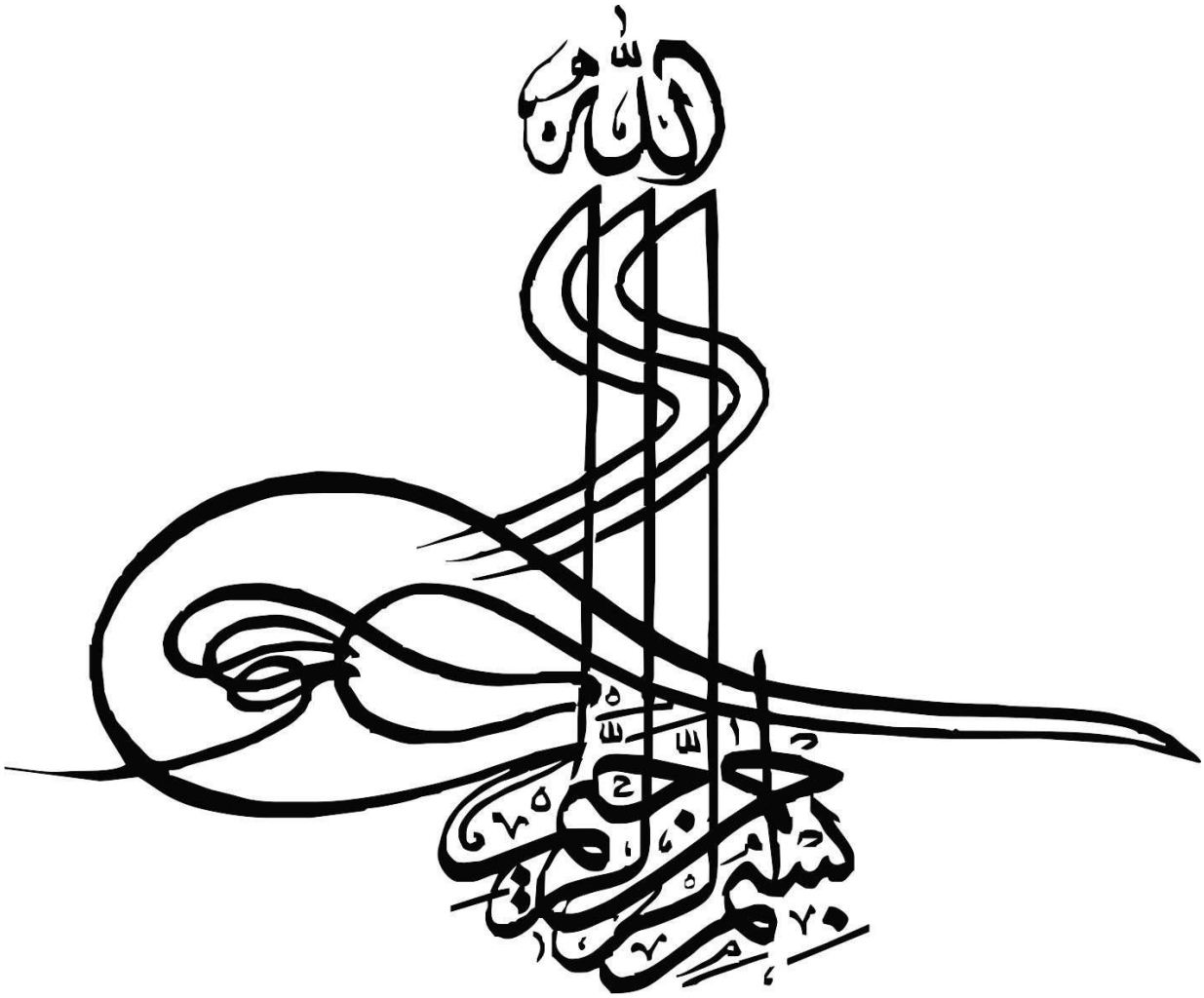

بسم الله الرحمن الرحيم

دانشکده: برق و رباتیک

گروه: الکترونیک

پایان‌نامه دوره‌ی کارشناسی ارشد مهندسی برق- الکترونیک

تعیین محل لحظه ای افراد صحبت کننده در یک اتاق سمینار

محمد حسام محمودی نژاد

استاد راهنما:

دکتر حسین مروی

استاد مشاور:

دکتر علیرضا احمدی فرد

شهریور ۱۳۸۹

تقدیم به پدر و مادرم

که هستی و همه ی وجودم از آنهاست.

تشکر و قدردانی

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

همچنین از دکتر احمدی فرد و دکتر معروضی به خاطر کمک‌های بی‌دریغ‌شان، بسیار متشکرم.

و نیز از تمامی همکاران و دوستانم در دانشگاه صنعتی شاهرود، تشکر می‌نمایم، کسانی که بدون حمایت و حضور آنها، انجام این تحقیق، میسر نبود.

از خانواده‌ام به خاطر حمایت‌های معنوی که در طی انجام این تحقیق و کل زندگی‌ام از من داشته‌اند، تشکر ویژه می‌نمایم.

دانشجو تأیید می نماید که مطالب مندرج دراین پـایان نامه (رساله) نتیجه تحقیقات خودش می باشد و در صورت استفاده از نتایج دیگران مرجع آن را ذکر نموده است.

کلیه حقوق مادی مترتب از نتایج مطالعات ، آزمایشات و نو آوری ناشی از تحقیق موضوع این پایان نامه (رساله) متعلق به دانشگاه صنعتی شاهرود می باشد .

ماه و سال: شهریور ۸۹

چکیده:

با بهبود تکنولوژی میکروالکترونیک و الگوریتم های تعیین محل و ردیابی صحبت کننده، امروزه می‌توان از اینگونه سیستم‌ها، برای مصرف کننده‌ها و بازارهای تجاری بهره گرفت. همچنین علاقمندی فراوانی برای یک کاربرد جدید در اتاق‌های مجهز به سنسورهای مختلف معروف به اتاق کنفرانس هوشمند در حال گسترش است.

در این تحقیق ما به دنبال تعیین محل لحظه‌ای افراد صحبت‌کننده در داخل یک اتاق سمینار می‌باشیم. باید در هر لحظه مشخص کنیم چند منبع فعال وجود دارد و سپس محل این منابع را در فضای فیزیکی تخمین بزنیم. تعیین محل در شرایط مختلفی نظیر: یک یا چند صحبت‌کننده به‌طورهمزمان و اینکه صحبت کننده‌ها به هر دو فرم ساکن و در حال حرکت، باشند و در حالت‌های نویزی مختلف، می‌تواند مورد بررسی قرار گیرد.

در ابتدا با شبکه‌بندی فضای اتاق سمینار و با استفاده از الگوریتم SRP-PHAT محل لحظه‌ای افراد صحبت‌کننده را مشخص می‌نماییم. از اطلاعات جهت سر صحبت کننده‌ها استفاده شده تا از آرایه‌های میکروفنی که در راستای مستقیم با جهت سر صحبت‌کننده قرار دارند استفاده کنیم، و تاثیر انعکاسات گرفته‌شده از سایر میکروفن‌ها را تضعیف نماییم. تابع استفاده شده در این الگوریتم را OPROD-PHAT می‌نامیم. در این بخش یک سیستم برای تعیین محل همزمان چندین صحبت‌کننده پیشنهاد می‌گردد، که در آن، توسط یک الگوریتم دومرحله‌ای ابتدا با استفاده از یک آستانه تطبیقی بر حسب انرژی هر فریم، فریم‌های بی‌صدا را جدا کرده و در ادامه توسط مشخصات همبستگی متقابل بین سیگنال یک جفت میکروفن، در مورد تعیین‌محل لحظه‌ای صحبت‌کننده‌ها تصمیم‌گیری می‌گردد. در ادامه برای کاهش تاثیر نویز زمینه، و افزایش سرعت جستجوی فضای اتاق سمینار و همچنین تعیین محل مناسب‌تر چند صحبت‌کننده به‌طورهمزمان، با بخش‌بندی کردن فضای اتاق، یک میزان فعالیت صوتی را در هر بخش تعیین نموده و در هر فریم زمانی، هدف تعیین بخش‌های شامل صحبت

کننده فعال می‌باشد. به دلیل وجود نویز متغیر، از یک آستانه وفقی با استفاده از الگوریتم EM برای تعیین فعال بودن هر بخش بهره می‌گیریم. سپس با استفاده از یکی از روش‌های غیرمستقیم TDE، به نام روش LI، در بخش‌های فعال محل صحبت‌کننده‌ها را تعیین می‌نماییم. توسط این ایده می‌توان روش‌های دو مرحله‌ای را برای تعیین چندین منبع صوتی به طور همزمان استفاده کرد.

در نهایت به ارزیابی الگوریتم‌های پیشنهادی می‌پردازیم. نتایج حاصله عملکرد مناسب الگوریتم‌های پیشنهادی را در بالا بردن دقت تعیین محل افراد صحبت‌کننده نشان می‌دهند.

کلمات کلیدی:

تعیین محل افراد صحبت‌کننده، آرایه میکروفنی، تخمین تاخیر زمانی (TDE)، اتاق سمینار، روش SRP-PHAT ، جهت سر صحبت‌کننده

فهرست مطالب

عنوان **صفحه**

فهرست *شکل* ها

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست اصطلاحات اختصاری

Term of acronym	Explanation
AED	adaptive eigenvalue decomposition
AMDF	Average magnitude difference function
AR	Autoregressive
ASL	Acoustic Source Localization
BSS	Blind Source Separation
CC	Cross-Correlation
CPS	crosspower spectrum phase
CSSP	Coherent Signal Subspace Processing
DMN	Distributed Microphone Network
DOA	direction of arrival
EKF	extended kalman filter
GCC	Generalized Cross-Correlation
GMM	Gaussian mixture model
GPS	Global positioning system
HI	Hyperbolic Intersection
HMM	Hidden markov model
ICA	Independent component analysis
IID	Independent and Identically Distributed
LS	least squares
MAMDF	modified Average magnitude difference function
MFCCs	Mel-Frequency Cepstral Coefficients
MIMO	Multiple Inputs Multiple Outputs

ML	maximum likelihood
MVDR	minimum variance distortionless response
PHAT	Phase Transform
PDF	probability density function
PLP	perceptual linear prediction
RADAR	Radio detection and ranging
SAD	Speech activity detection
SCG	Scaled conjugate gradient
SNR	Signal to noise
SNS	Speech non speech detection
SOE	Source Orientation Estimation
SONAR	Sound navigation and ranging
SRP	Steered Response Power
SVM	Support vector machine
TDE	Time Delay Estimation
TDOA	Time Delay Of Arrival
UART	universal asynchronous receiver/transmitter
UKF	Unscented Kalman Filter
VAD	Voice activity detection

فصل اول

مقدمه

۱-۱- مقدمه

در ابتدای این فصل اهداف و انگیزه‌های انجام این تحقیق آورده‌شده و سپس برخی از کاربردهای تعیین محل منبع صوتی مطرح می‌گردد. در ادامه تعاریف و مفاهیم پایه را مورد بحث قرار می‌دهیم و در نهایت ساختار کلی این تحقیق را مطرح می‌نماییم.

۱-۲- اهداف و انگیزش

جوامع امروزی به شدت متکی به کار با ماشین‌های مختلف است و مسیر تکنولوژی آینده در جهت خودکار نمودن تمامی سیستم‌ها پیش می‌رود. کلید ارتباط آسان بین انسان و ماشین استفاده از سیستم‌هایی است که در آن‌ها کاربر نیازمند نگه داشتن و یا پوشیدن قسمتی از تجهیزات ارتباطی، نباشد.[1]

بدین منظور از آرایه‌های میکروفنی بجای میکروفن‌های یقه ای[2] در سیستم‌های دریافت صوت استفاده می‌گردد؛ اما افزایش فاصله میان کاربر و میکروفن، باعث تضعیف سیگنال و تاثیر تداخل و نویز زمینه بروی آن می‌گردد، که می‌توان از سیستم‌های ترکیبی، که شامل الگوریتم‌های تخمین محل صحبت‌کننده و بهبود سیگنال صحبت است، برای رفع این مشکل یاری جست.

تعیین محل منابع صوتی در چند دهه گذشته، برای کاربردهای مختلفی مورد بررسی قرار گرفته است. در سالهای اخیر با رشد تکنولوژی و تمایل به خودکار انجام دادن کارهای مختلف، مساله آنالیز فعالیت‌های داخل اتاق کنفرانس بسیار مورد علاقه بوده است.

در این تحقیق ما به دنبال تعیین محل لحظه ای افراد صحبت کننده در داخل یک اتاق سمینار می‌باشیم. می‌بایست در هر لحظه مشخص کنیم چند منبع صوتی فعال وجود دارد و سپس محل این

[1] Hand free
[2] Lapels microphone

منابع را در فضای فیزیکی تخمین بزنیم. تعیین محل در شرایط مختلفی نظیر: یک یا چند صحبت کننده به طور همزمان و اینکه صحبت کننده‌ها به هر دو فرم ساکن و در حال حرکت باشند و در شرایط نویزی مختلف، می‌تواند مورد بررسی قرار گیرد. معمولا نویز ناشی از صداهای آمده از بیرون اتاق کنفرانس و یا صدای وسایل الکترونیکی نظیر فن کامپیوتر، پروجکتور و ... می‌باشد.

تعیین محل منبع صوتی در چند دهه گذشته برای بسیاری از کاربردها مورد بررسی قرار گرفته است، با پیشرفت تکنولوژی دیجیتال، حالتی خاصی از منابع صوتی یعنی مساله تعیین‌محل یک صحبت‌کننده در چند سال اخیر در داخل اتاق‌های هوشمند مطرح گردیده است و در ادامه تلاش‌هایی برای شرایطی جامع‌تر یعنی تعیین محل چندین فرد صحبت کننده در مقالات مورد بررسی قرار گرفته است.

در این حالت برخی از مشکلات اصلی تعیین محل و ردیابی صحبت کننده‌ها عبارتند از:

- همپوشانی صوتی: زمانی که چندین فرد با هم در حال صحبت کردن می‌باشند، مطرح می‌شود، که در [۱] در مورد مباحث مربوطه بررسی‌هایی صورت گرفته است.
- در اکثر مواقع تغییر سریع صحبت کننده‌ها را داریم که در جلسات و کنفرانس‌ها بسیار اتفاق می‌افتد، که عمل تعیین‌محل صحبت‌کننده را مشکل‌تر می‌سازد.
- همچنین اظهارات کوتاه در بین جملات افراد (معمولا کوتاه تر از یک ثانیه) اکثر تکنیک‌های ردیابی صحبت کننده را با مشکل مواجه می‌سازند.

۱-۳- کاربردها

تخمین موقعیت افراد یا اجسام کاربردهای بسیاری دارد، که در ادامه برخی از آنها معرفی می‌گردند:

- روباتیک: روبات‌ها از موقعیت خود برای دنبال کردن مسیر مورد نظر استفاده می‌کنند و همچنین با داشتن اطلاعات موقعیت محل منبع صوتی برای ارتباط و ردیابی حرکت افراد و سایر منابع صوتی بهره می‌گیرند.[۲]
- سیستم سیستم موقعیت‌یابی جهانی[۱] (GPS) برای هدایت وسایل نقلیه مختلف استفاده می‌گردد، سیستم‌های نظامی مانند رادار و سونار همچنین در سیستم‌های امنیتی در خیابان‌ها، برای تصادفات ماشین‌ها و یا شلیک گلوله و... از مشاهدات صوتی برای هدایت دوربین به سمت محل منبع صوتی استفاده می‌شود. مسایل امنیتی در[۳] بررسی شده است.[۲]
- تعیین محل صحبت کننده برای کاربردهایی نظیر سمعک[۳] ، [۴] و هدایت خودکار دوربین به سمت فرد در حال صحبت در یک اتاق سمینار به طوری که دوربین در مرکز میز داخل اتاق کنفرانس واقع شده و افراد در اطراف میز نشسته‌اند. برای این کاربرد نرخ بروزرسانی فریم‌ها از ۲۰۰-۳۰۰ms قابل قبول می‌باشد، در اکثر این موارد نیازمند سیستمی برای تعیین‌محل دقیق افراد صحبت کننده می‌باشیم. اگر این افراد در حال صحبت‌کردن حرکت کنند، باید بتوان از میزان حرکت آنها در هر فریم برای بررسی محل آنها چشمپوشی کرد. بنابراین طول هر فریم باید به اندازه ای باشد که صحبت کننده‌ها در آن ثابت فرض شوند و بعبارت دیگر کاهش اندازه فریم باعث افزایش دقت تعیین محل صحبت کننده‌ها می‌گردد.

از ترکیب آرایه میکروفنی و دانش پردازش سیگنال برای بهبود کیفیت سیگنالهای صوتی و شناخت صحبت استفاده نمود، برای مثال می‌توان سیگنال صحبت مربوط به یک گوینده را بهبود داده و سیگنالهای مربوط به سایر گوینده‌ها و منابع نویزی را حذف نمود. در برخی از کارها نیازمند سیستمی برای تمرکز بروی منبع صوتی خاص می‌باشیم و انجام هدایت الکترونیکی این کار، نیاز انجام دستی را

[1] Global positioning system

[۲] تمامی اصطلاحات اختصاری در فهرست مربوطه تعریف شده است.

[3] hearing aids

منتفی می‌سازد. از همین سیستم برای واحد شناخت صحبت راننده در یک ماشین به منظور انجام برخی ازفعالیت‌ها با فرمان صوتی داخل یک ماشین در [۵] مورد بررسی قرار گرفته است.

در سالهای اخیر تیم‌های تحقیقاتی بسیاری در زمینه آنالیز کنفرانس، محدوده گسترده‌ای از فعالیت‌ها را بررسی می‌کنند. این تیم‌ها اتاق‌های هوشمند مجهز به سنسورهای مختلف مانند انواع آرایه‌های میکروفنی و دوربین‌های ویدیویی دارند.

از میان این تیم‌های تحقیقاتی می‌توان موارد زیر را نام برد:

CHIL[۱],NIST[۲],AMI[۳],M4[۴],…

اکثر این تیم‌ها با همکاری وسرمایه‌گذاری کشورهای مختلف تشکیل شده اند، مثلا تیم CHIL خود ۱۵همکار از ۹ کشور دارد برخی از فعالیت‌های مربوط به این تیم عبارتند از:

- تعیین محل و ردیابی افراد در فضای سه بعدی به فرم صوتی و ویدیویی
- تشخیص جهت سر صحبت کننده
- تشخیص فعالیت صوتی[۵] و جداسازی صحبت از غیر صحبت
- Speaker diarization: که در آن باید مشخص گردد چه کسی چه موقع در حال صحبت‌کردن می‌باشد.
- شاخص‌گذاری و خلاصه سازی کنفرانس
- شناخت اتوماتیک صحبت

و سایر فعالیت‌ها که در حیطه کار این تحقیق قرار نمی‌گیرد.

[1] http://chil.server.de
[2] http://www.nist.gov/smartspace
[3] http://www.amiproject.org
[4] http://www.m4project.org
[5] Speech activity detection

تعیین‌محل افراد صحبت کننده پایه عملیات زیادی است و کاربردهای زیادی را می‌توان به همراه نتایج حاصل از تعیین محل مطرح نمود. در ادامه به طور خلاصه به برخی از این فعالیت‌ها اشاره خواهیم نمود:

- ردیابی صوتی: محل‌های تخمین زده شده اغلب شامل خطاهایی هستند که ناشی از انعکاس‌های صوتی و نویز زمینه و یا ضعف خود الگوریتم تعیین‌محل می‌باشند. اگر منبع‌صوتی متحرک باشد این خطاها بیشتر هم می‌شوند، و در کل استفاده از فیلترهای ردیابی باعث کاهش میزان خطا می‌گردد که شامل فیلتر کردن محل‌های تخمین زده شده به صورت لحظه‌ای می‌باشد. در ردیابی افراد صحبت‌کننده، از به هم پیوستن محلهای لحظه‌ای تخمین زده شده برای هر فرد صحبت‌کننده در مراحل قبلی برای دسته بندی کردن عبارات حرفی در طول زمان وتعیین مسیر حرکت صحبت کننده‌ها استفاده می‌گردد.
- پس از ردیابی می‌توان عمل صحبت از غیر صحبت[1] را برای هر دسته صوتی انجام داد، که بعد از این مرحله تنها دسته‌هایی که شامل گفتار هستند نگه داشته می‌شوند.
- تقسیم‌بندی صحبت[2]: که شامل تشخیص و جداسازی زمان صحبت و خاموشی برای هر صحبت کننده می‌باشد. این عملیات بروی دسته‌های ایجاد شده از مرحله ردیابی، اجرا می‌گردد.

در نگاه اول تقسیم‌بندی صحبت کار ساده ای به نظر می‌رسد. یک پیشنهاد ساده مقایسه انرژی در هر فریم صوتی است، اما در عمل جملات خیلی کوتاه هستند و در سمینارها اکثراً به طور همزمان صحبت می‌شود. علاوه بر این مسایل نویز غیر صحبتی نظیر حرکت بدن و صندلی، ورق زدن صفحات کتاب، صداهای ناشی از فن کامپیوتر و پروجکتور و... نیز وجود دارد. در عملیات تقسیم‌بندی صحبت نیازداریم اطلاعات کاملی از موارد زیر داشته باشیم:

[1] Speech non speech (SNS)
[2] Speech segmentation

(۱) در هر لحظه چند صحبت کننده به‌طور همزمان درحال صحبت می‌باشند. (۲) و نیز موقعیت قرارگیری این افراد در اتاق سمینار مشخص گردد، تا بتوان با دقت بیشتری عملیات تقسیم‌بندی صحبت را انجام داد.

- دسته‌بندی صحبت‌کننده‌ها[1]: در نهایت عمل دسته‌بندی صحبت‌کننده‌ها مطرح می‌گردد. در مرحله ردیابی صحبت‌های افراد در دسته‌های کوتاه قرار گرفته و در این مرحله با استفاده از الگوریتم‌هایی مثل MFCCs مشخص می‌شود که هر دسته متعلق به کدام فرد می‌باشد و به دسته‌های مختلف برچسب مربوط به فرد صحبت کننده خود را دریافت می‌کند.
- سیستم‌های صوتی- تصویری: که در آنها افراد در یک اتاق سمینار همواره توسط الگوریتم‌های مختلف ردیابی تصویری، دنبال می‌شوند. زمانی که افراد صحبت کنند، عمل ردیابی صوتی نیز اجرا می‌گردد، و هنگامی که افراد مختلف از نظر تصویری جلو هم قرار می‌گیرند، اطلاعات ناشی از ردیابی صوتی می‌تواند بسیار کمک کننده باشد. در سال‌های اخیر تلاش‌های زیادی در ردیابی صوتی –تصویری صورت گرفته است.

۱-۴- تعاریف و مفاهیم پایه در ASL

در این بخش مفاهیم پایه و تعاریف ریاضی مربوط به تعیین محل منبع صوتی[2] بررسی می‌شود.

