دانشکده مهندسی برق و رباتیک

رشته مهندسی برق گرایش مخابرات

پایان‌نامه کارشناسی ارشد

جداسازی گفتار دو گوینده همزمان مبتنی بر ویژگی‌های مناسب زمان- فرکانس

نگارنده :

رعنا دهقانی

استاد راهنما :

دکتر حسین مروی

بهمن ۱۳۹۵

تقدیم به

پدر و روح مادر همیشه خوبم

آنان که وجودم برایشان همه رنج بود و وجودشان برایم همه مهر. مویشان سپیدی گرفت تا رویم سپید بماند. آنان که فروغ نگاهشان، گرمی کلامشان و روشنی رویشان سرمایه جاودانی زندگیم است. آنان که راستی قامتم در شکستگی قامتشان تجلی یافت. در برابر وجود گرمشان زانوی ادب بر زمین می نهم و با دلی مملوء از عشق و محبت بر دستانشان بوسه می زنم.

تقدیر و تشکر

پیش از هر چیز سپاس گزار خداوند منانی هستم که هر چه داشته و دارم از اوست، بر آستانه بیکرانش سر تعظیم فرود آورده و به پاس هر آن چه به من بخشیده است، سجده شکر به جای می آورم.

سپاس فراوان نثار استادان بزرگواری که صادقانه و صمیمانه در راه پرورش فرزندان این آب و خاک قدم برمی دارند و در این راه از هیچ کوششی دریغ ندارند. سر تعظیم فرود می آورم بر این همه تواضع و بزرگ منشی و به این همه مناعت طبعی که دارند و عاشقانه در راه اعتلای فرهنگ این مرز و بوم قدم برمی دارند.

بانهایت تواضع و احترام از استاد راهنمای ارجمندم، **جناب آقای دکتر حسین مروی** تشکر و قدردانی می نمایم. راهنمایی های ایشان کمک بزرگی در انجام این پایان نامه برای اینجانب بود. مهم تر از آن نصیحت های ارزشمندی است که همواره آویزه گوش و چراغ راه آینده من است. تشکر صمیمانه خود را به محضر ایشان تقدیم می دارم و می دانم که بی شک انجام این پایان نامه بدون راهنمایی ها و تشویق های ایشان میسر نبود.

همچنین از اعضای محترم کمیته ارزشیابی، **جناب آقای دکتر امیدرضا معروضی** و **جناب آقای دکتر سید مسعود میررضایی** که بانهایت صبر و تأمل این پایان نامه را مورد بررسی قرار داده و نکات ارزشمندی را گوشزد نموده اند، قدردانی می نمایم.

آخر از همه اما نه کمتر از همه! بهترین سپاس و درود خود را تقدیم به پدر و مادر مهربانم، خواهر همواره همراهم و دو برادر خوبم می کنم که در هیچ یک از لحظات زندگی مرا تنها نگذاشتند و همواره دلم به بودنشان، کمک شان، هدایت شان و محبت شان گرم بود.

تعهد نامه

اینجانب **رعنا دهقانی** دانشجوی دوره کارشناسی ارشد رشته مهندسی برق/ مخابرات دانشکده مهندسی برق و رباتیک دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **جداسازی گفتار دو گوینده همزمان مبتنی بر ویژگی‌های مناسب زمان-فرکانس** تحت راهنمایی **دکتر حسین مروی** متعهد می‌شوم:

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

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

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

چکیده فارسی:

یکی از مباحث مهم در زمینه ارتباطات و مخابرات، جداسازی گفتار توسط ماشین می‌باشد که به‌دلیل مشکلات و ضعف‌های موجود در این سیستم، کماکان به‌عنوان یک چالش بزرگ باقی مانده است. این در حالی است که سیستم شنیداری انسان در مقایسه با سیستم جداسازی گفتار ماشین توانایی‌های قابل توجهی دارد. لذا با توجه به اختلاف در عملکرد سیستم جداسازی گفتار توسط ماشین و سیستم شنیداری انسان، نیاز به یک سیستم کارآمد برای جداسازی گفتار مورد نظر توسط ماشین در بسیاری از کاربردها محسوس می‌باشد. به‌عنوان‌مثال یکی از معایب سیستم تشخیص خودکار گفتار، کاهش عملکرد آن در صورت وجود صداهای مزاحم است. بنابراین چنین سیستمی می‌بایست به یک جداکننده گفتار خوب مجهز شود تا ضمن بهبود عملکرد آن در شرایط مختلف، در بهبود کیفیت گفتار و کاهش هزینه انتقال سیگنال غیرگفتار نیز نقش موثری داشته باشد. از آن‌جا که انتخاب دقیق واحدهای زمان-فرکانس تاثیر به‌سزایی در نتایج جداسازی دارد، در این پایان‌نامه یک روش با نظارت به‌منظور انتخاب واحدهای زمان-فرکانس مناسب بر پایه ویژگی‌های ضرایب تُنّک ارائه می‌شود. در روش پیشنهادی ابتدا نمونه‌ی صدای گوینده‌ها مدل‌سازی شده که این مدل‌سازی از طریق آموزش یک دیکشنری و به‌دست آوردن بردارهای پایه مربوط به هر کدام از گوینده‌ها انجام می‌گیرد. بعد از مدل‌سازی صدای هر یک از دو گوینده‌ی همزمان، برای هر صدا مطابق با مولفه‌های فرکانسی مربوط به آن عمل جداسازی انجام شده و به‌منظور کاهش نویز صدای خروجی و حذف مولفه‌های فرکانسی نامرتبط، یک مرحله‌ی پس‌پردازش نیز روی صدای خروجی انجام می‌شود. نتایج آزمایش‌های انجام شده با هدف جداسازی دو گوینده همزمان، نشان‌دهنده عملکرد قابل توجه روش پیشنهادی در مقایسه با روش‌های مبتنی بر ویژگی و مدل می‌باشد.

کلیدواژه: آنالیز ترکیب شنیداری محاسباتی، جداسازی گفتار دو گوینده همزمان، جداسازی با نظارت.

فهرست مطالب

عنوان **شماره صفحه**

فهرست اشکال

عنوان **شماره صفحه**

فهرست جداول

عنوان **شماره صفحه**

علائم و اختصارات

ASA……..Auditory Scene Analysis

ASR……..…….....Automatic Speech Recognition

BSS……...……Blind Source Separation

CASA……...……..Computational Auditory Scene Analysis

HMM……...……...Hidden Markov Model

HMS……..……...Harmonic Magnitude Suppression

ICA……..…….Independent Component Analysis

ISA……..…….Independent Subspace Analysis

LPC……...……...….Linear Predictive Analysis

MFCC……..…..…Mel Frequency Cepstral Coafficient

NMF……...……..……Non-negative Matrix Factorization

فصل اول

پیشگفتار

۱-۱- مقدمه

در زندگی روزمره زمان‌های زیادی پیش می‌آید که شما به‌همراه دوستتان در شهر قدم می‌زنید و با یکدیگر صحبت می‌کنید. در این حین شما صداهای دیگری را نیز از جمله همهمه، فریاد زدن و خندیدن را می‌شنوید. هنگامی‌که دوست شما صحبت می‌کند، گوش شما مجموعی از صدای دوستتان و صداهای دیگران را همزمان می‌شنود. اگر چه در اغلب موارد صدای بلندی از دیگران به گوش می‌رسد، اما باز هم شما به‌خوبی می‌توانید صدای دوستتان را بدون وقفه بشنوید. این‌طور که به‌نظر می‌رسد، سیستم شنوایی شما با کمی زحمت قادر به جداسازی گفتار دوستتان از صداهای دیگران است.

شرایط نوعی توضیح داده شده چیزی است که ما به‌طور دائم با آن برخورد می‌کنیم، هنگامی‌که یک نفر با ما صحبت می‌کند، چیزی که ما می‌شنویم تنها صدای آن شخص نیست، بلکه مخلوط (ترکیبی) از صدای او و صداهای تداخلی دیگر است. تداخل، هر نوع صدایی مانند صدای باد، موزیک و یا گوینده دیگری می‌تواند باشد. در این شرایط، ما نیاز داریم تا صدای هدف و اطلاعات مربوط به آن را از صدای مخلوط شده جدا نماییم. اشخاصی که سیستم شنوایی آن‌ها سالم است می‌توانند به‌راحتی عملیات جداسازی گفتار هدف را از انواع مختلف تداخل انجام دهند؛ البته باید یادآوری شود که در اکثر مواقع صداهای تداخلی آزاردهنده نمی‌باشند. اما در ارتباط‌های گفتاری با ماشین‌ها و رایانه‌ها، تداخل یک مشکل جدی است و در بسیاری از کاربردها سیستم جداسازی گفتار موثر، مورد نیاز می‌باشد. برای مثال، عملکرد سیستم بازشناسی خودکار گفتار[1] (ASR) به‌وسیله صداهای تداخلی شدیداً تخریب می‌شود و با به‌کارگیری یک سیستم جداسازی گفتار می‌توان عملکرد آن را بهبود بخشید. همچنین به‌کارگیری سیستم جداسازی گفتار در سیستم های مخابراتی باعث بهبود کیفیت گفتار و کاهش هزینه‌ها در ارسال سیگنال‌های غیرگفتاری می‌گردد. علاوه‌براین، صدای تداخلی یک مشکل جدی برای افرادی است که مشکل شنوایی دارند و برای دریافت گفتار هدف مجبورند از

[1] Automatic Speech Recognition

سمعک استفاده کنند. برای کمک به این افراد باید سمعک‌ها به‌نحوی طراحی شوند که بتوانـد گفتـار هدف را از مخلوط‌های صوتی جدا نماید.

پژوهش‌های زیادی برای گسترش سیستم‌های محاسباتی که به‌طور خودکار توانایی جداسازی و یا تضعیف تداخل از گفتار هدف را دارا می‌باشند، انجام گرفته است. بسیاری از این پژوهش‌ها بر روی موقعیت‌هایی متمرکز شده‌اند که منابع هدف و تداخل، در دو مکان متفاوت قرار گرفته‌اند و چندین میکروفون در دسترس می‌باشند. یکی از راه‌حل‌های ارائه شده در برخورد با این موضوع، تضعیف سیگنال تداخلی با استفاده از فیلترهای فضایی[1] است که در آن سیگنال‌های رسیده از جهت منبع هدف، استخراج و سیگنال‌های رسیده از جهت منبع تداخل، حذف می‌شود [۱]؛ اما در شرایطی‌که جهت منابع هدف و تداخل یکسان بوده و یا تنها یک میکروفون و یک صدای ضبط‌شده در دسترس باشد، نمی‌توان از این روش‌ها استفاده کرد. روش جداسازی کور منابع[2] (BSS) با استفاده از تکنیک آنالیز مؤلفه‌های مستقل[3] (ICA)، سیگنال‌های مخلوط را به اجزائی تجزیه می‌کند که از لحاظ آماری از یکدیگر مستقل می‌باشند. در شرایطی‌که جهت منابع هدف و تداخل متفاوت باشند و تعداد میکروفون‌ها از تعداد منابع بزرگ‌تر و یا مساوی آن‌ها باشد، روش ICA به‌خوبی کار می‌کند؛ اما در شرایطی‌که تنها یک میکروفون در دسترس باشد، نمی‌توان از این روش استفاده نمود.

در کاربردهایی مانند ارتباطات مخابراتی و بازیابی اطلاعات صوتی، نیاز به راه‌حـــل‌هـــایی بـــرای جداسازی گفتار در شرایط تک‌میکروفونه می‌باشد. راه‌حل‌های ارائه شده از خواص ذاتی سیگنال هدف و تداخل برای تشخیص و جداسازی آن‌هـا از یکدیگر استفاده می‌کننـد. الگـوریتم‌هـای بسیاری برای بهسازی گفتار تک‌میکروفونه ارائه گردیده اســت کـــه این الگوریتم‌ها به‌طور کلی بر پایه آنالیزگفتار و تداخل و در نتیجه تقویت گفتار و کاهش تداخل می‌باشند. به‌عنـوان‌مثـال در [۲]، الگوریتم‌هایی ارائه

[1] Spatial Filters
[2] Blind Source Separation
[3] Independent Component Analysis

شده است کــه در آن طیــف زمــان کوتــاه تداخل تخمین زده می‌شود و بر اساس آن تداخل تضعیف می‌یابد و یا در [۳] بر اساس مدل‌سازی گفتار، عمل‌جداسازی انجام می‌گیرد.

در راه‌های دیگر مخلوط آکوستیکی براساس تجزیه ویژه نمایش داده شده و سپس از آنالیز زیرفضا، برای از بین بردن تداخل استفاده می‌شود. مدل‌های مارکف مخفی روش دیگری است که بـــــــرای مدل‌سازی گفتار و تداخل سپس جداسازی آن‌ها به‌کار برده می‌شود [۴].

این مدل‌سازی بر اساس خواص معینی از تداخل انجام می‌گیرد و به‌همین منظور قابلیت برخورد با هر نوع تداخلی را ندارد، چرا که تنوع زیاد تداخل‌ها عمل مدل‌سازی و پیش‌بینی را مشکل می‌کند.

۱-۲- سیستم‌های‌جداسازی گفتار تک‌میکروفونه

در حالی‌که سیستم‌های جداسازی گفتار تک‌میکروفونه هنوز با مسائل حل نشـده‌ی بسیاری مواجه هستند، سیستم شنوایی انسان با توانـایی قابل توجهی، قادر به جداسازی گفتار می‌باشد. همین مسأله الهام‌بخش، باعث ایجاد یک راه‌حـل جدیـد و متفـاوت از دیگر روش‌های موجود برای جداسازی گفتار تک‌میکروفونه بر اساس فرآیند جداسازی سیستم شنوایی شد.

با توجه به فرآیند جداسازی در سیستم شنوایی انسان، در ســال Bregman۱۹۹۴ سیســتمی بــا موضــوع تحلیل صحنه شنیداری[۱](ASA) برای جداسازی تک‌میکروفونه مطرح کــرد [۵]. این سیستم جداسازی شامل دو مرحله می‌باشد: مرحله اول، بخش‌بندی[۲] نامیده می‌شود، واحدهای زمان-فرکانس[۳] (T-F) سیگنال مخلوط به بخش‌هایی تجزیه می‌شود که هر کدام از این بخش‌ها از یک منبع به‌وجـود آمـده‌انـد. مرحله دوم گروه‌بندی[۴] نامیده می‌شود که در آن بخـش‌های مربوط به یک منبع، با یکدیگر تشکیل یک گروه می‌دهند. بخش‌بندی و گروه‌بندی با توجه به اصول ادراکی (ویژگی‌های ASA) انجام می‌گیرد و بر اساس این ویژگی‌ها می‌توان فهمید که سیستم شنوایی چگونه سازماندهی شـــده است.

[1] Auditory Scene Analysis
[2] Segmentation
[3] Time-Frequency
[4] Grouping

این ویژگی‌ها خواص ذاتی گفتار را مشخص‌می‌کنند و عبارتند از: harmonicity، فراز[1] و فرود[2]، مکان و اطلاعات قبلی از صداهای خاص.

پـــژوهش‌هـــای انجـــام شـــده در زمینه ASA، [۶] الهام‌بخش کارهای قابل توجهی برای ساخت سیستم‌های آنالیز محاسـباتی نمـایش شـنوایی[3](CASA) برای جداسازی گفتار شده است. بسیاری از سیســـتم‌هـــای CASA برای شرایط دومیکروفونه پیشنهاد شده‌انـــد و مبنای آن‌هـــا بر این است که سیستم‌شنوایی انسان با توجه به متفاوت بودن مکان منـابع صـدا، قادر به مکان‌یابی منابع و در نتیجه جداسازی اصوات از جهت‌های گوناگون می‌باشد [۵،۷،۸]. سیستم شنوایی با مقایســـه ســـیگنال‌های رســـیده بـــه دو گـــوش (در نقش دو میکروفون) مکان منابع را پیدا کرده و با استفاده از جهت منابع، صدای هدف را از صدای تداخلی جدا می‌کند.

در واقع سیستم‌های دومیکروفونه در شرایطی‌که جهت منابع هدف و تداخل قابـل تفکیـک باشـد، دارای عملکرد قابل توجهی می‌باشند. توجه کنید در شرایطی‌که جهت منابع یکسـان باشـد و یـا یـک میکروفون در دسترس باشد، روش های دومیکروفونه قادر به جداسازی منابع نمی باشند.

در چندین سال اخیر تلاش‌های زیادی برای گسترش سیستم‌های CASA درشرایط تک‌میکروفونه انجـام گرفتـه اسـت. هدف این‌سیستم‌ها، جداسازی صدای هدف بدون در نظر گرفتن فرضیاتی درباره تداخل و افزایش‌کاربردهای آن نسبت به روش‌های بهسازی گفتار بوده است [۹،۱۰].

۱-۳- اهمیت سیستم جداسازی گفتار تک‌میکروفونه

در کاربردهای واقعی، گفتار توسط یک یا چند میکروفون ضبط گردیده و برای پردازش‌های چند بعدی به رایانه‌ها فرستاده می‌شود. البته اگر این امکان وجود داشته باشد، ضبط گفتار با چند میکروفون در اولویت است. در این حالت اطلاعات مکانی نیز می‌تواند ذخیره شود و به‌عنوان ویژگی‌های اضافی برای جداسازی گفتار به‌کار برده می‌شود. اگرچه در محیط‌هایی با چندین منبع

[1] Onset
[2] Offset
[3] Computational Auditory Scene Analysis

صدا، چنانچه گوینده هدف از قبل تعیین نشده باشد، آرایه‌های میکروفونی ممکن است مفید واقع نشوند و حتی در مواردی ممکن است شرایط محیطی، قابلیت استفاده از چندین میکروفون را فراهم نسازد. در بسیاری از این حالت‌ها استفاده از یک میکروفون، تنها انتخاب ممکن است.

یکی از کاربردهای جداسازی گفتار تک‌میکروفونه، در بازشناسی خودکار گفتار، پخش رادیویی است. در این موارد، سیگنال گفتار ارسال شده توسط کانال‌های رادیویی، جمع‌آوری می‌شود و در نتیجه هیچ‌گونه اطلاعات مکانی در دسترس نمی‌باشد. در طول ارسال گفتار هدف، برخی از قسمت-های آن به‌وسیله گفتار گوینده‌های تداخلی تخریب شده و این مسأله باعث پایین آمدن دقت سیستم بازشناسی گفتار می‌شود. یکی دیگر از کاربردهای سیستم جداسازی در سیستم بازشناسی گفتار، مکالمه‌های کنفرانسی از راه دور می‌باشد. در واقع وجود چندین گوینده تداخلی، باعث تخریب عملکرد سیستم بازشناسی گفتار می‌شود.

خروجی سیستم بازشناسی گفتار به‌عنوان ورودی برای سیستم‌هایی مانند جمع‌آوری اخبار، سیستم‌های مکالمه و سیستم‌های تبدیل متن به گفتار مورد استفاده قرار می‌گیرد. تمامی این سیستم‌ها برای بالا بردن دقت بازشناسی گفتار، نیاز به یک سیستم جداسازی گفتار تک‌میکروفونه با عملکرد خوب دارند؛ زیرا دقت پایین سیستم بازشناسی گفتار باعث افزایش جدی خطاها می‌شود. به‌همین دلیل وجود یک سیستم جداسازی گفتار با دقت بالا بسیار حائز اهمیت است.

همانطور که بیان شد، برخلاف سیستم چندمیکروفونه، سیستم تک‌میکروفونه به اطلاعات مکانی دسترسی ندارد و به‌همین دلیل تنها از ویژگی‌های آکوستیکی ذاتی مانند فرکانس گام، ساختار سیستم هارمونیکی و زمان محلی[1] یا مجاورت فرکانسی[2] برای جداسازی گفتار استفاده می‌کند. اطلاعات فرکانس گام به‌عنوان یک ویژگی خوب برای استخراج ساختار هارمونیکی مورد توجه قرار گرفته است. اما تخمین دقیق نمودارهای فرکانس گام گفتار هدف در حضور گفتار تداخلی دیگر، کاری

[1]Local Time
[2]Frequency Proximity

بسیار سخت است؛ زیرا رفتار گفتار تداخلی برخلاف نویزهای دیگر مانند نویز سفید/ رنگی و یا نویز ماشین‌ها بسیار شبیه گفتار هدف است.

۱-۴- ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل است که محتوای کلی آن به شرح زیر می‌باشد:

در فصل دوم، ابتدا روش‌های موجود برای جداسازی منابع گفتاری را دسته‌بندی کرده و بر روی روش‌های موجود در هر دسته، مروری مختصر داریم. سپس به بررسی روش‌های تک‌میکروفونه بر پایه ویژگی برای جداسازی منابع می‌پردازیم. در یک دسته‌بندی کلی می‌توان روش‌های جداسازی بر پایه ویژگی را به دو دسته، روش‌های CASA و روش‌های بر پایه آنالیز مدولاسیون و فیلتر نمودن طبقه-بندی نمود. ابتدا مروری بر روی سیستم‌ها و روش‌های ارائه شده در هر کدام از این دسته‌ها داشته و سپس مزایا و معایب هر کدام از آن‌ها را مورد بررسی قرار می‌دهیم.

در فصل سوم، تئوری الگوریتم‌های مورد استفاده برای استخراج ویژگی را توضیح می‌دهیم. در واقع روش‌های استخراج ویژگی را به دو دسته زمان و فرکانس تقسیم کرده و هر کدام را جداگانه بیان می-کنیم. در ادامه فصل به روش‌هایی می‌پردازیم که اطلاعات دو حوزه را با هم در بر می‌گیرد. به‌عنوان-مثال، روش‌های آنالیز پیش‌گویی خطی و آنالیز کپسترال را می‌توان نام برد. بعد از معرفی روش‌های استخراج ویژگی در مورد روش آموزش دیکشنری مورد استفاده در روش پیشنهادی توضیح می‌دهیم. این بخش شامل زیرمجموعه کدگذاری تُنک یک سیگنال صوتی و الگوریتم OMP می‌باشد که به شرح آن می‌پردازیم.

در فصل چهارم، روش پیشنهادی و پیاده‌سازی الگوریتم مورد استفاده در روش پیشنهادی خود را بیان می‌کنیم. ساختار این فصل شامل سه فاز می‌باشد. فاز آموزش، فاز تست و فاز پس‌پردازش. در هر بخش فاز مربوطه را کامل شرح می‌دهیم.

در فصل پنجم، الگوریتم‌های مطرح شده، در محیط نرم‌افزار MATLAB شبیه‌سازی شده و طی آن، روش پیشنهادی در پایان‌نامه را مورد ارزیابی قرار می‌دهیم. نتایج شبیه‌سازی نشان می‌دهد که

روش پیشنهادی توانسته است نسبت به سـایر روش‌هـا از SNR بـالاتر و بهتـری برخـوردار باشـد. در نهایت، نتایج کلی به‌دست آمده در این پایان‌نامه وپیشنهادهایی برای ادامه کار بیان شده است.

فصل دوم

مروری بر کارهای انجام شده و مفاهیم پایه

۲-۱- مروری بر روش های جداسازی منابع صوتی

در این فصل ابتدا روش‌های موجود برای جداسازی منابع را دسته‌بندی کرده و سپس بر روی روش‌های موجود در هر دسته، مروری مختصر داریم. هدف از انجام این پایان‌نامه، ارائه سیستمی با کارایی بالا برای جداسازی گفتار در شرایطی است که تنها یک میکروفون در دسترس باشد. به‌همین دلیل در ادامه این فصل به بررسی دقیق‌تر روش‌های جداسازی گفتار همزمان می‌پردازیم.

به‌طور کلی روش‌های موجود برای جداسازی گفتار همزمـــــان را می‌توان به‌صورت زیر دسته‌بندی کرد:

۱. آنالیز مؤلفه‌های مستقل (ICA)
۲. روش‌های مبتنی بر یادگیری آماری[1]
۳. روش‌های بر پایه مدل‌سازی
۴. روش‌های بر پایه ویژگی

۲-۲- آنالیز مؤلفه‌های مستقل

تصور کنید در یک اتاق که در آن دو نفر به‌طور همزمان در حال صحبت هستند، قرار دارید. دو میکروفون، که در مکان‌های مختلف گذاشته شده‌اند، موجود می‌باشد. میکروفون‌ها دو سیگنال ثبت شده زمانی را که با $x_1(t)$ و $x_2(t)$ نشان داده شده، در اختیار قرار می‌دهند. توجه داشته باشید xها دامنه و tها شاخص زمان هستند. هر یک از این سیگنال‌های ثبت شده، مجموع وزن سیگنال‌های گفتار ادا شده توسط دو گوینده هستند که با نماد s(t) نمایش داده می‌شوند. معادله خطی عبارت است از:

$$x_1(t) = a_{11}s_1 + a_{12}s_2 \quad (۱-۲)$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2 \quad (۲-۲)$$

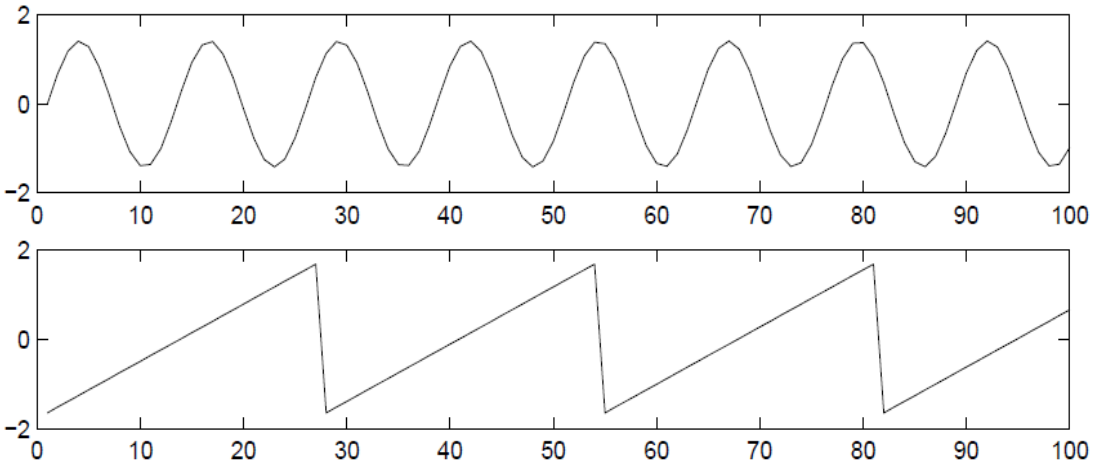
[1] Statistical Learning

a_{ij}ها پارامترهایی هستند که به فاصله میکروفون‌ها از گوینده‌ها بستگی دارند. اگر بتوان دو سیگنال گفتار اصلی، یعنی $s_i(t)$ها را با استفاده از سیگنال‌های ثبت شده $x_i(t)$ تخمین زد، بسیار مفید خواهد بود. این مسئله به‌عنوان کوکتل-پارتی[1] نامیده شده‌است. توجه گردد که هیچ‌گونه تأخیر یا فاکتور اضافی، در مدل آزمایشی وارد نمی‌شود.

به‌عنوان‌مثال، شکل موج‌های شکل (۲-۱) و شکل (۲-۲) را درنظر بگیرید، البته توجه گردد که این شکل موج‌ها سیگنال گفتار واقعی نیستند. سیگنال‌های گوینده در شکل (۲-۱) و مخلوط سیگنال‌ها در شکل (۲-۲) آورده شده است. حال مسئله جداسازی این شکل موج‌ها مطرح می‌باشد.

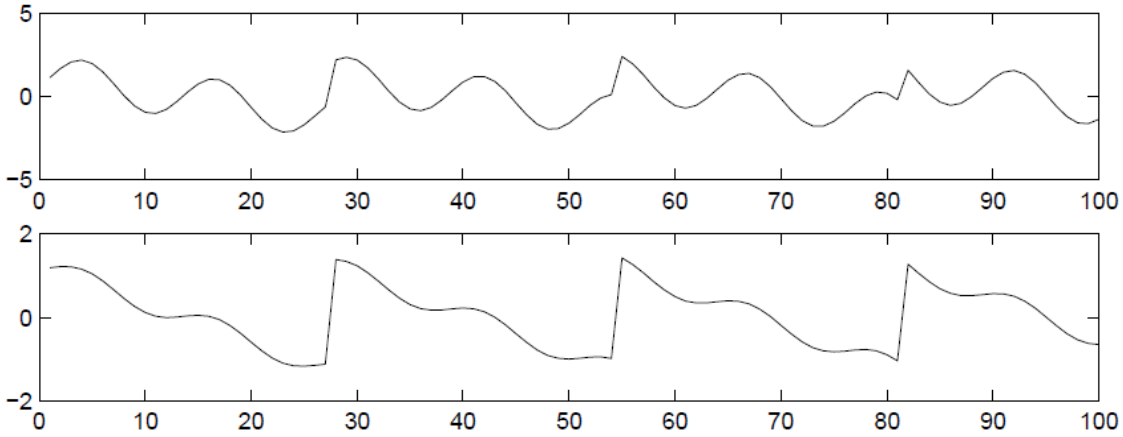
در واقع، اگر پارامترهای a_{ij} موجود باشند، مسئله حل شده است؛ در غیر اینصورت، مسئله پیچیده‌تر و شامل درنظر گرفتن موارد دیگر در حل این مسئله می‌باشد.

در سال‌های اخیر روش آنالیز مولفه‌های مستقل بسیار توسعه یافته است. به‌طوری‌که می‌توان با استفاده از تخمین a_{ij}ها مبتنی بر اطلاعات مستقل آن‌ها، امکان جداسازی دو سیگنال اصلی را از سیگنال‌های مخلوط شده انجام دهد. شکل (۲-۳) سیگنال‌های تخمین زده شده توسط روش ICA را نشان می‌دهد. همانطور که دیده می‌شود این سیگنال‌ها بسیار نزدیک به سیگنال اصلی هستند.

