی از ۱۳ گروه سنی (با بازه ی ۵ سال) تقسیم می گردد.

	1	2	3	4	5	6	7	8	9	10	11	12	13
Range	9..15	16..20	21..25	26..30	31..35	36..40	41..45	46..50	51..55	56..60	61..65	66..70	71..79
Train.	84	999	2507	1113	838	599	720	1020	755	514	287	202	192
Test	48	217	624	240	252	96	336	204	72	144	84	24	24

جدول ۲-۹ : تعداد داده های آموزشی و آزمایشی برای ۱۳ گروه سنی

- ۱۳ GMM در تقسیم گروه های سنی با بازه ی ۵ سال، آموزش داده شده اند.

نتایج بدست آمده، در تقسیم بندی به ۳ گروه سنی، ۶۰.۱۳ ٪ از کل داده ها به درستی رده بندی شده اند.

	جوان	میان سال	مسن
صحت	56.62%	47.56%	28.26%

جدول ۲-۱۰: نتیجه ی آزمایشات برای تقسیم بندی افراد به ۳ گروه سنی

و در تقسیم بندی به ۱۳ گروه سنی، ۲۳.۸۵ ٪ از کل داده ها به درستی رده بندی شده اند.

در کل، متوسط اختلاف بین سن تخمینی و سن واقعی افراد ۱۱.۳۸ سال است. [۱۲]

تحقیق دیگری در باره ی تخمین سن افراد از سیگنال صوت با استفاده از پایگاه داده ی فارسی farsdat انجام شده است [۱۳] که از SVM برای رده بندی استفاده کرده اند. در زیر مختصری از این مقاله آمده است.

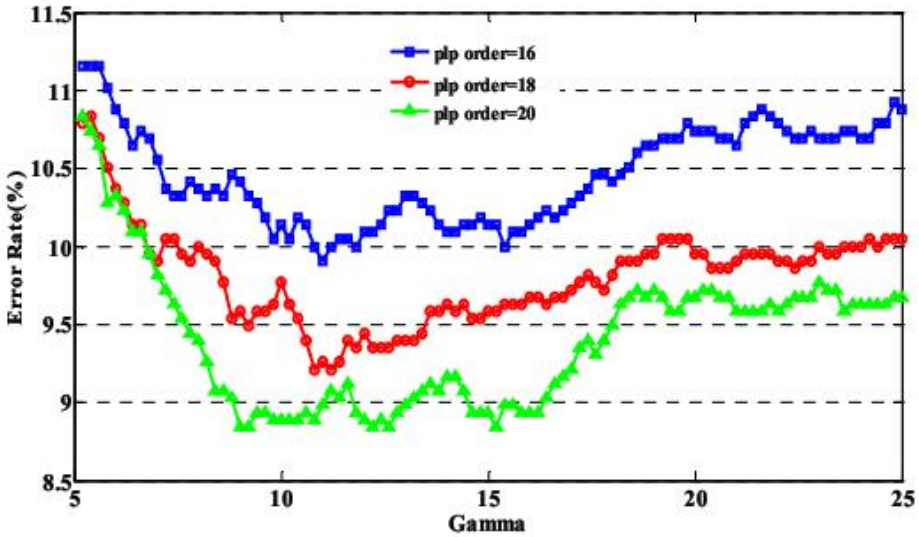
برای حفظ کلیت سیستم، نسبت تعداد مردها به زن ها و تعداد عبارات گفته شده توسط هر گوینده به طور اتفاقی انتخاب گردیده است. گوینده ها به ۶ گروه سنی تقسیم شده اند.

الف) ۱۵-۲۵ ساله : کلاس ۱ ب) ۲۶-۳۵ ساله : کلاس ۲ ج) ۳۶-۴۵ ساله : کلاس ۳

د) ۴۶-۵۵ ساله : کلاس ۴ ه) ۵۶-۶۵ ساله : کلاس ۵ د) ۶۶- ۷۳ ساله : کلاس ۶

ابتدا برای یکسان سازی، سطح دامنه ی سیگنال هر عبارت ضبط شده، برای تمامی نمونه ها، نرمالیزه شده است و سپس از هر نمونه ویژگی های PLP , MFCC استخراج شده است. از هر گروه سنی ۱۶۰ نمونه به طور اتفاقی انتخاب شده و ۶۲ ٪ از آنها به عنوان داده های آموزشی به کار رفته اند. از SVM برای رده بندی استفاده شده است. در روند آموزش و آزمایش رده بند از تابع kernel ، RBF(Gaussian radial basic function) استفاده شده است. و عملکرد سیستم با مقادیر مختلفی از پارامتر ¥ در تابع kernel مورد ارزیابی قرار گرفته است. هم چنین تاثیر تغییرات طول زمانی قاب ها در عملیات نمونه برداری، تعداد ضرایب MFCC و مرتبه یPLP ، بر بازدهی سیستم ارزیابی می شود.

بعد از نرمالیزه کردن نمونه ها، PLP به عنوان اولین ویژگی استخراج شده و پس از ارزیابی سیستم برای ¥ها و مرتبه های ۱۶ ، ۱۸ و ۲۰ برای PLP ، بهترین نتیجه را برای مرتبه ی ۲۰ برای PLP بدست آمده است. افزایش مرتبه ی PLP نرخ خطا را کاهش می دهد ولی نه به مقدار زیاد.



شکل ۲-۵ : نمودار نرخ خطا بر حسب پارامتر گاما در RBF kernel ، با استفاده از ویژگی PLP با ۳ مقدار متفاوت از مرتبه

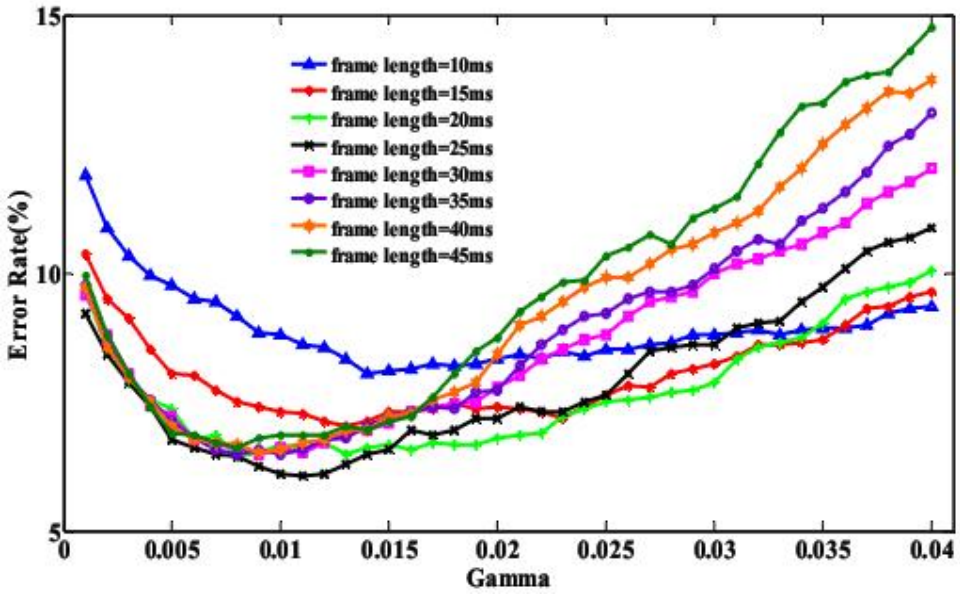
PLP	Gamma	Error Rate
16	11	9.91%
18	10.8	9.21%
20	9	8.84%

جدول ۲-۱۱: نتایج کمترین نرخ خطا برای مرتبه های مختلف PLP

همان طو ر که از جدول ۲-۱۱ مشاهده می شود، کمترین نرخ خطا با مرتبه ی ۲۰ در PLP و بازای ¥ ی بهینه، ۸.۸۴ ٪ است.

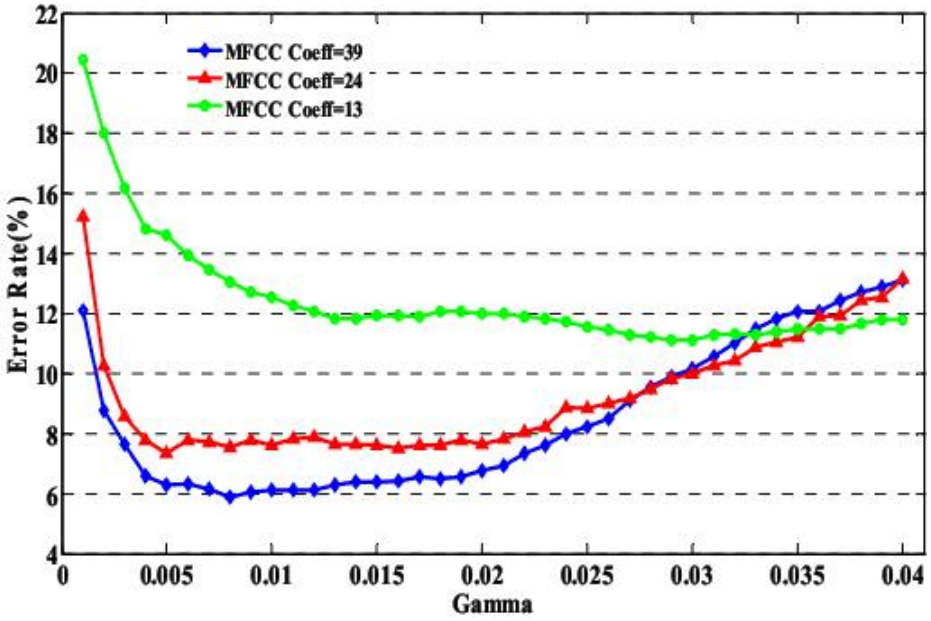
دومین ویژگی MFCC می باشد. نشان داده می شود که برای بدست آوردن یک عملکرد بهینه در سیستم های تخمین سن، انتخاب طول قاب مناسب در محاسبه ی ضرایب MFCC یک عامل مهم است.

در این مرحله عملکرد سیستم نسبت به تغییرات ¥ بسیار حساس بوده (در حد ۰.۰۰۱) از طرفی اگر یک طول قاب کوچک انتخاب شود، پاسخ قابل قبولی بدست نمی آید. همچنین اگر زمان طول قاب از 25ms افزایش یابد، تغییرات بسیار کمی در نرخ خطا داریم.



شکل ۲-۶ : نمودار نرخ خطا بر حسب پارامتر گاما در RBF kernel ، با به کارگیری مرتبه ی ثابت MFCC ، با طول قاب های زمانی مختلف

با انتخاب طول قاب برابر 25ms ، با تغییر تعداد ضرایب MFCC (۱۳و۲۴و۳۹) ویژگی، کمترین نرخ خطا (بهترین نتیجه) برای تعداد ویژگی برابر ۳۹ بدست می آید.



شکل ۲-۷ : نمودار نرخ خطا بر حسب پارامتر گاما در RBF kernel ، با به کارگیری مرتبه های مختلف MFCC ، با طول قاب زمانی برابر 25ms

در نهایت با انتخاب پارامترهای بهینه، کمترین نرخ خطای بدست آمده برابر ۵.۸۹ ٪ می باشد.

MFCC	frame length	Gamma	error rate
13	30ms	0.03	10.9
24	30ms	0.015	7.33
39	30ms	0.008	6.17
13	25ms	0.029	11.11
24	25ms	0.005	7.33
39	25ms	0.008	5.89

جدول ۲-۱۲ : نرخ خطا ی بدست آمده برای مقادیر مختلف گاما ، طول قاب و تعداد ضرایب MFCC

با بهره برداری از سیگنال گفتار می توان با تشخیص افراد مسن، راه را برای پاسخ به نیاز های ویژه ی آنان هموار کرد. این موضوع در [۱۴]مور بررسی قرار گرفته است که خلاصه ای از روش به کار برده شده در زیر آمده است.

انتخاب ویژگی ها ی صوتی jitter , shimmer و استخراج این ویژگی ها با استفاده از ابزار PRAAT که شامل الگوریتم های متعددی برا ی اندازه گیری ویژگی های jitter , shimmer در سیگنال می باشند. که ۵ مقدار برای jitter و ۳ مقدار برای shimmer محاسبه می کند.

Jitter : حداکثر مقدار انحراف از فرکانس پایه (F0)

مقادیر jitter به عنوان درصدی از طول دوره ی تناوب pitch ظاهر می شود. مقادیر بزرگ برای تغییرات jitter در صداهای آسیب دیده (مسن) دیده می شود. در صداهای نرمال و معمولی ،مقدار jitter از یک درصد دوره ی تناوب pitch هم کمتر است. در این تحقیق از ۵ الگوریتم متفاوت برای محاسبه ی jitter استفاده شده است. مانند JR(jitter ration) , PVI(period variability index) , RAP(relative average pertubation)

JR : تعیین تغییرپذیری دوره به دوره

PVI : محاسبه ی مقداری که به انحراف معیار یک دوره ی تناوب وابسته است.

RAP : مقایسه ی میانگین ۳ دوره با یک دوره تناوب داده شده

shimmer : حداکثر تغییرات در دامنه ی قله ها در دوره های تناوب pitch پشت سر هم، را نشان می دهد.

مقادیر بزرگ shimmer در صداهای آسیب دیده (مسن) دیده می شود. در صداهای معمولی و نرمال کمتر از 0.7 dB است.

تعدادی الگوریتم برای بازیابی چندین مقدار shimmer استفاده شده است که تفاوت آنها مانند تفاوت های بین الگوریتم های استفاده شده برای jitter می باشد. برای مثال ApQ(amplitude perturbation quatient) سعی بر کم رنگ کردن تغییرات طولانی دامنه دارد مانند آنچه RAP برای تغییرات فرکانس انجام می دهد. APQ از میانگین گیری ۱۱ نقطه ای (متوسط ۱۱ دوره) استفاده می کند.

دو پایگاه داده ی متفاوت استفاده شده است

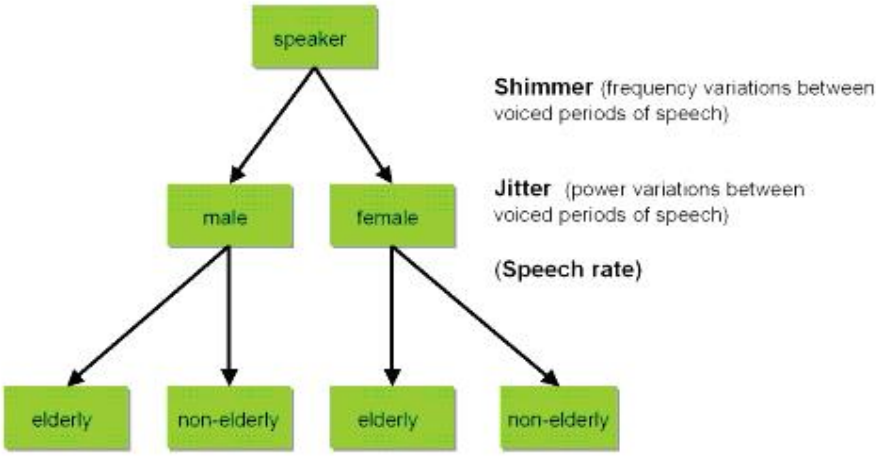
- SCANSOFT

شامل بیش از ۱۰۰۰۰ عبارت از ۳۴۷ گوینده با سن بیش از ۶۰ سال

- MSI

شامل حدود ۵۰۰۰ عبارت از ۴۶ گوینده با سن کمتر از ۶۰ سال

که در کل دو پایگاه داده ، ۱۶۲ نفر از گوینده ها زن و ۲۳۱ نفر مرد هستند.



شکل ۲-۸ : سلسله مراتب رده بندی از مرجع [۱۴]

از رده بند های زیر استفاده شده است:

۱- C4.5 decision tree induction (DT)

۲- Artificial neural network (ANN) ، با

نرخ آموزش learning rate = 0.15

مقدار حرکت momentum = 0.2

تعداد تکرار iteration = 500

۳- KNN(k=5 ,simple distance weighting)

۴- Naïve Bayes(NB)

۵- SVM(polynomial kernel with degree 4)

۶- BL(Baseline ,simple classifier)

نتایج بدست آمده بازای استفاده از رده بندهای متفاوت در جدول ۲-۱۳ نشان داده شده است.

	C4.5	ANN	kNN	NB	SVM	BL
gender	67.79	81.09	75.62	67.34	70.51	58.78
age	92.68	96.57	95.71	91.15	96.52	88.30

جدول ۲- ۱۳ : صحت پیش بینی سن و جنس (بر حسب درصد)

تعدادی از محققین دانشگاه های آلمان و فرانسه و آمریکا، ویژگی های صوتی توسعه یافته ای را برای تخمین سن افراد از سیگنال گفتار ضبط شده ارائه کرده اند.[۸] خلاصه ای از این روش به این صورت است که ویژگی های اصلی و پایه ، MFCC هستند که با تعدادی ویژگی های عروضی، pitch و فرکانس های فرمنت گسترش یافته اند. اضافه کردن این ویژگی ها به ویژگی های پایه ی MFCC باعث کاهش نسبی میانگین اندازه ی خطا به اندازه ی ۴-۲۰ ٪ می شود.

از طرف دیگر، نتایج قابل قبولی با میانگین اندازه ی خطای حدود ۱۲ سال، با استفاده از یک سیستم مستقل از جنس و تنها بر پایه ی MFCC بدست می آید.

این تحقیق بر آن است که یک رنج گسترده ای از انواع ویژگی های مختلف، نه تنها ویژگی ها cepstral کوتاه – مدت ، بلکه ویژگی های بلند – مدت عروضی نیز شامل هستند.

از پایگاه داده ی (university of Florida vocal aging database)UF-VAD استفاده شده است. که شامل گفتار های خواندنی (متن خوانی)توسط بزرگسالان می باشد (جوان- میان سال - مسن). برای هر گوینده در پایگاه داده، نه تنها سن واقعی بلکه سن حدسی برای آن ها نیز در دسترس است. این اطلاعات باعث می شود که محکی بین سرعت و صحت تشخیص سن توسط حدس انسان و رده بند های اتوماتیک بدست آید.

زبان پایگاه داده انگلیسی آمریکایی است .

دارای ۱۵۰ گوینده و۱۳۵۰ عبارت که از متن های شناخته شده ای مانند Rainbow passage , Grandfather passage و جملات SPIN جمع آوری شده اند. هر فرد حدود ۲ دقیقه از یک متن یکسان را می خواند.

۳ گروه سنی در پایگاه داده وجود دارد:

- جوان (۱۸-۲۹ سال) با میانگین سنی ۲۱ سال
- میان سال (۴۰-۵۵ سال) با میانگین سنی ۴۸ سال
- مسن (۶۲-۹۲ سال) با میانگین سنی ۷۹ سال

در هر گروه سنی ۵۰ نفر هستند که ۲۵ نفر زن و ۲۵ نفر مرد می باشند. در نهایت ۱۴۷ شنونده، سن افراد را تخمین می زنند که اطلاعات بدست آمده در پایگاه داده ثبت شده است. ویژگی های استفاده شده :

۱ – MFCC

۲ - سه گروه ویژگی از قسمت صدادار سیگنال استخراج شده اند (F0)pitch و اولین ۴ فرمنت F1-F4 ، ویژگی های عروضی)

در کل ۲۲۰ ویژگی بدست می آید. برای کاهش ابعاد ویژگی، از روش انتخاب ویژگی MAXR استفاده شده است.

الف) ۲۴ ویژگی MFCC شامل :

- log- energy
- static –MFCC
- ویژگی های دینامیک که با استفاده از یک روند تکراری بر ۵ قاب متوالی صورت می گیرد.

برای محاسبه ضرایب MFCC ، طول پنجره 16ms و شیفت پنجره 10ms در نظر گرفته شده است.

ب) pitch و فرمنت ها :

- f0 با استفاده از تابع cross correlation نرمالیزه شده و برنامه دینامیک، محاسبه می شود.
- فرمنت ها با استفاده از lpc تخمین زده می شوند و با برنامه ی دینامیک بهینه می گردند .

برای هر قاب ۴ فرکانس فرمنت F1-F4 با پهنای باند مربوطه ، B1-B4 ، استخراج می شود.

برای محاسبه ی pitch و فرمنت ها از شیفت قاب 10ms استفاده شده است.

ج) ویژگی های عروضی :

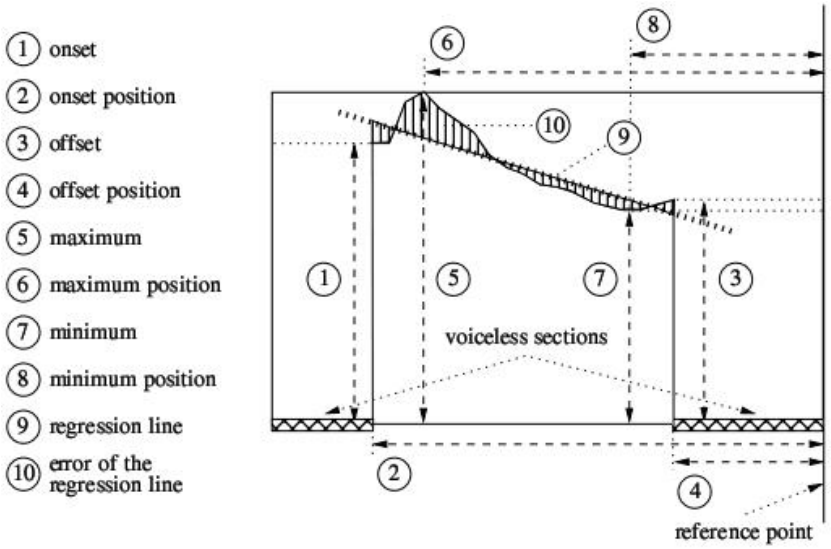
بعد از محاسبه ی ویژگی های pitch وانرژی برای هر قاب در سیگنال و دانستن اینکه هر قاب صوت صداردار است یا بی صدا ، یک بردار ویژگی با ابعاد زیاد برای هر قسمت صدادار سیگنال گفتار ایجاد می شود.

ویژگی های عروضی از ویژگی های پایه و اساسی استنتاج می شوند:

طول مدت هر قسمت، مکث ها ی گفتار، و کیفیت صدا (مانند jitter , shimmer)

صفات متعددی خصوصیات عروضی را برای هر قسمت صدادار سیگنال، مدل می کنند .

مثلا شکل ۲-۹ صفات مختلفی را برای F0 نشان می دهد. هر صفت یک ویژگی عروضی در نظر گرفته می شود.



شکل ۲-۹ : مدل منحنی pitch در یک قسمت صدادار از مرجع [۸]

در کل ۱۸۷ ویژگی عروضی متفاوت که هر کدام متعلق به یکی از گروه ویژگی های زیر هستند، استخراج می شود.(pitch ، انرژی، مکث، طول مدت و کیفیت)

به دلیل اینکه آزمایشات تخمین سن نیاز به یک بردار ویژگی با ابعاد ثابت و یکسان دارد که مستقل از طول مدت عبارات باشند و از ویژگی هایی که تنها از قسمت های خاصی از سیگنال قابل استخراج هستند نیز مستقل باشد، یک بردار ویژگی فراتر(meta)که تنها شامل میانگین ها و انحراف معیار های ویژگی های متفاوت برای هر گوینده است، تشکیل می شود. این نوع مدل کردن مربوط به GMM با یک مخلوط می باشد.

در نتیجه بعد یک بردار فرا ویژگی، دو برابر بردار ویژگی های اصلی می باشد.

یعنی با ۲۲۰ ویژگی متفاوت، برای هر گوینده یک بردار فرا ویژگی با ۴۴۰ بعد داریم که شامل :

- σ و μ مربوط به pitch
- σ و μ مربوط به فرمنت ها
- σ و μ مربوط به MFCC ها
- σ و μ مربوط به ویژگی های عروضی

نتایج بدست آمده از دو جنبه مورد بررسی قرار می گیرند:

- از یک جهت، گوینده ها بر اساس جنسیت تقسیم می شوند و سیستم SVR سن افراد را تخمین می زند. و یا نمونه های مربوط به مردها و زن ها برای تشکیل یک گروه مستقل از جنسیت، ترکیب می شوند.

- از جهت دیگر ویژگی ها جداگانه بررسی می شوند:

ابتدا ویژگی ها به طور دستی انتخاب می شوند (ا ز مجموعه ویژگی های pitch ، فرمنت ، MFCC ، عروضی)

بار بعد، با استفاده از الگوریتم MAXR به طور تحلیلی ویژگی ها را انتخاب می کنیم. (جدول ۲-۱۵) .

در جدول۲-۱۴ نتایج حاصل نشان داده شده است.

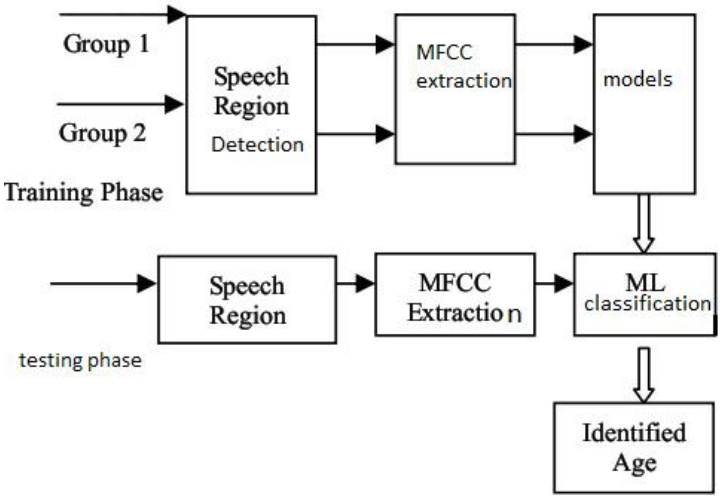
MAE(years)	F0	formants	MFCCs	prosodic	MAXR	All
females	11.3	9.5	9.2	7.3	6.9	6.2
males	13.2	11	8.6	9.5	7.6	7.9
combined	14.5	11	9	9.2	6.9	9.4

جدول ۲-۱۴ : MAE(Mean Absolute Error) ، میانگین خطای مطلقِ آزمایشات با مجموعه ویژگی های متفاوت . ستون MAXR نتایج را بعد از اعمال انتخاب ویژگی با MAXR نشان می دهد.

	female	male	Combined
MFCC	۵	۱	۴
Formant	۳	۱	۰
pitch	۴	۳	۲
energy	۲	۴	۳
Duration	۱	۳	۲
pause	۰	۰	۰
Quality	۰	۰	۱
Number of features	۱۵	۱۲	۱۲

جدول ۲-۱۵ : تعداد گروه های ویژگی مختلف . ویژگی هایی که بر اساس تحلیل MAXR انتخاب شده اند.

درتحقیق [۷] نیز مطالعه ای بر تخمین گروه سنی افراد انجام شده است. در این تحقیق صدا های افراد از دو گروه سنی کمتر از ۳۰ سال (۲۰-۴۰ ساله) و بیشتر از ۵۰ سال (۵۰-۶۰ ساله) ضبط شده است که در هر گروه سنی ۳۰ نفر وجود دارد. عبارت های گفته شده، متن های پیوسته ای هستند. دو روش رده بندی GMM و PNN آزمایش شده است. تنها از ویژگی های طیفی MFCC استفاده شده است. (۱۲ عدد) از هر 10ms از سیگنال، یک بردار ویژگی استخراج شده است. در شکل ۲-۱۰ بلوک دیاگرام مربوط به الگوریتم پیشنهادی در [۷] نشان داده شده است.



شکل ۲-۱۰ : بلوک دیاگرامی از تخمین زننده ی گروه سنی در مرجع [۷]

داده های آموزشی برای هر گروه سنی شامل ۴۰ ثانیه از سیگنال گفتار، که متعلق به صدای ضبط شده ی ۴ گوینده است، می باشد. داده های آزمایشی، شامل ۱۰۰۰ نمونه ی متفاوت از یک مدت زمان ۱۰ ثانیه ای که از ۵۰ گوینده (از هر گروه سنی ۲۵ نفر) گرفته شده است.

نتایج بدست آمده از این روش در جدول های ۲-۱۶ و ۲-۱۷ نشان داده شده است:

Input speaker	Age group 30 years	Age group 50 years
Age group 30 years	76%	24%
Age group 50 years	5%	95%

جدول ۲-۱۶: رده بندی تخمین گروه سنی بر اساس PNN

Input speaker	Age group 30 years	Age group 50 years
Age group 30 years	96.5%	3.5%
Age group 50 years	32%	68%

جدول ۲- ۱۷ : رده بندی تخمین گروه سنی بر اساس GMM

روش جدید دیگری برای تشخیص جنس و تخمین سن، استفاده از روشی با عنوان فاکتورگیری ماتریسی نا منفی است،[۱۵] که مختصرا در زیر به آن اشاره می کنیم.

بعد از تعیین یک مدل صوتی برای همه ی گوینده های پایگاه داده، وزن های مخلوط گوسی استخراج شده و برای تشکیل یک فرابردار برای هر گوینده، الحاق می گردند.

سپس با استفاده از فرابردارهای بدست آمده از داده های آموزشی ، یک ساختار چندگانه (hybrid) از WSNMF(weighted supervised non-negative matrix factorization) و GRNN(general regression neural network) ایجاد می گردد.

NMF(Non negative Matrix Factorization) یک الگوریتم آموزش ماشین (machine learning) مشهور است، که برای تشخیص کلمه، تفکیک منابع صدا بکار برده می شود.

در طول دهه ی گذشته بسط ها و تعمیم های گوناگونی از NMF مانند SNMF , WNMF ، برای حل مشکلات واقعی ایجاد شده است.

- SNMF :

مجموعه داده های آموزش عبارتند از :

$$S^{tr1} = \{(x_1, y_1),...,(x_n, y_n),...,(x_N, y_N)\}$$

x_n : بردار مشخصات بدست آمده (بردار ویژگی)

y_n : بردار برچسب، که اگر x_n متعلق به کلاس i ام باشد، المان i ام y_n برابر یک می شود؛ و بقیه ی المان های y_n برابر صفر هستند. تعداد سطرهای y_n برابر تعدا کلاس ها می باشد.

یک بردار ویژگی ممکن است به چندین کلاس متعلق باشد که در این صورت y_n چندین المان غیر صفر دارد.

هدف تعیین تابع کلاسه بندی (g) است به طوریکه برای هر مشاهده ی آزمایش x^{tst} ، $\hat{y} = g(x^{tst})$ تا حد امکان به برچسب صحیح ، نزدیک باشد.

فاز آموزش :

داده های آموزشی به دو دسته تقسیم شده اند :

- نمونه های آموزش اولیه، برای آموزش WSNMF
- نمونه های آموزش ثانویه، برای آموزش GRNN

۱- آموزش اولیه tr1 :

۶ کلاس متفاوت وجود دارد (با در نظر گرفتن جنسیت)

$$S^{tr1} = \{(x_1^{tr1}, u_1^{tr1}), ..., (x_n^{tr1}, u_n^{tr1}), ..., (x_N^{tr1}, u_N^{tr1})\}$$

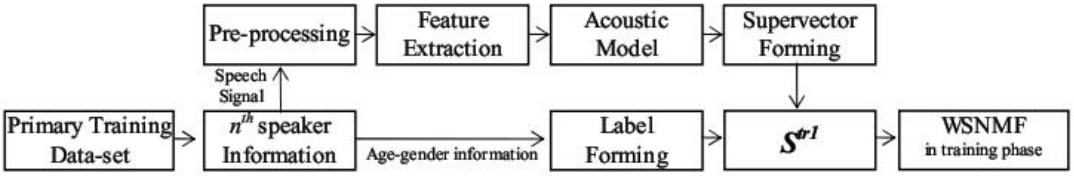
x_n^{tr1} : فرابردارها

u_n^{tr1} : بردار برچسب مربوط به شخص n ام. که اگر این شخص متعلق به کلاس i ام باشد، سطر i ام برابر یک است و بقیه ی المان ها صفر هستند.

$x_n = [(\mathrm{a}^1)^T ... (\mathrm{a}^s)^{/T} ... (\mathrm{a}^S)^T]^T$ که S برابر تعداد کل مخلوط ها می باشد.

$\mathrm{a}^s = N^s[w_1^s ... w_q^s ... w_Q^s]^T$ که N^s تعداد قاب ها در مخلوط s ام می باشد. و Q تعداد کل وزن ها در s امین مخلوط است.

ویژگی های صوتی شامل MFCC با میانگین نرمالیزه شده، طول لوله ی صوتی نرمالیزه شده، و مشتق های اول و دوم. این ویژگی ها با استفاده از یک الگوریتم LDA به یک فضای ۳۶ بعدی تصویر می شوند. که مستقل از هم هستند و همبستگی ندارند.



شکل ۲- ۱۱: بلوک دیاگرام روش مورد نظر در فاز آموزش اولیه در مرجع [۱۵]

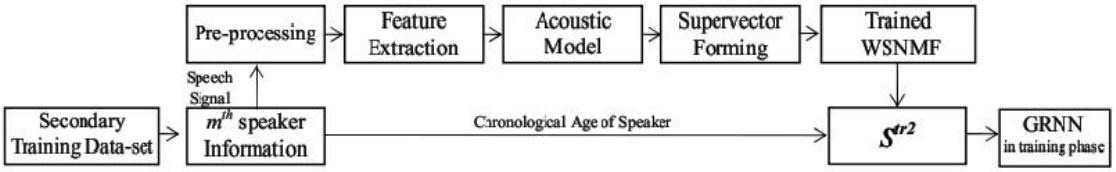
۲- آموزش ثانویه tr2 :

روند بدست آوردن فرابردارها و وزن های GMM برای هر گوینده تکرار می شود. (در مجموعه داده های آموزش ثانویه) سپس فرا بردار بدست آمده برای k امین گوینده از این مجموعه داده x_k^{tr2} ، به NMF آموزش داده شده در فاز اول اعمال می شود تا برچسب جنس – سن $\hat{u}_k^{tr2} = g(¥_k^{tr2})$ او تخمین زده شود.

در فاز آموزش سن دقیق هر گوینده را می دانیم. در نتیجه برای k امین برچسب تخمین زده شده ی جنس- سن، سن گوینده A_k^{tr2}، معلوم است. در نتیجه، بعد از تخمین برچسب جنس – سن همه ی M گوینده ی مجموعه آموزش ثانویه، یک مجموعه ی ورودی – خروجی ثانویه را می توان تشکیل داد:

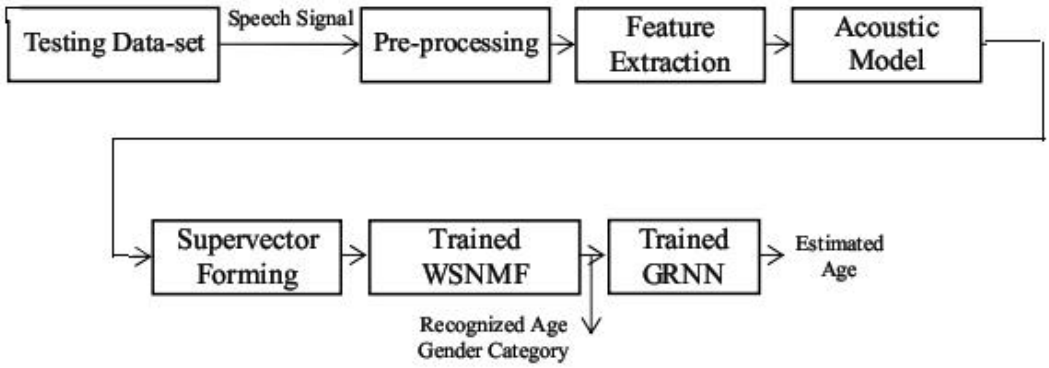
$$S^{tr2} = \{(\hat{u}_1^{tr2}, A_1^{tr2}), ..., (\hat{u}_k^{tr2}, A_k^{tr2}), ..., (\hat{u}_K^{tr2}, A_K^{tr2})\}$$

این مجموعه برای آموزش GRNN استفاده می شود.



شکل ۲- ۱۲ : بلوک دیاگرام روش مورد نظر در فاز آموزش ثانویه در مرجع [۱۵]

فاز آزمایش :



شکل ۲- ۱۳ : بلوک دیاگرام روش مورد نظر در فاز آزمایش در مرجع [۱۵]

در این تحقیق [۹] از پایگاه داده ی Dutch استفاده شده است. که شامل ۵۵۵ گوینده می باشد. که افراد، اخبار، مصاحبه ها، گزارش ها و گزارش های هوا شناسی را به طور زنده اجرا می کنند و یا از روی نوشته ای می خوانند.

افراد در این پایگاه داده به ۳ گروه سنی تقسیم شده اند. (جوان – میان سال - مسن)

نام دسته	Young male (YM)	Young female (YF)	Middle male (MM)	Middle female (MF)	Senior male (SM)	Senior female (SF)
سن	۱۸ - ۳۵	۱۸ - ۳۵	۳۶ - ۴۵	۳۶ – ۴۵	۴۶- ۸۱	۴۶- ۸۱
تعدادگوینده ها	۸۵	۵۳	۱۶۰	۴۱	۱۹۱	۲۵

جدول ۲-۱۸ : تعداد گوینده ها در دسته های مختلف سن – جنس

برای ارزیابی روش پیشنهادی، از شیوه ی 5-fold cross-validation استفاده شده است.

پس همه ی ۵۵۵ گوینده در پایگاه داده به ۵ دسته ی نا مرتبط و نا پیوسته تقسیم می شوند که در نتیجه هر دسته شامل ۱۱۱ گوینده است. سپس ۵ آزمایش مستقل اجرا می گردد. پس در هر آزمایش، ۴ دسته به عنوان داده ی آموزش و ۱ دسته ی باقی مانده به عنوان داده ی آزمایش به کار می رود. در همه ی ۵ آزمایش، ۳۴۴ نفر از ۴۴۴ گوینده ی مجموعه ی آموزش، به عنوان داده های آموزشی اولیه مورد استفاده قرار می گیرد.

صحت تشخیص جنس در همه ی ۵ آزمایش 96% می باشد .

در جدول۲-۱۹، صحت تخمین گروه سنی در همه ی آزمایش ها برای هر ۳ گروه سنی (جوان Y – میان سال M – مسن S)نشان داده شده است.

دسته ی سنی	Y	M	S
صحت شناسایی	38	40	65

جدول ۲- ۱۹ : صحت شناسایی گروه سنی بر حسب درصد

نتیجه ی نهایی برای ۶ گروه سنی – جنسیتی در جدول ۲- ۲۰ آورده شده است.

CL / AC	YM	YF	MM	MF	SM	SF
YM	13	03	58	0	26	0
YF	02	77	04	11	057	0
MM	06	01	44	01	47	0
MF	0	54	02	24	17	02
SM	03	01	19	0	76	0
SF	0	2	08	28	28	16

جدول ۲- ۲۰ : ماتریس ترکیبی نسبی برای روش مورد نظر در شناسایی ۶ گروه سن - جنسیت

که صحت تخمین سن با اندازگیری اندازه ی خطای میانگین محاسبه می شود، $MAE = \frac{\sum_{i=1}^{N} |w_i - \hat{w}_i|}{N}$ که

w_i سن حقیقی و $\hat{w}_i$ سن تخمینی برای iامین گوینده.

MAE(mean absolute error) برای روش پیشنهادی در همه ی ۵ آزمایش برابر ۷.۴۸ سال می باشد. [۱۵]

برای بهبود عملکرد سیستم تخمین سن و تعیین جنسیت افراد با سیگنال صوت، در طی مطالعه ای [۱۶] پنج روش متفاوت در سطح اکوستیک ترکیب شده اند. که در زیر نام برده شده است:

۱- سیستم Gaussian mixture model (GMM) بر اساس ویژگی های MFCC

۲- Support vector machine (SVM) بر اساس فرابردارهای میانگین GMM

۳- Support vector machine (SVM) بر اساس فرابردارهای ماتریس maximum likelihood linear regression (MLLR) در GMM

۴- Support vector machine (SVM) بر اساس فربردارهای جفت 'Tandem' در GMM

۵- سیستم پایه ی SVM بر اساس بردارهای ویژگی ۴۵۰ بعدی که شامل ویژگی های عروضی در سطح عبارات می باشند.

در شکل ۲- ۱۴ نمایی کلی از سیستم و چگونگی ترکیب روش های مختلف و به کارگیری آنها نشان داده شده است.