۱-۴-۱ موقعیت یابی فعال و غیر فعال[3]

موقعیت یابی منبع صوتی به دو نوع فعال و غیر فعال تقسیم می‌شود. در روش فعال، مانند رادارها، سیگنالی از گیرنده به هدف فرستاده شده و بر اثر تغییرات حاصل شده در سیگنال دریافتی از هدف،

[1] Speaker clustring
[2] Acoustic source localization (ASL)
[3] Active- passive localization

موقعیت آن تشخیص داده می‌شود. اما در روش پسیو سیگنالی برای تشخیص هدف منتشر نمی‌شود و موقعیت‌یابی براساس تغییرات سیگنال منتشرشده از هدف، در گیرنده‌ها صورت می‌گیرد.

روش پسیو به‌دلیل عدم ارسال سیگنال و مخفی بودن در سیستم‌های امنیتی و نیز به دلیل بی‌خطر بودن، در سیستم‌های پزشکی کاربرد بیشتری دارد. برای کار تعیین‌محل منبع صوتی نیز معمولا از روش‌های پسیو استفاده می‌گردد.

۲-۴-۱ انواع میکروفن

میکروفن‌های یقه‌ای: این امکان را می‌دهند که زمان دقیق شروع صحبت هر صحبت کننده را بدانیم و سیگنال تمیز و بهتری را نسبت به آرایه‌های میکروفنی داشته باشیم، اما محدوده کاربردهای کمتری دارد چراکه:

- هر فرد نیازمند یک میکروفن مجزا است
- آنها هیچ اطلاعاتی در مورد محل افراد صحبت کننده به ما نمی‌دهند.

آرایه‌های میکروفنی: به معنای فاصله بیشتر از ۳۰-۴۰cm بین صحبت کننده و میکروفن‌ها می‌باشد. در اغلب موارد از میکروفن‌های همه جهتی[1] استفاده می‌گردد. فرم هندسی موجود برای آرایه میکروفنی به صورت خطی و دایره ای تی-شکل و... می‌باشد.

برای استفاده صحیح در دریافت اطلاعات صوتی گرفته شده توسط جفت میکروفن‌ها در کار ASL، فاصله میکروفن‌ها نسبت به هم باید کمتر از نصف طول موج صوت باشد، $d < \lambda/2$، و ماکزیمم فرکانس نمونه‌برداری میکروفن‌ها هم می بایست برابر باشد با: $f_{\max} = c/2d$.

در روابط فوق d فاصله بین میکروفن‌ها، λ طول‌موج صحبت، c سرعت صوت را نشان می دهند.

[1] Omni directional microphones

۱-۴-۳ مراحل مختلف پروسه ASL

مساله ASL به چهار مرحله تقسیم می‌شود: ساطح شدن امواج، انتشار امواج، دریافت امواج، و الگوریتم‌های مختلف ASL

در ادامه هر بخش بررسی خواهیم نمود:

- *ساطح شدن موج*: سیستم‌های ASL معمولا با توجه به مشخصات نظیر پهنای باند و فرکانس مرکزی و مدل منبع صوتی طراحی می‌شوند. سیگنال صحبت از نوع باند پهن بوده و مدل منبع صوتی معمولا به فرم نقطه ای در نظر گرفته می‌شود.
- *انتشار امواج*: صوت بر اثر ارتعاش میکانیکی ذرات بوجود آمده و برای انتشار به هوا نیاز دارد. انتشار به دو فرم میدان دور[۱] و میدان نزدیک[۲] تقسیم می‌شود که در میدان دور، فاصله میکروفن‌ها تا منبع زیاد می‌باشد و انتشار امواج به صورت صفحه ای [۳] صورت می‌گیرد. در حالت میدان نزدیک فاصله میکروفن‌ها از منبع کم است و موج به صورت کروی به سمت آرایه میکروفنی حرکت می‌کند. در فرم میدان دور تنها جهت دریافت صوت[۴] قابل تشخیص است، در حالی در میدان نزدیک محل دقیق منبع نیز قابل تشخیص می‌باشد. اگر فاصله منبع تا آرایه میکروفنی r باشد و طول آرایه خطی L باشد و طول موج را هم با λ نشان دهیم در این صورت فرض میدان دور بودن با توجه به رابطه زیر مشخص می‌شود:

$$r > \frac{2L^2}{\lambda} \qquad (۱-۱)$$

ASL کاربردهایی در داخل و خارج اتاق دارد که در هر دو مورد امواج توسط نویز زمینه و انتشارات چند مسیری، مخدوش می‌شوند. در داخل اتاق چند مسیری شدن امواج بر اثر انعکاس دیواره‌ها و

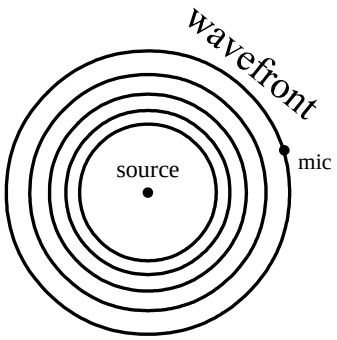
[1] Far-field
[2] Near-field
[3] Plane wave
[4] Direction of arrival (DOA)

اجسام بزرگتر از طول موج بوجود می‌آید. انعکاسات می‌تواند به فرم آینه بازتاب کنند و یا اینکه پخش‌شدگی امواج در جهات مختلف را داشته باشیم و باعث بوجود آمدن چندین تاخیر در سیگنال دریافتی توسط میکروفن‌ها گردد.

انعکاسات به ضریب جذب سطوح بستگی دارد، که خود آن نیز وابسته به عواملی چون زاویه‌برخورد، فرکانس موج و همچنین خصوصیات نوع ماده سطح، می‌باشد.

از طرف دیگر تضعیف امواج در هوا به عواملی نظیر دما، رطوبت و فرکانس موج بستگی دارد. سرعت صوت در هوا برابر $c = 20.1\sqrt{T}$ خواهد بود. مثلا در دمای $19^{\circ}C$سرعت صوت برابر ۳۴۲ m/s می‌باشد.[۹]

انتشار کروی امواج در شکل (۱-۱) نشان داده شده است.



شکل ۱-۱:انتشار امواج کروی از منبع صوتی نقطه‌ای به سمت میکروفن‌ها

مکان منبع و میکروفن‌ها را به صورت زیر نشان می‌دهند:

$$l^{ps} = [x^{ps}, y^{ps}, z^{ps}]^T$$
$$l_m = [x_m, y_m, z_m]^T \qquad (۲-۱)$$

همان طور که گفته شد در محیط واقعی، انعکاسات و جذب امواج از دیوارها پوست بدن، لباس و... وجود دارد، که در حیطه مدل‌های انعکاسی قرار داشته و هنوز باز برای تحقیق می‌باشد. با این وجود تاثیر انعکاسات به اندازه‌ای است که در تخمین موقعیت منبع در کنار دیواره‌ها همواره با خطای بیشتری همراه می‌باشد.

فرض مدل آزاد[1] در مطالعات بسیار معمول بوده، بدین معنی که فرض می‌شود که با هیچ مانع (میز و صندلی) و بازتاب (دیوار و...) مواجه نیستیم.

معمولاً از اثر دوپلر نیز چشم پوشی می‌شود، چراکه منبع صوتی می تواند متحرک باشد حال آنکه سرعت آن در برابر سرعت صوت قابل اغماض است.

سیگنال دریافتی توسط میکروفن‌ها به فرم زیر نشان داده می شود:

$$x_m(t) = A(l^{ps}, l_m).x^{ps}[t - TOF(l^{ps}, l_m)] \qquad (۳-۱)$$

$A(l^{ps}, l_m)$ فاکتور بهره دامنه و $TOF(l^{ps}, l_m)$ زمان پرواز موج از منبع به سمت میکروفن می‌باشد، و در ادامه معرفی می‌گردند.

در حوزه فرکانس برای سیگنال گرفته‌شده توسط میکروفن‌ها در هر فریم زمانی، داریم:

$$X_m^{(t)}(k) = A(l^{ps}, l_m).X^{(ps,t)}(k).e^{-j\pi\frac{k-1}{N_F}TOF(l^{ps}, l_m)} \qquad (۴-۱)$$

فاکتور بهره دامنه: اگر ما جذب انرژی صوتی در فضا را نادیده بگیریم کل انرژی انتقال یافته توسط امواج کروی ثابت می باشد . مساحت کره با شعاع آن متناسب می باشد با توجه به این خواهیم داشت:

[1] Free space

$$A(l^{ps}, l_m) = \frac{A_1}{\left\| l^{ps} - l_m \right\|} \quad (۵-۱)$$

که در آن A_1 یک مقدار ثابت مثبت و $\| \ \|$ نرم می باشد.

زمان پرواز [1]: با فرض موج کروی این زمان با فاصله میان میکروفن و منبع متناسب می‌باشد.

$$TOF(l^{ps}, l_m) = \frac{\left\| l^{ps} - l_m \right\|}{c} . f_s \quad (۶-۱)$$

- *دریافت توسط آرایه میکروفنی*: یک مبدل فشار صوت را به ولتاژ تبدیل کرده و سپس یک مبدل آنالوگ به دیجیتال، ولتاژ را دیجیتالی می‌کند. انواع میکروفن در بخش قبلی مطرح شد.
- *الگوریتم‌های تعیین محل*: در نهایت با توجه به نوع منبع صوتی، میزان نویز و انعکاسات زمینه و نحوه انتشار موج، الگوریتم مناسب تعیین محل منبع صوتی انتخاب می‌گردد. در فصل بعد، انواع الگوریتم‌های ASL بررسی می‌گردد.

در ادامه بحث مفاهیم پایه و ریاضیات مربوطه به بحث ASL را مطرح خواهیم نمود، به‌طوری‌که در فصل‌های بعدی در اجرای الگوریتم های مختلف از این مباحث استفاده می‌گردد.

۱-۴-۴ مفاهیم پایه

جفت میکروفن: اگر N_m میکروفن داشته باشیم، تعداد جفت میکروفن‌های ما برابر خواهد بود:

$$N_q = \frac{N_m \times (N_m - 1)}{2} \quad (۷-۱)$$

برای مثال در مورد جفت q خواهیم داشت: $x_{a_q}(t), x_{b_q}(t)$

[1] Time of flight

عملیات فریم کردن: الگوریتم‌های تعیین محل منبع‌صوتی بروی هر فریم به طور جداگانه اجرا می‌شود و همین‌طور که فریم‌ها در زمان پیش می‌روند قادر به دنبال کردن منبع صوتی خواهیم بود. سیگنالهای دریافتی توسط میکروفن‌ها به فریم‌هایی به طول L نمونه تقسیم می‌شوند.

$$x_{m,b}[n] = w[n]x[bA+n] \qquad \text{For M} = 1\ldots\text{m}, \text{N} = 1\ldots\text{L}-1 \qquad (۸-۱)$$

b تعداد فریم‌ها و A یک عدد ثابت مثبت است که پیشروی فریم را مشخص می‌کند . اگر A<L باشد فریم‌ها همپوشانی خواهند داشت و معمولاً A را برابر L/2 انتخاب می‌کنند، که معادل همپوشانی ۵۰٪ می‌باشد. یک انتخاب معمول برای نوع پنجره، همینگ خواهد بود که گلبرگ اصلی آن دو برابر وگلبرگ کناری آن db ۲۵ تضعیف بیشتر نسبت به پنجره مستطیلی خواهد داشت.

استفاده از این پنجرهایی که درانتهای خود باریک هستند[1] ،تاثیر ناپیوستگی در انتهای پنجره را از بین برده و تخمین طیفی بهتری را به‌همراه خواهد داشت.

تبدیل DFT: برای سیگنال میکروفن m و فریم b به صورت زیر تعریف می‌شود:

$$X_{m,b} = \sum_{n=0}^{L-1} x_{m,b}[n] e^{-jk\frac{2\pi}{k}n} \qquad 0 \le k \le K-1, m = 1..M \qquad (۹-۱)$$

توجه کنید که طول DFT برابر k است و $L \le k$ می‌باشد و زمانی‌که نیاز به صفرگذاری[2] باشد، K از L بزرگتر انتخاب می‌گردد. این کار برای رفع مشکلات ناشی از شیفت دایره‌ای استفاده می‌شود. گفتار در دوره‌های ms ۳۰ به فرم شبه‌ایستان فرض می‌شود . سیگنال پیوسته با فرکانس Fs نمونه‌برداری شده (با در نظر داشتن محدودیت نایکویست) و به یک سری از فریم‌های زمانی گسسته تقسیم می‌شود. هر فریم زمان گسسته شامل 2Nf نمونه می‌باشد:

$$X^{(t)} = [x^{(t)}(1)....x^{(t)}(n)....x^{(t)}(2N_F)]^T \qquad (۱۰-۱)$$

[1] tapered window
[2] Zero-padding

عملیات فریم کردن با استفاده از پنجره همینگ و با پیش تاکید ۰.۹۷ بروی ظرایب اجرا می‌شود:

$$x^{(t)}[n] = [0.54 - 0.46(\pi \frac{n-1}{N_F})].[x(t - N_F + n) - 0.97.x(t - N_F + n - 1)] \quad (۱۱-۱)$$

برای آنالیز DFT ما از ۵۰٪ همپوشانی میان فریم‌ها استفاده می‌کنیم تا دیتا با لبه‌های یک بلوک تنظیم شود با استفاده از پنجره نازک‌سازی شده (شبیه پنجره همینگ)که در مرکز بعدی قرار می‌گیرد و به تمام دیتاها وزنهای معادلی را خواهد داد، این کار صورت می‌گیرد.

برای آنالیز طیفی از DFT زمان کوتاه[1] استفاده می‌گردد. در هر فریم زمانی بردار DFT به صورت رابطه (۱-۱۱) نشان داده می‌شود.

$$X^{(t)} = [X^{(t)}(1)....X^{(t)}(k)....X^{(t)}(2N_F)]^T \quad (۱۲-۱)$$

صفرگذاری: برای جلوگیری از پدیده شیفت دایره‌ای معمولاً نیازمند عملیات صفرگذاری می‌باشیم، که در آن فریم‌های زمانی را با تعداد معادلی از صفرها به هم پیوند می‌دهیم.

(۱۳-۱)

$$\begin{cases} x^{(t)}[n] = [0.54 - 0.46(\pi \frac{n-1}{N_F})].[x(t - N_F + n) - 0.97.x(t - N_F + n - 1)] & if\ 1 \le n \le N_F \\ 0 & if\ .N_F \le n \le 2N_F \end{cases}$$

مواقعی که نیاز به صفرگذاری داریم مقدار N_f دو برابر در نظر گرفته می شود که نمونه‌های بدست آمده برای $x(t)$ با سایر مواقع تفاوتی نداشته باشد.

[1] Short-time windowed

تابع همبستگی در حوزه زمان[1] و فرکانس:

رولبط مربوط به توابع همبستگی در حوزه زمان و فرکانس به‌صورت زیر نشان داده می‌شود:

$$c_{x_1,x_1}(\tau) = \int_{\Re} x_1(\zeta).x_2(\zeta-\tau)d\zeta \qquad (۱۴-۱)$$

$$C_{x_1,x_2}(w) = \int_{\Re} c_{x_1,x_1}(\tau)e^{-jw\tau}d\tau a \qquad (۱۵-۱)$$

با ترکیب دو معادله فوق و استفاده از تعریف تبدیل فوریه داریم:

$$C_{x_1,x_2}(w) = X_1(w).(X_2(w))^* \qquad (۱۶-۱)$$

بنابراین در حوزه زمان برای تابع همبستگی متقابل، خواهیم داشت:

$$c_{x_1,x_1}(\tau) = \frac{1}{2.\pi}\int_{\Re} X_1(w).(X_2(w))^* e^{jw\tau}dw \qquad (۱۷-۱)$$

برای اطلاعات بیشتر از مباحث ریاضیاتی مطرح شده در این بحث می توانید به[۷] و برای مباحث پیشرفته‌تر، به [۸] و [۹] مراجعه نمایید.

۱-۵- ساختار پایان‌نامه

این تحقیق شامل پنج فصل می‌باشد: مفاهیم پایه و تعریف مساله در فصل جاری مورد بررسی قرار گرفت. در فصل دوم روش‌های مختلف تعیین محل منابع صوتی به‌طورکامل معرفی می‌گردد. اخیراً از روش پارتیکل فیلتر برای تقویت عملکرد تعیین محل منابع صوتی و ردیابی منابع‌صوتی استفاده شده است، که درپایان این فصل به طور خلاصه مورد بررسی قرار گرفته است.

[1] Continues time domain cross correlation function

در فصل سوم با انتخاب روش تخمین تاخیر زمانی[1] (TDE) برای تعیین محل منابع صوتی، به معرفی کامل این روش می‌پردازیم. سپس با تاکید بروی روش‌های مستقیم TDE، روش SRP-PHAT[2] را به عنوان روش اصلی تعیین محل صحبت کننده‌ها انتخاب می‌کنیم.

در فصل چهارم الگوریتم‌های پیشنهادی برای تعیین محل یک یا چند صحبت‌کننده را مطرح می‌کنیم. برای بهبود عملکرد تعیین محل صحبت‌کننده‌ها، تابعی به نام OPROD-PHAT معرفی شده، که در آن از اطلاعات جهت سر صحبت کننده برای تاکید بیشتر بروی جفت میکروفن‌هایی که در مسیر مستقیم جهت‌سر صحبت‌کننده قرار گرفته اند، با وزن بیشتری نسبت به سایر جفت میکروفن‌ها استفاده می‌گردد. با شبکه‌بندی فضای اتاق سمینار و با استفاده از توابع معرفی‌شده محل لحظه‌ای افراد صحبت‌کننده را در هر فریم‌زمانی مشخص می‌نماییم. از دو دیتابیس مختلف شامل چندین توالی برای انجام عملیات استفاده شده است. استفاده از اطلاعات جهت سر صحبت باعث می‌شود، که عملکرد بسیار مناسب‌تری در برابر انعکاسات داشته باشیم. در این بخش یک سیستم برای تعیین محل همزمان چندین صحبت‌کننده پیشنهاد می‌گردد، در یک الگوریتم دو مرحله‌ای ابتدا با استفاده از یک آستانه تطبیقی بر حسب انرژی هر فریم، فریم‌های بی‌صدا[3] را جدا کرده و در ادامه توسط مشخصات همبستگی متقابل بین سیگنال‌های یک جفت میکروفن، در مورد تعیین محل لحظه‌ای صحبت‌کننده‌ها تصمیم‌گیری می‌گردد.

برای افزایش سرعت و بهبود عملکرد در برابر نویز و در مساله تعیین محل همزمان چندین صحبت‌کننده، با تقسیم کردن فضای اتاق سمینار به چندین بخش متقارن، از الگوریتم‌های معرفی شده در بالا، برای تعیین میزان فعالیت صوتی در هر بخش استفاده نموده و سپس تعیین محل تنها در بخش های فعال اجرا می‌گردد.

[1] Time delay estimation (TDE)
[2] Steered response power- phase transform
[3] silence

با به‌کارگیری روش LI جهت قرارگیری صحبت‌کننده، تعیین می‌گردد. استفاده از یک آستانه تطبیقی برای تعیین اینکه در هر فریم کدام بخش فعال است، با استفاده از الگوریتم EM صورت می‌گیرد. در نهایت هم نتایج الگوریتم های پیشنهادی را بررسی خواهیم نمود.

در فصل پنجم، مرور کلی و نتیجه گیری بر این تحقیق و همچنین پیشنهادات برای کارهای آینده را خواهیم داشت.

فصل دوم

مروری بر روشهای تعیین محل منابع صوتی

۲-۱- مقدمه

در این فصل مروری بر روش‌های رایج در زمینه تعیین محل منابع صوتی خواهیم داشت و با توجه به مشخصات و کارایی‌های مربوط به هر روش، در فصل‌های آتی از روش‌های TDE، که یکی از مناسب‌ترین روش‌ها برای کار تعیین محل لحظه‌ای افراد صحبت کننده می‌باشد را در یک اتاق مجهز به آرایه‌ای از میکروفن‌ها انتخاب خواهیم کرد؛ و با جزییات بیشتری این روش‌ها را مورد بررسی قرار می‌دهیم. درانتهای این فصل، پارتیکل فیلتر را به عنوان یک روش مناسب در ردیابی صوتی، معرفی خواهیم کرد و در نهایت دریافت صوت با استفاده از پردازشگر DSP را مورد بررسی قرار می‌دهیم.

۲-۲- روشهای ASL

پنج روش کلی برای تعیین محل منابع صوتی وجود دارد، که در ادامه آنها را بررسی خواهیم نمود.

۲-۲-۱ روش‌های تخمین طیفی

تعیین محل صحبت‌کننده با استفاده از تخمین طیفی با وضوح بالا در سال‌های اخیر مورد توجه بوده است. تعدادی از تخمین‌های طیفی معروف عبارتند از: مدل Autoregressive (AR) ، تخمین طیفی مینیمم واریانس (MV)، و الگوریتم MUSIC .[۱۰]

اساساً این روش‌ها برای تخمین طیف فرکانس‌ها و مشخصات مربوط به آن به‌کار گرفته می‌شوند. علاوه بر آن برای سیگنال‌های باند باریک و تخمین طیفی برای تخمین DOA با استفاده از آرایه میکروفنی استفاده می‌شود. در این مورد تعیین‌محل با فرض یک منبع صوتی و نویز ایستا، انجام می‌گیرد. با به‌کارگیری ماتریس همبستگی سیگنال‌های ورودی از هر میکروفن، از تجزیه مقادیر ویژه برای جداسازی ماتریس همبستگی نویز و سیگنال استفاده می‌شود. روش جستجو برای تخمین دقیق

DOA نیازمند محاسبات پیچیده‌ای است. علاوه بر آن برای تخمین طیفی، باید دانش کاملی ازمشخصات کانال داشته باشیم.

۲-۲-۲ روشهایی بر پایه انرژی

حالتی را در نظر بگیرید که نویز زمینه قابل چشم پوشی است، منبع صوتی ایزوترپیک فرض شده و تضعیف امواج صوتی ساطح شده از منبع با فاصله رابطه عکس دارد. در این روش گیرنده‌ها توان سیگنال را اندازه‌گیری می‌کنند و آن را بر اساس توان دریافتی معلوم و تلفات انتشار موثر به فاصله تبدیل می‌نمایند. معمولا در این روش نیاز است که تلفات انتشار موثر قبلاً اندازه‌گیری گردد و به عنوان اطلاعات معلوم به گیرنده داده شود. نسبت انرژی سیگنال دریافتی توسط یک جفت میکروفن با توجه به فاصله آنها نسبت به منبع صوتی، مجموعه ای از نقاط را برای محل منبع صوتی تشکیل می‌دهد که به شکل یک دایره خواهد بود. سه میکروفن دو نقطه به ما می‌دهد که ناشی از تقاطع دو دایره می‌باشد، و در نهایت وجود چهار میکروفن محل دقیق منبع صوتی را مشخص می‌سازد.

این روش تنها در شرایط SNR پایین کاربرد دارد.