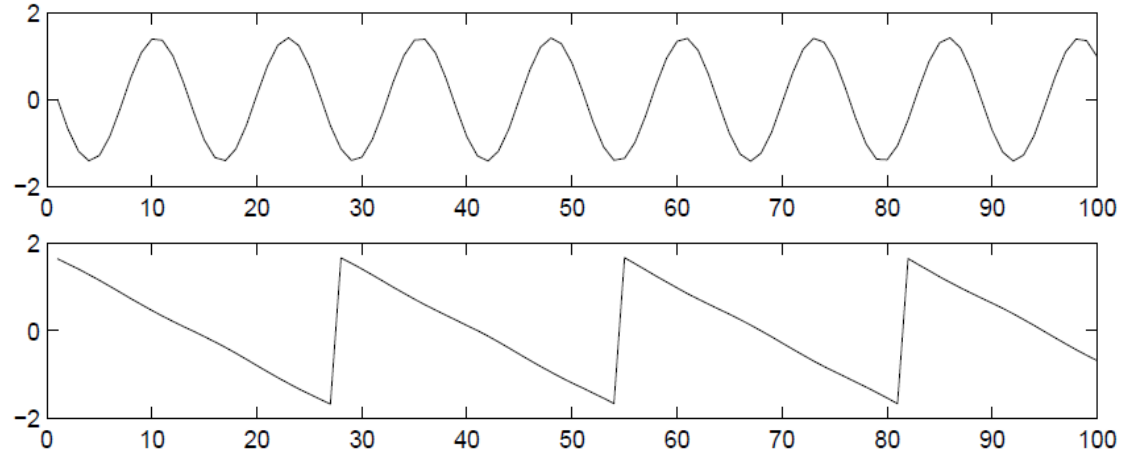


شکل (۲-۱) سیگنال‌های اصلی [۱۱]

[1] Cocktail-party



شکل (۲-۲) ترکیب شده سیگنال‌های اصلی با هم [۱۱]



شکل (۲-۳) سیگنال‌های تخمین زده شده توسط روش ICA [۱۱]

در آنالیز مولفه‌های مستقل بردار $x(k)\in R^m$ با یافتن ماتریس جداسازی $n\times m$ (که $m \geq n$ است) W بدست می‌آید.W از رتبه کامل[1] است و طوری به‌دست می‌آید که بردار سیگنال خروجی تخمین زده شده $y(k) = [y_1(k),\ldots,y_n(k)]^T$، تا حد امکان مستقل باشد.

$$y(k) = W.x(k) \qquad (۲-۳)$$

استقلال متغیرهای تصادفی، مفهوم کلی‌تری از ناهمبستگی است. می‌توان گفت متغیرهای تصادفی y_i و y_j از لحاظ آماری مستقل هستند اگر اطلاعات درباره مقادیر y_i هیچ اطلاعاتی را درباره مقادیر y_j فراهم نکند. از لحاظ ریاضی استقلال بین y_i و y_j را می‌توان با رابطه زیر بیان‌کرد [۱۱]:

$$p(y_i , y_j) = p(y_i) . p(y_j) \qquad (۲-۴)$$

[1]Full rank

p(y) نشان‌دهنده تابع چگالی احتمال متغیر تصادفی y است. به بیان دیگر سیگنال‌ها مستقل هستند، اگر تابع چگالی احتمال آن‌ها قابل جداسازی باشد.

ICA یک روش محاسباتی برای تجزیه یک سیگنال به زیرمؤلفه‌های جمع‌شونده با فرض استقلال آماری سیگنال‌های منابع غیرگوسی است. این روش، یک مورد خاص از روش‌های جداسازی کور منابع است. در پردازش‌های چند میکروفونه تحت شرایط خاص، امکان جداسازی کامل منابع از سیگنال‌های مخلوط با استفاده از روش ICA وجود دارد [۱۱]. این شرایط عبارتند از:

- تعداد میکروفون‌ها برابر و یا بیشتر از تعداد منابع باشد. به بیان دیگر تعداد ترکیبات خطی مشاهده شده نباید کمتر از تعداد مولفه‌های مستقل باشد.
- منابع باید از یکدیگر مستقل باشند. در واقع تمام مولفه‌های مستقل به جز یکی از این مولفه‌ها باید غیرگوسی باشند.
- تعداد منابع درسیگنال‌های مخلوط مشاهده شده معین باشد.
- ماتریس A باید از مرتبه کامل باشد.

۲-۲-۱- ابهامات ICA

در مدل ICA به‌صورت x = A.s این ابهامات قابل مشاهده است [۱۱]:

اول این‌که نمی‌توان انرژی مولفه‌های مستقل را تعیین کرد. رابطه $x = \sum_{i=1}^{n} a_i \, . \, s_i$ را در نظر بگیرید. از آن‌جا که s و A هر دو نامعلوم هستند هر اسکالری که در هریک از منابع s_i ضرب شود، می‌تواند توسط تقسیم ستون متناظر a_i در A با همان اسکالر حذف شود. در نتیجه می‌توان دامنه مولفه‌های مستقل را ثابت در نظر گرفت. با توجه به این‌که مولفه‌ها، متغیر تصادفی هستند بهترین راه برای انجام این کار فرض واریانس واحد $E\{s_i^2\} = 1$ برای هر یک از این مولفه‌هاست. سپس ماتریس A با درنظر گرفتن این محدودیت به‌دست می‌آید. این ابهام در بیشتر کاربردها قابل اغماض است.

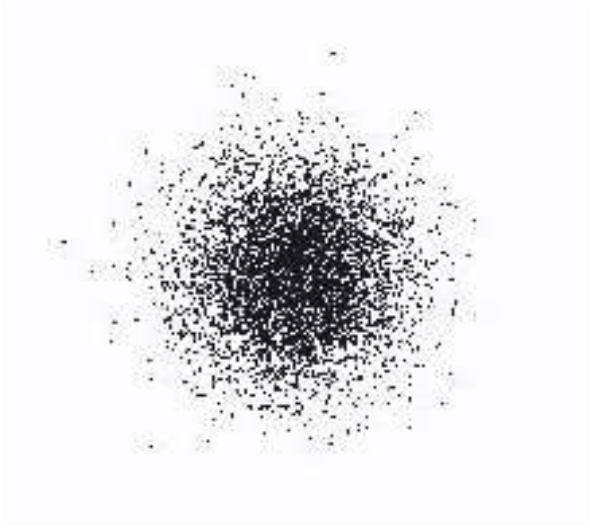
دوم این‌که نمی‌توان ترتیب مولفه‌های مستقل را تعیین کرد. به بیان ریاضی ماتریس جایگشت P و معکوس آن را می‌توان این‌گونه در مدل جایگزین کرد: $x=AP^{-1}Ps$. عناصر Ps متغیرهای مستقل منابع اصلی S_i در یک ترتیب جدید هستند و ماتریس AP^{-1} یک ماتریس ترکیب‌کننده نامعلوم جدید است که توسط الگوریتم‌های ICA به‌دست می‌آید.

۲-۲-۲- غیرگوسی بودن متغیرها در روش ICA

یکی از محدودیت‌های اساسی ICA این است که مولفه‌های مستقل نباید گوسی باشند. فرض‌کنید ماتریس ترکیب کننده متعامد و S_iها گوسی هستند و x_1 و x_2 گوسی، ناهمبسته و با واریانس واحد هستند. در این صورت چگالی توأم آن‌ها به شکل زیر است:

$$p(x_1, x_2) = \frac{1}{2\pi}\exp\left(\frac{-x_1^2 + x_2^2}{2}\right) \quad (۵-۲)$$

این توزیع در شکل (۲-۴) آورده شده است. با توجه به شکل مشخص است که چگالی کاملاً متقارن است. بنابراین شامل هیچ‌گونه اطلاعاتی در جهت ستون‌های ماتریس ترکیب کننده A نمی‌باشد. به‌همین دلیل A قابل تخمین نیست. اگر فقط یکی از مولفه های مستقل گوسی باشد مدل ICA قابل تخمین است.



شکل (۲-۴) چگالی توأم دو متغیر گوسی [۱۱]

۲-۲-۳- معیارهای غیرگوسی بودن

برای بررسی غیرگوسی بودن نیاز به یک معیار کمی است. دو تا از معیارهای غیرگوسی بودن را که در این‌جا معرفی می‌شود عبارتند از [۱۱]:

a) Kurtosis

b) Negentropy

۲-۲-۳-۱- Kurtosis

معیار کلاسیک غیرگوسی بودن kurtosis یا کومولانت مرتبه چهارم است و به این صورت تعریف می‌شود:

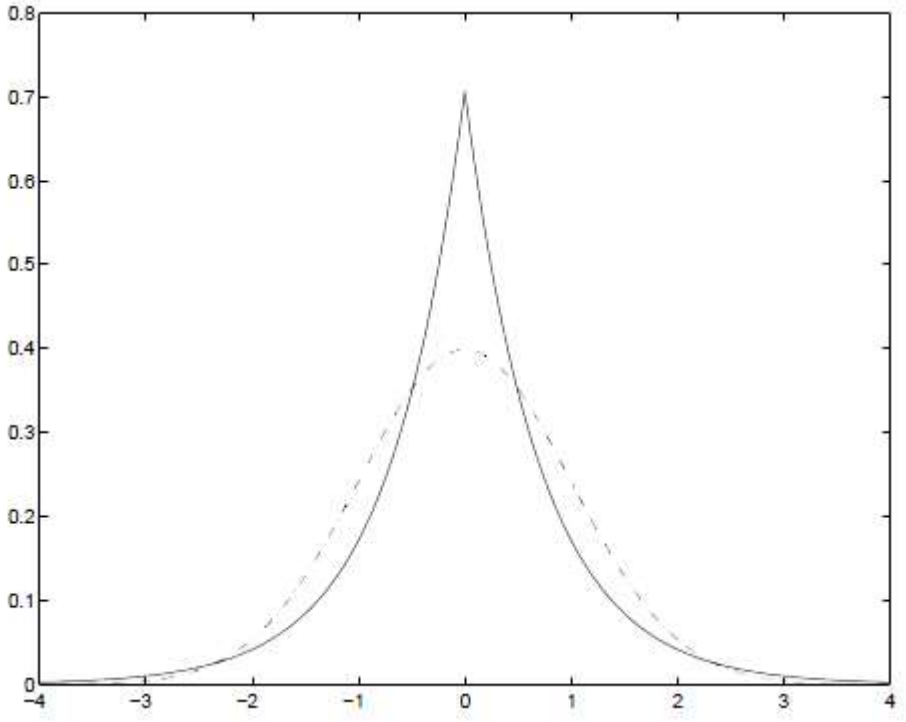
$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \qquad (۲-۶)$$

از آن‌جا که فرض می‌شود y دارای میانگین صفر و واریانس واحد باشد، بنابراین سمت راست رابطه بالا به‌صورت $E\{y^4\} - 3$ ساده می‌شود. این امر نشان می‌دهد که kurtosis شکل نرمالیزه شده ممان چهارم یعنی $E\{y^4\}$ می‌باشد. برای یک متغیر گوسی، kurtosis صفر و برای اکثر متغیرهای غیر گوسی، مخالف صفر می‌باشد. kurtosis می‌تواند مثبت یا منفی باشد. متغیرهای تصادفی که kurtosis آن‌ها منفی است، زیرگوسی[1] و متغیرهای تصادفی که kurtosis آن‌ها مثبت است، فراگوسی[2] نامیده می‌شوند. متغیرهای تصادفی فراگوسی نوعاً دارای تابع چگالی احتمال‌های نوک تیز هستند، به‌طوری‌که pdf آن‌ها در صفر دارای اندازه بزرگی می‌باشند؛ در حالی‌که در مقادیر متوسط مقدار pdf آن‌ها کم است. به‌عنوان‌مثال می‌توان به توزیع لاپلاس که pdf آن به واریانس ۱ نرمالیزه شده است، به‌صورت زیر اشاره کرد:

$$p(y) = \frac{1}{\sqrt{2}} exp(\sqrt{2}|y|) \qquad (۲-۷)$$

این pdf در شکل (۲-۵) رسم شده است. خطوط نقطه‌چین تابع چگالی توزیع گوسی را نشان می‌دهد. هر دو چگالی به واریانس یک نرمالیزه شده‌اند.

[1] Sub-Gaussian
[2] Super-Gaussian



شکل (۲-۵) تابع چگالی توزیع لاپلاس [۱۱]

متغیرهای تصادفی زیرگوسی نوعاً دارای یک pdf صاف می‌باشند. مقادیر این pdf نزدیک به صفر، ثابت است و در مقادیر بزرگتر، مقدار pdf بسیار کوچک است. مثال این pdf توزیع یکنواخت است.

چه از لحاظ محاسباتی و چه ازلحاظ تئوری معیار kurtosis ساده می‌باشد. از نظر محاسباتی kurtosis به‌سادگی توسط محاسبه ممان مرتبه چهارم قابل دستیابی است. همچنین از نظر آنالیزهای تئوری نیز ساده می‌باشد چرا که دارای دو خاصیت خطی زیر است:

$$kurt(x_1 + x_2) = kurt(x_1) + kurt(x_2) \quad (۸-۲)$$

$$kurt(ax_1) = a^4\, kurt(x_1) \quad (۹-۲)$$

که a یک اسکالر است. برای محاسبه مولفه‌های مستقل توسط این معیار، در عمل از یک بردار وزن W شروع می‌شود و جهتی را که kurt(y) در آن جهت ماکزیمم (در صورت مثبت بودن kurtosis) یا مینیمم (در صورت منفی بودن) می‌شود را پیدا می‌کند و از روش گرادیان و یا یکی از بسط‌هایش جهت محاسبه W جدید استفاده می‌کنیم. همچنین مشکل عمده این معیار ناپایداری آن است.

Negentropy -۲-۳-۲-۲

یک معیار مهم دیگر برای غیر گوسی بودن negentropy است. این معیار براساس کمیت تئوری اطلاعاتی آنتروپی تفاضلی می‌باشد. آنتروپی یک مفهوم اصلی در تئوری اطلاعات است. آنتروپی یک متغیر تصادفی، به‌صورت میزان اطلاعاتی که آن متغیر به ما می‌دهد، بیان می‌شود. هرچه متغیر تصادفی‌تر یعنی غیرقابل پیشگویی باشد، مقدار آنتروپی نیز بیشتر است.

آنتروپی H برای یک متغیر تصادفی گسسته به‌صورت زیر تعریف می‌شود:

$$H(Y) = -\sum_i P(Y = a_i).\log P(Y = a_i) \qquad (۱۰-۲)$$

که a_i مقادیر ممکن y می‌باشد. این تعریف را می‌توان به متغیرهای تصادفی پیوسته نیز تعمیم داد که در این حالت آنتروپی تفاضلی گفته می‌شود. آنتروپی تفاضلی بردار تصادفی y با چگالی f(y) به‌صورت زیر تعریف می‌شود:

$$H(y) = -\int f(y).\log f(y)dy \qquad (۱۱-۲)$$

یکی از نتایج اصلی تئوری اطلاعات این است که متغیر گوسی بیشترین آنتروپی را در میان متغیرهای تصادفی با واریانس یکسان دارد. این بدان معناست که می‌توان از این ویژگی جهت معیار غیرگوسی بودن استفاده کرد. شکل اصلاح شده، آنتروپی‌تفاضلی negentropy نامیده شده و به‌صورت زیر تعریف می‌شود:

$$J(y) = H(y_{gauss}) - H(y) \qquad (۱۲-۲)$$

که y_{gauss} یک متغیر تصادفی با ماتریس کوواریانس مشابه y می‌باشد. negentropy همیشه غیر منفی و صفر است اگر و تنها اگر y دارای توزیع گوسی باشد. از دیگر ویژگی جالب negentropy متغیر نبودن آن نسبت به تبدیلات خطی معکوس‌پذیر است. همچنین یکی از مشکلات این معیار، پیچیدگی محاسباتی آن است. تخمین negentropy با استفاده از تعریف، نیاز به تخمین تابع چگالی احتمال دارد. بنابراین تقریب‌های ساده‌تری از negentropy بسیار مفید خواهند بود.

در واقـــــــع شرایط بیان شده در بالا باعث ایجاد محدودیت‌هایی در استفاده از ICA به‌عنوان یک راه‌حل برای مسأله جداسازی گفتار تک‌میکروفونه می‌شـــود. در [۱۲] جهت گسترش روش جداسازی کور منابع برای مسأله جداسازی گفتار تک‌میکروفونه، از روش بیشـــــترین شباهت برای پیشنهاد یک الگوریتم ICA با نظـــارت، استفاده کـــرده است. روش پیشنهادی برای مخلوطی از سیگنال گفتار و موزیک به‌خوبی عمل می‌کند؛ اما در شرایطی‌که سیگنال‌های مخلوط گفتار باشند، عملکرد آن به‌شدت افت می‌کند [۱۲]. این به‌دلیل همپوشـانی زیـاد سیگنال‌های گفتار در حـوزه زمان و فرکانس‌است. در [۱۳]، روش ICA با یک ماسک زمان-فرکانس بـــرای جداســـازی گفتار با استفاده از دو میکروفون ترکیب می‌شود. در این روش فرض بر این است که هیچ اطلاعی از تعـــــداد منابع صوتی در دسترس نمی‌باشد.

نوع دیگری از روش ICA، روش آنالیز زیر فضای مستقل (ISA)[1]است که این روش، پردازش سیگنال‌های مشاهده شده را با افزایش بعد آن‌ها از ۱ به N انجام می‌دهد. عملکرد ISA بر روی سیگنال به‌صورت تجزیه آن به یک مجموعه با N پایه مستقل است. به‌طوری‌که باید این پایه‌ها با یکدیگر گروه‌بندی شوند و تشکیل زیرمجموعه‌هایی وابسته به هر منبع موجود در سیگنال مخلوط را بدهند. Davies و James نشان داده‌اند که ISA زمانی می‌تواند یک روش‌کارآمد باشد که طیف منابع از یکدیگر جدا باشند و این بدین مفهوم است که استقلال خطی پایه‌ها تضمین می‌شود [۱۴]. ISA با به‌کارگیری مؤلفه‌های دینامیکی برای نمایش سیگنال‌های غیرایستا، توسعه یافته روش ICA است. در روش ISA برخلاف ICA نیاز به داشتن میکروفون‌هایی حداقل برابر با تعداد منابع نمی‌باشد. بنابراین تعداد سینگال‌های مخلوط مورد نیاز از یک سیگنال به بالا می‌تواند باشد [۱۴].

۲-۳- روش‌های مبتنی بر یادگیری آماری

در روش‌های یادگیری آماری هیچ فرضی در مورد منابع در نظر گرفته نمی‌شود.

[1]Independent Subspace Analysis

در این‌روش‌ها، یادگیری بر اساس داده‌های مخلوط مشاهده شده و یا داده‌های قبلی موجود از هر کدام از منابع است. در ادامه به برخی کارهای انجام‌شده در این زمینه اشاره‌ای مختصر می‌گردد.

Roweis در [۱۵] یک مدل مارکف مخفی[1] (HMM)، برای داده‌های گفتار تمیز از یک گوینده را آموزش می‌دهد. در این روش با ترکیب HMM دو گوینده یک HMM وابسته تولید می‌شود که با استفاده از آن، احتمال این‌که یک نقطه در حوزه زمان-فرکانس سیگنال مخلوط مشاهده شده، مربوط به کدام گوینده است، تخمین زده می‌شود. جداسازی منابع به‌وسیله بازسازی سیگنال از نقاط زمان فرکانسی است که با احتمال بیشتری متعلق به گوینده مورد نظر می‌باشند و دیگر نقاط زمان-فرکانس صفر در نظر گرفته می‌شود.

دیدگاه Bach و Jordan به مسأله جداسازی منابع به‌صورت یک مسأله بخش‌بندی زمان-فرکانس است [۱۶]. در این روش به‌جای مدل کردن رفتار منابع به‌طور جداگانه، طیف سیگنال به دو یا چند مجموعه گسسته بخش‌بندی می‌شود. مجموعه‌ای از ویژگی‌ها مانند: فراز-فرود و فرکانس گام در حوزه زمان-فرکانس برای بخش‌بندی‌کننده[2] تولید می‌گردد. بخش‌بندی‌کننده طیفی، به‌وسیله ماتریس‌های مشابهت، پارامتری را برای هر ویژگی در فضای ویژگی‌ها تعریف می‌کند. مقادیر پارامترهای ماتریس مشابهت به‌وسیله داده‌های یادگیری و با استفاده از الگوریتم یادگیری طیف محاسبه می‌گردد.

روش دیگر برای جداسازی منابع در حوزه زمان-فرکانس، نمایش حوزه زمان-فرکانس سیگنال مخلوط مشاهده شده به‌وسیله یک ترکیب خطی از بردارهای پایه می‌باشد که این بردارهای پایه نمایشگر هر منبع هستند. دو روش معمول برای تخمین بردارهای پایه برای یک منبع، کوانتیزاسیون برداری [۱۷] و ماتریس فاکتوریل نامنفی [۱۸] است.

[1] Hidden Markov Model
[2] Segmenter

در [۱۷]، Ellis و Weiss یک کتاب کد کوانتیزاسیون برداری برای فریم‌های طیف زمان کوتاه داده‌های آموزشی یک منبع تولید کردند. در این روش سیگنال مخلوط مشاهده شده، به بردارهای کتاب کد منابع تصویر می‌شوند و در نهایت سیگنال جدا شده از بردارهایی که بیشترین تطابق را با یکدیگر دارند، ساخته می‌شوند. برای تکمیل این روش، یک HMM برای اعمال محدودیت‌های ترتیبی بین بردارهای کتاب کد، به آن افزوده شده است. این روش بر روی سیگنال گفتار مخلوط شده با نویزی شبیه گفتار، ارزیابی شده است. نتایج نشان می‌دهد که در برخی مواقع سیگنال گفتار به‌خوبی بهسازی نگردیده است.

در روش فاکتورگیری ماتریس نامنفی[1] (NMF) [۱۸]، دامنه تبدیل فوریه زمان کوتاه داده آموزشی از یک منبع، به‌وسیله حاصل‌ضرب یک ماتریس نامنفی الگوهای طیفی در یک ماتریس نامنفی الگوی زمانی مرتبط به‌دست می‌آید: $Y = HW$، که در این رابطه، ستون‌های ماتریس H بردارهای پایه‌ای هستند که با یکدیگر ساختار طیفی نمایش زمان-فرکانس سیگنال مخلوط را معین می‌کنند. همچنین سطرهای ماتریس W ساختار زمانی و وزن‌هایی را تعریف می‌کنند که به‌وسیله آن‌ها منابع در ماتریس مخلوط Y فعال هستند. سپس همین تکنیک بر روی سیگنال مخلوط مشاهده شده، اعمال می‌گردد. جداسازی به‌وسیله سنتز دامنه تبدیل فوریه زمان کوتاه از الگوهای طیفی و زمانی از منبع مطلوب به‌دست می‌آید.

روش NMF عمل جداسازی را با تصویر یک بردار ویژگی مخلوط به زیرفضای پیوسته منابع و سپس محاسبه قسمت‌های مربوطه به هر زیرفضا انجام می‌دهد [۱۸]. [۱۹] یک روش آنالیزی را جهت معین نمودن شرایطی‌که تحت آن تنها یک ماتریس NMF واحد وجود داشته باشد، به‌کار می‌گیرد. چندین نوع الگوریتم NMF برای یادگیری ماتریس H و W از روی ماتریس Y بر پایه قیدهای پیوستگی زمانی [۲۰] و یا بر پایه قیدهای تنکی [۱۸] جهت بهبود کیفیت سیگنال جدا شده پیشنهاد شده است. در [۲۱] یک الگوریتم تجزیه نامنفی تنک برای جداسازی منابع صوتی در شرایط

[1] Non-negative Matrix Factorization

تک‌میکروفونه ارائه شده است. در این روش، الگوریتم فیلتر وینر با تعمیم شرط ایستایی محلی به مدل منابع پارامتری غیرگوسی بهبود داده می‌شود. در [۲۲]، با تکیه بر پردازش گوسی اطلاعات قبلی، یک روش کلی برای به‌کارگیری اطلاعات قبلی برای NMF ارائه شده است. روش پیشنهادی در [۲۲] نشان می‌دهد که با انتخاب توزیع قبلی مناسب می‌توان به نتایج بهتری نسبت به روش NMF رسید.

NMF به‌منظور جداسازی گفتار تک‌میکروفونه به دو صورت به‌کار برده می‌شود: با نظارت[1] و بدون نظارت[2]. در روش بدون نظارت NMF به‌طور مستقیم و بدون داشتن اطلاعات قبلی از منابع، طیف سیگنال مخلوط را به طیف‌های سیگنال‌های موجود در سیگنال مخلوط تجزیه می‌کند. روش NMF با نظارت، این عمل را با سنتز منابع با استفاده از یک مجموعه آموزش داده شده بردارهای پایه، انجام می‌دهد. در [۲۳] یک روش جداسازی، با ترکیب روش NMF و مدل‌سازی منابع ارائه شده است. در واقع روش موجود به‌منظور حل پیچیدگی انتخاب بهینه بردارهای پایه در مرحله آموزش، ارائه شده است. در روش NMF اطلاعات فاز، مورد استفاده قرار نمی‌گیرد و تنها با استفاده از ماتریس طیف عمل جداسازی انجام می‌گیرد؛ اما در روش ارائه شده در [۲۳] این مسأله مورد بررسی قرار گرفته است.

۲-۴- مدل‌سازی منبع

در روش مدل‌سازی منبع برای جداسازی منابع، از یک مدل پارامتری برای هر منبع استفاده می‌شود. پارامترهای مدل برای هر منبع از مخلوط‌های مشاهده شده تخمین زده می‌شود. جدا کردن سیگنال‌ها به‌وسیله تولید سیگنال از مدل هر منبع، و یا به‌وسیله فیلتر کردن سیگنال مخلوط با فیلتری که توسط مقادیر پارامترهای مدل منبع کنترل می‌شود، صورت می‌پذیرد.

یکی از روش‌های اولیه مدل‌سازی منابع، روش انتخاب هارمونیک است که توسط Parsons ارائه شده است. در این روش یک مدل هارمونیکی برای گفتار صدادار فرض می‌شود. همچنین در این

[1] Supervised
[2] Unsupervised

روش، الگوریتم تخمین و ردیابی گام[1] بر پایه برداشتن قله‌های طیف، استفاده می‌شود. الگوریتم پیشنهادی با حذف هارمونیک‌های مربوط به گوینده تداخل کننده در حوزه زمان-فرکانس و بازسازی هارمونیک‌های گوینده مورد نظر در حوزه زمان، سیگنال گفتار دوگوینده را از یکدیگر جدا می‌سازد. نتایج به‌دست آمده نشان دهنده عملکرد خوب این روش بوده؛ اما این روش با آزمون شنوایی معمول ارزیابی نگردیده است.

تکنیک حذف دامنه هارمونیک (HMS)[2]بیان شده توسط Hanson و Wong نیز استراتژی‌هایی شبیه روش قبل دارد. در این روش پارامترهای هارمونیک‌های گفتار تداخلی حدس زده می‌شوند. در واقع در یک سری موقعیت‌های عملی قوی‌تر از گفتار هدف، تخمین زده می‌شوند. سپس هارمونیک‌های سیگنال تداخلی از سیگنال مخلوط کم می‌شود تا سیگنال گفتار گوینده هدف بازسازی شود. Morgan در [۲۴] با به‌کارگیری تخمین‌گر گام ML [۱۲] این تکنیک را بهبود داده است. همچنین با به‌کارگیری یک مرحله بهسازی باعث بهبود هارمونیک‌ها و فورمنت[3]های‌گوینده هدف شده است.

Stubbs و Summerfield دو روش جداسازی گفتار بر پایه مدل هارمونیک را مورد ارزیابی قرار داده‌اند [۲۵]: ۱) فیلتر کپسترال[4]، و ۲) تکنیک ترکیبی که در آن، روش فیلتر کپسترال با انتخاب هارمونیک ترکیب شده است. در روش فیلتر کپسترال، کپستروم[5] مخلوط دو سیگنال گفتار دارای پیک‌هایی در محل گام دو گفتار است. با حذف یکی از پیک‌ها می‌توان عمل‌جداسازی دو گوینده را انجام داد. در روش هیبرید [۲۶]، انتخاب هارمونیک برای بالا بردن کیفیت جداسازی در هنگامی‌که گفتار هدف بی‌صدا و گفتار تداخلی صدادار است، به فیلتر کپستروم اضافه می‌شود. این الگوریتم جداسازی بر روی انسان با شنوایی نرمال و سمعک‌دار، ارزیابی شده است. عملکرد هر دو سیستم

[1]Pitch
[2]Harmonic Magnitude Suppression
[3]Formants
[4]Cepstral Filter
[5]Cepstrum

هنگام ارزیابی با سیگنال‌های گفتار با گام تک‌تن، خوب بوده است. اما هنگامی‌که با سیگنال‌های گفتار نرمال ارزیابی شده، تنها روش فیلتر کپستروم و تنها برای انسان با سیستم شنوایی نرمال به‌خوبی کار کرده است.

با توجه به نظرات Hu و Wang در [۲۷]، مشکل اصلی برای همه روش‌های ذکر شده، عدم توانایی آن‌ها در برخورد با قسمت‌های فرکانس بالای گفتار است. چنانچه یک هارمونیک یک جزء غالب در یک کانال بانک فیلتر باشد، آن را قابل تفکیک می‌نامند؛ در غیر این صورت، آن هارمونیک غیرقابل تفکیک نامیده می‌شود. برای بانک فیلتر شنوایی استفاده شده روش Hu و Wang، فرکانس‌های پایین قابل تفکیک و فرکانس‌های بالا غیر قابل تفکیک می‌باشند. در این روش پارامترهای هارمونیک‌های قابل تفکیک به‌وسیله تجزیه AM-FM کانال بانک فیلتر مربوطه، تخمین زده می‌شوند. همچنین مدولاسیون دامنه هارمونیک‌های غیر قابل تفکیک برای تخمین پارامتر آن‌ها مورد ارزیابی قرار می‌گیرند. هارمونیک‌های تخمین زده شده برای گوینده‌های متفاوت بر اساس پیوستگی زمانی، همبستگی متقابل و مدولاسیون دامنه برچسب‌گذاری می‌شوند.

تعدادی از روش‌های جداسازی، منابع را به‌صورت یک مدل سینوسی کلی، مدل‌سازی می‌نمایند [۲۸]. در اغلب این روش‌ها یک محدودیت هارمونیک برای بهبود عملکرد تخمین پارامترها، به مدل سینوسی اضافه می‌شود. روش بازسازی منابع جدا شده در تکنیک مدل‌سازی سینوسی مشکل می‌باشد. در این روش برای بازسازی سیگنال جدا شده، یک بانک از نوسان‌سازهای متغیر با زمان با پارامترهای دامنه، فرکانس و فاز هارمونیک‌های منبع سیگنال جدا شده، استفاده می‌شوند.