برای پیاده سازی این روش در [۱۶] از پایگاه داده ی

2010 Interspeech Paralinguistic Challenge aGender database استفاده شده است.

که گوینده ها شامل :

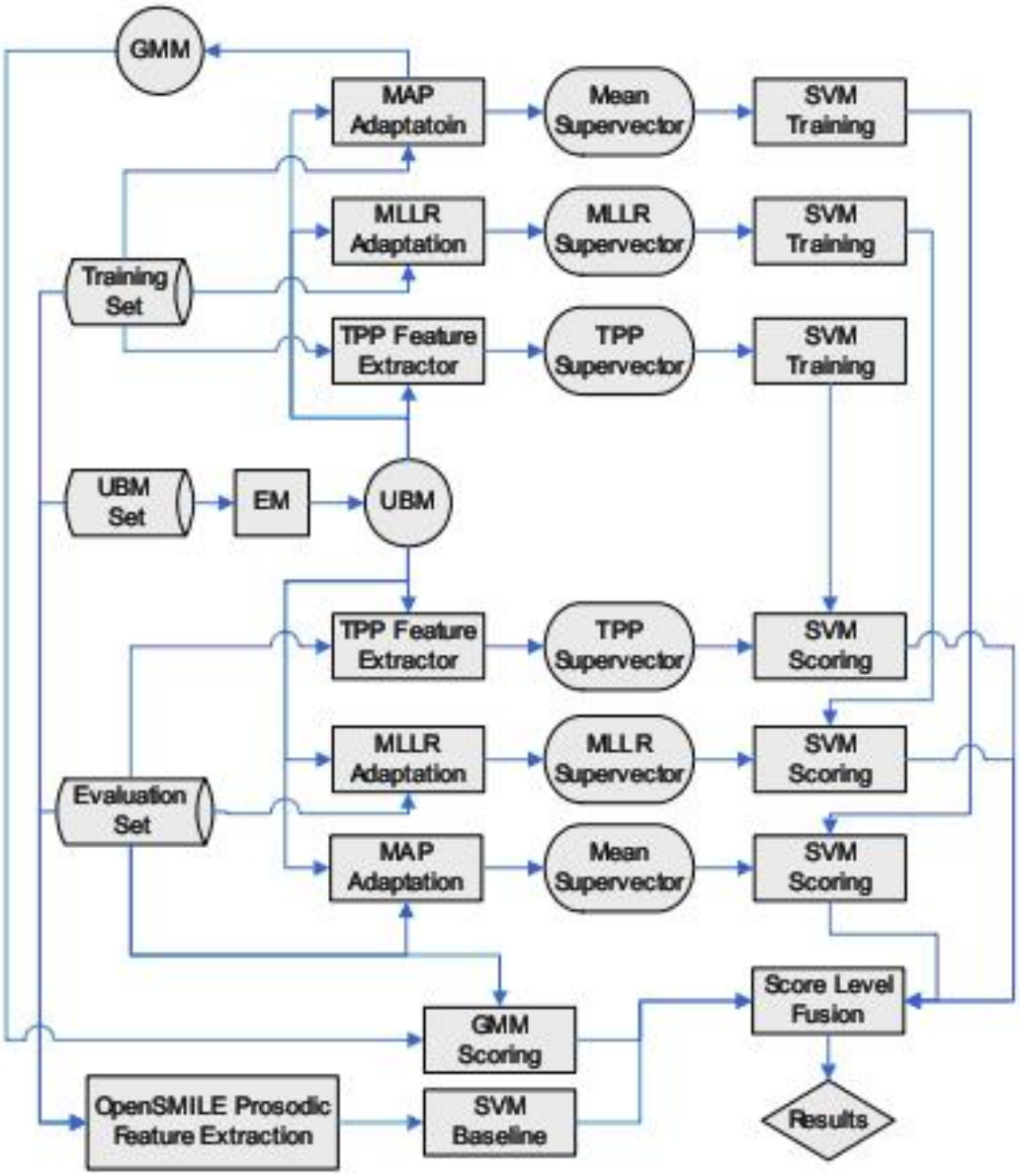
- کودکان زیر ۱۳ سال , C
- نوجوانان ۱۴- ۱۹ ساله ،YF/YM
- بزرگسالان ۲۰- ۵۴ ساله ، AF/AM
- افراد مسن با ۵۵ سال به بالا سن .SF/SM

می باشند. نتایج بدست آمده بر حسب صحت تخمین در جدول ۲-۲۲ آمده است.

	C	YF	YM	AF	AM	SF	SM
C	**67.0**	13.9	5.8	5.2	1.1	6.4	0.7
YF	16.4	**59.0**	0.5	16.0	0.1	7.7	0.2
YM	0.2	0.4	**57.1**	1.8	17.6	3.3	19.7
AF	4.0	25.0	0.9	**37.9**	0.2	31.8	0.1
AM	0.2	0.0	34.5	1.5	**25.4**	1.3	37.1
SF	4.7	8.9	0.9	27.9	0.3	**56.4**	0.9
SM	0.1	0.0	15.3	0.5	15.6	2.6	**65.9**

جدول ۲- ۲۲ : ماتریس ترکیبی برای ۷ دسته ی سن - جنسیت

بعلاوه این سیستم ترکیبی دارای صحتی برابر % 52.7 برای کلاسه بندی همزمان جنس- سن می باشد.



شکل ۲ – ۱۴ : چشم انداز سیستم در مرجع [۱۶]

فصل سوم :

مروری بر انواع روش های

استخراج ویژگی

۳-۱ مقدمه

هدف از استخراج ویژگی، تبدیل شکل موج گفتار به انواع نمایش پارامتری (که به طور قابل توجهی دارای نرخ اطلاعاتی کمتری هستند) برای تحلیل و پردازش بیشتر می باشد، که به عنوان مقدمه ی اصلی پردازش سیگنال شناخته می شود. سیگنال گفتار، هنگامی که بر یک دوره ی زمانی به اندازه ی کافی کوتاه (5~10 ms) در نظر گرفته می شود، سیگنالی است که به آرامی با زمان تغییر می کند(شبه-متناوب)؛ به این معنا که مشخصات سیگنال تقریبا ثابت و بدون تغییر می ماند. با این حال، بر یک دوره ی زمانی بلند مدت (در مرتبه ی $\frac{1}{5}$ ثانیه و یا بیشتر) مشخصات سیگنال تغییر می کنند تا بتوانند صوت های متنوعی که گفته می شود را منعکس کنند. در نتیجه، تحلیل طیفی زمان- کوتاه معمول ترین راه برای توصیف سیگنال گفتار می باشد.

هر قاب زمان- کوتاه گفتار به یک بردار ویژگی تبدیل می شود.

بردار ویژگی باید شامل اطلاعاتی باشد که

- برای شناسایی و تمایز صدا های گفتار کارآمد باشد
- نسبت به ویژگی های فردی گوینده و دیگر فاکتورهای نامربوط، حساس نباشد.

بردار ویژگی یک توصیف اختصاصی از طیف توان سیگنال است.

می دانیم که حساسیت گوش های انسان به گفتار غیر خطی می باشد. یک مقیاس غیر یکنواخت بر اساسMel، معمولا برای شبیه سازی این حساسیت غیر خطی گوش ها برای اهداف استخراج ویژگی گفتار به کار می رود.

۳-۲ استخراج ویژگی در برابر کلاسه بندی:

استخراج ویژگی روندی برای تبدیل یک رشته از نمونه های گفتار به مجموعه ای از بردارهای مشاهده می باشد که رویداد ها را در فضای احتمالاتی که عمل کلاسه بندی در آن صورت می گیرد، نمایش می دهد.

استخراج ویژگی به عنوان تصویر کردن یک بردار n بعدی x به یک بردار m بعدی y نیز تعریف می شود که $m<n$ به گونه ای است که یک رابطه ی معیار مناسب، بهینه شود. و کلاسه بندی به عنوان تصویر کردن یک بردار مشاهده ی x به یک کلاس ωi که $i= 1,2,.., L$ و L تعداد کلاس ها می باشد.

همبستگی بالایی بین روش های استخراج ویژگی و طراحی کلاسه بند برای شناسایی گفتار وجود دارد. بسته به مسئله ی کلاسه بندی، پیچیدگی طراحی یک سیستم شناسایی گفتار بین گام استخراج ویژگی و گام کلاسه بندی توزیع می شود. یک کلاسه بند کارامد و موثر، این پیچیدگی را بین گام های استخراج ویژگی و کلاسه بندی به طور مناسبی توزیع می کند.

سه مسئله ی مهم در طراحی سیستم های استخراج ویژگی عبارتند از :

۱- استخراج ویژگی هایی که از لحاظ ادراکی، با معنی هستند.

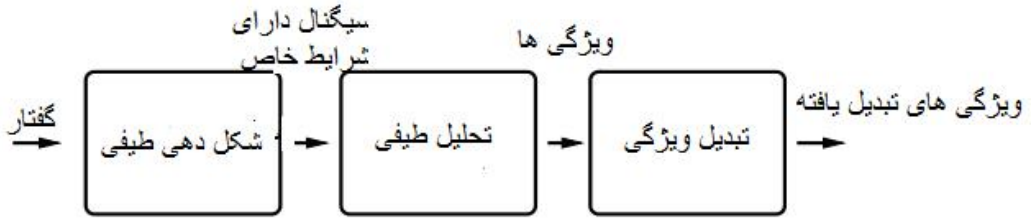
۲- استخراج ویژگی های تغییر ناپذیر و یا پایدار

۳- استخراج ویژگی هایی که مبین همبستگی زمانی طیف هستند.

۳-۳ مراحل استخراج ویژگی از سیگنال گفتار

یک سیستم استخراج ویژگی معمول شامل ۳ گام اصلی است:

۱- شکل دهی طیفی ۲- تحلیل طیفی ۳- تبدیل پارامتری



شکل ۳- ۱ : بلوک دیاگرام نشان دهنده ی ۳ گام اصلی در استخراج ویژگی برای شناسایی گفتار

۳-۳-۱ شکل دهی طیفی:

سیگنال گفتار طبیعی به صورت آنالوگ در دسترس است. شکل دهی طیفی سیگنال گفتار طبیعی با یک تبدیل آنالوگ به دیجیتال سیگنال شروع می شود. میکروفن استفاده شده برای ضبط سیگنال گفتار، اعوجاج غیرخطی و نویز فرکانس خط را ایجاد می کند. علاوه بر این، عملکرد تبدیل A_D (آنالوگ به دیجیتال) اعوجاج خود را به دلیل تابع تبدیل غیر خطی مبدل A-D ، دارد. نرخ نمونه برداری 8,10,12 , 16 K معمولا برای دیجیتال کردن سیگنال گفتار به کار می رود. این نرخ های نمونه برداری دقت و تفکیک پذیری زمانی و فرکانس خوبی بدست می دهند. سیگنال دیجیتال شده از یک فیلتر با پاسخ ضربه محدود (FIR) با تابع تبدیل (۳-۱)عبور می کند.

(۱-۳)

$$H(z) = \sum_{k=0}^{k=N-1} a(k) z^{-k}$$

که N طول فیلتر است. در عمل، یک فیلتر دیجیتال با یک ضریب به صورت (۳-۲)

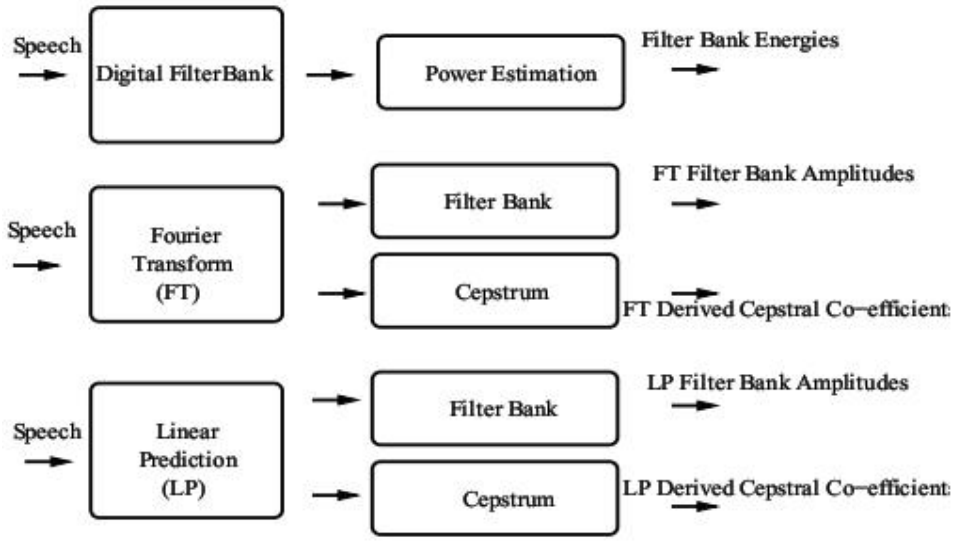
$$H(z) = 1 + az^{-1} \qquad (۲-۳)$$

که فیلتر پیش تاکید نامیده می شود، به کار می رود. مقدار ~ معمولا از 1- تا 0.4- تغییر میکند.

گفتارطبیعی یک شیب (tilt) طیفی به اندازه ی -20 dB/octave دارد. فیلتر پیش تاکید برای جبران این شیب طیفی گفتار طبیعی به کار می رود. شکل دهی فرکانسی به طور کلی با تحلیل طیفی دنبال می شود.

۳-۳-۲ تحلیل طیفی:

الگوریتم های تحلیل طیفی را می توان به ۶ کلاس اصلی تقسیم کرد که دسته بندی آنها در شکل ۳-۲ نشان داده شده است.



شکل ۳ - ۲ : دسته بندی الگوریتم های تحلیل طیفی

روش های تحلیل طیفی اولیه، انرژی های بانک فیلتر را پس از عبور طیف گفتار از یک بانک فیلتر محاسبه می کنند. تکنیک های بر اساس پیش بینی خطی (linear prediction - LP) به طور گسترده در دهه های ۱۹۷۰ تا ۱۹۸۰ به کار می رفتند. اگرچه انواع تکنیک های LP مانند perceptual linear prediction (PLP) هنوز هم کاربرد دارند، تکنیک های تبدیل فوریه، به طور وسیع تری به کار برده می شود.

فرکانس صوتی طبیعی f به یک مقیاس فرکانسی غیرخطی تصویر می شود که عموما یا مقیاس Bark است و یا مقیاس Mel . مقیاس Bark به صورت (۳-۳) تعریف می شود:

$$B_f = 13\ tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\ tan^{-1}\left(\frac{f^2}{(7500)^2}\right) \qquad (۳-۳)$$

و یا

$$\text{Bark}(f) = 7\,\text{arcsinh}\left(\frac{f}{650}\right) \qquad (۳-۴)$$

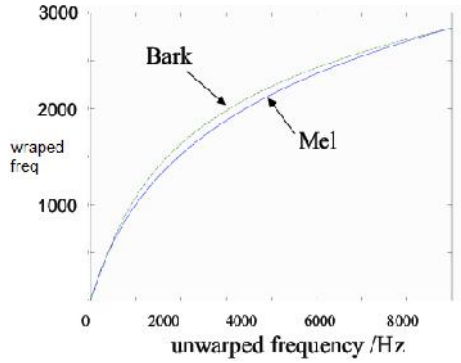
که B_f مقیاس فرکانسی Bark است.

مقیاس Mel به صور ت (۳-۵) تعریف می شود :

$$M_f = 2595\ log_{10}\,(1 + f/700) \qquad (۵\text{-}۳)$$

که M مقیاس فرکانس Melمی باشد.

فرکانس صوتی طبیعی f با عبور طیف گفتار از یک بانک فیلتر که در مقیا س Bark و یا Mel می باشد ، به یک مقیاس فرکانسی غیرخطی تصویر می شود .[۱۷]



شکل ۳-۳ : مقایسه ی مقیاس Bark , Mel از مرجع [۱۹]

ویژگی های Cepstral به کار رفته در شناسایی گفتار بر اساس تکنیک های پردازش سیگنال homomorphic می باشند .این ویژگی ها به دلیل توانایی شان در تفکیک منبع از سیستم (deconvolution)، در حوزه ی استخراج ویژگی گفتار مهم می باشند. ضرایب Cepstral از تبدیل فوریه ی انرژی های بانک فیلتر بدست می آیند که cepstral frequency co-efficients (MFCC) نامیده می شوند.

ضرایب LP با به کارگیری تکنیک های مدل کردن پارامتری، که طیف گفتار را در یک پردازش بازگشتی مدل می کند ،استخراج می شوند. هم چنین این ضرایب را می توان از دامنه های بانک فیلتر LP استخراج کرد که از نمونه برداری طیف LP در فرکانس های بحرانی بانک فیلتر و یا ضرایب cepstral ِLP که ضرایب cepstral ی هستند که از طیف LP محاسبه شده اند. ضرایب cepstral ی که از باند بحرانی طیف LP استخراج می شود، ضرایب perceptual linear prediction co-efficients (PLPCC) نامیده می شوند. تحلیل طیفی، یک سری ویژگی های همبسته را نتیجه می دهد، که می توانند تنها اطلاعات ایستا (static) در سیگنال گفتار را به دام اندازند، به همین دلیل معمولا این تحلیل با یک مرحله ی تبدیل ویژگی همراه است.

۳-۳-۳ تبدیل ویژگی:

سیگنال گفتار با تغییرات زمانی توصیف می شود که مشخص کردن آنها با معیارهای اندازه گیری مطلق دشوار می باشد. بنابراین مشتق های زمانی مرتبه بالا معمولا به بردار های ویژگی مطلق اضافه می شوند. مشتق های زمانی مرتبه اول و دوم بردارهای ویژگی مطلق به طور رایج درشناسایی گفتار به کار می روند. مشتق های زمانی مرتبه ی اول و دوم با استفاده از اختلاف بین قاب های همسایه ی گفتار، با استفاده از تحلیل بازگشتی محاسبه می شوند.

۳-۴ دسته بندی تکنیک های استخراج ویژگی گفتار

به طور کلی تکنیک های استخراج ویژگی گفتار به ۷ دسته تقسیم می شوند که هر کدام به طور مختصر توضیح داده می شود.

- **روش حوزه ی شکل موج:**

این روش از دامنه، طول مدت (duration) ،نرخ عبور از صفر و نرخ عبور از سطحِ سیگنال گفتار به عنوان ویژگی استفاده می کند. اِشکال این روش نرخ داده ی بالایی است که درگیر می باشند.

- **روش بر اساس مدل طیفی صوتی**

این روش شبیه سازی واضحی از مدل صوتی انسان در مورد فعالیت زنگ صدا (طنین) در حلقه ی گوش، را درگیر می کند. اشکال این روش عملکرد نسبتا ضعیف در شناسایی و عدم فهم نقش مغز در ادراک انسان.

- **روش حوزه ی تبدیل :**

ویژگی های گفتار در حوزه ی تبدیل یافته استخراج می شوند، که اطلاعات مناسب گرفته می شوند و اطلاعات اضافی از سیگنال گفتار حذف می شود. بعضی از روش های رایج حوزه ی تبدیل عبارتند از بانک های فیلتر، تبدیل فوریه ی گسسته (DFT) ، ضرایب پیش بینی خطی ((linear prediction co-efficients (LPC) و تبدیل های متعامد مانند تبدیل Karhunen-Loeve و تبدیل کسینوسی گسسته (DCT) .

- **روش پردازش سیگنال Homomorphic**

معمولا در علوم مهندسی با سیگنال هایی که درک و تحلیل آنها مشکل است، مواجه می شویم. استراتژی پردازش homomorphic ، تبدیل این وضعیت غیر قابل کنترل به یک سیستم خطی ساده می باشد، که تکنیک های تحلیل آنها به خوبی قابل درک است. این کار با اعمال تبدیل ها و طرح های ریاضی که برای یک کاربرد خاص مناسب باشد، صورت می گیرد.

برای مثال، کاربرد کلاسیک پردازش homomorphic ، جداسازی سیگنال هایی است که درهم ضرب می شوند، مانند $a(t) = b(t) \times c(t)$. که این سیستم را می توان با لگاریتم گرفتن به یک سیستم خطی تبدیل کرد، یعنی سیگنال هایی که با هم جمع می شوند: $log[a(t)] = log[b(t)] + log[c(t)]$

این روش رایج ترین روش پردازش سیگنال برای استخراج ویژگی گفتار به کار می رود. ویژگی های cepstral با استفاده از این روش استخراج می شوند.

- **روش فشرده سازی طیفی:**

تحقیقات اولیه توسط Stevens and Volkman بر سیستم شنوایی انسان که با طیف Mel معرفی می شود، صورت گرفته است.

- **روش حوزه ی مشترک زمان – فرکانس :**

طیف نگاره ها spectogram از اولین نمونه های توزیع های زمان– فرکانس

time frequency distributions -TFD) مورد استفاده در مدل کردن سیگنال هستند.

- **روش شکل دهی طیفی :**

بدلیل مشخصه های فیزیولوژیکی مکانیزم تولید گفتار، گفتار صدادار یک شیب طیفی منفی 20- dB دارد. شیب منفی یا با به کارگیری فیلتر پیش تاکید و یا با اضافه کردن توان به بعضی از ضرایب، مستقیما در طیف، در طی عملیات نمایش گفتار ظاهر می شود.

۳-۵ روش های رایج استخراج ویژگی:

با ظهور مدل های Markov در شناسایی گفتار، ویژگی هایی که با معنا و مفهوم هستند و نسبت به محیط صوتی محدود، تغییر ناپذیر می باشند، به سرعت معمول و عمومی شدند.

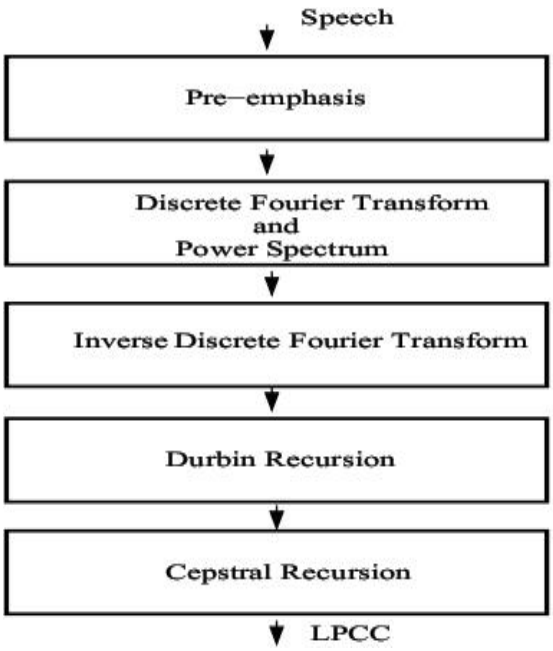
ویژگی های اخیر استفاده شده در شناسایی گفتار ضرایب frequency cepstral coefficients (MFCC) ، perceptual linear prediction cepstral coefficients (PLPCC) ، perceptual critical bank features (PCBF) ، انرژی ، توان ، منحنی های F0 و ویژگی های منبع می باشند.

MFCC شاید معمول ترین ویژگی استفاده شده در شناسایی گفتار باشند. ویژگی های طیفی مانند MFCC به طور کلی از طیف دامنه ی تبدیل فوریه ی زمان – کوتاه محاسبه می شوند.

ویژگی های مدل – مبنا مانند LPCC و PLPCC با به کارگیری تحلیل پیش بینی خطی محاسبه می شوند.

۳-۵-۱ ضرایب Linear Prediction Cepstral

این ضرایب از دهه ی ۱۹۷۰ تا ۱۹۸۰ به طور وسیعی در شناسایی گفتار استفاده شد که متعلق به روش های مدل – مبنای کلی (generic model based) می باشد. Makhoul روش های پیش بینی خطی را برای شناسایی گفتار معرفی کرد. در شکل ۳-۴ نحوه ی محاسبه یLPCC نشان داده شده است.



شکل ۳ – ۴ : گام های محاسبه ی LPCC

مراحل محاسبه ی LPCC :

- اعمال قاب بندی و پنجره گذاری برسیگنال گفتار
- محاسبه ی DFT و مربع دامنه ی آن.
- اعمال عکس تبدیل فوریه گسسته
- استنتاج یک مدل خود بازگشتی با به کارگیری تحلیل بازگشتی
- استفاده از یک تبدیل متعامد مانند KLT و یا DCT برای محاسبه ی LPCC ناهمبسته [۱۷]

پیش بینی ای که LPC از s[n] می کند، $\sum a_k s[n-k]$ می باشد. که خطا برابر است با :

$$E_n = \sum_{m=-\infty}^{\infty} \left(s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right)^2 \qquad (۶-۳)$$

اگر a_k ها معادلات نرمال (۳-۷) را حل کنند، خطا حداقل می شود:

$$R_n(i) = \sum_{k=1}^{p} a_k R_n(|i-k|) \qquad (۷-۳)$$

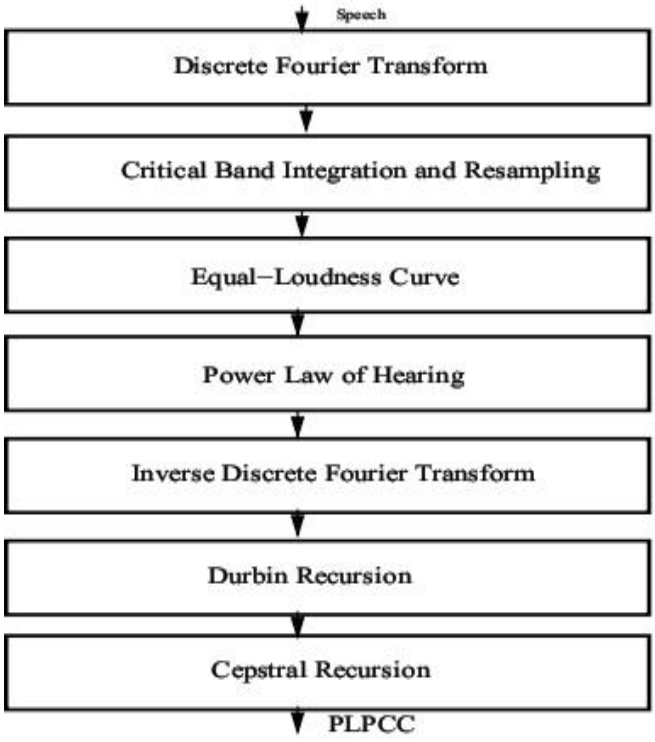
$$R_n(i) \equiv \sum_{m=-\infty}^{\infty} s_n(m) s_n(|m-i|)$$

که تابع خود همبستگی به صورت (۳-۸) تعریف می شود:[۱۸]

$$R_n(k) = \sum_{m} x_n(m) x_n(m-k) \qquad (۸-۳)$$

۳-۵-۲ ضرایب Perceptual Linear Prediction

perceptual linear prediction (PLP) طیف linear prediction (LP) را به یک مقیاس فرکانس غیر خطی از گوش انسان، تصویر می کند. ضرایب perceptual linear prediction co-efficients (PLPCC) ، تعمیم یاقته ی LPCC هستند. PLPCC در مقایسه با LPCC عملکرد نسبتا بهتری در شناسایی گفتار دارد.



شکل ۳ -۵ : گام های محاسبه ی PLPCC

مراحل محاسبه ی PLPCC :

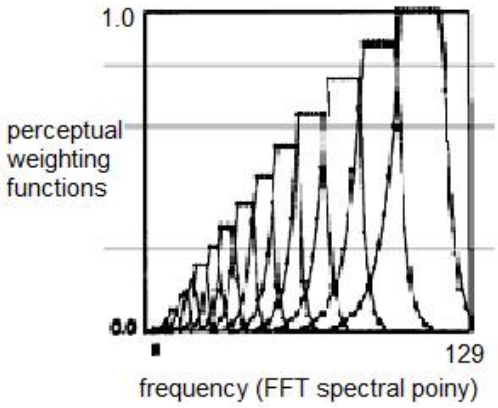
- اعمال قاب بندی و پنجره گذاری بر سیگنال گفتار
- محاسبه ی DFT و مربع دامنه ی آن
- انتگرال گیری از طیف توان
- فیلتر پیش تاکید کردن طیف برای شبیه سازی حساسیت نابرابرگوش انسان به فرکانس های مختلف
- فشرده کردن دامنه های طیف با ریشه ی سوم گرفتن بعد از انتگرال گیری
- اعمال عکس تبدیل فوریه ی گسسته
- اعمال صاف سازی (یکنواخت سازی) طیفی درطیف های با باندهای بحرانی critical band spectra با استفاده از یک مدل خود بازگشتی حاصل از یک تحلیل بازگشتی
- استفاده از یک تبدیل متعامد مانند KLT و یا DCT برای محاسبه یPLPCC
- اعمال اختیاری lifter برای برابرکردن (equalize) واریانس ضرایب cepstral مختلف [۱۷]

ابتدا طیف فرکانسی را به مقیاس Bark تبدیل می شود. (۳-۹)

$$\tilde{S}(b) = \sum_{k=0}^{N-1} |H_b(k)|^2 |X(k)|^2, \quad b = 1, \ldots, K \qquad (۹-۳)$$

فیلتر ها ،Hb(k) ، به طور یکنواخت در فرکانس Bark قرار می گیرند.(شکل ۳-۶)

دامنه ی آنها با منحنی بلندی صدای – برابر (equal-loudness contour) اندازه گیری می شود.(تخمینی از مقدار بلندی صدا در هر فرکانس)



شکل ۳ -۶: از مرجع [۱۸]

ریشه ی سوم طیف توان محاسبه می شود :

- ریشه ی سوم جایگزین لگاریتم که در MFCC بکار می رود، میشود.
- بلندی یک آهنگ صدا با ریشه ی سوم توانش نسبت دارد:

$$Y(b) = S(b)^{0.33} \qquad (۱۰-۳)$$

سپس، عکس تبدیل فوریه برای یافتن perceptual autocorrelation محاسبه می شود:

$$\begin{aligned} \tilde{R}(m) &= \frac{1}{2K}\sum_{b=0}^{2K} Y(b) e^{\frac{j2\pi bm}{2K}} \\ &= \frac{1}{K}\sum_{b=1}^{K} Y(b)\cos\left(\frac{\pi bm}{K}\right) + \frac{(-1)^m}{2K} Y(K) \end{aligned} \qquad (۱۱-۳)$$

سپس از معادلات نرمال (۳-۱۲) برای یافتن ضرایب perceptual LPC (PLP) استفاده می شود :

$$\tilde{R}(m) = \sum_{k=1}^{p} \tilde{a}_k \tilde{R}(|m-k|) \qquad (۱۲-۳)$$

در نهایت، از رابطه ی بازگشتی (۳-۱۳) برای یافتن perceptual LPC cepstrum (PLP CC) استفاده می شود.[۱۸]

$$\tilde{c}(m) = \tilde{a}_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \tilde{c}(k)\tilde{a}_{m-k}, \quad 1 \le m \le p \qquad (۱۳-۳)$$

۳-۵-۳ ضرایب Linear Frequency Cepstral

LFCC ویژگی های cepstral بودند که قبل از ظهور MFCC به کار می رفتند. که به صورت زیر محاسبه می شوند:

- سیگنال گفتار x(n) با یک تابع پنجره ی w(n) پنجره گذاری می شود و تبدیل فوریه ی زمان – کوتاه گسسته ی آن $X(n,\mu_k)$ محاسبه می شود :

$$X(n,\omega_k)=\sum_{m=-\infty}^{+\infty}x(m)w(n-m)e^{-j\omega_k m} \qquad (۱۴-۳)$$

$$\omega_k=\frac{2\pi}{N}k$$

که N طول DFT و n,k اعداد صحیح هستند.

- مربع دامنه ی $X(n,\mu_k)$ ، سپس با یک سری پاسخ فرکانسی فیلترهایی که بر مقیاس خطی توزیع شده اند، وزن دهی میشود.

- در نهایت LFCC به صورت (۳-۱۵) بدست می آیند :

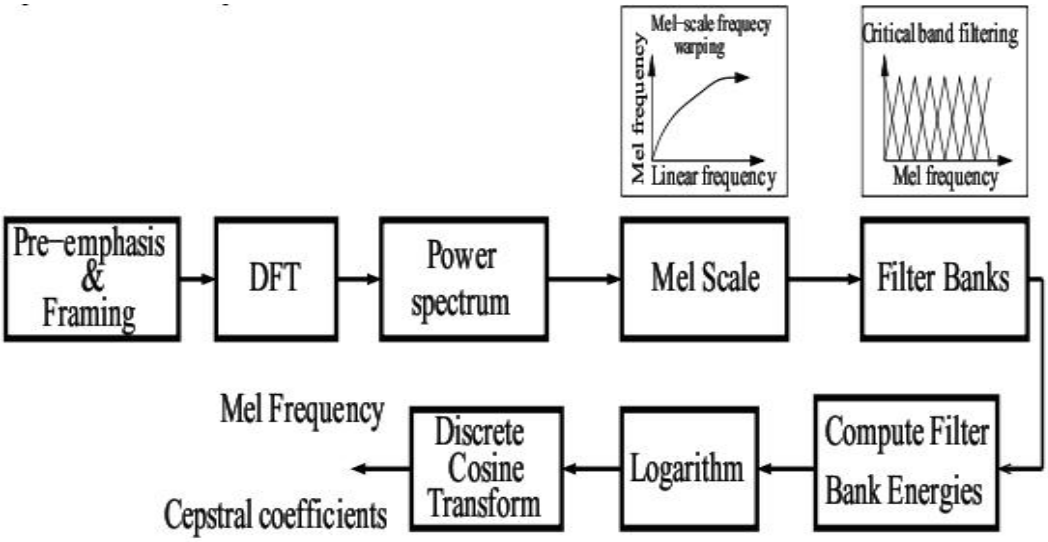
$$C_{lin}[n,m]=\frac{1}{R}\sum_{l=0}^{R-1}log\{E_{lin}(n,l)\}cos(\frac{2\pi}{R}lm) \qquad (۱۵-۳)$$

که R تعداد فیلتر و عکس تبدیل به صورت تبدیل کسینوسی گسسته DCT و $E_{lin}(n,l)$ انرژی در هر قاب سیگنال گفتار در زمان n و برای l امین فیلتر مقیاس خطی است.

LFCC را همچنین می توان مستقیما با تبدیل کسینوسی گسسته گرفتن از طیف تبدیل فوریه ی گسسته ی دامنه لگاریتمی محاسبه کرد. [۱۷]

۳-۵-۴ ضرایب Mel Frequency Cepstral

امروزه شاید پرکاربردترین ویژگی در شناسایی گفتار، MFCC می باشد. اساس MFCC بر تغییرات شناخته شده ی پهنای باند های بحرانی گوش انسان با فرکانس ها، و فیلترهایی که در فرکانس های پایین به صورت خطی و در فرکانس های بالا به صورت لگاریتمی قرار گرفته اند، می باشد. MFCC یک کاهش داده ی قابل توجهی را ایجاد می کند، زیرا تعداد کمی از ضرایب برای نمایش Cepstrum سیگنال صوتی کافی است. Stevens و Volkman مقیاس Mel را به عنوان نتیجه ی مطالعه ی ادراکی صوت و شنوایی انسان، گسترش دادند. مقیاس Mel توسط Mermelstein و Davis، برای استخراج ویژگی از سیگنال گفتار برای بهبود عملکرد شناسایی به کار برده شد.



شکل ۳- ۷: بلوک دیاگرام نشان دهنده ی گام های محاسبه ی MFCC از مرجع [۱۷]

مراحل محاسبه ی MFCC :

- سیگنال گفتار x(n) با یک پنجره ی w(n) پنجره گذاری می شود و تبدیل فوریه ی زمان – کوتاه گسسته ی آن $X(n,\mu_k)$ محاسبه می شود.

$$X(n,\omega_k) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\omega_k m} \qquad (۱۶-۳)$$

$$\omega_k = \frac{2\pi}{N}k$$

که N طول DFT و n,k اعداد صحیح هستند.

- مربع دامنه ی $X(n,\mu_k)$ با یک سری پاسخ فرکانسی فیلترهایی که در مقیاس Mel توزیع شده اند، وزن دهی می شود.

- انرژی در هر قاب ازسیگنال گفتار در زمان n و برای l امین فیلتر در مقیاس mel محاسبه می شود:

$$E_{Mel}(n,l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l} |V_l(\omega_k)X(n,\omega_k)|^2 \qquad (۱۷-۳)$$

که L_1, U_1 قطع بالا و پایین هر فیلتر می باشد. و

$$A_l = \sum_{k=L_l}^{U_l} |V_l(\omega_k)|^2 \qquad (۱۸-۳)$$

در نهایت ضرایب MFCC محاسبه می شوند:

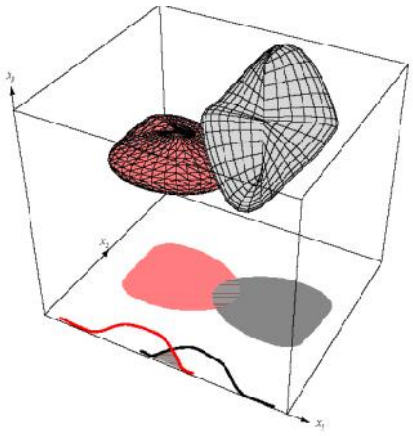
$$C_{Mel}[n,m] = \frac{1}{R}\sum_{l=0}^{R-1} log\{E_{Mel}(n,l)\}cos(\frac{2\pi}{R}lm) \qquad (۱۹-۳)$$

که R تعداد فیلتر ها می باشد و عکس تبدیل به صورت تبدیل کسینوسی گسستهDCT می باشد.

- استفاده از DCT یک مجموعه بردار ناهمبسته را نتیجه می دهد که استفاده از کوواریانس های قطری را در مدل کردن توزیع بردار ویژگی، ممکن می سازد. [۱۷]

۳-۶ کاهش ویژگی :

- در کاربرد مسائل چندکلاسه، مواجه شدن با مسائلی که ده ها یا صد ها ویژگی دارند غیر معمول نیست.
- ظاهرا، به نظر می رسد که هر ویژگی حداقل در بعضی رده بندی ها، مفید هستند.
- به طور کلی، اگر عملکرد بدست آمده از یک مجموعه ویژگی نامناسب باشد، طبیعی به نظر می رسد که تعداد جدیدی ویژگی اضافه شود.
- اگرچه افزایش تعداد ویژگی ها باعث افزایش پیچیدگی کلاسه بند می شود، برای بهبود عملکرد سیستم، قابل قبول است.



شکل ۳ -۸ : در فضای یک بعدی x1 و یا فضای دو بعدی x1,x2 یک خطای بیزین غیر صفر وجود دارد. درحالی که در فضای x1,x2,x3 به دلیل چگالی های بدون اشتراک ،خطای بیزین به صفر رسیده است. از مرجع [۲۰]

مسئله ی ابعاد ویژگی ها :

- متاسفانه، به طور مکرر در عمل مشاهده شده است که، افزایش ویژگی های جدید باعث تضعیف عملکرد می شود تا بهبود عملکرد.
- دو نکته ی مهم باید مورد توجه قرار گیرد:
  - صحت کلاسه بندی چگونه با ابعاد ویژگی متاثر می گردد(بسته به مقدار داده های آموزشی)؟
  - پیچیدگی کلاسه بند چگونه با ابعاد ویژگی متاثر می شود؟
- دلایلی بالقوه برای افزایش خطا :
  - فرضیات نادرست در انتخاب مدل
  - خطای تخمین به دلیل تعداد محدود نمونه های آموزشی برای مشاهدات با ابعاد زیاد (overfitting)

- راه حل های بالقوه :

- کاهش ابعاد ویژگی ها و - ساده سازی تخمین

- کاهش ابعاد را می توان با :

- طراحی مجدد ویژگی ها

- انتخاب زیرمجموعه ای مناسب از میان ویژگی های موجود

- ترکیب ویژگی های موجود

بدست آورد.

- ساده سازی خطاهای تخمین را می توان با :

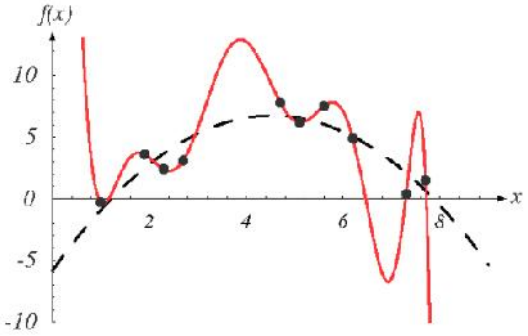
- فرض کوواریانس برابر برای همه ی کلاس ها (در مورد گوسی)

- استفاده از رگولارسیون

- استفاده از اطلاعات اولیه و یک تخمین Bayes

- استفاده از نوآوری هایی مانند استقلال شرطی و ...