۲-۲-۳ ASL با استفاده از پاسخ ضربه

مسیر طی شده موج صوتی از منبع به سمت میکروفن‌های مختلف، بستگی به محل منبع و جذب و انعکاس امواج در اتاق دارد.

اگر مشخصات پاسخ ضربه اتاق بخوبی مشخص باشد، از طریق یک پروسه کالیبره کردن موقعیت منبع صوتی قابل استنتاج خواهد بود. اما در عمل کالیبره کردن کار بسیار مشکل و نامطلوبی می‌باشد. بنابراین پاسخ ضربه نیاز به تخمین زدن دارد، که به این کار تشخیص کانال MIMO کور[1] می‌گویند و نیازمند محاسبه مدل کاملی از اتاق دارد که هنوز این مساله باز برای تحقیق می‌باشد.[۱۱]

[1] Blind Multiple Inputs Multiple Outputs

۴-۲-۲ Beamforming

به روش‌هایی که در آنها آرایه‌ها بروی محل یا جهت مشخصی از فضا متمرکز می شوند، در کل Beamforming گفته می شود. زمانی که محل منبع مشخص نیست، یک Beamformer می‌تواند برای پویش یا هدایت یک ناحیه از پیش تعیین شده توسط تغیرات روی تاخیرهای موجود عمل کند. زمانی که از این روش استفاده می‌کنیم، خروجیBeamformer را پاسخ هدایت شونده می‌گوییم. توان پاسخ هدایت شونده[1] (SRP)، ممکن است مقادیر ماکزیمم زیادی داشته باشد ولی در شرایط مطلوب، مقادیر ماکزیمم در هنگام تمرکز در جهت خاص معادل محل یک منبع صوتی می‌باشد. اغلب در زمان پردازش Beamforming فیلترهایی تطبیق داده می شوند تا سیگنال منبع را بهبود داده و سایرین را تضعیف کند. و بنابراین این فیلترها نقش حذف نویز زمینه را به عهده دارند. و زمانی که Beamformer برای تعیین محل منابع بکار برده می شود، وظیفه فیلترها تقویت توان سیگنال منبع مطلوب در خروجی Beamformer در زمان تمرکز آرایه بروی آن می‌باشد.

ساده ترین نوع پاسخ هدایت شونده از خروجی یک Delay-and-sum beamformer بدست می‌آید Delay-and-sum beamformers یک شیفت زمانی را بروی آرایه ای از سیگنالها اعمال کرده تا زمان تاخیر انتشار سیگنال‌ها از منابع صوتی به هر میکروفن را جبران کند. این سیگنالها با هم جمع می‌شوند تا یک سیگنال تک خروجی را تشکیل دهند. همان طور که گفته شد بیشتر Beamformerها علاوه بر این از یک فیلتر هم استفاده می‌کنند و در کل filter-and-sum beamformers را تشکیل می‌دهند.

یکی از روشهای بهینه Beamforming روش (MVDR)[2] می‌باشد که از فرکانس‌های وابسته به کواریانس ماتریس تخمینی از سیگنال‌ها، استفاده می‌کند.[۱۲]

اما اگر سیگنال بین فریم‌های همسایه ایستا نباشد، مانند سیگنال صحبت، این تخمین با مشکل مواجه

[1] Steered response power(SRP)
[2] minimum variance distortionless response

خواهد بود [۱۳]. برای تخمین کواریانس ماتریس در سیگنال‌های باند پهن، می‌بایست پهنای باند را به زیر باندهایی تقسیم کنیم، که بعد از تخمین ماتریس، این زیرباندها با هم ترکیب می‌گردند.

۵-۲-۲ ASL بر پایه روشهای TDE

رایج ترین روش‌های ASL روشهایی بر پایه TDE است که برای سیگنال‌های باند پهن، مانند سیگنال صحبت مناسب می‌باشد. زمان پرواز موج صوتی از منبع صوتی نسبت به میکروفن‌های مختلف در یک آرایه با توجه به محل قرار گرفتن میکروفن‌ها، مختلف می‌باشد؛ اغلب روشهای تعیین‌محل از اختلاف کوچکی که موج صوتی در رسیدن به میکروفن‌ها دارد، استفاده می‌کنند؛ به‌طوری که با مقایسه میان سیگنال‌های دریافتی میکروفن‌های مختلف در یک آرایه میکروفنی و در نظر داشتن فاصله میان هر دو جفت میکروفن، با توجه به چیدمان هندسی میکروفن‌ها، می‌توان محل لحظه ای چندین منبع صوتی را مشخص نمود. بنابراین روشهایی که به این غیر همزمانی بودن استناد می‌کنند می‌بایست از موقعیت دقیق هر یک از میکروفن‌ها در آرایه میکروفنی اطلاع داشته باشند، اما در مورد فرم اتاق نیاز به اطلاعاتی نمی‌باشد.

روشهای تعیین محل میدام نزدیک بر پایه TDE به دو بخش روش‌های مستقیم و روشهای غیرمستقیم یا دومرحله‌ای تقسیم می‌شوند. در روشهای TDE دو مرحله ای از تخمین زمان تاخیر ورود[1] (TDOA) در مرحله اول استفاده کرده و تاخیر زمانی ورود سیگنال به جفت میکروفن‌ها بدست می‌آید. مرحله بعدی تعیین محل منبع صوتی است که تکنیک‌های زیادی برای این کار وجود دارد و در بخش‌های بعدی مورد بررسی قرار خواهند گرفت.

یکی ازمشکلات روشهای دو مرحله ای این است که تخمین اشتباه تاخیر زمانی در محیط‌های پرانعکاس و نویزی در مرحله اول موجب افزایش خطا در مرحله بعدی‌که تعیین محل منبع صوتی

[1] Time delay of arrival

است، می‌گردد. و همچنین استفاده از اطلاعات یک فریم زمانی برای روش‌های دومرحله‌ای TDE پیشنهاد نمی‌شود، چراکه به مقدار بیشتری از دیتا نیاز داریم و بنابراین از متوسط گیری بروی چندین فریم متوالی استفاده می‌گردد، که این مساله در برابر طبیعت متغیر گفتار و همچنین حرکت انسان در حال صحبت ایجاد مشکل می‌نماید.

در روشهای مستقیم یا یک مرحله ای مستقیماً از اندازه‌های TDE بدست آمده و در فضای فرکانس تعیین‌محل منابع صوتی انجام می‌شود. این روش‌ها در برابر نویز و انعکاسات عملکرد بهتری را از خود نشان می‌دهند. در کل مزیت روشهای دومرحله‌ای به روشهای مستقیم، زمان کوتاه محاسباتی آنها و عیب آنها ناتوانی آنها در اداره مسایل مربوط به چندین منبع به طور همزمان است.

البته اخیراً روشهایی برای استفاده تکنیک‌های غیرمستقیم در تعیین‌محل همزمان چندین منبع صوتی پیشنهاد شده است. که بر پایه جداسازی منابع کور[1] (BSS) می‌باشد.[۱۴]

BSS به جداسازی سیگنال‌های چندین منبع از یک مجموعه سیگنال مختلط، بدون داشتن دانشی درباره خود سیگنال‌های منبع می‌پردازد. سیگنال دریافتی توسط یک میکروفن می‌تواند مجموعه‌ای از نسخه‌های فیلتر شده سیگنال‌های انتشار یافته باشد. هدف اصلی BSS بازسازی هر کدام از این سیگنال‌های انتشار یافته می‌باشد. با تشخیص پاسخ ضربه کانال و پروسه فیلترینگ، تاخیر زمانی برای هر جفت میکروفن و هر منبع می‌تواند از این فیلترها بدست آید.

آنالیز مولفه های غیر مستقل[2] (ICA) یک تکنیک رایج برای این هدف می‌باشد.

در حالت میدان دور تعیین محل با توجه به مقادیر DOA یک یا چند آرایه میکروفنی بدست می‌آید. در فضای باز مساله با توجه به محدودیت سرعت انشار صوت،پیچیده تر می‌شود.

[1] Blind Source Separation
[2] Independent component analysis

۲-۳- معرفی روشهای TDOA

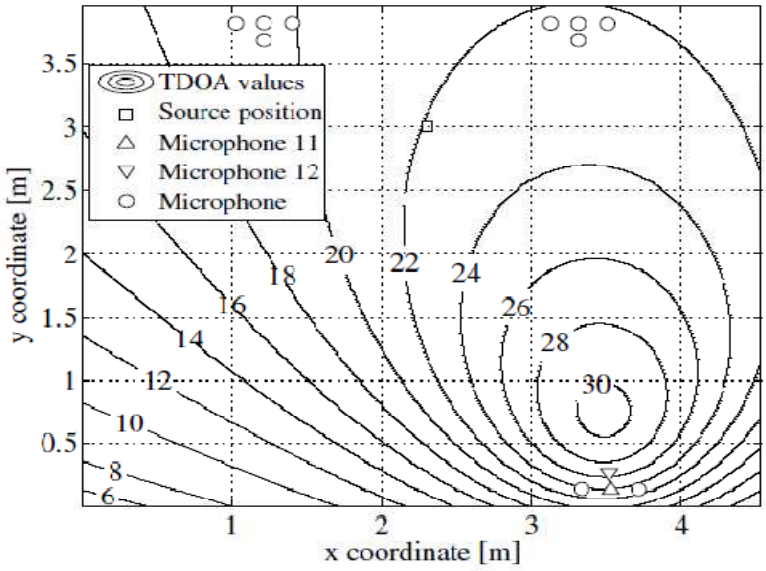
در این بخش به معرفی روشهای TDOA موجود خواهیم پرداخت. قسمت اول تمامی روشهای TDE احتیاج به محاسبه تاخیر زمانی ورود سیگنال بین جفت میکروفن‌ها دارند.

اگر منبع در L^{PS} واقع شده باشد و میکروفن‌هایی را در L_1, L_2 داشته باشیم، TDOA تئوری به فرم زیر تعریف می‌شود:

$$\tau_q^{th}(L^{PS}) = TOF(L^{PS} - L_{a_q}) - TOF(L^{PS} - L_{b_q}) \qquad (۱-۲)$$

روش TDOA به آرایه‌ای از میکروفن‌ها احتیاج دارند. در این روش، یکی از گیرنده‌ها به عنوان گیرنده مبنا درنظرگرفته شده و اختلاف زمان دریافت سیگنال در سایر گیرنده‌ها با این گیرنده محاسبه می‌شود. ابتدا می‌بایست TDE هر یک از جفت میکروفن‌ها از طریق آنالیز DFT بدست آورده و سپس به تعیین محل منابع با در نظر داشتن محل دقیق میکروفن‌ها بپردازیم.

افزایش تعداد تخمین‌های TDOA موجب می‌شود که از یک میانگین آماری برای تعیین محل گوینده استفاده شود، اما انعکاس محیط هم تا اندازه زیادی در آن تاثیر می گذارد.

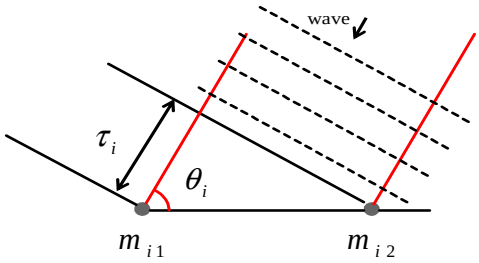


شکل ۱-۱ : TDOA میان یک جفت میکروفن در ارتفاع ۱.۱۲۸ m در فضای دوبعدی نشان داده شده است.

برای هر جفت میکروفن مقادیر محاسبه شده برای TDOA به صورت صفحات هیپربولیک نشان داده می‌شود. محل منبع بروی یکی از این صفحات قرار می‌گیرد و با استفاده از چندین جفت میکروفن در روشهای مختلف، محل دقیق منبع توسط تقاطع این صفحات هیپربولیک مشخص خواهد شد. در فضای سه بعدی با استفاده از چهار میکروفن محل منبع مشخص می‌شود. در شکل ۱-۱ مقادیر TDOA میان یک جفت میکروفن نشان داده شده است.

اگر حالت میدان دور را درنظر بگیریم، موج به فرم صفحه‌ای انتشار یافته و هیپربولیک‌ها می‌توانند به فرم خط راست تخمین زده شوند که شیب خط نشانگر جهت یا زاویه رسیدن سیگنال می‌باشد. رابطه میان τ_i, θ_i و به عبارتی DOA, TDOA به صورت زیر تخمین زده می‌شود:

$$\theta_i = \arccos\left(\frac{c\tau_i}{\|m_{i1} - m_{i2}\|}\right) \qquad (۲-۲)$$



در شکل ۱-۲ رابطه بین TDOA,DOA برای حالت far-field نشان داده شده است.

تعیین محل و دنبال کردن چندین صحبت کننده به دلیل طبیعت دینامیک سیگنال گفتار و حرکت افراد در حال صحبت، نیازمند نرخ بروز رسانی ۲۰-۳۰ ms برای فریم‌های زمانی می‌باشد. و زمانی که فریم‌ها تا این اندازه کوچک شوند، انعکاسات تاثیر زیادی بروی دیتاها می‌گذارد به طوری که

تکنیکهای زیادی برای بهبود عملکرد پیشنهاد شده است که در ادامه بررسی خواهند شد.

از جمله معروف‌ترین روش‌های تعیین TDOA بین جفت میکروفن‌ها، برپایه استفاده از تابع همبستگی بین سیگنال‌های یک جفت میکروفن استوار است، که به تابع GCC-PHAT معروف می‌باشد. پیک ماکزیمم تابع همبستگی جایی است که دو سیگنال حداکثر شباهت را نسبت به‌هم دارند؛ از این خاصیت برای تعیین میزان تاخیر ورود سیگنال بین هر دو جفت میکروفن استفاده می‌گردد.

کلیه روش‌های تخمین میزان تاخیر و روش‌های مستقیم TDE به طور کامل در فصل سوم بررسی می‌گردند. در ادامه با فرض بدست‌آوردن میزان تاخیر بین جفت‌میکروفن‌ها، باید در مرحله دوم محل دقیق منبع صوتی را تخمین بزنیم. در بخش بعدی به‌طور خلاصه به معرفی روش‌های موجود می‌پردازیم.

۲-۳-۱ تعیین محل closed-form بر پایه TDOA

همان‌طورکه در بخش‌های قبلی آورده شد، با استفاده از روش‌های مختلف TDOA تاخیر زمانی بین جفت میکروفن‌های مختلف بدست می‌آید؛ سپس با استفاده از مقادیر بدست آمده از مرحله قبل به معرفی روش‌هایی می‌پردازیم که برای تعیین محل منابع صوتی بکار می‌روند.

فرض کنید یک میکروفن به عنوان مرجع در $M_0 = [0,0,0]$ قرار گرفته است،و M میکروفن دیگر در محل‌های مشخص m_k و.منبع در محل نامشخص r قرار دارد.

$$\Delta\tau_k \cdot c = d_k = \|r - m_k\| - \|r\| \qquad (۲-۳)$$

مقدار $\Delta\tau_k$ می‌تواند از تخمین TDOA بدست آید. با ساده سازی رابطه فوق داریم:

$$\|r - m_k\|^2 = (d_k - \|r\|)^2$$
$$d_k \|r\| + m_k{}^T r = b_k \qquad (۴-۲)$$
$$b_k = \frac{\|m_k\| - d_k}{2}$$

با ادامه بحث در[۱۵] به معیار حداقل مربعات می‌رسیم:

$$J = \sum_{k=1}^{M} (d_k \|r\| + m_k^T r - b_k)^2 \qquad (۵-۲)$$

با مینیمم کردن معادله فوق در مرجع آورده شده، محل منبع با استفاده از روش Unconstrained LS بدست می‌آید.

بهبودهای زیادی در مقالات برای روش LS صورت گرفته است:

- در [۱۶] با معرفی روش Extended Unconstrained LS روش قبلی را با چندین میکروفن مرجع در نظر می‌گیرد، چراکه در نظر گرفتن مقادیر TDOA میان میکروفن‌های غیر مرجع و وجود اندازه‌های بیشتر موجب بهبود تخمین می‌گردد.
- در[۱۷] تحت روشی به نام pre-multiplying میکروفن مرجعی در نظر گرفته نمی‌شود و معادلات برای جفت میکروفن {i,j} نوشته می‌شود.
- Approximate LS در [۱۵] که در آن با استفاده از بسط سری تیلور معادلات را خطی کرده و سپس با انتخاب یک حدس اولیه و روش بازگشتی، الگوریتم را تا جایی ادامه می‌دهند که جواب به مقدار خاصی همگرا شود. در این روش مشکلاتی همچون حدس اولیه مناسب و همگرایی جواب وجود دارد.
- Weighted constrained LS در [۱۸]. بررسی شده است.
- Constrained LS در [۱۹].مورد بررسی قرار گرفته است.

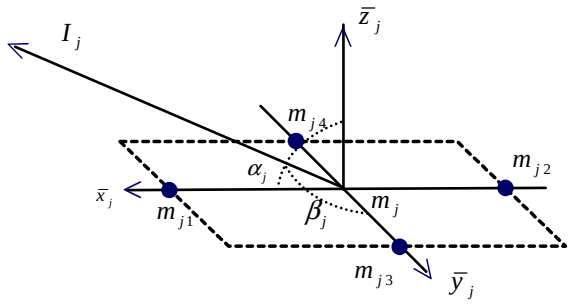
- Spherical interpolation(SI) و سپس بهبود آن با تصحیح خطی در [۲۰] مطرح شده است.
- Spherical Intersection (SX) در [۲۱] مورد بحث قرار گرفته است.

در روش maximum likelihood فرض می‌شود که مقادیر TDOA از جفت میکروفن‌ها نسبت به هم مستقل هستند و توسط نویز گوسی مخدوش شده اند. با توجه به تابع چگالی احتمال داده‌ها نسبت به موقعیت نامعلوم منبع صوتی، در نهایت با مینیمم کردن تابعی که مجموع مربعات خطا بین مقادیر TDOA و منبع نامعلوم را دارد، به محل منبع صوتی می‌رسیم.

روش‌های پیشنهادی توسط Fang و Chan، از حل دقیق معادلات، بدون عمل خطی‌سازی، برای تخمین مکان منبع استفاده می‌کنند که البته اشکال روشی چون روش Fang، عدم امکان استفاده از تأخیرهای اضافی در محاسبه مکان است. در واقع روش Chan، معروف به تکنیک HI در [۲۲] معرفی شده است.

هر کدام از تخمینگرها بر پایه LS با مینیمم کردن یک اندازه خطا، که یک تابع غیرخطی بر حسب محل منبع صوتی می‌باشد، درگیر هستند. در ادامه قصد داریم روش LI را که توسط[۲۳] معرفی شده است، را بررسی کنیم. این روش بر این پایه بنا شده است که اگر TDOA جفت‌میکروفن‌ها مشخص است، محل منبع توسط هیپربولیک قابل تشخیص است. اگر فرض شود که منبع در فاصله دور از میکروفن‌ها قرار دارد، این هیپربولیک می‌تواند توسط یک مخروط با زاویه و راس زیر نشان داده شود:

$$\theta_i = \cos^{-1}\left(\frac{c\tau_i}{\left\|x_{i,1} - x_{i,2}\right\|}\right)$$
$$m_i = \frac{(x_{i,1} - x_{i,2})}{2} \qquad (۶-۲)$$



شکل ۱-۳: تخمین هیپربولیک با مخروط و اشتراک مخروط‌های فرضی با خط I_j نشان داده شده است.

همانطور که درشکل ۲-۳ دیده می‌شود تقاطع خط واصل جفت میکروفن‌ها در نقطه m_i است که می‌توان یک سیستم مختصات کارتزین محلی در این نقطه به فرم زیر در نظر گرفت:

$$X = \frac{x_{i,1} - x_{i,2}}{\left\| x_{i,1} - x_{i,2} \right\|}$$
$$Y = \frac{x_{i,3} - x_{i,4}}{\left\| x_{i,3} - x_{i,4} \right\|} \qquad (۷-۲)$$
$$Z = X \times Y$$

از TDOA اولین جفت میکروفن یک مخروط با جهت ثابت، زاویه α_j وابسته به محورX و به همین ترتیب زاویه β_j وابسته به محور Y را خواهیم داشت. تقاطع دو مخروط برای دو جفت میکروفن آورده شده یک خط خواهد بود:

$$I_j = r_j a_j + m_j \qquad for \quad m = 1...M \qquad (۸-۲)$$

که در آن پارامتر a برابر است با :

$$a = \begin{pmatrix} \cos(\alpha_j) \\ \cos(\beta_j) \\ \cos(\gamma_j) \end{pmatrix} \quad (۹-۲)$$

در روابط فوق r_j فاصله از یک نقطه بروی خط می‌باشد. محل منبع صوتی می‌تواند در محیط‌های بدون نویز از تقاطع این خطوط بدست آید؛ اما در محیط‌های نویزی این خطوط معمولا اشتراک نخواهند داشت. زوایای آورده شده در رابطه (۲-۹) با استفاده از رابطه (۲-۶) برای هر یک از جفت میکروفن‌ها بدست می‌آید.

در ادامه بحث سایر روش‌های بر پایه TDE را مورد بررسی قرار می‌دهیم.

۲-۴- رهیافت آماری ASL

به دلیل اینکه رهیافت آماری قابلیت این را داردکه زمانی که انعکاسات غالب هستند، به طور موثر با آنها مواجه شود، در سال‌های اخیر مورد علاقه بوده است. همچنین در برابر نویز قوی‌تر عمل کرده و حساسیت کمتری به خطاهای ناشی از مکان قرارگیری میکروفن‌ها دارد.

در[۲۴] یک کلاسیفایر TDOA طراحی شده و با استفاده از جهت دریافت سیگنال تابع پگالی احتمال مربوط به TDOA تخمین زده می‌شود. کلاسه‌بندی سپس برای مقایسه مدل توزیع احتمالی با یک توزیع احتمالی گرفته‌شده از سیگنال‌های ورودی، انجام می‌شود.

رهیافتی بر پایه HMM برای مدل‌کردن و کلاسی‌بندی کردن مسیر حرکت منبع در[۲۵] صورت گرفته است.

۲-۵- تعیین محل ترتیبی بر پایه TDE

اگر منبع صوتی در میان فریم‌های متوالی، فعال باشد و حرکت آن قابل مدل کردن باشد؛ از روش‌های تخمین ترتیبی[1] می‌توان برای بهبود عملکرد ASL استفاده نمود. به‌طوری‌که از اطلاعات فریم‌های‌زمانی گذشته، برای تعیین محل فریم حال استفاده می‌شود؛ ایده اصلی این است که، نویز با پیک غالب در فریم حال، لزوماً در تخمین محل توسط این روش، نمی‌تواند غالب باشد .