همانطور که بیان کردیم هدف از این پایان نامه، جداسازی گفتار تک‌میکروفونه می‌باشد. در قسمت اول این فصل، به بیان راه‌حل‌های موجود برای جداسازی دو گوینده پرداختیم. همانطور که بیان شد، در روش فیلتر فضایی حداقل نیاز به دو میکروفون می‌باشد. همچنین این روش قادر به جداسازی سیگنال‌های رسیده به میکروفون‌ها از یک مسیر نمی‌باشد.

روش‌های موجود برای جداسازی منابع صوتی در شرایط تک‌میکروفونه، به دو دسته کلی روش‌های بر پایه ویژگی[1] و روش‌های بر پایه مدل‌سازی[2] تقسیم می‌شوند. در روش‌های بر پایه مدل‌سازی، هر کدام از سیگنال‌های هدف و تداخل با استفاده از مدل‌های آماری، مدل می‌شوند. مدل به‌کار رفته برای جداسازی دو گوینده مورد استفاده قرار می‌گیرد. مشکل این روش برخورد با تداخل‌های جدید است زیرا که برای سیگنال‌های تداخلی جدید مدلی در نظر گرفته نشه است. همچنین در این روش، برای مدل‌سازی سیگنال‌های هدف و تداخل نیاز به داده‌هایی تمیز از هر دو سیگنال داریم. در روش بر پایه ویژگی، با توجه به ویژگی‌های موجود در سیگنال گفتار و نیز با توجه به سیستم شنوایی انسان، عمل جداسازی انجام می‌گیرد. در ادامه به بررسی روش‌های بر پایه ویژگی موجود برای جداسازی گفتار تک‌میکروفونه می‌پردازیم.

۲-۵- روش‌های بر پایه ویژگی

همانطور که می‌دانیم انسان در محیطی زندگی می‌کند که در برگیرنده منابع متعدد صوتی بوده و بنابراین صدای رسیده به گوش انسان، ناشی از چندین منبع صوتی است. در حالی‌که موضوع بهسازی گفتار با استفاده از یک میکروفون یک موضوع بحث برانگیز شده است، سیستم شنوایی انسان توانایی قابل توجهی را برای جداسازی گفتار از خود نشان داده است. بر همین اساس، Bregman در سال ۱۹۹۴ سیستم جداسازی شنوایی را با عنوان ASA در [۵] مطرح کرد. با الهام گرفتن از اصول ASA، سیستم محاسباتی ASA برای جداسازی گفتار ارائه شد. این سیستم که CASA نامیده شد، بدون استفاده از فرضیات در خصوص خواص گفتار و تداخل و تنها با استفاده از اصول ادراکی و خواص ذاتی صدا مانند هارمونیک بودن[3]، فراز، فرود، محل منبع صدا[4]و اطلاعات قبلی[5] از صداهای خاص، عمل

[1]Feature-Based
[2]Model-Based
[3]Harmonicity
[4]Location
[5]Prior knowledge

جداسازی گفتار را انجام می‌دهد. در ادامه به بررسی کامل سیستم CASA و روش‌های موجود در این زمینه می‌پردازیم.

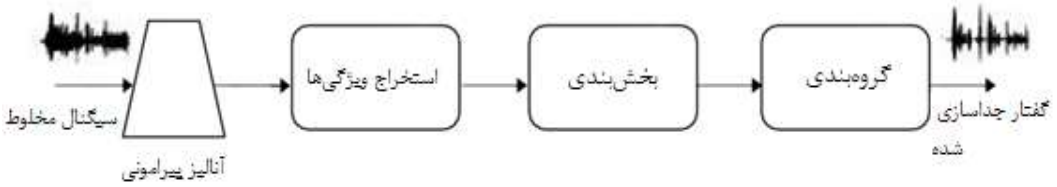
۲-۵-۱- روش‌های CASA

سیستم‌های CASA به‌طور گسترده برای جداسازی منابع گفتاری به‌کار برده می‌شوند. در سیستم CASA مجموعه‌ای وسیع از پیاده‌سازی محاسباتی سیستم آنالیز شنوایی (ASA) به‌کار برده می‌شود. قبل از توصیف سیستم CASA، ابتدا به بررسی سیستم آنالیز شنوایی و کاربردهای اصلی آن می‌پردازیم. بسیاری از دانشمندان عقیده دارند که سیستم شنوایی و بینایی بسیار به یکدیگر شباهت دارند. سیستم شنوایی گفتار را به سیستم عصبی منتقل نموده و در مغز عملیات پردازشی شبیه پردازش تصویر بر روی آن انجام می‌شود. کلیه عملیات سیستم شنوایی را می‌توان به دو مرحله دسته-بندی نمود: بخش‌بندی و گروه‌بندی. در مرحله اول، ابتدا گفتار به فضای دو بعدی زمان-فرکانس برده می‌شود و سپس واحدهای زمان-فرکانس شبیه به یکدیگر در یک بخش یا ناحیه قرار داده می‌شوند و به این ترتیب کل فضای زمان-فرکانس بخش‌بندی می‌شود. در مرحله دوم، بخش‌ها بر اساس ویژگی-های آکوستیکی و دیگر اطلاعات، به دسته‌های متفاوتی گروه‌بندی می‌شوند. در نهایت سیگنال گفتار هدف و یا تداخل و یا هر دوی آن‌ها برای اهداف گوناگون بازسازی می‌شوند.

به‌طور کلی سیستم CASA از روش‌های محاسباتی برای تولید یک سیستم ادراکی ماشینی شبیه سیستم شنوایی انسان، استفاده می‌کند. Wang و Brown در [۹]، CASA را زمینه مطالعاتی محاسباتی تعریف کردند که هدف آن به‌دست آوردن کارایی انسان در ASA، به‌وسیله یک یا دو میکروفون ضبط صدا است.

شکل (۲-۶) بلوک دیاگرام ساده‌ای از سیستم CASA را نشان می‌دهد. در ابتدا سیگنال گفتار وارد مرحله تابع تبدیل می‌گردد. در اغلب موارد، این تابع، سیگنال گفتار تک‌بعدی زمان را به‌صورت نمایش دو بعدی معروف زمان-فرکانس تبدیل می‌نماید. این انتقال توسط روش استاندارد تبدیل فوریه زمان

کوتاه[1] (STFT) و یا فیلتربانک گاماتون[2] (که از مشاهدات روان-آکوستیکی[3] محیط شنیداری به‌دست آمده، مدلی استاندارد از گوش درونی است [۲۹]) صورت می‌گیرد. در مرحله بعد همه واحدهای زمان-فرکانس به ناحیه‌هایی متفاوت بخش‌بندی می‌شوند. همه واحدهای زمان-فرکانس قرار گرفته شده در یک ناحیه، مربوط به یک منبع گفتار می‌باشند. الگوریتم‌های استخراج ویژگی متفاوتی در این مرحله برای بهینه‌سازی نتایج بخش‌بندی، ارائه شده است. در مرحله آخر، بخش‌های به‌دست آمده در مرحله قبل، گروه‌بندی می‌شوند. این مرحله بر اساس فرمت گفتار[4]، به‌منظور استخراج بخش‌های مربوط به منبع گفتار هدف و حذف دیگر بخش‌ها پردازش می‌شود. یکی از معمول‌ترین روش‌های به‌کار رفته در مرحله آخر، شناسایی گوینده[5] است. در نهایت همه واحدهای زمان-فرکانس استخراج شده، برای بازسازی و سنتز گفتار هدف به‌کار برده می‌شوند.



شکل (۲-۶) بلوک دیاگرام ساده‌ای از سیستم CASA [۶]

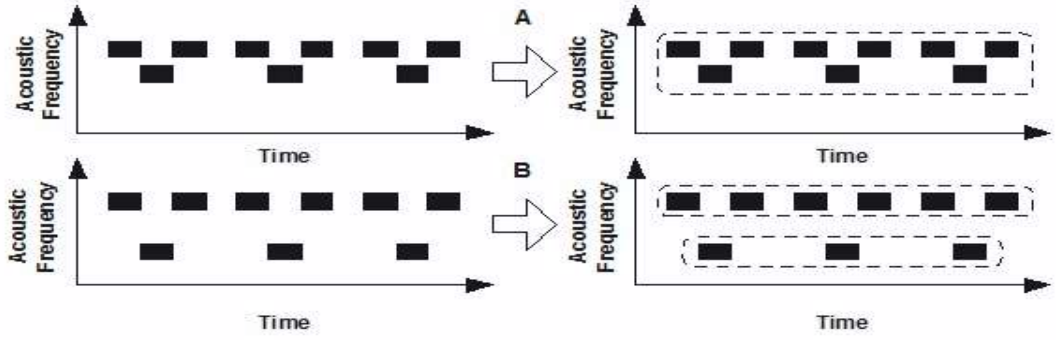
سیگنال گفتار، سیگنالی است که هم در بعد زمان و هم در بعد فرکانس، دارای محدوده دینامیکی بالایی می‌باشد. این ساختار غیرایستایی[6] گفتار باعث می‌گردد که بازشناسی و جداسازی گفتار در حضور سیگنال تداخلی گفتار دیگر، عملی پیچیده و مشکل گردد. به‌همین دلیل است که در سیستم‌های CASA به مسأله تصویر سیگنال گفتار یک بعدی به ابعاد بالاتر، توجه زیادی شده است.

در خیلی از اوقات سیگنال گفتار تخریب شده به‌وسیله تداخل، در حوزه زمان-فرکانس به‌دلیل پراکندگی انرژی[7] در این حوزه، قابل جداسازی از سیگنال تداخلی است. برای درک این موضوع با

[1] Short Time Fourier Transform
[2] Gammatone Filterbank
[3] Psychoacoustic Observations
[4] Utterance-Based Format
[5] Speak Identification
[6] Nonstationary
[7] Energy Sparsity

بررسی اسپکتروگرام سیگنال گفتار به‌راحتی می‌توان دید که همه واحدهای زمان-فرکانس نقش مهمی در نمایش اطلاعات مربوط به گفتار ندارند. درصد بالایی از واحدهای زمان-فرکانس، دارای انرژی کم و بسیار کم اهمیت‌تر از درصد کمی از واحدها هستند که دارای انرژی زیادی می‌باشند. تبدیل زمان-فرکانس باعث کاهش ابهام سیگنال گفتار نسبت به قبل از تبدیل می‌گردد. در نتیجه واحدهای زمان-فرکانس که کمتر به‌وسیله سیگنال گفتار تداخلی مورد تخریب قرار گرفته‌اند، به‌عنوان واحدهایی پایه، برای بازسازی سیگنال گفتار هدف مورد استفاده قرار می‌گیرند.

با توجه به مزایای تصویر دو بعدی، تئوری دو مرحله‌ای ASA می‌تواند به مسأله جداسازی منابع کمک زیادی نماید. شکل (۲-۷) شماتیک ساده‌ای را جهت اثبات این تئوری نشان می‌دهد. در شکل (۲-۷)-a ناحیه‌های زمان-فرکانس بدست آمده به‌علت نزدیک بودن به یکدیگر در محور فرکانس، در یک گروه دسته‌بندی می‌شوند؛ این در حالی است که در شکل (۲-۷)-b ناحیه‌هایی که در محور فرکانس از یکدیگر دور می‌باشند در دو گروه جدا از هم دسته‌بندی شده‌اند [۹،۳۰].



شکل (۲-۷) شماتیک ساده‌ای از روش گروه‌بندی ناحیه‌های به‌دست آمده در حوزه زمان-فرکانس در روش‌های CASA. a) گروه‌بندی ناحیه‌ها در یک دسته، به‌علت نزدیکی ناحیه‌ها در محور فرکانس. b) گروه‌بندی ناحیه‌ها در دو دسته، به علت دوری ناحیه‌ها در محور فرکانس [۹،۳۰].

۲-۵-۱-۱- مروری بر روش‌های CASA برای جداسازی گفتار تک‌میکروفونه

بیشتر سیستم‌های CASA به‌منظور جداسازی سیگنال گفتار از سیگنال تداخلی، توسعه یافته‌اند. تنها هدف تعداد محدودی از این سیستم‌ها، جداسازی انواع دیگر صداها است؛ مانند سیستم Mellinger که برای جداسازی موزیک و سیستم Li و Wang که برای جداسازی آواز از موزیک است.

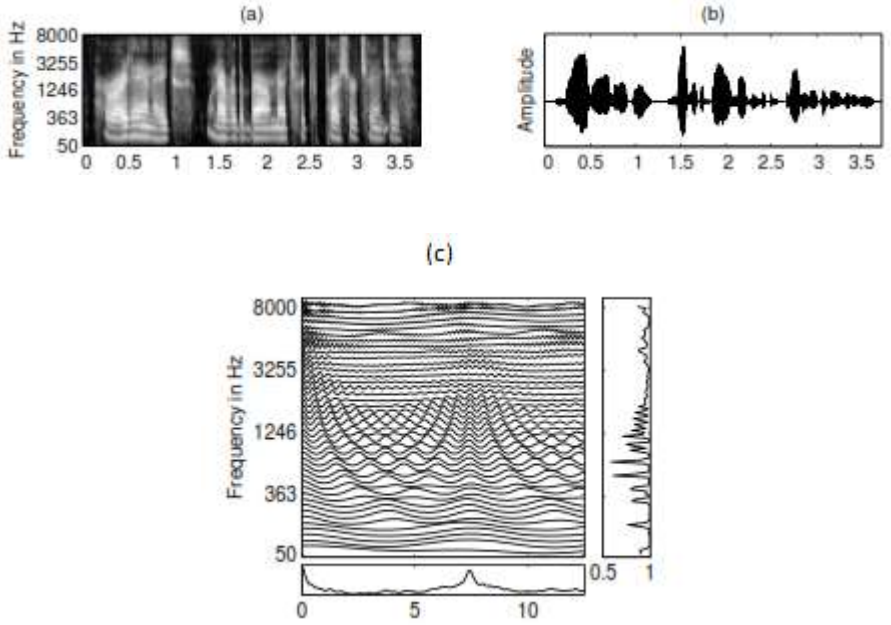
همانطور که بیان شد سیستم‌های CASA در ابتدا با استفاده از فیلتربانک، سیگنال را از حوزه زمان به فضای زمان-فرکانس منتقل می‌کنند. خروجی آنالیز اولیه سیستم CASA (خروجی فیلتربانک)، برای استخراج ویژگی‌هایی متناسب با ASA به‌کار برده می‌شود. بسیاری از سیستم‌های CASA برای جداسازی گفتار تک‌میکروفونه، بر روی جداسازی گفتار صدادار[1]، با استفاده از ویژگی تناوب، تمرکز یافته‌اند. بهترین نمایش برای آنالیز تناوب، همبستگی نگاشت[2] است که در بسیاری از سیستم‌های CASA به‌کار گرفته می‌شود. همبستگی نگاشت، نشان‌دهنده خودهمبستگی سیگنال‌های هرکانال از بانک فیلتر در محدوده زمانی یک دوره است. تناوب سیگنال به‌وسیله تابع خودهمبستگی[3] (ACF) نمایش داده می‌شود [۳۱].

جداسازی گفتار در حضور نویز و سایر تداخل‌ها، یکی از مسائل چالش برانگیز به‌ویژه در مورد جداسازی تک‌میکروفونه می‌باشد. گرچه پیشرفت‌های قابل ملاحظه‌ای در زمینه‌ی جداسازی گفتار صدادار به‌دست آمده است، با این وجود عملکرد سیستم‌های جداسازی CASA در بخش‌های فرکانس بالا مناسب نیست. پس برچسب‌گذاری دقیق واحدها در هر دو ناحیه‌ی فرکانس پایین و بالا می‌تواند تاثیر قابل توجهی در نتایج جداسازی داشته باشد. در [۳۲] روش‌های جدیدی برای برچسب‌گذاری واحدهای زمان-فرکانس[4] پیشنهاد می‌شود. به بیان دقیق‌تر، در مرحله‌ی جداسازی اولیه، برای انتخاب واحدهای زمان-فرکانس، همبستگی بین کانالی پوش پاسخ در تمام فرکانس‌ها به‌کار رفته است. همچنین، در مرحله ردیابی گام و برچسب گذاری واحدها در فرکانس‌های بالا، از تابع خودهمبستگی پوش بهبودیافته[5] (EEACF) برای حذف پیک‌های نادرست و یا مضارب پیک اصلی از منحنی تابع خودهمبستگی پوش[6] (EACF)، استفاده شده است که نشان‌دهنده نتایج بهتر جداسازی نسبت به روش‌های معمول می‌باشد.

---

[1] Voice Speech
[2] Correlogram
[3] Autocorrelation Function
[4] Time-Frequency (T-F) Units
[5] Enhanced Envelope Autocorrelation Function
[6] Envelope Autocorrelation Function

شـــکل (۲-۸) نمونه‌ای از همبستگی نگاشت را برای سیگنال گفتار مـــرد با طول ۰.۷ ثانیه نشان می‌دهد. هر کدام از نمودارهای نشان داده شده برای هر فرکانس، ACF بـــه‌دست آمده برای سیگنال باند باریک مربوط به کانال فیلتربانـــک بـــا همـــان فرکانس را نشان می‌دهد. ماکزیمم‌های نمودارهای ACF، دوره تناوب سیگنال مربوطه را نشان می‌دهد. با توجه به این‌که سیگنال گفتار در طول این بازه زمانی صـــدادار است، همه نمودارهای ACF، ماکزیمم‌هایی را در تأخیر ۵.۸۷ میلی‌ثانیه نشان می‌دهد که بیانگر دوره تناوب سیگنال گفتار در آن بازه زمانی است.



شکل (۲-۸) a) نمودار گوینده مرد در حوزه زمان-فرکانس. b) شکل‌موج گوینده مرد. c) همبستگی نگاشت برای سیگنال گوینده مرد با طول ۰.۷ ثانیه. پانل پایینی همبستگی نگاشت مختصر و پانل سمت راست تابع همبستگی متقابل بین کانال‌ها را نشان می‌دهد [۶].

این مسأله باعث ایجاد انگیزه‌هایی برای ایجاد الگوریتم تعیین گام به‌وسیله نمودارهای ACF گردیده که همبستگی نگاشت مختصر[1] نامگذاری شده‌اند. همبستگی نگاشت مختصر برابر مجموع نمودارهای ACF در همبستگی نگاشت بر روی محور فرکانس است که در پایین شکل (۲-۸) نشان داده شده است. ماکزیمم به‌دست آمده در همبستگی نگاشت مختصر، دوره تناوب سیگنال گفتار را نشان می‌دهد. در سال Weintraub۱۹۸۵، روشی برای جداسازی دو گوینده همزمان زن و مرد ارائه

[1]Summary Correlogram

کرد [۳۳]. در سیستم پیشنهادی وی، ابتدا نمودار گام هر گوینده، تخمین زده شده و سپس با تخمین دامنه‌های طیف هر منبع، سیگنال گفتار بر اساس تناوب و پیوستگی‌زمانی، دو سیگنال گفتار از یکدیگر جدا می‌شوند. سیستم جداسازی Weintraub شامل سه مرحله به شرح زیر است:

در مرحله اول، نمودارهای گام مربوط به هر گوینده تخمین زده می‌شود. در این مرحله، سیستم از تابع خودهمبستگی برای پیدا کردن تناوب سیگنال استفاده می‌کند. بر اساس این تابع، سیستم یک گام غالب در هر فریم زمانی به طول ۱۰ میلی‌ثانیه را معین می‌کند. گام غالب ممکن است به هر یک از دو گوینده تعلق داشته باشد. گام غالب با اطلاع از این‌که یک گوینده مرد و دیگری زن است، به گوینده صحیح نسبت داده می‌شود. با استفاده از گام غالب، سیستم نمودار گام هر گوینده را بر اساس پیوستگی زمانی نمودارهای گام ردیابی می‌نماید. خروجی این مرحله دو دوره تناوب تخمین زده شده در هر فریم زمانی است، بدون توجه به این‌که گفتار مربوطه کجا صدادار، بی‌صدا و یا سکوت است.

در مرحله دوم سیستم نوع سیگنال گفتار را برای هر گوینده تعیین می‌کند. Weintraub در سیستم پیشنهادی خود هفت نوع سیگنال گفتار را مورد بررسی قرار می‌دهد: سکوت، متناوب، غیر متناوب، فراز، فرود، افزایش به سمت تناوب[۱] و کاهش یافتن تناوب[۲]. یک مدل مارکف برای مدل‌سازی رابطه ترتیبی انواع سیگنال گفتار، هنگامی‌که دو گوینده همزمان وجود داشته باشد، آموزش داده می‌شود. گام و دامنه سیگنال گفتار ویژگی‌هایی است که برای آموزش مدل مارکف به‌کار برده می‌شود. سیستم Weintraub بر پایه مدل مارکف، الگوریتم ویتربی را برای تعیین نوع سیگنال گوینده در هر فریم زمانی به‌کار می‌برد. خروجی این مرحله، باعث ایجاد توصیف دقیق‌تری از نوع سیگنال گفتار در هر فریم زمانی می‌گردد. در نتیجه دوره‌های گام تخمین زده شده برای فریم‌هایی که سیگنال گفتار در آن‌ها متناوب نیست، حذف می‌گردند.

[1] Increasing Periodicity
[2] Decreasing Periodicity

در مرحله سوم، با تخمین دامنه طیف مربوط به هر منبع صدا، با توجه به دوره گام و نوع گفتار برای هر گوینده، عمل جداسازی دو گوینده همزمان انجام می‌گیرد. در نهایت دامنه طیف تخمین زده شده برای بازسازی سیگنال گفتار هدف، مورد استفاده قرار می‌گیرد.

محاسبه دوره گام در سیستم Weintraub بر این اساس بود که دوره گام گوینده زن از دوره گام گوینده مرد بالاتر است. به‌همین دلیل در سیستم Weintraub امکان جداسازی دوگوینده همزمان مرد و یا زن وجود نداشت. به‌طور کلی این سیستم تنها برای جداسازی گفتار دو گوینده زن و مرد طراحی گردیده بود و قادر به جداسازی سیگنال گفتار از دیگر سیگنال‌های تداخلی (غیر از سیگنال گفتار) نبود.

در سال ۱۹۹۳، Cooke توانست اولین سیستم جداسازی سیگنال گفتار از انواع سیگنال‌های تداخلی را ارائه نماید [۳۴]. سیستم پیشنهادی او، ابتدا سیگنال مخلوط از فیلتربانک را عبور داده و سپس فرکانس لحظه‌ای مربوط به هر سیگنال باند باریک به‌دست آمده از هر کانال فیلتربانک را محاسبه می‌کند. کانال‌های مجاور با یکدیگر در هر فریم زمانی که دارای تغییرات آرام فرکانس لحظه‌ای در طول کانال‌ها هستند، در کنار یکدیگر گذاشته می‌شوند. این کانال‌های کنار هم گذاشته شده، شامل یک یا چند هارمونیک از یک منبع صدا هستند. کانال‌های کنار هم گذاشته در طول فرکانس، در فریم‌های زمانی مجاور نیز به یکدیگر متصل شده و بخش‌ها را به‌وجود می‌آورند. بخش‌هایی که با یکدیگر همپوشانی داشته باشند و دارای فرکانس گام یکسان باشند، تشکیل یک گروه می‌دهند. در نهایت نمودارهای گام برای هر گروه تخمین زده شده و گروه‌هایی که دارای نمودارهای گام یکسان باشند، در یک رشته قرار می‌گیرند که هر رشته مربوط به یک منبع صدا می‌باشد.

یک مجموعه داده‌های گفتاری توسط Cooke برای ارزیابی عملکرد سیستم جداسازی گفتار تک‌میکروفونه تهیه گردیده است. این مجموعه شامل ۱۰ صدای گفتار زن و مرد و ۱۰ نمونه سیگنال تداخلی است. با استفاده از سیگنال‌های این مجموعه می‌توان ۱۰۰ نمونه سیگنال مخلوط با نسبت

سیگنال به نویز[1] (SNR) مختلف تهیه نمود. سیگنال‌های تداخلی که تنوع بسیاری نیز دارند، عبارتند از: ۱- تک‌تن[2] با فرکانس 1kHz، ۲- نویز سفید، ۳- نویز ضربه‌ای[3]، ۴- نویز همهمه، ۵- موزیک rock، ۶- سوت کارخانه، ۷- زنگ تلفن، ۸- گفتار زن، ۹- گفتار مرد، ۱۰- گفتار زن. این مجموعه هم‌چنین برای ارزیابی دیگر سیستم‌های جداسازی گفتار نیز به‌کار رفته است [۳۳،۳۴،۳۵].

در ارزیابی عملکرد سیستم جداسازی گفتار Cooke، به‌طور میانگین ۷۰٪ از سیگنال گفتار جدا شده مربوط به گفتار هدف و در حدود ۱۰٪ از آن مربوط به سیگنال تداخلی است. سیستم Cooke، گفتار صدادار در فرکانس‌های پایین را بهتر از فرکانس‌های بالا جداسازی می‌نماید. در واقع این سیستم حتی در زمانی‌که سیگنال تداخلی نیز وجود نداشته باشد، قادر به بازسازی بسیاری از قسمت‌های هدف در فرکانس بالا نمی‌باشد.

در سال ۲۰۰۵، Brown و Cooke سیستم جداسازی تک‌میکروفونه قبلی را بهبود دادند [۳۵]. هدف از مدل پیشنهادی، تخمین یک ماسک زمان-فرکانس باینری برای گفتار هدف بود . گفتار هدف با به‌کارگیری ماسک و سنتز آن به‌وسیله روش سنتز Weitraub بازسازی می‌گردد [۳۳]. این سیستم پیشنهادی در بسیاری از جهات متفاوت از سیستم Cooke است. در این سیستم همبستگی نگاشت برای نمایش تناوب محاسبه می‌گردد. برای بخش‌بندی فضای زمان-فرکانس در این سیستم، کانال‌های همسایه بر اساس تابع همبستگی متقابل بین کانال‌ها با یکدیگر ادغام می‌شوند.

تابع همبستگی متقابل شباهت بین تابع خودهمبستگی کانال‌های مجاور را با یکدیگر مقایسه می‌کند. تابع خودهمبستگی متقابل بالاتر نشان‌دهنده شباهت بیشتر بین دو کانال مجاور است. یک نمونه از تابع همبستگی متقابل در شکل (۲-۸) نمودار سمت راست نشان داده شده است. همانطور که در شکل نشان داده شده، کانال‌های همسایه‌ای که دارای محتوای فرکانسی یکسان می‌باشند (هارمونیک سیگنال گفتار)، دارای تابع خودهمبستگی مشابه بوده و درنتیجه تابع همبستگی متقابل

[1] Signal-to-Noise Ratio
[2] Pure Tone
[3] Noise Bursts

آن‌ها زیاد می‌باشد. در نتیجه در این سیستم بر اساس تابع همبستگی متقابل، کانال‌هایی که دارای محتوای فرکانسی یکسان هستند، با یکدیگر ادغام می‌شوند. کانال‌های ادغام شده در فریم‌های زمانی پشت سر هم بر اساس پیوستگی زمانی به یکدیگر متصل‌شده و تشکیل بخش می‌دهند. برای گروه‌بندی، سیستم ابتدا نمودار گام هر بخش را با استفاده از تابع خودهمبستگی تخمین می‌زند. سپس بخش‌های دارای نمودار گام یکسان را در یک رشته، گروه‌بندی می‌کند. همچنین در این سیستم از ویژگی‌های فراز و فرود برای گروه‌بندی بخش‌ها استفاده می‌شود [۹،۳۵].

برای ارزیابی سیستم Brown و Cooke از مجموعه داده‌های Cooke استفاده شده و نتایجی شبیه نتایج سیستم Cooke به‌دست آمد. در این سیستم مقدار زیادی از انرژی سیگنال تداخلی از رشته به‌دست آمده برای سیگنال هدف حذف گردید؛ اما مقدار قابل توجهی از انرژی سیگنال هدف، به خصوص در فرکانس‌های بالا نیز در سیگنال هدف جدا شده، از بین رفته است. نتایج ارزیابی نشان می‌دهد که به‌کارگیری ویژگی‌های اضافی فراز و فرود برای گروه‌بندی نیز نتوانسته باعث بهبود قابل توجهی در عملکرد سیستم شود.

در سال ۱۹۹۱، Wang و Brown یک سیستم CASA شبیه سیستم Brown و Cooke پیشنهاد دادند. تفاوت اصلی میان این دو سیستم این است که در سیستم Wang و Brown از یک شبکه نوسان‌ساز دو لایه برای جداسازی گفتار استفاده شده است. در اولین لایه، بخش‌ها براساس تابع همبستگی متقابل و پیوستگی زمانی تشکیل می‌شوند. در هر فریم زمانی واحدهای زمان-فرکانس همسایه در صورتی در بخش‌ها با یکدیگر ادغام می‌شوند که تابع همبستگی متقابل آن‌ها از یک حد آستانه بالاتر باشد. به‌عبارت دیگر، دو واحد زمان-فرکانس در یک کانال و در دو فریم مجاور در صورتی با یکدیگر ادغام می‌شوند که تابع خود همبستگی متقابل میان آن دو واحد بالاتر از حد آستانه باشد. در لایه دوم شبکه Wang و Brown با استفاده از گام کلی تخمین زده شده در هر فریم زمانی، بخش‌ها در دو رشته، یکی برای سیگنال هدف و دیگری برای تداخل، گروه‌بندی می‌شوند. برای تخمین گام کلی، ابتدا سیستم بزرگ‌ترین بخش را به‌عنوان بخش پایه پیدا می‌کند. سپس سیستم

فریم‌های هر بخش را که تناوب گام آن‌ها با تناوب گام کلی بخش پایه یکی است، تعیین می‌نماید. اگر تناوب گام بیش از نیمی از فریم‌های بخش با بخش پایه یکی باشد، آن بخش با بخش پایه در یک رشته قرار داده می‌شود. در غیر این صورت، این بخش در رشته دیگر گذاشته می‌شود. با وجود این‌که سیستم Wang و Brown از لحاظ محاسباتی ساده‌تر از سیستم Brown و Cooke بود، اما دارای عملکردی شبیه به آن بود [۹].