بدست آورد.



شکل ۳- ۹ : مسئله ی داده های ناکافی مشابه مسائل مطرح شده در curve fitting می باشد. داده های آموزشی (نقطه های سیاه رنگ) از یک تابع درجه دوم بعلاوه ی نویز گوسی انتخاب شده اند. یک چند جمله ای درجه ۱۰، به طور کامل بر داده ها قرار می گیرد ولی ما یک چند جمله ای درجه دو را برای عمومیت بهتر مسئله، ترجیح می دهیم.[۲۰]

همه ی کلا سه بند های رایج می توانند با مشکل ابعاد ویژگی روبرو شوند. در حالی که ایجاد ارتباطی دقیق بین احتمال خطا، تعداد نمونه های آموزشی، تعداد ویژگی ها، و تعداد پارامترها بسیار مشکل است، تعدادی راهبرد پیشنهاد می شود :

- به طور کلی، استفاده ی حداقل تعداد نمونه های آموزشی ۱۰برابر تعداد ویژگی ها ($n/d > 10$) .
- هرچه کلاسه بند پیچیده تر باشد، نسبت تعداد نمونه ها به ابعاد ویژگی باید بزرگتر باشد.

کاهش ویژگی :

یک روش برای مقابله با مسئله ی ابعاد زیاد ، کاهش ابعاد ویژگی با ترکیب ویژگی ها می باشد.

- مسائل مطرح شده در کاهش ویژگی :
  - انتقالات خطی در برابر غیر خطی
  - استفاده و یا عدم استفاده از برچسب کلاس ها (بسته به میزان دسترسی داده های آموزشی)
  - هدف از آموزش دهی
    - حداقل کردن خطای کلاسه بندی(آموزش تمایزی)
    - حداقل کردن خطای بازسازی (PCA)
    - حداکثر کردن تفکیک پذیری کلاس ها (LDA)
    - حفظ جهت های قابل توجه (projection pursuit)
    - مستقل کردن ویژگی ها تا حد امکان (ICA)

ترکیب های خطی بدلیل اینکه به سادگی محاسبه می شوند، معمولا جذاب هستند. روش های خطی، داده های با ابعاد بالا را بر یک فضای با ابعاد کمتر تصویر می کند. فواید این تصویر کردن ها عبارت است از:

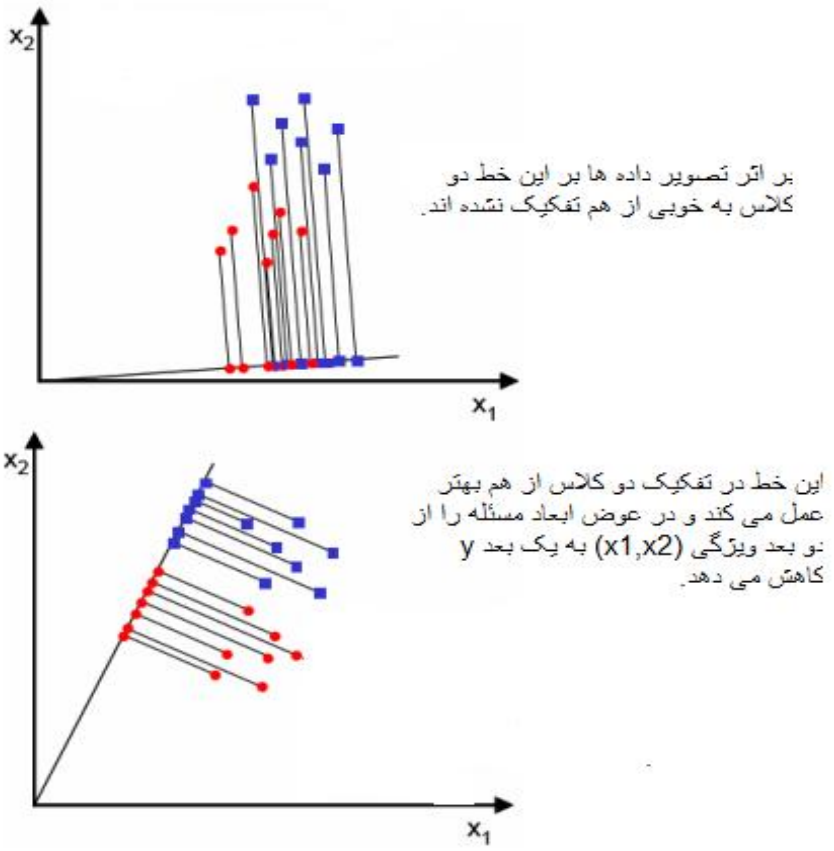
- کاهش پیچیدگی در کلاسه بندی و تخمین
- توانایی تست دیداری داده های چند متغیره در دو یا سه بعد.

- با داشتن $X \in R^d$ ، هدف یافتن یک تبدیل خطی A است که $y = A^T X \in R^{d'}$ که $d' < d$

- دو روش کلاسیک برای یافتن تبدیل خطی بهینه عبارتند از:
- Principal Components Analysis (PCA) : در پی یافتن تصویری است برای بهترین نمایش از داده ها در جهت کمترین مربعات.
- Linear Discriminant Analysis (LDA) : در پی یافتن تصویری است برای بهترین تفکیک داده ها در جهت کمترین مربعات. [۲۰]

۳-۶-۱ LDA:

LDA ماتریس تبدیل بهینه ای را می یابد که اکثر اطلاعاتی را که می توان برای تفکیک بین کلاس ها به کار برد، را نگه می دارد. در نتیجه، این تحلیل نیاز دارد که داده ها دارای برچسب کلاس مناسب باشند.



شکل ۳- ۱۰ : نمایشی از عملکرد روش LDA

برای فرموله کردن روند بهینه سازی به صورت ریاضی، باید بردار میانگین و ماتریس کوواریانس را برای هر کلاس بدست می آوریم:

$$\bar{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i \qquad (۲۰-۳)$$

$$\bar{W}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}(x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$$

و برای کل مجموعه ی داده ها (برای همه ی کلاس ها):

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T . \qquad (۲۱-۳)$$

در فرمول های بالا N تعداد کل داده های آموزشی و Nj تعداد داده های آموزشی از کلاس j . (J کلاس وجود دارد.)

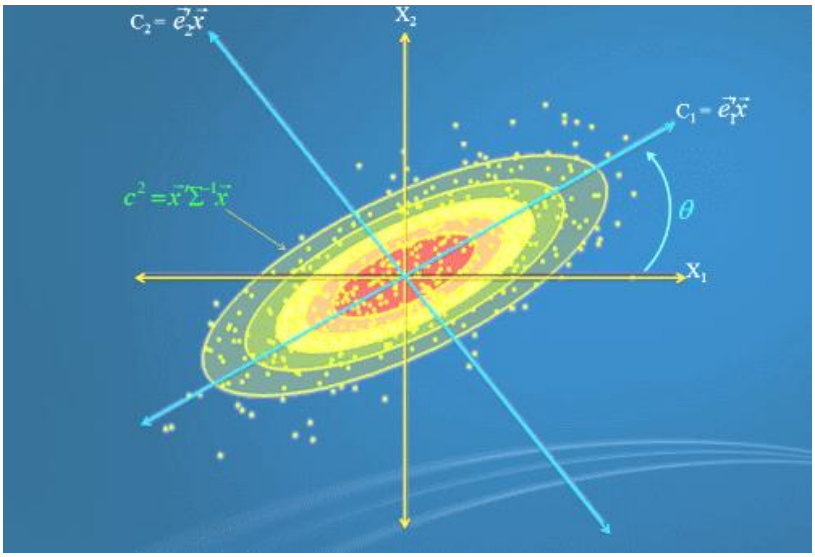
$$\sum_{j=1}^{J} N_j = N . \qquad (۲۲-۳)$$

با این تعاریف، به راحتی می توانیم شرط بهینه سازی را فرموله کنیم :

$$\hat{\theta} = \arg\max_{\theta_p} \frac{|\theta_p^T \bar{T} \theta_p|}{|\theta_p^T \bar{W} \theta_p|},$$
$$\bar{W} = \frac{1}{N}\sum_{j=1}^{J} N_j \bar{W}_j . \qquad (۲۳-۳)$$

اگر چه در نگاه اول این شرط پیچیده به نظر می رسد، به راحتی قابل در ک است. صورت کسر کوواریانس داده های آموزشی در فضای ویژگی تبدیل یافته است. مخرج کسر کوواریانس متوسط در هر کلاس در فضای ویژگی تبدیل یافته است. در نتیجه، این شرط سعی دارد که فاصله ی بین کلاس ها حداکثر شود در حالی که همزمان اندازه ی هر کدام از کلاس ها حداقل گردد. این دقیقا همان چیزی است که می خواستیم زیرا این شرط تضمین می کند که اکثر اطلاعات تفکیکی در فضای ویژگی تبدیل یافته حفظ می شوند. با توجه به فرمول بالا ماتریس بهینه را می توان به راحتی بدست آورد. نتیجه ی یک مشتق گیری طولانی نشان می دهد که $\hat{}$ شامل آن بردارهای ویژه ای از $\overline{W}^{-1}\overline{T}$ هستند که مربوط به p تا از بزرگترین مقادیر ویژه می باشند. [۲۱]

۳-۶- ۲ تحلیل مؤلفه‌های اصلی (PCA)



شکل ۳ - ۱۱ : نقاط سبز رنگ، نمونه‌هایی از توزیع نرمال دومتغیره‌اند و محور آبی رنگ، مختصات جدید در راستای قرار گرفتن بیشترین تغییرات نمونه بر روی مؤلفه‌های اصلی است.از مرجع[۲۲]

تحلیل مولفه‌های اصلی (Principal Component Analysis - PCA) تبدیلی درفضای برداری است، که غالباً برای کاهش ابعاد مجموعه ی داده ها مورد استفاده قرار می‌گیرد. تحلیل مولفه‌های اصلی در سال ۱۹۰۱ توسط کارل پیرسون ارائه شد. این تحلیل شامل تجزیه مقدارهای ویژه ی ماتریس کوواریانس می‌باشد.

تحلیل مولفه‌های اصلی در تعریف ریاضی یک تبدیل خطی متعامد است که داده را به دستگاه مختصات جدید می‌برد به طوری که بزرگترین واریانس داده بر روی اولین محور مختصات، دومین بزرگترین واریانس بر روی دومین محور مختصات قرار می‌گیرد و همین طور برای بقیه. تحلیل مولفه‌های اصلی می‌تواند برای کاهش ابعاد داده مورد استفاده قرار بگیرد، به این ترتیب مولفه‌هایی از مجموعه داده را که بیشترین تاثیر در واریانس را دارند حفظ می‌کند. برای ماتریس داده X^1با میانگین تجربی صفر، که هر سطر یک مجموعه مشاهده و هر ستون داده‌های مربوط به یک شاخصه است، تحلیل مولفه‌های اصلی به صورت زیر تعریف می‌شود:

$$Y^T = X^T W = V\Sigma \qquad (۲۴-۳)$$

به طوری که $V\Sigma W^T$ تجزیه ی مقدارهای منفرد ماتریس X^T می‌باشد.

محدودیتهای تحلیل مولفه‌های اصلی

استفاده از تحلیل مولفه‌های اصلی منوط به فرضهایی است که باید در نظر گرفته شود. از جمله:

- فرض خطی بودن (ما فرض می کنیم مجموعه داده ترکیب خطی پایه‌هایی خاص است.)
- فرض بر این که میانگین و کواریانس از نظر احتمالاتی قابل اتکا هستند.
- فرض بر این که واریانس شاخصه اصلی داده است.

محاسبه مولفه‌های اصلی با استفاده از ماتریس کواریانس:

بر اساس تعریف ارائه شده از تحلیل مولفه‌های اصلی، هدف از این تحلیل انتقال مجموعه داده X با ابعاد M به داده Y با ابعاد L است. بنابرین فرض بر این است که ماتریس X از بردارهای $X_1 \cdots X_N$تشکیل شده است که هر کدام به صورت ستونی در ماتریس قرار داده شده است. بنابرین با توجه به ابعاد بردارها (M) ماتریس داده‌ها به صورت $M \times N$است.

محاسبه میانگین تجربی و نرمال سازی داده‌ها

نتیجه میانگین تجربی، برداری است به صورت:

$$u[m] = \frac{1}{N} \sum_{i=1}^{N} X[m, i] \qquad (۲۵-۳)$$

که به طور مشخص میانگین تجربی روی سطرهای ماتریس اعمال شده است.

سپس ماتریس انحراف از میانگین به صورت (۳-۲۶) به دست می‌آید:

$$B = X - uh \qquad (۲۶-۳)$$

که h برداری با اندازه $1 \times N$با مقدار ۱ در هرکدام از درایه‌ها است.

محاسبه ماتریس کواریانس

ماتریس کواریانس C با ابعاد $M \times M$ به صورت (۳-۲۷) به دست می‌آید:

$$C = \mathbb{E}[B \otimes B] = \mathbb{E}[B \cdot B^*] = \frac{1}{N} B \cdot B^* \qquad (۲۷-۳)$$

که، $\mathbb{E}$ میانگین حسابی است. $\otimes$ ضرب خارجی است. B^* ماتریس ترانهاده مزدوج ماتریس B است.

محاسبه مقادیر ویژه ماتریس کواریانس و بازچینی بردارهای ویژه

در این مرحله، مقادیر ویژه و بردارهای ویژه ماتریس کواریانس، C، به دست می‌آید.

$$V^{-1}CV = D \qquad (۲۸-۳)$$

V ماتریس بردارهای ویژه و D ماتریس قطری است که درایه‌های قطر آن مقادیر ویژه هستند. آنچنان که مشخص است، هر مقدار ویژه متناظر با یک بردار ویژه است. به این معنا که ماتریس V ماتریسی $M \times M$ است که ستونهای آن بردارهای ویژه می‌باشند و بردار ویژه V_q در ستون qام قرار دارد و مقدار ویژه qام یعنی درایه ی $\lambda_q = D_{q,q}$ متناظر با آن است. بازچینی بردارهای ویژه بر اساس اندازه ی مقادیر ویژه متناظر با آنها صورت می‌گیرد. یعنی بر اساس ترتیب نزولی مقادیر ویژه، بردارهای ویژه بازچینی می‌شوند. یعنی $p \leq q \Rightarrow \lambda_p \leq \lambda_q$

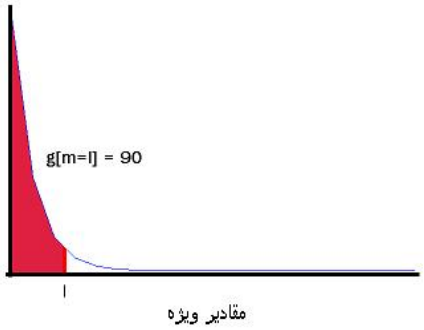
انتخاب زیرمجموعه‌ای از بردارهای ویژه به عنوان پایه

انتخاب زیرمجموعه‌ای از بردارهای ویژه با تحلیل مقادیر ویژه صورت می‌گیرد. زیرمجموعه نهایی با توجه به بازچینی مرحله قبل به صورت $V_1 \cdots V_l$ انتخاب می‌شود. در اینجا می‌توان از انرژی تجمعی استفاده کرد که طبق آن

$$g[m] = \sum_{q=1}^{m} \lambda_q \qquad (۲۹-۳)$$

انتخاب l باید به صورتی باشد که حداقل مقدار ممکن را داشته باشد و در عین حال g مقدار قابل قبولی داشته باشد. به طور مثال می‌توان حداقل l را انتخاب کرد که $g[m = l] \leq 90\%$ (۳۰-۳)

بنابرین خواهیم داشت: $W[p,q] = V[p,q], p = 1 \dots M, q = 1 \dots l$



شکل ۳- ۱۲: تحلیل مقادیر ویژه ماتریس کواریانس. محور عمودی g است. از مرجع[۲۲]

انتقال داده به فضای جدید

برای این کار ابتدا تبدیلات زیر را انجام می دهیم:

ماتریس $S_{M,1}$ انحراف معیار مجموعه داده است که می‌تواند به صورت (۳-۳۱) به دست بیاید:

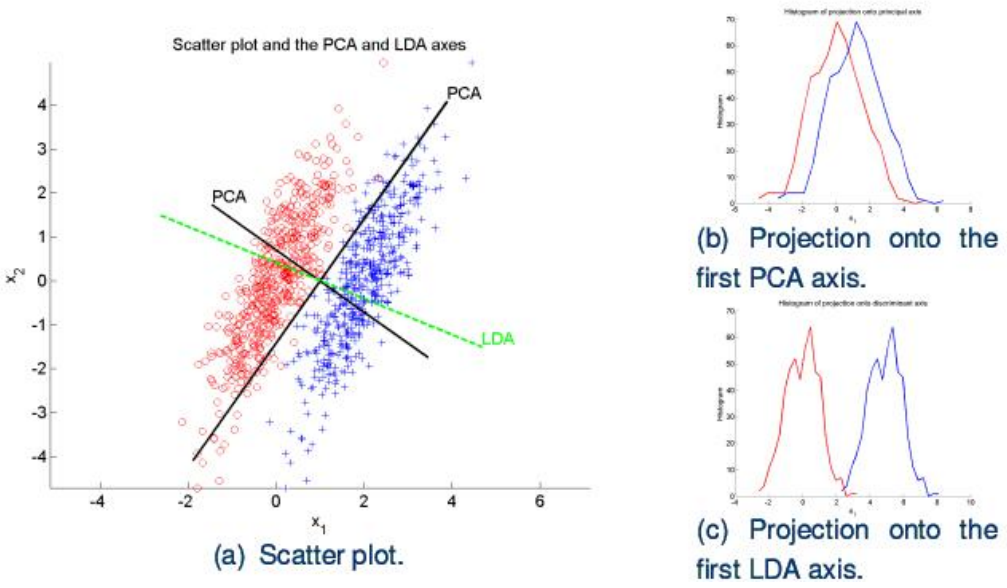
$$s[i] = \sqrt{C[i,i]} \qquad (۳۱-۳)$$

سپس داده به صورت $Z = \frac{B}{s}$ تبدیل می‌شود. که ماتریسهای C و Bدر بالا توضیح داده شده اند.

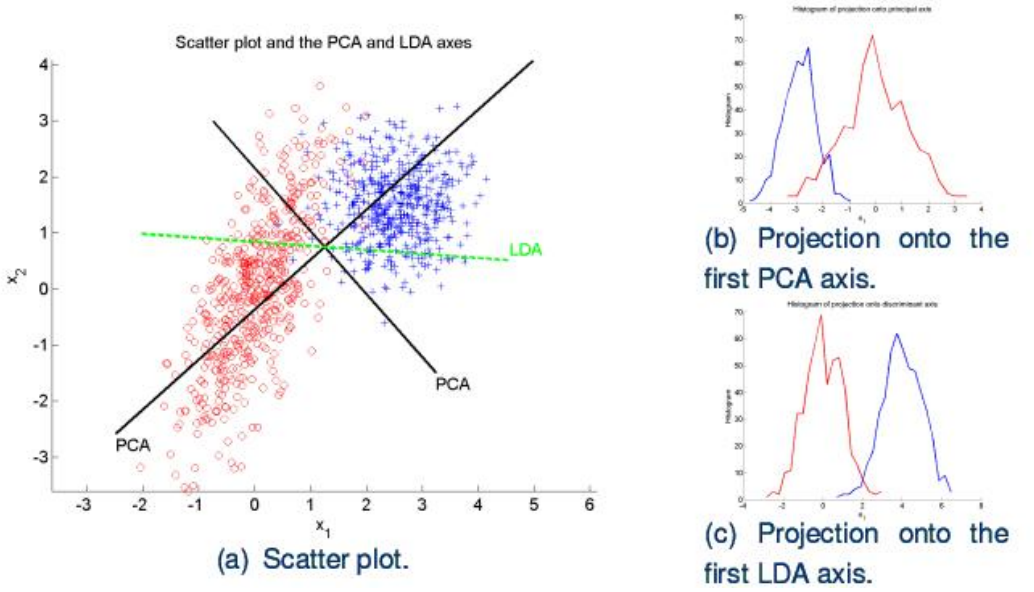
داده‌ها می‌توانند به ترتیب (۳-۳۲) به فضای جدید برده شوند: [۲۲]

$$Y = W^*.Z \qquad (۳۲-۳)$$

۳-۶-۳ مقایسه ی تصویری روش PCA و LDA :



شکل ۳-۱۳: از مرجع [۲۰]



شکل ۳ -۱۴: از مرجع [۲۰]

در شکل های ۳-۱۳ و ۳-۱۴ نمودار پراکندگی و محورهای LDA , PCA برای نمونه های دو متغیره برای دو کلاس نشان داده شده است. هیستوگرام مربوط به تصویر نمونه ها بر اولین محور LDA تفکیک بهتری از تصویرنمونه ها بر اولین محور PCA نشان می دهد.

۳-۷ روش های تعمیم ویژگی MFCC برای کاربردهای مختلف :

۳-۷-۱ استخراج ویژگی از ضرایب قسمت بالایی خود همبستگی (higher – lag autocorrelation) ، در شناسایی گفتار پایدار:

استخراج ویژگی از ضرایب قسمت بالایی خود همبستگی، یک روش استخراج ویژگی که در نویز پس زمینه ی جمع پذیر پایدار است، برای شناسایی گفتار اتوماتیک می باشد. نویز پس زمینه ضرایب خودهمبستگی سیگنال گفتار را اکثرا درlower-time lags خراب می کند، در حالی که ضرایب خود همبستگی higher-time lags کمترین تاثیر را می پذیرند؛ به همین دلیل، این روش ضرایب خود همبستگی lower-time lags را کنار می گذارد و تنها ضرایب خودهمبستگی higher-time lags را برای تخمین طیفی به کار می برد. طیف دامنه ی رشته ی خودهمبستگی higher-lag پنجره گذاری شده، به عنوان تخمینی از طیف توان سیگنال گفتار به کار می رود.

این تخمین طیف توان، پیش تر نیز به کار می رفته است. مثلا در MFCC، توسط بانک فیلتر Mel ، عمل لگاریتم و تبدیل کسینوسی گسسته، این کار صورت می گرفته است.

این ضرایب cepstral به عنوان .autocorrelation Mel frequency cepstral coe cients (AMFCCs) می باشند.

یک شناسنده ی گفتار در یک محیط صوتی آموزش داده می شود و معمولا در محیط دیگری آزمایش می شود. بنابراین، همواره عدم تطابقی بین محیط آموزش و آزمایش وجود دارد. این عدم تطابق، باعث کاهش عملکرد شدیدی در شناسایی گفتار می شود. یک عامل مهم و موثر در عدم تطابق بین دو محیط آموزش و آزمایش، نویز پس زمینه ی جمع پذیر (ناهمبسته با گفتار) است. تعدادی روش برای غلبه بر این عدم تطابق محیطی پیشنهاد شده است.

- روش های استخراج ویژگی پایدار - روش های تقویت گفتار
- روش های جبران ویژگی - روش های چند-باندی
- روش های ویژگی ناپیدا(از دست رفته) - روش های جبران کانال

این روش در [۲۳] مورد بررسی قرار گرفته است که بر پایه ی روش استخراج ویژگی ای است که بتواند اعوجاج نویز پس زمینه ی جمع پذیر را به صورت پایداری تحمل کند.

یک قاب از سیگنال نویزی $x(n), n= 0,1,\ldots,N$ ، که N طول قاب (برحسب تعداد نمونه ها)می باشد را داریم.

طیف توان سیگنال نویزی $x(n)$ را با $Pxx(\omega)$ و تابع خودهمبستگی اش را با $r_{xx}(n)$ نشان میدهیم.

همین طور برای سیگنال پاک (بدون نویز) $s(n)$ ، نماد های $Pss(\omega)$ و $r_{ss(n)}$ را به کار می بریم.

برای سیگنال نویز d(n) ، این نمادها به ترتیب به صورت $P_{dd}(\mu)$ و $r_{dd(n)}$ هستند.

$$x(n) = s(n) + d(n), \qquad (۳۳-۳)$$

هدف، استخراج ویژگی های شناسایی از سیگنال گفتار نویزی x(n) به صورتی است که آنها مشخصه های طیفی سیگنال پاک s(n) را به طور دقیق بدست آورند و کمترین تاثیر را از نویز بپذیرند.

در استخراج ویژگی MFCC ، سیگنال نویزی x(n) در طی مراحل زیر پردازش می شود:

۱- تحلیل زمان-کوتاه فوریه ی x(n) با بکارگیری پنجره ای با طول محدود (مانند یک پنجره ی Hamming به طول ۳۲ میلی ثانیه) و استفاده از روش periodogram برای محاسبه ی تخمین طیف توان $P^{\wedge}_{xx}(\omega)$ سیگنال x(n) .

۲- اعمال بانک فیلتر mel بر $P^{\wedge}_{xx}(\omega)$ برای بدست آوردن انرژی های بانک فیلتر

۳- محاسبه ی تبدیل کسینوسی گسسته ی DCT انرژی های بانک فیلتر لگاریتمی برای بدست آوردن MFCC

ویژگی های MFCC در شناسایی سیگنال های پاک به طور قابل قبولی عمل می کنند ولی عملکرد آنها برای گفتارهای نویزی بسیار ضعیف می باشد. این اتفاق به این دلیل است که تخمین طیف توانی که بر اساس periodogram می باشد و در محاسبه ی MFCC به کار می رود، به شدت تحت تاثیر نویز پس زمینه ی جمع پذیر می باشد و در نتیجه عملکرد شناسایی MFCC را برای گفتار نویزی کاهش می دهد.

در[۲۳] ، از پردازش در حوزه ی خودهمبستگی برای تخمینِ پایداری از طیف توان برای گفتار نویزی، استفاده شده است. تابع خودهمبستگی یک سیگنال از طریق تبدیل فوریه به طیف توان سیگنال مرتبط می شود و دو خصوصیت قابل توجه زیر را دارد:

۱- خصوصیت جمع پذیری:اگر دو سیگنال ناهمبسته باشند، تابع خودهمبستگی مجموع آنها برابر است با مجموع توابع خودهمبستگی آنها. در نتیجه :

$$r_{xx}(n) = r_{ss}(n) + r_{dd}(n) \qquad (۳۴-۳)$$

که s(n) و d(n) سیگنال های نا همبسته هستند.

۲- خصوصیت پایداری :

تابع خودهمبستگی سیگنال نویز اتفاقی سفید در همه جا به جز صفرامین time-lag ، صفر می باشد و برای سیگنال های نویز پهن باند، عمدتا به lower-time lags محدود می شود و برای higher-time lags بسیار کوچک و یا صفر است. در نتیجه، نویز جمع پذیر d(n) ، higher- lag های تابع خودهمبستگی را تحت تاثیر قرار نمی دهد. پس ضرایب خودهمبستگی higher- lag نسبتا در برابر اعوجاج نویز جمع پذیر پایدار هستند.

به دلیل این خصوصیات جذاب، پردازش حوزه ی خود همبستگی در گذشته برای تخمین طیفی خود بازگشتی autoregressive (AR) (و یا تحلیل پیش بینی خطی - linear prediction (LP)) سیگنال های نویزی استفاده می شده است.

اولین تلاش در این مسیر بر اساس استفاده از معادلات Yule–Walker مرتبه – بالا بود. که در ضرایب خود همبستگی که در مجموعه ی معادله ی Yule–Walker درگیر هستند، ضریب zero-lag حذف می شود.

سایر روش های مشابه، یا از ضریب zero-lag صرف نظر می کنند و یا سهم تعداد محدودی (کمی) از اولین ضرایب را کاهش می دهند.

همه ی این روش ها بر اساس مدل تمام-قطب قسمت علی رشته ی خودهمبستگی سیگنال x(n) ، می باشند.

به دلیل اینکه اعوجاج نویز پهن باند تنها بر ضرایب خود همبستگی lower-lag تاثیر می گذارد، از آنها صرف نظر می کنیم و تنها از ضرایب خودهمبستگی higher-lag (>2 ms) بهره می گیریم.

ضرایب خودهمبستگی higher- lag سیگنال x(n) را در یک تابع پنجره ی مناسب ضرب می کنیم. و طیف دامنه ی رشته ی پنجره گذاری شده، به عنوان تخمینی از طیف توان سیگنال محاسبه می شود. این روشِ تخمین طیف توان به عنوان روش higher-lag autocorrelation spectrum estimation (HASE) معرفی می شود.

گام های ۲و۳ که در روند استخراج ویژگی MFCC در بالا توضیح داده شد، بر این تخمین طیف توان مذکور، برای بدست آمدن ویژگی های ضرایب cepstral اعمال می شود.

این ویژگی ها، autocorrelation Mel frequency cepstral coe □ cients (AMFCCs) نامیده می شوند.

باید توجه شود که روش HASE برای تخمین طیف توان (برای محاسبه ی AMFCC) بر اساس خواص جمع پذیری و پایداری تابع خودهمبستگی سیگنال می باشد. این دو خصوصیت تنها در حالت asymptotic (با فرض ergodicity) معتبر می باشد. یعنی در زمانی که ضرایب خود همبستگی از سیگنال های نا محدود (بسیار طولانی) محاسبه می شوند.

برای محاسبه ی MFCC ، تنها تحلیل زمان- کوتاه سیگنال گفتار (با طول سیگنال حدود 32 ms) در نظر گرفته شده است. این تابع خودهمبستگی آن دو خصوصیت را دقیقا تامین نمی کند. تاثیر تحلیل زمان- کوتاه بر این خواص و عملکرد شناسایی گفتار بحث خواهد شد.

بعضی خواص تابع خود همبستگی زمان - کوتاه

تابع خود همبستگی یک سیگنال شامل اطلاعات مشابهی از سیگنال است که در طیف توان سیگنال جود دارد. در حوزه ی طیف توان، اطلاعات به صورت تابعی از فرکانس ظاهر می شوند و در حوزه ی خود همبستگی، اطلاعات به صورت تابعی از زمان می باشند. بعضی از خواص تابع خودهمبستگی زمان – کوتاه مرتبط با روش AMFCC توضیح داده می شود.

بدلیل اینکه روش AMFCC ضرایب خودهمبستگی zeroth , lower-lag را حذف می کند و تنها از ضرایب خودهمبستگی higher-lag برای تخمین طیفی استفاده می کند، دانستن این نکته که این ضرایب چگونه شامل اطلاعات طیفی ضروری برای شناسایی گفتار می باشند، لازم است.

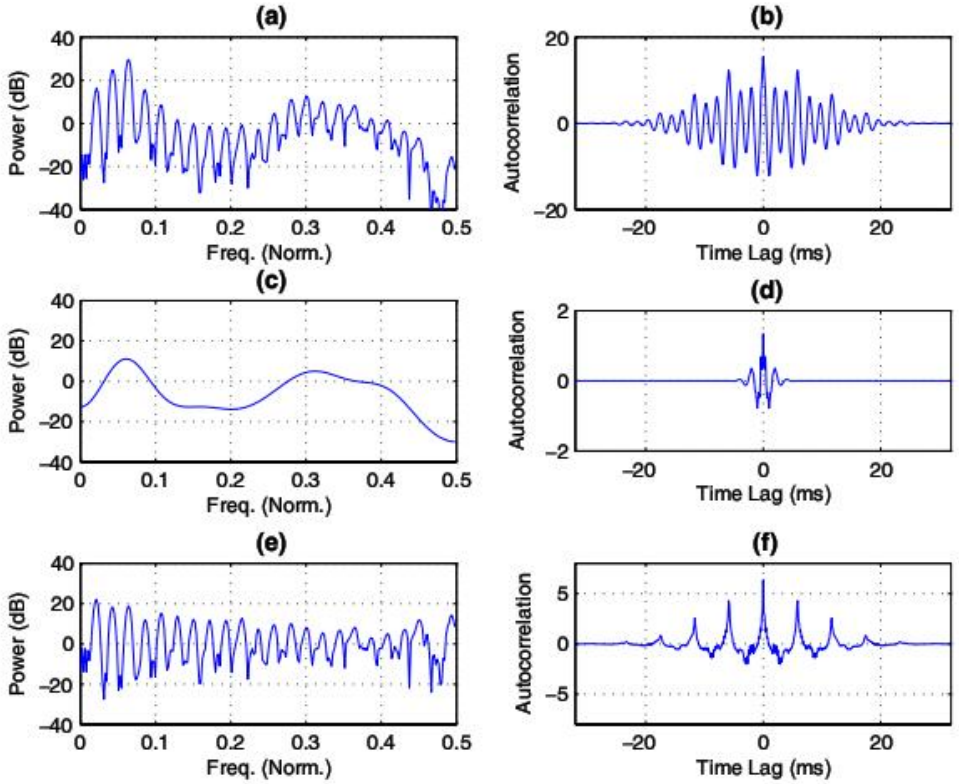
هم چنین، بدلیل اینکه AMFCC به عنوان روش استخراج ویژگی پایدار بر پایه ی اینکه اعوجاج نویز جمع پذیر بیشترین ضرایب خود همبستگی اش در نزدیکی lower-time-lags می باشد و ضرایب خودهمبستگی higher-lag آن صفر هستند (یا بسیار کوچک اند)، پیشنهاد شده است، می خواهیم بدانیم کدام نوع از سیگنال های نویز این خصوصیت را دارند و تا چه حدی این خصوصیت را تامین می کنند.(تحت شرایط تحلیل زمان- کوتاه)

سیگنال های گفتار

مدل منبع- سیستم رایج برای تولید گفتار، سیگنال گفتار را به عنوان خروجی یک سیستم خطی، متغیر با زمان معرفی می کند که یا با یک منبع نویز سفید (برای گفتار بی صدا) و یا با یک منبع قطار ضربه ی متناوب (برای گفتار صدادار) تحریک می شود. برای شناسایی گفتار، معمولا مایلیم که پاسخ دامنه ی سیستم متغیر با زمان را به عنوان تابعی از زمان استخراج کنیم.(با این فرض که پاسخ دامنه ی سیستم، اطلاعات گفتار کافی را برای شناسایی دقیق داشته باشد.)

برای روشن ساختن اینکه چگونه ضرایب خود همبستگی higher-lag شامل اطلاعات لازم برای شناسایی گفتار هستند ، یک قاب 32 ms از یک سیگنال پاک صدادار با صدای / ey / از یک خانم را در نظر می گیریم.

طیف توان و تابع خود همبستگی را در شکل ۳-۱۵ (a,b) نشان داده شده است.(تابع خودهمبستگی از معکوس تبدیل فوریه ی طیف توان بدست می آید. بنابراین، یک تخمین بایاس شده از تابع خود همبستگی برای یک قاب فراهم می کند.)



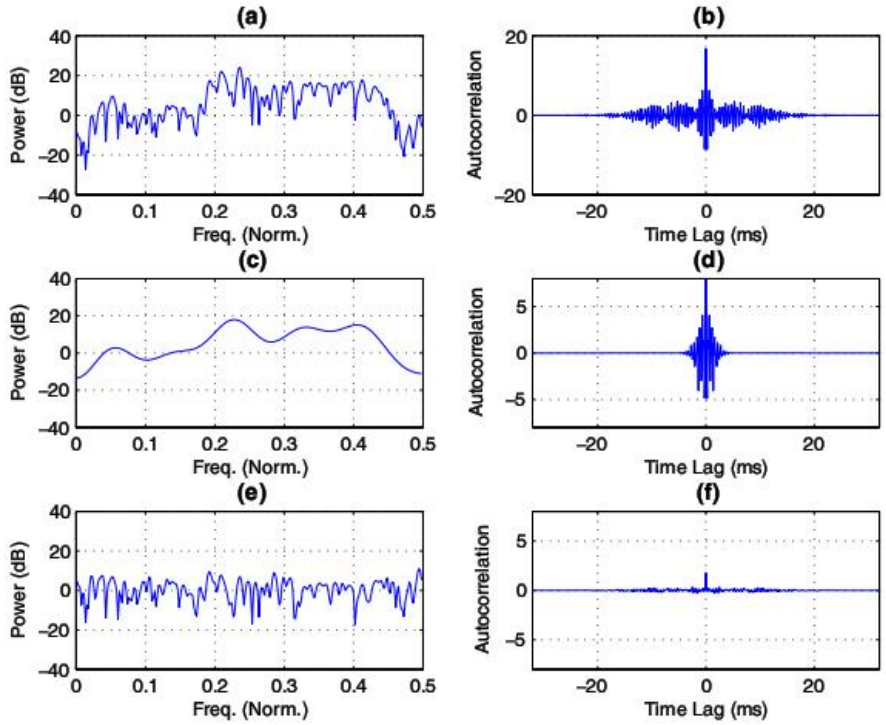
شکل ۳-۱۵ : طیف توان و تابع خودهمبستگی برای یک قاب ۳۲ میلی ثانیه ای از گفتار 'ey' voiced یک خانم . (ادامه توضیح شکل در صفحه بعد) a) سیگنال طیف توان بر حسب dB . b) رشته ی خودهمبستگی مربوط به طیف سیگنال در a . c) منحنی صاف (روان) شده ی طیف بر حسب dB که با نگه داشتن اولین ۱۲ ضریب cepstral مربوط به طیف سیگنال در a محاسبه شده است. d)رشته ی خودهمبستگی مربوط به طیف هموار شده در c

e) طیف توان محرکه بر حسب dB ، که با کم کردن طیف هموار شده لگاریتمی در c از طیف سیگنال لگاریتمی در a تخمین زده شده است. f)رشته خود همبستگی مربوط به طیف محرک در e

برای نشان دادن اجزای سیستم و منبع برای این سیگنال، از تحلیل homomorphic استفاده می کنیم و با نگه داشتن ۱۲ ضریب cepstral نخست، طیف صاف و روان شده (smooth) را محاسبه می کنیم.

این به عنوان تخمینی از طیف توان جزء سیستم به کار برده می شود. باقی مانده ی طیف توان (که با تقسیم طیف توان سیگنال بر طیف smooth) به عنوان تخمین طیف توان منبع استفاده می شود. (شکل ۳- ۱۵ (c,e)) توابع خود همبستگی آنها در شکل ۳-۱۵ (d,f) نشان داده شده است. در تصاویر ۳-۱۵ (d,f) می توان مشاهده کرد که تابع خود همبستگی جزء سیستم تنها به lower-lags محدود می شود و تابع خود همبستگی جزء منبع تناوب سیگنال گفتار صدادار را نشان می دهد. بدلیل اینکه تابع خود همبستگی سیگنال (شکل ۳-۱۵ b) کانولوشن توابع خودهمبستگی

سیستم و منبع می باشد، این تابع به صورت متناوب است که اطلاعات طیف smooth شده متناوبا در همه ی بازه های زمانی تکرار می شود و مقادیر آن برای higher – lag ها بزرگ می باشد. به همین دلیل، می توان از قسمت پایین تابع خود همبستگی سیگنال صرف نظر کرد و هنوز بتوانیم تخمین خوبی از طیف توان، با استفاده از ضرایب خودهمبستگی higher-lag برای شناسایی گفتار داشته باشیم. تحلیل مشابهی برای یک قاب 32 ms از یک سیگنال پاک بی صدا از صدای یک خانم با صدای /s/ در شکل ۳-۱۶ نشان داده شده است.