فرض کنید که X_t نشانگر متغیر حالت در فریم زمانی t مطابق با رابطه زیر در فضای حالت باشد:

$$X_t = [x_t, y_t, \dot{x}_t, \dot{y}_t]^T \qquad (۱۰-۲)$$

فرض کنید Y_t متغیر مشاهداتی که از الگوریتم‌های ASL بدست می‌آید، باشد؛ بدین ترتیب با استفاده از فیلتر بیزین، این سیستم را می‌توان توسط معادلات زیر بیان کرد:

$$\begin{aligned} X_t &= g(X_{t-1}, u_t) \\ Y_t &= h(X_t, v_t) \end{aligned} \qquad (۱۱-۲)$$

که در معادله فوق، g(.) , h(.) تابع غیر خطی و متغیر با زمان که مشاهدات را حالت منبع صوتی می‌نگارد. و u_t و v_t نویز سفید IID[2] می‌باشد. هدف نهایی محاسبه $P(X_t | Y_{1t})$ تابع چگالی احتمال پسین [3]است.

فیلتر بیزین برای حل این مساله، دو مرحله پیشگویی و آپدیت را در نظر گرفته است:

$$\begin{aligned} P(X_t | Y_{1:t-1}) &= \int P(X_t | X_{t-1}) P(X_{t-1} | Y_{1:t-1}) dx_{t-1} \\ P(X_t | Y_{1:t}) &\propto P(Y_t | X_t) P(X_t | Y_{1:t-1}) \end{aligned} \qquad (۱۲-۲)$$

[1] sequential
[2] Independent and Identically Distributed
[3] posterior PDF

که در آن $P(X_t \mid X_{t-1})$ بیانگر چگالی انتقالی(یا مدل حرکتی منبع) و $P(Y_t \mid X_t)$ به تابع احتمال معروف می‌باشد. و برای هر فریم زمانی، این دو مرحله تکرار می‌شوند.

یکی از معروفترین فیلترهای تخمینگر، کالمن فیلتراست[۲۶] که در شرایط خطی و گوسی کاربرد دارد؛ ولی تخمین TDOA به طور خطی به موقعیت منبع وابسته نیست. همچنین اظهارات کوتاه در جملات (که معمولا کوتاه تر از یک ثانیه می‌باشند) و تغییر سریع صحبت کننده و همپوشانی صحبتی نیز کالمن را با مشکل مواجه می‌سازد.

با بهبود فیلتر کالمن به EKF با استفاده از یک مرحله خطی سازی با استفاده از سری تیلور مشکل غیر خطی بودن را بر طرف کرد. اما تظیم پارامترهای این فیلتر در کل مشکل می‌باشد، و تخمین‌گر بهینه‌ای نیست. اخیرا UKF در[۲۷] پیشنهاد شده است که در آن مشکل غیر گوسی بودن فیلتر کالمن مورد بررسی قرار گرفته است.

این روش هنوز در مورد همزمانی صحبت کننده‌ها و پدیده همپوشانی، تغییرات سریع صحبت کننده‌ها و اظهارات کوتاه هنوز با مشکل مواجه است.

یکی از فیلترهای مطرح برای ردیابی روش مونت کارلو ترتیبی[۱] (SMC) معروف به پارتیکل‌فیلتر[۲] می‌باشد که با بیان کردن هر توزیع احتمالی از طریق مجموعه محدودی از پارتیکل‌ها یک فیلتر بهینه بیزین را تخمین می‌زند. کاربرهایی از تعیین محل یک منبع صوتی و ردیابی آن در[۲۸] دیده می‌شود. تغییر سریع صحبت کننده‌ها نیازمند یک مدل چند منبعی دارد و با وجود اینکه پارتیکل‌فیلتر با استفاده از یک توزیع چند قله‌ای[۳] این مشکل را بررسی کند اما اگر تعداد آبجکت‌ها در طول زمان تغییر زیادی داشته باشد و باز هم تغییر سریع صحبت کننده‌ها را داشته باشیم، پارتیکل‌فیلتر با مشکلاتی مواجه می‌شود.

[1] Sequential Monte-Carlo
[2] Particle filter

[۳] به ازای تعداد افراد، قله‌ایی در نظر گرفته می‌شود.

در مساله ردیابی صوتی باید به این نکته توجه نمود که در گفتگوها معمولا صحبت‌ها بسیار کوتاه می‌باشد و به طور مرتب صحبت کننده‌ها تغییر می‌کنند و ما بهتراست بجای استفاده از فیلترهای معرفی شده که در پریودهای زمانی طولانی (چند ثانیه یا بیشتر) کار می‌کنند، از روشی استفاده کنیم که عمل ردیابی را در دوره‌های کوتاه تر(کمتر از(۲۵۰ ms) انجام دهد.

در ادامه بحث به معرفی کامل‌تر پارتیکل‌فیلتر پرداخته و به علت عملکرد مناسب این فیلتر برای ردیابی منابع صوتی اخیراً در مقالات بسیاری از این روش استفاده شده است.

۲-۶- پارتیکل فیلتر

این فیلتر برای نمایش تابع چگالی احتمال پسین بوسیله مجموعه‌ای از نمونه‌های وزن‌دار به‌نام پارتیکل و محاسبه تخمین حالت مطلوب بر اساس این پارتیکل‌ها به‌کار برده می‌شود؛ در این فیلتر با استفاده از مجموعه‌ای از پارتیکل‌ها، و بر اساس مدل حرکتی انتخاب شده برای منبع صوتی عملیات انتشار پارتیکل‌ها صورت می‌گیرد و بر اساس مجموعه‌ای از مشاهدات صوتی (ناشی از اطلاعات الگوریتمTDOA) وزندهی می‌شوند.

در ادامه بر اساس یک توزیع احتمالی ناشی از وزن‌های پارتیکل‌ها عملیات Resampling انجام می‌شود. در نهایت موقعیت منبع با یک میانگین‌گیری از موقعیت پارتیکل‌ها بدست می‌آید. در عمل با داشتن یک مدل احتمالی بر اساس maximum likelihood (ML)، بعد از چند تکرار پارتیکل‌ها در نواحی که احتمال بالاتری دارند، متمرکز می‌شوند.

یکی از مشکلات پارتیکل‌فیلتر این است که، اگر تابع احتمال خیلی دقیق و تیز باشد، پارتیکل‌ها دور محل منبع گسترش می‌یابند؛ در این حالت به برخی از عملیات هموار سازی نیاز است.[۲۹]. بنابراین در کل یک روش ساده برای حل معادلات پیشگویی و آپدیت آورده شده فیلتر بیزین، استفاده از پارتیکل‌فیلتر برای تخمین عددی انتگرال‌ها می‌باشد.

پارتیکل‌فیلتر با استفاده از مجموعه‌ای از N_j مجموعه وزندار تصادفی، برای هر فریم زمانی به تخمین چگالی احتمال پیشین می‌پردازد:

$$P(X_t|Y_{1:t}) \approx \sum_{n=1}^{N} w_t^{(n)} \delta(X_t - X_t^{(n)}) \qquad (۱۳-۲)$$

که مجموع وزن‌های اسکالر $w_t^{1,\dots,N}$ واحد است وتابع $\delta(.)$ دلتای دیراک می‌باشد. یک روش ساده برای پیاده‌سازی پارتیکل فیلتر، روش SIR می‌باشد. در این الگوریتم پارتیکل‌ها بر اساس یک مدل حرکتی انتشار می‌یابند؛ این مدل بر اساس دانشی از حرکت افراد صحبت کننده انتخاب می‌شود. می‌توان نشان داد که واریانس وزن‌ها در طول زمان افزایش می‌یابد، که باعث کاهش دقت الگوریتم می‌گردد؛این مساله به انحطاط[1] معروف است که برای رفع این مشکل از Resampling استفاده می‌شود؛ که در آن پارتیکل‌ها با وزن‌های کوچک با پارتیکل‌های با وزن بیشتر جایگزین می‌شوند. موقعیت منبع صوتی در فریم زمانی t با رابطه زیر بدست می‌آید:

$$\hat{X}_t \approx \sum_{n=1}^{N} w_t^{(n)} . X_t^{(n)} \qquad (۱۴-۲)$$

بعد از تخمین توزیع چگالی، نقاط تخمینی برای نمایش موقعیت منبع انتخاب می‌شوند و یک فیلتر میانگین‌گیر در ادامه به‌کار گرفته می‌شود.

$$X_t^{ME} = median\{x_t^1, x_t^2, \dots x_t^N\} \qquad (۱۵-۲)$$

فرض: در فریم زمانی t-1 مجموعه‌ای از پارتیکل‌ها $X_{t-1}^{(n)}$ و وزن‌ها $W_{t-1}^{(n)}$ به صورت گسسته‌ای $P(X_{t-1}|Y_{1:t-1})$ را به نمایش می‌گذارند.

[1] degeneracy

تکرار: مشاهده Y_t از فریم زمانی t و آپدیت پارتیکل‌ها با توجه به روابط زیر:

۱- پیشگویی: انتشار پارتیکل‌ها از طریق معادله انتقالی

۲- آپدیت: تخصیص دادن یک وزن احتمالی به هر پارتیکل و نرمالایز کردن وزن‌ها.

For j=1 to N do

$\hat{X}_t^j \sim P(X_t \mid X^j_{t-1})$

Calculate $w_t^j = P(Y_t \mid X^j_t)$

END

$w_t^{1:N} / \sum_{j=1}^{N} w_t^j$

۳- Resampling: بدست آوردن N نمونه جدید $X_t^{(n)}$ از پارتیکل‌های موجود $\{\hat{X}_t^j\}_{j=1}^N$بر طبق وزن‌های آن‌ها، سپس ریست کردن وزن‌ها $w_k^{(n)} = 1/N$

$c_1 = 0$

For j=2 to N do

Calculate CDF: $c_j = c_{j-1} + w_t^j$

END

j=1

Starting point: $u_1 = U(0, N_j^{-1})$

For k=1 to N do

$u_k = u_1 + (k-1).N_j^{-1}$

While $u_k > c_j$ do

j=j+1

END

$X_t^{(k)} = X_t^j, w_k^{(n)} = 1/N$

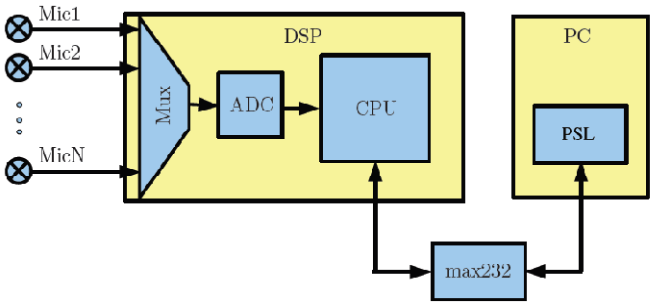
نتیجه: مجموعه $\{X_t^{(n)}, w_t^{(n)}\}_{n=1}^N$ توزیع چگالی $P(X_t \mid Y_{1t})$ را تخمین می‌زند.

الگوریتم ۱: SIR الگوریتم برای پارتیکل فیلتر

در انتهای این فصل دریافت اطلاعات صوتی را با استفاده از پردازشگر DSP مورد بررسی قرار می‌دهیم.

۲-۷- استفاده از DSP در ASL

سیگنال‌های آنالوگ دریافتی توسط جفت میکروفن‌ها نیاز به یک سیستم ارزان قیمت و دقیق برای تبدیل به سیگنال دیجیتال دارند. و در نهایت این سیگنال‌های دیجیتال به کامپیوترها فرستاده شده تا عملیات ASL بروی آن‌ها اجرا گردد. DSP پروسسوری است که برای پردازش سیگنال‌های دیجیتال طراحی شده است؛ و معمولا شامل پورت‌های آنالوگ به دیجیتال در ورودی و پورت‌های دیجیتال به آنالوگ در خروجی می‌باشد. DSP معمولا از یک حافظه کوچک برای عملیات‌های همزمان مانند، عملیات حسابی، ضرب سریع و بهینه‌سازی محاسبات پردازش سیگنال برپایه الگوریتم‌های FFT ، بهره می‌گیرد. این خواص باعث می‌شود که DSP یک ابزار قدرتمند برای ASL بر پایه آرایه‌های میکروفنی باشد. DSP سیگنال‌ها را دیجیتالی کرده و به یک PC انتقال می‌دهد، تا الگوریتم‌های ASL بروی آن‌ها اجرا گردد.



شکل ۱-۴: سخت افزار مورد نظر برای دریافت سیگنال و استفاده از ان برای الگوریتم‌های ASL در کامپوتر

میکروکنترلرهای جدید معمولا یک یا چند پورت UART برای ارتباط سریال با وسایل جانبی و کامپیوترها دارد. پروتکل استاندارد استفاده شده برای یک پورت سریال 232(PS) می‌باشد.

سرعت انتقال پورت سریال معمولا محدود به پورت سریال کامپیوتر وبرابر ۱۱۵۲۰۰ bauds/s می‌باشد.

ماکزیمم نرخ تبدیل AD از رابطه زیر پیروی می‌کند:

$$f_{s\max} = \frac{f_{\max}}{N} \qquad (۱۶-۲)$$

که در آن N تعداد پورت‌های انتخاب شده در ورودی می‌باشد. و بنابراین با افزایش استفاده از پورت‌های ورودی هر DSP ، فرکانس نمونه‌برداری دیتاها نیز کاهش می‌یابد.

می‌توان برای هر جفت میکروفن از یک DSP مجزا استفاده کرد و سپس تمامی آن‌ها را به یک DSP مرجع متصل کرد؛ البته در این صورت تمامی DSPها باید توسط DSP مرجع همزمان شوند. در این ساختار نیاز به حافظه چندانی برای ذخیره دیتاها توسط DSP نداریم وبنابراین پردازشگر می‌تواند با سرعت بالاتری عملیات خود را انجام دهد و پیاده‌سازی آنی[1] داشته باشد.

و نیز می‌توان توسط یک DSP کل میکروفن‌ها را اداره کرد؛ مزیت استفاده از این ساختار، عدم نیاز به همزمانی، قیمت کمتر و پیچیدگی کمتر مدار ارتباطی می‌باشد. اما با توجه به رابطه آورده شده برای دستیابی به فرکانس نمونه‌برداری مناسب تعداد DSPها باید کم باشد. همچنین اجرا محاسبات با حجم بالاتر ما را از پیاده‌سازی آنی دور می‌کند.

[1] Real time

۲-۸- جمع‌بندی

در این فصل روش‌های رایج در زمینه ASL مطرح شد،در ادامه توضیحات کاملی در مورد روش‌های غیرمستقیم یا TDOA آورده شد. سپس در مورد مرحله دوم این روش یعنی تعیین محل منبع صوتی تحت عنوان روش‌های closed-form صحبت شد. پارتیکل فیلتر به عنوان یکی از موثرترین فیلترها برای ردیابی منبع صوتی مورد بحث قرار گرفت، و در نهایت به طور خلاصه در مورد تهیه دیتای صوتی با استفاده از DSP بحث کردیم.

در فصل سوم و با تاکید بیشتر بروی روش‌های TDE بحث خود را با انتخاب روش‌های مستقیم TDE به دلایل برتری‌هایی که نسبت به سایر روش‌های غیر مستقیم داشت، ادامه می دهیم. جزییات و پیاده‌سازی این روش در ادامه مورد بررسی قرار خواهد گرفت.

فصل سوم

تعیین محل منابع صوتی توسط روش‌های تخمین تاخیر زمانی

۳-۱- مقدمه

در فصل دوم کلیه روش‌های ASL مورد بررسی قرار گرفت. در این تحقیق ما روش‌های بر حسب TDE را انتخاب می‌کنیم، که در این فصل به طور مفصل مورد بررسی قرار خواهد گرفت. کلیه روش‌های بر حسب TDE به تخمین تاخیر زمانی بین جفت میکروفن‌ها نیازمند هستند. در روش های غیرمستقیم بعد از محاسبه این مقادیر تاخیر از آنها در مرحله دوم برای تعیین محل دقیق منابع صوتی استفاده می‌شود، اما در روش‌های مستقیم، در یک مرحله و با استفاده از تاخیرهای بدست آمده در مورد محل تخمینی منبع مستقیماً تصمیم‌گیری می‌شود.

در ادامه بحث ابتدا درمورد روش‌های تعیین TDOA، بحث خواهیم نمود و در ادامه به بررسی کامل روش‌های مستقیم خواهیم پرداخت. در انتها دیتابیس‌های استفاده شده در این تحقیق را معرفی می‌نماییم.

۳-۲- روش GCC-PHAT

اصول کاری اغلب روشهای TDE بر پایه تابع همبستگی متقابل تعمیم‌یافته[1] (GCC) می‌باشد. CC میان دو سیگنال میزان شباهت سیگنال‌ها را نسبت به هم بدست آورده و می‌تواند نشانگر یک پیک بر حسب زمان تاخیر باشد، که سیگنال شیفت یافته مطابق با تاخیر زمان رسیدن سیگنال بین دو میکروفن ایجاد می‌کند.

تابع CC ماکزیمم مقدار را در زمان تاخیر تئوری TDOA خواهد داشت. استفاده از روشCC برای محاسبه مقدار تاخیر بین دو سیگنال، ایده‌آل نبوده و در بسیاری از موارد و در صورت وجود نویز و یا انعکاسات محیطی دقت پایینی دارد. زمانی که توان انعکاسات بالا رود و یا مثلاً جهت سر گوینده بر خلاف میکروفن‌ها و به طرف دیوار باشد، چندین پیک ثانویه بوجود می‌آید که ناشی از امواج صوتی

[1] Generalized Cross-Correlation

بازگشتی می‌باشد؛ بنابراین برای رفع پاره ای از مشکلات و با توجه شرایط محیطی، عمل پیش فیلتر، یعنی اعمال چندین تابع وزن استفاده می‌شود؛ تا بتوان اثرات نویز و انعکاسات ناخواسته محیطی را کاهش داد. این فیلترها براساس نوع شرایط و کاربردها، از یکدیگر متمایز می‌گردند. این روش را GCC یا همبستگی متقابل تعمیم یافته می‌نامند.

از معروفترین این فیلترها می توان به ML اشاره کرد که از نظر تئوری زمانی که مقدار انعکاسات کم باشد، روش بهینه ای است، اما با افزایش انعکاسات دچار مشکل می‌گردد. یکی دیگر از توابع وزنی، فیلتر Phase Transform (PHAT) می‌باشد که در برابر نویز و انعکاسات عملکرد خوبی از خود نشان‌می‌دهد. سایر توابع وزنی Roth, SCOT, Eckart, and the Hassab-Boucher method در[۳۰] معرفی شده‌است. در یک فریم زمانی با استفاده از روش GCC-PHAT ما به دنبال پیدا کردن مقدار تاخیری هستیم که تابع GCC-PHAT را در حوزه زمان ماکزیمم کند. در ادامه بحث معادلات مربوط به این روش را بررسی می‌کنیم.

برای بدست آوردن TDE بین یک جفت میکروفن از روابط زیر استفاده می کنیم که در آن c سرعت صوت بوده و معادل c = 342 m/s در نظر گرفته می‌شود.

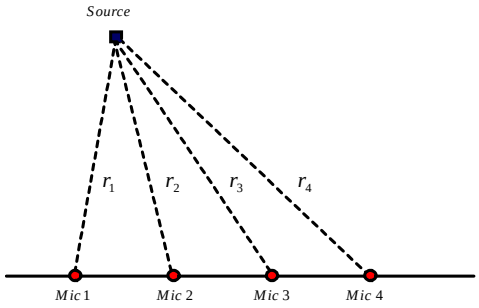
$$\tau_m = \frac{r_m^{(s)}}{c} \qquad (۱-۳)$$

$$\tau_{a,b} = \tau_a - \tau_b = \frac{r_a^{(s)} - r_b^{(s)}}{c} \qquad (۲-۳)$$

در شکل ۳-۱ یک آرایه خطی از میکروفن‌ها را به همراه محل منبع صوتی نشان داده ایم.

با توجه به معادلات آورده شده برای تابع CC تابع GCC را در حوزه فرکانس به فرم زیر معرفی می‌کنیم:

$$G_{x_1,x_2}^{(t)}(k) = (G_1(k)X_1^{(t)}(k))(G_2(k)X_2^{(t)}(k))^* \quad (۳-۳)$$



شکل ۳-۱: تعیین تاخیر زمانی امواج از منبع به سمت جفت میکروفن‌ها

$$G_{x_1,x_2}^{(t)}(k) = (G_1(k)G_2(k)^*)(X_1^{(t)}(k)(X_2^{(t)}(k))^*) \quad (۴-۳)$$

$$G_{x_1,x_2}^{(t)}(k) = (\psi_{1,2}(k))X_1^{(t)}(k)(X_2^{(t)}(k))^* \quad (۵-۳)$$

با استفاده از یک تابع وزنی مناسب، GCC می‌تواند پیکی را نشان دهد که معادل TDE میان میکروفن ۱و۲ خواهد بود. این روش به روش crosspower spectrum phase (CPS) نیز معروف می‌باشد.

در کل تابع ما چندین ماکزیمم دارد که دامنه و تاخیر زمانی معادل این نقاط ماکزیمم بستگی به فاکتورهایی نظیر فاصله میان میکروفن‌ها ،طبیعت سیگنال صوتی و سیگنال نویز و نیز انتخاب تابع وزنی دارد.

سیگنال‌های یک جفت میکروفن به فرم زیر نشان داده می شود:

$$x_1(n) = a_1 s(n-D) + h_1(n) * s(n) + n_1(n) \quad (۶-۳)$$
$$x_2(n) = a_2 s(n) + h_2(n) * s(n) + n_2(n)$$

معمولا در مدلهای تقریبی فرض می شود که:

- سیگنال و نویز ناهمبسته هستند.
- انعکاس وجود ندارد.
- نویز دو سیگنال نسبت به هم ناهمبسته اند.

با توجه به فرضیات فوق داریم:

$$R^{(t)}_{x_{a_q},x_{b_q}}(\tau) = F^{-1}.[R^{(t)}_{s,s}[w]e^{-jwD}] = R_{ss}(\tau - D) \quad (۷-۳)$$

در عمل دو فرض آخری همواره صحیح نیست و تداخل می‌تواند تاثیر گذار باشد بنابراین:

$$R^{(t)}_{x_{a_q},x_{b_q}}(\tau) = \sum_i R_{ss}(\tau - D_i) \quad (۸-۳)$$

در این حالت اگر چندین پیک برای تابع خود همبستگی منبع داشته باشیم، همپوشانی و تداخل این پیک‌ها در صورت نزدیک بودن به هم می‌تواند مشکل‌ساز باشد.