علاوه بر سیستم‌های بالا، مطالعات زیادی در زمینه CASA برای جداسازی گفتارتک‌میکروفونه انجام گرفته است. در بسیاری از این سیستم‌ها نیز از ویژگی هارمونیکی گفتار برای‌جداسازی استفاده گردیده است. به‌عنوان‌مثال، Parsons سیستمی را برای جداسازی دو گوینده همزمان به‌وسیله نمودارهای گام دو گوینده از روی طیف سیگنال مخلوط پیشنهاد کرد که در آن هارمونیک‌های گفتار هدف با توجه به نمودارهای گام تخمین زده شده انتخاب می‌شوند [۳۶]. چندین روش برای جداسازی دو صدای همزمان، بر اساس تناوب سیگنال، با هدف مدل‌سازی رفتار سیستم شنوایی انسان ارائه شده است. علاوه بر تناوب، دیگر ویژگی‌های ASA نیز درسیستم‌های جداسازی گفتار به‌کار گرفته شده است. Abe و Ando سیستمی بر مبنای مدولاسیون فرکانس و دامنه برای جداسازی گفتار ارائه کردند. Masuda-Katsuse و Kawahara یک سیستم CASA پیشنهاد دادند که در آن برای منابع صوتی متفاوت، رشته‌هایی با توجه به ردیابی تغییرات در شکل طیف، تولید می‌نمایند.

برخی از پژوهش‌ها برای جداسازی گفتار دو گوینده همزمان، از مدل های وابسته بـه گوینـده‌ی از پیش آموزش داده شده[۱]، استفاده می‌کنند. در [۳۷]، یک روش بدون نظارت برای جداسازی گفتار دو گوینده پیشنهاد داده شده اسـت. روش موجـود دو مرحلـه اصـلی CASA (آنـالیز ترکیـب شـنیداری محاسباتی) را دنبال می‌کند: مرحله اول، بخش‌بندی[۲] نامیده شده که ترکیب شنیداری را به عناصر (یا بخش‌های) حسی تجزیه می‌کند؛ به‌طوری‌که هر یک از آن‌ها مربوط به یک منبع یکتـا باشـد. مرحلـه دوم، گروه‌بندی نامیده شده و در آن بخش‌های مربوط به یک منبع در یک گروه قرار می‌گیرند.

[1] Pretrained
[2] Segmentation

در [۳۷]، سیستم پیشنهادی به‌منظور جداسازی گفتار صدادار[۱]، از یک الگوریتم tandem [۳۸]، برای گروه‌بندی همزمان و سپس خوشه‌بندی[۲] بدون نظارت برای گروه‌بندی متوالی استفاده می‌کند. به منظور جداسازی گفتار بی‌صدا[۳]، در مرحله اول بخش‌های گفتار بی‌صدا مبتنی بر تجزیه و تحلیل فراز-فرود تولید می‌شوند. بخش‌های گروه‌بندی شده با استفاده از ماسک‌های باینری مکمل[۴] گفتار صدادار جدا می‌شوند. در واقع صداهای همزمان توسط ماسک باینری ایده آل[۵] (IBM) تخمین زده می‌شوند [۹]. یک ماسک IBM، با استفاده از یک SNR محلی[۶] به‌دست می‌آید. به‌طور خاص، به یک واحد زمان-فرکانس (T-F) که سیگنال هدف در آن غالب است، عدد یک و به غیر آن، عدد صفر نسبت داده می‌شود [۳۹]. ارزیابی اصولی و مقایسه‌ها نشان می‌دهند که این روش با وجود بدون نظارت بودن، عملکرد قابل توجهی در مقایسه با روش‌های مبتنی بر مدل و وابسته به گوینده داشته است.

همچنین در سال ۲۰۱۳، Wang و Hu علاوه بر مقاله [۳۷]، یک روش با نظارت را نیز برای جداسازی دو گوینده همزمان ارائه دادند. در این روش مشخصات دو گوینده داده شده است. در روش-های مبتنی بر مدل، عدم تطابق بین ترازهای آموزش[۷] و تست[۸] سیگنال وجود دارد. به‌همین منظور در [۴۰] توسط یک الگوریتم تکراری[۹] به‌منظور تطابق مدل‌های گفتار با ترازهای سیگنال در مرحله تست ارائه شده است. این الگوریتم در مرحله اول، تخمین‌های اولیه منبع سیگنال را با استفاده از مدل‌های گفتار HMM نامنطبق به‌دست می‌آورد، سپس SNR ورودی از صدای مخلوط را تشخیص می‌دهد. SNR ورودی به‌منظور تطابق مدل‌های گفتار برای تخمین درست استفاده شده است. از مرحله دوم به بعد که نسبت سیگنال به نویز به عدد خاصی همگرا می‌شود، این الگوریتم تکرار می‌شود. بعد از این

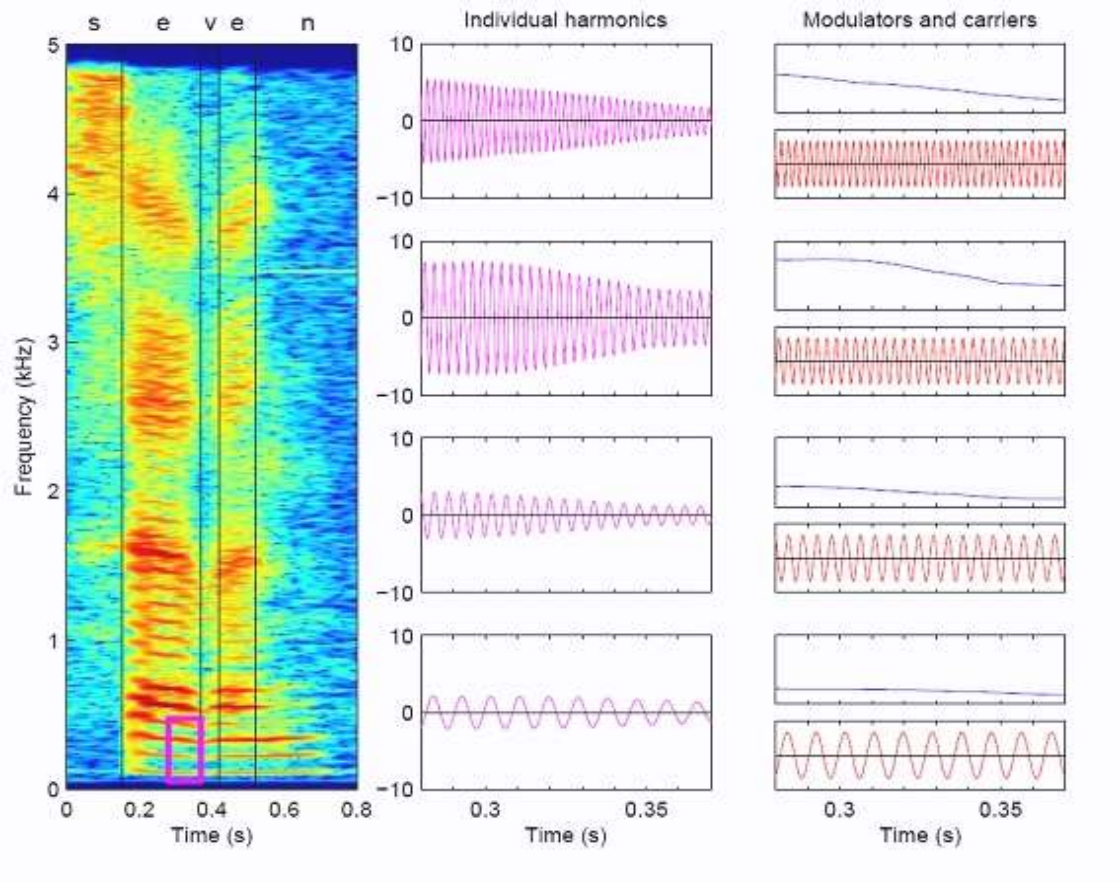
[1] Voiced Speech
[2] Clustering
[3] Unvoiced Speech
[4] Complementary Binary Masks
[5] Ideal Binary Mask
[6] Local SNR
[7] Training
[8] Test
[9] Iterative Algorithm

مرحله ماسک‌های باینری مناسب برای جداسازی دو گوینده همزمان به‌دست می‌آید. مقایسه‌ها نشـان می‌دهند روش پیشنهادی عملکرد خوبی در مقایسه با روش‌های مبتنی بر مدل داشته است.

۲-۵-۲- آنالیز مدولاسیون و فیلتر نمودن

سیگنال‌های طبیعی مانند گفتار و موزیک را می‌توان توسط سیگنال مدولاتور فرکانس پایین که سیگنال حامل فرکانس بالا را مدوله کرده است، نمایش داد [۴۱]. این مفهوم به‌وسیله اسپکتروگرام یک نمونه سیگنال گفتار که در شکل (۲-۹) نشان داده شده است، قابل توصیف می‌باشد. طیف نشان داده شده در شکل، توزیع انرژی سیگنال را در فضای زمان-فرکانس نشان می‌دهد. رنگ روشن نشان‌دهنده نقاطی با انرژی کم و رنگ تیره برای نقاطی با انرژی زیاد است.

همانطور که اسپکتروگرام سیگنال نشان می‌دهد، بیشتر انرژی سیگنال در خطوط افقی متمرکز شده‌اند که دارای فاصله مساوی در طول محور فرکانس است. هر کدام از این خطوط افقی مربوط به یکی از هارمونیک‌های سیگنال گفتار می‌باشد. سیگنال‌های جداگانه چهار هارمونیک اول در یک فریم زمانی کوتاه (به‌وسیله یک مستطیل کوچک در پایین اسپکتروگرام نشان داده شده است) در پانل وسطی شکل (۲-۹) نشان داده شده است. با توجه به سیگنال‌های نشان داده شده برای هر هارمونیک، می‌توان فهمید که هر سیگنال از حاصل ضرب دو سیگنال حامل و پوش تشکیل شده است. حامل، یک سیگنال فرکانس بالا است که شامل ساختار زمانی سیگنال هارمونیک است و پوش، سیگنال فرکانس پایینی است که تغییرات آهسته آهسته سیگنال دامنه سیگنال هارمونیک را دنبال می‌کند. تجزیه هر سیگنال هارمونیک، به سیگنال حامل و پوش (یا سیگنال مدولاتور) در پانل سمت راست شکل (۲-۹) نشان داده شده است.



شکل (۲-۹) نمایش اسپکتروگرام سیگنال گفتار و چهار هارمونیک اول آن که به سیگنال‌های حامل و مدولاتور تجزیه شده است. پانل سمت چپ اسپکتروگرام سیگنال گفتار یک گوینده مرد برای بیان کلمه "seven" را نشان می‌دهد. پانل وسطی سیگنال چهار هارمونیک اول و پانل سمت راست سیگنال مدولاتور و حامل حاصل از تجزیه هر سیگنال هارمونیک را نشان می‌دهد [۴۱].

بسیاری از مطالعات نشان داده است که سیگنال مدولاتور در درک یک سیگنال گفتار بسیار با اهمیت است. به‌عنوان مثال، هنگامی که یک سیگنال مدولاتور، با یک مقدار ثابت جایگزین شود، گفتار غیر قابل درک می‌شود. در مقابل، اگر سیگنال حامل با یک نویز سفید جایگزین شود اما سیگنال مدولاتور دست نخورده باقی بماند، گفتار همچنان قابل درک است. اهمیت سیگنال مدولاتور در درک سیگنال گفتار سبب شد تا این سیگنال به‌عنوان یک ویژگی برای جداسازی منابع صدا، در شرایطی که تنها یک میکروفون در دسترس است، مورد استفاده قرار گیرد.

برای آنالیز و پردازش سیگنال مدولاتور، ابتدا سیگنال باند وسیع به‌وسیله فیلتربانک به سیگنال‌های باند باریک تجزیه می‌گردد. سپس هر سیگنال باند باریک به سیگنال‌های مدولاتور و حامل تجزیه

می‌شود. در نهایت با توجه به سیگنال مدولاتور به‌دست آمده، روش‌های استخراج ویژگی متفاوتی برای جداسازی منابع صوتی از یکدیگر مورد استفاده قرار می‌گیرد.

روش‌های موجود برای تخمین سیگنال مدولاتور به دو دسته طبقه‌بندی می‌شوند: ناهمدوس و همدوس. در روش ناهمدوس تخمین سیگنال مدولاتور یا پوش یک سیگنال باند باریک حقیقی با استفاده از تبدیل پوش هیلبرت[1] و برای سیگنال باند باریک با مقدار مختلط، به‌وسیله عملگر تخمین دامنه[2] انجام می‌گیرد. در روش همدوس ابتدا سیگنال حامل با استفاده از یک تخمین‌گر حامل[3]، بر اساس خواص فرکانس لحظه‌ای[4] (IF) سیگنال تخمین زده شده و سپس با توجه به سیگنال حامل بدست آمده سیگنال مدولاتورتخمین زده می‌شود.

Schimmel الگوریتمی بر مبنای روش ناهمدوس برای جداسازی دو گوینده همزمان، درشرایطی‌که تنها یک میکروفون در دسترس است، پیشنهاد کرد [۴۱]. در این الگوریتم با فرض معلوم بودن محدوده گام هر گوینده و ثابت بودن آن در همه کانال‌های فرکانسی، یک ماسک برای جداسازی دو گوینده همزمان در فضای طیف مدولاسیون ارائه شده است. اما همانطور که می‌دانیم در شرایط واقعی و برای گوینده‌های متفاوت و ناشناس، این محدوده گام معین نمی‌باشد و باید توسط روش‌های موجود به‌منظور تخمین فرکانس گام دو گوینده همزمان، تخمین زده شود.

Won و دیگر همکارانش الگوریتمی بر اساس روش همدوس برای بهبود سیگنال نویزی در سمعک و کسانی که مشکل شنوایی دارند، ارائه کرده‌اند. در این روش، با به‌کارگیری فیلتر وفقی بر روی سیگنال مدولاتور به‌دست آمده از روش همدوس و با فرض دانستن سیگنال کانال مرجع (سیگنال نویز)، سیگنال نویزی بهبود می‌یابد؛ اما در شرایطی‌که تنها یک میکروفون و یک صدای ضبط شده در دسترس باشد، امکان دسترسی به سیگنال نویز وجود ندارد. در نتیجه الگوریتم پیشنهادی تنها در شرایطی قادر به بهسازی سیگنال نویزی است که سیگنال کانال مرجع در دسترس باشد [۴۲].

---

[1] Hilbert Envelope
[2] Magnitude Operator
[3] Carrier Estimator
[4] Instantaneous Frequency

در [۴۳] یک سیستم جداسازی تک‌میکروفونه بر پایه روش دمدولاسیون ناهمدوس ارائه شـده کـه در آن ابتدا محدوده فرکانس گام یک، یا دو گوینده همزمان، تخمین زده می‌شود و سپس با اسـتفاده از آن، گفتار هدف از تداخل جدا می‌شـود. محدوده گام با استفاده از الگوریتم فراز و فرود و با توجه به توزیع انرژی گوینده‌ها در حوزه طیف مدولاسیون، تخمین زده شده است. زمانی‌که گوینده‌های هدف و تداخل هر دو مرد و یا هر دو زن باشند، روش‌های تخمین فرکانس گام به دلیل نزدیکی فرکانس گام دو گوینده با خطای زیادی مواجه می‌شوند. در نتیجه سیستم‌های CASA که فرکانس گام را به‌عنوان یک ویژگی مهم در جداسازی گفتار به‌کارمی‌برند، دچار مشکل می‌شوند. در مقابل یکی از نوآوری‌های مهم سیستم پیشنهادی، تخمین محدوده فرکانس گام دو گوینده همزمان، در فریم‌های زمانی کوچک اسـت. در این سیستم همچنین ماسک تولید شده برای جداسازی، وابسته به محدوده گام تخمین زده شده در هر کانال فرکانسی مـــی‌باشـــد. با توجه به نتایج بـــه‌دست آمده، سیستم پیشنهادی قادر به جداسازی قسمت عمده قسمت‌های صدادار سیگنال گفتار هدف از سیگنال تداخلی است. عـلاوه‌براین، قسمت‌هایی از گفتار بی‌صـــدا که به‌علت نزدیکی به قسمت‌های صـــدادار شبه پریودیک می‌باشند نیز توسط این سیستم جداشده است.

در واقع نتایج موجود در [۴۳] نشان می‌دهد که سیستم پیشنهادی در مقابل تداخل مقاوم می‌باشد و در شرایطی‌که انرژی سیگنال تداخلی زیاد باشد نیز قادر به تخمین خوبی از محدوده فرکانس گام و سیگنال گفتار صدادار است. همچنین نتایج ارزیابی نشان می‌دهد که عملکرد سیستم پیشنهادی در جداسازی گفتار در شرایط تک‌میکروفونه در مقایسه با روش CASA بیان شده، بهتر عمل می‌کند.

۲-۶- تخمین فرکانس گام

یکی از مسائل اولیه در آنالیز گفتار، تعیین فرکانس گام است. به‌عبارت دیگر، فرکانس گام یکی از ویژگی‌های کلیدی برای سیستم‌های جداسازی گفتار بر پایه گام است. الگوریتم تعیین PDA[1] مستقیماً از مدل‌های ادراک گام الهام گرفته است. در این مدل‌ها فرض می‌شود که گام به‌عنوان تناوب

[1] Pitch Detemination Algorithm

الگوهای عصبی در حوزه زمان قابل مشاهده است [۴۴] و یا این‌که درحوزه فرکانس اجزا هارمونیک توسط حلزونی گوش[1] قابل تفکیک است.

مکانیسم‌های زمانی سرنخ‌هایی را برای ردیابی تناوب‌ها در حوزه زمان فراهم می‌کنند. از جمله این مکانیسم‌ها می‌توان به محل قطع صفر[2][۴۵]، فیلتر پایین‌گذر [۲۹]، تابع‌خودهمبستگی [۳۲]، توابع تفاضلی و ترکیبات آن‌ها اشاره نمود. از طرف دیگر، مکانیسم‌های طیفی برای درک گام، آنالیز مؤلفه‌های فرکانسی جداگانه را پیشنهاد می‌کنند که از آن‌ها برای تخمین بهترین جاسازی[3] "الگوی گام[4]" استفاده می‌شود. روش‌های مختلفی بر پایه تطبیق الگوها مانند: شانه طیفی[5]، الک طیفی[6] و مجموع زیر هارمونیک‌ها پیشنهاد گردیده است. با ترکیب دو دسته روش بالا جهت درک گام، یک دسته روش به نام روش‌های زمان-طیفی ایجاد می‌شود که از تئوری Licklider [۴۶] منشأ گرفته شده و توسط دیگران گسترش یافته است.

در کاربردهای واقعی، یک ردیاب چند گام[7]، برای زمانی که صدای تداخلی شامل یک ساختار هارمونیکی (مانند موزیک یا گفتار دیگر) باشد، مورد نیاز است. این عمل به دلیل همپوشانی هارمونیک منابع که باعث ضعیف شدن نشانه‌های گام منابع می‌شود، بسیار مشکل‌تر از تخمین یک گام است. تاکنون چندین روش برای ردیابی دو گام همزمان ارائه شده است. Toloen و Karjalainen یک تحلیل‌کننده چند گام را با به‌کارگیری دو باند وابسته به محدوده فرکانس بالا و پایین و نیز با استفاده از تابع خودهمبستگی مجموع بهبودیافته، طراحی کرده‌اند [۴۷]. Wu و دیگر همکارانش آماره‌های دوره تناوب گام را بر روی یک مکانیسم انتخاب کانال مدل کردند و از یک HMM برای استخراج منحنی‌های گام پیوسته استفاده نمودند. Jordan یک مدل بر اساس مدل آماری طیف سیگنال با استفاده از یک HMM برای رفتار گام ارائه کرده است [۴۸].

---

[1] Cochleagram
[2] Zero-crossing
[3] Best-Fitting
[4] Pitch template
[5] Spectral Comb
[6] Harmonic Sieve
[7] Multipitch

در [۴۹] طیف توان سیگنال مخلوط، به‌صورت مدل‌های منابع پارامتری که با استفاده از قسمت‌های صدادار گفتار آموزش داده شده‌اند، مدل شده است. Klapuri در واقع یک مدل تخمین و حذف پیشنهاد داده است که با توجه به یک الگوریتم تکراری، نقاط گام را درموزیک دارای چندین صدا و سیگنال گفتار ردیابی می‌کند. Hu و Wang در الگوریتمی برای تخمین گام و جداسازی گفتار به‌صورت همزمان و تکراری ارائه کردند [۴۰].

۲-۷- جمع‌بندی و نتیجه‌گیری

در دنیای واقعی ترکیب گفتار و دیگر صوت‌ها توسط یک یا چند میکروفون، برای انجام پردازش گردآوری می‌شود. در محیط‌هایی نظیر مهمانی‌ها، استفاده از روش‌هایی مانند آنالیز مولفه‌های مستقل (ICA) یا فیلترینگ فضایی که در آن‌ها آرایه‌ای از میکروفون‌ها برای جداسازی گفتار مورد نظر استفاده می‌شود، نتایج مناسبی ندارند. همچنین در برخی محیط‌ها استفاده از چندین میکروفون امکان‌پذیر نبوده و تنها انتخاب، حالت تک‌میکروفونه خواهد بود.

در این فصل روش‌های موجود برای جداسازی همزمان گفتار تشریح شد. معایب و مزایای روش‌ها گفته و مقایسه شدند. در واقع یکی از چالش‌های مهم در بهسازی گفتار، جداسازی سیگنال گفتار هدف از سیگنال تداخلی شبیه به گفتار است. همانطور که در این فصل به آن اشاره شد، دقت روش‌های CASA در جداسازی گفتار تک‌میکروفونه، وابسته به دقت تخمین فرکانس گام دو گوینده همزمان است. این بدین دلیل است که در این روش‌ها، یک ماسک مناسب در حوزه T-F برای جداسازی گفتار با توجه به فرکانس گام تخمین زده شده، تولید می‌شود.

فصل سوم

تئوری الگوریتم های مورد استفاده

۳-۱- روش‌های استخراج ویژگی

۳-۱-۱- روش‌های حوزه زمان

روش‌های پردازش سیگنال حوزه زمان تنها از داده‌های زمانی به‌منظور تحلیل و آنالیز بهره می‌گیرند و هیچ‌گونه تبدیل دیگری بر داده‌ها اعمال نمی‌شود. در این روش‌ها عمدتاً از ویژگی‌های آماری سیگنال زمانی استفاده می‌شود.

۳-۱-۲- روش‌های حوزه فرکانس

در هنگام مطالعه رفتار یک سیگنال در حوزه زمان، تنها اطلاعات محدودی از آن به‌دست می‌آید و بسیاری از مشخصات مهم آن پنهان می‌ماند. بدین منظور سیگنال از فضایی به فضای دیگر انتقال داده می‌شود تا بعضی از اطلاعات مفید آن در حوزه جدید آشکار گردد. ابزار انتقال از حوزه زمان به فرکانس، تبدیل فوریه پیوسته و گسسته است. تبدیل فوریه که به‌صورت زیر تعریف می‌گردد، تابع $x(t)$ در حوزه زمان را به یک تابع فرکانسی $X(\omega)$ تبدیل می‌نماید.

$$\mathrm{X}(\omega) = \int_{-x}^{x} x(t) e^{-j\omega \mathrm{t}}\, dt \qquad (۱-۳)$$

تبدیل فوریه گسسته به‌صورت زیر تعریف می گردد.

$$\mathrm{X}(\mathrm{K}) = \sum X(n) e^{\frac{-2\pi j n}{N} K} \qquad (۲-۳)$$

به‌منظور افزایش سرعت عملیاتی، تبدیل فوریه سریع معرفی گردید. در واقع می‌توان گفت که در حال حاضر مهم‌ترین ابزار در پردازش و تحلیل سیگنال‌ها، این روش می‌باشد. تبدیل فوریه معمولاً برای پنجره‌ای کوچک از سیگنال اعمال می‌گردد تا سیگنال پایا باشد.

پس از محاسبه تبدیل فوریه یک سیگنال گسسته، بردارهای مختلفی به‌دست می‌آیند که شامل اطلاعات دامنه و فاز تبدیل فوریه می‌باشند.

۳-۱-۳- روش‌های حوزه زمان- فرکانس

پس از بیان روش‌های استخراج ویژگی در دو حوزه زمان و فرکانس، به روش‌هایی می‌پردازیم که اطلاعات دو حوزه را با هم در نظر می‌گیرند. یک مشکل اساسی در همه‌ی روش‌های حوزه فرکانس این است که سیگنال مورد بررسی را ایستا در نظر گرفته و محتوای فرکانسی را نسبت به زمان، ثابت می‌گیرند. این در حالی است که در اکثر کاربردهای عملی، سیگنال‌های صوتی غیر ایستا و محتوای فرکانسی آن‌ها، تابع زمان می‌باشند. بنابراین روش‌های جدیدی برای برطرف کردن این کاستی ارائه شده است که مهم‌ترین آن‌ها تبدیل فوریه زمان کوتاه و تبدیل موجک می باشند. در تبدیل فوریه پیوسته و گسسته بخشی از سیگنال در نظر گرفته می‌شود و محتوای فرکانسی آن به همه‌ی سیگنال تعمیم داده می‌شود. در محاسبه طیف توان متوسط نیز پنجره‌ای با طول مشخص برای سیگنال در نظر گرفته می‌شود و طیف فرکانسی برای آن محاسبه می‌گردد. سپس این پنجره در طول سیگنال حرکت داده می‌شود و پس از محاسبه طیف برای همه پنجره‌ها، مقادیر به‌دست آمده میانگین‌گیری شده و مقدار متوسط آن‌ها به‌عنوان طیف فرکانسی سیگنال در نظر گرفته می‌شود. در این روش قطعه‌ای از سیگنال که اقتباس می‌گردد، همیشه دارای طول یکسانی است. بنابراین تکه‌ی زمانی به‌کار رفته برای تحلیل یک جزء با فرکانس بالا، دارای همان طول زمانی به‌کار رفته، برای نوع با فرکانس پایین می‌باشد. در هر دو حالت ذکر شده، اطلاعات زمانی سیگنال از بین می‌رود و تخصیص یافتگی مناسبی در حوزه زمان وجود ندارد. برای رفع این مشکل راه‌حل اولیه‌ای پیشنهاد شده است که در نهایت منجر به ابداع تبدیل موجک گشته است. ساده‌ترین کار ممکن این است که هر بار که پنجره با طول ثابت روی سیگنال باز شده و طیف فرکانسی آن محاسبه می‌گردد، زمانی متناسب با آن پنجره به طیف به‌دست آمده اختصاص داده شود. با این کار در حین تغییر موقعیت پنجره، می‌توانیم تغییرات طیف فرکانسی سیگنال را نیز برحسب زمان داشته باشیم، چنین تبدیلی تبدیل فوریه زمان کوتاه نامیده می‌شود.

۳-۱-۴- آنالیز پیش‌گویی خطی[1] (LPC)

یکی از مشهورترین تکنیک‌های آنالیز و سنتز گفتار، پیش‌گویی خطی است که به‌دلیل توانایی آن در نمایش شکل موج بر حسب پارامترهای متغیر با زمان، در بسیاری از کاربردهای کدگذاری گفتار به‌کار می‌رود. تکنیک‌های پیش‌گویی خطی از نظر محاسباتی بسیار سریع هستند و در تخمین پارامترهای سیگنال گفتار همانند دوره‌ی اصلی تناوب و فورمنت‌ها، ابزار ضروری محسوب می‌شوند. ایده‌ی اصلی، مبتنی بر این نظریه است که می‌توان طیف سیگنال گفتار را با پاسخ فرکانسی یک فیلتر تمام قطب، تقریب زد و ضرایب این فیلتر را به‌عنوان بردار مشخصه‌ی آن مورد استفاده قرار داد. در آنالیز پیش‌گویی خطی، یک نمونه از سیگنال گفتار در یک زمان مشخص، با ترکیب خطی از تعدادی از نمونه‌های قبلی آن، تخمین زده می‌شود. اگر $S(n)$ نمونه‌ای در زمان n باشد، می‌توان آن را با P نمونه‌ی قبلی آن به‌صورت زیر تقریب زد:

$$S(n) \cong \ a_1 S(n-1) + a_2 S(n-2) + \cdots + a_p S(n-p) \qquad (۳-۳)$$

اگر ضرایب a_1 تا a_p در طول سیگنال گفتار ثابت فرض شوند، معادله‌ی (۳-۴) را می‌توان به‌صورت زیر نوشت:

$$S(z) = \sum_{i=1}^{p} a_i \, z^{-i} S(z) + GU(z) \qquad (۳-۴)$$

$$V(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i \, z^{-i}} = \frac{G}{1 - P(z)} \qquad (۳-۵)$$

به‌طوری‌که $V(z)$ یک فیلتر تمام قطب است که درجه‌ی آن P می‌باشد و مدل تقریباً خوبی برای مجرای گفتار[2] و محفظه‌ی دهان می‌باشد. هدف در آنالیز پیش‌گویی خطی، به‌دست آوردن ضرایب این فیلتر یعنی a_1 تا a_p است. هر قدر که مرتبه‌ی مدل (P) بیشتر باشد، شکل دقیق‌تری از طیف سیگنال گفتار به‌دست می‌آید و مشخصه‌های بیشتر و دقیق‌تری توسط ضرایب a_1 تا a_p نشان داده

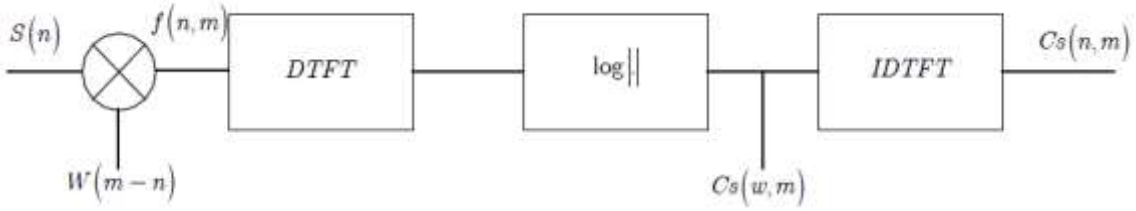
[1] Linear Predictive Analysis
[2] Vocal Tract

می‌شوند؛ البته زمان و پیچیدگی محاسبات نیز به همان اندازه بیشتر خواهد شد. مقدار P را اغلب ۸ (دو فورمنت اول)، ۱۲ (سه فورمنت اول) و یا ۱۶ (چهار فورمنت اول) در نظر می‌گیرند. ضرایب پیش-گویی خطی برای هر قاب پنجره شده، محاسبه می‌شود و به‌عنوان ویژگی، مورد استفاده قرار می‌گیرد.