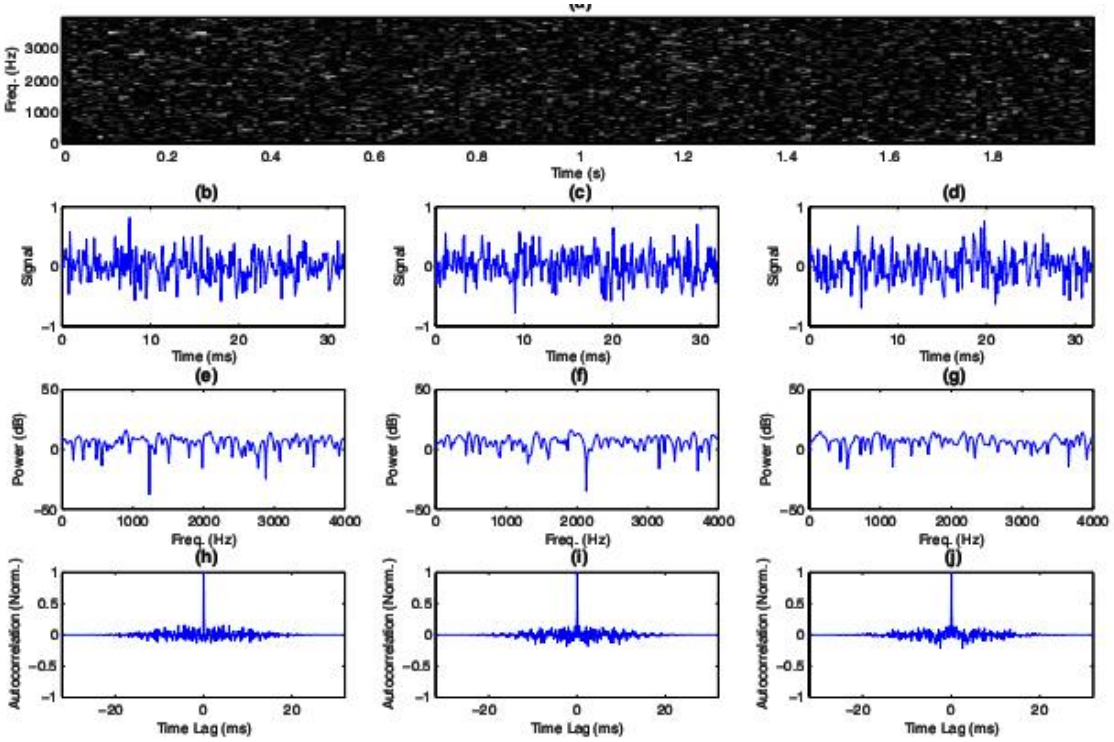
شکل ۳-۱۶ : طیف توان و تابع خودهمبستگی برای یک قاب ۳۲ میلی ثانیه ای از گفتار 's' unvoiced یک خانم . a) سیگنال طیف توان بر حسب dB . b) رشته ی خودهمبستگی مربوط به طیف سیگنال در a . (c) منحنی صاف (روان) شده ی طیف بر حسب dB که با نگه داشتن اولین ۱۲ ضریب cepstral مربوط به طیف سیگنال در a محاسبه شده است. d)رشته ی خودهمبستگی مربوط به طیف هموار شده در c . (e) طیف توان محرکه بر حسب dB ، که با کم کردن طیف هموار شده لگاریتمی در c از طیف سیگنال لگاریتمی در a تخمین زده شده است. f)رشته خود همبستگی مربوط به طیف محرک در e

در اینجا، تابع خود همبستگی جزء سیستم (۳-۱۶ d) به lower-lag محدود شده است و تابع خود همبستگی جزء منبع (۳-۱۶ f) متناوب نمی باشد و در higer-lag نیز دارای مقادیر غیر صفر است. در نتیجه ی کانولوشن، تابع خودهمبستگی سیگنال (۳-۱۶ b) شامل اطلاعات طیفی درباره ی جزء سیستم در higher –lag است (اگرچه به اندازه ی آنچه در مورد سیگنال گفتار صدادار دیدیم نمی باشد). بنابراین، می توان، ضرایب خودهمبستگی higher – lag سیگنال را برای تخمین طیف توان گفتار بی صدا نیز بکار بریم.

سیگنال های نویز

۱- سیگنال نویز سفید اتفاقی

در تئوری، تابع خود همبستگی باید برای همه ی lag ها به جز صفرامین lag، صفر باشد. می خواهیم بدانیم که آیا این موضوع در تحلیل زمان – کوتاه هم صدق می کند یا خیر ؟ 2s از نویز سفید گوسی مصنوعی (با کامپیوتر تولید شود) را در نظر می گیریم و تحلیل زمان-کوتاه (با پنجره hamming) را در هر 10ms با قابی به طول 32ms را انجام می دهیم.



شکل ۳- ۱۷ : تحلیل زمان کوتاه سیگنال نویز سفید اتفاقی مصنوعی با استفاده از قاب های ۳۲ میلی ثانیه ای.

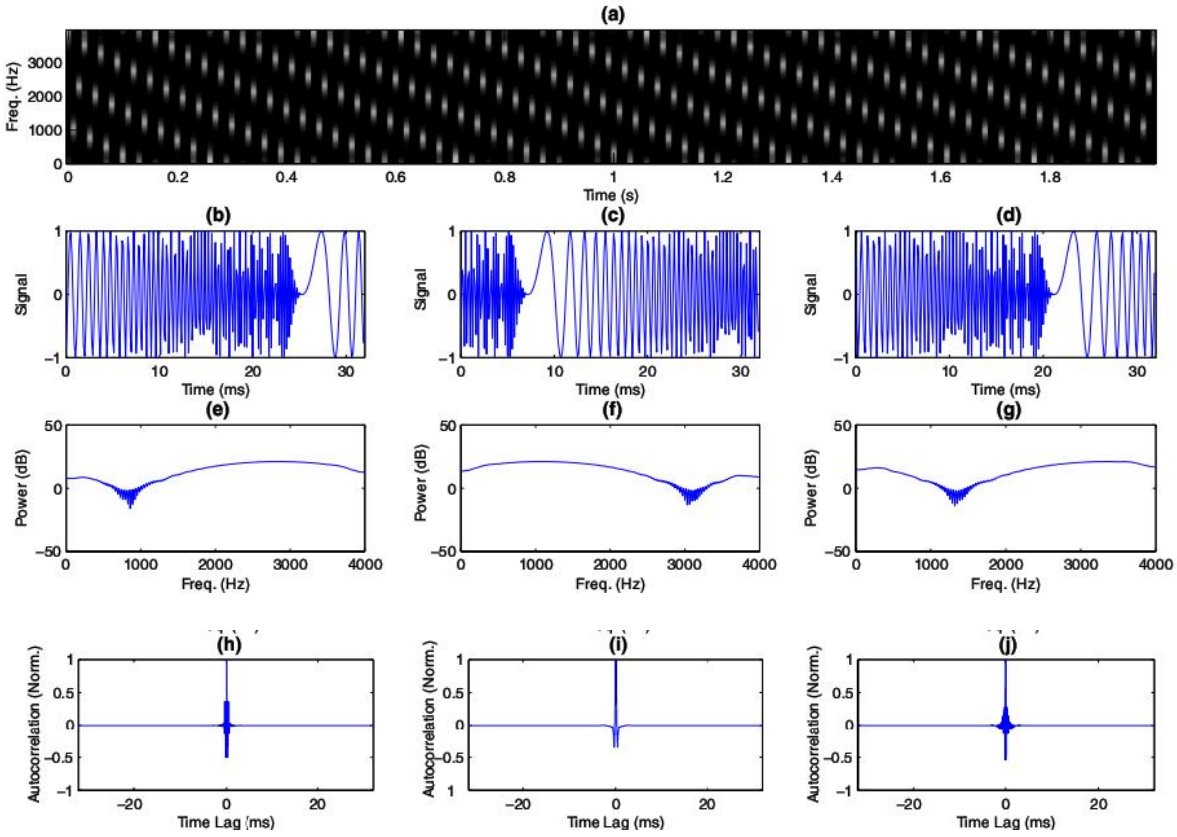
a) طیف نگاره ی یک نمونه ی ۲ ثانیه ای از سیگنال طویل نویز b-d) شکل موج قاب های نویز به ترتیب در ۰.۵ و ۱ و ۱.۵ ثانیه. e-g) طیف های توان قاب های نشان داده شده در b-d به ترتیب . h-j) رشته های خودهمبستگی مربوط به طیف های توان در e-g به ترتیب

همان طور که انتظار می رفت، ضرایب خودهمبستگی higher –lag از لحاظ دامنه از صفرامین ضریب خود همبستگی کوچک تر اند، ولی به دلیل تحلیل زمان – کوتاه مقادیر غیر صفر دارند.

به طور مشابه، می توان نشان داد سیگنال های نویز اتفاقی پهن باند ایستا (۶ نویز سفید اتفاقی را می توان به عنوان نوع خاصی از نویز اتفاقی پهن باند در نظر گرفت که باند آن از صفر تا نصف فرکانس نمونه برداری می باشد) این خصوصیت را که دامنه ی ضرایب خودهمبستگی lower – lag بزرگ و ضرایب higher – lag بسیار کوچک هستند را تامین می کند.

۲- نویز chrip (جیر جیر)

مانند صدای آژیر های خطر و جیک جیک پرندگان



شکل ۳- ۱۸ : تحلیل زمان کوتاه سیگنال نویز chrip مصنوعی با استفاده از قاب های ۳۲ میلی ثانیه ای.

a) طیف نگاره ی یک نمونه ی ۲ ثانیه ای از سیگنال طویل نویز

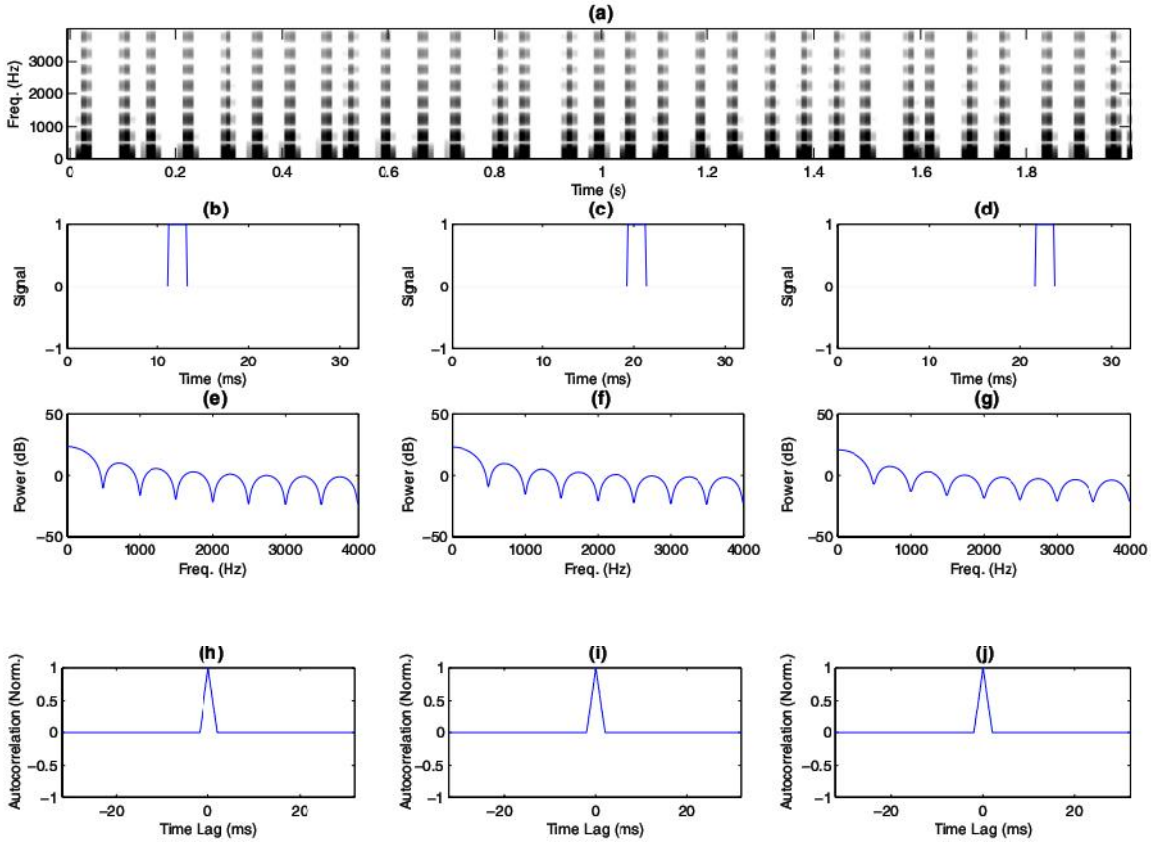
b-d) شکل موج قاب های نویز به ترتیب در ۰.۵ و ۱ و ۱.۵ ثانیه.

e-g) طیف های توان قاب های نشان داده شده در b-d به ترتیب.

h-j) رشته های خودهمبستگی مربوط به طیف های توان در e-g به ترتیب

۳ – نویز ضربه ای

مانند صدای تایپ کردن بر روی صفحه کلید ، صدای شلیک تیربار



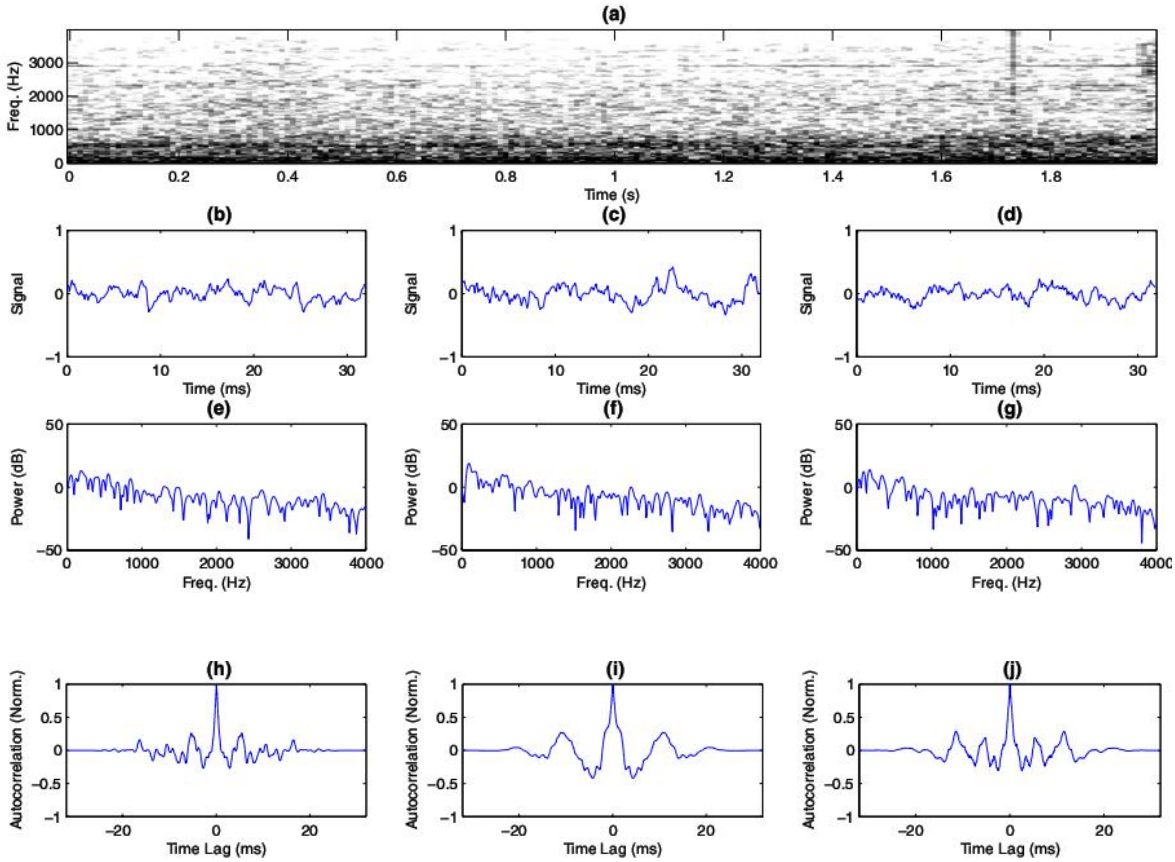
شکل ۳- ۱۹ : تحلیل زمان کوتاه سیگنال نویز ضربه ای مصنوعی با استفاده از قاب های ۳۲ میلی ثانیه ای.

a) طیف نگاره ی یک نمونه ی ۲ ثانیه ای از سیگنال طویل نویز

b-d) شکل موج قاب های نویز به ترتیب در ۰.۵ و ۱ و ۱.۵ ثانیه. e-g) طیف های توان قاب های نشان داده شده در b-d به ترتیب. h-j) رشته های خودهمبستگی مربوط به طیف های توان در e-g به ترتیب

همان طو ر که مشاهده می شود، اکثر ضرایب خودهمبستگی در نزدیکی lower –time lag می باشند و در higher – lag صفر می باشند.

تا اینجا از نویز های مصنوعی استفاده شده است؛ و دیدیم که شرایط مناسبی برای استفاده از روش AMFCC را فراهم می کنند. اکنون نویز های واقعی را مد نظر قرار می دهیم. مثلا نویز اتومبیل، که از پایگاه داده ی Aurora-2 بدست آمده است. و تحلیل زمان – کوتاه آن را در نظر می گیریم.(شکل ۳-۲۰)



شکل ۳-۲۰ : تحلیل سیگنال نویز اتومبیل با استفاده از قاب های ۳۲ میلی ثانیه ای.

a) طیف نگاره ی یک نمونه ی ۲ ثانیه ای از سیگنال طویل نویز . b-d) شکل موج قاب های نویز به ترتیب در ۰.۵ و ۱ و ۱.۵ ثانیه. e-g) طیف های توان لگاریتمی قاب های نشان داده شده در b-d به ترتیب، که با استفاده از پنجره ی hamming محاسبه شده اند . h-j) رشته های خودهمبستگی مربوط به طیف های توان در e-g به ترتیب

از شکل نشان داده می شود که نویز اتومبیل نوعی نویز پهن باند (پایین گذر) می باشد. دامنه ی ضرایب خودهمبستگی در $lower - lag$ بزرگتر از $higher - lag$ ها هستند ولی ضرایب در $higher - lag$ ها هم به اندازه ی نویز های مصنوعی (ساختگی) قبل، کوچک نمی باشند. بنابراین نویز اتومبیل خصوصیت پایداری لازم را تامین می کند، ولی نه به اندازه ی نویز های مصنوعی.

در نتیجه ، دیدیم که :

۱- ضرایب خود همبستگی $higher - lag$ سیگنال گفتار $s(n)$ شامل اطلاعات درباره ی طیف توان سیگنال $Pss(u)$ می باشد و

۲- دامنه ی ضرایب خودهمبستگی higher – lag سیگنال نویز d(n) برای بعضی از نویز ها بسیار کوچک هستند.

اگر فرض شود که خصوصیت جمع پذیری در حوزه ی خود همبستگی صادق است (ممکن است در تحلیل زمان – کوتاه کاملا صدق نکند.) ضرایب خودهمبستگی higher – lag سیگنال گفتار نویزی y(n) ، نسبت به اعوجاج نویز پس زمینه ی جمع پذیر حساس نخواهد بود.

پس، از ضرایب خودهمبستگی lower – lag صرف نظر می کنیم و تنها ضرایب خود همبستگی higher – lag (lag های بزرگتر از 2ms) سیگنال گفتار نویزی برای تخمین طیفی پایدار، استفاده می شوند.

ضرایب lower – lag را می توان با به کارگیری یک تابع پنجره ی نوک تیز مخروطی (tapered) تضعیف کرد. تاثیر دیگراین کار، تضعیف ضرایب خودهمبستگی very high-lag ، که واریانس نسبتا بالایی دارند، می باشد.

تخمین طیف از روی ضرایب خودهمبستگی higher - lag

در روش HASE ، تنها از قسمت higher –lag ِ یک طرفِ (طرفِ علی) رشته ی خود همبستگی استفاده می کنیم. با در نظر گرفتن تنها قسمت علی رشته ی خود همبستگی اطلاعاتی را از دست نمی دهیم زیرا رشته ی خود همبستگی متقارن می باشد. از طیف دامنه ی رشته ی خودهمبستگی پنجره گذاری شده ی یک طرفه به عنوان طیف توان سیگنال استفاده می کنیم. اگرچه، قبل از این، از تعداد کمی از نخستین ضرایب خودهمبستگی lower – lag صرف نظر می کنیم چراکه بیشتر تحت تاثیر نویز جمع پذیر قرار می گیرند. و تنها از ضرایب خودهمبستگی higher – lag (بیشتر از 2ms) استفاده می کنیم. پنجر ه را در رشته ی خود همبستگی، در یک higher- starting lag قرار می دهیم.

روش HASE برای محاسبه ی طیف توان سیگنال x(n) , n = 0 , 1, …, N-1 (که N برای سیگنال گفتاری به طول ۳۲ میلی ثانیه با فرکانس نمونه برداری 8kH z ، ۲۵۶ می باشد.) به صورت زیر می باشد:

۱- ضرب x(n) با تابع پنجره $W_s(n)$ (پنجره hamming) و بدست آوردن

$$x_w(n) = x(n)w_s(n), n = 0,1,..., N-1 \qquad (۳۵-۳)$$

۲- محاسبه ی تخمین بایاس شده ی ضرایب خودهمبستگی یک طرفه، با استفاده از روش محاسباتی مناسب مانند تبدیل فوریه ی سریع (fast fourier transform – FFT algorithm)

$$R(i)= \frac{1}{N}\sum_{n=0}^{N-i-1} x_w(n)x_w(n+i), \qquad i = 0,1,\ldots,N-1 \qquad (۳۶-۳)$$

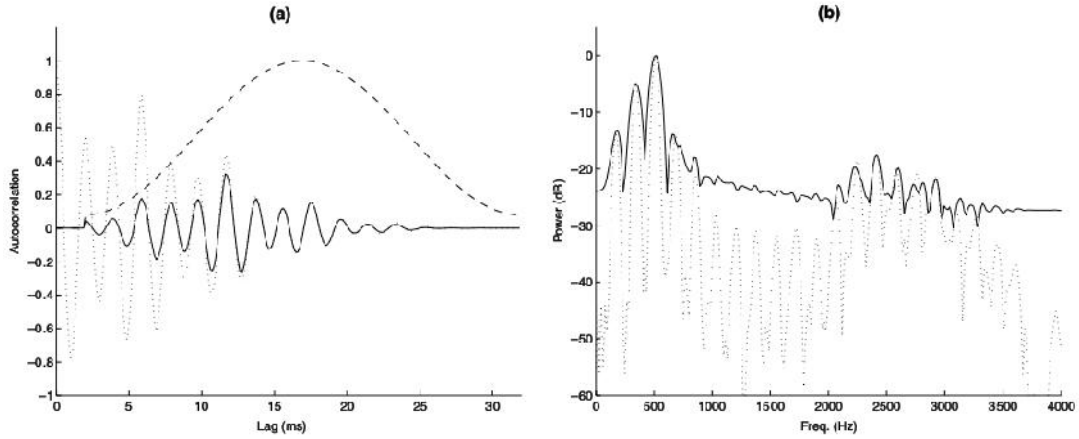
۳- حذف اولین L(=16) تای ضرایب خود همبستگی lower – lag و ضرب باقی مانده ی رشته ی خودهمبستگی higher – lag با یک تابع پنجره $Wr(n)$ برای دست آوردن $R_w(n) = R(L-n)w_r(n), n = 0,1,...,M-1$ که M=N-L

۴- افزودن 2N-M صفر در انتهای $R_w(n)$ و n=0,1,…,M-1 و محاسبه ی تبدیل فوریه ی گسسته DFT این سیگنال بدست آمده با استفاده از الگوریتم FFT. و استفاده از اندازه ی این تبدیل به عنوان تخمین طیف توان $\hat{P}_{xx}(\omega)$.

تابع پنجره $Wr(n)$ که در رشته ی خودهمبستگی یک طرفه (علی) به کار می رود، نقش مهمی در روش تخمین طیفی دارد. بنابراین بحث مختصری درباره ی این تابع پنجره انجام می دهیم :

طراحی تابع پنجره :

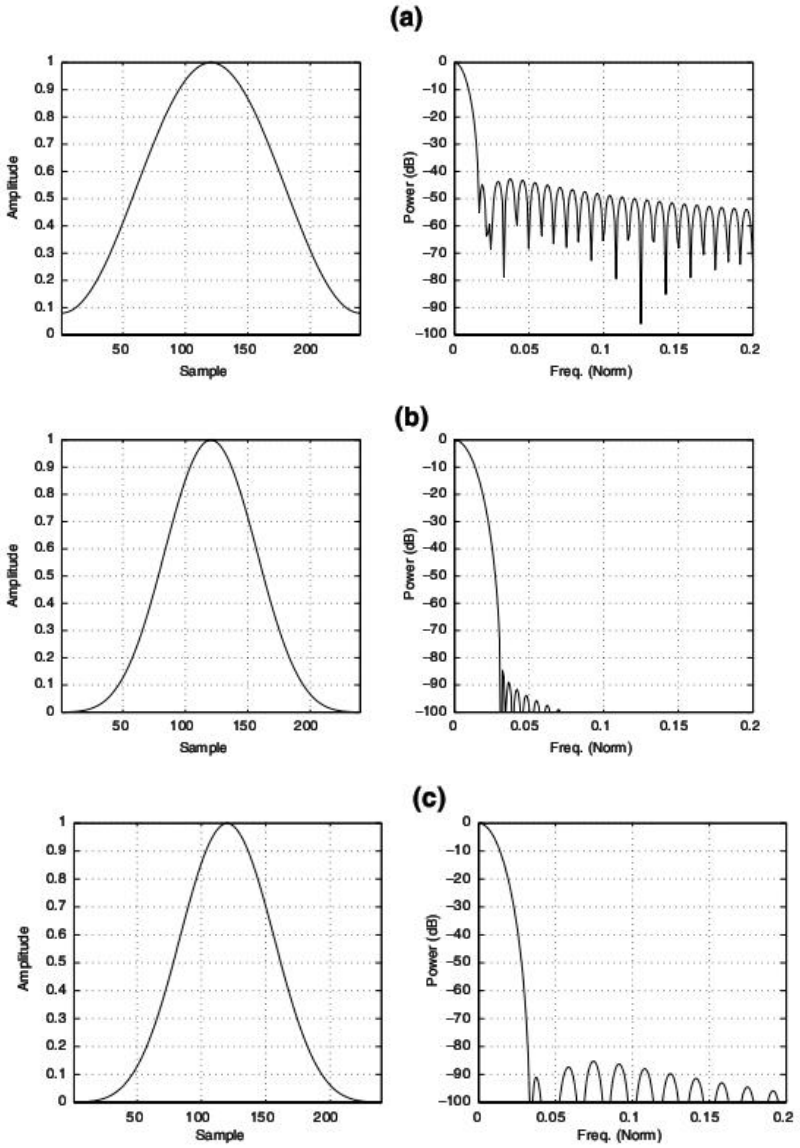
در روش HASE ، از پنجره ی hamming، $W_s(n)$ ، برای پنجره کردن سیگنال استفاده می کنیم ولی برای $Wr(n)$ ، نمی دانیم چه نوع تابع پنجره ای را برای پنجره کردن رشته ی خود همبستگی higher – lag یک طرفه به کار بریم ؟ برای یافتن تاثیر پنجره بر عملکرد تخمین طیفی در روش HASE ، ابتدا از تابع پنجره ی hamming برای $Wr(n)$ استفاده می کنیم. این روش HASE را بر سیگنال مطرح شده در شکل ۳-۱۵ ، اعمال می کنیم. عملیات پنجره گذاری بر رشته ی خودهمبستگی higher – lag در شکل ۳- ۲۱ نشان داده شده است.



شکل ۳- ۲۱ : نمایش روش HASE (با استفاده از رشته ی خود همبستگی higher – lag ، پنجره گذاری شده با hamming) بر یک قاب ۳۲ میلی ثانیه ای از گفتار voiced یک خانم. 'ey'

a) رشته ی یک طرفه (علّی) خودهمبستگی (خط نقطه چین)، یک تابع پنجره ی hammming به طول ۳۰ میلی ثانیه که از زمان ۲ میلی ثانیه آغاز شده (خط تیره)، رشته خودهمبستگی پنجره گذاری شده(خط توپر) (b) طیف توان تخمینی به روش HASE (خط توپر) و به روش periodogram (خط نقطه چین)

تخمین طیفی توان قاب گفتار با روش HASE در شکل ۳-۲۱ b به همراه تخمین periodogram آن نشان داده شده است. از شکل ۳- ۲۱ b می توان فهمید که روش periodogram می تواند جزئیات طیفی هارمونیک ها به کوچکی 43dB- را دنبال کند، یعنی رنج دینامیکی آن حدود 43 dB می باشد. روش HASE هارمونیک های دامنه بالا را به خوبی بدام می اندازد، ولی برای هارمونیک های با دامنه ی کمتر از 22 dB- نمی تواند بدام اندازد. به بیان دیگر، رنج دینامیکی آن به 22 dB محدود می شود. برای اهمیت این رنج دینامیکی، تصویر تابع پنجره ی hamming و طیف توان آن در شکل ۳- ۲۲ a نشان داده شده است.



شکل ۳-۲۲ : تابع پنجره و طیف توان آن برای

a) پنجره ی hamming و b) پنجره kaiser (α=۱۱.۳) و c) پنجره ی DDR hamming

می بینیم که دامنه ی بلندترین گوشه ی کناری side-lobe در حدود 43dB کمتر از گوشه ی اصلی main lobe است. یعنی رنج دینامیکی پنجره ی hamming برابر 43 dB است.

در اینجا باید توجه شود که پنجره ی hamming در روش periodogram تنها بر سیگنال اعمال می شود ولی در روش HASE هم بر سیگنال و هم بر رشته ی خودهمبستگی higher – lag اعمال می شود.

می دانیم که طیف توان رشته ی خود همبستگی یک سیگنال دارای رنج دینامکی ای دو برابر طیف توان سیگنال می باشد. روش HASE از پنجره ی hamming بر رشته ی خود همبستگی استفاده می کند، پس، تخمین طیف توان منتجه ی از سیگنال دارای رنج دینامیکی 22 dB می باشد.

زمانی که رشته ی خود همبستگی یک طرفه را پنجره گذاری می کنیم، لازم است تابع پنجره ای که رنج دینامیکی آن دو برابرتابع پنجره ای که معمولا برای سیگنال اصلی در حوزه ی زمان استفاده میکنیم، به کار برده شود. .

مثلا، پنجره ی hamming که معمولا در حوزه ی زمان بر سیگنال برای شناسایی گفتار استفاده می شود، دارای رنج دینامیکی 43 dB- می باشد. برای تولید یک تخمین طیفی با رنج دینامیکی کافی، که از رشته ی خودهمبستگی پنجره شده بدست آمده است، نیاز به تابع پنجره ای با رنج دینامیکی 86dB می باشد.

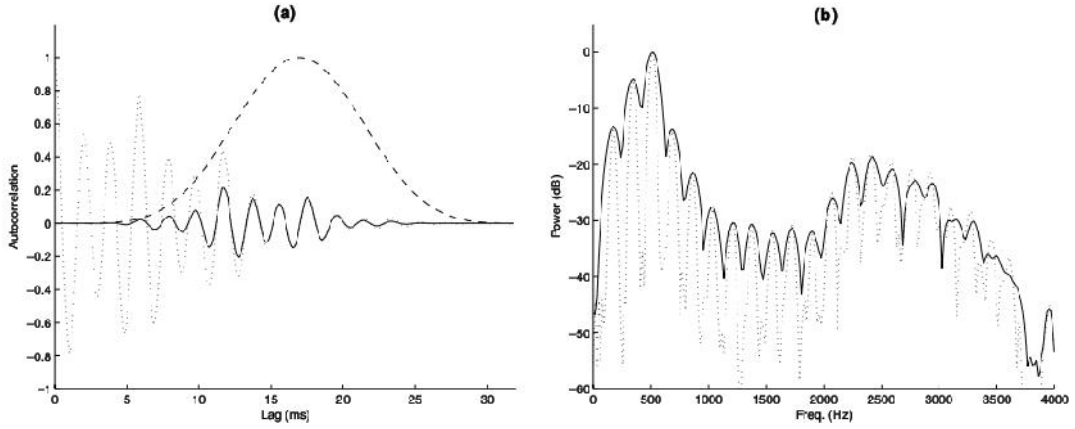
تابع پنجره ی Kaiser چنین شرایطی را تامین می کند. تابعی که برای تولید یک پنجره به طول M به کار می رود به صورت (۳-۳۶) است:

$$w(n) = \begin{cases} \dfrac{I_0\left(2\alpha\sqrt{\frac{n}{M-1}-\left(\frac{n}{M-1}\right)^2}\right)}{I_0(\alpha)}, & 0 \leqslant n < M, \\ 0, & \text{otherwise}, \end{cases} \qquad (۳۶-۳)$$

که $I_0(x)$ تابع بسل اصلاح شده نوع اول و α پارامتر طراحی می باشد که رنج دینامیکی پنجره را تنظیم می کند.

برای بدست آوردن تابع پنجره ی Kaiser با رنج دینامیکی 86dB ، α=11.3 می باشد. تابع پنجره ی منتجه و طیف توانش در شکل ۳-b۲۲ نشان داده شده است. از این پنجره برای پنجره گذاری رشته ی خودهمبستگی higher – lag برای قاب گفتار نشان داده شده در شکل ۳-۲۱ به روش HASE استفاده می کنیم .

عمل پنجره ی Kaiser را بر رشته ی خود همبستگی و نتیجه ی آن در شکل ۳-۲۳ a نشان داده شده است. و تخمین طیف توان قاب گفتار بدست آمده هم در شکل ۳-۲۳ b نشان داده شده است.



شکل ۳-۲۳ : نمایش روش HASE (با استفاده از رشته ی خود همبستگی higher – lag ، پنجره گذاری شده با kaiser) بر یک قاب ۳۲ میلی ثانیه ای از گفتار voiced یک خانم. 'ey'

a) رشته ی یک طرفه (علی) خودهمبستگی (خط نقطه چین) ، یک تابع پنجره ی kaiser به طول ۳۰ میلی ثانیه که از زمان ۲ میلی ثانیه آغاز شده (خط تیره) ، رشته خودهمبستگی پنجره گذاری شده(خط توپر) (b) طیف توان تخمینی به روش HASE (خط توپر) و به روش periodogram (خط نقطه چین)

با مقایسه ی شکل ۳-۲۳ b با شکل ۳-b۲۱ ، می بینیم که تابع پنجره ی Kaiser مشکل رنج دینامیکی که در تابع پنجره ی hamming با آن روبرو بودیم را حل می کند. ولی هارمونیک های کمی پهن تر را تولید می کند. این به دلیل این است که تابع پنجره ی Kaiser نسبت به تابع پنجره ی hamming دارای main lobe پهن تری می باشد.

مشکلات تابع پنجره ی Kaiser :

۱- از لحاظ محاسباتی تابع پنجره ی Kaiser به دلیل داشتن تابع بسل از توابع پنجره ی کسینوسی مانند hamming ،پیچیده تر هستند.

۲ - هر گاه تصمیم بگیریم که از یک تابع پنجره ای متفاوت $w_s(n)$ (به غیر ازپنجره ی hamming) برای اعمال بر سیگنال در حوزه ی زمان، استفاده شود (مانند پنجره ی Blackman) ، باید نرم افزار استخراج ویژگی را به گونه ای اصلاح کنیم که تابع پنجره ی Kaiser دارای رنج دینامیکی ای دو برابر رنج دینامیکی تابع پنجره ی جدید باشد.

برای حل این مسائل، یک روش طراحی ویژه برای محاسبه ی تابع پنجره ای، پیشنهاد می شود که دارای ۲ برابر رنج دینامیکی تابع پنجره ی ws(n)، که بر سیگنال در حوزه ی زمان اعمال می شود، باشد و هر وقت تابع پنجره ی(ws(n مشخص شد، بتوان آن را به طور اتوماتیک در نرم افزار تولید کرد.

با توجه به اینکه می خواهیم پنجره ی Hamming در حوزه زمان، بر سیگنال اعمال کنیم و از رشته ی خود همبستگی از 2 ms تا 32 ms (با طول M=240 نمونه با نرخ نمونه برداری (8kHz)استفاده کنیم، می خواهیم تابع پنجره ای به طول M طراحی کنیم که رنج دینامیکی دو برابر پنجره ی Hamming داشته یاشد .

روند طراحی به صورت زیر است :

۱- ساختن یک پنجره hamming به طول M/2

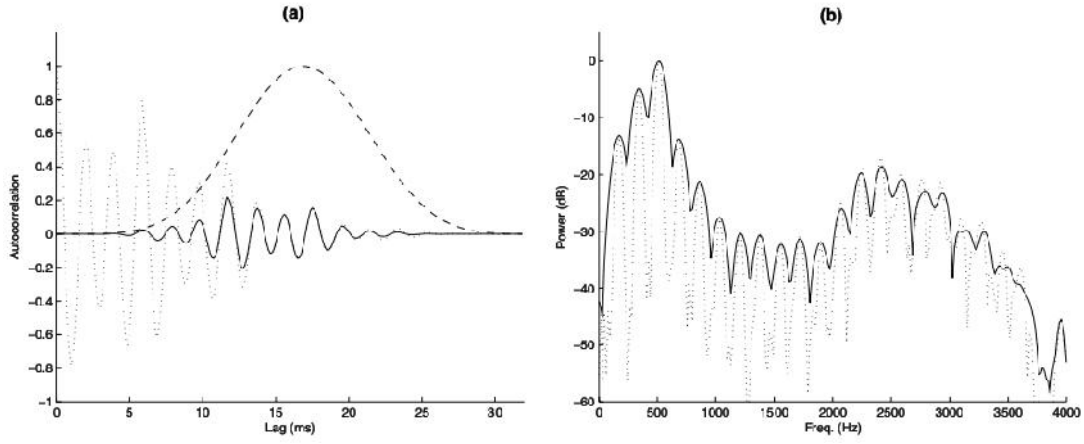
۲- محاسبه ی رشته ی خود همبستگی دو طرفه (بایاس شده) به طول M-1 که یک ماکزیمم در صفرامین lag در مرکز داشته باشد.

۳- اضافه کردن یک صفر در انتهای رشته ی خود همبستگی به طول M-1 ، تا پنجره ی دلخواه به طول M بدست آید.

این پنجره دارای رنج دینامیکی برابر 86 dB می باشد.

این تابع پنجره را به عنوان پنجره ی double-dynamic-range (DDR) Hamming مطرح می کنیم. این تابع پنجره به همراه طیف توانش در شکل ۳-۲۲ c نشان داده شده است.

عملکرد این پنجره را بر رشته ی خود همبستگی higher –lag مربوط به سیگنال شکل ۳-۲۱ بررسی می کنیم که تخمین طیف توان منتجه در شکل ۳-۲۴ a نشان داده شده است.



شکل ۳-۲۴ : نمایش روش HASE (با استفاده از رشته ی خود همبستگی higher – lag ، پنجره گذاری شده با DDR hamming) بر یک قاب ۳۲ میلی ثانیه ای از گفتار صدادار یک خانم. 'ey'

(a) رشته ی یک طرفه (علی) خودهمبستگی (خط نقطه چین) ، یک تابع پنجره ی hammingبه طول ۳۰ میلی ثانیه که از زمان ۲ میلی ثانیه آغاز شده (خط تیره) ، رشته خود همبستگی پنجره گذاری شده(خط توپر) (b) طیف توان تخمینی به روش HASE (خط توپر) و به روش periodogram (خط نقطه چین)

با مقایسه ی شکل ۳-۲۴ b با شکل ۳-۲۳ b می بینیم که پنجره ی DDR hamming از لحاظ عملکرد تخمین طیفی به خوبی تابع پنجره ی Kaiser می باشد .

- **عملکرد تخمین طیفی به روش HASE**:

همان طور که پیش تر گفته شد، روش HASE از طیف دامنه ی رشته ی خودهمبستگی higher – lag یک طرفه ی پنجره گذاری شده، به عنوان تخمینی از طیف توان سیگنال استفاده می کند.

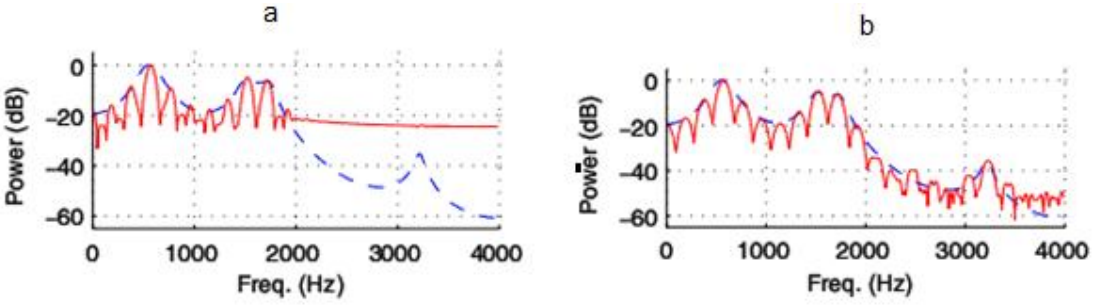
برای دانستن عملکرد تخمین طیفی این روش، برای تحلیل، ابتدا از یک سیگنال گفتار صدادار ساختگی (ترکیبی) استفاده می شود، زیرا طیف توان آن مشخص است و می تواند به عنوان طیف مرجع به کار رود. این کار در ارزیابی عملکرد روش تخمین طیف مذکور، کمک می کند.