برای رفع این مشکل، از تابع وزنی PHAT استفاده می‌کنیم که معادل سفید کردن در حوزه فرکانس سیگنالهای میکروفن است و به فرم زیر نشان داده می شود:

$$\psi_{1,2}[k] = \frac{1}{\left|X_1^{(t)}[k].(X_2^{(t)}[k])^*\right|} \quad (۹-۳)$$

$$G^{\wedge(PH)}_{x_1,x_2}[w] = \frac{R^{\wedge}_{ss}(w)e^{-jwD}}{\left|R^{\wedge}_{ss}(w)e^{-jwD}\right|} = e^{-jwD}$$
$$g^{\wedge}_{x_{a_q},x_{b_q}}(\tau) = \delta(\tau - D) \quad (۱۰-۳)$$

در این حالت سیگنال خود همبستگی منبع، ناپدید شده و برای تابع همبستگی متقابل خروجی پیک تیزتری را خواهیم داشت. لازم به ذکر است که اطلاعات مربوط تاخیر زمانی سیگنال‌ها نسبت به‌هم در فاز سیگنال موجود بوده و اعمال تابع وزنی فوق تاثیری بروی آن ندارد. بنابراین برای تابع GCC-PHAT خواهیم داشت:

$$G_{x_1,x_2}^{(PH,t)}[k]=\frac{(X_1^{(t)}[k]X_2^{(t)}[k])^*}{\left|X_1^{(t)}[k].(X_2^{(t)}[k])^*\right|} \quad (۱۱-۳)$$

می‌توان تابع وزنی PHAT را با توان اعمال نمود که در این صورت در شرایط نویزی ضعیف با انتخاب توان بین (۰.۵و۰.۶) بهترین نتایج را خواهیم داشت.[۳۱]

عملیات صفرگذاری برای جلوگیری از مقوله شیفت دایره‌ای بروی سیگنال‌ها انجام می‌گیرد. برای بهبود عملکرد GCC می‌بایست بروی چندین فریم متوالی متوسط گیری صورت گیرد چراکه این عمل باعث حذف نویز و انعکاسات می‌گردد. این متوسط گیری را بروی I فریم متوالی انجام می‌شود، ولی باید به این نکته توجه کرد که اگر برای حذف نویز تعداد فریم‌ها را برای متوسط گیری افزایش دهیم، دقت تعیین محل در صورتیکه صحبت کننده متحرک باشد کاهش می یابد.

$$C_{x_1,x_2}^{t}[k]=\frac{1}{I}\sum_{i=t}^{t+I-1} X_{1,i}[k].(X_{2,i}[k])^* \quad (۱۲-۳)$$

بنابراین می‌توان تابع GCC-PHAT را به فرم زیر بهبود داد:

$$G_{x_1,x_2}^{(PH,t)}[k]=\frac{C_{x_1,x_2}^{t}[k]}{\left|C_{x_1,x_2}^{t}[k]\right|} \quad (۱۳-۳)$$

$$g_{x_{aq},x_{bq}}^{(t)}(\tau)=F^{-1}.G_{x_1,x_2}^{(PH,t)}[k]=\frac{1}{k}\sum_{k=0}^{K-1}\frac{C_{x_1,x_2}^{t}[k]}{\left|C_{x_1,x_2}^{t}[k]\right|}.e^{jk\frac{2\pi}{K}\tau} \quad (۱۴-۳)$$

و در ادامه می‌توان تاخیر زمانی بین جفت میکروفن‌ها را به فرم زیر تخمین زد:

$$\tau_q^{(t)}=\arg\max[g_{x_{aq},x_{bq}}^{(t)}(\tau)] \quad (۱۵-۳)$$

برای افزایش اطمینان از صحت عملکرد سیستم فوق، معیارهایی را برای هر فریم در نظر گرفته می‌شود که در صورت تایید آن معیارها بروی آن فریم عمل تعیین محل صورت گیرد. دو تا از معیارهای مناسب برای این کار پیک تابع GCC و نسبت اولین پیک به دومین پیک است، که در صورتی که از یک مقدار تعیین شده تجربی توسط کاربر بیشتر باشد، عملیات تعیین محل بروی آن فریم صورت می‌گیرد.

۳-۳- روش AMDF

رابطه Average magnitude difference function (AMDF) برای سیگنالهایی از یک جفت میکروفن به صورت زیر در[۳۲] تعریف شده است:

$$R_P^{AMDF}(\tau) = \frac{1}{L}\sum_{t=0}^{L-1}|x_L(t) - x_K(t+\tau)| \qquad (۱۶-۳)$$

که در آن L طول فریم است. با تعریف AMSF به طور مشابه در مرجع و ترکیب آنها روش modified Average magnitude difference function (MAMDF) را نیز معرفی کرده و برای تعیین میزان تاخیر جفت میکروفن‌ها از آن کمک گرفته است.

۴-۳- روش TDOA وفقی

وجود انعکاسات در محیط اتاق موجب شده تا نتوان از روش GCC تخمین دقیقی بدست آورد. روش adaptive eigenvalue decomposition (AED) برای محیط‌های پر انعکاس مطرح شده است. هدف این روش تخمین کور پاسخ ضربه و سپس بدست آوردن میزان تاخیر می‌باشد. در مواردی که چندین منبع و یک جفت میکروفن داریم، با استفاده از ماتریس کواریانس دو سیگنال در یک جفت میکروفن مقادیر TDOA را تخمین می‌زنند[۳۲]، اما زمانی که چندین جفت میکروفن و چندین منبع

وجود دارد، مشخص نیست که چطور از این تاخیرهای زمانی بدست آمده برای تخمین محل دقیق منبع صوتی استفاده گردد.

۳-۵- روش‌های مستقیم

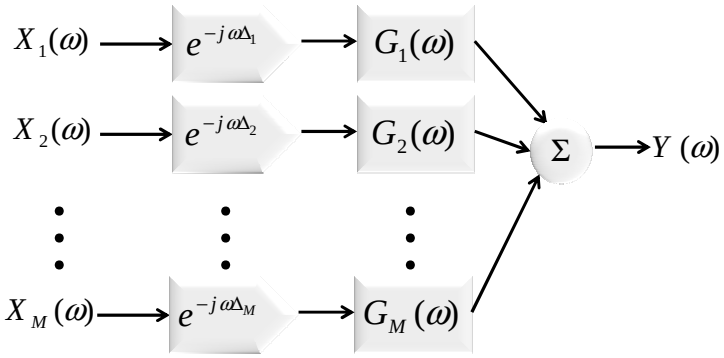
در این بخش قصد داریم روش‌های مستقیم TDE را بررسی نماییم. همان‌طور که در فصل دوم بحث شد، در سال‌های اخیر و با رشد تکنولوژی از روش‌های مستقیم باوجود محاسبات پیچیده‌تر، بیشتر استفاده می‌گردد. دراین تحقیق روش اصلی استفاده شده بر مبنای روش‌های مستقیم می‌باشد که در ادامه به‌طور کامل بررسی می‌گردد.

الگوریتم‌های این کلاس اساساً تلاش می‌کنند تا پویشی را بروی شبکه‌ای از نقاط اجرا کنند، که تابعی را ماکزیمم یا مینیمم می‌کند. روش‌های مستقیم بر خلاف روش‌های دومرحله‌ای حجم بالای محاسبات را نیاز دارند. امروزه با رشد تکنولوژی، این روش‌ها بسیار مورد توجه هستند. بزرگترین مزیت این روش‌ها این است که نیازی به فرض نوع انتشار صوت و مدل میکروفن و سیگنال صوت، ندارد.

روش‌های مستقیم به دو دسته اصلی Coherent Signal Subspace Processing (CSSP) و SRP تقسیم می‌شوند: روشCSSP معمولاً در زمینه رادار و مخابرات استفاده می شود. این روش برای سیگنال‌های باند باریک طراحی شده است البته تلاش‌هایی برای تبدیل این روش برای سیگنالهای باندپهن نیز در آن‌ها صورت گرفته است. این روش در برابر انعکاسات از حساسیت بالایی برخوردار است و همچنین نیازمند محتوای مناسبی از اطلاعات می‌باشد، که به معنی فریم‌های با زمان طولانی و یا متوسط گیری در چندین فریم زمانی است. در کل با وجود پیشرفت‌هایی که برای تعیین محل منابع صوتی داشته، هنوز روش مناسبی در این حیطه بشمار نمی‌رود.

ایده اصلی در روش Steered Response Power (SRP) تخمین توان یک نقطه در فضا توسط هدایت تاخیرها می‌باشد که این کار توسط اختلاف زمان پرواز نسبت به میکروفن‌های مختلف انجام می‌شود و محل تخمینی از طریق یافتن ماکزیمم محلی توان یک Beamforming بدست می‌آید. چندین منبع صوتی به طور همزمان چند توان ماکزیمم را در فضای جستجو به همراه دارد اما انعکاسات نیز مانند یک توان ماکزیمم عمل می‌کند که برای رفع این مشکل مانند روش GCC ، از تابع وزنی PHAT استفاده می‌گردد.[۳۴].

سیگنال‌هایی که از منابع به میکروفن‌ها می‌رسد شامل تاخیر است و در فضای انتشار توسط نویز و انعکاسات فیلتر می‌شود؛ نسخه‌های تاخیر یافته منبع، سیگنالهای حاصل می‌توانند با هم جمع شده، به‌طوری‌که سیگنال نویز ناهمبسته را حذف کنند. یک filter-and-sum beamformer می‌تواند به فرم شکل ۳-۲ نشان داده شود:



شکل ۳-۲ : filter and sum beamformer

تاخیرهای هدایت شونده [1] $\Delta_1...\Delta_m$ برای متمرکز شدن و هدایت آرایه به سمت مکان منابع صوتی بکار گرفته می‌شوند. Filter and sum beamformer درحوزه فرکانس به صورت زیر تعریف می‌شود:

$$Y_t(w, X_1^t ... X_M^t) = \sum_{m=1}^{M} G_m(w) X_m(w) e^{-jw\Delta_m} \quad (۱۷-۳)$$

زمانی که توسط Beamformer بروی جهت منبع صوتی متمرکز شویم، توان steered response ماکزیمم می‌گردد، SRP می‌تواند به صورت توان خروجی یک filter-and-sum beamformer نشان داده شود.

$$\begin{aligned} &P_{SRP}(L, X_1^t ... X_M^t) = \int_{\Re} Y_t(w, X_1^t ... X_M^t)(Y_t(w, X_1^t ... X_M^t)^*) dw \\ &P_{SRP}(L, X_1^t ... X_M^t) \\ &= \int_{\Re} (\sum_{a=1}^{M} G_a(w) X_a(w) e^{-jw\Delta_a})(\sum_{b=1}^{M} G_b^*(w) X_b^*(w) e^{jw\Delta_b}) dw \end{aligned} \quad (۱۸-۳)$$

رفتار تاخیرهای هدایت شونده‌ای که معادله فوق را ماکزیمم کنند، شبیه به رفتار تاخیرهای زمانی معادل TDE میان جفت میکروفن‌ها می‌باشد که ماکزیمم پیک را در روش GCC به ما می‌داد. SRP به طور ذاتی از سیگنالهای چندین میکروفن موجود میانگین می‌گیرد. می‌توان نشان داد که SRP از M آرایه میکروفنی معادل با GCC گرفتن از تمامی جفت میکروفن‌ها می‌باشد.

$$\begin{aligned} &P_{SRP}(L, X_1^t ... X_M^t) \\ &= \sum_{a=1}^{M} \sum_{b=1}^{M} \int_{\Re} G_a(w) G_b^*(w) X_a(w) X_b^*(w) e^{jw(\Delta_b - \Delta_a)} dw \end{aligned} \quad (۱۹-۳)$$

در بخش قبل برای GCC داشتیم:[2]

$$g_{1,2}(\tau) = \frac{1}{2\pi} \int_{\Re} G_1(w) G_2^*(w) X_1(w) . (X_2(w))^* e^{jw\tau} dw \quad (۲۰-۳)$$

[1] Steering delays

با ترکیب معادلات فوق داریم:

$$P_{SRP}(L, X_1^t ... X_M^t) = 2\pi \sum_{a=1}^{M} \sum_{b=1}^{M} g_{a,b}(\Delta_a - \Delta_b) \qquad (۲۱-۳)$$

این معادل مجموع تمام جفت‌های ممکن از GCC است که بوسیکه اختلاف در steering delays شیفت زمانی داده می‌شوند. با دیدن معادله فوق می‌توان فهمید که SRP با استفاده از سیگنالهای مربوط به میکروفن‌های مختلف و متوسط‌گیری روی آنها به نحوی عملکرد بهتری را در شرایط مختلف از خود نشان می‌دهد. این عمل معادل متوسط گیری مکانی می‌باشد، چراکه هر میکروفن مکان مختلفی را از سایرین دارد؛ درحالی‌که در GCC ما متوسط زمانی بروی چند فریم می‌گرفتیم و چون نیازمند دنبال کردن منبع صوتی متحرک می‌باشیم، بنابراین روش SRP عملکرد بهتری را نسبت به GCC در تعیین محل افراد صحبت کننده از خود نشان می‌دهد. در برخی مواقع این امکان وجود دارد که با توجه به جهت سر گوینده، سیگنال منعکس شده از دیواره‌ها پیک بیشتری نسبت به خود سیگنال اصلی، در موقعیت یک جفت میکروفن ایجاد کند. در این صورت استفاده از متوسط گیری از کل میکروفن‌ها بسیار کمک کننده خواهد بود.

به بیان دیگر در این روش ما برای تعیین محل منبع صوتی از یک تک فریم استفاده می‌کنیم و برای نمونه مثل روش CC از متوسط گیری بروی چندین فریم زمانی استفاده نمی‌کنیم. این انتخاب به خاطر طبیعت پویا و متغیر گفتار توجیه می‌شود.

در این روش به طور ذاتی کمبود دیتا در فریم‌های کوچک گرفته شده، توسط سیگنالهای چندین میکروفن جبران می‌شود. و یکی دیگر از مزیت‌های این روش نسبت به روشهای دومرحله ای این است که می‌توان چندین صحبت کننده را به طور همزمان تعیین محل کرد. به طوری که توان سیگنال خروجی می‌تواند چندین ماکزیمم داشته باشد، که هر ماکزیمم معادل محل یک صحبت کننده خواهد بود. برتری متوسط‌گیری مکانی در روش SRP-PHAT نسبت به متوسط‌گیری زمانی در

GCC-PHAT در زمانی که صحبت کننده در حال صحبت به طور نا گهانی با سرعت بالایی حرکت کند، نمایان می‌گردد.

برای بهبود عملکرد SRP در برابر نویز و انعکاسات مانند قسمت قبلی، از فیلتر PHAT استفاده می‌کنیم:

$$G_1(w)G_2^*(w) = \frac{1}{|X_1(w).(X_2^*(w))|} = \frac{1}{|X_1(w)||(X_2^*(w))|}$$

$$\frac{1}{|X_1(w)|} = \frac{1}{|(X_2^*(w))|} \qquad (۲۲-۳)$$

$$G_1(w) = \frac{1}{|X_1(w)|} \qquad G_2(w) = \frac{1}{|(X_2(w))|}$$

بنابراین برای M میکروفن خواهیم داشت:

$$G_m(w) = \frac{1}{|X_m(w)|} \qquad (۲۳-۳)$$

این روش سفید کردن بروی تیز شدن پیک‌ها در تبدیل فاز تاثیر می‌گذارد.

رابطه نهایی توان SRP-PHAT در حوزه فرکانس و برای فرکانسهای مثبت به فرم زیر است:

$$P_{SRP-PHAT}(L, X_1^t \ldots X_M^t) = \sum_{k=2}^{N_F+1} \left| \sum_{m=1}^{M} \frac{X_m^{(t)}(k)}{|X_m^{(t)}(k)|} e^{j\pi \frac{k-1}{N_F} TOF(L, L_m)} \right|^2 \qquad (۲۴-۳)$$

$$L^{(t)} = \arg\max[P_{SRP-PHAT}(L, X_1^t \ldots X_M^t)] \qquad (۲۵-۳)$$

$$P_{SRP-PHAT}(L, X_1^t \ldots X_M^t) = \sum_{k=2}^{N_F+1} P_{SRP-PHAT}(k, L, X_1^t \ldots X_M^t) \qquad (۲۶-۳)$$

$$P_{SRP-PHAT}(k,L,X_1^t...X_M^t)=\left|\sum_{m=1}^{M}\frac{X_m^{(t)}(k)}{\left|X_m^{(t)}(k)\right|}e^{j\pi\frac{k-1}{N_F}TOF(L,L_m)}\right|^2$$
$$=\left|\sum_{m=1}^{M}\exp[j(\angle X_m^{(t)}(k)+\pi\frac{k-1}{N_F}TOF(L,L_m))]\right|^2 \quad (۲۷-۳)$$

با استفاده از رابطه $|z|^2=z.z^*$ داریم:

(۲۸-۳)

$$=N_m+\sum_{m=1}^{M}\sum_{\substack{m'\neq m\\ m'=1}}^{M}\exp\{j[\angle X_m^{(t)}(k)-\angle X_{m'}^{(t)}(k)+\pi\frac{k-1}{N_F}(TOF(L,L_m)-TOF(L,L_{m'}))]\}$$

و با استفاده از $z+z^*=2\Re\{z\}$ داریم:

(۲۹-۳)

$$P_{SRP-PHAT}(k,L,X_1^t...X_M^t)$$
$$=N_m+2.\sum_{m=1}^{M}\sum_{m'=m+1}^{M}\cos[\angle X_m^{(t)}(k)-\angle X_{m'}^{(t)}(k)+\pi\frac{k-1}{N_F}(TOF(L,L_m)-TOF(L,L_{m'}))]$$

$$=N_m+2.\sum_{q=1}^{N_q}\cos[\angle X_m^{(t)}(k)-\angle X_{m'}^{(t)}(k)+\pi\frac{k-1}{N_F}(TOF(L,L_m)-TOF(L,L_{m'}))] \quad (۳۰-۳)$$
$$=N_m+2.\sum_{q=1}^{N_q}\cos[\quad u_q^{(t)}(k)\quad+\quad u_q^{(th)}(k,L)\quad]$$

که $u_q^{(t)}(k)$ را فاز مشاهداتی و $u_q^{(th)}(k,L)$ را فاز تئوری می نامیم.

با استفاده از رابطه $\cos(x)=1-2\sin^2(\frac{x}{2})$ داریم:

$$P_{SRP-PHAT}(k,L,X_1^t...X_M^t) = N_m + 2.N_q - 4.\sum_{q=1}^{N_q}\sin^2\left(\frac{u_q^{(t)}(k)-u_q^{(th)}(k,L)}{2}\right) \quad (۳۱-۳)$$
$$= N_m + 2.N_q - 4.N_q d^2(u^{(t)}(k),u^{(th)}(k,L))$$

تابع d بکار رفته در رابطه فوق به فرم زیر معرفی می شود:

$$d(a,b) = \sqrt{\frac{1}{N_q}\sum_{q=1}^{N_q}\sin^2\left(\frac{a-b}{2}\right)} \quad (۳۲-۳)$$

برای تکمیل کردن روابط فوق داریم:

$$P_{SRP-PHAT}(L,X_1^t...X_M^t) = N_m + 2.N_q - 4.N_q\sum_{k=2}^{N_F+1} d^2(u^{(t)}(k),u^{(th)}(k,L)) \quad (۳۳-۳)$$

در رابطه فوق محل L که تابع $P_{SRP-PHAT}$ را ماکزیمم کند، معادلاً ترم سوم را مینیمم می‌کند؛ که در محاسبات از پیچیدگی کمتری برخوردار است. و به عبارت دیگر ماکزیمم کردن $P_{SRP-PHAT}$ معادل مینیمم کردن تابع آورده شده در فوق می‌باشد.

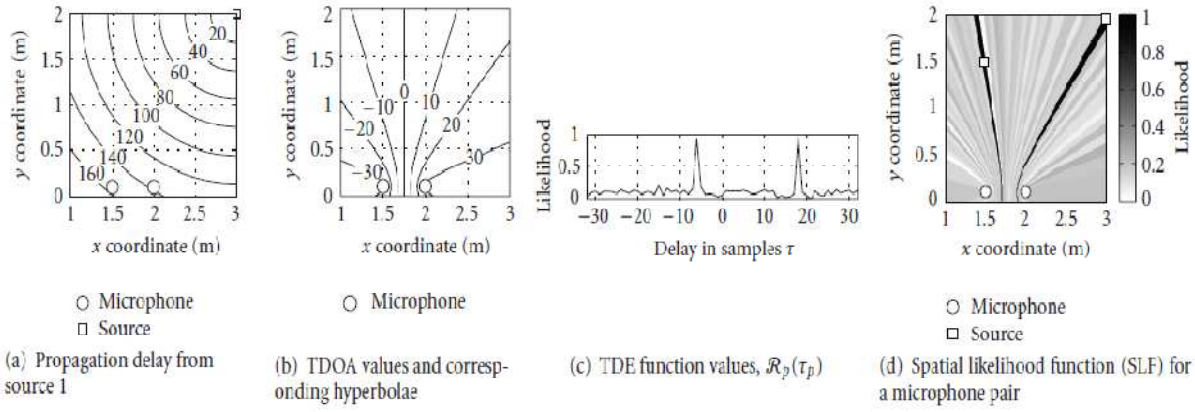
۱-۵-۳ استفاده از توابع SLF در تعیین محل بر پایه TDE

تخمین TDOA زمانی که SNR ضعیف باشد، با خطا مواجه می‌گردد و در انعکاسات قوی دیگر به موقعیت منبع صوتی بستگی ندارد. از مقادیر TDOA بدست آمده از جفت میکروفن‌های مختلف می‌توان توابع احتمال مکانی Spatial likelihood function (SLF) را بوجود آورد، بدین معنی که احتمال وجود منابع صوتی در یک محل دلخواه را با توجه به هر مقدار تاخیر موجود از جفت میکروفن‌ها نشان می‌دهد و هر مقدار تاخیر بدست آمده از جفت میکروفن‌ها می‌تواند یک نقطه احتمالی برای منبع صوتی باشد. ماکزیمم شدن این تابع SLF در هر نقطه، بیانگر وجود منبع در آن نقطه می‌باشد. [۳۵]

محل منبع صوتی r به مقدار TDOA ، $\Delta\tau_{p,r}$ ارجاع داده می‌شود.

در شکل ۳-۳ (a) انتشار امواج از یک منبع صوتی (b) مقادیر تاخیر با توجه به صفحات هیپربولیک، (c)استفاده از تابع GCC-PHAT برای نشان دادن مقدار تاخیر بین سیگنال‌های یک جفت میکروفن و (d) در نهایت تابع SLF با توجه به تاخیرهای بدست امده در مرحله قبل که می‌توان از مجموعه ای از مقادیر TDOA از چندین جفت میکروفن محل دقیق منابع را تخمین زد.

استفاده از عملگرهای مختلف برای ترکیب مقادیر چندین جفت میکروفن باید در جهت ایجاد ماکزیمم کردن شکل تابع SLF به بهترین نحو ممکن باشد.