۳-۱-۵- آنالیز کپسترال

یکی از مشخصه‌هایی که از سیگنال گفتار استخراج می‌شود و در بسیاری از کاربردها مورد استفاده قرار می‌گیرد، ضرایب کپسترال می‌باشد. این ضرایب نه تنها اطلاعات فیلتر مجرای گفتار را در خود دارند، بلکه حاوی اطلاعات سیگنال تحریک نیز می‌باشند. بنابراین مشخصه‌های مناسبی را برای تشخیص گفتار در اختیار قرار می‌دهند.

آنالیز کپسترال[۱] به دو دسته‌ی کپستروم حقیقی[۲] و کپستروم مختلط[۳] تقسیم می‌شود. تفاوت کپستروم حقیقی و مختلط این است که آنالیز کپستروم مختلط، حاوی اطلاعات فاز سیگنال صحبت نیز هست؛ ولی از آن‌جا که اطلاعات فاز در شنوایی انسان اهمیت کمی دارد، معمولاً از آنالیز کپستروم حقیقی استفاده می‌شود. شکل (۳-۱) بیانگر چگونگی محاسبه‌ی ضرایب کپسترال است.



شکل (۳-۱) نحوه‌ی محاسبه‌ی ضرایب کپسترال [۵۰]

ضرایب کپسترال با استفاده از ضرایب پیش‌گویی خطی نیز می‌توانند محاسبه شوند. اگر ضرایب پیش‌گویی خطی، ضرایب فیلتر رابطه‌ی (۳-۵) و از مرتبه‌ی P باشند، ضرایب LPCC به صورت رابطه‌ی (۳-۶) محاسبه می‌شوند:

[1] Cepstral
[2] Real Cepstrum
[3] Complex Cepstrum

$$C(n) = \begin{cases} a(n) + \sum_{k=1}^{n-1} \frac{n-k}{n} a_k C_{n-k} & 1 \le n \le p \\ \sum_{k=1}^{p} \frac{n-k}{n} a_k C_{n-k} & p < n \le \infty \end{cases} \quad (۶-۳)$$

که در رابطه a(n) ضرایب LPC و C(n) ضرایب کپسترال می‌باشند.

۳-۱-۶- استفاده از مقیاس MEL در آنالیز کپسترال

یکی از راه‌های بهبود دقت تشخیص سیستم، هنگامی‌که از ضرایب کپسترال به‌عنوان بردارهای ویژگی استفاده می‌شود، استفاده از یک تبدیل غیرخطی است که حساسیت گوش را نسبت به حوزه‌های مختلف فرکانس مدل کند. مسئله‌ی قابل توجه این است که باندهای مختلف فرکانس از نظر اطلاعات شنوایی، دارای ارزش‌های مختلفی هستند. خاصیت مقیاس مل[1] این است که به اطلاعات حوزه‌ی پایین فرکانس ارزش بیشتری می‌دهد.

مل واحد ارزیابی صدای درک شده یا فرکانس آهنگ[2] است. مقیاس MEL با رابطه‌ی زیر چنین تعریف می‌شود:

$$F_{mel} = 1000 log_2(1 + F_{KHZ}) \quad (۷-۳)$$

ثابت شده است که درک یک فرکانس ویژه، توسط سیستم شنوایی، تحت تاثیر انرژی یک باند بحرانی از فرکانس‌های پیرامون آن قرار دارد. پهنای باند بحرانی برای فرکانس‌های کمتر از یک کیلوهرتز برابر با ۱۰۰ هرتز است و برای فرکانس‌های بیش از یک کیلوهرتز، به صورت لگاریتمی افزایش می‌یابد. بنابراین برای محاسبه‌ی ضرایب مل کپستروم می‌توان لگاریتم انرژی کل در باندهای بحرانی حول فرکانس‌های مل را به عنوان ورودی تبدیل فوریه گسسته معکوس به کار گرفت و با محاسبه‌ی این تبدیل ضرایب مل کپستروم را محاسبه نمود. ضرایب مل کپستروم به‌صورت معادله (۳-۸) تعریف می‌شود.

[1] Mel
[2] Tone

$$C_s(n,m) = \frac{1}{N'} \sum_{k=0}^{N'-1} \bar{Y}(k) \cos(k \frac{2\pi}{N'} n) \quad (۸-۳)$$

ضرایب مل کپستروم به اختصار MFCC نامیده می‌شوند.

۳-۲- آموزش دیکشنری

بیشتر سیستم‌های CASA وابسته به ویژگی‌های مبتنی بر فرکانس گام می‌باشند. فرکانس گام برای دو گوینده زن و مرد در قسمت‌های زیادی همپوشانی دارد. بنابراین دنبال کردن چندین گام در یک سیگنال صدادار، کار بسیار مشکلی می‌باشد. در حالتی‌که دو گوینده هم‌جنس (زن-زن و یا مرد-مرد) باشد، به‌مراتب کار سخت‌تر خواهد شد. بنابراین به‌جای استفاده از گام باید به دنبال ویژگی بهتری بود تا بتوان مولفه‌های فرکانسی هر کدام از گوینده‌ها را با دقت بیشتری جداسازی نمود.

در این پایان‌نامه، یک روش جداسازی تک‌میکروفونه مبتنی بر ویژگی‌های تُنّک ارائه خواهیم داد. روش پیشنهادی برای هر کدام از گوینده‌ها یک دیکشنری از ویژگی‌های مناسب آموزش می‌دهد. سپس با استفاده از دیکشنری‌های آموزش دیده، صدای هر کدام از دو گوینده همزمان، بازسازی می‌شود. همچنین باید متذکر شد که روش پیشنهادی در این پایان‌نامه تلفیقی از روش یادگیری آماری و روش مبتنی بر ویژگی به‌منظور جداسازی تک‌میکروفونه می‌باشد.

دیکشنری از مجموعه‌ای از بردارهای پایه (اتم) تشکیل شده است. هر ستون ماتریس دیکشنری یک بردار پایه می باشد. ایده‌ی این روش بر این اساس استوار است که می‌توان یک بردار را به‌صورت ترکیب خطی از بردارهای دیگر بیان کرد. برای مثال اگر بردارP به‌صورت ترکیب خطی از اتم‌ها بیان شود و n تعداد اتم‌ها در دیکشنری باشد، خواهیم داشت:

$$P = \propto_1 p_1 + \propto_2 p_2 + \propto_3 p_3 + \cdots = \sum_{j=1}^{n} \propto_j p_j \quad (۹-۳)$$

$$P = Dq \quad (۱۰-۳)$$

به‌طوری‌که p_iها بردارهای پایه می باشند و $q = (\propto_1 \propto_2 \ldots + \propto_n)^T$ بردار ضرایب است که اهمیت هر اتم را برای ساخت بردار P، تعیین می‌کند. $D = (p_1\ p_2\ \ldots + p_n)$ دیکشنری مورد

استفاده برای به‌دست آوردن ارائه‌ی تنک بردار P می‌باشد. بردار بازسازی شده باید تا حد ممکن به بردار مورد نظر نزدیک باشد. بردار ضرایب q به‌وسیله‌ی حل معادله‌ی بهینه‌سازی زیر به‌دست می‌آید:

$$\mathrm{q} = \mathrm{argmin}_q \parallel Dq - \mathrm{P} \parallel + \gamma \|\alpha\|_0 \qquad (۱۱-۳)$$

$$\mathrm{D} = \mathrm{argmin}_D \parallel Dq - \mathrm{P} \parallel + \gamma \|\alpha\|_0 \qquad (۱۲-۳)$$

آموزش دیکشنری یک مرحله مهم در روش پیشنهادی می‌باشد. یک دیکشنری مناسب می‌تواند عملکرد سیستم را بالا ببرد و بالعکس. یکی از الگوریتم‌های ارائه شده برای آموزش دیکشنری، الگوریتم K-SVD می‌باشد. روش K-SVD از تجزیه‌ی مقادیر منفرد برای آموزش دیکشنری استفاده می‌نماید. این روش حالت توسعه یافته‌ای از الگوریتم K-means می‌باشد. این الگوریتم با مینیمم کردن معادله‌ی (۳-۱۱) نسبت به q ابتدا مقادیر کدگذاری تُنک (q) برای دیکشنری اولیه را محاسبه می‌نماید، سپس با ثابت در نظر گرفتن کدگذاری تُنک به‌دست آمده، با استفاده از معادله (۳-۱۲)، سعی می‌کند اتم‌های موجود در دیکشنری را برای تطبیق بهتر، به‌روزرسانی نماید. دیکشنری اولیه معمولاً از مقادیر تصادفی به‌دست می‌آید. این الگوریتم تکراری آنقدر ادامه پیدا می‌کند تا خطای بازسازی تمام نمونه‌های آموزشی با دیکشنری به‌دست آمده مینیمم شود [۵۱].

۳-۲-۱- کدگذاری تُنک یک سیگنال صوتی

کد گذاری تُنک ارائه‌ی سیگنال مورد نظر با کمترین تعداد ممکن، از اتم‌های یک دیکشنری می‌باشد. در ارائه‌ی فرکانسی-زمانی سیگنال X (تبدیل فوریه زمان کوتاه)، هر ستون بیانگر ویژگی استخراج شده از یک پنجره‌ی زمان کوتاه از سیگنال ورودی است. مشاهده‌ی X_i (یک ستون از ماتریس تبدیل فوریه زمان کوتاه) می‌تواند با استفاده از دیکشنری D به‌صورت تُنک رمزگذاری شود. بنابراین:

$$X_i = D\alpha \qquad (۱۳-۳)$$

به‌طوری‌که D یک ماتریس m×p (m>p)، X_i یک ماتریس m×1 و α یک ماتریس p×1 می‌باشد. بنابراین در این مسئله، تخمین α را طوری می‌یابیم که شرط تُنک بودن α همواره برقرار باشد. شرط

تُنک بودن باعث می‌شود که فقط تعداد کمی از عناصر موجود در α غیر صفر باشند و بقیه عناصر باید صفر باشند. به هر کدام از ستون‌های ماتریس D، یک اتم گفته می‌شود. در مسئله‌ی کدگذاری تُنک، هر سیگنال X_i به‌صورت ترکیب خطی تعداد کمی از اتم‌های موجود در دیکشنری D نوشته می‌شود. بنابراین مسئله‌ی تخمین تُنک نمونه X_i به‌صورت زیر قابل بیان می‌باشد:

$$min_{\alpha \in R^{p\times 1}} \|\alpha\|_0 \qquad such\ that\ \ X_i = D\alpha \tag{۱۴-۳}$$

نرم l_0 یک شبه نرم[1] می‌باشد که عناصر غیر صفر α را می‌شمارد. حل این مسئله بسیار مشکل می‌باشد. الگوریتم‌های زیادی برای حل این مسئله پیشنهاد شده است. در این پایان‌نامه از OMP[2] برای حل این مسئله استفاده شده است[۵۲].

۳-۲-۲- الگوریتم OMP

هدف اصلی الگوریتم OMP به‌دست آوردن تخمینی از مسئله‌ی کدگذاری تُنک به‌وسیله‌ی قید تُنک بودن می‌باشد. مسئله‌ی کدگذاری تُنک به‌وسیله‌ی معادله‌ی زیر بیان می‌شود:

$$\alpha = argmin_\alpha \|X_i - D\alpha\|_2^2 \qquad subject\ to\ \|\alpha\|_0 < k \tag{۱۵-۳}$$

به‌طوری‌که k بیشترین تعداد اتم‌های شرکت کننده در فرایند بازسازی سیگنال X_i را دارا می باشد. در الگوریتم OMP اتمی که بیشترین همبستگی را با باقی‌مانده هر مرحله داشته باشد، به‌عنوان یکی از اتم‌های شرکت کننده در بازسازی سیگنال در نظر گرفته می‌شود. بعد از چندین تکرار، این الگوریتم سیگنال ورودی را بر حسب ترکیب خطی اتم‌های موجود در دیکشنری D، بیان می‌کند. الگوریتم OMP در زیر بیان شده است [۵۲].

در این الگوریتم، متغیر I اندیس اتم‌های انتخاب شده در هر مرحله را نگهداری می‌کند، r مقدار باقی‌مانده در هر مرحله است و α خروجی ارائه تنک می‌باشد. در این الگوریتم ابتدا اتمی که بیشترین همبستگی را با باقی‌مانده هر مرحله داشته باشد، انتخاب می‌شود (مرحله ۵). سپس شماره‌ی اتم

[1] pseudo-norm
[2] Orthognal matching pursuit

انتخاب شده در مرحله قبل، به متغیر I اضافه می‌شود (مرحله‌ی ۶). درمرحله‌ی ۷ اتم‌های انتخاب شده به فضای سیگنال X_i برده می‌شوند و در مرحله‌ی ۸ مقدار باقی‌مانده به‌روز می‌شود [۵۳].

الگوریتم OMP

۱- ورودی: دیکشنری D، سیگنال ورودی X_i و معیار تُنک بودن k

۲- خروجی: ارائه‌ی تُنک α به‌صورتی‌که $X_i=D\alpha$

۳- مقداردهی اولیه: $I=[]$ ، $r=X_i$ ، $\alpha=0$

۴- تا زمانی‌که شرط تُنک بودن $\|\alpha\|_0 < k$ محقق نشده است مراحل زیر انجام شود:

۵- $\hat{u} = argmax_k |d_k^T r|$

۶- $I=[I; \hat{u}]$

۷- $\alpha_I = (D_I)^+ X_i$

۸- $r = X_i - D_I \alpha_I$

۹- پایان

۳-۲- جمع‌بندی و نتیجه‌گیری

در این فصل به تشریح استخراج ویژگی‌های مورد نظر برای یک سیگنال صوتی پرداختیم. در واقع یک سیگنال صوتی دارای ویژگی‌های گسترده‌ای است که نمی‌توان همه‌ی آن‌ها را استخراج کرد. به‌همین منظور روی سیگنال صوتی تبدیلاتی انجام می‌دهیم تا بتوانیم از این طریق روی قسمت‌هایی از اطلاعات تمرکز کرده و آن‌ها را به‌عنوان ویژگی سیگنال به‌منظور مقایسه‌ی سیگنال‌ها با یکدیگر استخراج کنیم. همچنین بعد از تشریح استخراج ویژگی‌ها، به آموزش دیکشنری و زیرمجموعه‌های مربوط به آن برای استفاده در روش پیشنهادی خود پرداختیم.

فصل چهارم

روش پیشنهادی و پیاده‌سازی آن

۴-۱- مقدمه

در این فصل روش پیشنهادی برای جداسازی صدای دو گوینده‌ی همزمان از یک صدای مخلوط بیان می‌شود. یکی از چالش‌های موجود در این حوزه جداسازی صدای گوینده‌های هم‌جنس می‌باشد. با توجه به این‌که اکثر روش‌های موجود از تخمین گام برای جداسازی صدا استفاده می‌نمایند و فرکانس گام گوینده‌های هم‌جنس بسیار به‌هم نزدیک می‌باشد، بنابراین در جداسازی صدای هم‌جنس روش‌های موجود با مشکل روبرو می‌باشند. همانطور که در بخش نتایج خواهید دید روش پیشنهادی قادر است که صدای گوینده‌های هم‌جنس را نیز به‌خوبی گوینده‌های غیر هم‌جنس جدا کند.

برخلاف روش‌هایی مانند [۳۷]، که در آن‌ها عملیات جداسازی از طریق ردیابی گام گوینده‌های مختلف انجام می‌شود، در روش پیشنهادی عملیات جداسازی از طریق مدل‌سازی صدای گوینده‌های مختلف صورت می‌گیرد. در حقیقت نوآوری روش پیشنهادی مدل‌سازی صدای گوینده‌های مختلف با استفاده از آموزش دیکشنری می‌باشد. در روش پیشنهادی هر بردار ویژگی استخراج شده از صدای مخلوط بر حسب بردارهای پایه‌ی گوینده‌های مختلف تجزیه می‌شود. و در نهایت صدای جداسازی شده از طریق ترکیب بردارهای پایه‌ی یک گوینده مشخص به‌دست می‌آید.

در روش پیشنهادی، برخلاف روش‌های پیشین مانند [۳۷] و [۴۰]، فرکانس جداسازی، محدود نمی باشد. همچنین یکی از محاسن روش پیشنهادی محدود نبودن به دو گوینده می‌باشد، یعنی می‌توان با تعمیم روش پیشنهادی صدای مخلوط تعداد بیشتری گوینده را نیز جداسازی کرد. در روش پیشنهادی از ویژگی و خواص تبدیل فوریه زمان کوتاه برای جداسازی سیگنال صوتی استفاده شده است، درحالی‌که اکثر روش‌هایی که تاکنون پیشنهاد شده است از کاکلیگرام و یا فیلتربانک گاماتون برای جداسازی استفاده کرده‌اند. سادگی ویژگی مورد استفاده در روش پیشنهادی باعث کاهش زمان لازم برای جداسازی می‌شود. روش پیشنهادی یک روش با نظارت می‌باشد. در روش پیشنهادی ابتدا نمونه‌های صدای دو گوینده همزمان مدل‌سازی می‌شود. این مدل‌سازی به‌وسیله‌ی آموزش یک دیکشنری و به‌دست آوردن بردارهای پایه‌ی مربوط به هر کدام از گوینده‌ها انجام می‌شود. بعد از

مرحله‌ی مدل‌سازی، برای هر صدا مولفه‌های فرکانسی مربوط به هر کدام از دو گوینده همزمان جداسازی می‌شود. به‌منظور کاهش نویز صدای خروجی و حذف مولفه‌های فرکانسی سیگنال‌های تداخل، یک مرحله‌ی پس‌پردازش[1] نیز روی صدای خروجی انجام می‌شود. نتایج خروجی نشان می‌دهد که روش پیشنهادی طبق معیار SNR (نسبت سیگنال به نویز) و MSE (خطای میانگین مربعات)، که در فصل بعد به آن‌ها اشاره می‌شود برتری قابل توجهی نسبت به سایر روش‌های موجود دارد. روش پیشنهادی شامل سه بخش آموزش، تست و پس‌پردازش می‌باشد. در بخش بعد هر کدام از این مراحل در الگوریتم‌های پیشنهادی ۱، ۲ و ۳ توضیح داده شده‌اند.

۴-۲- الگوریتم‌های پیشنهادی

الگوریتم پیشنهادی ۱: فاز آموزش

۱- برای هر کدام از گویندگان مراحل ۲ تا ۶ اجرا شود.
۲- برای تمام صداهای آموزشی مراحل ۳ و ۴ اجرا شود.
۳- بخش‌بندی صدای ورودی
۴- استخراج ویژگی تبدیل فوریه، MFCC و LPC از فریم‌های موجود در صدای ورودی.
۵- تشکیل ماتریس ویژگی گوینده.
۶- استفاده از الگوریتم K-SVD برای به‌دست آوردن بردارهای پایه و مدل‌سازی صدای گوینده.

الگوریتم پیشنهادی ۲: فاز تست

۱- کنار هم قرار دادن دیکشنری‌های تمام گویندگان و تشکیل دیکشنری جامع.
۲- برای هر صدای تست مراحل ۳ تا ۱۰ اجرا شود.
۳- بخش‌بندی صدای ورودی.
۴- استخراج ویژگی تبدیل فوریه از هر فریم صدای ورودی.
۵- به‌ازای هر فریم موجود در صدای ورودی مراحل ۶ تا ۸ اجرا شود.
۶- پیدا کردن ضرایب اسپارس مربوط به ویژگی استخراج شده از این فریم.
۷- جداسازی ضرایب مربوط به هر کدام از گویندگان.
۸- به‌دست آوردن بردار ضرایب تبدیل فوریه صدای جداسازی شده در این فریم.

[1] Postprocessing

۹- به‌دست آوردن ماتریس تبدیل فوریه زمان کوتاه هر گوینده.

۱۰- محاسبه‌ی عکس تبدیل فوریه و بدست آوردن صدای جداسازی شده.

الگوریتم پیشنهادی ۳: فاز پس پردازش

۱- برای هر صدای جداسازی شده مراحل ۲ تا ۷ اجرا شود.

۲- بخش‌بندی صدا.

۳- استخراج ویژگی MFCC و LPC از هر فریم.

۴- تشکیل بردار ویژگی $F = \begin{bmatrix} MFCC \\ LPC \end{bmatrix}$.

۵- پیدا کردن نزدیک‌ترین بردار به بردار F در ماتریس ویژگی C.

۶- جایگزین کردن بردار حوزه‌ی زمان مربوط به این ویژگی.

۷- میانگین‌گیری در مناطقی‌که همپوشانی دارد.

در ادامه هر کدام از مراحل بالا با جزئیات بیشتر شرح داده می‌شوند:

۴-۳- فاز آموزش

۴-۳-۱- بخش‌بندی

سیگنال صوتی یک سیگنال غیر ایستا می‌باشد. برای توصیف بهتر سیگنال‌های غیر ایستا از بخش‌بندی استفاده می‌شود. بخش‌بندی به عملیات تبدیل سیگنال اصلی به تکه‌های کوتاه زمان گفته می‌شود. در این پایان‌نامه هر سیگنال صوتی به فریم‌های کوتاه زمان ۲۵ میلی ثانیه‌ای تقسیم می‌شود. به‌منظور استخراج جزئیات کامل سیگنال، فریم‌های کوتاه زمان با همپوشانی انتخاب می‌شوند. بدیهی است هر چه همپوشانی بین فریم‌ها بیشتر باشد اطلاعات استخراج شده کامل‌تر می‌باشد. اما باید در نظر داشت که همپوشانی زیاد بین فریم‌های استخراج شده باعث افزایش پیچیدگی محاسباتی الگوریتم می‌شود. بنابراین در این پایان‌نامه از همپوشانی ۵۰ درصد بین فریم‌ها برای استخراج اطلاعات استفاده می‌شود.

۴-۳-۲- استخراج ویژگی

در مرحله قبل هر سیگنال صوتی به تعدادی فریم با تعداد نمونه‌های برابر تقسیم شد. اگر هر کدام از فریم‌های استخراج شده را به‌صورت برداری بنویسیم، داریم:

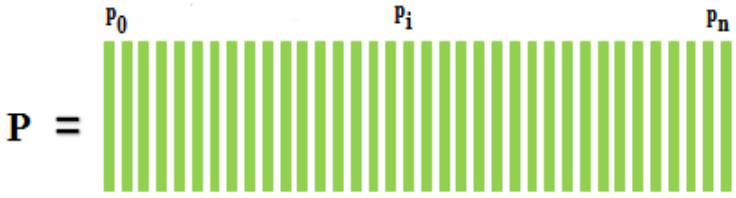
$$P_i = vec(S_i) \qquad (۱-۴)$$

به‌طوری‌که S_i فریم i-ام از سیگنال S می‌باشد. عملگر vec، عملگر تبدیل یک فریم سیگنال صوتی به بردار ستونی می‌باشد.

در این مرحله لازم است اطلاعات موجود در تمام سیگنال‌های صوتی مربوط به یک گوینده‌ی مشخص در یک ماتریس جمع‌آوری شود. اگر تمام P_iهای یک گوینده مشخص کنار یکدیگر قرار داده شوند ماتریس P به‌صورت زیر به‌وجود می‌آید:

$$P = [P_0, P_1, P_2, \dots, P_i, \dots, P_n] \qquad (۲-۴)$$

در شکل (۴-۱) نحوه تشکیل ماتریس P نشان داده شده است.



شکل (۴-۱) نحوه‌ی تشکیل ماتریس P

در این مرحله از هر کدام از ستون‌های ماتریس P سه ویژگی استخراج می‌شود. ۱) تبدیل فوریه، ۲) MFCC، ۳) LPC.

از ویژگی تبدیل فوریه در قسمت آموزش و تست استفاده خواهد شد و از ویژگی MFCC و LPC برای بهبود عملکرد الگوریتم پیشنهادی استفاده می‌شود.

- **تبدیل فوریه**

تبدیل فوریه مهم‌ترین ویژگی مورد استفاده در این پایان‌نامه می‌باشد. تبدیل فوریه دارای خواص منحصر به‌فردی می‌باشد که آن را از سایر ویژگی‌های مرسوم سیگنال‌های صوتی متمایز کرده است. ویژگی خطی بودن و معکوس پذیر بودن تبدیل فوریه کارایی منحصر به‌فردی به آن برای جداسازی سیگنال صوتی می‌دهد. اگر دو سیگنال به نام‌های s_1 و s_2 داشته باشیم تبدیل فوریه آن‌ها به‌صورت زیر می‌باشد:

$$S_1 = F(s_1) \quad (۳-۴)$$

$$S_2 = F(s_2) \quad (۴-۴)$$

به‌صورتی‌که S_1 و S_2 به‌ترتیب تبدیل فوریه سیگنال s_1 و s_2 می‌باشند.

حال اگر سیگنال s_3 را به‌صورت زیر تعریف نماییم:

$$s_3 = s_1 + s_2 \quad (۵-۴)$$

تبدیل فوریه سیگنال s_3 به‌صورت زیر به‌دست می‌آید:

$$S_3 = F(s_3) = F(s_1 + s_2) \quad (۶-۴)$$

با توجه به خاصیت خطی بودن تبدیل فوریه خواهیم داشت:

$$F(s_1 + s_2) = F(s_1) + F(s_2) \quad (۷-۴)$$

بنابراین:

$$S_3 = S_1 + S_2 \quad (۸-۴)$$

بدین معنی که اگر سیگنال s_3 ترکیب خطی سیگنال s_1 و s_2 باشد، تبدیل فوریه S_3 نیز ترکیب خطی تبدیل فوریه‌ی این دو سیگنال می‌باشد. این خاصیت تبدیل فوریه را به ویژگی منحصر به‌فردی برای جداسازی سیگنال صوتی تبدیل می‌نماید. با توجه به مقدمه بالا اگر صدای دو گوینده در حوزه‌ی زمان با هم مخلوط شده باشند، می توانیم عملیات جداسازی را در حوزه‌ی فرکانس انجام دهیم.

خاصیت مهم دیگر تبدیل فوریه قابلیت معکوس‌پذیری آن می‌باشد. بعد از جداسازی مولفه‌های فرکانسی سیگنال می‌توان از قابلیت معکوس‌پذیری تبدیل فوریه استفاده نمود و سیگنال حوزه زمان آن را به‌دست آورد.

با توجه به مطالب گفته شده، در این پایان‌نامه از خواص تبدیل فوریه زمان کوتاه برای جداسازی صدای گوینده‌ها استفاده می‌شود. اگر p_i یک ستون از ماتریس P باشد، ضرایب تبدیل فوریه‌ی f_i به‌صورت زیر به‌دست می‌آید:

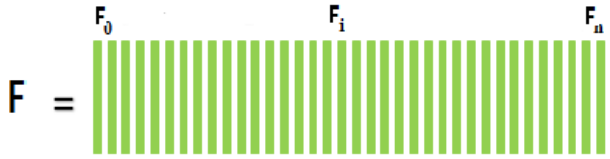
$$f_i = abs(F\ (p_i)) \qquad (۹-۴)$$

به‌صورتی‌که F عملگر استخراج ضرایب تبدیل فوریه از سیگنال P_i می‌باشد. عملگر abs(.) مقدار قدر مطلق یک بردار را برمی‌گرداند. f_i اندازه‌ی ضرایب تبدیل فوریه‌ی سیگنال P_i می‌باشد.

اگر از تمام فریم‌های موجود در ستون‌های ماتریس P تبدیل فوریه گرفته شود ماتریس حاصل به‌صورت زیر خواهد بود:

$$F = [f_0, f_1, f_2, \dots, f_i, f_n] \qquad (۱۰-۴)$$

در شکل (۴-۲) ماتریس اندازه‌ی تبدیل فوریه زمان کوتاه مربوط به یک گوینده خاص نشان داده شده است.



شکل (۴-۲) ماتریس اندازه‌ی تبدیل فوریه زمان کوتاه مربوط به یک گوینده خاص

- **ویژگی MFCC و LPC**

بعد از این‌که صدای گوینده‌های مختلف توسط عکس تبدیل فوریه جداسازی شد، ممکن است تعدادی از مولفه‌های فرکانسی به اشتباه به یکی از گوینده‌ها نسبت داده شوند، در صورتی‌که

متعلق به گوینده دیگر باشند. بنابراین برای بهبود عملکرد الگوریتم و همچنین افزایش دقت جداسازی نیاز به یک مرحله پس‌پردازش می‌باشد. با توجه به عملکرد مناسبی که ویژگی‌های MFCC و LPC در شناسایی صداهای مختلف از خود نشان داده‌اند، در این پایان‌نامه ویژگی‌های MFCC و LPC را از هر بردار موجود در ماتریس P به‌صورت زیر استخراج می‌نماییم:

$$m_i = MFCC(p_i) \quad (۱۱-۴)$$

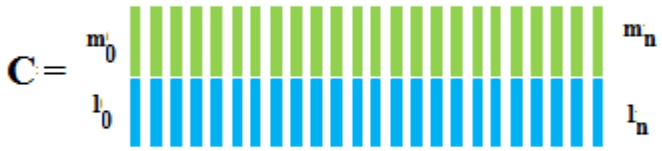
$$l_i = LPC(p_i) \quad (۱۲-۴)$$

به‌صورتی‌که p_i بردار i-ام موجود در ماتریس P می‌باشند. MFCC و LPC به ترتیب عملگرهای استخراج ویژگی MFCC و LPC می‌باشند. m_i و l_i به ترتیب ویژگی‌های MFCC و LPC استخراج شده از بردار p_i می‌باشند.

اگر از تمام بردارهای موجود در ماتریس P به‌صورت بالا ویژگی MFCC و LPC استخراج شود، می‌توان ماتریس ویژگی به‌منظور عملیات پس‌پردازش را به‌صورت زیر تعریف کرد:

$$C = \begin{bmatrix} m_0 & m_1 & \dots & m_n \\ l_0 & l_1 & \dots & l_n \end{bmatrix} \quad (۱۳-۴)$$

در شکل (۴-۳) نحوه‌ی تشکیل ماتریس ویژگی به‌منظور عملیات پس‌پردازش نشان داده شده است.