برای تولید این سیگنال ترکیبی، یک قسمت از صدای واقعی /r/ را در نظر می گیریم و سپس با استفاده از روش خودهمبستگیِ تحلیل LP ، ضرایب LP آن را محاسبه می کنیم.

یک قاب 32ms سیگنال گفتار ساختگی را به کار می بریم و با استفاده از دو روش تخمین طیفی HASE، طیف توان آن را محاسبه می کنیم.

روش ۱- روش HASE با پنجره گذاری رشته ی خود همبستگی higher – lag یک طرفه با استفاده از پنجره ی hamming

روش ۲- روش HASE با پنجره گذاری رشته ی خود همبستگی higher – lag یک طرفه با استفاده از پنجره ی DDR hamming



شکل ۳-۲۵ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدادار ساختگی پاک . خط تیره در هر نمودار طیف توان اصلی سیگنال ساختگی می باشد.خط توپر در هر نمودار تخمین طیف توان می باشد. a)روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با hamming پنجره گذاری شده و b) روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با DRR hamming پنجره گذاری شده

شکل ۳-۲۵ b تخمین طیفی بدست آمده از روش ۱ را نشا ن می دهد. طیف توان اصلی سیگنال ساختگی ۴ قله دارد، ولی تخمین طیفی توان با این روش تنها ۳ قله دارد. کاهش رنج دینامیکی طیفی که نتیجه ی استفاده از پنجره ی hamming برای پنجره گذاری رشته ی خود همبستگی است، بسیار واضح است. طیف اصلی رنج دینامیکی 45 dB- را نشان می دهد. رنج دینامیکی بدست آمده با این روش حدود نصف این مقدار است که نقطه ضعف مطرح شده را نشان می دهد.

شکل ۳-۲۵ c تخمین طیفی بدست آمده از روش ۲ را نشان می دهد. بدلیل اینکه پنجره ی DDR hamming دارای رنج دینامیکی بزرگی است، ۴ قله ی موجود در طیف اصلی، در تخمین طیفی با این روش نیز بدست آمده است. همچنین، همخوانی بین طیف اصلی و تخمین طیفی عالی است. این شکل نشان می دهد که تخمین طیف توان خیلی خوبی از گفتار را می توان با استفاده از رشته ی خودهمبستگی یک طرفه، ساخت به شرط آنکه تابع پنجره به طور مناسب طراحی شده باشد.

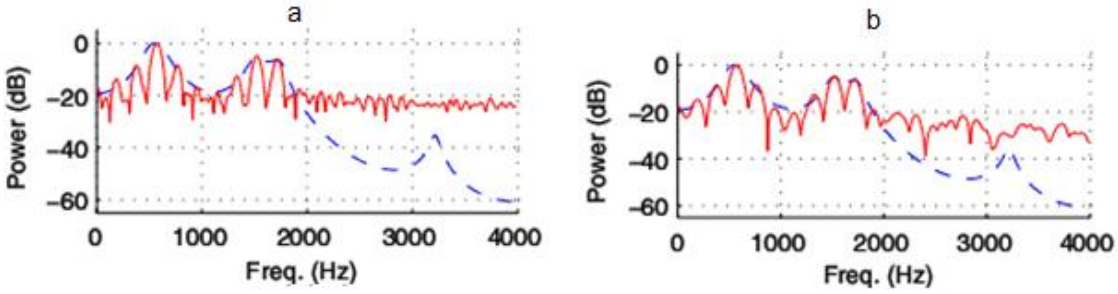
در شکل ۳ -۲۵ عملکرد دو روش تخمین طیفی برای سیگنال گفتار ساختگی پاک (بدون نویز) بررسی شد.

اکنون عملکرد دو روش بالا را برای سیگنال گفتار ساختگی نویزی در نظر می گیریم.

برای این منظور یک قاب 32 ms از گفتار صدادار ساختگی را با اضافه کردن ۴ نویز زیر تخریب می کنیم تا SNR برابر 10 dB بدست آید:

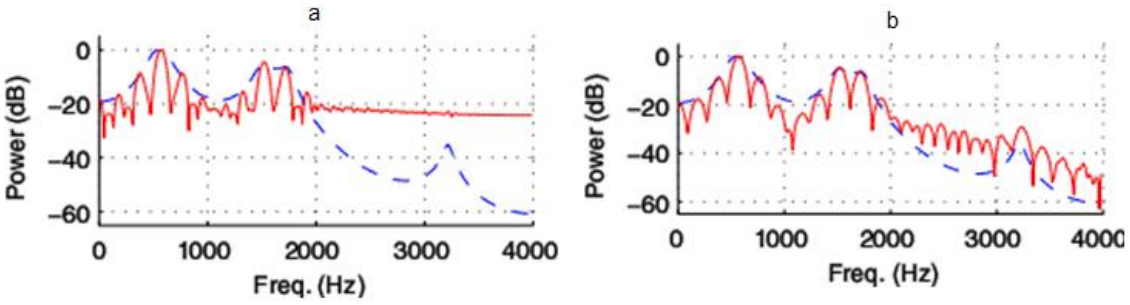
- نویز سفید گوسی اتفاقی - نویز ساختگی chrip - نویز ساختگی ضربه ای - نویز واقعی اتومبیل

نتایج عملکرد طیفی دو روش بالا برای هر نویز در شکل های ۳- ۲۶ تا ۳- ۲۹نشان داده شده است.



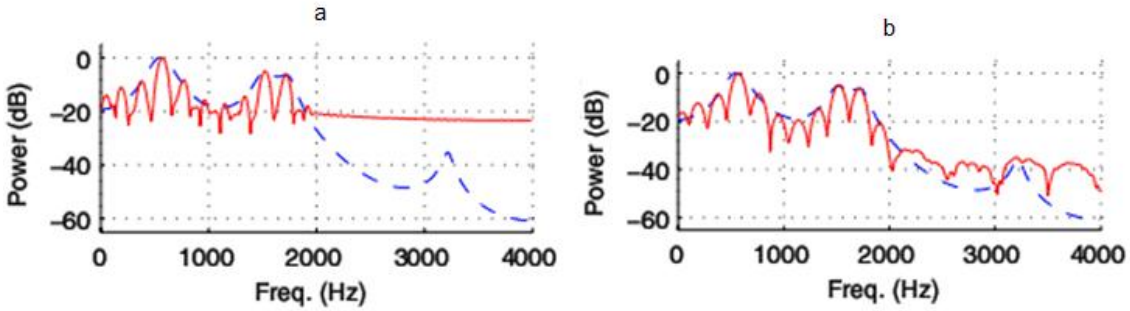
شکل ۳- ۲۶ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز سفید اتفاقی ساختگی با 10 dB SNR تخریب شده است . خط تیره در هر نمودار طیف توان اصلی سیگنال ساختگی می باشد.خط توپر در هر نمودار تخمین طیف توان می باشد.

a)روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با hamming پنجره گذاری شده و b) روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با DRR hamming پنجره گذاری شده



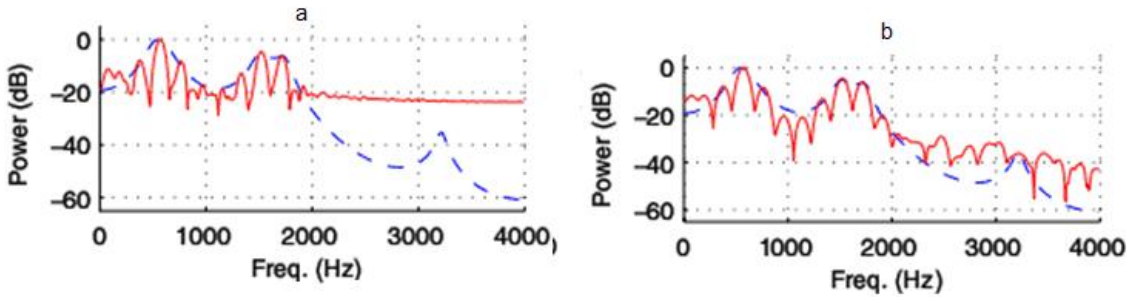
شکل ۳- ۲۷ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز chrip ساختگی با 10 dB SNR تخریب شده است. خط تیره در هر نمودار طیف توان اصلی سیگنال ساختگی می باشد.خط توپر در هر نمودار تخمین طیف توان می باشد.

a)روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با hamming پنجره گذاری شده و b) روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با DRR hamming پنجره گذاری شده



شکل ۳-۲۸ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز ضربه ای اتفاقی ساختگی با 10 dB SNR تخریب شده است. خط تیره در هر نمودار طیف توان اصلی سیگنال ساختگی می باشد.خط توپر در هر نمودار تخمین طیف توان می باشد.

a)روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با hamming پنجره گذاری شده و b) روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با DRR hamming پنجره گذاری شده



شکل ۳-۲۹ : مقایسه ی روش های تخمین طیفی با به کارگیری یک قاب ۳۲ میلی ثانیه ای از گفتار صدای ساختگی که با نویز واقعی اتومبیل با 10 dB SNR تخریب شده است. خط تیره در هر نمودار طیف توان اصلی سیگنال ساختگی می باشد.خط توپر در هر نمودار تخمین طیف توان می باشد.

a)روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با hamming پنجره گذاری شده و b) روش HASE با استفاده از رشته ی خود همبستگی higher- lag که با DRR hamming پنجره گذاری شده

با مقایسه ی این تصاویر با شکل ۳-۲۵ می بینیم که عملکرد دو روش بالا تحت تاثیر نویز قرار می گیرد. برای چگونگی عملکرد این روش ها در حضور نویز، عملکرد تخمین فرمنت (اندازه گیری شده بر حسب تعداد فرمنت هایی که به خوبی تخمین زده شده اند)را برای هر دو روش محاسبه می کنیم. زمانی یک فرمنت به خوبی تخمین زده شده است که با یکی از ۴ فرمنت موجود در طیف توان اصلی سیگنال ساختگی تطابق داشته باشد.

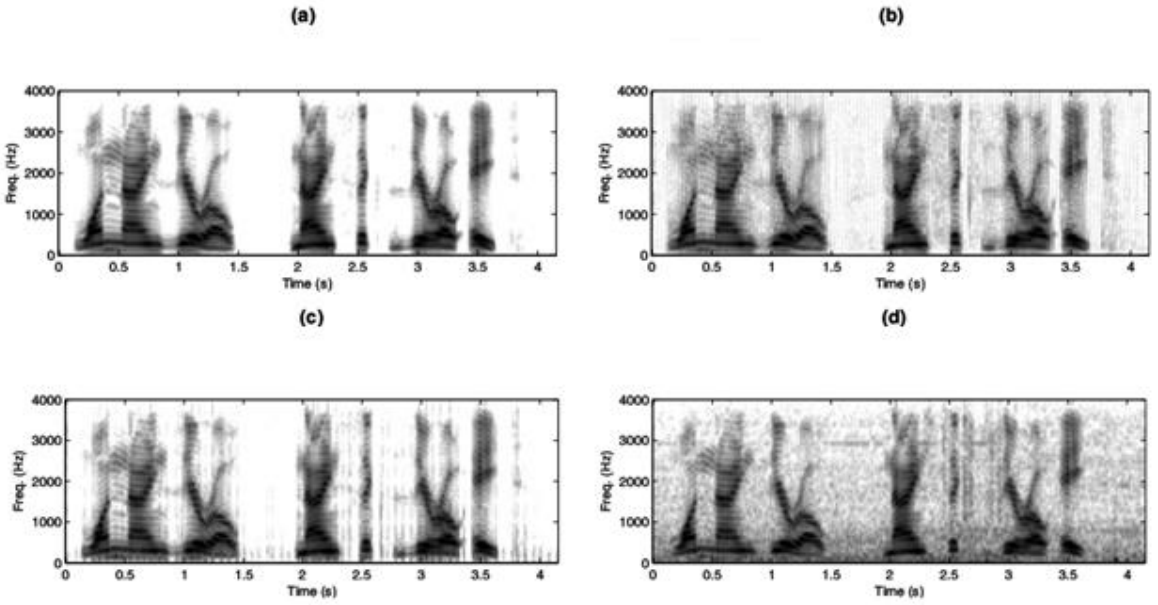
در جدول ۳-۱ عملکرد تخمین فرمنت برای دو روش HASE ، برای هر کدام از انواع نویز ها خلاصه شده است.

Method	Number of well-estimated formants			
	White	Chirp	Impulse	Car
HASE method – Hamming	3	3	3	3
HASE method – DDR Hamming	3	4	3	3

جدول ۳- ۱ : عملکرد تخمین فرمنت با استفاده از دو روش ذکر شده ، بر یک قاب ۳۲ میلی ثانیه ای از گفتار ساختگی نویزی

حال، عملکرد این دو روش را برای یک سیگنال گفتار واقعی بررسی می شود.

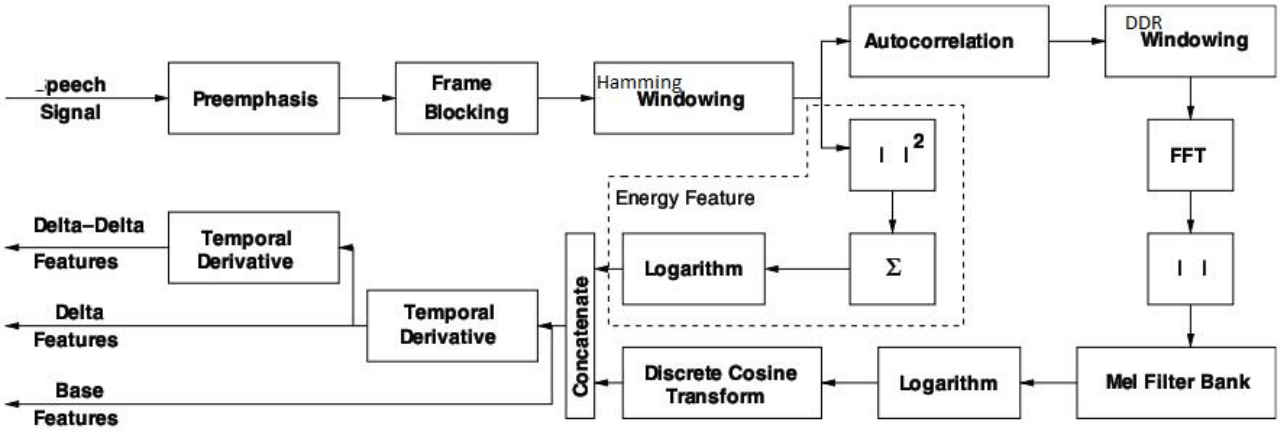
یک عبارت گفتار را از پایگاه داده ی. (Aurora (MAL_19Z96Z8A انتخاب می کنیم. طیف نگاره ی (spectogram) این سیگنال گفتار واقعی را محاسبه می کنیم.(با جابجایی قاب 10 ms و طول قاب 32 ms)



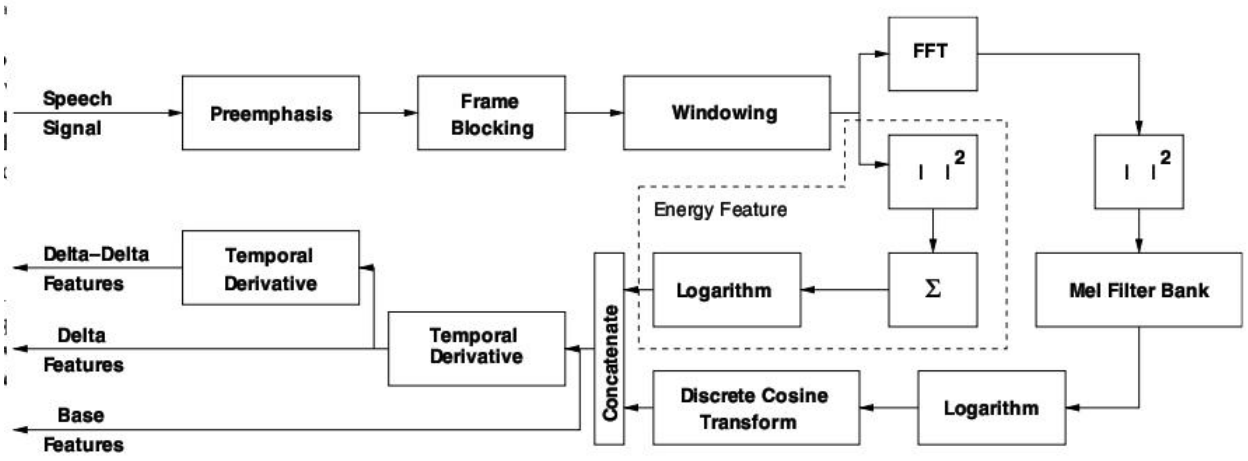
شکل ۳-۳۰ : طیف نگاره ی گفتار واقعی (عبارت 'MAL_19Z96Z8 A' از پایگاه داده ی Aurora) با به کارگیری روش priodogram(نمودار فوقانی در هر نیم تصویر) و روش HASE با پنجره گذاری مقادیر زمان – بالای رشته ی خود همبستگی با DDR hamming (نمودار تحتانی در هر نبم تصویر) با نویز های مختلف . a) عبارت پاک b)نویز c . chrip) نویز ضربه ای . d) نویز اتومبیل

روش استخراج ویژگی AMFCC

از ضرایب خودهمبستگی برای time lag های کمتر از 2ms صرف نظر می کنیم و پنجره ی DDR را بر باقی مانده ی رشته ی خود همبستگی (higher - lag) اعمال می کنیم. (2-32 ms)



شکل ۳-۳۱ : بلوک دیاگرام الگوریتم استخراج ویژگی در AMFCC



شکل ۳- ۳۲ : بلوک دیاگرام الگوریتم استخراج ویژگی در MFCC

۳-۷-۲ ضرایب MFCC دو طیفی برای شناسایی گوینده ی پایدار :

روش دوطیفی در زمینه ی شناسایی گفتار در [۲۴] مطرح شده است که نتایج حاصل از آزمایش های انجام شده را به طور مختصر اشاره می کنیم.

در سیستم های تشخیص گوینده، سیگنال های صحبت ضبط شده و در یک پایگاه داده ذخیره میشود. مجموعه ی آموزش، که شامل بردارهای ویژگی است، با مجموعه ی آزمایش مقایسه می شوند. شرایط بدست آوردن داده های مجموعه ی آموزش و مجموعه آزمایش می تواند بسیار متفاوت باشد. در حالی که مجموعه آموزش را معمولا می توان در یک محیط بدون نویز بدست آورد، ممکن است مجموعه آزمایش را نتوان با این شرایط تهیه کرد. این امر ممکن است به کاهش عملکرد یک سیستم تشخیص گوینده منجر شود. روشهای متعددی برای جلوگیری از این مشکل در مقالات ذکر شده است که عمدتا شامل استخراج ویژگی پایدار، تکنیک های تقویت صحبت و جبران نویز می باشند.

در این مقاله، کاربرد بخش دو- طیفی برای محاسبه ی MFCC به عنوان ویژگیهای پایدار درسیستم تشخیص گوینده مستقل از متن، معرفی می گردد.

مهم ترین مزیت کاربرد دو- طیفی قابلیت کاهش نویز گوسی افزایشی در عین حفظ اطلاعات فاز سیگنال میباشد. استخراج ویژگی های یک سیگنال گفتار عمدتا براساس طیف آن صورت می گیرد زیرا هر مشخصاتی در باره ی لوله ی صوتی را می توان از طیف سیگنال بدست آورد. گرچه طیف یک سیگنال گفتار را می توان با مدل های متفاوتی تعریف کرد، استفاده از بانک های فیلتر به جای تحلیل پیش بینی خطی (linear prediction) ، ویژگی های پایدارتری را بدست می دهد. در این مقاله [۲۴] ، استفاده از قطعه دو-طیفی(bispectrum slice) به جای ضرایب MFCC رایج که با بانک های فیلتر بدست آمده اند، بررسی می شود.

bispectrum slice :

اگر برای هر سیگنال گسسته $x(n)$ ، خود همبستگی سگانه (autotripplecorrelation) به صورت(۳-۳۷) باشد:

$$c(\tau_1,\tau_2) = E[x(n)x(n+\tau_1)x(n+\tau_2)] \qquad (۳۷-۳)$$

دوطیفی $B(\mu_1,\mu_2)$ به عنوان تبدیل فوریه ی دو بعدی تابع خودهمبستگی سیگنال تعریف میشود.

$$B(\mu_1,\mu_2) = F\{c(\tau_1,\tau_2)\} \qquad (۳۸-۳)$$

که $F\{.\}$ مقدار مورد انتظار(امید ریاضی) و $E[.]$ تبدیل فوریه می باشد.

عکس تبدیل فوریه یک بعدی دو طیفی، $q(n)$ ، بر روی خط $\mu_1 = \mu_2$ به عنوان مجموع توده ای(sum of cumulants) تعریف می شود:

$$q(n) = x(n) * x(n) * k(n) \qquad (۳۹-۳)$$

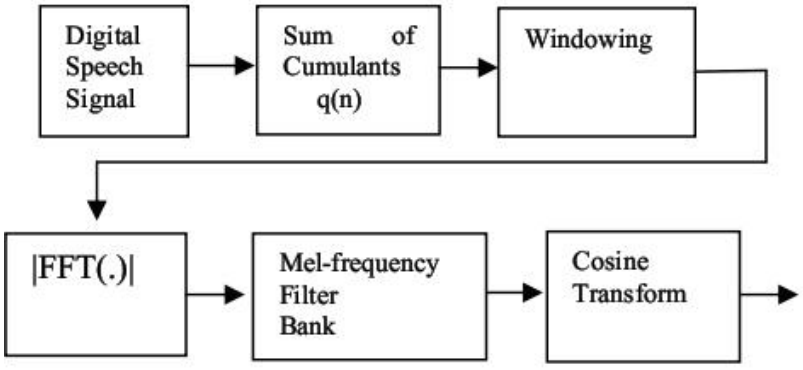
که * عملگر کانولوشن را نشان می دهد.و

$$k(n) = \begin{cases} x(N-1-N/2) \\ 0 \end{cases} \qquad (۴۰-۳)$$

برای سیگنال با $n=0,1,\ldots,N-1$. رشته $q(n)$، دارای $4N-3$ نمونه با $n = -2(N-1),\ldots,-1,0,1,\ldots,2(N-1)$ می باشد.

برای استخراج ویژگی، سیگنال به قسمت هایی به طول 16ms و با همپوشانی 10ms تقسیم شده است. بعد از تخمین مجموع توده ای با استفاده از فرمول (۳-۳۹)، برای این قسمت ها، mfcc محاسبه می گردد.

در شکل ۳- ۳۳ مراحل استخراج ویژگی نشان داده شده است.



شکل ۳- ۳۳ : بلوک دیاگرام مراحل استخراج ویژگی در [۲۴]

از GMM برای رده بندی استفاده شده است. اگر M مرتبه ی GMM باشد، GMM به صورت (۳-۴۱) تعریف می گردد: (۳-۴۱) و $p(x|\text{a}) = \sum_{i=1}^{M} w_i b_i(x)$

$$b_i(x) = \frac{1}{(2®)^{D/2}|C_i|^{1/2}} \exp(-0.5(x-«_i)^T C_i^{-1}(x-«_i)) \qquad (۳-۴۲)$$

که x یک بردار اتفاقی D بعدی، $w_i (i = 1,\ldots,M)$ ضرایب وزن و $b_i(x)$ توابع چگالی اجزا می باشند.

ضرایب وزن مخلوط باید در معادله (۳-۴۳) صدق کنند :

$$\sum_{i=1}^{M} w_i = 1 \qquad (۴۳-۳)$$

GMM با $\mathrm{a} = \{w_i, \ll_i, C_i\}$ تعریف می گردد، که $\ll_i$ بردارهای میانگین و C_i ماتریس کوواریانس می باشد. برای سیستم های تشخیص گوینده، a میتواند به عنوان مدل گوینده ظاهر شود. سپس پارامترهای مدل با روش حداکثر کردن امید ریاضی که پیشامد مدل را حداکثر می کند، تخمین زده می شود.

از پایگاه داده TIMIT استفاده شده است. مجموعه داده های آموزش شامل ۵۰ گوینده مرد با لهجه ی یکسان با طول سیگنال های مختلف و مجموعه داده های آزمایش شامل ۵ جمله مختلف از قسمت های مختلف صحبت هر عضو مجموعه آموزش می باشد. مرتبه GMM ، M=40 در نظر گرفته شده است.

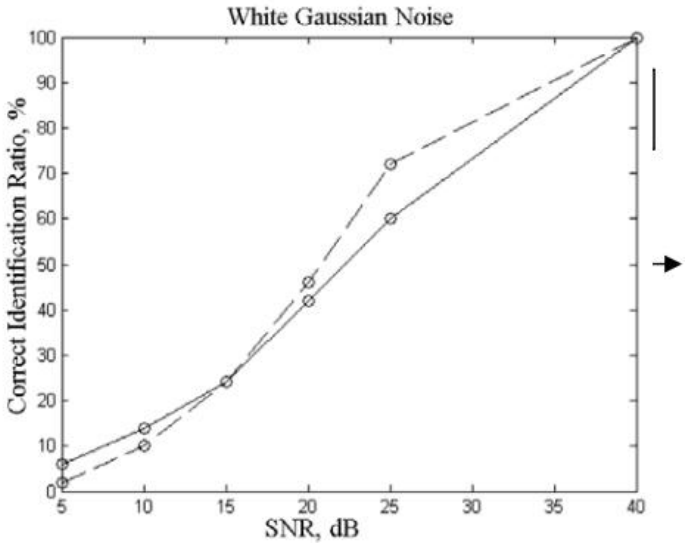
ارزیابی به وسیله ی امتیاز کلی نرمالیزه شده صورت می گیرد که به صورت استخراج لگاریتمی پیشامد گوینده حقیقی از حداکثر پیشامد گوینده های دیگر بجز گوینده حقیقی تعریف می گردد:

$$\log(L(X)) = \log p(X|S = S_c, \mathrm{a}) - \max(p(X|S \neq S_c, \mathrm{a})) \qquad (۴۴-۳)$$

که X ویژگی ها در مجموعه ی آزمایش می باشد. S گوینده در مجموعه ی آموزش و S_c گوینده حقیقی با پارامترهای مدل a. اگر امتیاز نرمالیزه شده مثبت باشد ، گوینده به درستی تخمین زده شده است.

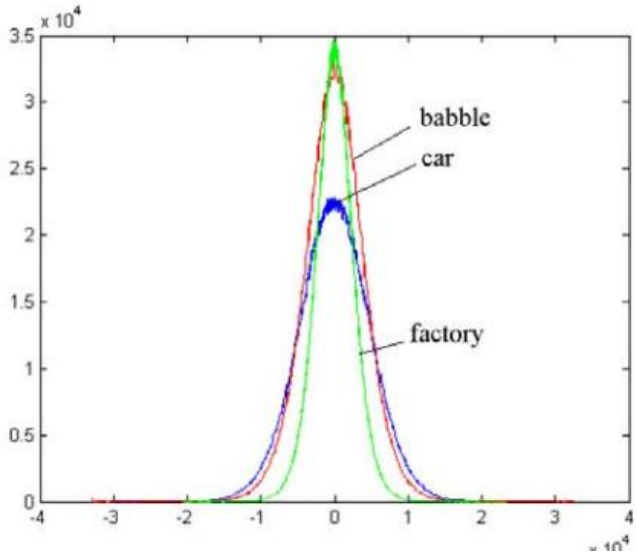
ابتدا، سیستم تشخیص گوینده را با مجموعه داده ی آموزش بدون نویز و مجموعه آزمایش نویز دار شبیه سازی می کنیم. سپس، برای مقایسه ی پایداری دوطیفی نسبت به تک طیفی، نویز گوسی سفید با SNR=40,20,15,10,5 dBs را به سیگنال گفتار اضافه می کنیم.

عملکرد شناسایی گوینده برای دو روش مذکور در شکل ۳- ۳۴ نشان داده شده است. برای نسبت های سیگنال به نویز کمتر از 10dB، نتیجه ی عملکرد برای هر دو روش طیفی و دو طیفی پایین است ولی برای بیشتر از 10 dB ، نرخ شناسایی گوینده در زمانی که از روش دوطیفی استفاده می شود، بهتر است .



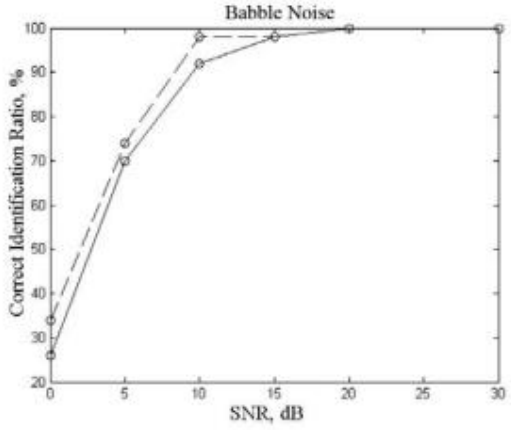
شکل ۳-۳۴ : نسبت تشخیص صحیح بر حسب SNR برای روش طیفی (خط توپر) و روش دو طیفی (خط چین) در زمانی که نویز سفید گوسی اضافه شود.[۲۴]

در مجموعه ی آزمایش برای نشان دادن تاثیر روش های پیشنهادی، نویزهای دیگری نیز به سیگنال گفتار اضافه می شود. نمونه های نویز واقعی، مانند babble و اتومبیل و کارخانه، از پایگاه داده ی NOISEX از مجموعه ی سیگنال گفتار NATO ، جمع آوری شده است. اگرچه ممکن است توزیع این نویزها گوسی نباشد، ولی متقارن هستند زیرا خطوط انحراف (اریب) آنها به صفر نزدیک تر است. هیستوگرام این نویزها در شکل ۳-۳۵ نشان داده شده است .

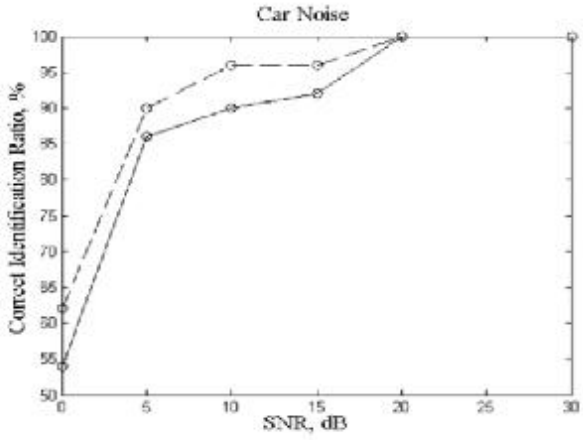


شکل ۳-۳۵: هیستوگرام نویز های babble و اتومبیل و کارخانه[۲۴]

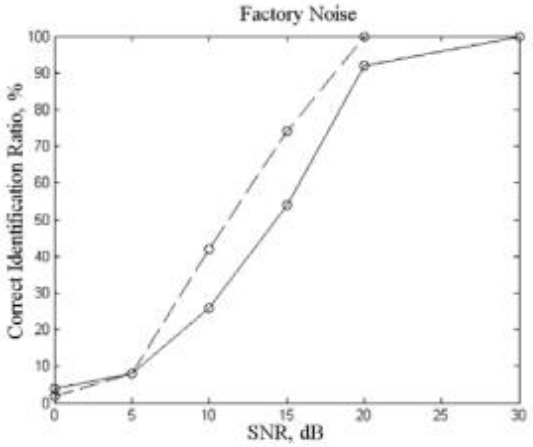
نتایج آزمایش برای نویزهای واقعی در شکل های ۳-۳۶ و ۳-۳۷ و ۳-۳۸ آورده شده است. که نشان می دهند بـرای هـر SNR بیشتر از 0 dB ، زمانی که از روش دوطیفی برای استخراج ویژگی در سیستم شناسایی گوینده استفاده می شـود، عملکرد بهتری داریم.



شکل ۳-۳۶:[۲۴]



شکل ۳-۳۷



شکل ۳-۳۸

۳-۷-۳ غیر حساس کردن MEL- Cepstrum نسبت به اجزای طیفی نادرست (بدلی) (RootMFCC)، برای شناسایی گفتار پایدار

واضح است که قله ها در طیف بانک فیلتر Mel لگاریتمی نشانه های مهمی در توصیف صفات صوت گفتار هستند. اما، اختلال و آشفتگی در طیف توان بعد از فشرده سازی لگاریتمی می تواند از لحاظ عددی قابل توجه باشد. نشان می دهیم که حتی اگر قله های طیف ثابت نگه داشته شوند، اختلالات دارای انرژی کم در طیف توان می تواند تغییرات بزرگی در ضرایب cepstral ایجاد کند. هم به صورت تحلیلی و هم به صورت آزمایشی نشان داده می شود که با به توان رساندن طیف بانک فیلترMel لگاریتمی، قبل از محاسبات مربوط به cepstrum، می توان حساسیت ضرایب cepstra نسبت به اختلالات و آشفتگی های بدلی دارای انرژی کم، را به طور قابل ملاحظه ای کاهش داد. طیف مدولاسیون -Mel cepstrum از cepstra های بدست آمده محاسبه می شود که پایداری بیشتری برای بردار ویژگی ترکیبی نتیجه می دهد. در آزمایشات انجام گرفته با سیگنال های صحبت، نشان داده میشود که تکنیک پیشنهادی براساس ویژگی ها، افزایش چشمگیری درعملکرد تشخیص گفتار در شرایط نویز نا ایستا نتیجه می دهد. (در مقایسه با ویژگی های MFCC , RASTA-PLP)

آنچنان که واضح است، در حضور سطوح نویز جمع پذیر رایج ، فرمنت ها در مقایسه با دره های طیفی که نمایش دهنده ی ریپل های جعلی هستند، کمتر تحت تاثیر قرار می گیرند.

تبدیل کسینوسی طیف بانک فیلتر Mel لگاریتمی (log-MelFBS) که معمولا به عنوان بردار ویژگی های MFCC شناخته شده اند، نسبت به ریپل ها در دره های طیفی که مشخصه ای از صوت گفتار را تعیین نمی کنند، حساس اند. این یکی از دلایل عملکرد ضعیف ویژگی های MFCC در شرایط نویز جمع پذیر می باشد.

با مشاهده ی اینکه قسمت های دارای دامنه ی بیشتر (مانند فرمنت ها) در طیف، تاثیر نسبتا کمتری از نویز می پذیرند ، Paliwal مرکز ثقل زیرباند های طیفی (Spectral Subband Centroids-SSC) را به عنوان ویژگی ها پیشنهاد کرد.

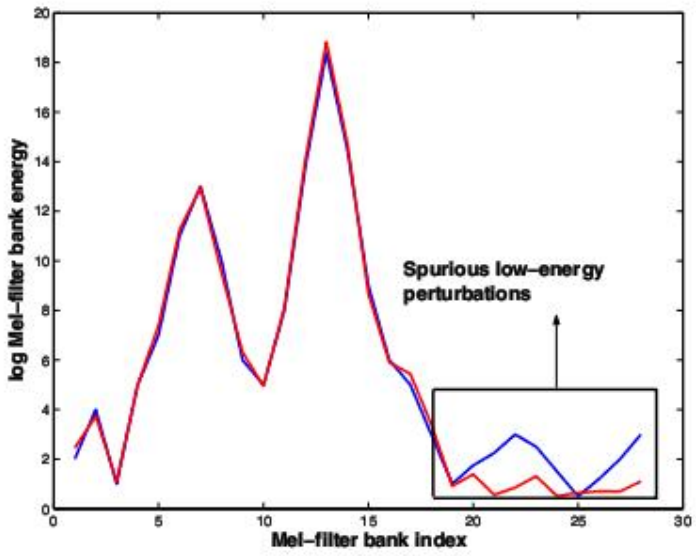
در [۲۵]، به طور تحلیلی نشان داده شده است که به توان رساندن logMelFBS می تواند باعث کاهش حساسیت ضرایب cepsrum به اختلالات جعلی در دره های logMelFBS می شود.(در مقایسه با قله ها)

Lim استفاده از سیستم root homomorphic deconvolution system (SRDS) را به عنوان نمونه ی نسبتا کلی تری از Logarithmic homomorphic deconvolution system-LHDS ، پیشنهاد می کند.

SRDS از فشرده سازی ریشه ای root ، $1 > ¥, ^{¥}(.)$ انرژی های بانک فیلتر mel به جای فشرده سازی لگاریتمی استفاده می کند.

یکی از نتایج اعمال فشرده سازی لگاریتمی بر انرژی های بانک فیلتر mel، کاهش رنج دینامیکی دامنه های طیف می باشد. در نتیجه، اختلالات جعلی (ساختگی) که در حوزه ی طیف توان از لحاظ عددی نا چیز هستند، ممکن است بعد از فشرده سازی لگاریتمی انرژی های بانک فیلتر mel از لحاظ عددی قابل توجه شوند.

در شکل ۳-۳۹، این مسئله نمایش داده شده است. منحنی آبی و قرمز دو نمونه از یک logMelFBS با فرمت های یکسان هستند ولی دارای اختلالات متفاوتی در انرژی پایین می باشند.

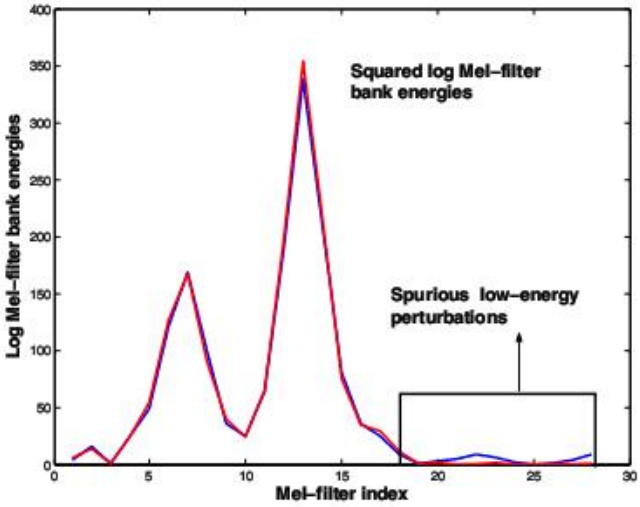


شکل ۳-۳۹ : انرژی های بانک log mel-filter از یک گفتار پاک و نویزی (مشوش)[۲۵]

این اختلالات تقریبا 10^{-12}% از انرژی طیف توان (قبل از فشرده سازی log) می باشند که توصیفی (مشخصه ای) از صوت گفتار را ندارند. هر چند DCT یک تبدیل خطی است، وزن دهی یکسانی برای فرمنت ها و خروجی های انرژی پایین بانک فیلتر می دهد. درنتیجه نسبت به ریپل های جعلی حساس است.

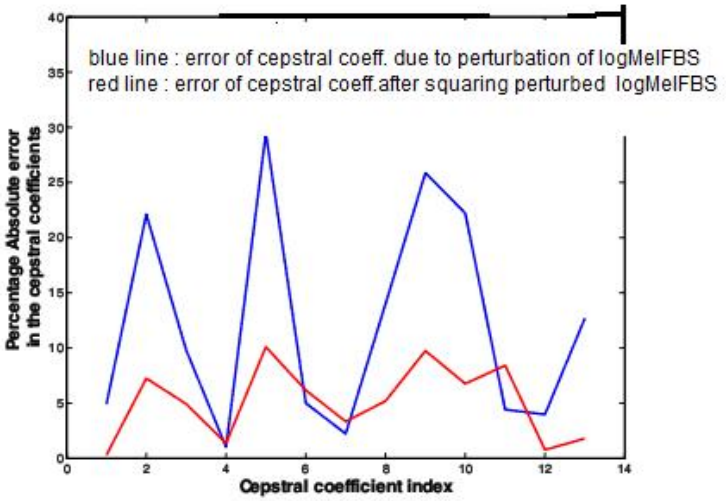
یک راه حل طبیعی برای این مسئله، وزن دهی به logMelFBS به گونه ای است که فرمنت ها نسبت به نمونه های انرژی کم بانک فیلتر، مهم تر شوند. برای این منظور، کپی ای از خود logMelFBS ،گزینه ی مناسبی برای lifter می باشد چراکه تاکید بیشتری بر فرمنت ها نسبت به خروجی های انرژی پایین logMelFBS دارد. این کار همانند به توان رساندن logMelFBS با یک توان P که P>1 می باشد.