(a) Propagation delay from source 1 (b) TDOA values and corresponding hyperbolae (c) TDE function values, $\mathcal{R}_p(\tau_p)$ (d) Spatial likelihood function (SLF) for a microphone pair

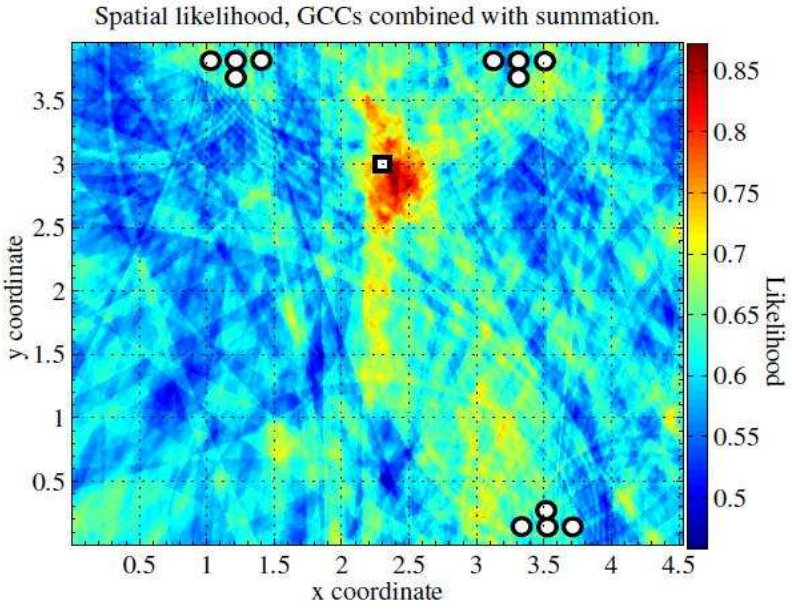
شکل ۳-۳ : دو منبع در r_1=(3, 2) و r_2 = (1.5, 1.5) قرار گرفته اند و مقادیر TDOA میباشند.[۳۵] $\Delta\tau1 = 18, \Delta\tau2 = -6$

یکی از روش‌های مطرح در ASL روش SRP-PHAT است که در بخش قبلی مورد بررسی قرار گرفت. با استفاده از این ایده و با اضافه کردن چندین جفت تابع TDE از میکروفن‌های مختلف می‌توان ابهام محل منبع و پیک مربوط به تابع SLF را کاهش داد؛ و در نهایت یک جستجو برای نقطه ماکزیمم ما را به محل منبع می‌رساند.

$$P_{SRP-PHAT}(R_{[1:S]} \mid r) = \sum_{P=1}^{S} R_P^{GCC-PHAT}(\Delta\tau_{P,r}) \qquad (۳۴-۳)$$

در شکل ۳-۴ استفاده از اجتماع در ترکیب تاخیر بدست آمده از جفت میکروفن‌های مختلف در تابع SLF را می‌بینیم. اما یک سری دنباله، ناشی از جفت میکروفن‌های مختلف وجود دارد که ناشی از عمل اجتماع می‌باشد و در جستجو برای مقدار ماکزیمم در شکل ما را دچار مشکل می‌کند. بنابراین برای از بین بردن این دنباله‌های مزاحم کافی است از اشتراک گرفتن از جفت میکروفن‌های مختلف استفاده کنیم. بدین ترتیب که به جای عمل جمع، از ضرب استفاده کنیم.

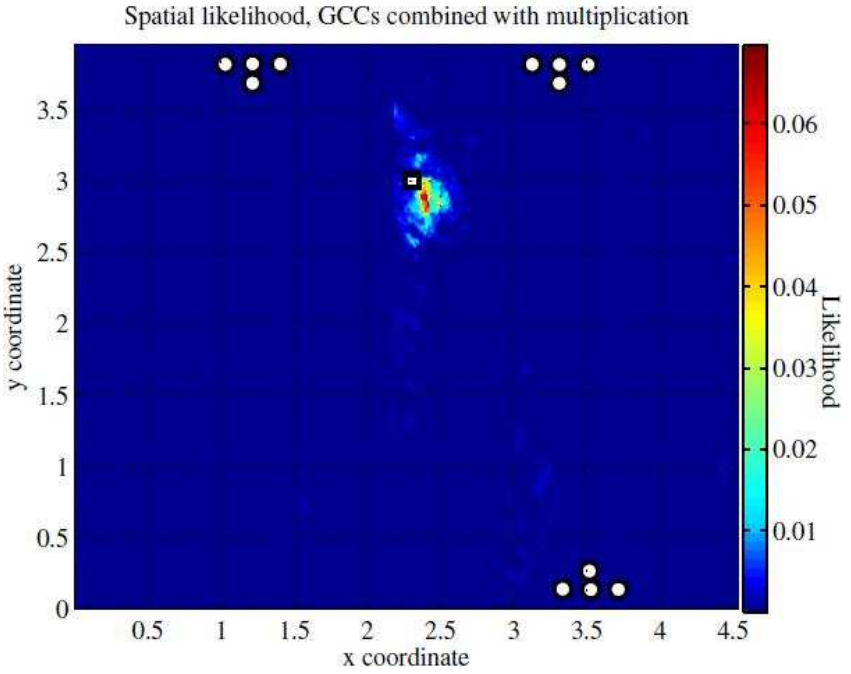
$$P_{PROD-PHAT}(R_{[1:S]} \mid r) = \prod_{P=1}^{S} R_P^{GCC-PHAT}(\Delta\tau_{P,r}) \qquad (۳۵-۳)$$



شکل ۳-۴ :استفاده از اجتماع برای ترکیب تمامی جفت میکروفن‌ها در تابع SLF [۳۵]

عمل ضرب تنها اطلاعاتی را که توابع SLF در آن با هم مشترک هستند را باقی نگه می‌دارد. در شکل ۳-۵ تابع اشتراک دیده می‌شود.

محل منابع صوتی در معادل ماکزیمم پیک تابع SLF می‌باشد. و اگر چندین منبع صوتی به طور همزمان داشته باشیم، معادلاً چندین پیک ولی با ارتفاع کمتر خواهیم داشت.



شکل ۳-۵ :استفاده از اشتراک در توابع SLF که تنها اطلاعات مشترک باقی می‌ماند. [۳۵]

در انتهای بحث در این فصل دیتابیس‌های استفاده شده در قسمت پیاده‌سازی‌ها معرفی می‌گردد.

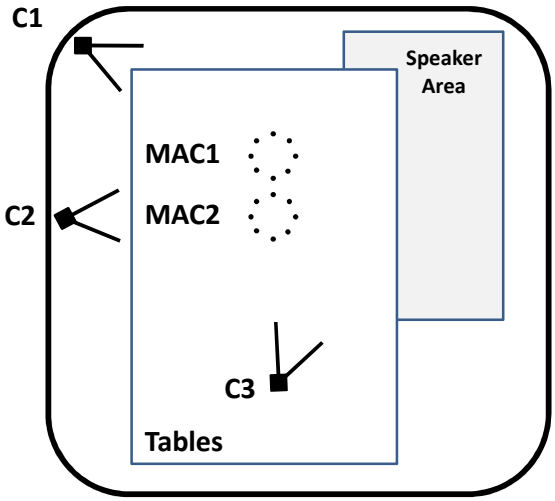
۳-۶- معرفی دیتابیس

در این بخش قصد داریم دو تا از دیتابیس‌هایی را که در این تحقیق از آنها استفاده شده است را به طور مختصر معرفی نماییم.

۳-۶-۱ دیتابیس با استفاده از آرایه میکروفنی دایره‌ای در مرکز اتاق

این دیتابیس رنج گسترده‌ای از فعالیت‌های صوتی، تصویری و صوتی-تصویری را با استفاده از شانزده میکروفن در دو آرایه میکروفنی و سه عدد دوربین پوشش می‌دهد؛ در زمینه صوتی مواردی مانند: یک فرد صحبت کننده و ساکن تا چندین فرد در حال صحبت و متحرک را به همراه مساله همپوشانی صحبتی در این دیتابیس می‌بینیم.

با استفاده از دوربین‌های کالیبره شده، محل هر صحبت کننده در فضای سه بعدی مشخص شده است که برای ارزیابی نتایج تعیین محل و ردیابی افراد صحبت‌کننده از آن می‌توان بهره برد.



شکل ۳-۶: سه دوربین و دو آرایه دایره‌ای و تنها ناحیه L شکل تحت پوشش سه دوربین هستند.

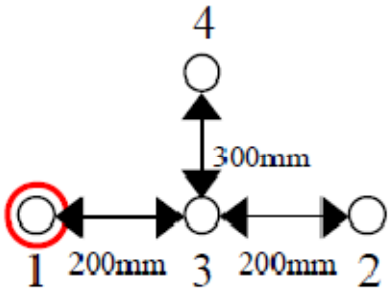
همانطور که در شکل دیده می‌شود، از سه دوربین و دو آرایه میکروفنی UCA هشت‌تایی با شعاع ۱۰ cm استفاده شده و فاصله بین دو آرایه هم ۰.۸ m می‌باشد.فرکانس نمونه برداری میکروفن‌ها ۱۸ Khz می‌باشد. اطلاعات کاملی از ابزارهای موجود در اتاق‌های کنفرانس را در[۳۶] داریم.

۳-۶-۲ دیتابیس UKA CHIL

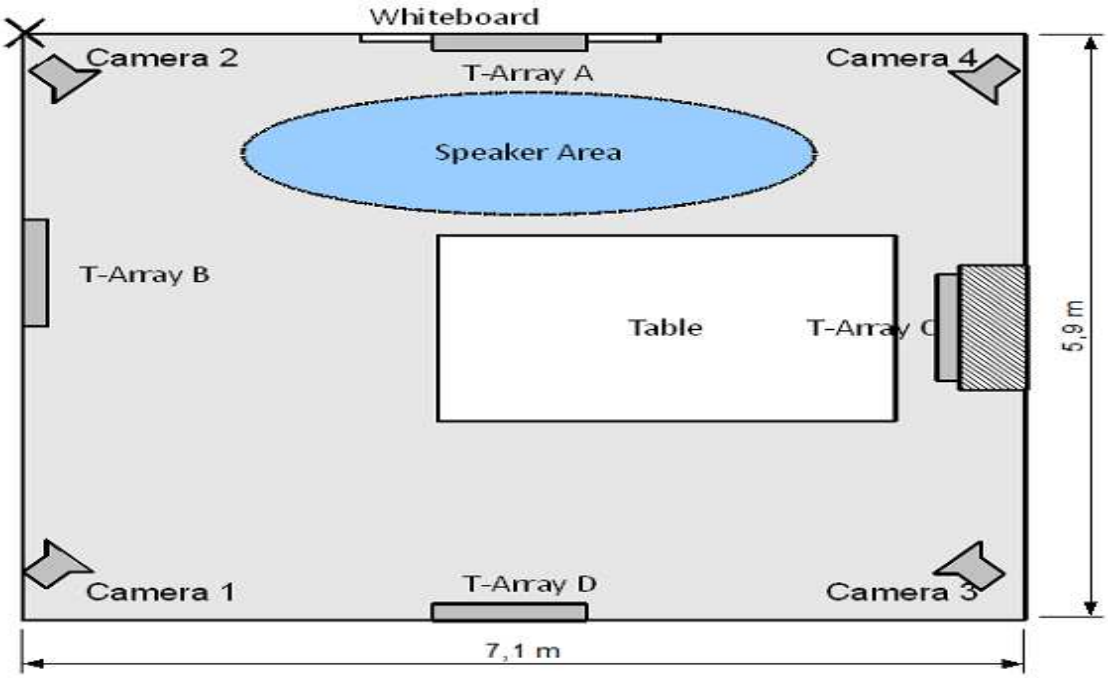
استفاده از شبکه‌ای از میکروفن‌ها، فضای اطراف اتاق سمینار را بخوبی پوشش داده، وسیگنال با کیفیت بالایی را در اختیار ما می‌گذارد. حداقل سنسور بکار رفته در هر یک از اتاق‌های هوشمند CHIL عبارتند از:

- آرایه میکروفنی ۶۴ کانال (NIST MARK III)
- حداقل ۳ آرایه میکروفنی به فرم معکوس T شکل
- میکروفن‌های یقه‌ای

میکروفن‌های رومیزی بجز سنسورهای صوتی، معمولا از چندین دوربین هم برای پوشش دادن فعالیت‌های تصویری استفاده می‌گردد. آرایه T شکل در شکل زیر نشان داده شده است، که فرم خاص آن برای تشخیص بهتر جهت و تخمین سه‌بعدی بهتر منبع صوتی، طراحی گردیده است:



شکل ۳-۷: شکل آرایه Tشکل در UKA CHIL room



شکل ۳-۸: شکل اتاق سمینارUKA CHIL موجود در University of Karlsruhe

فاصله میکروفن‌ها از هم با توجه به توازن میان تفکیک بالا و کاربرد در روش‌های GCC باید در نظر گرفته شود؛ افزایش فاصله میکروفن‌ها از هم باعث افزایش تفکیک تخمین TDOA و کاهش حساسیت DOA نسبت به نویز می‌گردد.

از NIST MARK III معمولا برای شناسایی صحبت و از میکروفن‌های یقه‌ای برای آوانویسی و از میکروفن‌های رومیزی معمولا برای شناسایی صحبت کننده و Speech activity detection (SAD) استفاده می‌گردد.

شکل فوق Distributed Microphone Network (DMN) درمربوط به اتاق UKA را نشان می‌دهد که ابعاد اتاق ۳×۵.۹×۷.۱ متر می‌باشد.

سخت افزار بکار رفته در این لابراتوار عبارتند از:

- ۴ عدد آرایه معکوس T شکل که هرکدام تشکیل شده است از ۴ عدد SHURE microflex 391W/O میکروفن تک‌جهتی
- ۴ میکروفن رومیزی SHURE microflex 391W/O
- آرایه میکروفنی خطی ۶۴ کانال NIST MARK III
- ۴ عدد میکروفن یقه‌ای Countryman E6WD5

تعداد بالای میکروفن‌ها، نیازمند یک سیستم جهت اداره دریافت این حجم بالای اطلاعات دارد؛ همزمانی سیگنال‌های گرفته شده هم مقوله مهمی در عملکرد الگوریتم‌ها بشمار می‌رود، و در این لابراتوار سیستم کاملی بدین منظور طراحی شده است که برای اطلاعات بیشتر می‌توانید به مقالات مربوط به پروژه CHIL-CLEAR مراجعه شود.

برای تعیین‌محل منبع صوتی برای هر فریم زمانی می توان تنها از جفت میکروفن‌های افقی استفاده کرد. فرکانس نمونه‌برداری KHZ ۴۸ می‌باشد.

۳-۷- جمع‌بندی

در این فصل در مورد روش‌های TDE بحث گردید. در ابتدا روش‌های تعیین تاخیر زمانی بین جفت میکروفن‌ها به طور کامل معرفی شدند و درادامه روش‌های مستقیم TDE مانند روش SRP-PHAT معرفی شدند که از جمله روش‌های مورد نظر برای پیاده‌سازی دز این تحقیق می‌باشند. دلایل برتری روش‌های مستقیم به روش‌های TDOA مورد بررسی قرار گرفت.

در انتهای فصل دیتابیس‌های استفاده شده در این تحقیق معرفی شدند. در فصل بعدی با تاکید بروی روش‌های مستقیم TDE الگوریتم‌های پیشنهادی برای بهبود عملکرد سیستم‌های ASL ارایه می‌گردد و در نهایت پیاده‌سازی روش‌های مختلف پیشنهادی را خواهیم داشت.

فصل چهارم

الگوریتم های پیشنهادی برای تعیین محل صوتی چندین صحبت کننده به طور همزمان

۴-۱- مقدمه

در فصل قبل مرور کاملی بر روش‌های بر پایه TDE داشتیم. در سال‌های اخیر این روش‌ها بیشتر مورد توجه بوده و همان‌طور که دیدیم شامل دو دسته، دو مرحله‌ای(غیر مستقیم) و یک‌مرحله‌ای (مستقیم) بودند. روش‌های مستقیم عملکرد بهتری در برابر نویز و انعکاسات از خود نشان می‌دادند و به دلیل کار کردن با فریم‌های زمانی کوتاه‌تر، از قدرت بیشتری در ردیابی صوتی بهرمند هستند و قابلیت تعیین محل چندین منبع صوتی را دارا می‌باشند.

در این فصل بحث را با تاکید بروی روش‌های مستقیم، برای تعیین محل لحظه‌ای چندین منبع صوتی ادامه می‌دهیم. کلیه نوآوری‌ها و الگوریتم‌های پیشنهادی و پیاده‌سازی این الگوریتم‌ها و نتایج حاصله در این فصل مطرح می گردد.

از اطلاعات جهت سر صحبت‌کننده برای بهبود عملکرد سیستم تعیین محل صحبت‌کننده در برابر انعکاسات استفاده می‌شود و تابعی پیشنهادی را OPROD-PHAT می‌نامیم.

پیاده‌سازی‌ها شامل دو بخش اصلی، پیاده‌سازی شبکه‌ای و پیاده سازی بخش‌بندی می‌باشد. در ابتدا فضای اتاق سمینار به فرم شبکه‌ای از نقاط گسسته در نظر می‌گیریم، که هر نقطه می‌تواند یک محل تخمینی برای صحبت کننده باشد. تعیین محل لحظه‌ای صحبت‌کننده در هر فریم‌زمانی، با استفاده از تابع پیشنهادی انجام می‌شود. در ادامه برای تعیین محل همزمان چندین صحبت‌کننده سیستمی پیشنهاد می‌گردد که در آن با استفاده از اطلاعات ناشی شده از توابع بهبود یافته همبستگی بین سیگنال‌های گرفته‌شده از یک جفت میکروفن، تعداد افراد در هر صحبت در هر فریم‌زمانی تعیین شده سپس با این دانش به تعیین محل می‌پردازیم. به عبارت دیگر سیستم تعیین محل لحظه‌ای را چندین بار مطابق با تعداد افراد فعال تکرار می‌کنیم.

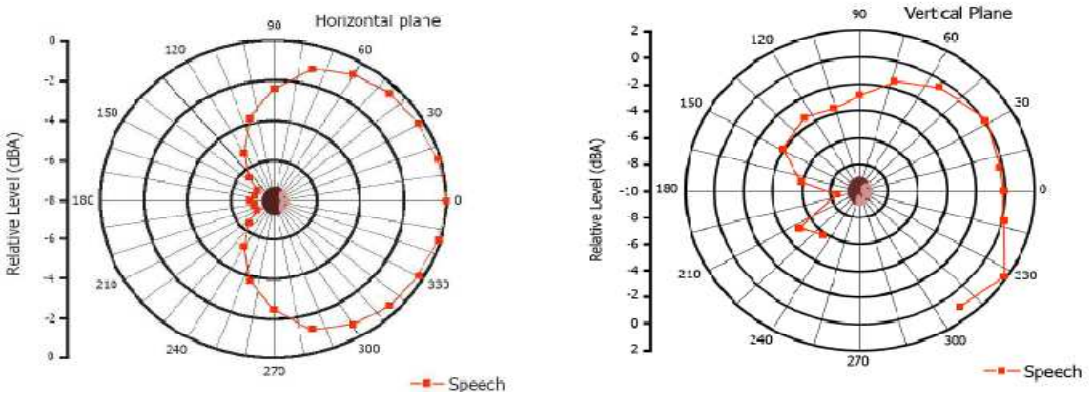
در بخش دوم پیاده‌سازی‌ها، برای افزایش عملکرد سیستم در برابر نویز و انعکاسات و افزایش سرعت جستجو برای تعیین محل، با تقسیم فضای اتاق به چندین بخش، یک میزان فعالیت صوتی را برای هر

بخش با استفاده از توابع مختلف، تعریف می‌کنیم. در ادامه با استفاده از یک آستانه تطبیقی در هر فریم زمانی بخش‌های فعال را مشخص می‌کنیم و تنها در قسمت‌های فعال به تعیین محل می‌پردازیم. و در نهایت هم با معرفی متریک‌های مختلف به ارزیابی الگوریتم‌های مختلف می‌پردازیم.

۴-۲- تقویت تعیین محل منبع صوتی با اطلاعات جهت‌یابی منبع صوتی

یکی از مسایلی که در کار ASL می‌تواند حائز اهمیت باشد، جهت‌یابی منبع صوتی می‌باشد، که اجازه می‌دهد تا در یک اتاق سمینار بهتر بفهمیم که فرد صحبت کننده چه کاری می‌کند و یا اینکه به کجا اشاره دارد. داشتن دانشی در مورد جهت منبع صوتی می‌تواند تخمین بهتری برای محل منبع صوتی از اطلاعات بسیاری که از شبکه‌ای از میکروفن‌ها که در کل اتاق سمینار قرار گرفته‌اند، فراهم آورد.

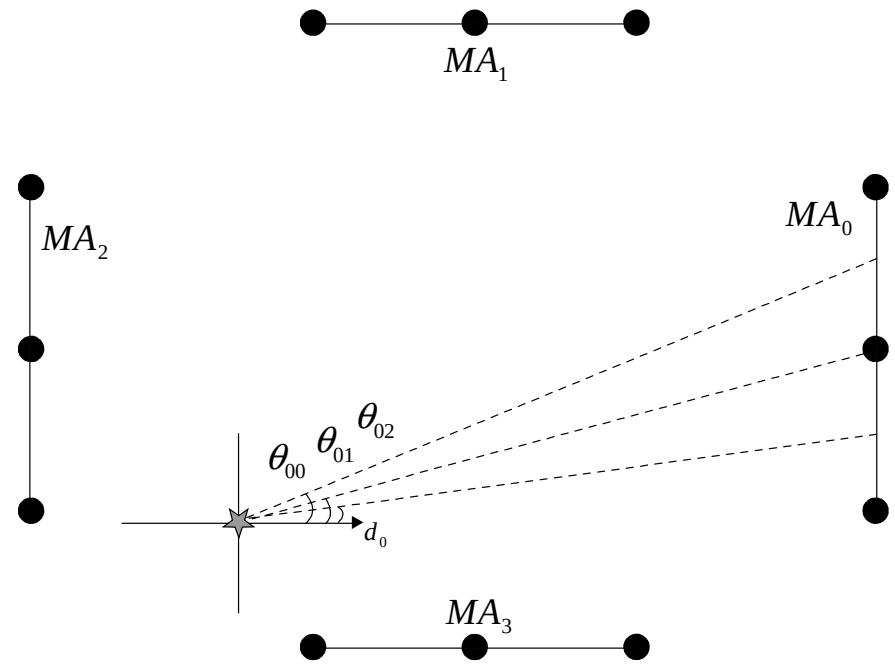
در[۳۷] رهیافتی برای Source Orientation Estimation (SOE) بر اساس انرژی صورت گرفته است، و کلیت آن بر پایه این مساله است که انرژی تشعشع یافته در پشت سر گوینده، کمتر از انرژی جلوی صورت آن است. همان‌طور که در شکل زیر دیده می‌شود، نشان داده است که صوت به طور یکنواخت در تمام جهات منتشر نمی‌شود و بیشتر انرژی آن به سمت روبرو خواهد بود:



شکل ۴-۱: انتشار امواج صوتی در جهت افقی و عمودی[۳۸]

۴-۲-۱ سیستم پیشنهادی برای بهبود عملکرد تابع تعیین محل با استفاده از اطلاعات ناشی از جهت سر صحبت‌کننده (تابع OPROD-PHAT)

در روش‌های مستقیم ما تنها از روی مقدار ماکزیمم ناشی از جفت میکروفن‌ها در تابع SLF مربوطه موقعیت منبع را تخمین می‌زنیم، و هیچ دانشی در مورد جهت منبع در این خصوص به کار نمی‌بریم، ولی اگر تاکید بیشتری روی آن دسته از میکروفن‌هایی که موج را به طور مستقیم دریافت می‌کنند داشته باشیم و سیگنال ناشی از سایر میکروفن‌ها که انعکاسات را دریافت می‌کنند را تضعیف کنیم، بسیار سودمندتر خواهد بود و با دقت بالاتری محل منابع صوتی را تخمین می‌زنیم؛ چراکه با این روش تاثیر انعکاسات را تا حدودی از بین برده‌ایم. با توجه به این نکته، در ادامه در مورد سیستم پیشنهادی برای بهبود عملکرد الگوریتم ASL با استفاده از اطلاعات ناشی از SOE بحث خواهیم نمود.