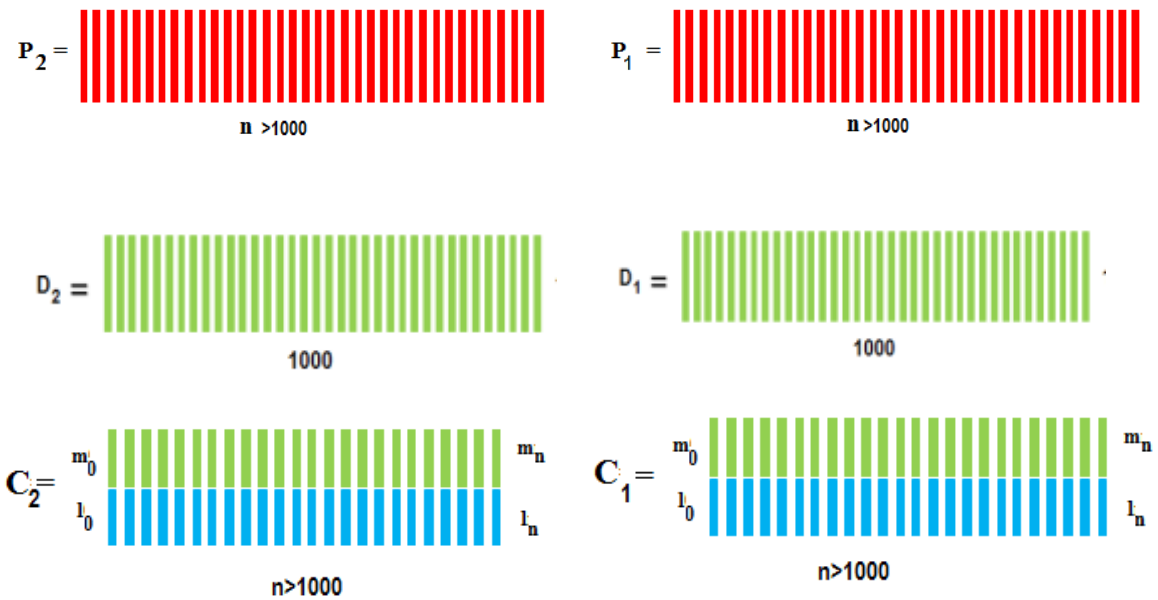


شکل (۴-۳) نحوه‌ی تشکیل ماتریس ویژگی به‌منظور عملیات پس‌پردازش

۴-۳-۳- آموزش دیکشنری

مرحله بعد، ساخت یک دیکشنری برای هر کدام از گوینده‌ها با استفاده از ماتریس F استخراج شده در مرحله قبل می‌باشد. همانطور که در فصل قبل توضیح داده شد، الگوریتم K-SVD دیکشنری مورد نیاز برای ارائه‌ی هر بردار دلخواه بر حسب ترکیب خطی بردارهای پایه را می‌سازد. دیکشنری آموزش-یافته قادر است هر بردار دلخواه را بر حسب ترکیب خطی تعداد اندکی از اتم‌های موجود در دیکشنری بیان کند. دیکشنری مورد استفاده در الگوریتم پیشنهادی شامل ۱۰۰۰ اتم برای هر گوینده می‌باشد. طول بردار هر اتم ۵۱۳ نمونه می‌باشد. بنابراین بعد از مرحله‌ی آموزش برای هر گوینده یک دیکشنری با ابعاد ۱۰۰۰×۵۱۳ خواهیم داشت.

تا این مرحله برای هر گوینده سه ماتریس (P و D و C) ساخته شده است که در مراحل بعدی الگوریتم، از آن‌ها استفاده خواهد شد. شکل (۴-۴) تمام ماتریس‌هایی که در مرحله‌ی آموزش ساخته شده‌اند را نشان می‌دهد.



شکل (۴-۴) تمام ماتریس‌های ساخته شده در مرحله‌ی آموزش برای دو گوینده

D_1 و D_2 به‌ترتیب دیکشنری‌های مربوط به گوینده‌ی اول و دوم می‌باشند که از طریق نتیجه‌ی الگوریتم K-SVD به‌دست آمده‌اند.

۴-۴- فاز تست

در فاز تست یک سیگنال مخلوط از صدای دو گوینده همزمان به الگوریتم داده خواهد شد و با توجه به مدل‌سازی که در مرحله قبل انجام شده است، الگوریتم سعی در جداسازی صدای دو گوینده دارد.

۴-۴-۱- بخش‌بندی سیگنال تست و استخراج ویژگی

این مرحله مشابه مرحله‌ی بخش‌بندی انجام شده در مرحله‌ی آموزش می‌باشد. سیگنال ورودی را به بخش‌هایی به طول ۲۵ میلی ثانیه با همپوشانی ۵۰ درصد تقسیم می‌کنیم. سپس از هر فریم استخراج شده از سیگنال تست، اندازه‌ی تبدیل فوریه گرفته می شود. تبدیل فوریه مورد استفاده در این پایان‌نامه شامل ۵۱۳ ضریب می‌باشد.

۴-۴-۲- تشکیل دیکشنری جامع

فلسفه‌ی ساخت دیکشنری، ارائه‌ی یک سیگنال بر حسب ترکیب خطی بردارهای پایه‌ی موجود در دیکشنری می‌باشد. با توجه به این‌که هر کدام از دیکشنری‌های ساخته شده، قادر می‌باشند صدای یکی از گویندگان را بر حسب ترکیب خطی بردارهای پایه بیان کنند، بنابراین یک دیکشنری جامع که شامل بردارهای پایه‌ی هر دو دیکشنری باشد، قادر است سیگنال صوتی مخلوط را بر حسب اتم‌های موجود در این دیکشنری ارائه دهد. برای تشکیل دیکشنری جامع به‌صورت زیر عمل می‌نمائیم:

$$D = [D_1, D_2] \qquad (۱۴-۴)$$

به‌صورتی‌که D_1 و D_2 دیکشنری مربوط به گوینده‌ی اول و دوم می‌باشند و D دیکشنری جامع است.

۴-۴-۳- محاسبه‌ی ضرایب تُنک

بعد از این‌که از هر فریم صدای تست ویژگی تبدیل فوریه استخراج شد، لازم است ضرایب اسپارس این ویژگی برای عملیات جداسازی محاسبه شود. در حقیقت با محاسبه‌ی ضرایب اسپارس یک بردار

ویژگی، سهم هر دیکشنری از مولفه‌های بردار ویژگی مشخص می‌شود. اگر f_i بردار اندازه‌ی تبدیل فوریه زمان کوتاه مربوط به فریم i-ام صدای تست باشد، داریم:

$$\alpha = argmin_{\alpha}\|f_i - D\alpha\| + \gamma\|\alpha\|_0 \quad (۱۵-۴)$$

با حل این معادله، ضرایب α به‌گونه‌ای به‌دست می‌آید که بردار f_i توسط دیکشنری D توصیف شود. همانطور که در معادله‌ی بالا مشخص است، $\|\alpha\|_0$ شرط تُنک بودن را به معادله اضافه می‌نماید. این معادله توسط الگوریتم OMP قابل حل خواهد بود.

۴-۴-۴- جداسازی مولفه‌های فرکانسی

بعد از این‌که ضرایب α به‌دست آمد، می‌توان سهم هر دیکشنری از این ضرایب را جداسازی کرد؛ بدین معنی که بردار ضرایب α به دو قسمت ۱۰۰۰ تایی تقسیم می‌شود، به‌گونه‌ای که ۱۰۰۰ ضریب اول سهم دیکشنری اول و ۱۰۰۰ ضریب دوم سهم دیکشنری دوم می‌باشد.

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_{1000} \\ \alpha_{1001} \\ \dots \\ \alpha_{2000} \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \quad (۱۶-۴)$$

$\hat{\alpha}_1$ ضرایب اسپارس دیکشنری D_1 و $\hat{\alpha}_2$ ضرایب اسپارس دیکشنری D_2 می‌باشد. عملیات جداسازی مولفه‌های فرکانسی توسط ضرب ضرایب اسپارس هر دیکشنری در همان دیکشنری انجام می‌شود.

$$h_1 = D_1\hat{\alpha}_1 \quad (۱۷-۴)$$

$$h_2 = D_2\hat{\alpha}_2 \quad (۱۸-۴)$$

h_1 مولفه‌ی فرکانسی مربوط به گوینده‌ی اول و h_2 مولفه‌ی فرکانسی مربوط به گوینده دوم می‌باشد. مراحل بالا برای تمام فریم‌های صدای مخلوط انجام می‌شود تا مولفه‌های فرکانسی تمام فریم‌ها جداسازی شود.

۴-۴-۵- تبدیل فوریه معکوس

اگر عملیات جداسازی را برای تمام فریم‌های موجود در صدای مخلوط انجام دهیم، در نهایت دو ماتریس شامل مولفه‌های فرکانسی گوینده‌ی اول و دوم خواهیم داشت. حال اگر از هر کدام از ماتریس‌های به‌دست آمده عکس تبدیل فوریه زمان کوتاه بگیریم، صدای جداسازی شده مربوط به هر کدام از گوینده‌ها به‌دست می‌آید. با توجه به این‌که گوش انسان به فاز موجود در صدا حساس نمی‌باشد، برای انجام عملیات معکوس تبدیل فوریه از فاز صدای مخلوط استفاده می‌نمائیم.

۴-۵- پس‌پردازش

بعد از این‌که صدای هر کدام از دو گوینده همزمان توسط روش پیشنهادی جداسازی شد، برای افزایش دقت جداسازی و همچنین کاهش نویز، از یک مرحله پس‌پردازش روی صداهای جداسازی شده استفاده می‌شود. در پس‌پردازش پیشنهادی، هر فریم از صدای جداسازی شده، با نزدیک‌ترین فریم از فریم‌های آموزشی (از لحاظ فاصله‌ی اقلیدسی بردارهای هر دو فریم) جایگزین می‌شود. بعد از این عملیات می‌توان اطمینان حاصل کرد که تمام مولفه‌های فرکانسی صدای جداسازی شده به‌درستی شناخته شده است.

با توجه به عملکرد بسیار خوبی که ویژگی MFCC و LPC در شناسایی صدای گویندگان از خود نشان داده‌اند، در این پایان‌نامه از این دو ویژگی برای یافتن نزدیک‌ترین فریم به خروجی استفاده می‌نمائیم. مراحل اجرای پس‌پردازش به شرح زیر است:

۴-۵-۱- بخش‌بندی صدای جداسازی شده و استخراج ویژگی

همانند بخش‌بندی انجام شده در مرحله‌ی آموزش، در این مرحله نیز صدای خروجی مرحله‌ی تست را بخش‌بندی می‌نمائیم. هر دو صدای خروجی مرحله‌ی تست را به فریم‌های ۲۵ میلی ثانیه‌ای با همپوشانی ۵۰ درصد تقسیم می‌نمائیم. در مرحله‌ی بعد، از هر فریم صدا، ویژگی‌های MFCC و LPC را همانند آنچه در بخش قبل گفته شد استخراج می‌نمائیم. اگر $\overline{m}$، بردار ویژگی MFCC

استخراج شده از این فریم و $\bar{l}$ بردار ویژگی LPC استخراج شده از این فریم باشد، در مرحله‌ی بعد به دنبال نزدیک‌ترین فریم در فریم‌های آموزشی با توجه به این ویژگی می‌باشیم.

۴-۵-۲- انتخاب بهترین فریم

نزدیکترین فریم به‌نحوی انتخاب می‌شود که بردار ویژگی فریم مورد نظر، به بردار ویژگی فریمی که قصد پس‌پردازش آن را داریم، به لحاظ فاصله‌ی اقلیدسی نزدیک باشد، یعنی:

$$C_j = argmin_j \left\| C_j - \left[\frac{\bar{m}}{\bar{l}}\right] \right\|_2^2 \quad (۱۹-۴)$$

از آن‌جایی که تعداد بردارهای ماتریس C و ماتریس P یکسان هستند و بردارهای موجود در این دو ماتریس متناظر می‌باشند، بنابراین بعد از پیدا کردن نزدیک‌ترین فریم درون ماتریس C می‌توان بردار متناظر با آن را در ماتریس P جایگزین کرد. در حقیقت، با این روش نویزها و مولفه‌های فرکانسی که به صدا اضافه شده‌اند حذف خواهد شد.

۴-۶- جمع‌بندی و نتیجه‌گیری

در این فصل روش پیشنهادی در سه مرحله‌ی آموزش، تست و پس‌پردازش به‌صورت کامل بیان شد. ابتدا، نمونه صداهای دو گوینده همزمان را از طریق آموزش یک دیکشنری مدل‌سازی کردیم. سپس در مرحله‌ی تست، جداسازی دو گوینده همزمان بر اساس ویژگی تُنّک (ضرایب اسپارس) ارائه شد. همچنین، به‌منظور کاهش نویز صدای خروجی و حذف مؤلفه‌های فرکانسی سیگنال‌های تداخل، یک مرحله پس‌پردازش روی صدای خروجی انجام شد.

فصل پنجم

نتایج تجربی شبیه‌سازی

۵-۱- مقدمه

در فاز آموزش روش ارائه شده، هر سیگنال صوتی به فریم‌های ۲۵ میلی ثانیه‌ای با همپوشانی ۵۰ درصد تقسیم می‌شود. سیگنال صوتی ورودی با نرخ بیت ۲۵ کیلو نمونه بر ثانیه ضبط شده است. در هر ۲۵ میلی ثانیه، ۶۲۵ نمونه موجود می‌باشد. بنابراین بردار $p_i \in R^{625\times1}$ می‌باشد. برای استخراج ویژگی تبدیل فوریه، از ۱۰۲۴ ضریب استفاده شده است که با حذف ضرایب فوریه فرکانس‌های منفی، برداری به طول ۵۱۳ ضریب متناظر با هر بردار p_i وجود خواهد داشت. در استخراج ویژگی MFCC از ۱۳ ضریب و در استخراج ویژگی LPC از ۱۶ ضریب استفاده شده است. بنابراین بردار $C_i \in R^{29\times1}$ می‌باشد. دیکشنری گوینده به‌گونه‌ای آموزش داده می‌شود که هر اتم آن، بر حسب حداکثر ۲۵ اتم قابل ارائه می‌باشد. به‌منظور سرعت بخشیدن به فرآیند آموزش و تست، دیکشنری ساخته شده در روش ارائه شده دارای ۱۰۰۰ اتم می‌باشد. از آن‌جا که طول هر اتم ۵۱۳ می‌باشد، بنابراین $D \in R^{513\times1000}$ خواهد بود. در فاز تست هر دو دیکشنری مربوط به دو گوینده همزمان صدای مخلوط را به‌هم متصل می‌کنیم. دیکشنری جامع دارای ۲۰۰۰ اتم می‌باشد. الگوریتم OMP را به‌گونه‌ای اجرا می‌کنیم که هر بردار ویژگی استخراج شده از صدای تست را با حداکثر ۵۰ اتم بیان نماید. چون دیکشنری جامع ۲۰۰۰ اتم دارد، بنابراین بردار ارائه‌ی تُنک نیز دارای ۲۰۰۰ عنصر می‌باشد که ۱۰۰۰ عنصر اول مربوط به دیکشنری D_1 و ۱۰۰۰ عنصر دوم آن مربوط به دیکشنری D_2 می‌باشد.

۵-۲- پایگاه داده آموزشی

الگوریتم ارائه شده یک الگوریتم با نظارت می‌باشد؛ بنابراین نیاز به یک پایگاه داده با تعداد زیادی صدای ضبط شده وجود دارد. از طرفی، چالش اصلی موجود در جداسازی صدای تک‌میکروفونه، مربوط به جداسازی صدای هم‌جنس می‌باشد. بنابراین، لازم است پایگاه داده‌ی انتخاب شده هم شامل صدای مرد و هم زن باشد. پایگاه داده‌ی (SSC)[1] [۵۴] شامل ۱۷۰۰۰ صدای ضبط شده‌ی گوینده مرد و زن

[1] Speech separation chalenge

می‌باشد. در این پایگاه داده، ۳۴ گوینده موجود می‌باشد که حدوداً نیمی از آن‌ها مرد و نیمی دیگر زن می‌باشند. برای هر کدام از این ۳۴ گوینده ۵۰۰ صدای ضبط شده موجود است. هر فایل صوتی در این پایگاه داده شامل ۶ کلمه می‌باشد که کلمات بیان شده در نام فایل صوتی نوشته شده است. برای مثال فایل صوتی "bbaf2n" شامل جمله‌ی "bin blue at F 2 now" می‌باشد. جدول (۵-۱)، لیست تمام انتخاب‌های ممکن کلمات در این پایگاه داده را نشان می‌دهد. جملات موجود برای هر کدام از گوینده‌ها ترکیبی از تمام حالات ممکن در این جدول می‌باشد.

جدول (۵-۱) لیست انتخاب‌های ممکن در پایگاه داده‌ی مورد نظر

Command	Color	Preposition	Letter	Number	Adverb
bin(b)	blue (b)	at (a)	A-Z	0-9	again(a)
Lay(l)	green(g)	by(b)			now(n)
Place(p)	red (r)	in(i)			please(p)
Set(s)	white(w)	with(w)			soon(s)

در این پایان‌نامه، برای آموزش الگوریتم پیشنهادی از ۴ گوینده موجود در این پایگاه داده استفاده شده است. از بین این چهار گوینده، دو گوینده مرد و دو گوینده زن می‌باشند. گفتنی است صداهایی که برای تست استفاده شده‌اند، در فرآیند آموزش به الگوریتم داده نشده‌اند. همچنین صداهای انتخابی برای تست، به‌صورت کاملاً تصادفی انتخاب شده است.

۵-۳- معیار ارزیابی

برای مقایسه روش پیشنهادی با سایر روش‌ها از معیار SNR و MSE استفاده شده است. معیار SNR به‌صورت زیر محاسبه می‌شود:

$$SNR = 10\ log_{10}(\sum_{t} x_a^2\ [t]\ /\ \sum_{t}(x_a[t] - \widehat{x_a}[t])^2) \qquad (۱-۵)$$

به‌صورتی‌که $x_a[t]$ و $\widehat{x_a}[t]$ به‌ترتیب سیگنال تمیز و سیگنال تخمین زده شده خروجی می‌باشند.

همچنین، معیار MSE به‌صورت زیر محاسبه می‌شود:

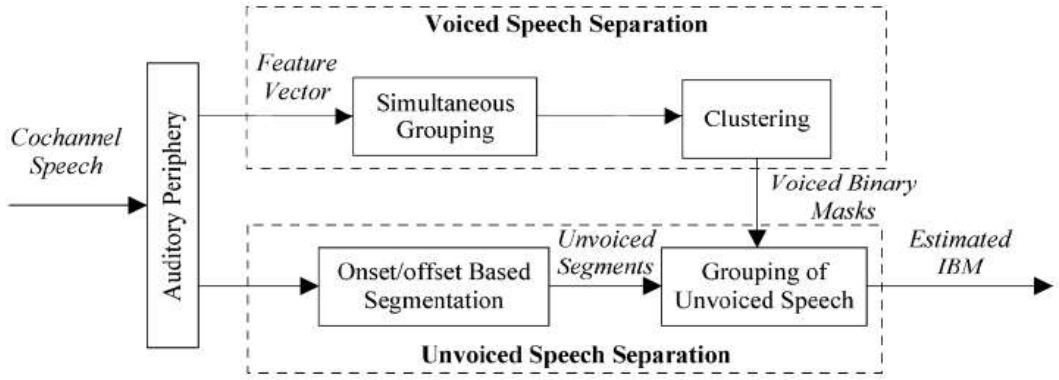
$$MSE = \frac{1}{T}\sum_{t}(x_a[t] - \widehat{x_a}[t])^2 \qquad (۲-۵)$$

به‌صورتی‌که $x_a[t]$ و $\widehat{x_a}[t]$ به‌ترتیب سیگنال تمیز و سیگنال تخمین زده شده خروجی می‌باشند و T تعداد کل نمونه‌های موجود در سیگنال $x_a[t]$ می‌باشد.

۵-۴- نتایج شبیه‌سازی

سه حالت مختلف برای آزمایش الگوریتم پیشنهادی در نظر گرفته شده است؛ بدین صورت که گوینده‌ها (مرد-مرد)، (زن-زن) و یا (زن-مرد) می باشند. در هر حالت، دو آزمایش مختلف با صداهای متفاوت انجام می‌شود.

در این قسمت، روش پیشنهادی با دو روش [۳۷] و [۴۰] (که در فصل دوم نیز به آن‌ها اشاره شده است) مقایسه می شود. مرجع [۳۷] به جداسازی بدون نظارت دو گوینده همزمان پرداخته است که بلوک دیاگرام این روش در شکل (۵-۱) نشان داده شده است؛ این روش مبتنی بر ویژگی و بدون نظارت می باشد. مطابق با اطلاعات موجود در [۳۷] و بلوک دیاگرام شکل (۵-۱)، از روش مبتنی بر ویزگی CASA و ماسک‌های باینری تخمین زده شده برای جداسازی استفاده شده است. همانطور که قبلاً نیز اشاره شد، CASA در قسمت آنالیز پیرامونی اغلب از دو حالت تبدیل فوریه یا فیلتربانک گاماتون برای جداسازی استفاده می‌کند که در [۳۷] از فیلتربانک گاماتون به‌عنوان روش استخراج ویژگی استفاده کرده است. مرجع [۴۰] از روش مبتنی بر مدل و از طریق یک الگوریتم تکراری، دو گوینده همزمان را جدا می‌کند. البته در این روش هم از یک ماسک باینری به‌منظور تخمین سیگنال خروجی استفاده شده است.



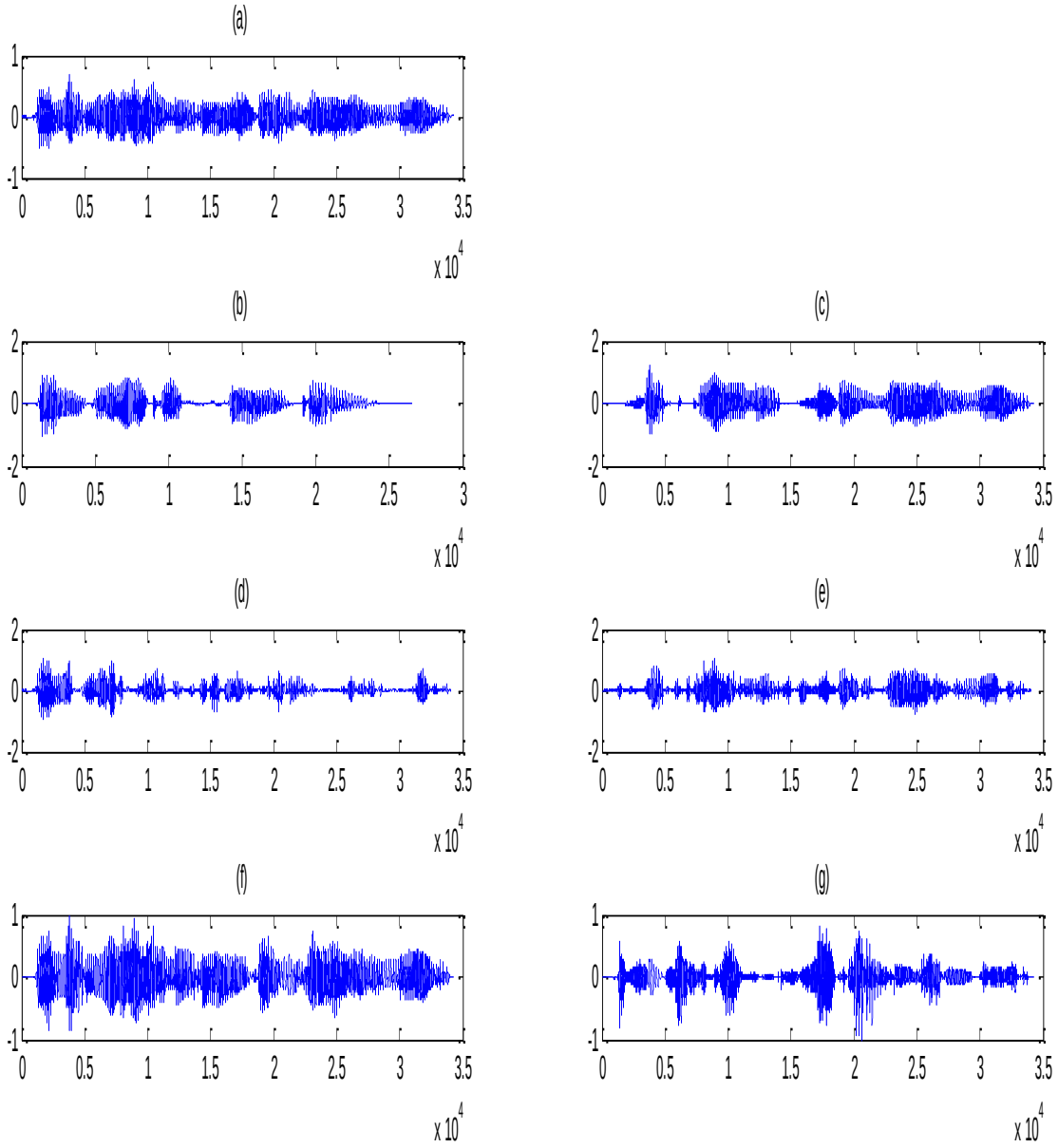
شکل (۵-۱) بلوک دیاگرام پیشنهادی برای سیستم جداسازی دو گوینده همزمان [۳۷]

۵-۴-۱- آزمایش اول و دوم دو گوینده هم‌جنس (مرد-مرد)

در این بخش، از دو گوینده هم‌جنس مرد برای انجام شبیه‌سازی به‌منظور جداسازی روش پیشنهادی و روش‌های موجود در [۳۷] و [۴۰] استفاده شده است. شکل‌های (۵-۲) و (۵-۳) نتایج مقایسه شکل موج خروجی روش پیشنهادی با الگوریتم‌های [۳۷] و [۴۰] را نشان می‌دهند. تفاوت آزمایش‌ها در جملات گفته شده و گویندگان آن‌ها می باشد، به‌طوری‌که در هر آزمایش گوینده و جمله‌ی گفته شده متفاوت می‌باشد. جملاتی که هر کدام از گویندگان به زبان می‌آورند، در قسمت زیرنویس شکل به‌صورت خلاصه آورده شده است.

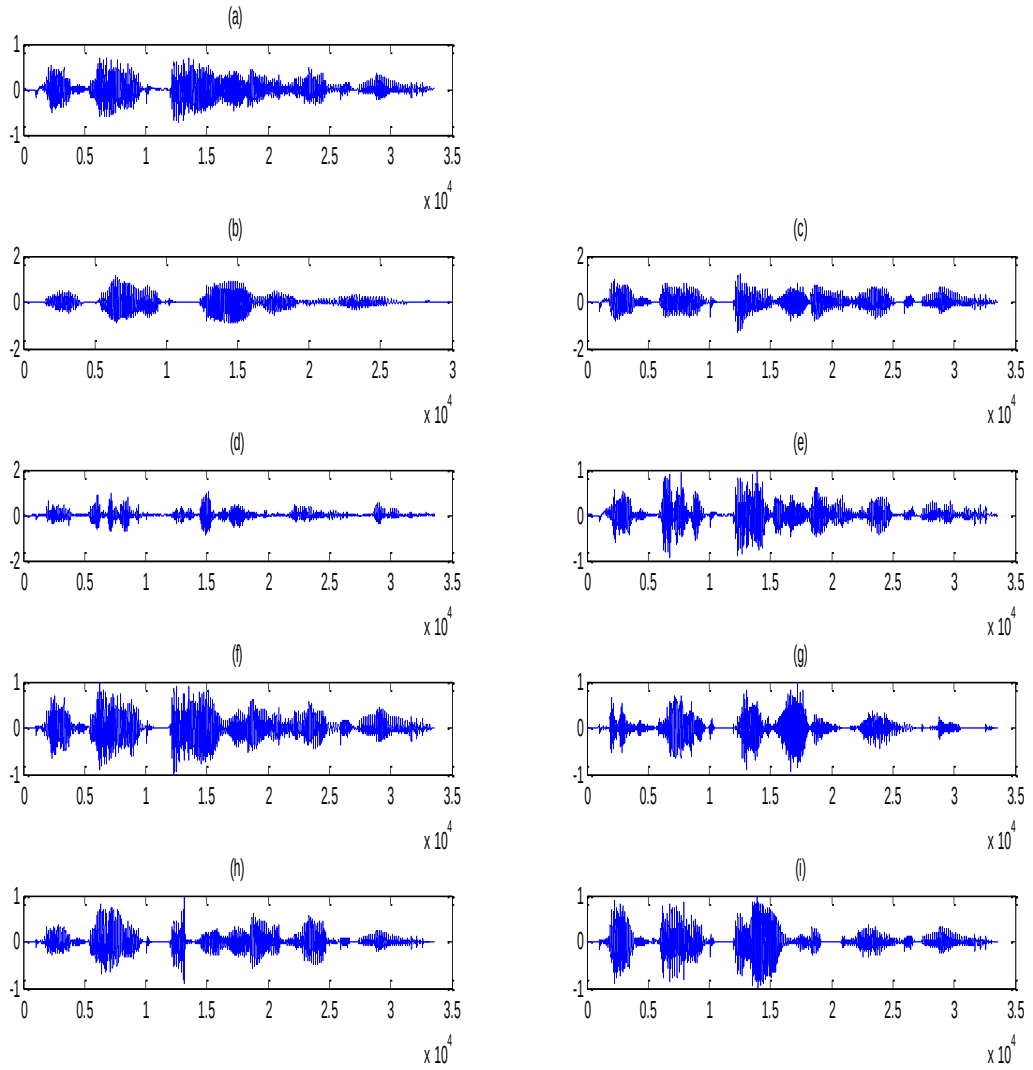
به‌صورت کلی در قسمت (a) هر کدام از شکل‌ها، صدای مخلوط دو گوینده نشان داده شده است. در قسمت (b) و (c)، صداهای تمیز ورودی مربوط به هر کدام از گویندگان قبل از مخلوط شدن آورده شده است. قسمت (d) و (e)، نتایج روش پیشنهادی را نشان می‌دهد و قسمت (f) و (g)، نتایج [۳۷] را نشان می‌دهد. در قسمت (h) و (i)، نتایج جداسازی روش [۴۰] آورده شده است.