شکل ۳-۴۰ ، مربع (توان ۲) logMelFBS دو نمونه ی شکل ۳-۳۹ را نشان می دهد. همان طور که از منحنی ها در شکل ۳-۴۰ مشخص است، فرمنت ها در مقایسه با ریپل های جعلی برجسته تر شده اند.



شکل ۳-۴۰ : مربع انرژی های بانک log mel-filter برای گفتار پاک و نویزی [۲۵]

در شکل ۳-۴۱ ،منحنی آبی مربوط است به اندازه ی تفاوت بین اولین ۱۳ ضریب DCT برای دو نمونه ی logMelFBS نشان داده شده در شکل ۳-۳۹ برحسب درصد؛ و منحنی قرمز مربوط است به اندازه ی تفاوت بین اولین ۱۳ ضریب DCT برای توان دوم logMelFBS (شکل ۳-۴۰). اینکه منحنی قرمز کمتراز منحنی آبی است، نشان دهنده ی این است که به توان دو رساندن logMelFBS حساسیت ضرایب DCT پایین تر را به سوی ریپل های جعلی در محدوده ی انرژِ ی کم، کاهش می دهد.



شکل ۳-۴۱: [۲۵]

k امین ضریب DCT برای یک رشته ی N تایی x را در نظر بگیرید. برای مثال اگر N=10 , k=5 باشد :

$$\begin{aligned} X_{DCT}(k) &= \sum_{n=0}^{N-1} cos(\pi kn/N)x(n) \\ &= \sum_{n=0}^{9} cos(\pi 5n/10)x(n) \\ &= x(0)/2 + \sum_{n=1}^{4}(-1)^n \frac{x(2n)-x(2n-2)}{(2n)-(2n-2)} + x(8)/2 \\ &\simeq \sum_{n=1}^{4}(-1)^n x'(2n-1), \end{aligned} \qquad (۴۵-۳)$$

که $x'(n)$ مشتق گسسته ی x می باشد.

در نتیجه حساسیت DCT در logMelFBS را می توان تقریبا به صورت حساسیت مشتقات logMelFbs اندازه گرفت.

نماد حساسیت $\circ(a,b)$ را به عنوان نسبت مشتقات تابع log(x) در یک انرژی mel -فرمنت x=a و یک مقدار انرژی پایین بانک فیلتر mel ، x=b تعریف می کنیم.

با توجه به فرمول (۳-۴۵)، انتظار داریم $\circ(a,b)$ سهم نسبی یک قله ی logMelFBS و انرژی های پایین بانک فیلتر mel ، را در یک ضریب DCT که همان ضریب cepstrall است، را اندازه گیری نماید.

$$\begin{aligned} \rho(a,b) &= \frac{log'(x)|_{x=a}}{log'(x)|_{x=b}} = \frac{1/a}{1/b} \\ &= b/a \qquad \text{که} \quad a \gg b \\ &\Rightarrow \rho(a,b) \ll 1.00 \end{aligned} \qquad (۴۶-۳)$$

به طور مشابه ضریب حساسیت $\pm(a,b)$ را به صورت نسبت مشتقات تابع $sign(\log(x))[\log(x)]^P$ در یک انرژی mel – فرمنت x=a و یک مقدار انرژی پایین بانک فیلتر mel، x=b، تعریف می کنیم.

$$\begin{aligned} \sigma(a,b) &= \frac{P[sign(log(a))][log(a)]^{P-1}/a}{P[sign(log(b))][log(b)]^{P-1}/b} \\ &= \frac{[sign(log(a))][log(a)]^{P-1}}{[sign(log(b))][log(b)]^{P-1}}(b/a) \\ &= \frac{[sign(log(a))][log(a)]^{P-1}}{[sign(log(b))][log(b)]^{P-1}}\rho(a,b) \qquad \text{که} \quad a \gg b \\ &\Rightarrow \sigma(a,b) > \rho(a,b) \; where \; a \gg b, P > 1.0 \end{aligned} \qquad (۴۷-۳)$$

مقدار $1.0 << a$ در فرمول $\pm(a,b)$ به این معنی است که یک واحد تغییر در مقدار انرژی کم بانک فیلتر mel ، یعنی "b" ، در مقایسه با یک واحد تغییر در انرژی mel – فرمنت , "a" ، تاثیر بسیار بیشتری بر محاسبه ی تبدیل کسینوسی گسسته ی logMelFBS دارد. در نتیجه، می توان از فرمول $\pm(a,b)$ دریافت که تبدیل کسینوسی گسسته logMelFBS نسبت به اختلالات در محدوده ی انرژی پایین در مقایسه با محدوده های اطراف فرمنت ها، بسیارحساس اند. به هر حال، برای $1.0 \le b << a < \infty$ و P>1 ، $\pm(a,b)$ همواره از $\pm(a,b)$ بزرگتر است.

این را می توان با به کارگیری $(\log(x+1))^P$ که x انرژی طیف توان است که هیچ گاه مقدار منفی را نمی گیرد ، تحصیل کرد. این حقیقت که $\pm(a,b)$ همیشه از $\pm(a,b)$ بزرگتر است به این معنی است که توانسته ایم حساسیت ضرایب cepstrum را نسبت به اختلالات کم انرژی کاهش دهیم. یک پارامتر مهم در فرمولی که در بالا بیان شده است، نمای (توان) P می باشد. همان طور که از فرمول $\pm(a,b)$ می توان فهمید، حساسیت نسبت $\pm(a,b)$ با افزایش P به طورنمایی افزایش می یابد. با این وجود، مقدار بزرگ P باعث می شود که مدولاسیون طیفی بزرگترین فرمنت ها مقادیر عددی بسیار بزرگی را داشته باشند که مدولاسیون طیفی ناچیزی برای سایر فرمنت ها از لحاظ عددی نسبت به بزرگترین فرمنت ها ارائه می دهد. در نتیجه یک مقدارمتوسط P برای این گونه پردازش مناسب ترین است. [۲۵]

نتایج آزمایشات برمجموعه OGI Numbers95 در جدول ۳- ۲ آورده شده است.

SNR	MFCC	RMFCC	ExpoMFCC
Clean	6.1	6.1	6.2
Fact SNR 12	14.0	12.0	11.6
Fact SNR 6	31.5	20.6	20.3
Fact SNR 0	75.7	45.7	44.3
F16 SNR 12	15.8	12.3	12.1
F16 SNR 6	32.8	20.8	20.9
F16 SNR 0	75.1	44.2	43.4

جدول ۳-۲ : نرخ خطای کلمه برای نویز کارخانه و f16 . بهترین نتایج برای RMFCC(R=0.10) و ExponentiatedMFCC(P=2.7) گزارش شده است.

۳-۷-۴ روش استخراج ویژگی (A-MFCC)،

Autocorrelation Mel Frequency Cepstral Coefficients

این روش برای تشخیص گوینده، در زمانی که سیگنال گفتار با نویز جمع پذیر خراب شده باشد، مناسب می باشد که در اینجا مختصری به معرفی این روش می پردازیم.[۲۶]

عدم مطابقت شرایط محیطی بین داده های آموزش و آزمایش، عملکرد سیستم تشخیص گوینده را تنزل می دهد. تنزل عملکرد در درجه اول بدلیل حضور نویز پس زمینه در زمانی است که سعی بر تطابق گوینده ای به مجموعه گوینده های شناخته شده در پایگاه داده می باشد.

MFCC ها شاید پرکاربردترین وسیله در سیستم های تشخیص گوینده می باشند. یکی از موضوعات مهم در رابطه با MFCC ، حساسیت زیاد آن نسبت به نویز جمع پذیر می باشد. برای غلبه بر این مشکل، روش فیلتر کردن زمانی بررشته ی خود همبستگی، برای حداقل کردن تاثیر نویز جمع پذیر پیشنهاد شده است. ویژگی پیشنهادی Autocorrelation Mel Frequency Cepstral Coefficients (A-MFCC) نامیده می شود که از فیلتر کردن زمانی رشته ی خودهمبستگی یک طرفه کوتاه مدت بدست می آید. این عمل فیلتر کردن اثر نویز جمع پذیر را حداقل می کند. هیچ اطلاعات اولیه ای از مشخصات نویز لازم نمی باشد. نویز جمع پذیر می تواند یک نویز رنگی هم باشد.

ورودی یک سیستم تشخیص گوینده نمونه های صحبت هستند و خروجی شاخص (نماد) گوینده ی مشخص شده می باشد. سه جزء اصلی در سیستم تشخیص گوینده وجود دارد: استخراج ویژگی، مدل گوینده، و الگوریتم تطابق.

جزء استخراج ویژگی، مجموعه برداری از مشخصه های گوینده است که از سیگنال ورودی بدست می آید.

سپس از رو ی همین بردارها مدل هر گوینده تولید می شود.

روند تطابق بین مدل گوینده ها مقایسه ای انجام می دهد. انتظار می رود که استخراج ویژگی بحرانی ترین جزء سیستم باشد ولی این قسمت از لحاظ طراحی بسیار پیچیده تر از روند تطابق می باشد.

عدم مطابقت شرایط محیطی بین داده های آموزش و آزمایش به طور جدی عملکرد سیستم های تشخیص گفتار و گوینده را کاهش می دهد. عدم تطابق بین محیط های آموزش و آزمایش ناشی از نویز پس زمینه و اعوجاج کانال می باشد.

پایداری سیستم تشخیص گوینده می تواند از سه راه انجام گیرد: اولا استفاده از تقویت گفتار برای افزایش نسبت سیگنال به نویز. ثانیا استخراج پارامترهای پایدار سیگنال های گفتار برای حداقل کردن تاثیر نویز برسیگنال گفتار. ثالثا استفاده از تکنیک جبران مدل برای تنظیم دینامیکی مدل گفتار پاک (بدون نویز) برای محیط های نویزی.

تکنیک های زیادی برای غلبه بر این مسئله تنزل عملکرد پیشنهاد شده است مانند

Parallel Model Compensation (PMC) و stochastic matching (SM) و .... گرچه تکنیک های متعددی می توانند عملکرد قابل مقایسه ای را ارائه دهند ولی نقاط ضعفی کاربرد عملی آن ها را محدود می کند.

برای مثال تکنیک Spectral Subtraction و Parallel Model Compensation (PMC) نیاز به آگاهی اولیه در باره ی مشخصات نویز دارند.

در مقاله ی[۲۶] روشی برای حذف تاثیر نویز برپایه ی ایده ی فیلتر کردن زمانی در حوزه ی خودهمبستگی پیشنهاد شده است و ضرایب Mel frequency cepstral از آن بدست می آید که A-MFCC نامیده می شوند.

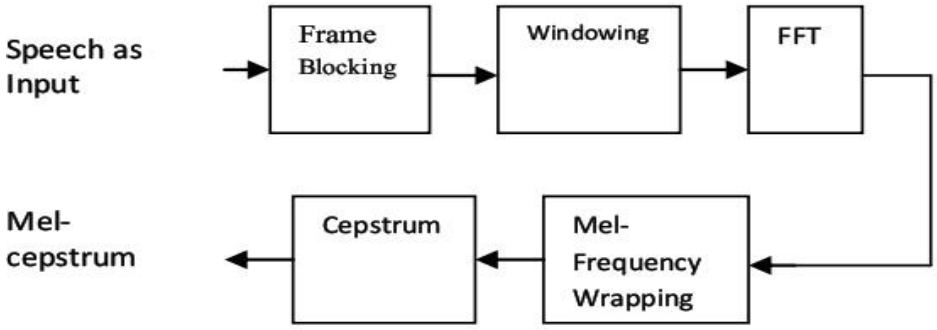
هنگامی که یک گفتار با نویز جمع پذیر خراب می شود، نویز نه تنها در حوزه ی طیف توان با گفتار جمع پذیر است بلکه در حوزه ی خودهمبستگی نیز جمع پذیر می باشد. به جای کاهش (تفریق) نویز در محدوده ی طیف توان، در محدوده ی خودهمبستگی بر اساس روش فیلتر زمانی کردن ، نویز را حذف می کنیم.

- MFCC

MFCC شناخته شده ترین و محبوب ترین ویژگی می باشد. MFCC با فیلترهایی که به طور خطی در فرکانس های پایین و به طور لگاریتمی در فرکانس های بالا، قرار گرفته اند، برپایه ی تغییرات شناخته شده ی پهنای باندهای بحرانی گوش انسان با فرکانس هایش می باشد. این موضوع درمقیاس Mel-frequency به صورت فاصله گذاری خطی برای فرکانس های کمتر از 1000 hz و فاصله گذاری لگاریتمی برای فرکانس های بیشتر از 1000 hz ظاهر شده است.

MFCC داده ها را به طور قابل توجه ای کاهش می دهد؛ زیرا ضرایب کمی برای نمایش cepstrum سیگنال صوتی کافی می باشد. در شکل۳- ۴۲ محاسبه ی MFCC نشان داده شده است. سیگنال گفتار به قاب هایی که هرکدام شامل ۲۵۶ نمونه می باشد تقسیم می شود و یک فیلتر پیش – تاکید برهر قاب اعمال می گردد.

ضریب پیش – تاکید استفاده شده، 0.9375 می باشد. یک پنجره Hamming برای حداقل کردن ناپیوستگی های سیگنال در ابتدا و انتهای هرقاب به کاربرده شده است و سپس FFT محاسبه می شود. در نهایت، با عبور دامنه ی بدست آمده از بانک فیلتر mel و سپس تبدیل کسینوسی گسسته گرفتن از آن، ضرایب MFCC بدست می آید.

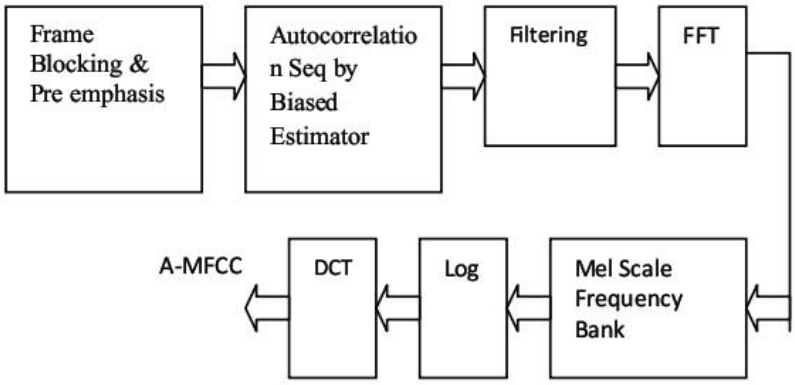


شکل ۳-۴۲: محاسبه ی MFCC در [۲۶]

- استخراج ویژگی پایدار(A-MFCC)

در [۲۶] روش جدیدی پیشنهاد شده، که مبتنی بر بهینه کردن قله های بدست آمده از طیف خودهمبستگی سیگنال گفتار می باشد. این روش قله های طیفی خودهمبستگی را حفظ می کند.

ابتدا، خودهمبستگی سیگنال نویزی محاسبه می شود. بدلیل اینکه خودهمبستگی زمانی نویز یک DC و یا سیگنالی با تغییرات بسیار آرام و اندک می باشد، تاثیر آن با یک فیلتر بالا گذر حذف می گردد. رشته ی خودهمبستگیِ سیگنال قاب ، با استفاده از یک تخمین زننده ی بایاس شده بدست می آید. سپس یک فیلتر در حوزه ی زمانی به رشته ی خودهمبستگی اعمال می شود تا رشته ی خودهمبستگی نسبی برای حذف نویز جمع پذیر بدست آید. مجموعه ای از ضرایب MFCC پایدار از دامنه ی(اندازه) طیف توانی خودهمبستگی نسبی، با اعمال آن به بانک فیلتر های mel-frequency مرسوم و عبور لگاریتم آن از بلوک DCT، بدست می آیند.



شکل ۳- ۴۳ : بلوک دیاگرام پردازشگر A-MFCC در [۲۶]

مجموعه ی داده ها ونتایج آزمایشی در [۲۶] : یک پایگاه داده ی دیجیتال که از ۲۰۰ کلمه Hindi که توسط ۳۰ گوینده بیان گردیده، استفاده شده است. نمونه های گفته شده توسط ۱۵ مرد، ۱۰ زن و ۵ کودک ضبط شده است. هر گوینده ۵ تکرار از کلمات را تلفظ می کند.

ضبط صداها برروی یک ضبط صوت در یک استادیو صورت گرفته است. و SNR>40dB

ابتدا عملکرد MFCC , A-MFCC را در زمانی که داده های آموزش و آزمایش در محیط بدون نویزی (پاک) هستند (با فرض حدود 40 dB نسبت سیگنال به نویز) ارزیابی می کنیم. مشاهده می کنیم که نرخ شناسایی برای MFCC و A-MFCC تقریبا یکسان است.

Feature type	Recognition rate (%)
MFCC	98.24
A-MFCC	99.27

جدول ۳- ۳ : نرخ شناسایی گوینده (٪) برای گفتار پاک (بدون نویز)

سپس آزمایش را بر داده های نویزی تکرار می کنیم:

عبارت ها ی نویزی شده ی آزمایش با اضافه کردن نویز مصنوعی در ۵ سطح مختلف SNR تولید می شوند. نویز سفید با یک برنامه ی تولید عدد تصادفی (random number generator program)، ایجاد می گردد و سایر نویزهای رنگی مانند نویز کارخانه، نویز F16 از مجموعه ی NATO RSG-10 استخراج شده اند. نویز ها به سیگنال پاک (عاری از نویز) برای SNR های 20,15,10,5,0 dB اضافه می شود. نتیجه ی سیستم شناسایی در جدول ۳-۴ نشان داده شده است. از جدول ها مشهود است که A-MFCC در مقابل نویز جمع شونده کاملا پایدارتر است. [۲۶]

Feature type	Noise levels (dB)					
	40	20	15	10	5	0
MFCC	98.2	83.8	55.8	29.3	10.5	3.7
A-	99.2	85.8	58.9	34.8	15.0	7.4

Feature type	Noise levels (dB)					
	40	20	15	10	5	0
MFCC	99.24	84.10	56.17	30.18	11.50	4.20
A-	99.31	86.11	58.90	35.16	16.15	8.20

Feature type	Noise levels (dB)					
	40	20	15	10	5	0
MFCC	98.0	83.2	57.1	31.4	10.8	3.7
A-	98.9	85.9	59.1	35.7	14.9	7.2

جدول ۳-۴) نرخ شناسایی گوینده (٪) برای گفتار آموزشی خراب شده با نویزهای، به ترتیب سفید و کارخانه و F16

۳-۷-۵ Relative MFCC برای شناسایی گفتار تلفنی پایدار بر اساس جبران کانال

ثابت شده است که تکنیک جبران کانال یک روش موثر برای شناسایی گفتار پایدار می باشد. در [۲۷]، عملکرد روش پیشنهادی (RMFCC-relative MFCC) با سایر روش های سابق جبران کانال CMS(Cepstral mean subtraction),2-levelCMS,RASTA(RelAtive SpecTrAl) برای شناسایی گفتار تلفنی پایدار مقایسه شده است.

استفاده از RMFCC باعث کاهش ۳۹.۸ درصدی نرخ خطای کلمه ای نسبت به سیستم HMM مرسوم، شده است. آزمایشات نشان می دهند که در مقایسه با RASTA ، RMFCC بدون از دست دادن دقت، پیچیدگی محاسبات را کاهش می دهد. و نیز عملکرد آن نسبت به CMS,2-level CMS بهتر می باشد.

سیگنال گفتار نه تنها پیام های زبان شناسی را حمل می کند بلکه تحت تاثیر سایر منابع اطلاعاتی نیز می باشد. یکی از مخرب ترین منبع تغییرات غیر – زبان شناسی، محیط ارتباطات می باشد، که معمولا شامل اتاق ضبط، میکروفن و مهم ترین جزء یعنی کانال ارتباطی مانند خط تلفن می باشد. عملکرد یک سیستم شناسایی اتوماتیک (ASR) زمانی که شناسنده دریک محیطی به غیر از محیطی که درآن آموزش داده شده است، به کار گرفته می شود، به شدت کاهش می یابد. با وجود آن که ممکن است کاهش عملکرد معمولا به تاثیرات غیر خطی محیط و نویز جمع پذیر در سیگنال نسبت داده شود ولی مشخصه ی فرکانسی کانال ارتباطی به تنهایی می تواند طیف زمان – کوتاه سیگنال را شدیدا تحت تاثیر قرار دهد. بدلیل اینکه بیشترین معیارهای همانندی (همسانی) ای که در سیستم های ASR اعمال می شوند، به طور مستقیم و یا غیر مستقیم براساس طیف زمان – کوتاه گفتار می باشند، عملکرد ASR می تواند به طور قابل توجهی تحت تاثیر مشخصه ی فرکانسی کانال ارتباطی قرار گیرد.

طبق بعضی از گزارشات، نرخ خطا در یک شناسنده ی گفتار هنگامی که داده های آزمایش با یک فیلتر قطب/صفری که یک خط تلفن بلند را مدل می کند، فیلتر می شوند، می تواند از % 1.3 تا % 44.6 افزایش یابد. بنابراین، یافتن روش ها ی جبران کانال پایدار برای کاربردهای عملی شناسایی گفتار تلفنی، یک عامل مهم می باشد.

پایداری شناسایی گفتار به طور گسترده ای مورد تحقیق و مطالعه قرار گرفته است، و تعدادی از روش های جبران کانال پیشنهاد شده است.

CMS، که ابتدا میانگین cepstral هر عبارت را محاسبه می کند و سپس میانگین cepstral از ضرایب cepstral هر قاب کم می شود، از لحاظ سادگی الگوریتم موثری است. با این حال، کارآیی CMS در زمانی که کانال را نمی توان به طور مناسبی به صورت خطی مدل کرد، به شدت محدود می شود.

برای پردازش کانال های غیر خطی،2-level CMS پیشنهاد شده است، که ابتدا سیگنال گفتار ورودی به دو قسمت دسته بندی می شود و میانگین cepstral برای هر قسمت محاسبه می شود، سپس میانگین های cepstral متفاوت از ضرایب cepstral هر قسمت کم می شود. این روش به طور موثری برای جبران کانال در شناسایی رقم (اعداد) برای کاربردهای موبایل و شناسایی گوینده بکار رفته است. این روش در عملکرد از CMS بهتراست؛ با این حال، این روش به دسته بندی سیگنال نیازمند است و عملکرد آن به نتیجه ی کلاسه بندی بستگی دارد.

پردازش RASTA ، که از یک فیلتر میان گذر با فرکانس قطع بسیار کم استفاده می کند، یک روش جبران کانال کارآمد می باشد. و می تواند اعوجاج های کانال که تغییرات آرامی دارند را حذف کند و عملکرد خوبی داشته باشد. پردازش RASTA ِ مرسوم بر طیف لگاریتمی PLP(perceptual linear predictive) اعمال می شود. اما به هر حال PLP محاسبات پیچیده ای را می طلبد.

در کارهای قبلی، به جای PLP برای کاهش محاسبات از تحلیل طیفی mel استفاده شده است. براساس ارتباط خطی بین طیف لگاریتمی mel-frequency و ضرایب cepstral در mel-frequency (MFCC) ، پردازش RASTA را از طیف لگاریتمی mel-frequency به MFCC توسعه داده می شود. و یک فیلتر میان گذر مانند RASTA برای شناسایی گفتار پایدار پیشنهاد شده است. سپس، پارامتر قطب فیلتر را با انجام آزمایشات و انتخاب مقدار اولیه ی انتگرال گیر، انتخاب می گردد.

Cepstral mean subtraction (CMS)

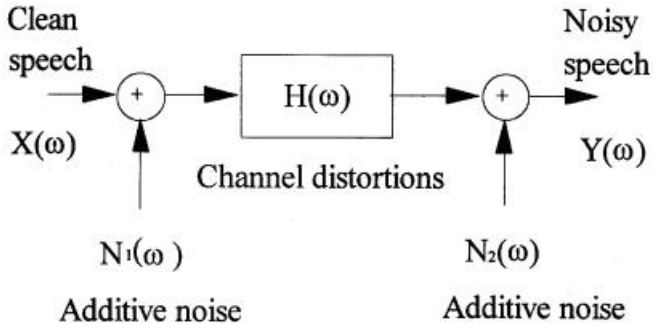
در این روش فرض براین است که مجموع متوسط رشته ی ویژگی گفتار ورودی، صفر است، و ضرایب cepstral را برای حداقل کردن عدم تطابق ناشی از اعوجاج کانال، در داده های آموزش و آزمایش، اصلاح می کند. CMS معمولا به عنوان یک روش استاندارد برای جبران کانال می باشد. که در آن میانگین بردار cepstral از ضرایب cepstral هر عبارت، کم می شود :

$$\mathrm{C}_t = \mathbf{c}_t - \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{c}_t \quad (t = 0, 1, 2, \ldots, T-1), \tag{۴۸-۳}$$

که c_t و C_t بردارهای cepstral در قاب t، به ترتیب قبل و بعد از پردازش CMS می باشند. و T تعداد کل قاب ها در عبارت هستند.

two-level CMS :

به طور کلی سیگنال گفتار نه تنها با اعوجاج کانال خراب می شود بلکه نویز جمع پذیر هم قبل از ورود سیگنال به ASR ، باعث تخریب می شود.



شکل ۳- ۴۴ : نمایش اعوجاج سیگنال گفتار در [۲۷]

در طیف توان، گفتار نویزی $Y(w)$ برابر است با :

$$Y(\omega) = [X(\omega) + N_1(\omega)]\, H(\omega) + N_2(\omega), \qquad (۴۹-۳)$$

که $X(w)$ گفتار ورودی، $N_1(\mu), N_2(\mu)$ نویز محیط و $H(w)$ اعواج های کانال می باشند.

برای شرایط آموزش و آزمایش پاک، $N_1(\mu), N_2(\mu)$ قابل چشم پوشی هستند، اعوجاج ها در حوزه ی cepstral جمع پذیر می شوند و درنتیجه CMS قسمت تغییرناپذیر با زمانِ اعوجاج های کانال را حذف می کند.

هنگامی که نویز جمع پذیر زیادی ظاهر می شود، تنها می توان جزئی از نویز در قسمت هایی از گفتار با SNR بالا را نادیده گرفت و CMS را در آن قسمت ها به کار برد. در موارد مشابه، می توان از جزئی ازگفتار در قسمت هایی با SNR خیلی پایین (مثلا گفتار با سطح کم و یا قسمت هایی که هیچ گفتاری وجود ندارد) چشم پوشی کرد. بنابراین، عملکرد شناسایی را می توان با استفاده از two-level CMS که جبران کانال جداگانه ای برای قسمتهایی که به عنوان گفتار و قسمتهایی که به عنوان پس زمینه دسته بندی شده اند، اعمال می کند؛ بهبود داد.

با داشتن مشاهدات cepstral یک قاب : $C = c_0, c_1, ..., c_{T-1}$ هر قاب به عنوان یک قابِ گفتار و یا پس زمینه دسته بندی می شود.

اگر $E = E_0, E_1, \ldots, E_{T-1}$ رشته لگاریتم انرژی مشاهدات باشند، تابع پس زمینه به صورت (۳-۵۰) می باشد:

$$bck(t) = \begin{cases} 1 & E_t < \alpha E_{\max} \\ 0 & \text{otherwise} \end{cases} \quad (t = 0, 1, 2, \ldots, T-1), \tag{۵۰-۳}$$

که $E_{\max}$ بیشترین لگاریتم انرژی مشاهدات می باشد، و پارامتر ~ثابتی است که به طور تجربی انتخاب شده است.

برای two-level CMS ، بردارهای cepstral جبران شده ی C_t محاسبه می شوند :

$$\mathbf{C}_t = \begin{cases} \mathbf{c}_t - \mathbf{C}^b & bck(t) = 1 \\ \mathbf{c}_t - \mathbf{C}^y & \text{otherwise} \end{cases} \quad (t = 0, 1, 2, \ldots, T-1) \tag{۵۱-۳}$$

که C^b, C^y به ترتیب بردارهای میانگین cepstral قاب پس زمینه و گفتار می باشند.

RASTA-PLP:

آزمایشات ادراکی نشان می دهد که درک گفتار توسط انسان ممکن است بتواند پس زمینه ی غیر- زبان شناسی ایستا را حذف کند و پیام های زبان شناسی متغیر را تقویت نماید. بنابراین، اتخاذ ویژگی بر اساس شنیده های انسان برای شناسایی گفتار پایدار مفید است.

در تکنیک RASTA-PLP ، تعدادی از خصوصیات شنیدن با تقریب ها ی مهندسی عملی، شبیه سازی شده و سپس یک فیلتر میان گذر به صورت (۳-۵۲) به نمایش cepstral لگاریتمی از گفتار اعمال می شود:

$$H(Z) = 0.1 \frac{Z^4(2 + Z^{-1} - Z^{-3} - 2Z^{-4})}{1 - 0.94Z^{-1}}. \tag{۵۲-۳}$$

صورت این فیلتر یک تخمین خطی از مشتق زمانی است، در حالی که مخرج یک انتگرال گیر ساده می باشد.

پردازش RASTA به طور موثری اعوجاج های کانال که تغییرات آرامی دارند را حذف می کند.

روش RMFCC(Relative MFCC) :

در روش RASTA - PLP ، روش RASTA تنها برای حذف اعوجاج ها ی کانال که تغییرات آرامی دارند، به کار می رود. در حالی که PLP برای شبیه سازی خصوصیات شنیداری انسان بکار می رود. تحلیل طیفی Mel نیز یک روش برای شبیه سازی خصوصیات شنیداری انسان می باشد که از PLP آسان تر می باشد.

اگر H(Z) نمایش دهنده ی فیلتر میان گذر RASTA باشد، $Y_{z,i}, \bar{Y}_{z,i}$ نشان دهنده ی i امین طیف لگاریتمی فرکانس mel در حوزه ی تبدیل Z ، به ترتیب قبل و بعد از پردازش با RASTA می باشند.

$$\bar{Y}_{Z,i} = H(Z) \cdot Y_{Z,i} . \qquad (۵۳-۳)$$

MFCC ، که به عنوان ویژگی ها در اکثر شناسنده های گفتار رایج به کار می روند، با استفاده از تبدیل کسینوسی گسسته گرفتن از طیف لگاریتمی mel-frequency محاسبه می شود.

$$\begin{aligned} C_z(k) &= \sum_{i=1}^{B} \cos\left[k(i-0.5)\frac{\pi}{B}\right] \bar{Y}_{z,i} \\ &= H(Z) \sum_{i=1}^{B} \cos\left[k(i-0.5)\frac{\pi}{B}\right] Y_{z,i} \\ &= H(Z) c_z(k) \quad (k = 1,2,3,\ldots,K), \end{aligned} \qquad (۵۴-۳)$$

که $C_z(k), c_z(k)$ k امین MFCC در حوزه ی تبدیل Z ، به ترتیب با استفاده و بدون استفاده از پردازش RASTA ، B تعداد باندهای فرکانس mel و K بُعد MFCC می باشد.

از معادله بالا منطقی به نظر می رسد که پردازش RASTA از طیف لگاریتمی به MFCC طی شود. (یعنی ابتدا MFCC محاسبه شود سپس با یک فیلتر میان گذر پردازش شوند.) معمولا B از K بزرگتراست. (در مقاله [۱۸] B=40 , K=12) و در نتیجه این نوع از MFCC نسبی (Relative MFCC) پیچیدگی های محاسباتی را کاهش می دهد.

مهم ترین قسمت پردازش RASTA فیلتر IIR می باشد:

$$H(Z) = G \frac{Z^4(2 + Z^{-1} - Z^{-3} - 2Z^{-4})}{1 - \rho Z^{-1}} . \qquad (۵۵-۳)$$

در اینجا نیز از این فیلتر استفاده شده است و باید پارامترهای این فیلتر را برای RMFCC انتخاب شود.

هنگامی که یک سیگنال ورودی X[t] از H(Z) در معادله ی بالا عبور می کند، خروجی Y[t] عبارت است از :

$$Y[t] = G \sum_{n=0}^{4} (n-2) X[t+n] + \rho\, Y[t-1], \qquad (۵۶-۳)$$

که t=0,1,2,…,T-1 برچسب قاب ها هستند و مقدار اولیه ی Y[-1] باید انتخاب شود.

پایگاه داده :

داده های آموزش دارای عبارت هایی از ۴۰ گوینده (۲۲ مرد و ۱۸ زن) و داده های آزمایش عبارت هایی از ۴۰ گوینده ی متفاوت (۲۲ مرد و ۱۸ زن) می باشد(سیستم مستقل از گوینده است.) هر گوینده ۹۳ جمله را چندین بار می خواند و سپس ۸۴ کلمه ی جدای کُره ای به طور دستی جدا می شود و برچسب گذاری می شوند. تعداد کل عبارت ها برای داده های آموزش ۱۱۳۸۱و برای داده های آزمایش ۸۰۳۶ می باشد.

برای ارزیابی پایگاه داده از SNR به عنوان معیار اندازه گیری عینی استفاده می شود.

Measurement	Training database	Testing database
SNR	14.07 dB	13.95dB
SEGSNR	13.79 dB	13.85dB
MAXSNR	19.78 dB	19.00dB

جدول ۳- ۵ : اندازه های مختلف SNR برای پایگاه داده

ابتدا سیگنال گفتار با فرکانس نمونه گیری 8KHz دیجیتال می شود، یک فیلتر پیش- تاکید $H(z)=1-0.95z^{-1}$ به نمونه های گفتار اعمال می شود و یک پنجره Hamming با ۲۴۰ نمونه (30 ms) بر هر 15ms اعمال می شود. طیف توان سیگنال پنجره گذاری شده، در هر قاب با استفاده از یک تبدیل فوریه گسسته ۲۵۶ نقطه ای، محاسبه می شود. و 40 ضریب طیفی $\text{mel}-\text{frequency}$ براساس فیلترهای میان گذر mel-frequency بدست می آید. سپس ۱۲ MFCC با استفاده از DCT محاسبه می شود.

در نهایت، یک شناسنده ی HMM مرسوم با چگالی پیوسته، برای مقایسه عملکرد سیستم به کار می رود.

نتایج بدست آمده در جدول ۳-۶ آمده است : [۲۷]

Method	Training database	Testing database
Baseline	6.5%	11.8%
Delta-MFCC	3.4%	9.9%
CMS	2.7%	7.8%
Two-level CMS	2.5%	7.2%
RASTA	2.1%	7.1%
RMFCC	2.3%	7.1%
CMS + RMFCC	2.3%	7.1%
Two-level CMS + RMFCC	2.3%	7.1%

جدول ۳-۶ : نرخ خطای کلمه با به کارگیری انواع روش های جبران کانال

فصل چهارم:

روش پیشنهادی برای

تخمین سن افراد با استفاده از گفتار

۴-۱ : مقدمه

در این فصل الگوریتم پیشنهادی جهت تخمین سن افراد معرفی می گردد. این الگوریتم بر اساس روش استخراج ویژگی RootMFCC می باشد. هم چنین در الگوریتم مورد نظر سعی بر جداسازی واج های (phoneme) سیگنال گفتار و در نهایت استخراج ویژگی از سیگنال مربوط به هر واج می باشد.

هدف از ارائه ی این الگوریتم مقایسه ی عملکرد سیستم های تخمین سن سابق با عملکرد سیستم پیشنهادی، و بررسی تاثیر جداسازی واج های سیگنال گفتار و هم چنین به کارگیری روش استخراج ویژگی RootMFCC در بهبود نتایج بدست آمده در سیستم های تخمین سن افراد با استفاده از گفتار، می باشد.

در قسمت استخراج ویژگی، به منظور مقایسه عملکرد سیستم به ازای روش های مختلف استخراج ویژگی، علاوه بر روش RootMFCC از روش های MFCC , LPC , PLP هر کدام به طور مستقل نیز استفاده شده است.

در قسمت کلاسه بندی نیز، در حالت کلی از رده بند خطی ساده ای که در نرم افزار Matlab موجود است، استفاده شده است. علاوه بر روش linear classifier از کلاسه بندی به روش Mahalanobis distance نیز استفاده شده است تا با مقایسه نتایج بدست آمده کلاسه بند مناسب انتخاب گردد.

در حالت کلی تمام افراد در ۳ گروه سنی با بازه ی ۲۰ سال تقسیم بندی شده اند.

در نهایت برای مشاهده ی عملکرد سیستم بازای ۵ گروه سنی با بازه ی ۱۲ سال و ۲ گروه سنی (کمتر و بیشتر از ۳۵ سال) نیز آزمایشات صورت گرفته است .

۴-۲: معرفی پایگاه داده :

در این پایان نامه از پایگاه داده فارس‌دات (FarsDat) دادگان گفتاری استاندارد برای زبان فارسی استفاده شده است. در زیر مختصری درباره ی این پایگاه داده آورده ایم.

یکی از مهم‌ترین ابزارها برای طراحی و ساخت سیستم‌های هوشمند شناسایی گفتار، تهیه دادگان گفتاری برای آموزش مدل‌های آکوستیکی صداها و آواهای تشکیل‌دهنده یک زبان است. برای تعلیم صداها و آواهای تشکیل‌دهنده یک زبان، تعداد بسیاری زیادی نمونه از این آواها که توسط گویندگان مختلف ادا شده است، باید توسط انسان و بصورت دستی و با دقت میلی‌ثانیه از گفتار آنها استخراج شود.

دادگان فارس‌دات (FarsDat)، مجموعه‌ای از عبارات و جملات است که توسط گویندگان فارسی‌زبان 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. بدون تهیه این دادگان، شروع تحقیقات وسیع و گسترده در زمینه بازشناسی گفتار پیوسته فارسی با دایره لغات بزرگ (LVCSR) در دانشگاه‌های تراز اول و مراکز تحقیقاتی کشور میسر نبوده است. این دادگان، به عنوان دادگان استاندارد گفتاری زبان فارسی در داخل و خارج کشور شناخته شده و برای آموزش سیستم‌های هوشمند تشخیص گفتار استفاده می‌شود. اهمیت طراحی و ساخت این دادگان در زبان فارسی، همانند اهمیت دادگان TIMIT در زبان انگلیسی است. در زبان فارسی، دادگانی با تنوع، وسعت و کیفیت برچسب‌زنیِ دادگان فارس‌دات وجود ندارد. البته این مهم با تلاش و کوشش مجموعه‌ای از محققین دانشگاه‌های برتر کشور میسر گردیده که اولین و مهم‌ترین گام را در شناسایی گفتار پیوسته در زبان فارسی برداشته‌اند.

دادگان فارس‌دات دارای ویژگیها و قابلیت‌های زیر است:

۱- استخراج پرکاربردترین کلمات زبان فارسی از روزنامه‌ها

۲-طراحی ۳۸۶ جمله با استفاده از ۱۰۰۰ کلمه شامل کلیه دنباله‌های دوآوایی در زبان فارسی

۳-متوازن بودن ۳۸۶ جمله مذکور از لحاظ آوایی (Phonetically-Balanced)

۴- انتخاب ۳۰۴ گوینده بر حسب جنسیت، سن، میزان تحصیلات و لهجه از نقاط مختلف کشور

۵-پوشش ۱۰ لهجه رایج فارسی در کشور (تهرانی، ترکی، اصفهانی، جنوبی، شمالی، خراسانی، بلوچی، کردی، لری و یزدی)

۶- تولید ۲۰ جمله در ۲ جلسه توسط هر گوینده با کیفیت صوتی بسیار بالا در اتاقک ضد صدا

۷- تقطیع دستی جملات به آواهای تشکیل‌دهنده آنها و تهیه برچسب‌های واجی و آوایی برای آن

۸- ضبط جملات با فرکانس 22050 HZ و با کیفیت بسیار بالا (نسبت سیگنال به نویز ۳۴ دسی‌بل)

۹- تهیه نرم‌افزار مخصوص استفاده از دادگان فارس‌دات به منظور استفاده آسان‌تر توسط کاربران

۱۰-استفاده از نشانه‌گذاری بین‌المللی IPA برای برچسب‌زنی آواها و واج‌های زبان فارسی

۱۱- تأیید کیفیت دادگان توسط مؤسسه بین‌المللی ELRA

۱۲-برطرف کردن نقاط ابهام در برچسب‌زنی توسط دستگاه Sona-Graph

۱۳-دارا بودن ساختاری شبیه به ساختار دادگان انگلیسی TIMIT (استاندارد ترین و معتبرترین دادگان گفتاری در زبان انگلیسی).