شکل ۴-۲: نمایش چگونگی عملکرد تابع ORIENTED-SLF برای تشخیص جهت منبع

فرض کنید M جفت میکروفن داریم، با در نظر گرفتن N جهت از پیش تعیین شده $d = 0,...N-1$ φ_d و با توجه به شکل ۴-۲، برای یک نقطه تخمینیP برای منبع، M خط، از محل منبع به سمت جفت میکروفن‌ها کشیده می‌شود. زاویه تشکیل شده بین هر جهت d و خطوط کشیده به سمت جفت‌میکروفن‌ها را با θ_{id} نشان می‌دهیم. تابع تعیین کننده جهت با فرض استفاده از SRP-PHAT برای SLF برای نقطه p و جهت d به فرم زیر خواهد بود:

$$OSRP-PHAT(t,p,d)=\sum_{i=0}^{M-1} R_P^{GCC-PHAT}(\Delta\tau_{i,P})\bullet\omega(\theta_{id}) \qquad (۱-۴)$$

و اگر از تابع PROD-PHAT استفاده کنیم خواهیم داشت:

$$OPROD-PHAT(t,p,d)=\prod_{i=0}^{M-1} R_P^{GCC-PHAT}(\Delta\tau_{i,P})\bullet\omega(\theta_{id}) \qquad (۲-۴)$$

که در آن‌ها $\omega(.)$ یک تابع وزنی و θ_{id} یک زاویه میان خط عبوری از P به سمت جفت میکروفن ها و جهت d مطابق شکل می‌باشد. تابع وزنی به فرم یک تابع گوسی برای تاکید بیشتر روی جهت مورد نظر و به صورت زیر تعریف می‌شود:

$$\omega(\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-\theta^2}{2\sigma^2}} \qquad (۳-۴)$$

ما برای تعیین اینکه جهت سر صحبت‌کننده کدام است با استفاده از تابع وزنی گوسین برای هر نقطه تخمینی منبع در تمامی جهات از پیش تعیین شده مقدار خروجی تابع مربوطه را بدست می‌آوریم. بنابراین جهت سر صحبت کننده به صورت زیر معرفی می‌شود:

$$\hat{d}=\arg\max\left[OSRP-PHAT(P,d)\right] \qquad (۴-۴)$$

همان طور که آورده شد، با استفاده از جهت‌یابی منبع صوتی، تنها بروی آن میکروفن‌هایی تاکید بیشتری می‌شود که در مسیر مستقیم انتشار موج قرار گرفته‌اند و سایر میکروفن‌ها را که تحت تاثیر انعکاسات قرار دارند را تضعیف می‌کنیم؛ حال محل تخمینی منبع صوتی را بر حسب یک تابعی وزنی SLF معرفی می‌کنیم.

$$\hat{P} = \arg\max_{p\in L}\{\arg\max_{d\in\theta} OPROD - PHAT\,(P,d)\} \qquad (۵-۴)$$

بنابراین با استفاده از این تابع هم می‌توان با اطلاع از موقعیت منبع، جهت آن را مشخص کرد؛ و هم‌اینکه تخمین بهتری برای منبع صوتی داشته باشیم.

در این میان نیازمند انتخاب پارامترهایی نظیر σ , N می‌باشیم. انتخاب پارامتر N تا حدود زیادی به ابعاد اتاق کنفرانس بستگی دارد، تعداد جهات در نظر گرفته شده، باید به‌گونه‌ای باشد که با توجه به محل‌های قرار گیری جفت میکروفن‌ها و سیگمای انتخاب شده، کلیه جهات پوشش داده شود. اگر تعداد جهات خیلی زیاد باشد، برای داشتن تمایز با سایر جهات مقدار سیگما باید کوچک در نظر گرفته شود و تنها جفت میکروفن‌های که در راستای مستقیم سیگنال صحبت هستند تقویت شوند. و اگر تعداد N کوچک باشد باید مقدار پارامتر سیگما بزرگ انتخاب گردد، تا بتوانیم کل جهات را پوشش داده و از طرفی بتوانیم از اطلاعات جفت میکروفن‌های بیشتری استفاده نماییم.

در کل در انتخاب مقدار N و پارامتر سیگما، باید از طرفی تفکیک بین جهات در نظر گرفته شده، وجود داشته باشد و ما تنها از جفت میکروفن‌های مسیر مستقیم استفاده نماییم. و از طرف دیگر میزان اطلاعات دریافتی از جفت میکروفن‌ها به اندازه کافی باشد و تعداد زیادی از اطلاعات سیگنال‌های سایر جفت میکروفن‌ها را تضعیف نکرده باشیم.

ما با توجه به اتاق CHIL UKA مورد نظر در دیتابیس تعداد جهاتN برای این اتاق با توجه به موقعیت قرارگیری آرایه میکروفنی و محدوده انتخابی برای سیگما در نظر می‌گیریم.

هرچه مقدار واریانس بزرگتر باشد، میزان مشارکت جفت میکروفن‌های کناری را، زمانی که بروی یک جفت میکروفن تاکید می‌کنیم، افزایش می‌دهد. در کل با استفاده از آزمایش میزان این پارامترها را می‌توان به فرم زیر تعیین نمود:

$$1.5 \le \sigma \le 3 \quad N = 16 \qquad (۶-۴)$$

در جمع‌بندی این بخش می‌توان گفت: در فرکانس‌های مختلف، از متوسط گیری وزن‌دار جفت میکروفن‌ها که تمامی جهات را پوشش داده‌اند استفاده خواهیم کرد. و برای هر بخش تنها جهتی را که ماکزیمم مقدار را به ما بدهد، تقویت می‌کنیم و اطلاعات مربوط به جفت میکروفن‌هایی که در سایر جهات قرار گرفته‌اند را تضعیف می‌کنیم.

بنابراین با بکارگیری این روش ما به طور خودکار امواج انعکاسی را تضعیف کرده و تنها از اطلاعات مربوط به جفت میکروفن‌هایی که در مسیر مستقیم امواج رسیده از منابع صوتی بوده‌اند استفاده می‌کنیم.

۴-۳- تعیین محل افراد صحبت کننده با شبکه کردن فضای اتاق سمینار

در این بخش، هدف پیاده‌سازی روش‌های مستقیم TDE معرفی شده در فصل قبل می‌باشد. برای پیاده سازی عملی ناچاریم فضای پیوسته را با شبکه‌ای از نقاط، که به صورت متقارن کل فضای اتاق را پوشش می‌دهند، به فضای گسسته تبدیل کنیم. هر چه تعداد نقاط بیشتر باشد، تخمین دقیق‌تری خواهیم داشت و به همان نسبت زمان بیشتری را صرف خواهیم کرد تا تاخیر زمانی را از هر نقطه نسبت به کلیه جفت میکروفن‌ها محاسبه کنیم. این شبکه می‌تواند، در فضای کروی یا کارتزین در داخل اتاق سمینار در نظر گرفته شود.

در فصل قبل روابط کاملی را از تابع SRP-PHAT بیان کردیم، که در ادامه هدف پیاده سازی این

روابط و تعیین محل افراد با توجه به معادله (۳-۲۱) می‌باشد.در ادامه قصد داریم به پیاده‌سازی الگوریتم‌های رایج و پیشنهادی در مرحله قبل بپردازیم. در ابتدا عملیات با توجه به توابع موجود برای شرایطی که یک فرد در حال صحبت است، اجرا می‌گردد.

برای حالت‌هایی که چندین فرد به‌طور همزمان صحبت می‌کنند، الگوریتمی پیشنهاد می‌گردد، که در آن در ابتدا با استفاده از یک سیستم SNS فریم‌های silence توسط یک آستانه تطبیقی از سایر فریم‌ها جدا می‌گردد، که تا مقدار زیادی از حجم محاسبات مربوطه می‌کاهد.

در ادامه با استفاده از اطلاعات ناشی از تابع همبستگی متقابل بین سیگنال‌های دریافتی از یک جفت میکروفن، تعداد افراد صحبت‌کننده را مشخص کرده و در نهایت عملیات تعیین محل یک صحبت کننده را چندین بار با توجه به تعداد افراد در حال صحبت در هر فریم زمانی، تکرار می‌کنیم. جزییات و نحوه پیاده سازی‌ها در ادامه بررسی می‌شود.

۴-۳-۱ پیاده سازی شبکه‌ای تعیین محل صوتی یک فرد در اتاق سمینار

همان طور که در بخش‌های قبلی صحبت شد، در این قسمت می‌خواهیم با استفاده از نرم‌افزار مطلب الگوریتم‌های SRP-PHAT,PROD-PHAT را با استفاده از دیتابیس دایره‌ای معرفی‌شده در فصل سوم، مورد بررسی قرار دهیم. .

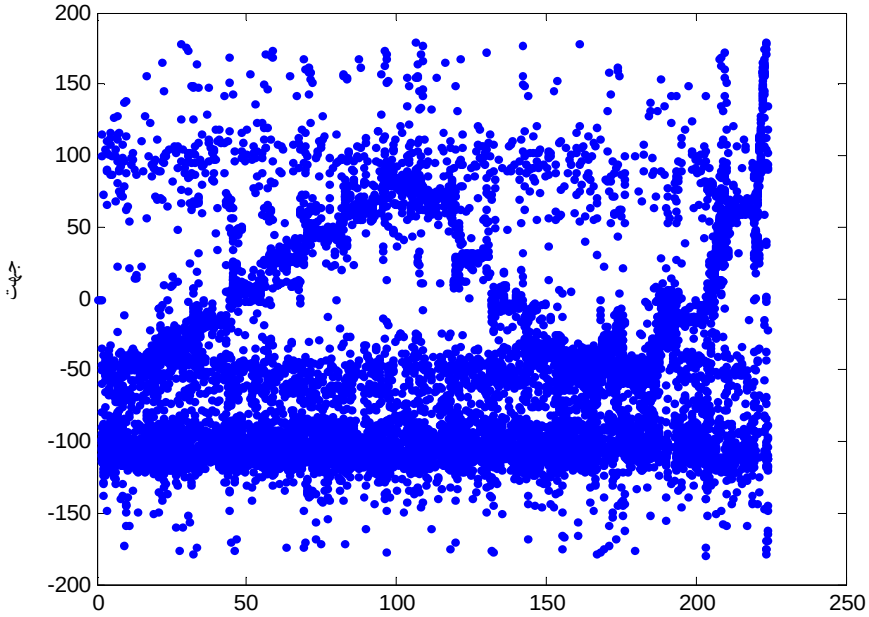
با توجه به محل قرارگیری آرایه میکروفنی در هر یک از دیتابیس‌ها، نیازمند در نظر گرفتن یک مبدا مختصات می‌باشیم؛ ما در هر دو مورد وسط میز واقع در اتاق را به‌عنوان مبدا در نظر می‌گیریم و شبکه‌ای از نقاط را در ارتفاع بالاتر از ان در ایجاد می‌کنیم، چراکه در ارتفاع کمتر از یک متر صحبت کننده‌ای قرار نخواهد گرفت و این کار باعث افزایش سرعت جستجو می‌گردد.

سرعت صوت را ۳۴۲ m/s در نظر می‌گیریم؛ با توجه به فرکانس نمونه‌برداری و پنجره۵۱۲ نمونه‌ای طول هر فریم زمانی تعریف می‌گردد. از فریم شیفت ٪۵۰ استفاده می‌کنیم. با توجه به رابطه TDOA

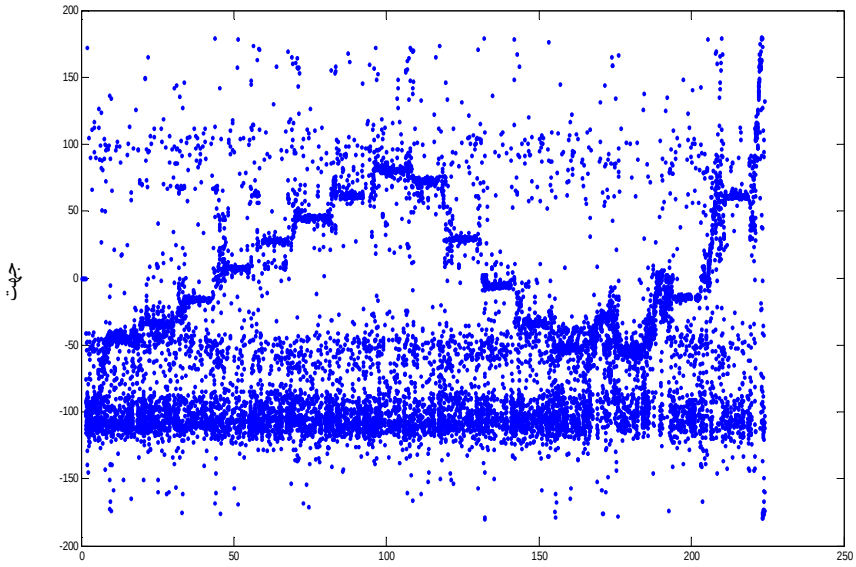
آورده شده در (۲-۱)، تاخیر زمانی ورود سیگنال صوتی از هر نقطه در شبکه مفروض را نسبت به هر یک از جفت میکروفن‌ها محاسبه می‌کنیم. سیگنال‌های دریافتی را فریم به فریم خوانده و رابطه (۳-۲۴) را برای بدست آوردن خروجی SRP-PHAT پیاده‌سازی می‌کنیم؛ و به همین ترتیب نقاط تخمینی را فریم به فریم ذخیره می‌کنیم. شبکه در نظر گرفته‌شده شامل ۱۸۰۰۰۰ نقطه در فضای کروی با مبدا مختصات معرفی‌شده می‌باشد؛ به‌طوری‌که در مختصات کروی و در فضای سه‌بعدی برای (r,φ,θ) به ترتیب ($25\times120\times60$) نقطه را برای شعاع ۳متری و ارتفاعی که تا ۸۰ درجه را پوشش می‌دهد و در نظر گرفتن ۱۲۰ صفحه با اختلاف ۳ درجه در میان، تا کل ۳۶۰ درجه را پوشش داده باشیم. در ادامه برای دو تابع SRP-PHAT,PROD-PHAT نتایج تخمینی را نشان خواهیم داد.

اولین آزمایش را با استفاده از دیتابیس دایره‌ای برای حالتی که تنها یک فرد در فضای اتاق ایستاده و پس از صحبت‌کردن، عمل خود را در نقاط مختلف اتاق تکرار می‌کند، اجرا می‌کنیم.

در نمودارهای آورده‌شده محور عمودی جهت و محور افقی فریم‌های زمانی را نشان می‌دهند. محل قرارگیری صحبت کننده در هر فریم‌زمانی با نقطه‌هایی بر حسب جهت قرارگیری نشان داده شده است. همان‌طور که شکل‌های زیر نشان می‌دهند تابع MULTI-PHAT خروجی مناسب‌تری را به نسبت SRP-PHAT داشته است، که ناشی از پیک‌های با ارتفاع بیشتر و تاثیرپذیری کمتر نویز زمینه می‌باشد. در شرایطی که یک صحبت کننده داریم تنها محل ماکزیمم پیک همبستگی ناشی شده از جفت میکروفن‌ها محل صحبت کننده را مشخص می‌کند، اما در مواردی که چندین فرد در حال صحبت به طور همزمان می‌باشند، نیازمند الگوریتمی می‌باشیم که در آن از طریق سایر پیک‌های تابع همبستگی میان جفت میکروفن‌ها، وجود چند صحبت‌کننده را تشخیص دهیم. مشکل اصلی مجهول بودن تعداد افراد صحبت‌کننده است، و در هر زمان نمی‌توانیم مثلاً به ازای چهار فرد، چهار پیک ماکزیمم را انتخاب کنیم؛ چراکه ممکن است افراد در برخی از فریم ها صحبت نکنند و ما پیک‌های ناشی شده از نویز و انعکاسات را برای تخمین محل صحبت‌کننده‌ها استفاده نماییم.



شکل ۴-۳: تعیین‌محل لحظه‌ای صحبت کننده توسط نقاط در هر فریم نشان داده شده است. محور افقی زمان و محور عمودی جهت را نشان می‌دهند. از تابع SRP-PHAT استفاده شده است.



شکل ۴-۴: تخمین لحظه‌ای محل صحبت‌کننده. یک صحبت کننده به طور ساکن در بخش‌های مختلف ایستاده و صحبت می‌کند.از تابع PROD-PHAT استفاده شده است.

برای رفع این مشکل الگوریتمی راپیشنهاد می‌دهیم که در آن فریم‌های که شامل صحبت هستند از فریم‌های silence جدا شوند. بنابراین تنها بروی فریم‌هایی عملیات تعیین محل را انجام می‌دهیم که شامل صحبت باشند. و سپس تعداد افراد صحبت‌کننده در هر فریم‌زمانی را با استفاده یک الگوریتم پیشنهادی بر مبنای همبستگی بین سیگنال‌های یک جفت میکروفن تشخیص می‌دهیم.

۴-۳-۲ جداسازی فریم‌های صحبت از غیر صحبت و الگوریتم پیشنهادی برای تعیین محل همزمان چندین صحبت‌کننده

در ابتدای این بخش به‌طورخلاصه سیستم‌های SNS رایج را مورد بررسی قرار می‌دهیم و در ادامه به معرفی سیستم پیشنهادی خود می‌پردازیم، که علاوه بر جداسازی سیگنال silence و غیر صحبت برای تعیین محل چندین صحبت‌کننده هم به‌کار می‌رود.

Voice activity detection (VAD) نقش مهمی را در بسیاری از سیستم‌های کاربردی ایفا می‌کند؛ در مخابرات و موبایل‌ها از آن می‌توان برای کاهش پهنای باند در شبکه برای ارسال بسته‌های صوتی تنها در زمانی که فعالیت صوتی وجود دارد، استفاده کرد؛ علاوه بر آن در سیستم‌های شناخت صحبت و صحبت کننده، در سیستم‌های تعیین محل صحبت کننده کاربردهای بسیاری دارد.

از مشخصات سیگنال صوتی می‌توان برای جداسازی صحبت از غیرصحبت استفاده کرد، مشخصاتی از قبیل: انرژی، محاسبه مجموع مربعات دامنه سیگنال در یک فریم زمانی، میزان عبور از صفر که در سیگنال‌های صوتی نرخ عبور از صفر، پایین‌تر می‌باشد.

سیگنال‌های صحبتی معمولا سیگنال‌هایی با فرکانس پایین می‌باشند و انتظار می رود که سهم بیشتر انرژی را در فرکانس‌های پایین داشته باشند،کمتر از ۲ Khz،در حالی که سیگنال‌های غیر صحبت توضیع یکنواخت‌تری دارند.بنابراین از نسبت انرژی فرکانس‌های پایین به انرژی کل باند فرکانسی می‌توان برای عملیات SNS استفاده کرد.

در طراحی یک سیستم SNS مشخص کردن ویژگی که به طور موثر فریم‌ها را به دو بخش صحبت و غیرصحبت تبدیل کند بسیار مهم است. در یک اتاق کنفرانس وجود صداهایی مانند: صدای زنگ تلفن همراه، صدای بسته شدن در، صداهای ناشی شده از افراد حاضر مثل خندیدن، سرفه کردن، نفس نفس زدن و... باعث می‌شود که سیستمی برای جداسازی صحبت از غیر صحبت برای کار تعیین محل استفاده کنیم.

در انتخاب یک سیتم SNS ، انتخاب یک ویژگی مناسب با توجه به ساختار زمانی یا طیفی سیگنال باید صورت گیرد؛ در کل دو این سیستم‌ها را می‌توان به دو بخش تقسیم کرد:

- *انتخاب یک مقدار آستانه تعریف شده توسط کاربر*: ویژگی‌های صوتی از هر فریم سیگنال صحبت استخراج شده و سپس با یک مقدار آستانه برای طبقه‌بندی فریم زمانی استفاده می‌شود. این ویژگی‌ها شامل: انرژی، نرخ zero-crossing ، pitch، و استفاده از اطلاعات تابع همبستگی و ...می‌باشد.[۳۹]
- *الگوریتم‌های آموزش ماشین و شناسایی الگو:* در این پروسه به تخمین پارامترهای مدل برای سیگنال‌های صحبت و غیرصحبت با استفاده از دیتاهای آموزشی می‌پردازیم . ویژگی معمول استفاده از ظرایب طیفی شبیه MFCCs, PLPs می‌باشد.

  همچنین از ترکیب سیستم‌های SNS و استفاده از SVM، ANN GMM, HMM برای جداسازی پدیده‌های مختلف صوتی در داخل یک اتاق سمینار استفاده می‌شود.[۴۰]

این نوع از الگوریتم‌های جداکننده به دو فرم پارامتریک و غیر پارامتریک وجود دارند. در نوع پارامتریپک ما نیازمند دانستن فرم توزیع دیتاها می باشیم، مثلا با فرض اینکه دیتاها توزیع گوسی دارند، از فاصله ماهالونوبیسی می‌توان استفاده کرد؛ که درآن دیتاهای ورودی متعلق به کلاسی هستند که کمترین فاصله را از مرکز آن کلاس داشته باشد.

انواع غیر پارامتریک مانند ANN,SVM,HMM,GMM در ادامه به طور خلاصه تعریف می‌شوند:

ANN شامل واحدهای محاسباتی معروف به نرون‌ها می‌باشد که چندین سیگنال ورودی گرفته و اگر مجموع وزن‌دار سیگنال‌های ورودی از یک مقدار آستانه بیشتر باشد مقدار خروجی برابر یک می‌گردد.

SVM از دیتاهای اموزشی برای پیدا کردن یک صفحه بهینه که جداکننده کلاس‌های مختلف باشد، استفاده می‌کند. این صفحه ماکزیمم فاصله میان کلاس‌های مختلف آموزشی را دارد. اگر دیتاهای ورودی به طور خطی جداپذیر نباشند، نمونه‌های آموزشی ورودی به بعد بالاتری انتقال داده می‌شود، تا بتوان به طور خطی آن‌ها را از یکدیگر جدا نمود.