در این بخش در آزمایش اول، جمله گفته شده گوینده اول (مرد)، "bin blue at F 4 again" و جمله گفته شده گوینده دوم (مرد)، "set green at C 2 now" می‌باشد. شکل موج خروجی در شکل (۵-۲) آورده شده است.



شکل (۵-۲) مقایسه شکل موج صدای جداسازی شده آزمایش اول (مرد-مرد)، جمله‌ی گفته شده توسط گوینده اول"bbaf4a" و جمله‌ی گفته شده توسط گوینده دوم "sgaczn" می باشد. در تمام شکل‌ها محور افقی نمونه و محور عمودی دامنه می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰]. در این آزمایش روش [۳۷] نتوانست به‌درستی جداسازی را انجام دهد.

در آزمایش دوم، جمله گفته شده گوینده اول (مـرد)، "lay blue with R 1 now" و جملـه گفتـه شده گوینده دوم (مرد)، "place red by C 8 now" می‌باشد. شـکل مـوج خروجـی در شـکل (۵-۳) آورده شده است.



شکل (۵-۳) مقایسه شکل موج صدای جداسازی شده آزمایش دوم (مرد-مرد)، جمله‌ی گفته شده توسط گوینده اول"lbwr1n" و جمله‌ی گفته شده توسط گوینده دوم "prbc8n" می باشد. در تمام شکل‌ها محور افقی نمونه و محور عمودی دامنه می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰] h) صدای جداسازی شده نخست توسط روش [۳۷] i) صدای جداسازی شده دوم توسط روش [۳۷]

همانطور که از این آزمایش‌ها پیداست، شکل موج خروجی روش پیشنهادی به مقدار بسیار زیادی به صداهای صاف ورودی نزدیک‌تر است. این نکته در تمام شکل موج‌های خروجی بخش‌های بعدی نیز صادق است. در واقع، روش پیشنهادی در مقایسه با دو روش موجود، عملکرد بهتری را از خود نشان داده است. همچنین قابل ذکر است که در آزمایش اول، روش بدون نظارت در [۳۷] نتوانست عملیات جداسازی را انجام دهد.

۵-۴-۲- آزمایش سوم و چهارم دو گوینده غیر هم‌جنس (مرد-زن)

در این بخش از دو گوینده غیر هم‌جنس مرد و زن برای انجام آزمایش‌ها و نتایج شبیه‌سازی استفاده شده است.

در این بخش در آزمایش سوم، جمله گفته شده گوینده اول (مرد)، "bin blue at F 4 again" و جمله گفته شده گوینده دوم (زن)، "lay white at D 6 place" می‌باشد.

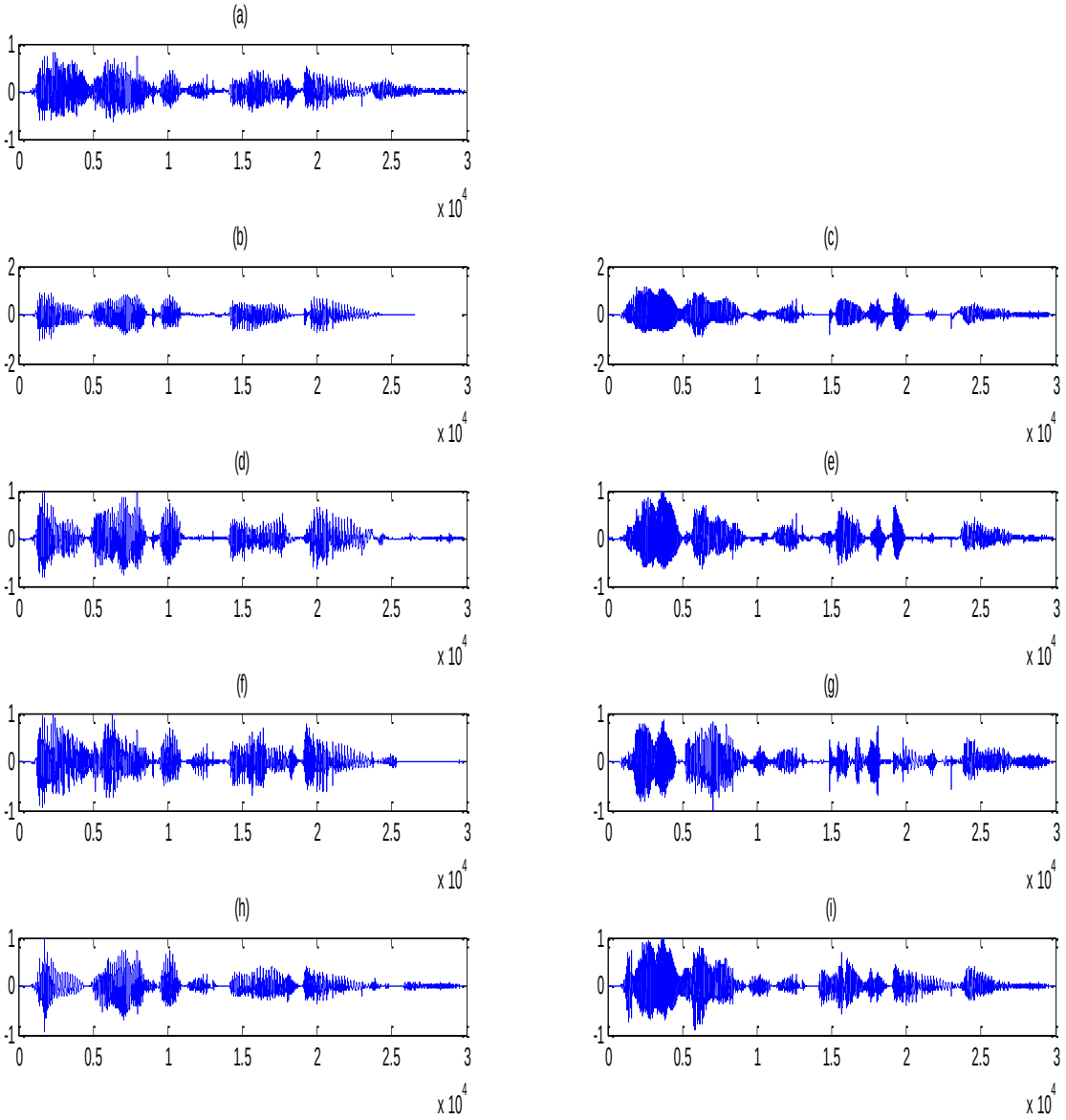
در آزمایش چهارم، جمله گفته شده گوینده اول (مرد)، "place red by C 8 now" و جمله گفته شده گوینده دوم (زن)، " bin red in R 3 soon" می‌باشد. نتایج در شکل‌های (۵-۴) و (۵-۵) آورده شده است.

۵-۴-۳- آزمایش پنجم و ششم دو گوینده هم‌جنس (زن-زن)

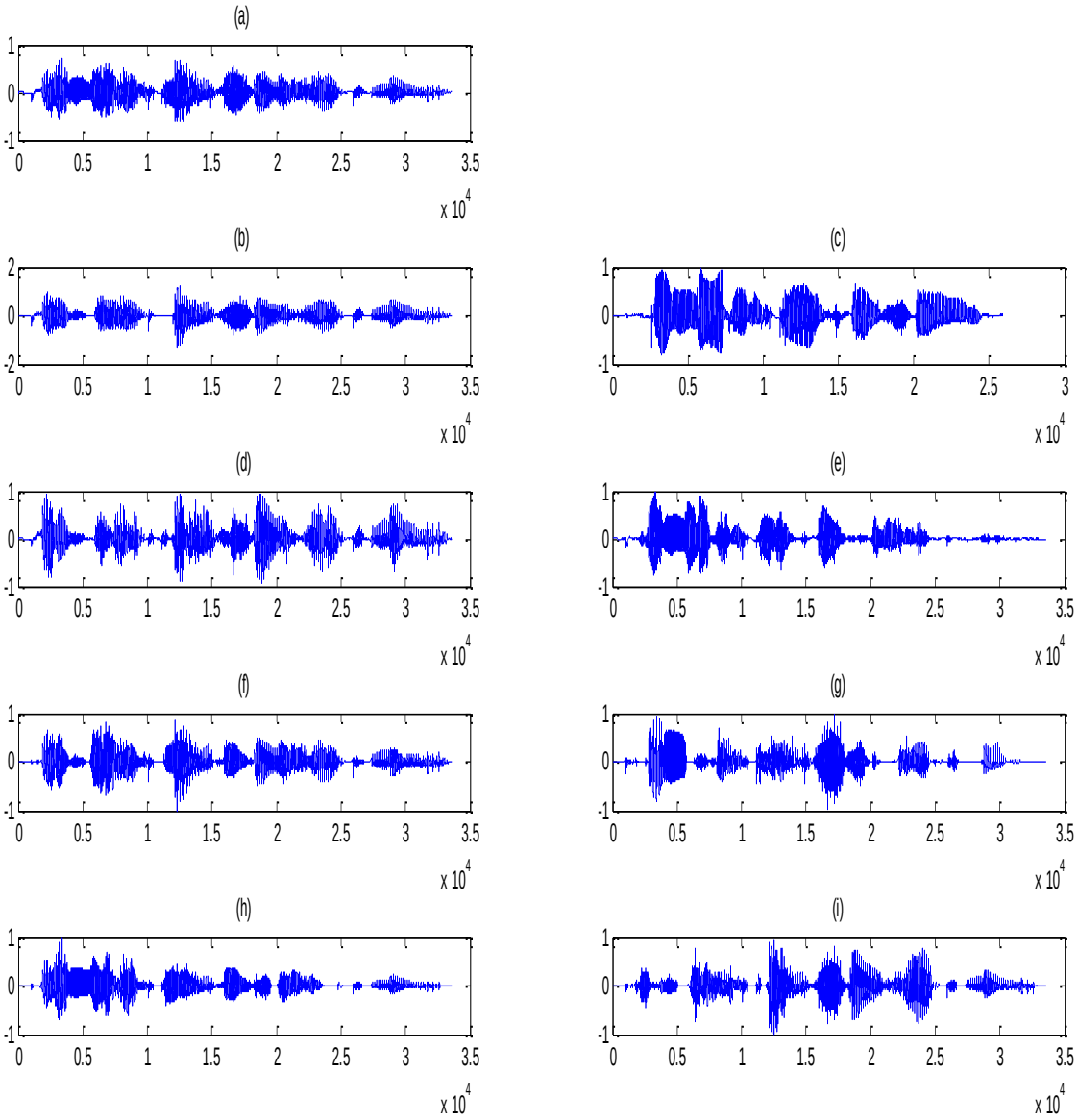
در این بخش از دو گوینده هم‌جنس زن برای انجام آزمایش‌ها و نتایج شبیه‌سازی استفاده شده است.

در این بخش در آزمایش پنجم، جمله گفته شده گوینده اول (زن)، " bin red in R 3 soon " و جمله گفته شده گوینده دوم (زن)، "lay white at D 6 place " می‌باشد.

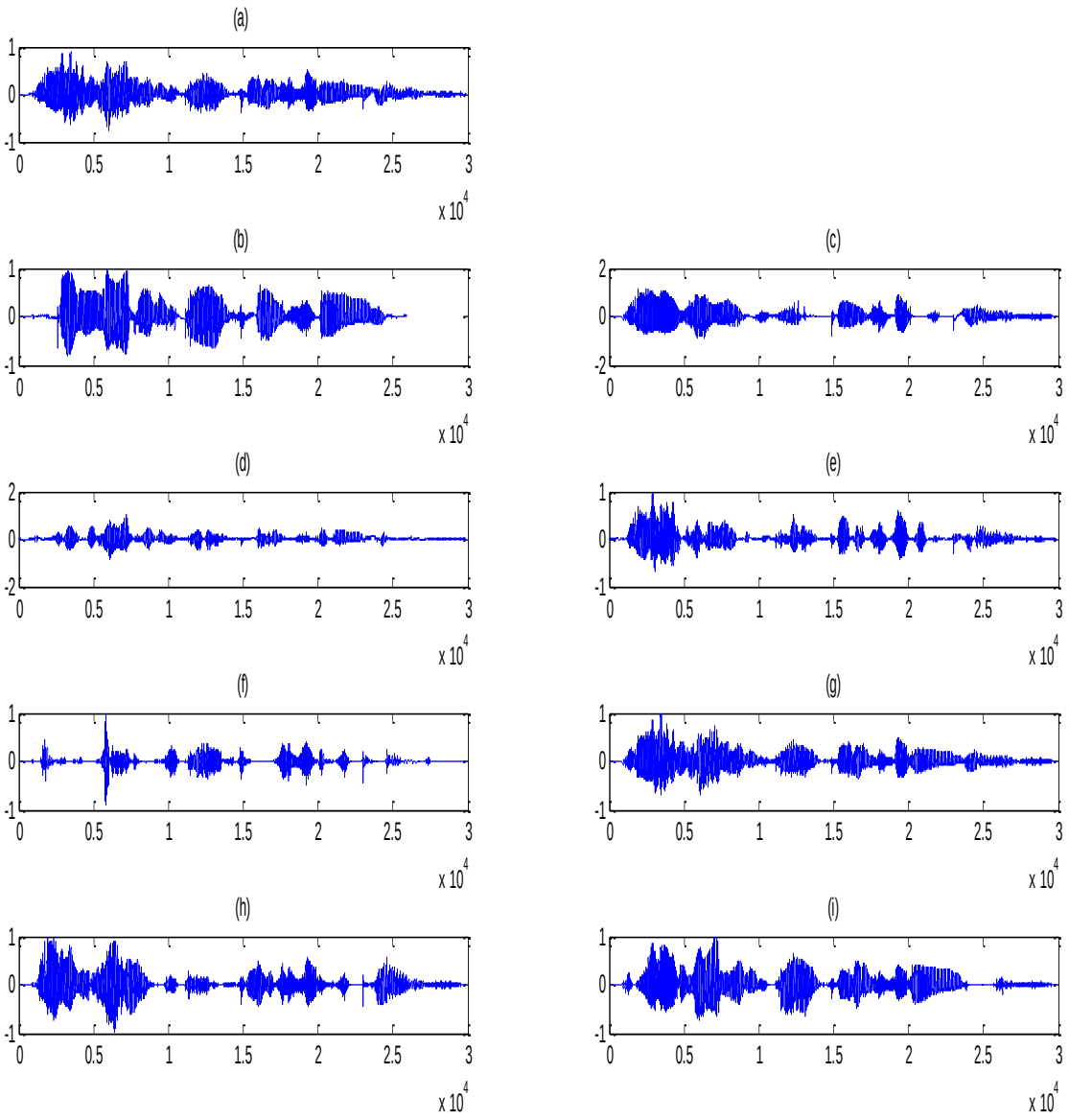
در آزمایش ششم، جمله گفته شده گوینده اول (زن) "set blue by M 9 again" و جمله گفته شده گوینده دوم (زن) "set white with U 2 now" می‌باشد. نتایج در شکل‌های (۵-۶) و (۵-۷) آورده شده است. در ادامه شکل موج‌های مربوط به هر آزمایش به‌ترتیب نشان داده شده است.



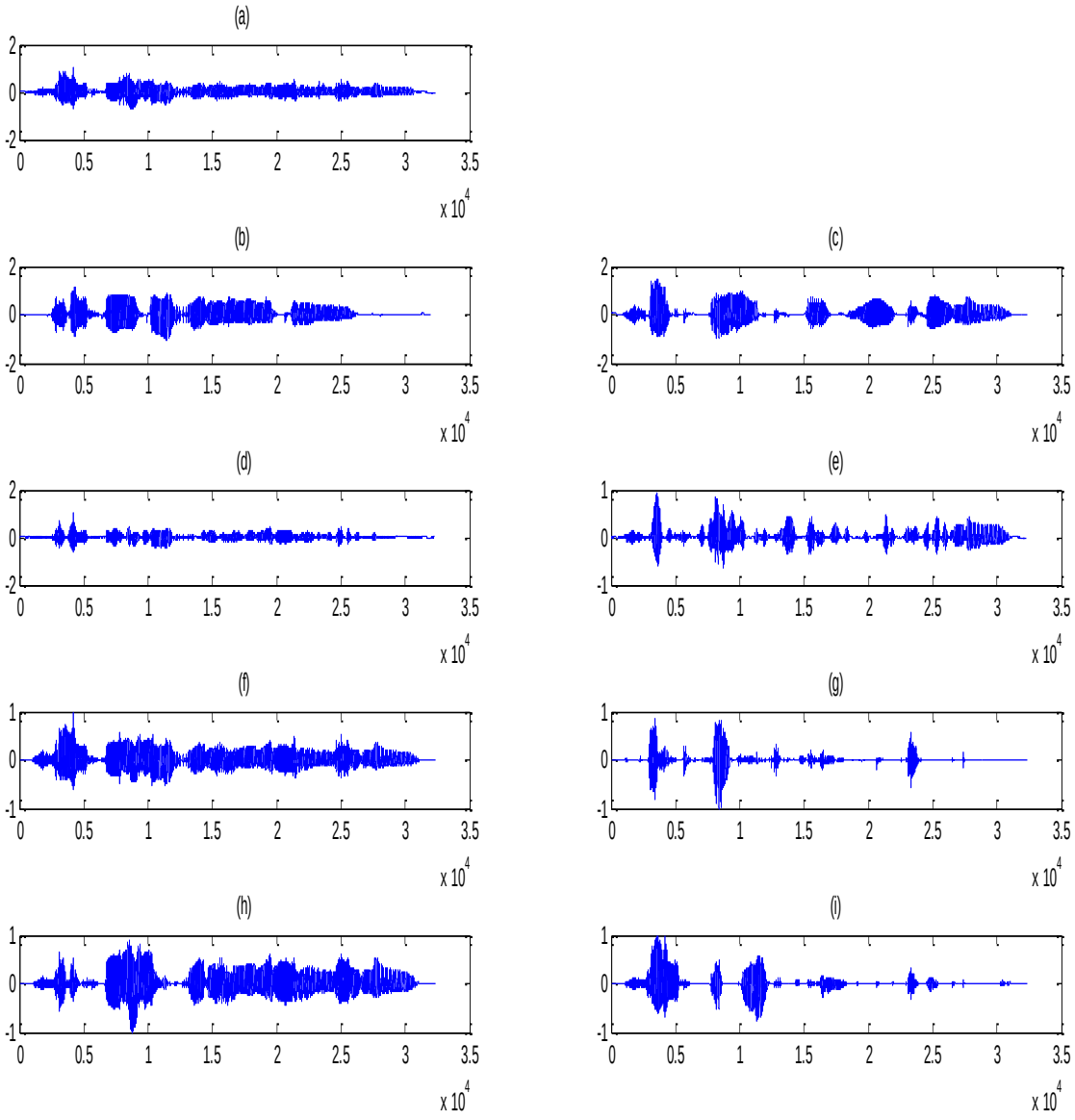
شکل (۵-۴) مقایسه شکل موج صدای جداسازی شده آزمایش سوم (مرد-زن) ، جمله‌ی گفته شده توسط گوینده اول"bbaf4a" و جمله‌ی گفته شده توسط گوینده دوم "lwad6p" می باشد. در تمام شکل‌ها محور افقی نمونه و محور عمودی دامنه می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰] h) صدای جداسازی شده نخست توسط روش [۳۷] i) صدای جداسازی شده دوم توسط روش [۳۷]



شکل (۵-۵) مقایسه شکل موج صدای جداسازی شده آزمایش چهارم (مرد-زن) ، جمله‌ی گفته شده توسط گوینده اول"prbc8n" و جمله‌ی گفته شده توسط گوینده دوم "brir3s" می باشد. در تمام شکل‌ها محور افقی نمونه و محور عمودی دامنه می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰] h) صدای جداسازی شده نخست توسط روش [۳۷] i) صدای جداسازی شده دوم توسط روش [۳۷]

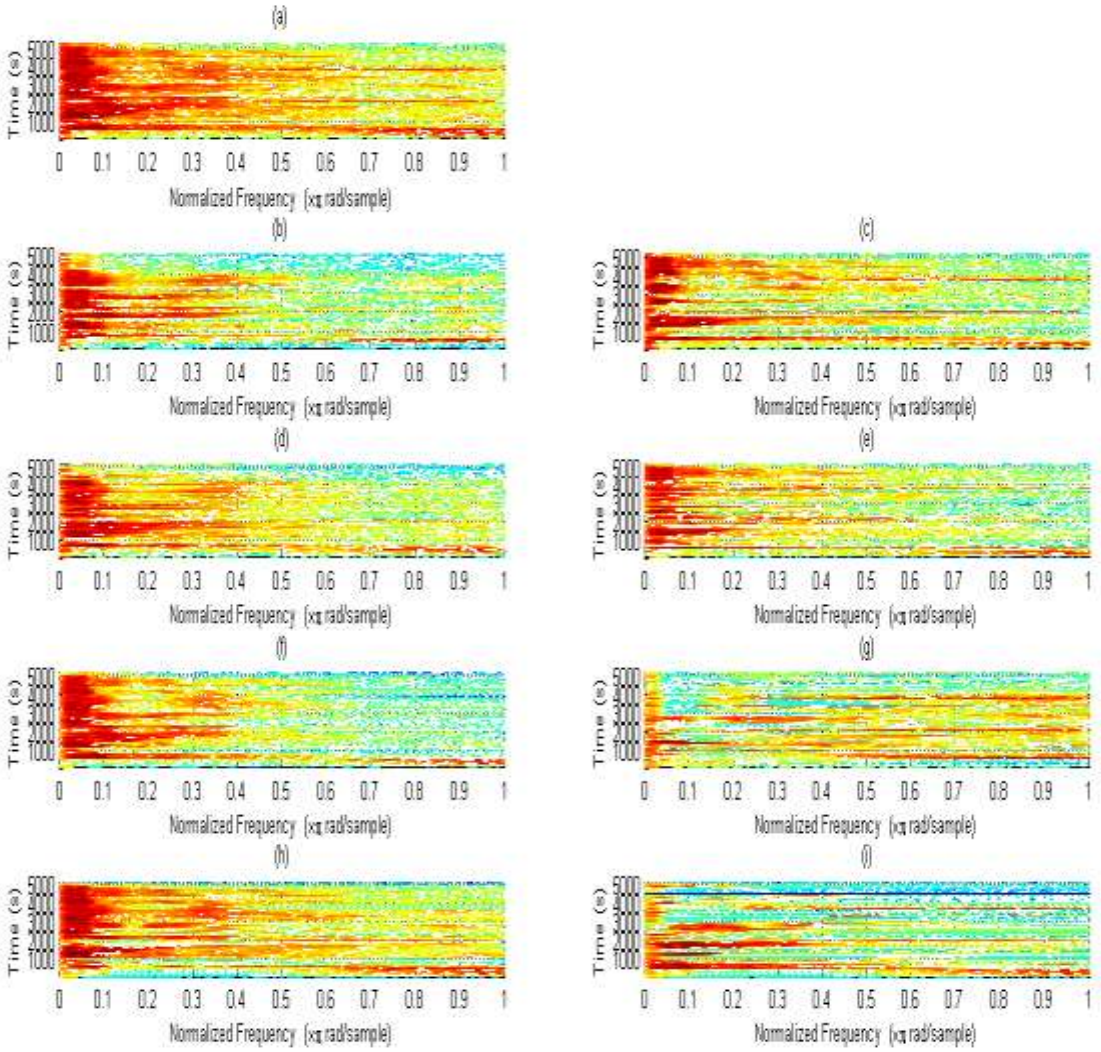


شکل (۵-۶) مقایسه شکل موج صدای جداسازی شده آزمایش پنجم (زن-زن) ، جمله‌ی گفته شده توسط گوینده اول"brir3s" و جمله‌ی گفته شده توسط گوینده دوم "lwad6p" می باشد. در تمام شکل‌ها محور افقی نمونه و محور عمودی دامنه می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰] h) صدای جداسازی شده نخست توسط روش [۳۷] i) صدای جداسازی شده دوم توسط روش [۳۷]



شکل (۵-۷) مقایسه شکل موج صدای جداسازی شده آزمایش ششم (زن-زن) ، جمله‌ی گفته شده توسط گوینده اول"sbbm9a" و جمله‌ی گفته شده توسط گوینده دوم "swwu2n" می باشد. در تمام شکل‌ها محور افقی نمونه و محور عمودی دامنه می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰] h) صدای جداسازی شده نخست توسط روش [۳۷] i) صدای جداسازی شده دوم توسط روش [۳۷]

در شکل (۵-۸) اسپکتروگرام صدای خروجی تمام روش‌های آزمایش ششم (زن-زن) آورده شده است. با توجه به شکل مشاهده می‌شود که روش پیشنهادی، واحدهای زمان-فرکانسی را بسیار بهتر از سایر روش‌ها جداسازی کرده است. در واقع اسپکتروگرام جمله جداسازی شده توسط روش پیشنهادی، نسبت به روش [۳۷،۴۰]، شباهت بیشتری به شکل‌های مربوط به جمله‌ی تمیز ورودی دارد.

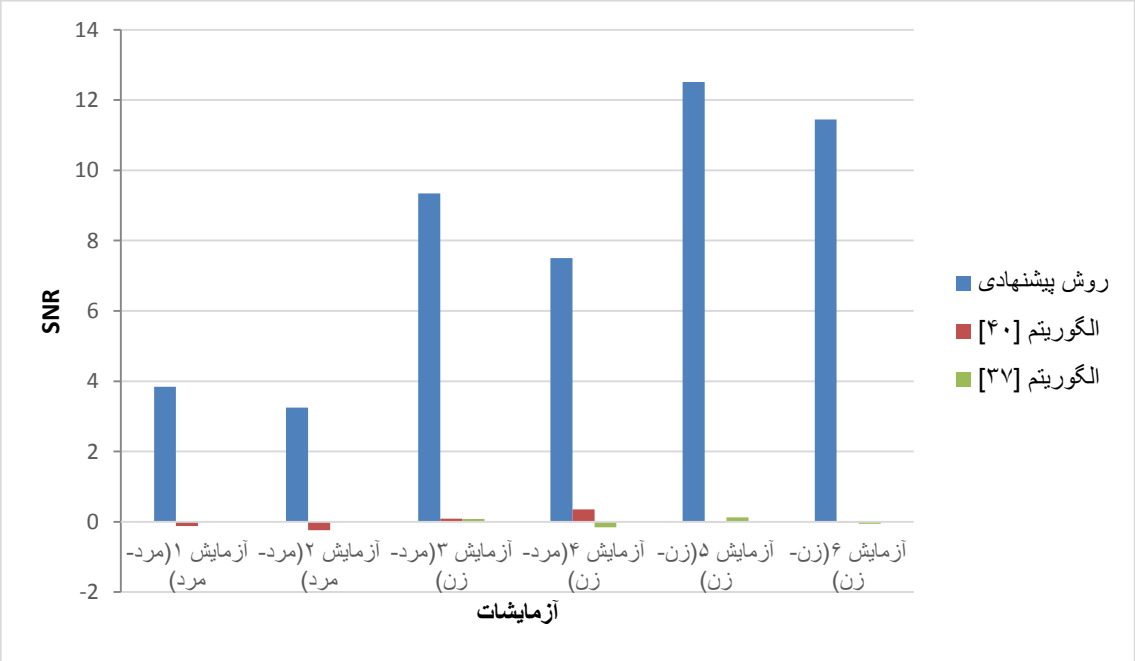


شکل (۵-۸) مقایسه اسپکتروگرام صدای جداسازی شده آزمایش ششم (زن-زن) ، جمله‌ی گفته شده توسط گوینده اول"sbbm9a" و جمله‌ی گفته شده توسط گوینده دوم "swwu2n" می باشد. a) صدای مخلوط b) صدای تمیز اول c) صدای تمیز دوم d) صدای جداسازی شده نخست توسط روش پیشنهادی e) صدای جداسازی شده دوم توسط روش پیشنهادی f) صدای جداسازی شده نخست توسط روش [۴۰] g) صدای جداسازی شده دوم توسط روش [۴۰] h) صدای جداسازی شده نخست توسط روش [۳۷] i) صدای جداسازی شده دوم توسط روش [۳۷]

۵-۴-۴- مقایسه روش پیشنهادی با سایر روش‌ها

در این بخش نتایج حاصل از مقایسه روش پیشنهادی با روش‌های Hu & Wang در [۳۷،۴۰]، به‌ازای انجام هر شش آزمایش آورده شده است. همچنین لازم به ذکر است، برای انجام آزمایش‌های مدل Hu & Wang، از صداهای تست به‌کار برده شده در روش پیشنهادی خود، استفاده کردیم؛ به‌دلیل این‌که به‌صداهای تست استفاده شده در اصل مقاله دسترسی نداشتیم.

در شکل (۵-۹) نمودار مقایسه‌ی SNR خروجی آورده شده است. همانطور که ملاحظه می‌شود، افزایش قابل توجهی در SNR خروجی سیستم پیشنهادی نسبت به [۴۰،۳۷] در تمام آزمایش‌ها دیده می‌شود. در این نمودار کاملاً مشخص است که روش پیشنهادی در تمام مراحل، بسیار بهتر از سایر روش‌ها عمل کرده است.



شکل (۵-۹) نمودار مقایسه SNR روش پیشنهادی در مقایسه با سایر روش‌ها

در جدول (۵-۲)، مقدار عددی معیارهای ارزیابی SNR و MSE در آزمایش‌های مختلف برای مقایسه نشان داده شده است.