۱۴-دارای مقاله مرجع چاپ شده در کنفرانس SST'1994 و قابل ارجاع برای پژوهشگران.

۱۵-اولین و تنها دادگان برچسب‌دار در سطح آوا برای بازشناسی گفتار پیوسته فارسی در کشور

از مشخصات فنی پایگاه داده، سازگاری برنامه استفاده از فارس‌دات با سیستم عامل windows است.

- **کاربردهای farsdat:**

۱- استفاده در طراحی و ساخت سیستم‌های هوشمند تشخیص گفتار (تبدیل گفتار به متن فارسی) و تعیین هویت گوینده

۲- استفاده برای کاربردهای پژوهشی و دانشگاهی در تمامی مراکز و دانشگاه‌های فعال در زمینه پردازش گفتار

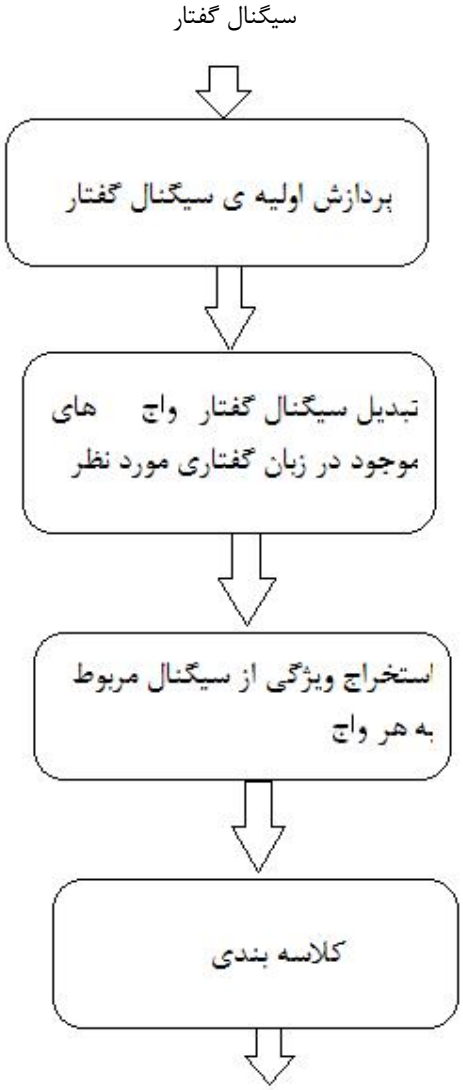
۳- استفاده در دانشکده‌های فنی-مهندسی و نیز دانشکده‌های زبان‌شناسی به عنوان دادگان استاندارد

- **افتخارات farsdat:**

- رتبه سوم تحقیقات کاربردی جشنواره بین المللی خوارزمی در سال ۱۳۷۴.

۴-۳ : معرفی الگوریتم پیشنهادی

در شکل ۴- ۱ ، به طور اجمالی روش پیشنهادی برای تخمین سن افراد نشان داده شده است.



شکل ۴-۱

۴-۴ :پیاده سازی الگوریتم پیشنهادی

در این پایان نامه از قسمت اول پایگاه داده استفاده شده است که شامل ۱۰ جمله متفاوت برای هر شخص میباشد. از کل داده های در دسترس (۳۰۴ سیگنال مربوط به ۳۰۴ گوینده) ۶۰ در صد داده ها به عنوان داده های آموزش و ۴۰ درصد با قی مانده داده های آزمایش انتخاب شده اند. کل فراد را از لحاظ گروه سنی به ۳ دسته تقسیم می کنیم ؛

گروه ۱ (h) افراد بین ۱۳_۳۳ ساله (۹۸ نفر در مجموعه آموزش و ۶۵ نفر در مجموعه آزمایش)

گروه۲ (g) افراد ۳۴-۵۴ ساله (۶۲ نفر در مجموعه آموزش و۴۲ نفر در مجموعه آزمایش)

گروه ۳(w) افراد ۵۵-۷۵ ساله (۲۲ نفر در مجموعه آموزش و ۱۵ نفر در مجموعه آزمایش)

برای هر شخص سیگنال مربوطه (که شامل ۱۰ جمله می باشد) در طی برنامه ای به واج هایphoneme موجود در زبان فارسی تفکیک شده است. تعداد واج های متفاوت تعریف شده در این پایگاه داده ۴۲ عدد می باشد که در یک فایل pdf. توسط تهیه کننده های پایگاه داده ارائه شده است.

مراد از تفکیک سیگنال صوتی افراد به واج ها این است که با توجه به اطلاعات موجود در پایگاه داده که شامل شماره ی نمونه زمانی آغازین و پایانی مربوط به هر واج در هر سیگنال است، سیگنال زمانی مربوط به هر واج که توسط گوینده بیان شده است، استخراج می گردد. در حقیقت داده های اصلی ما سیگنال های مربوط به واج ها، برای هر شخص، می باشند. برای استخراج ویژگی ها ، چندین حالت را در نظر گرفتیم :

۴-۴ – ۱ : بکارگیری روش استخراج ویژگی MFCC

الف : حالت اول

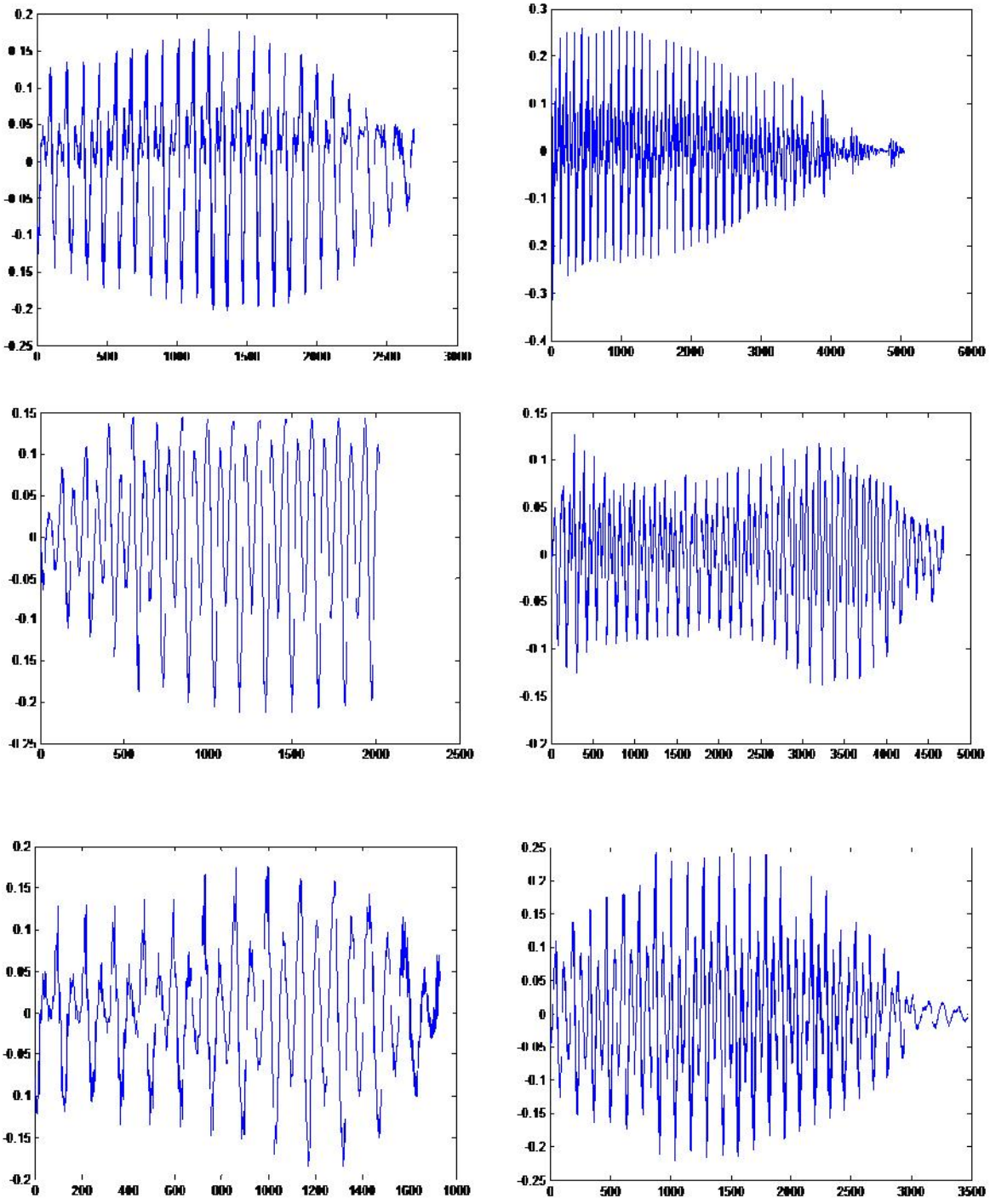
هر شخص امکان دارد یک واج مشخص را چندین بار ادا کرده باشد و یا اینکه یک واج را اصلا نگوید.

برای هرشخص، برای هر بار تکرار یک واج، سیگنال زمانی برای آن واج، که برای هر تکرار، سیگنال متفاوتی است، را بدست می آوریم. حال برای اینکه یک نماینده از این واج برای شخص مورد نظر داشته باشیم می توانیم :

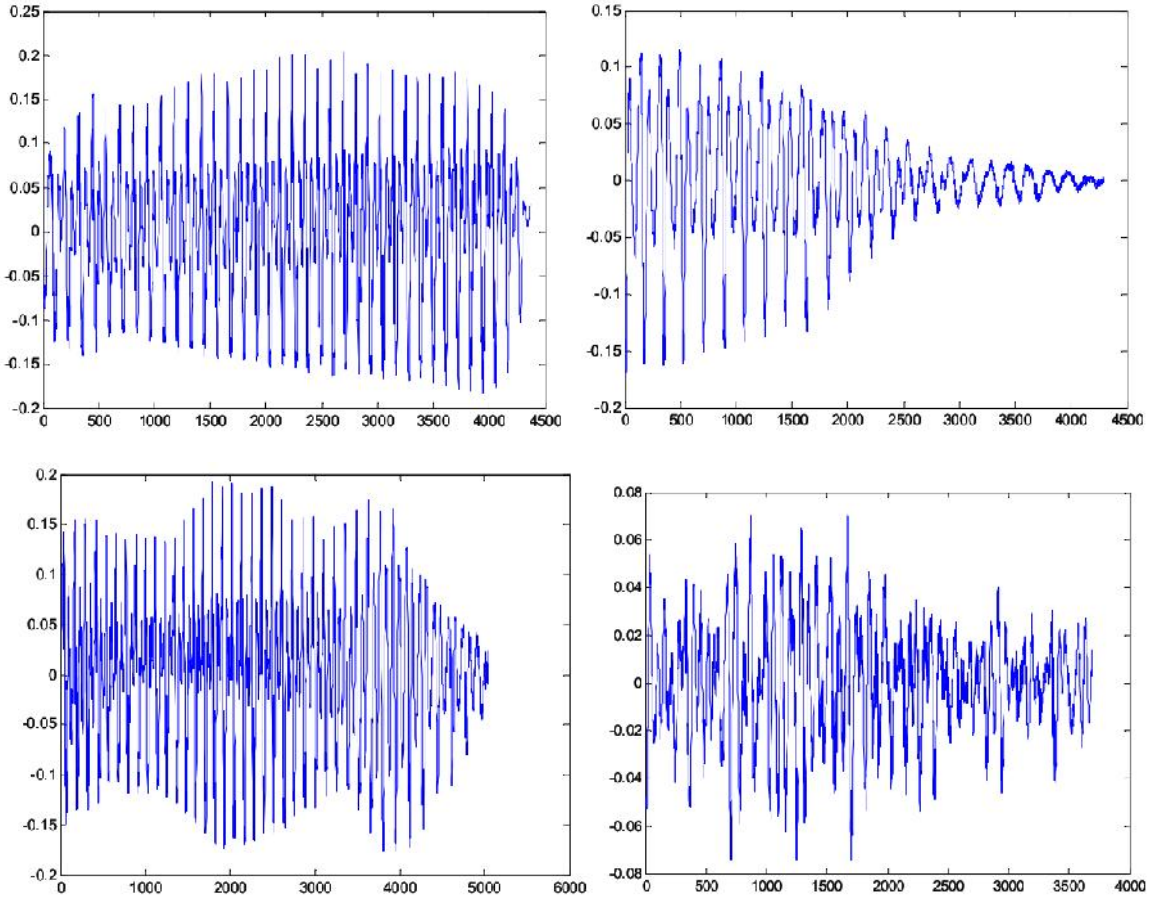
سیگنال های زمانی متفاوت بدست آمده برای یک واج را در حوزه ی زمانی میانگین گیری کنیم،

به این صورت که، متوسط طول زمانی سیگنال ها را بدست آوریم و به انتهای سیگنال هایی با طول زمانی کوچک تر از طول متوسط، صفر اضافه کنیم و از انتهای سیگنالهایی با طول زمانی بزرگتر از طول متوسط، نمونه هایی را حذف کنیم سپس از کل سیگنال های بدست آمده برای این واج مشخص در حوزه ی زمان میانگین بگیریم؛ و در نهایت یک سیگنال، که نماینده ی یک واج معین برای یک شخص است را بدست آوریم.

مثلا در سیگنال گفتار شخص شماره ۱۰۰ ، واج " ای " ۹ بار تکرار شده است که سیگنال های مربوطه و سیگنال میانگین نهایی در شکل ۴-۲ نشان داده شده است :



شکل ۲-۴ : ادامه در صفحه بعد



شکل ۴ـ ۲ : ۹ تکرار مربوط به واج "ای" که توسط شخص شماره ۱۰۰ در پایگاه داده ، ادا شده است. آخرین شکل سمت راست مربوط است به میانگین این ۹ سیگنال مربوط به واج "ای" ، که برای استخراج ویژگی از این سیگنال میانگین استفاده می شود. محور عمودی نمودار ها ، دامنه ی سیگنال و محور افقی تعداد نمونه ها(زمان) می باشد.

حال می توانیم از سیگنال زمانی متوسط بدست آمده، ضرایب MFCC و یا هر نوع دیگری از ویژگی ها را استخراج نماییم. برای بدست آوردن MFCC سیگنال را به قاب هایی به طول 25ms تقسیم می کنیم. ابتدای هر قاب به اندازه 10ms از ابتدای قاب قبلی فاصله دارد. اندازه طول قاب را به روش سعی و خطا بدست آورده ایم. سپس از ماتریس ضرایب MFCC بدست آمده که دارای۱۳ردیف و تعداد ستون هایش برابر تعداد قاب های موجود در سیگنال واج میانگین، می باشد، میانگین می گیریم تا تنها یک بردار ۱۳ تایی از ویژگی ها را داشته باشیم. پس برای هر واج، در سیگنال هر شخص یک بردار ویژگی بدست می آید.۴۲ واج متفاوت داریم، که برای هر واج از یک رده بند استفاده می کنیم.

در حقیقت برای هر نمونه ی آزمایش، ابتدا باید بردار ویژگی مربوط به واج های آن استخراج شود و سپس هر واجی در رده بند مخصوص خود دسته بندی شود. و در نهایت، با در نظر گرفتن دسته های مشخص شده برای همه ی ۴۲ واج

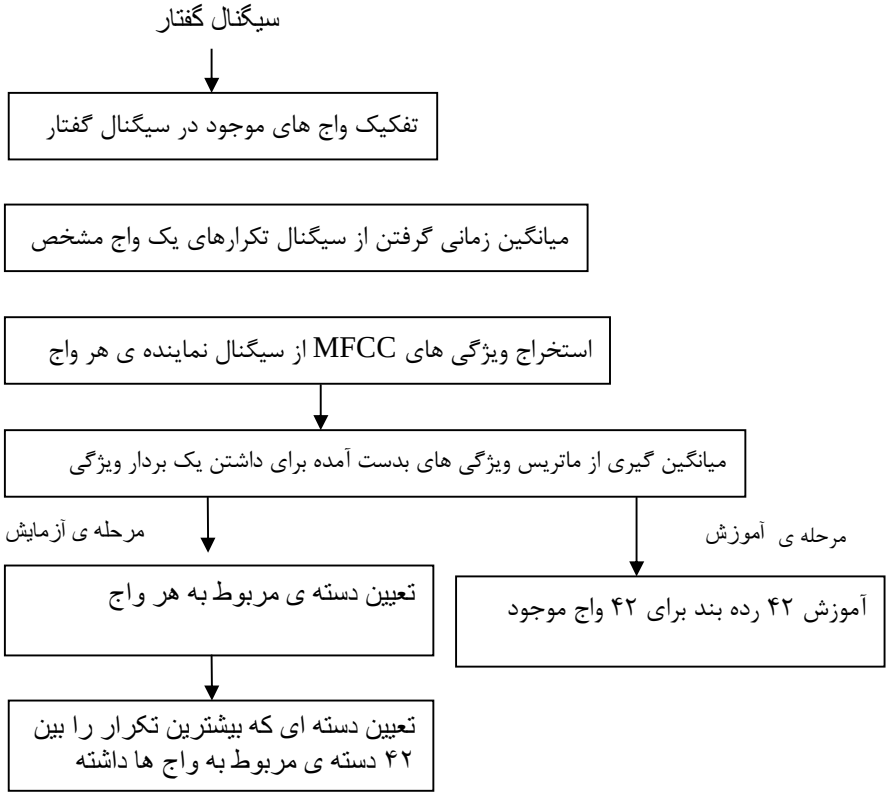
برای شخص مورد نظر، دسته ای که بیشترین تکرار را بین دسته های واج ها داشته است را به عنوان دسته سنی شخص ، در نظر می گیریم.

برای یافتن خطا ی آزمایش، تعداد نمونه های آزمایشی را که به درستی دسته بندی شده اند از کل داده های آزمایش موجود کم می کنیم . درصد خطا را هم به طور کلی و هم برای هر گروه سنی به طور جداگانه محاسبه می کنیم.

Total error : درصد خطای کل که برابر است با نسبت تعداد نمونه های آزمایشی که نادرست دسته بندی شده اند به کل داده های آزمایشی .

errorh : درصد خطا در دسته ی h ، که برابر است با نسبت نمونه های آزمایشی در دسته h که نادرست کلاسه بندی شده اند به کل نمونه های آزمایشی در دسته h . به همین ترتیب errorw , errorg نیز تعریف می شوند.

خلاصه ای از روش بیان شده در حالت اول در شکل ۴-۳ ارائه شده است :



شکل ۴- ۳ : بلوک دیاگرام مربوط به روش حالت اول

روش	errorh	Errorg	Errorw	total error
حالت اول	29.2308	26.1905	73.3333	33.61

جدول ۴-۱ : نتایج بدست آمده برای حالت اول (خطا ی آزمایش بر حسب درصد)

ب : حالت دوم

- برای هر گوینده (نمونه) سیگنال مربوط به هر واج استخراج می شود.(از هر واج ممکن است تعدادی(بیش از یک بار) موجود باشد و ممکن است هیچ گاه ازیک واج مشخصی توسط گوینده بیان نشده باشد.)
- طول زمانی سیگنال واج های مختلف و حتی واج های یکسان متفاوت است.
- برای هر سیگنال واج به طور مجزا ۱۳ ویژگی MFCC استخراج می شود.
- بدلیل اینکه طول زمانی سیگنال های مربوط به واج ها متفاوت است، ابعاد ماتریس های ویژگی حتی برای یک واج مشخص هم متفاوت است(تعداد سطرهای هر ماتریس برابر تعداد قابهای موجود در هر سیگنال واج و تعداد ستون ها برابر تعداد ویژگی ها ۱۳ میباشد.). در نتیجه از هر ماتریس ویژگی میانگین گرفته می شود و یک بردار با ۱۳ ویژگی که میانگین ویژگی ها در قاب های frame موجود در سیگنال واج است بدست می آید.

برای واج هایی که گوینده آن ها را ادا نکرده است یک بردار ۱۳ تایی با درایه های صفر در نظر می گیریم.

اگر یک گوینده، یک واج مشخص را چندین بار تکرار کرده باشد، در نهایت از بردارهای ویژگی مربوط به هر تکرار آن واج میانگین گرفته می شود.

- در نتیجه هر گوینده دارای ۴۲ بردار ویژگی است که هر کدام مربوط به یک واج مشخص است.
- بعد از مرحله ی استخراج ویژگی، نوبت به دسته بندی (classification) داده ها می رسد. در اینجا ۴۲ کلاسیفایر، برای هر واج یک کلاسیفایر، استفاده می شود.

به این ترتیب که بردارهای ویژگی مربوط به هر واج، برای همه ی افراد در مجموعه ی آموزش به عنوان داده های ورودی هر کدام از کلاسیفایرها استفاده می شوند. مثلا ۳۰۴ تا بردار ویژگی مربوط به واج (ای i) داریم که با کلاسیفایر مربوط به واج (i) دسته بندی می شوند.و به همین ترتیب برای سایر واج ها.

برای هر گوینده در مجموعه ی آزمایش، ۴۲ واج مربوط به او دسته بندی می شوند.

برای تشخیص گروه سنی گوینده مورد نظر، دسته ی مشخص شده برای هر واج را در نظر می گیریم. گروه سنی شخص برابر است با بیشترین گروه سنی ای که در ۴۲ واج مربوط به گوینده تکرار شده است. مثلا اگر برای شخص x ، ۳۰ واج متفاوت مربوط به او در دسته ۱ ، ۷ واج در دسته ۲ و ۵ واج در دسته ۳رده بندی شده باشند، در نهایت دسته (گروه سنی) شخص، دسته ۱ اعلام می گردد.

برای کاهش خطا در داده های آموزش، اگر بردار ویژگی ای صفر بود (مثلا یک گوینده یک واج مشخص را ادا نکرده باشد) در کلاسیفایر مر بوط به آن واج، این بردار ویژگی از مجموعه داده های آموزشی حذف می شود.

همچنین در داده های آزمایش، اگر گوینده ای واجی را ادا نکرده باشد، دسته مربوطه را (صفر) در نظر می گیریم که در نتیجه در انتخاب دسته ی گوینده تاثیر ی ندارد.

حالت دوم – a) در نهایت امکان دارد برای یک شخص در داده های آزمایش، به تعداد مساوی واج در ۲ یا ۳ دسته تقسیم بندی شوند که در این صورت دسته گوینده مورد نظر را، طبق اطلاعاتی که در پایگاه داده در مورد گروه سنی او وجود دارد، در نظر می گیریم.

مثلا اگر گوینده x در داده های آزمایش، از ۴۲ واج موجود در سیگنال صحبت او، ۱۷ واج در دسته ۱ ، ۱۷ واج در دسته ۲ و ۸ واج در دسته ۳ باشند، و شخص x طبق اطلاعات موجود در پایگاه داده در گروه سنی ۲ باشد، برای محاسبه خطای نهایی، شخص x را در گروه سنی ۲ در نظر می گیریم.

حالت دوم – b)و یا اینکه کلا می توانیم با حذف نمونه هایی با این مشکل از مجموعه داده های آزمایش، درصد خطا را محاسبه کنیم.

برای محاسبه ی خطا، مانند حالت اول عمل می کنیم.

روش	errorh	Errorg	Errorw	Total error
حالت دوم a	26.1538	35.7143	60	33.61
حالت دوم b	29.3103	39.4737	64.2857	37.27

جدول ۴-۲ : نتایج بدست آمده برای حالت دوم (خطای آزمایش بر حسب درصد)

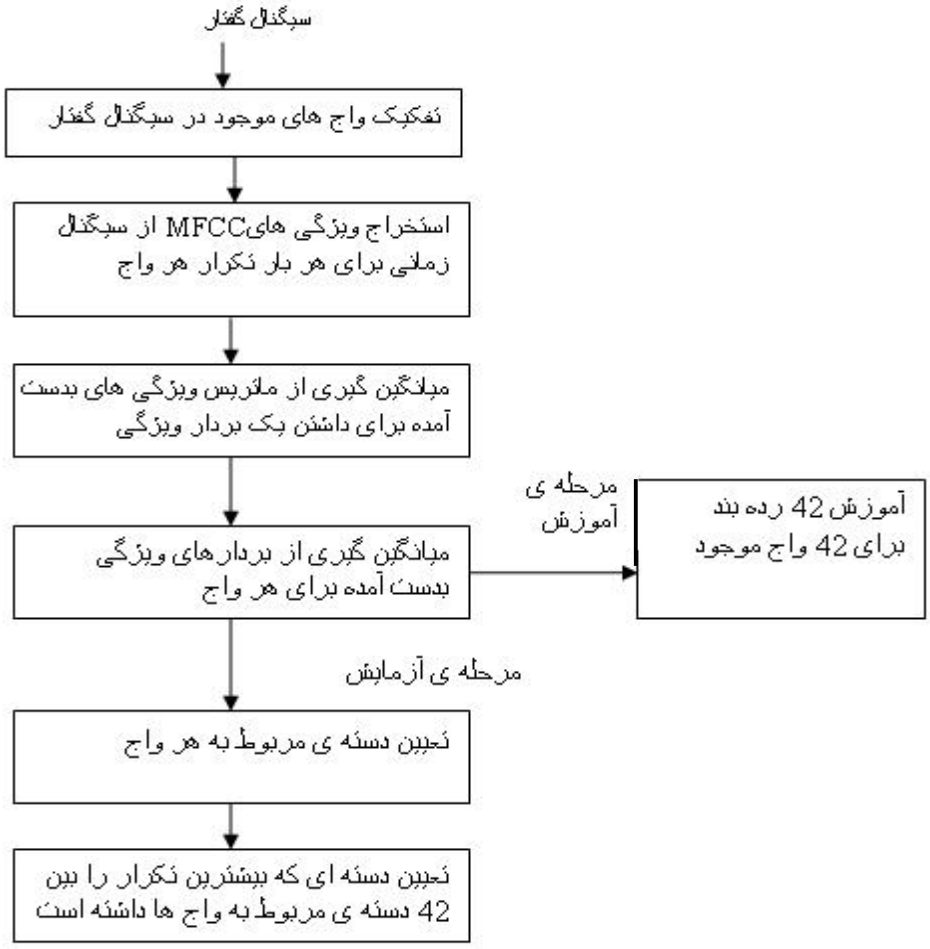
حالت دوم – c) اگر نمونه هایی را که برای یک واج مشخص اطلاعاتی ندارند (گوینده آن واج را ادا نکرده باشد)را از کل مجموعه ی آموزش و مجموعه آزمایش حذف نکنیم درصد خطا به صورت زیر تغییر خواهد کرد:

روش	Errorh	Errorg	Errorw	total error
حالت دوم c	26.1538	45.2381	53.3333	36.07

جدول ۴- ۳ :

با توجه به نتایج بدست آمده کمترین خطا برای حالت a بدست آمده است. به بیان دیگر کمترین خطا در حالتی است که اولا نمونه هایی را که واج مشخصی را ادا نکرده اند را از مجموعه های آموزش و آزمایش حذف کنیم و ثانیا در مرحله ی آزمایش، دسته ی نمونه هایی که دسته ی آنها به طور صریح مشخص نشده است، را با توجه به پایگاه داده تعیین کنیم.

خلاصه ای از روش بیان شده در حالت دوم در شکل ۴-۴ ارائه شده است :



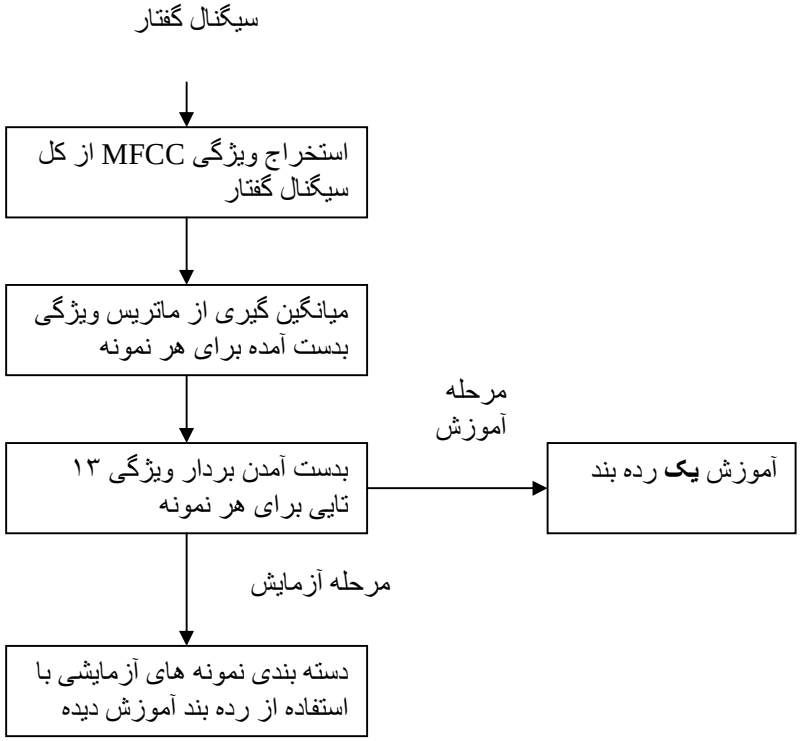
شکل ۴-۴ : بلوک دیاگرام مربوط به حالت دوم

ج : حالت سوم

استخراج ویژگی های MFCC از کل سیگنال برای هر شخص و در نهایت میانگین گیری ازماتریس ویژگی ها و بدست آوردن یک بردار ۱۳ تایی از ویژگی ها برای هر نمونه .

روش	Errorh	Errorg	errorw	total error
حالت سوم	47.6923	61.9048	80	56.56

جدول ۴- ۴ : نتایج بدست امده برای حالت سوم (خطای آزمایش بر حسب درصد)



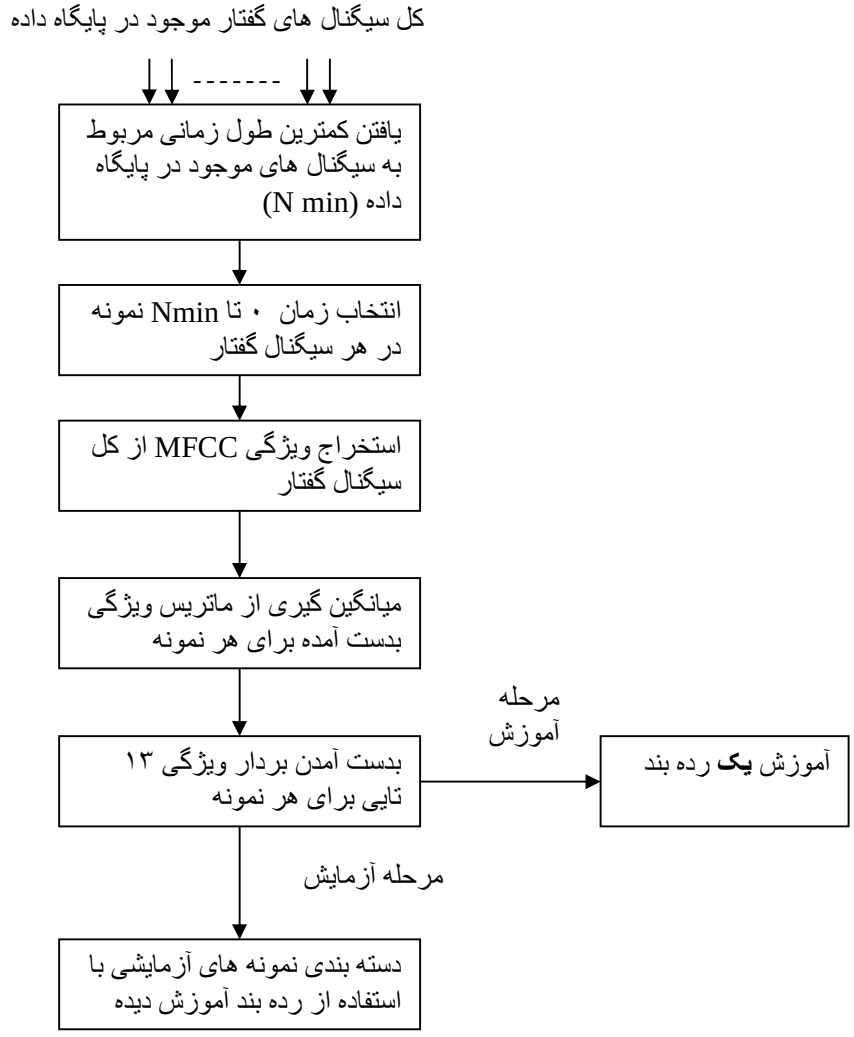
شکل ۴- ۵ : بلوک دیاگرام مربوط به روش حالت سوم

د : حالت چهارم

دراین حالت کمترین طول سیگنال موجود در پایگاه داده را در نظر می گیریم، از هر سیگنال های گفتار موجود در پایگاه داده، تنها به اندازه ی کمترین طول زمانی موجود انتخاب می کنیم و سپس مانند حالت قبل از کل سیگنال بدست آمده ویژگی های MFCC را استخراج می کنیم.

روش	errorh	errorg	Errorw	total error
حالت چهارم	49.2308	59.5238	73.3333	55.74

جدول ۴- ۵ : نتایج بدست آمده برای حالت چهارم (خطای ازمایش بر حسب درصد)



شکل ۴-۶ : بلوک دیاگرام مربوط به حالت چهارم

از نتایج بدست آمده در قسمت **۴-۴ – ۱** مشاهده می شود که :

- اگر سیگنال گفتار را به واج ها، که اجزای تشگیل دهنده اش می باشند، تفکیک کنیم و سپس از سیگنال واج ها بردار ویژگی ها را استخراج کنیم، نتیجه ی بهتری نسبت به زمانی که از کل سیگنال گفتار ویژگی ها را استخراج کنیم، بدست می آید. دلیل این امر، بدست آمدن ویژگی ها به طور دقیق تر می باشد. در واقع با عمل تفکیک سیگنال گفتار به واج هایش، ویژگی های استخراج شده اطلاعات بیشتر و دقیق تری را در اختیار ما قرار می دهند.
- همچنین، زمانی که برای یک واج مشخص در یک سیگنال گفتار معین، عمل میانگین گیری زمانی صورت می گیرد، و سپس از سیگنال متوسط منتجه، ویژگی ها را استخراج می کنیم، اطلاعات بیشتری را نسبت به حالتی که از ویژگی های بدست آمده از سیگنال هر تکرار یک واج مشخص، متوسط گیری می کنیم، از دست می دهیم. به همین دلیل نتیجه ی حاصله از حالت دوم بهتر است.

۴-۴-۲ : بکارگیری روش استخراج ویژگی LPC

الف) حالت اول

در این بخش ضرایب lpc را به عنوان ویژگی مورد نظر می گیریم. و از کل سیگنال مربوط به هر شخص ۱۲ ضریب LPC را استخراج می کنیم. (p=12)

روش	errorh	errorg	Errorw	Total error
حالت اول LPC	53.8462	59.5238	53.3333	55.74

جدول ۴- ۶ : نتایج بدست آمده برای حالت اول LPC (خطا ی آزمایشی بر حسب درصد)

ب) حالت دوم

اگر برای سیگنالهای هر واج مربوط به هر شخص ضرایب lpcرا بدست آوریم و مانند حالات بیان شده در ۴-۴–۱حالت (ب) که قبلا توضیح دادیم، از ۴۲ رده بند متفاوت، هر کدام برای ۴۲هر واج، استفاده کنیم نتیجه ی زیر بدست می آید.

a) با در نظر گرفتن دسته ی مقادیر صفر در ماتریس های آزمایش برابر دسته ی واقعی آنها در اطلاعات پایگاه داده.

b) با حذف نمونه هایی که دسته ی آنها صفر انتخاب شده است از مجموعه ی آزمایش

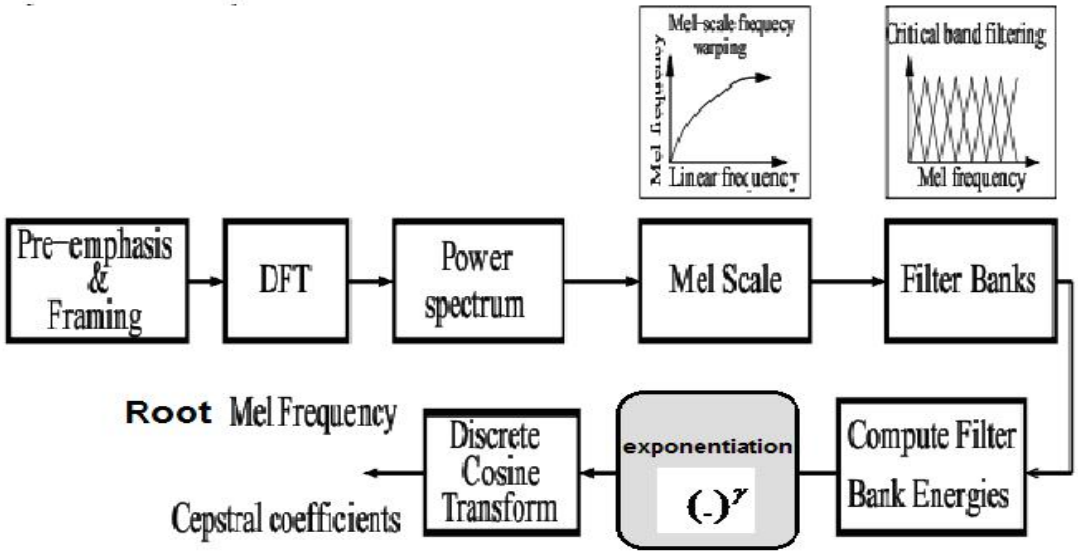
روش	errorh	errorg	Errorw	total error
حالت دوم a	23.0769	30.9524	60	30.33
حالت دوم b	25.8621	37.1429	64.2857	34.58

جدول ۴- ۷ : نتایج بدست آمده برای حالت دوم LPC (خطای آزمایشی بر حسب درصد)

۴-۴-۳ بکارگیری روش استخراج ویژگی RootMFCC

در این قسمت برای استخراج ویژگی، از روش Rootmelfcc استفاده می کنیم .

این روش همانند روش MFCC می باشد با این تفاوت که در مرحله ی بعد از اعمال بانک فیلتر بر طیف توان سیگنال ، که از بانک انرژی بدست آمده لگاریتم گرفته می شود، در روش RootMFCC به جای اعمال لگاریتم، بانک انرژی به توان رسانده می شود. این پارامتر توان را gama می نامیم.



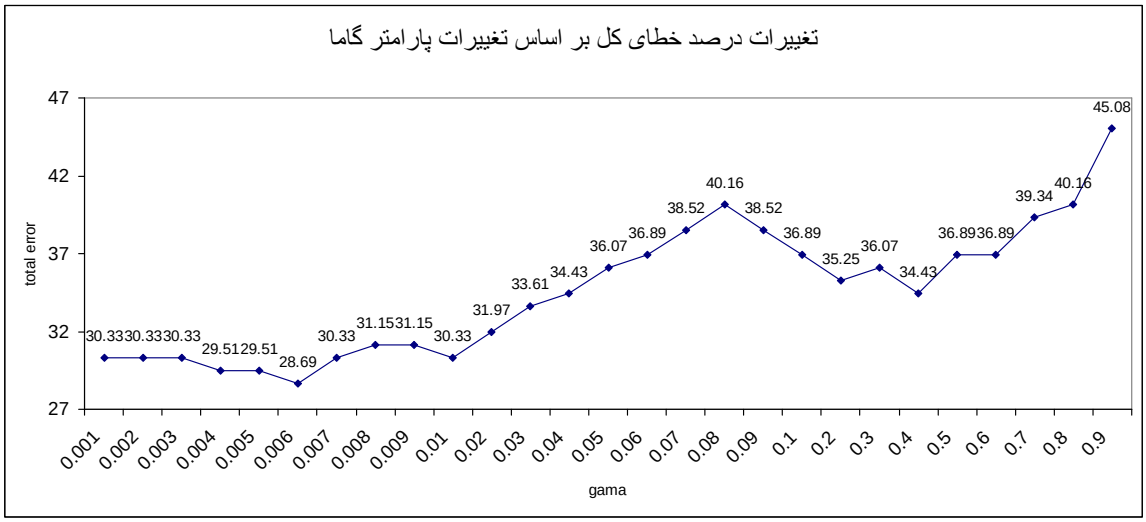
شکل ۴- ۷ : بلوک دیاگرام مربوط به محاسبه ی ضرایب ویژگی RootMFCC از مرجع [۱۱]

در اینجا نیز همانگونه که در قسمت ۴-۴-۱ حالت (ب a) صورت گرفت، ۱۳ ویژگی RootMFCC را بدست می آوریم .