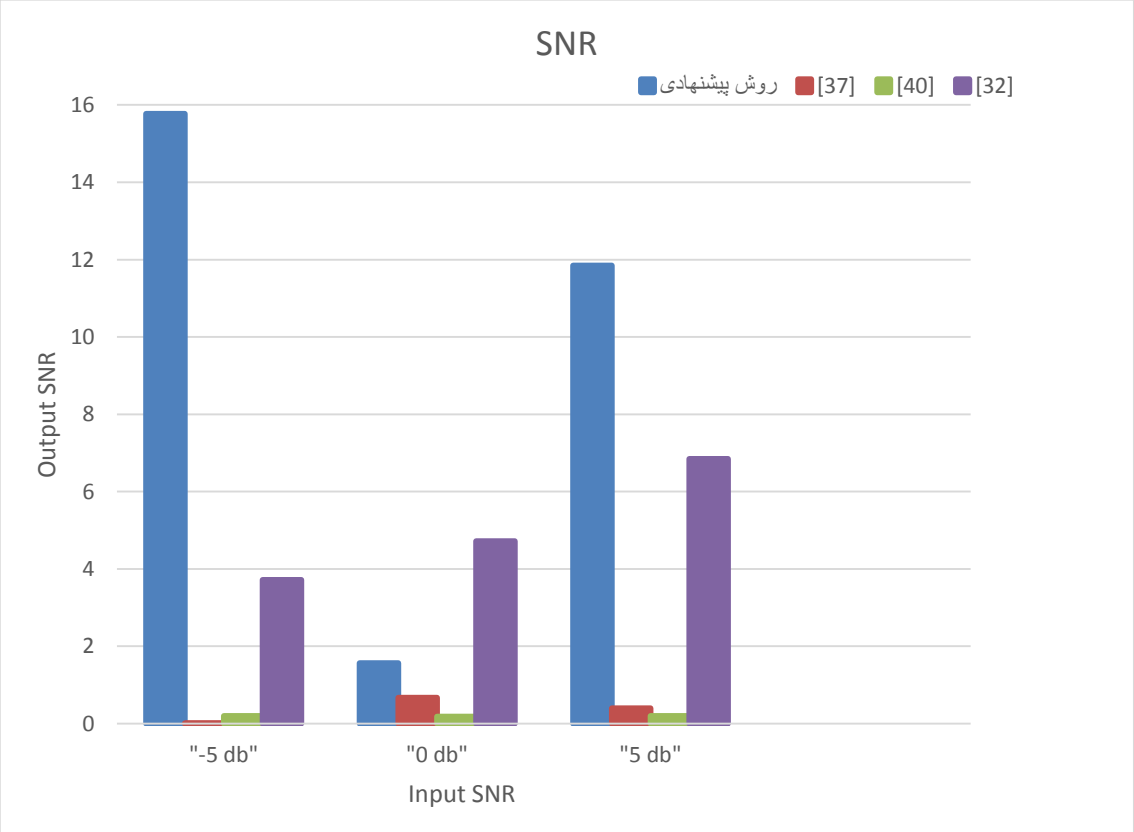
جدول (۵-۲) مقایسه MSE و SNR روش پیشنهادی با سایر روش‌ها

	روش پیشنهادی		الگوریتم [۴۰]		الگوریتم [۳۷]	
	SNR	MSE	SNR	MSE	SNR	MSE
آزمایش ۱(مرد-مرد)	**3.8143**	0.0277	-0.1173	0.0629	-	-
آزمایش ۲(مرد-مرد)	**3.2505**	0.0277	-0.2433	0.0574	0.0105	0.0584
آزمایش ۳(مرد-زن)	**9.3489**	0.0111	0.0882	0.1128	0.0827	0.1124
آزمایش ۴(مرد-زن)	**7.5055**	0.0103	0.3569	0.0633	-0.1571	0.0686
آزمایش ۵(زن-زن)	**12.5181**	0.0119	-0.0140	0.1054	0.1291	0.1050
آزمایش ۶(زن-زن)	**11.4532**	0.0317	-0.0150	0.2388	-0.0529	0.2408

۵-۴-۵- انجام آزمایش‌ها تحت شرایط نویزی

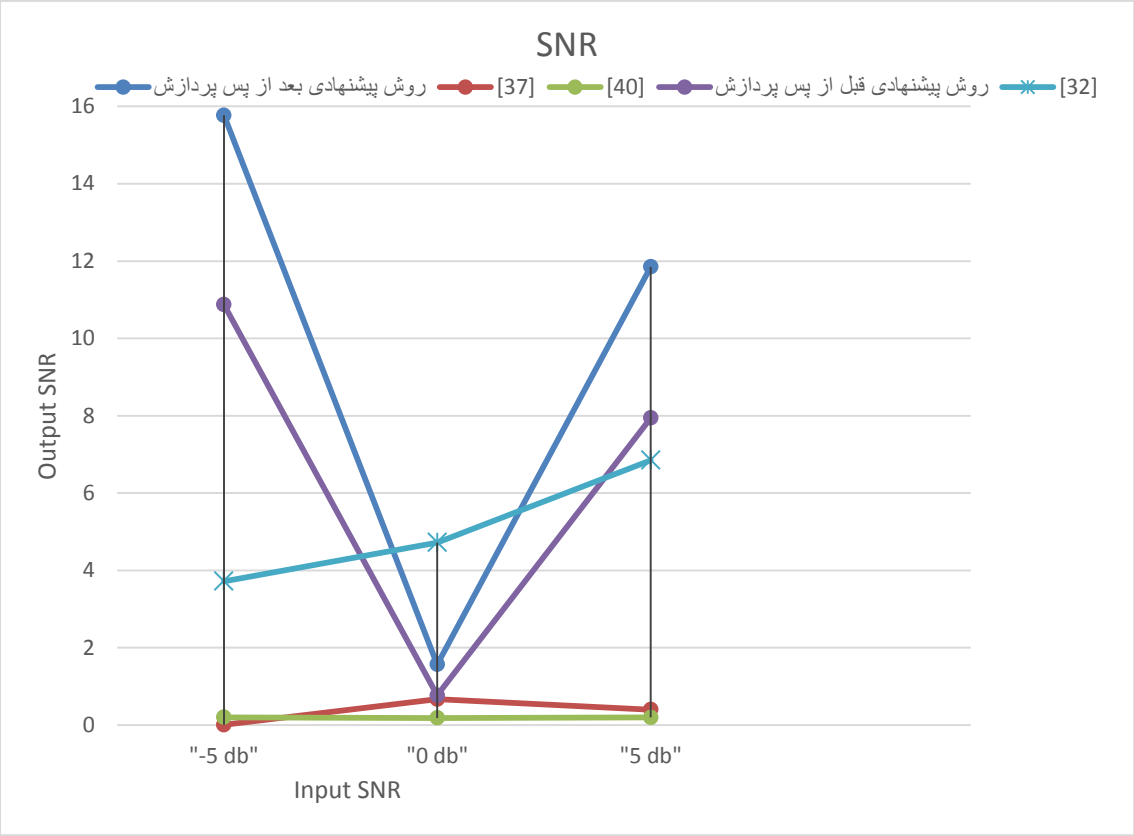
به‌منظور مقایسه رفتار روش پیشنهادی با سایر روش‌ها، در حالتی‌که صدای ورودی نویزی باشد، آزمایش جدیدی را طراحی می‌کنیم. در این آزمایش جمله‌ی گفته شده توسط گوینده اول،"prbc8n" و جمله‌ی گفته شده توسط گوینده دوم، "brir3s" می‌باشد. در این آزمایش گوینده اول، مرد و گوینده دوم، زن می باشد. به صدای مخلوط نویز سفید گوسی با SNR -۵ دسی‌بل ، ۰ دسی‌بل و ۵ دسی‌بل اضافه شده است. در این بخش علاوه بر روش‌های [۴۰،۳۷]، برای مقایسه از مرجع [۳۲] نیز استفاده کرده‌ایم. البته لازم به ذکر است که، این قسمت در آزمایش‌های ما انجام نشده است و فقط از نتایج آن استفاده کرده‌ایم. در [۳۲] برای جداسازی از پایگاه داده Cooke استفاده کرده است؛ درحالی‌که ما از پایگاه داده SSC به‌منظور جداسازی استفاده نمودیم. جملات ادا شده متفاوت می‌باشند. شایان ذکر است که فقط به‌منظور مقایسه و استناد، (البته در دو شرایط متفاوت) این قسمت از مرجع [۳۲] اضافه شده است. مطابق اطلاعات درج شده در [۳۲]، تحت شرایط نویزی -۵ دسی‌بل، ۰ دسی‌بل و ۵ دسی‌بل، نتایج حاصل از SNR خروجی به ترتیب: ۳/۷۲ دسی‌بل، ۴/۷۵ دسی‌بل و ۶/۸۵ دسی‌بل شده است.

نتایج آزمایش تحت شرایط نویزی، قبل و بعد از پس‌پردازش به‌منظور مقایسه روش پیشنهادی با مدل‌های Hu & Wang [۳۷،۴۰] و [۳۲]، در شکل‌های (۵-۱۰) و (۵-۱۱) نشان داده شده است. شکل (۵-۱۰)، نمودار میله‌ای حاصل از نتایج عددی روش پیشنهادی بعد از پس‌پردازش با روش‌های [۳۲،۳۷،۴۰] را نشان می‌دهد.



شکل (۵-۱۰) مقایسه SNR ورودی برحسب خروجی تحت شرایط نویزی (میله‌ای)

در شکل (۵-۱۱)، نمودار مقایسه SNR ورودی نسبت به SNR خروجی (که مطابق با فرمول (۵-۱) به‌دست آمده‌اند)، به‌صورت خطی نشان داده شده است. توجه گردد در این حالت، نتایج روش پیشنهادی قبل از پس‌پردازش نیز در نظر گرفته شده است. همانطور که مشاهده می‌شود، روش پیشنهادی نسبت به سایر روش‌ها از جایگاه بهتری برخوردار است.



شکل (۵-۱۱) نمودار مقایسه SNR ورودی برحسب خروجی تحت شرایط نویزی (خطی)

جدول (۵-۳) نتایج آزمایش این بخش را به‌صورت عددی نشان می‌دهد. با توجه به جدول (۵-۳) مشاهده می‌شود که روش پیشنهادی با توجه به معیارهای MSE و SNR از جایگاه بهتر و بالاتری نسبت به سایر روش‌ها برخوردار است.

جدول (۵-۳) مقایسه MSE و SNR صدای جداسازی شده‌ی روش پیشنهادی با سایر روش‌ها

الگوریتم [۴۰]		الگوریتم [۳۷]		روش پیشنهادی قبل از پس‌پردازش		روش پیشنهادی بعد از پس‌پردازش		Input SNR
MSE	SNR	MSE	SNR	MSE	SNR	MSE	SNR	
0.111	0.201	0.113	0.005	0.056	10.873	0.0125	**15.77**	-5 db
0.110	0.183	0.103	0.668	0.090	0.765	0.0706	**1.56**	0 db
0.112	0.194	0.108	0.40	0.067	7.941	0.0188	**11.85**	5 db

۵-۵- نتیجه‌گیری

در یک مهمانی، ما می‌توانیم به یک صدای خاص توجه داشته باشیم و دیگر صداهای تداخلی موجود در محیط اطراف خود را فیلتر نماییم. این قابلیت ادراکی، باعث ایجاد انگیزه‌ای برای پدید

آوردن یک زمینه مطالعاتی جدید گردید. هدف این زمینه مطالعاتی، طراحی سیستم‌های جداسازی گفتار بر اساس اصول سیستم شنوایی انسان است. در بسیاری از کاربردها نظیر بازشناسی گفتار اتوماتیک و مخابرات راه دور، به یک سیستم موثر که توانایی جداسازی سیگنال گفتار هدف از سیگنال تداخلی را در شرایط تک‌میکروفونه داشته باشد، نیاز می‌باشد.

در این پایان‌نامه یک روش تلفیقی از روش یادگیری آماری و مبتنی بر ویژگی، برای جداسازی صدای دو گوینده‌ی همزمان از یک صدای مخلوط بیان شد. یکی از چالش‌های موجود در این حوزه، جداسازی صدای گوینده‌های هم‌جنس می‌باشد. با توجه به این‌که اکثر روش‌های موجود، از تخمین فرکانس گام برای جداسازی صدا استفاده می‌نمایند و گام گوینده‌های هم‌جنس بسیار نزدیک به‌هم می‌باشد، بنابراین در این حالت روش‌های موجود با مشکل روبرو می‌شوند. در روش پیشنهادی ما محدودیت فرکانس گام وجود ندارد. همانطور که در بخش نتایج شبیه‌سازی دیده شد، روش پیشنهادی قادر است صدای گوینده‌های هم‌جنس را نیز به‌خوبی گوینده‌های غیر هم‌جنس جدا کند. نوآوری روش پیشنهادی، مدل‌سازی صدای گوینده‌های مختلف با استفاده از آموزش دیکشنری است. از مزایای روش پیشنهادی در این پایان‌نامه، محدود نبودن فرکانس جداسازی می‌باشد. در واقع روش موجود، به فرکانس‌های بالا و پایین، وابسته نیست. البته یکی از مشکلات موجود در این پایان‌نامه در دسترس نبودن پایگاه‌های داده با صدای ضبط شده بسیار زیاد از هر گوینده خاص است. به‌دلیل وجود الگوریتم یادگیری و آموزش، ما برای هر گوینده به تعداد زیادی صدای ضبط شده از همان گوینده نیاز داریم که فقط پایگاه داده SSC این شرایط را دارا می‌باشد.

در روش پیشنهادی، از خواص تبدیل فوریه زمان کوتاه برای جداسازی سیگنال صوتی استفاده شده است. روش پیشنهادی یک روش با نظارت می‌باشد. همانطور که توضیح داده شد، در روش پیشنهادی ابتدا نمونه‌های صدای گوینده‌ها مدل‌سازی می‌شود. این مدل‌سازی به‌وسیله‌ی آموزش یک دیکشنری و به‌دست آوردن بردارهای پایه‌ی مربوط به هر کدام از گوینده‌ها انجام می‌شود. بعد از مدل-سازی صدای هر کدام از گوینده‌ها، برای هر صدا، مولفه‌های فرکانسی مربوط به هر کدام از دو گوینده

همزمان جداسازی می‌شود. به‌منظور کاهش نویز صدای خروجی و حذف مولفه‌های فرکانسی سیگنال-های تداخل، یک مرحله‌ی پس‌پردازش نیز روی صدای خروجی انجام می‌شود. نتایج خروجی نشان می‌دهد که روش پیشنهادی، طبق معیار ارزیابی SNR (نسبت سیگنال به نویز) و MSE (خطای میانگین مربعات) برتری محسوسی نسبت به سایر روش‌های مبتنی بر ویژگی و مدل بررسی شده در این پایان‌نامه دارد.

۵-۶- کارهای آتی

- در روش پیشنهادی از تبدیل فوریه زمان کوتاه به‌عنوان استخراج ویژگی در جداسازی استفاده شده است. به‌نظر می‌رسد استفاده از تبدیلات دیگر مثل موجک هم بتواند در این زمینه مفید باشد.
- پایگاه داده‌های آموزشی موجود بسیار کوچک می‌باشند، تهیه‌ی پایگاه داده‌ای بزرگ، به‌منظور آموزش بهتر دیکشنری و جامع‌سازی الگوریتم، ضروری به‌نظر می‌رسد.
- هر چند معیار SNR معمول‌ترین روش مقایسه‌ی عددی در الگوریتم‌های جداسازی صدا می-باشد، اما این الگوریتم قادر نیست تمام خواص دو سیگنال را مقایسه کند. بنابراین ارائه‌ی یک معیار مناسب برای مقایسه روش‌های جداسازی صدا ضروری می باشد.
- در سیستم‌های جداسازی، اصطلاح "هدف" به سیگنال‌های گفتاری نسبت داده می‌شود که، هدف ما جداسازی آن‌ها از بقیه سیگنال‌ها می‌باشد. در عمل، تصمیم‌گیری در مورد این‌که کدام منبع صوتی به‌عنوان منبع هدف در نظر گرفته شود، وابسته به سیستم می‌باشد. حتی هنگامی‌که هدف از قبل نیز مشخص باشد، تصمیم‌گیری در مورد این‌که کدام سیگنال جدا شده متعلق به منبع هدف می‌باشد، برای سیستم‌ها مسأله ساده‌ای نمی‌باشد. به‌عنوان‌مثال اگر هدف، جداسازی سیگنال گفتار یک گوینده از گوینده دیگری باشد، باید مشخص گردد که

کدام گوینده، هدف مورد نظر است. یک موضوع برای تحقیقات آینده به‌کارگیری یک سیستم بازشناسی گوینده برای حل این مشکل است.

- یکی دیگر از موضوعات مهمی که باید قبل از به‌کارگیری سیستم‌های جداسازی گفتار در محیط‌های واقعی مورد بررسی قرار گیرد، انعکاس است. انعکاس باعث ایجاد چالش‌های بسیاری در زمینه جداسازی می‌شود. بنابراین، درنظرگرفتن انعکاس در زمینه جداسازی می‌تواند یک موضوع برای تحقیقات آینده محسوب شود.

مراجع

[1] Benesty, J., Chen, J., Huang, Y., & Dmochowski, J. (2007), "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), 1053-1065.

[2] Ephraim, Y., & Malah, D. (1984), "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6), 1109-1121.

[3] Jensen, J., & Hansen, J. H. (2001), "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, 9(7), 731-740.

[4] Ephraim, Y., Malah, D., & Juang, B. H. (1989), "On the application of hidden Markov models for enhancing noisy speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12), 1846-1856.

[5] Bregman, A. S. (1994), "Auditory scene analysis: The perceptual organization of sound," MIT press.

[6] Hu, G., & Wang, D. (2006), "An auditory scene analysis approach to monaural speech segregation," *Topics in acoustic echo and noise control*, 485-515.

[7] Madhu, N., & Martin, R. (2011), "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 1900-1912.

[8] Roman, N., Wang, D., & Brown, G. J. (2003)," Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, 114(4), 2236-2252.

[9] Wang, D., & Brown, G. J. (2006), *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE Press, NewYork.

[10] Shao, Y., Srinivasan, S., Jin, Z., & Wang, D. (2010)," A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language*, 24(1), 77-93.

[11] Hyvärinen, A., & Oja, E. (2000). "Independent component analysis: algorithms and applications, "*Neural networks*, *13*(4), 411-430.

[12] Jang, G. J., & Lee, T. W. (2003). "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, *4*(Dec), 1365-1392.

[13] Pedersen, M. S., Wang, D., Larsen, J., & Kjems, U. (2005, September). "Overcomplete blind source separation by combining ICA and binary time-frequency masking," In *IEEE International Workshop on Machine Learning for Signal Processing* (pp. 15-20).

[14] Davies, M. E., & James, C. J. (2007). "Source separation using single channel ICA," *Signal Processing*, *87*(8), 1819-1832.

[15] Roweis, S. T. (2000, November). "One microphone source separation," In *NIPS* (Vol. 13, pp. 793-799).

[16] Jordan, F. R. B. M. I. (2005). "Blind one-microphone speech separation: A spectral learning approach," In *Advances in Neural Information Processing Systems 17: Proceedings of the Conference* (Vol. 17, p. 65). MIT Press.

[17] Ellis, D. P., & Weiss, R. J. (2006, May). " Model-based monaural source separation using a vector-quantized phase-vocoder representation," In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 5, pp. V-V). IEEE.

[18] Schmidt, M. N., & Olsson, R. K. (2006, September). "Single-channel speech separation using sparse non-negative matrix factorization," In *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*.

[19] Laurberg, H. (2007, August). "Uniqueness of non-negative matrix factorization," In *IEEE/SP 14th Workshop on Statistical Signal Processing*.

[20] Virtanen, T. (2007). "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 1066-1074.

[21] Benaroya, L., Bimbot, F., & Gribonval, R. (2006). "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 191-199.

[22] Schmidt, M. N., & Laurberg, H. (2008). "Nonnegative matrix factorization with Gaussian process priors," *Computational intelligence and neuroscience*,3.

[23] King, B., & Atlas, L. (2010, March). "Single-channel source separation using simplified-training complex matrix factorization," In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4206-4209). IEEE.

[24] Morgan, D. P., George, E. B., Lee, L. T., & Kay, S. M. (1997). "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Transactions on Speech and Audio Processing*, *5*(5), 407-424.

[25] Stubbs, R. J., & Summerfield, Q. (1990). "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, *87*(1), 359-372.

[26] Eldin, A. M. M., & Youssif, A. A. (2013). "A Hybrid Approach for Co-Channel Speech Segregation based on CASA, HMM Multipitch Tracking, and Medium Frame Harmonic Model," *arXiv preprint:1312.4127*.

[27] Hu, G., & Wang, D. (2004). "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, *15*(5), 1135-1150.

[28] Virtanen, T., & Klapuri, A. (2001). "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on the* (pp. 83-86).

[29] Holdsworth, J., Nimmo-Smith, I., Patterson, R., & Rice, P. (1988). "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, *1*, 1-5.

[30] Mahmoodzadeh, A., Abutalebi, H. R., Soltanian-Zadeh, H., & Sheikhzadeh, H. (2012). "Single channel speech separation in modulation frequency domain based on a novel pitch range estimation method," *EURASIP Journal on Advances in Signal Processing*, (1), 1-10.

[31] Slaney, M., & Lyon, R. F. (1993). "On the importance of time-a temporal representation of sound," *Visual representations of speech signals*, 95-116.

[۳۲] گراوانچی‌زاده، م، ایمانی شاملو، ص، (۱۳۹۳). "جداسازی تک‌گوشی گفتار صدادار مبتنی بر روش‌های جدید انتخاب واحدهای زمان-فرکانس در فرکانس‌های پایین و بالا" *مجله مهندسی برق دانشگاه تبریز*، (۱)۴۳، ۵۱-۶۳.

[33] Weintraub, M. (1985). "A theory and computational model_ of auditory monaural sound separation," (Doctoral dissertation, Stanford University).

[34] Cooke, M. (2005). "Modelling auditory processing and organisation (Vol. 7)," Cambridge University Press.

[35] Brown, G. J., & Cooke, M. (1994). "Computational auditory scene analysis," *Computer Speech & Language*, *8*(4), 297-336.

[36] Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, *60*(4), 911-918.

[37] Hu, K., & Wang, D. (2013). "An unsupervised approach to cochannel speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(1), 122-131.

[38] Hu, G., & Wang, D. (2010). "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 2067-2079.

[۳۹] گراوانچی‌زاده، م، دادور، پ، (۱۳۹۵). " تخمین SNR ورودی با استفاده از ماسک باینری در سیستم‌های مبتنی بر آنالیز ترکیب شنیداری محاسباتی" *مجله مهندسی برق دانشگاه تبریز*.

[40] Hu, K., & Wang, D. (2013). "An iterative model-based approach to cochannel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, (1), 1-11.

[41] Schimmel, S. M. (2007). "Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices," *(Doctoral dissertation, University of Washington).*

[42] Won, J. H., Schimmel, S. M., Drennan, W. R., Souza, P. E., Atlas, L., & Rubinstein, J. T. (2008). "Improving performance in noise for hearing aids and cochlear implants using coherent modulation filtering," *Hearing research*, *239*(1), 1-11.

[۴۳] محمودزاده، آ، (۱۳۹۱)، رساله دکتری: " توسعه روش های مبتنی بر ویژگی برای جداسازی دو گوینده همزمان"، دانشکده برق و کامپیوتر، دانشگاه یزد

[44] Meddis, R., & Hewitt, M. J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *The Journal of the Acoustical Society of America*, *89*(6), 2866-2882.

[45] Lee, Y. K., Kwak, C., Lee, I. S., & Kwon, O. W. (2010, June). "Single-channel speech separation using zero-phase models," In *IEEE International Symposium on Consumer Electronics (ISCE)* (pp. 1-4). IEEE.

[46] Licklider, J. C. R. (1951). "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, *23*(1), 147-147.

[47] Tolonen, T., & Karjalainen, M. (2000). "A computationally efficient multipitch analysis model," *IEEE transactions on speech and audio processing*, *8*(6), 708-716.

[48] Bach, F. R., & Jordan, M. I. (2005, March). "Discriminative training of hidden markov models for multiple pitch tracking [speech processing examples]," In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing,* (Vol. 5, pp. v-489). IEEE.

[49] Klapuri, A. (2008). "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 255-266.

[50] Deller Jr, J. R., Proakis, J. G., & Hansen, J. H. (1993). "Discrete time processing of speech signals," Prentice Hall PTR.

[51] Aharon, M. Elad, A. Bruckstein, M. (2006). "The K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representations," *IEEE Trans. Signal Process, vol. 54, no. 11, pp. 4311-4322*.

[52] Donoho, D.L, Tsaig, Y., Drori, I., & Starck, J.L., (2012). "Sparse Solution of Underdetermined Systems of Linear Equations by Stagewise Orthogonal Matching Pursuit," *IEEE Trans. Inf. Theory, vol. 58, no. 2, pp. 1094-1121*.

[53] Pati, Y. C., Rezaiifar, R., & Krishnaprasad, P. S. (1993, November). "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition", In Signals, Systems and Computers, *Conference Record of The Twenty-Seventh Asilomar Conference on (pp. 40-44). IEEE*.

[54] Cooke, M. Lee, T., Speech Separation Challenge (21 September 2006). [http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparation Challenge.htm]

واژه‌نامه انگلیسی به فارسی

A

Automatic Speech Recognition ..... سیستم بازشناسی خودکار گفتار

Auditory Scene Analysis ..... تحلیل صحنه شنیداری

Autocorrelation Function ..... تابع خودهمبستگی

B

Best-Fitting ..... بهترین جاسازی

Blind Source Separation ..... جداسازی کور منابع

C

Carrier Estimator ..... تخمین‌گر حامل

Cepstral Filter ..... فیلتر کپسترال

Cepstrum ..... کپستروم

Clustering ..... خوشه‌بندی

Cochleagram ..... گوش حلزونی

Cocktail-party ..... کوکتل-پارتی

Complementary Binary Masks ..... ماسک های باینری مکمل

Complex Cepstrum ..... کپستروم مختلط

Computational Auditory Scene Analysis ..... آنالیز محاسباتی نمایش شنوایی

Correlogram ..... همبستگی نگاشت

D

Decreasing Periodicity ..... کاهش یافتن تناوب

E

Energy Sparsity ..... پراکندگی انرژی

Enhanced Envelope Autocorrelation Function ..... تابع خودهمبستگی پوش بهبودیافته

Envelope Autocorrelation Function ..... تابع خودهمبستگی پوش

F

Feature-Based ..... مبتنی بر ویژگی

Formants ..... فورمنت‌ها

Frequency Proximity ..... مجاورت فرکانسی

Full rank ..... رتبه کامل

G

Gammatone Filterbank ..... فیلتربانک گاماتون

Grouping ..... گروه‌بندی

H

Harmonic Magnitude Suppression ..... تکنیک حذف دامنه هارمونیک

Harmonic Sieve ..... الک طیفی

Harmonicity ..... هارمونیک بودن

Hidden Markov Model ..... مدل مارکف مخفی

Hilbert Envelope ..... پوش هیلبرت

I

Ideal Binary Mask ..... ماسک باینری ایده‌آل

Increasing Periodicity ..... افزایش به سمت تناوب

Independent Component Analysis ..... آنالیز مولفه‌های مستقل

Independent Subspace Analysis ..... آنالیز زیرفضای مستقل

Instantaneous frequency ..... فرکانس لحظه‌ای

Iterative Algorithm ..... الگوریتم تکراری

L

Linear Predictive Analysis ..... آنالیز پیش‌گویی خطی

Local SNR ..... نسبت سیگنال به نویز محلی

Local Time ..... زمان محلی

Location ..... محل منبع صدا

M

Magnitude Operator ..... عملگر تخمین دامنه

Mel ..... مل

Model-Based ..... مبتنی بر مدل‌سازی

Multipitch ..... چند گام

N

Noise Bursts ..... نویز ضربه‌ای

Non-negative Matrix Factorization ..... فاکتورگیری ماتریس نامنفی

Nonstationary ..... غیرایستایی

O

Offset ..... فرود

Onset ..... فراز

P

Pitch Detemination Algorithm ..... الگوریتم تعیین گام

Pitch template ..... الگوی گام

Pitch ..... گام

Pretrained ..... از پیش آموزش داده شده

Prior knowledge ..... اطلاعات قبلی

Psychoacoustic Observations ..... مشاهدات روان – آکوستیکی

Pure Tone ..... تک تن

R

Real Cepstrum ..... کپستروم حقیقی

S

Segmentation ..... بخش‌بندی

Segmenter ..... بخش بندی کننده

Short Time Fourier Transform ..... تبدیل فوریه زمان کوتاه

Signal-to-Noise Ratio ..... نسبت سیگنال به نویز

Spatial Filters ..... فیلترهای فضایی

Speak Identification ..... شناسایی گوینده

Spectral Comb ..... شانه طیفی

Statistical Learning ..... یادگیری آماری

Sub-Gaussian ..... زیرگوسی

Summary Correlogram ..... همبستگی نگاشت مختصر

Super-Gaussia ..... فراگوسی

Supervise ..... با نظارت

T

Test ..... تست

Time-Frequency (T-F) Units ..... واحدهای زمان – فرکانس

Time-Frequency ..... زمان – فرکانس

Tone ..... آهنگ

Training ..... آموزش

U

Unsupervised Approach ..... روش بدون نظارت

Unvoiced Speech ..... گفتار بی‌صدا

Utterance-Based Format ..... بر اساس فرمت گفتار

V

SVocal Tract ..... مجرای گفتار

Voiced Speech ..... گفتار صدادار

Z

Zero-crossing ..... محل قطع صفر

Abstract

One of the most important subjects in the field of communication is separating speech by machine that remains as a challenge due to the problems and weaknesses in this system. This is while human hearing system have remarkable abilities in comparison to speech separator machines. Consequently, due to the differences in the operation of speech separator system in machine and human hearing systems, an efficient system for speech separation by machines is indispensable. For instance, one of the disadvantages of automated speech recognizer is its performance decrement in noisy environment. Therefore, such a system have to be equipped with a proper speech separator, not only to improve its performance in different conditions, but also to improve speech quality and reduce non-speech signal transfer cost. As accurate Time-Frequency units selection has a significant impact on separation results, in this thesis we propose a supervised method to select appropriate frequency-time units. In the proposed approach, speakers' speech samples are first modeled via training a dictionary and obtaining fundamental vectors related to each speaker. After modeling two synchronous speakers' speeches, separation operation is done for each speech according to its frequency component. Furthermore, to reduce noise and to remove irrelevant frequency components, a post processing step is also carried out on the output speech.

As can be seen from the systems' output results, our proposed method has a considerable improvement in its operation compared to other feature-based methods.

Keywords- Computational auditory scene analysis (CASA)- cochannel speech separation- supervised segregation.

Shahrood University of Technology

Faculty of Electrical Engineering and Robotics

M.Sc. Thesis in Communication

Speech Separation Of Two Speakers Based on Appropriate Time-Frequency Features

By: **Rana Dehghani**

Supervisor:

Dr. Hossein Marvi

February 2017