در جدول ۴-۶ درصد خطا به ازای gama های مختلف، پارامتر موجود در روش RootMFCC، محاسبه شده است تا بهترین مقدار برای gama انتخاب گردد.

مقادیر مختلف پارامتر گاما	errorh	errorg	Errorw	Total error
0.001	18.4615	38.0952	60	30.33
0.002	20	35.7143	60	30.33
0.003	20	35.7143	60	30.33
0.004	20	33.3333	60	29.51
0.005	20	33.3333	60	29.51
0.006	18.4615	33.3333	60	28.69
0.007	18.4615	38.0952	60	30.33
0.008	20	38.0952	60	31.15
0.09	20	38.0952	60	31.15
0.01	18.4615	38.0952	60	30.33
0.02	20	38.0952	66.6667	31.97
0.03	23.0769	38.0952	66.6667	33.61
0.04	24.6154	40.4762	60	34.43
0.05	26.1538	45.2381	53.3333	36.07
0.06	26.1538	45.2381	60	36.89
0.07	30.7692	45.2381	53.3333	38.52
0.08	35.3846	42.8571	53.3333	40.16
0.09	35.3846	40.4762	46.6667	38.52
0.1	32.3077	35.7143	60	36.89
0 .2	27.6923	38.0952	60	35.25
0.3	33.8462	30.9524	60	36.07
0.4	29.2308	30.9524	66.6667	34.43
0.5	30.7692	38.0952	60	36.89
0.6	32.3077	30.9524	73.3333	36.89
0.7	33.8462	35.7143	73.3333	39.34
0.8	36.9231	33.3333	73.3333	40.16
0.9	36.9331	45.2381	80	45.08

جدول ۴-۸ : درصد خطای آزمایش برای روش استخراج ویژگی RootMFCC بازای گامای مختلف.



شکل ۴-۸ : نمودار تغییرات خطای کل بر حسب تغییرات گاما

در این پروژه به طور دستی پارامتر گاما را تغییر دادیم تا بتوانیم کمترین خطا را بدست آوریم. همان طور که در نمودار ۴-۸ مشاهده می شود به ازای $\gamma = 0.006$ کمترین خطا را داریم.

۴-۴-۴ بکارگیری روش استخراج ویژگی plp

در این قسمت نیز مانند روش ۴-۴-۱ حالت (ب) عمل می کنیم با این تفاوت که بجای ویژگی MFCC، ویژگی PLP را استخراج می کینم، مرتبه ی plp مورد استفاده ۹ می باشد.

انتخاب حالت (ب) در قسمت ۴-۴-۱به این دلیل است که از بین تمام حالات بیان شده در قسمت ۴-۴-۱، حالت (ب) کمترین خطا را داشته است، درنتیجه از سایر حالات ذکر شده در ۴-۴- ۱ در این قسمت استفاده نمی شود.

a) با در نظر گرفتن دسته ی مقادیر صفر در ماتریس های آزمایش برابر دسته ی واقعی آنها در اطلاعات پایگاه داده

b) با حذف نمونه هایی که دسته ی آنها صفر انتخاب شده است از مجموعه ی آزمایش

روش	errorh	Errorg	Errorw	total error
Plp a	29.2308	35.7143	60	35.25
Plp b	31.1475	44.1176	60	39.09

جدول ۴-۹ : نتایج بدست آمده برای قسمت ۴-۴-۴ (خطای آزمایش بر حسب درصد)

۴-۴-۵ اعمال روش PCA برای کاهش ابعاد ویژگی

تحلیل مولفه های اصلی یک روش کاهش متغیر می باشد. این روش زمانی مفید است که داده هایی با تعدادی متغیر (معمولا تعداد زیادی از متغیرها) بدست آورید، و معتقد باشید که تعدادی متغیر زائد در آن مجموعه ی متغیر ها وجود داشته باشد. در این زمینه، زائد بودن به این معنی است که بعضی از متغیرها به متغیر دیگری همبسته می باشند، شاید این همبستگی به دلیل این باشد که آنها ساختار یکسانی را اندازه گیری می کنند. این افزونگی (زائد بودن)، این فکر را به ذهن می رساند که شاید امکان کاهش متغیرهای مشاهده شده به تعداد کمتری مولفه های اصلی (متغیرهای ساختگی) وجود داشته باشد که به کمک متغیرهای مشاهده شده ای که بیشترین واریانس را دارند، محاسبه شوند.

مولفه های اصلی را می توان به عنوان ترکیب خطی از متغیرهای مشاهده شده ای که به طور بهینه وزن دهی شده اند، تعریف کرد .

" ترکیب خطی " به این واقعیت اشاره دارد که امتیاز یک مولفه، از جمع امتیازهای متغیرهای مشاهده شده ی مورد مطالعه با یکدیگر، ایجاد می شود. و "وزن دهی بهینه " نیز به این حقیقت بر می گردد که متغیرهای مشاهده شده به صورتی وزن دهی می شوند که مولفه های منتجه بیشترین مقدار واریانس را در مجموعه داده ها داشته باشند.

در حقیقت تعداد مولفه ها ی استخراج شده در یک تحلیل مولفه ی اصلی برابر تعداد متغیرهای مشاهده شده ای است که تجزیه و تحلیل می شوند. اگرچه، در اکثر تحلیل ها، تنها چند مولفه ی اول برای ایجاد یک واریانس قابل قبول استفاده می شوند. در نتیجه تنها این چند مولفه ی اول نگه داشته می شوند، تفسیر می شوند، و در تحلیل های بعدی استفاده می شوند.

اولین مولفه ی اصلی استخراج شده از تحلیل مولفه های اصلی، برای تخمین حداکثر مقدار واریانس کلی در بین متغیرهای مشاهده شده می باشد. تحت شراطی معمولی، به این معناست که اولین مولفه ی اصلی حداقل با بعضی از متغیرهای مشاهده شده همبسته خواهد بود.

دومین مولفه ی اصلی استخراج شده، دارای دو مشخصه ی مهم است. اولا، این مولفه برای تخمین حداکثر مقدار واریانس در بین مجموعه داده هایی است که توسط مولفه ی اول به حساب نیامده اند. تحت شرایط معمولی، به این معناست که دومین مولفه ی اصلی با بعضی از متغیر های مشاهده شده ای همبسته است که همبستگی زیادی با اولین مولفه ی اصلی نداشته اند. ثانیا، دومین مولفه با اولین مولفه ناهمبسته خواهد بود. در حقیقت، اگر همبستگی بین مولفه ۱ و مولفه ۲ را محاسبه کنیم، برابر صفر خواهد بود.

سایر مولفه های اصلی که استخراج می شوند نیز همین دو مشخصه را دارند :

- هر مولفه ای برای تخمین حداکثر واریانس در متغیرهای مشاهده شده ای به کار می رود که توسط مولفه های قبلی به حساب نیامده اند. و
- هر مولفه ی اصلی با همه ی مولفه های قبلی اش ناهمبسته می باشد. [۲۸]

با ذکر این مقدمه درباره ی روش کاهش ویژگی PCA، تاثیر اعمال این تابع را بر ویژگی های بدست آمده در قسمت ۴-۴-۱ حالت (ب a) مورد بررسی قرار می دهیم. همان طور که گفته شده در این قسمت از ویژگی های MFCC استفاده کردیم.

- ماتریس ویژگی های مربوط به داده های آموزش و آزمایش را بدست می آوریم.

- ابعاد ماتریس داده های آموزش برابر است با تعداد داده های آموزش در ۱۳(تعداد ویژگی ها)

- ابعاد ماتریس داده های آزمایش برابر است با تعداد داده های آزمایش در ۱۳

حال می توانیم با اعمال تابع PCA به این ماتریس ها، تعداد ابعاد ویژگی ها را کاهش دهیم.(یک ، ۲ ،۳ و ۴ بعدی)

روش	Errorh	Errorg	Errorw	total error
pca 1	18.3673	73.8095	86.6667	40
pca 2	22.449	78.5714	80	43.23
pca 3	15.3846	59.5238	60	36.07
pca 4	13.8462	69.0476	93.3333	42.62

جدول ۴-۱۰ :نتایج حاصل از اعمال تابع PCA بر ماتریس ویژگی های بدست آمده در قسمت ۴-۴-۱ ، حالت (ب a)(درصد خطای ازمایش)

۴-۴-۶ اعمال روش LDA برای کاهش ابعاد ویژگی

Linear Discriminant Analysis (LDA) روش شناخته شده ای برای کاهش ابعاد و کلاسه بندی می باشد که داده های ابعاد-بالا را بر یک فضای ابعاد – پایین تصویر می کند که داده ها در آن به گونه ای هستند که بیشترین تفکیک بین کلاس ها ایجاد می شود. ویژگی های بدست آمده از LDA، ترکیب خطی ویژگی های اصلی هستند که ضرایب این ترکیب خطی از ماتریس تبدیل بدست می آید. تصویر و یا تبدیل بهینه در LDA کلاسیک، با حداقل کردن فاصله ی درون کلاسی و حداکثر کردن فاصله ی بین کلاس ها به طور همزمان، و در نتیجه حداکثر کردن تفکیک کلاس ها، بدست می آید. تبدیل بهینه با حل یک مسئله ی مقادیر ویژه ی کلی، محاسبه می شود. این روش در بسیاری از کاربردها به خوبی اعمال شده است؛ مانند شناسایی چهره .

در این قسمت هم مانند همان روش ۴-۴-۱ حالت (ب a) عمل می کنیم و بعد ازاستخراج ماتریس ویژگی های MFCC مربوط به داده های آموزش و آزمایش، تابع LDA را به منظور کاهش ابعاد ویژگی ها بر داده ها اعمال می کنیم.

LDA (1) : تصویر کردن داده ها به یک فضای یک بعدی از ویژگی ها

LDA (2) : تصویر کردن داده ها به یک فضای دو بعدی از ویژگی ها

روش	Errorh	Errorg	Errorw	total error
lda 1	20.4082	66.6667	26.6667	33.55
lda 2	27.6923	45.2381	53.3333	36.89

جدول ۴- ۱۱ : نتایج مربوط به اعمال تابع LDA بر روش ذکر شده در قسمت ۴-۴-۱ ،حالت (ب a)

۴-۴-۷ بررسی تاثیر تفکیک داده ها بر اساس جنسیت بر نتایج بدست آمده

می توانیم گوینده ها را براساس جنسیت آنها تفکیک کنیم و برای هر گروه(گروه مردان(۲۰۵ نفر) و گروه زنان(۹۹ نفر)) به طور جداگانه سیستم تخمین سن را به کار بریم.

اگر از همان روش بیان شده در قسمت ۴-۴- ۱ حالت (ب a) برای استخراج ویژگی در هر گروه جنسی، استفاده کنیم، نتایج زیر بدست می آید.

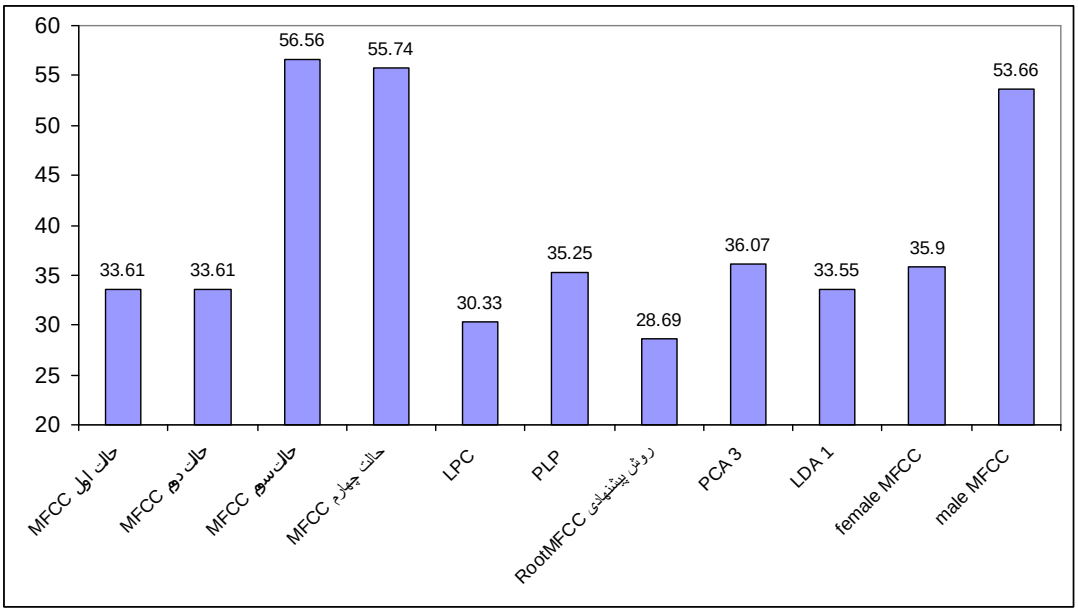
هم چنین از روش RootMFCC که قبلا توضیح داده شده نیز به ازای گاما برابر ۰.۰۰۶ برای هر دو گروه زنان و مردان ، استفاده می کنیم .

	Errorh	Errorg	Errorw	total error
Male(MFCC)	60	29.0323	100	53.66
Female(MFCC)	16	70	75	35.9
Male(RootMFCC)	57.5	35.4839	100	54.88
Female(RootMFCC)	28	70	75	43.59

جدول ۴-۱۲ : نتایج حاصل از تفکیک داده ها بر اساس تفکیک جنسیت

۴-۵ نتایج نهایی بدست آمده در پروژه

در نمودار ۴-۹ درصد خطای کل برای روش های مختلفی که بیان شد، نشان داده شده است. برای رسم نمودار و مقایسه ی نتایج روش های مختلف، از هر ۷ قسمت ذکر شده ،حالتی انتخاب شده است که در آن حالت روش مذکور کمترین خطا ی کل را داشته باشد.



شکل ۴- ۹ : نمودار مقایسه نتایج درصد خطای کل برای روش های مختلف

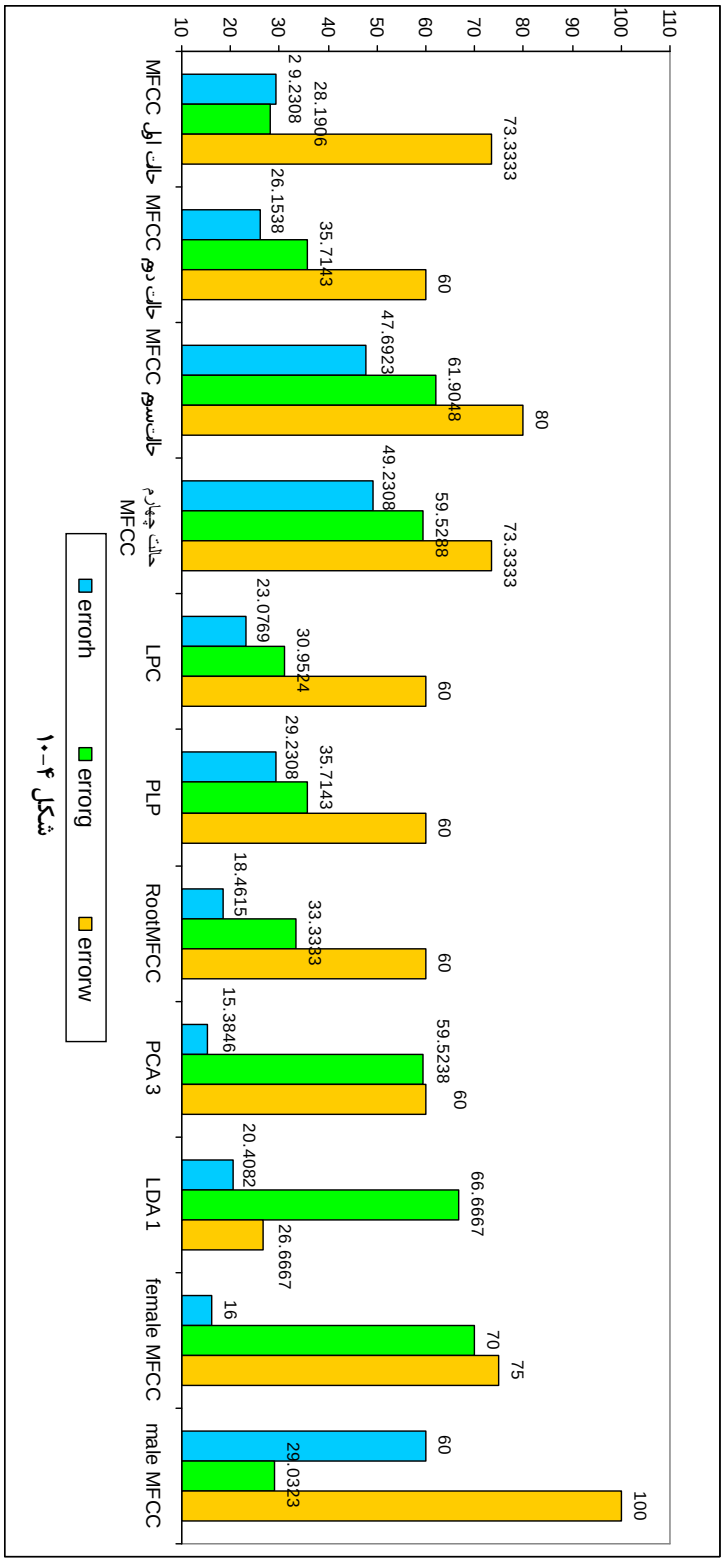
همان طور که از نمودار ۴-۹ آشکار است، درصد خطای کل مربوط به روش پیشنهادی، یعنی استفاده از روش استخراج ویژگی RootMFCC با گامای 0.006 نسبت به سایر روش های آزمایش شده، کمتر می باشد.

هم چنین تفکیک سیگنال گفتار به واج های تشکیل دهنده ی آن و سپس اعمال روش استخراج ویژگی مورد نظر بر سیگنال مربوط به واج ها، تاثیر خوبی بر کاهش درصد خطای کل داشته است. (حالت اول و دوم MFCC در مقایسه با حالت های سوم و چهارم MFCC) .

در نمودار شکل ۴- ۱۰ درصد خطا برای هر ۳گروه سنی مختلف (گروه hو گروه g و گروه w) برای همه ی روش های گفته شده در نمودار ۴-۹ ، نشان داده شده است .

گروه سنی h، که پایین ترین رنج سنی را دارا است، کمترین خطا را در اکثر روش ها دارد و گروه سنی w با بالاترین رنج سنی، بیشترین خطا را دارد.

درصدخطای کمتر در گروه سنی h ، به دلیل تعداد داده های بیشتر موجود در پایگاه داده در این رنج سنی نسبت به دو گروه دیگر می باشد.



شکل ۴-۱۰

۴-۶ تاثیر تعداد گروه های سنی برنتایج آزمایش:

- **الف**) اگر اشخاص را به ۲ گروه سنی تقسیم کنیم ،

گروه h : افراد با سن کمتر از ۳۵ سال ، ۱۸۳ نفر

گروه g : افراد با سن بالاتر از ۳۵ سال. ۱۲۱ نفر

- **ب**) اگر اشخاص را به ۵ گروه سنی تقسیم کنیم :

گروه h : افراد بین ۱۳-۲۵ سال ، ۱۱۰ نفر

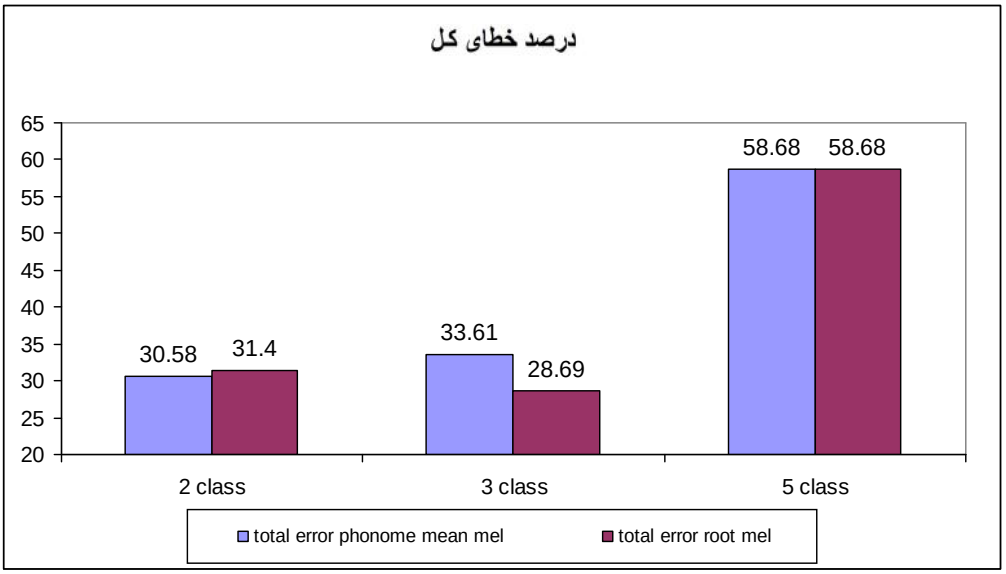
گروه g : افراد بین ۲۶ – ۳۷ سال ، ۸۳ نفر

گروه w : افراد بین ۳۸ – ۴۹ سال ، ۵۳ نفر

گروه k : افراد بین ۵۰ – ۶۱ سال ، ۴۲ نفر

گروه z : افراد بین ۶۲- ۷۳ سال ، ۱۶ نفر

نتایج بدست آمده با استفاده از روش استخراج ویژگی RootMFCC با گامای 0.006 و همچنین روش MFCC حالت (ب a) ، در مقایسه با تقسیم افراد به ۳ گروه سنی ، در نمودار زیر نشان داده شده است:

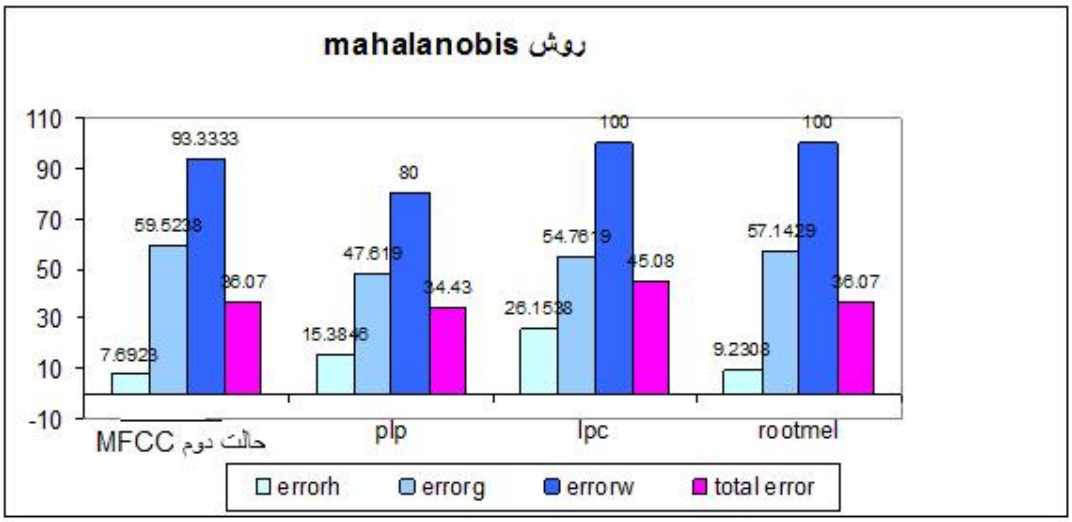


شکل ۴- ۱۱ : درصد خطای کل برای تقسیم داده ها به ۲ گروه سنی ، ۳ گروه سنی و ۵ گروه سنی

همان طور که در شکل ۴- ۱۱ مشخص است، در تقسیم داده ها به ۳ گروه سنی و استفاده از روش پیشنهادی RootMFCC ، با گامای ۰.۰۰۶ کمترین درصد خطا ی کل بدست می آید.

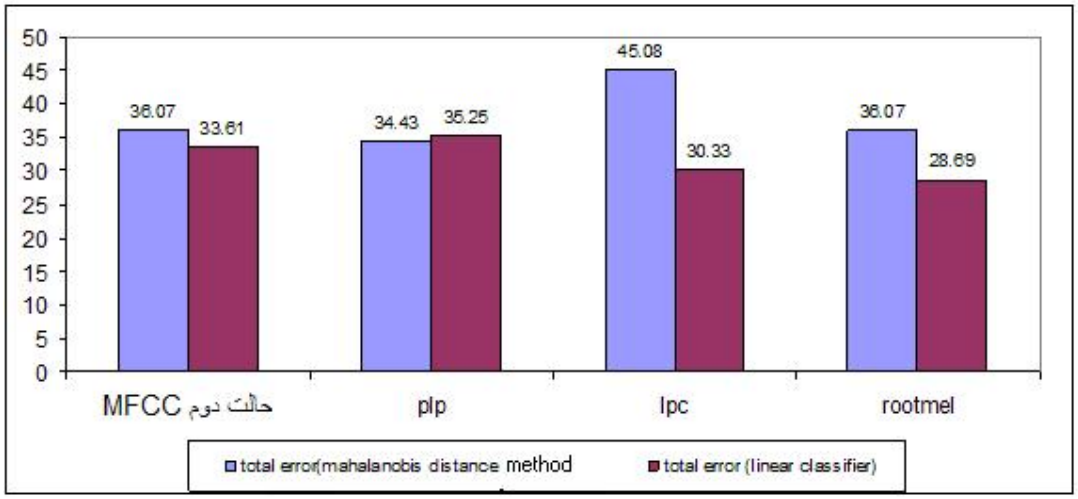
۴-۷- مقایسه رده بند به روش mahalanobis distance و linear

در نرم افزار Matlab دستوری به نام classify وجود دارد. با انتخاب نوع Linear ، برای هر گروه، با تخمین مشترک کوواریانس، توابع چگالی نرمال چند متغیره را که مناسب باشد تطبیق می دهد. فاصله ی اقلیدسی که در کلاسه بند linear از آن استفاده شده است، بر اساس همبستگی مجموعه ی داده ها (نمونه ها) می باشد. در روش Mahalanobis distance، از فاصله ی mahalanobis با تخمین های کوواریانس چند لایه ای استفاده می شود. فاصله ی Mahalanobis فاصله ای بر اساس همبستگی بین متغیر هایی (ویژگی هایی) است که به کمک آنها می توان الگوهای مختلف را تشخیص داد و تحلیل کرد. این فاصله میزان شباهت بین یک مجموعه نمونه های ناشناخته را با مجموعه ای معلوم، اندازه گیری می کند.



شکل ۴- ۱۲ : مقایسه درصد خطا در هر گروه سنی h,g,w ، برای ۴ روش نشان داده شده در شکل، با استفاده از روش mahalanobis distance

همان طور که مشاهده میشود با به کارگیری روش mahalanobis distance علی رغم اینکه کمترین خطای کل (۳۴.۴۳) مربوط به روش استخراج ویژگی PLP می باشد، ولی روش پیشنهادی نیز با این روش کلاسه بندی نیز دارای درصد خطای کل قابل مقایسه ای (۳۶.۰۷) می باشد.



شکل ۴- ۱۳: مقایسه درصد خطای کل بدست آمده برای رده بند Linear و روش Mahalanobis برای۴ روش استخراج ویژگی مذکور

در نمودار ۴- ۱۳درصد خطای کل برای ۴ روش استخراج ویژگی نشان داده شده، با استفاده از رده بند linear و روش mahalanobis برای کلاسه بندی، مقایسه شده است. همان طور که می بینیم رده بند linear دارای درصد خطای کل کمتری نسبت به روش mahalanobis میباشد. و روش پیشنهادی نیز کمترین درصد خطای کل را داراست.

۴-۸ ارزیابی نتایج حاصل از الگوریتم های مختلف با روش پیشنهادی

در پایان برای اینکه نتیجه بدست آمده از روش پیشنهادی را با نتایج سایر مطالعاتی که در زمینه ی تخمین سن افراد با استفاده از گفتار، صورت گرفته است را مقایسه کنیم، مختصرا این نتایج را بیان می کنیم. (جدول ۴-۱۳)

با توجه به این که در روش های مختلف از پایگاه داده های متفاوت و همچنین تعداد گروه های سنی مختلف با روش های استخراج ویژگی متفاوت و کلاسه بند های مختلف استفاده شده است، هدف از جمع آوری اطلاعات در یک جدول تنها آگاهی از نتایج بدست آمده در زمینه ی تخمین سن با استفاده از گفتار می باشد و اینکه نتیجه ی حاصل از روش پیشنهادی را می توان با انتخاب تعداد ویژگی های بیشتر و استفاده از کلاسه بند دقیق تری ارتقا داد.

۴-۹ نتیجه گیری کلی

در این تحقیق هدف اصلی تخمین سن افراد با استفاده از سیگنال صوتی گفتار آنها به روش RootMFCC می باشد. از چند روش پردازش سیگنال برای استخراج ویژگی (MFCC,PLP, LPC, RootMFCC) استفاده شده است.

برای استخراج ویژگی های گفتار مربوط به هر شخص، سیگنال صوت او را به سیگنال های واج های موجود در آن تجزیه می کنیم. زبان فارسی ۴۲ واج دارد. در نتیجه هر سیگنال گفتار تبدیل به ۴۲ بردار ویژگی می شود که هر بردار نماینده ی اطلاعات موجود در سیگنال های یک واج مشخص در سیگنال گفتار مورد نظر می باشد. برای هر واج مشخص یک رده بند مجزا به کار برده ایم. در نهایت، برای هر سیگنال گفتار ورودی برای هر گوینده در بخش نمونه های آزمایش، کلاس مربوط به همه ی واج های موجود در سیگنال را تخمین زده و سپس با بررسی کلاس های تخمین زده شده برای ۴۲ واج موجود در سیگنال گفتار فرد مورد نظر، کلاسی را که بیشترین تکرار را دارد به عنوان گروه سنی تخمین زده شده انتخاب می کنیم. در این تحقیق افراد را در سه کلاس جوان، میانسال و مسن کلاس بندی می کنیم.

تعداد ویژگی های RootMFCC ۱۳ می باشد. در روش استخراج ویژگی RootMFCC ،پارامتر ریشه یا همان گاما را، برای بدست آوردن کمترین نرخ خطا تغییر می دهیم. در نهایت به ازای $\gamma = 0.006$ کمترین نرخ خطا با به کارگیری کلاسه بند Linear، برابر ٪ ۲۸.۶۹ بدست می آید.

	روش استخراج ویژگی استفاده شده	تعداد گروه سنی	پایگاه داده	نوع رده بند	صحت رده بندی(٪)
مرجع [۹]	MFCC	۳ (بازه ی ۲۵ سال)	DARPA switch board	HMM	۶۸.۴
	MFCC + pitch				۷۰.۲
مرجع [۱۰]	Formant Pitch Energy spectral	۲ (سن آستانه ۴۵)	DES	PNN	۹۰.۴۳
				Bayes - SFFS	۸۳.۹۲
				SVM(guassian kernel)	۸۴.۹۰
				SVM(linear kernel)	۷۹.۸۷
			ELSDSR	PNN	۸۹.۱۷
				Bayes – SFFS	۷۱.۱۸
				SVM(guassian kernel)	۷۴.۶۸
				SVM(linear kernel)	۷۱.۴۶
مرجع [۱۱]	MFCC , MFCC 's first and second derivatives	۵ (کودکان زیر ۱۰ سال)	PLAKSS	SVM/SVR	۸۳
مرجع [۱۲]	MFCC , MFCC 's first and second derivatives	۳ (با بازه ی ۲۳ سال)	Czech Speech Dat	GMM	۶۰.۱۳
		۱۳ (با بازه ی ۵ سال)			۲۳.۸۵
مرجع [۱۳]	39 MFCC	۶ (با بازه ی ۱۰ سال)		SVM	۹۴.۱۱
	20 PLP				۹۱.۱۶

مرجع[۱۴]	Jitter , shimmer	۲ (با سن آستانه ۶۰)		ANN	۹۶.۵۷
				KNN	۹۵.۷۱
				NB	۹۱.۱۵
				SVM	۹۶.۵۲
				BL	۸۸.۳۰
مرجع [۸]	MFCC , pitch , formants , prosodic fitures	۳	UF-VAD	GMM	۹۳.۱
مرجع [۷]	12 MFCC	کمتر از ۳۰ سال		PNN	۷۶
				GMM	۹۶.۵
		بیشتر از ۵۰ سال		PNN	۹۵
				GMM	۶۸
مرجع [۱۵]	MFCC, prosodic features	۴	Interspeech paralinguistic challenge	سیستم ترکیبی SVM , GMM	۵۲.۷
روش پیشنهادی	13 ROOtMFCC	۳ گروه سنی (بازه ی ۲۰ سال)	FARSDAT	Linear discriminant	۷۱.۳۱

جدول ۴-۱۳

فهرست مراجع :

[1] Susanne Schötz , "Perception, Analysis and Synthesis of Speaker Age" , Department of Linguistics and Phonetics Centre for Languages and Literature ,Lund University, 2006

[2] Susan Hughes, Bradley C. Rhodes, "Making Age Assessments Based On Voice: The Imoact Of The Reproductive Viability Of The Speaker", Department of Psychology, Albright College and Department of Biology, Bucknell University

[3] Markus Iseli, "Dependencies of Voice Source Measures on Age, Sex, Vowel Context, and Prosodic Features", University of California, Los Angeles , 2007

[4] Ulrich Reubold, Jonathan Harrington, Felicitas Kleber, "Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers", Institute of Phonetics and Speech Processing (IPS), University of Munich, 4 December 2009.

[5] Markus Iseli, Yen-Liang Shue, and Abeer Alwan, "Age- And Gender-Dependent Analysis Of Voice Source Charactiristics", University of California Los Angeles Dept. of Electrical Engineering.

[6] Gurpinder Kaur, Arjan Singh, "Rule Based Age Detection System", Department of Computer Science & Engg., SUSCET, Tangori,Department of Computer Science & Engg., BBSBEC.

[7] Devendra Kumar Yadav, Kamini Malhotra , Anu Khosla, " Age Group Estimation based on Acoustic Analysis of Speech ", Scientific Analysis Group,DRDO, Metcalfe House.

[8] Werner Spiegl, Elmar N¨ oth, Chair of Pattern Recognition (LME), University Erlangen-Nuremberg, Germany

, Georg Stemmer, SVOX Deutschland GmbH, Munich, Germany

, Eva Lasarcyk, Dep. of Computational Linguistics and Phonetics, Saarland University, Germany

, Varada Kholhatkar4, Dep. of Computer Science, University of Minnesota Duluth, USA

,Andrew Cassidy, Puyang Xu , The Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA

, Blaise Potard, CRIN, Nancy, France

, Stephen Shum, International Computer Science Institute, University of California at Berkeley, USA

, Young Chol Song , Dep. of Computer Science, Stony Brook University, USA

,Peter Beyerlein , Dep. Bioinformatics, University of Applied Sciences Wildau, Berlin, Germany
, James Harnsberger , Speech Perception Laboratory, University of Florida, USA.
"Analyzing Features for Automatic Age Estimation on Cross-Sectional Data".

[9] Izhak Shafran, Michael Riley, and Mehryar Mohri, "Voice Signatures"
, AT&T Labs – Research , Florham Park, USA 07932

[10] M. H. Sedaaghi , "A Comparative Study of Gender and Age Classification in Speech Signals" , Iranian Journal of Electrical & Electronic Engineering, Vol. 5, No. 1, March 2009 .

[11] Tobias Bocklet, Andreas Maier, Elmar Noth , "Age Determination of Children in Preschool and Primary School Age with GMM-based Supervectors and Support Vector Machines/Regression" , Institute of Pattern Recognition, University of Erlangen-Nuremberg, Germany

[12] Valiantsina Hubeika ," Estimation of Gender and Age from Recorded Speech ", Faculty of Infromation Technology, Brno University of Technology

[13] Davood Mahmoodi ," Age Estimation Based on Speech Features and Support Vector Machine",Department of Electrical Engineering and Robotics ,Shahrood University of Technology ,Shahrood, Iran, 2011

[14]Christian M¨ uller, Frank Wittig, J¨ org Baus , "Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs" ,Department of Computer Science , Saarland University, Germany

[15] Mohamad Hasan Bahari, Hugo Van hamme , "Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization", Centre for Processing Speech and Images ,Katholieke Universiteit Leuven ,Leuven, Belgium.

[16] Ming Li , Kyu J. Han, Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,University of Southern California, Los Angeles, USA
, Chi-Sang Jung , School of Electrical and Electronic Engineering, Yonsei University, Korea , "Combining Five Acoustic Level Modeling Methods for Automatic Speaker Age
and Gender Recognition" .

[17] Rajesh Mahanand Hegde,"Fourier Transform phase-based features for speech recognition", Doctor of Philosophy,Department of computer science and engineering Indian institute of thechnoligy Madras,july 2005.

[18] Mark Hasegawa-Johnson, "Landmark-Based Speech Recognition:Spectrogram Reading, Support Vector Machines,Dynamic Bayesian Networks, and Phonology", University of Illinois at Urbana-Champaign, USA

[19] Tan Lee, "Feature Extraction", Page 1 of 11, February 28, 2005

[20] Selim Aksoy, "Feature Reduction and Selection" ,Department of Computer Engineering ,Bilkent University,CS 551, Spring 2010.

[21] Stefan Geirhofer, "Feature Reduction with Linear Discriminant Analysis and its Performance on Phoneme Recognition", University of Illinois at Urbana-Champaign , Department of Electrical and Computer Engineering ,May 8, 2004.

[22] http://support.sas.com/publishing/pubcat/chaps/55129.pdf ,"Principal Component Analysis", chapter 1

[23] Benjamin J.Shannon, Kuldip K.Paliwal,"Feature extraction from higher-lag autocorrelation coe□cients for robust speech recognition", School of Microelectronic Engineering, Gri□th University, Nathan Campus, Brisbane, QLD 4111, Australia, July 2005

[24] Ufuk Ülüğ , Tolga Esat Özkurt , Tayfun Akgül, "Bispectrum Mel-frequency Cepstrum Coefficients for Robust Speaker Identification", Department of Electronics and Communications Engineering, Istanbul Technical University, Istanbul, Turkey, Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA May 22-25, 2007

[25] Vivek Tyagi, Christian Wellekens ,"On Desensitizing the Mel-Cepstrum to Spurious Spectral Components for Robust Speech Recognition" ,20th September 2004

[26] Amita Dev , "A Novel Feature Extraction Technique for Speaker Identification" , Bhai Parmanand Institute of Business Studies, Department of Training & Technical Education ,India - New Delhi-110092 ,International Journal of Computer Applications (0975 – 8887) ,Volume 16– No.6, February 2011

[27] Jiqing Han, Wen Gao, "Robust telephone speech recognition based on channel Compensation", Department of Computer Science and Engineering, Harbin Institute of Technology, 16 October 1997

[28] Lindsay I Smith , "A tutorial on Principal Components Analysis", February 26, 2002

Department of Electricity and Robatic

Thesis submitted
for the degree of M.Sc.

Title:

Age estimation from recorded speech by using RootMFCC feature extraction method

Supervisor:Dr. Hossein Marvi

Advisor:Dr. Ali Soleimani

By: Atefeh Dehghanian

Sep 2012

Abstract

Age estimation based on human's speech features, is a considerable subject in automatic speech recognition (ASR) systems. Although some researches have been done in speaker age estimation, but more studies, especially in Farsi Language speech is required. Like other speech processing systems, we are face with two main challenges in age estimation: finding an appropriate method for feature extraction, and choosing a reliable classification method.

The main goal of this research is using Root Mel Ferequency Cepstral Coefficients features in an age estimation system and finding an optimum root (gama)to have a less error; and also comparing the performance of an age estimation system when using this feature, with other common features like MFCC, PLP and LPC.

For feature extraction, the whole of speech signal is separated to phonemes (smallest part of speech). these phoneme signals is used for feature extraction and classification.

Linear classification and Mahalanobis distance method are used for classifying each phoneme signal in one of the age groups.

Expriments which has been done on FARSDAT database, shows that at gama=0.006 (the root in RootMFCC method) there is the least error rate (28.69 %). besides, separating speech signal to its phonemes and extracting features from each phoneme signal, improves the results and decreases the error rate.

Keywords: age estimation, speech, phoneme, RootMFCC feature extraction, FARSDAT