HMM ,GMM از جداکننده‌های آماری می‌باشند. در HMM احتمال اینکه فریم حال متعلق به کلاس مشخصی باشد با توجه به کلاس مربوط به فریم‌های قبلی تشخیص داده می‌شود. این احتمال به ماتریس حالت احتمال، انتقال از یک کلاس به کلاس دیگر و احتمال مقادیر ویژگی مشاهده شده بستگی دارد. در GMM فرض می‌شود که بردار ویژگی یک توزیع گوسی دارد و در آن چگالی احتمال هر کلاس با توجه به برخی از ویژگی‌های دیتاهای آموزشی تخمین زده می‌شود و از آن برای جداسازی کلاس‌ها استفاده می‌شود.

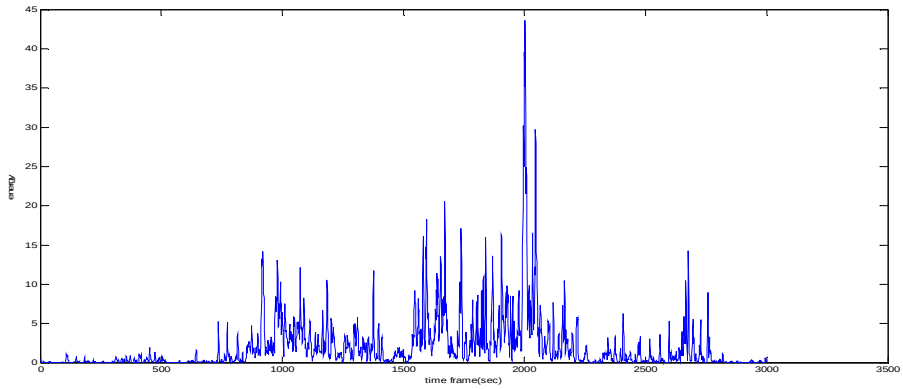
عملکرد سیستم‌های SNS به شرایط اتاق سمینار از نظر انتشار امواج صوتی، تعداد صحبت‌کننده‌ها و نوع و محل میکروفن‌ها بستگی دارد.

روشهای معرفی شده دو مشکل اصلی دارند: در شرایط ضعیف برای SNR روش آستانه چندان مناسب نمی‌باشد و تکنیک‌هایی که برای شناسایی الگو بکار می‌روند، نیازمند دیتاهای زیادی برای آموزش مدل و همچنین داشتن یک دانش اولیه در مورد نویز می‌باشد. و در هر فریم زمانی با تغییر شرایط محیطی نیازمند آموزش جدید می‌باشیم.

در ادامه بحث در مورد الگوریتم پیشنهادی SNS و تعیین تعداد افراد صحبت‌کننده به‌طور همزمان، بحث می‌گردد.

با توجه به آنکه تعداد زیادی از فریم ها در یک کنفرانس می‌توانند بی‌صدا باشند، با جداسازی آنها می‌توان تا مقدار زیادی از حجم محاسبات کاست و از انجام عملیات ASL را بروی فریم‌هایی که شامل صحبت‌کننده نمی‌باشند، خودداری کرد. بنابراین در ابتدا از ویژگی انرژی برای جداسازی فریم‌های بی‌صدا استفاده می‌شود. میزان آستانه برای جداسازی این قسمت باید در حد پایینی انتخاب گردد، یکی از مشکلات استفاده از انرژی در عملیات SNS ، عملکرد ضعیف این سیستم در برابر نویز محیط می‌باشد؛ در ادامه یک آستانه تطبیقی به‌کار گرفته می‌شود، که ما از میانگین وزن‌دار چند فریم بی‌صدای قبلی برای آپدیت کردن مقدار آستانه تطبیقی استفاده می‌کنیم.

در انتخاب مقدار آستانه تطبیقی در این مرحله در ابتدا یک مقدار اولیه برای آستانه در نظر گرفته می‌شود سپس در انتهای یک بلوک از فریم‌های زمانی، از مقدار انرژی چندین فریم که بی‌صدا تشخیص داده شده اند، استفاده کرده و با استفاده از تابع وزندار زیر برای انتخاب مقدار جدید آستانه استفاده می‌گردد:

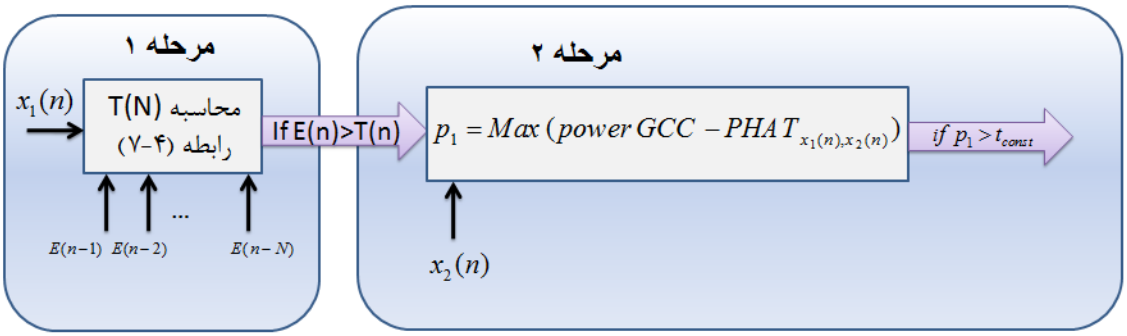


شکل ۴-۵ :انرژی سیگنال صحبت. در فریم‌هایی که بی‌صدا می‌باشند میزان انرژی کم‌تر است.

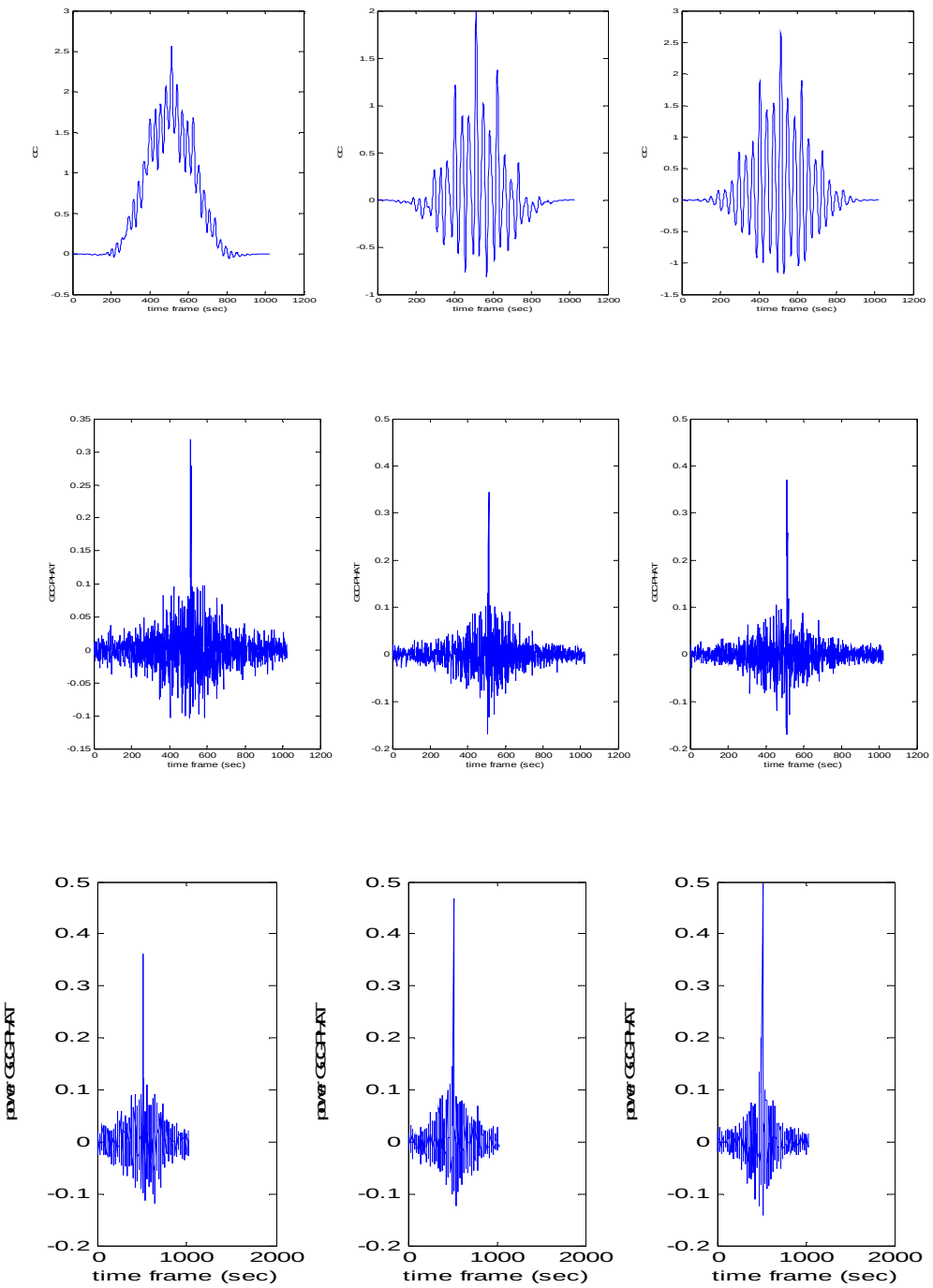
درشکل ۴-۵ انرژی مربوط به فریم‌های مختلف درشرایطی که افراد مختلف در اتاق کنفرانس در حال صحبت‌کردن می‌باشند نشان داده شده است. مقدار اولیه آستانه انرژی را می‌توان با توجه به شکل درنظر گرفت.

$$E(n)=\sum_{m=0}^{N-1}[x^2(n+m)]$$
$$T(n)=\sum_{i=1}^{N}\beta^i E(n-i)+(1-\beta)E(n) \qquad (۴-۷)$$

انتخاب مقدار β وابسته به میزان مشارکت فریم‌های قبلی برای تعیین تطبیقی مقدار آستانه در فریم زمانی حال، می‌باشد. برای‌اینکه انتخاب این فاکتور تنها به انرژی فریم بی‌صدا زمان حال وابسته نباشد، ما با استفاده از فاکتور β تعدادی از فریم‌های بی‌صدا اخیر را در انتخاب این مقدار آستانه تطبیقی سهیم می‌کنیم، تا از حساسیت سیستم نسبت به نویز لحظه‌ای کاسته باشیم. مقدار این فاکتور می‌تواند $0 \le \beta \le 1$ در نظر گرفته شود. هر چه به صفر نزدیک‌تر انتخاب گردد، از میزان مشارکت سایر فریم‌ها کاسته می‌گردد.



شکل ۴-۶: الگوریتم پیشنهادی SNS و تعیین محل چندین صحبت‌کننده به طور همزمان



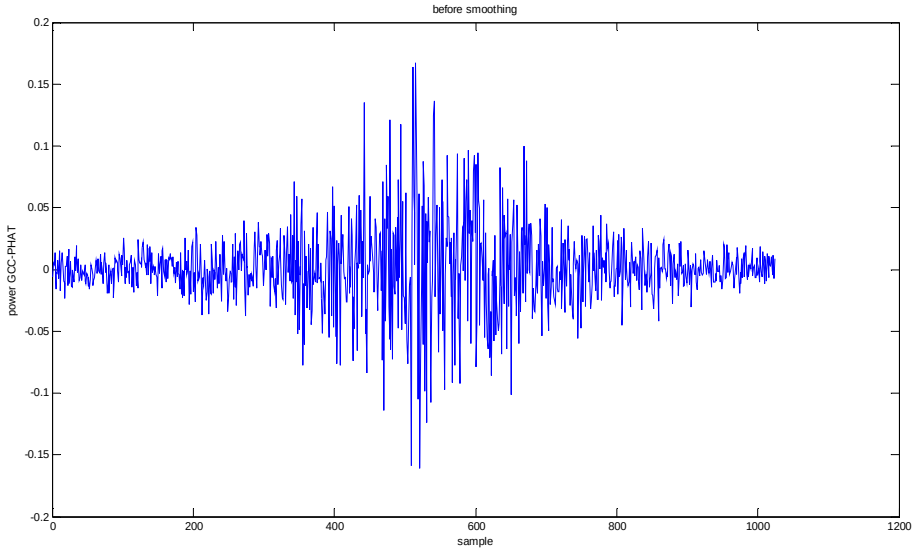
شکل ۴-۷: تابع همبستگی متقابل(CC), تابع GCC-PHAT و تابع power GCC-PHAT به ترتیب برای سه فریم از بالا به سمت پایین، نشان داده شده است.

پس از جداسازی فریم‌های بی‌صدا ، فریم‌های باقی مانده انرژی بالاتری را نسبت به فریم‌هایی که بی‌صدا انتخاب شده‌اند، داشته‌اند. این انرژی می‌تواند از نویز و انعکاسات ناشی شده باشد؛ بنابراین در ادامه برای تقویت عملیات SNS می‌توانیم از پیک همبستگی بین جفت میکروفن‌ها نیز برای تشخیص و جداسازی صحبت از نویز استفاده نمود.

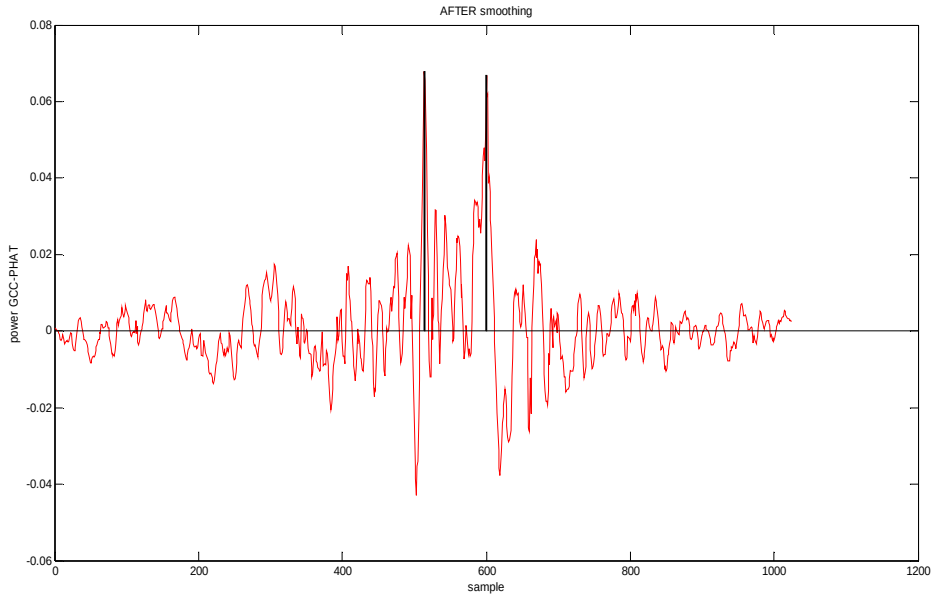
به عنوان جمع‌بندی الگوریتم پیشنهادی، همان طور که در شکل ۴-۶ دیده می‌شود، در کل شامل دو مرحله می‌باشد، در مرحله اول با استفاده از یک آستانه تطبیقی توسط تابع انرژی تنها فریم‌های بی‌صدا را جدا نمودیم. فریم‌های باقی‌مانده شامل صحبت و نویز محیطی می‌باشد که برای جداسازی آنها از پیک تابع POWER GCC-PHAT مربوط به یکی از جفت میکروفن‌ها استفاده خواهیم کرد.. برای تقویت الگوریتم SNS مرحله قبل از خاصیت CC برای حذف سیگنال‌های غیر صحبت باقی‌مانده استفاده می‌کنیم. درشکل ۴-۷ برای سه فریم متوالی در ردیف بالا CC توسط دو میکروفن گرفته شده است. در صورت وجود صحبت در آن فریم زمانی، پیک تابع همبستگی نسبت به سایر پیک‌های ناشی‌شده از نویزهای زمینه قابل تشخیص خواهد بود. همان‌طور که از فصل قبلی در مورد تابع وزنی PHAT صحبت شد، این تابع به خوبی از تاثیر انعکاسات که پیک‌های مشابه پیک اصلی را تولید می‌کنند، می‌کاهد. بنابراین در ردیف دوم از تابع GCC-PHAT استفاده شده است.

همان‌طور که دیده می‌شود، پیک اصلی به خوبی از سایرین جدا شده است. با استفاده از یک آستانه ثابت به‌آسانی می‌توان در مورد وجود سیگنال صحبت در فریم‌های باقی‌مانده در مرحله دوم تصمیم‌گیری نمود. در ادامه برای تقویت پیک‌های تابع GCC-PHAT پیشنهاد می‌شود از توان دوم سیگنال صحبت برای تابع همبستگی استفاده گردد.

$$G_{P\ x_1,x_2}^{(PH,t)}[t] = \frac{(X_1^{(t)}[t])^2.((X_2^{(t)}[t])^2)^*}{\left|(X_1^{(t)}[t])^2.((X_2^{(t)}[t])^2)^*\right|} \qquad (۸-۴)$$

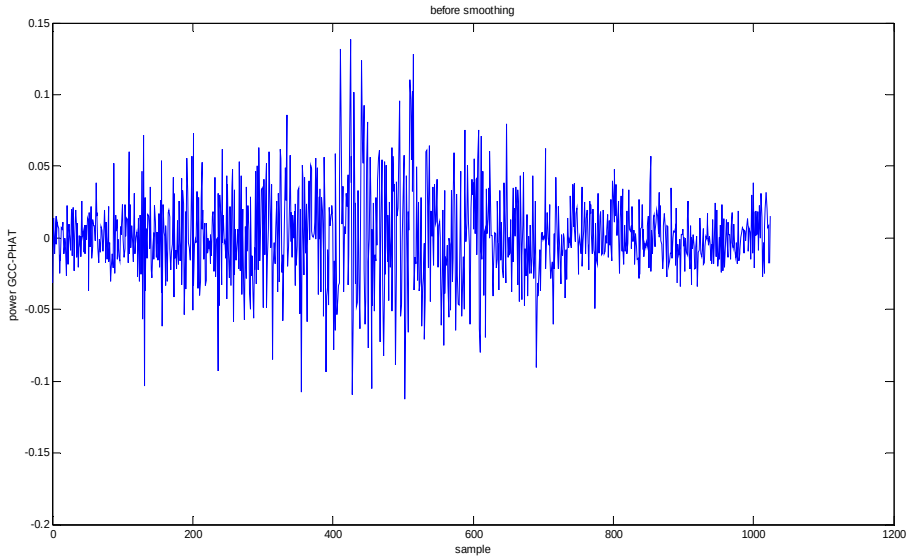


(الف)

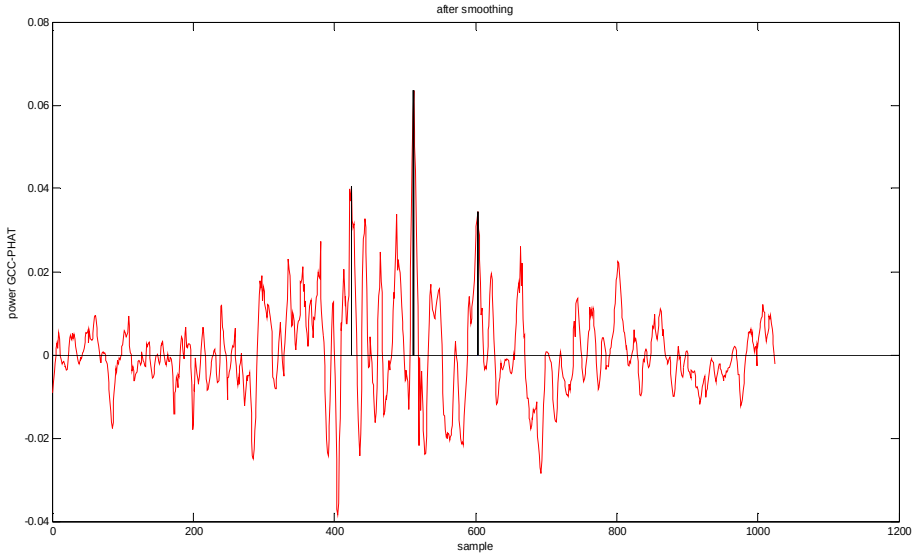


(ب)

شکل ۴-۸: نمودار (الف) مربوط به تابع Power GCC-PHAT می‌باشد، برای حالتی که دو نفر به‌طور همزمان در حال صحبت می‌باشند. نمودار (ب) شکل smooth شده فوق است.



(الف)



(ب)

شکل ۴-۹: نمودار (الف) مربوط به تابع Power GCC-PHAT می‌باشد، برای حالتی که سه نفر به‌طور همزمان در حال صحبت می‌باشند. نمودار (ب) شکل smooth شده فوق است.

اگر میزان ماکزیمم پیک POWER GCC-PHAT بین یک جفت میکروفن از یک آستانه بیشتر باشد، فرض بر آن است که آن پیک مربوط به یک صحبت‌کننده می‌باشد. حال برای تعیین این مساله که سایر پیک‌ها ناشی از نویز و انعکاسات بوده و یا مربوط به سایر افراد صحبت‌کننده می‌باشد، از یک نسبت بر حسب مقدار ماکزیمم پیک تعیین شده در مرحله قبل می‌توان استفاده نمود.

در شکل ۴-۸ و شکل ۴-۹ تابع همبستگی مربوط به دو فریمی که چندین فرد در حال صحبت‌کردن می‌باشند، نشان داده شده است، که در آن چندین پیک power GCC-PHAT را مشاهده می‌کنیم، که می‌تواند نشانگر وجود چندین صحبت‌کننده باشد.
همان‌طور که در شکل دیده می‌شود، ابتدا تابع Power GCC-PHAT را مشاهده می‌نماییم که با استفاده از توابع هموارساز، کاملاً هموار شده است، و در نهایت برای تعیین تعداد افراد صحبت‌کننده فعال در هر فریم‌زمانی، از یک نسبت بر حسب ماکزیمم پیک تابع استفاده شده است. همچنین قید فاصله بین دو پیک هم در نظر گرفته شده، چراکه پیک‌های نزدیک به‌هم در تابع همبستگی می‌تواند از یک صحبت‌کنند ناشی شده باشد.

هدف یافتن تعداد افراد فعال در هر فریم زمانی می‌باشد، که پس از تعیین این مساله، در چنین فریم‌هایی برای روش ASL استفاده شده، برای نمونه روش SRP-PHAT، پس از بدست آوردن محل اولین صحبت‌کننده،که معادل بزرگترین پیک در تابع SRP-PHAT است، محدوده مربوط به آن را حذف کرده و با استفاده از یک مقدار آستانه که نسبتی را بر حسب اولین پیک دارد، به دنبال سایر پیک‌های تابع SRP-PHAT هستیم، که محل سایر صحبت‌کننده‌ها را مشخص می‌نماید.

بنابراین پس از تعیین تعداد افراد صحبت کننده، که از طریق تعداد پیک‌های تابع Power-GCC-PHAT محاسبه می‌شود، محل سایر صحبت‌کننده ها را با حذف محل مورد نظر اولین صحبت کننده، محل سایرین را دقیقاً با تکرار الگوریتم بدست می‌آوریم. سایر پیک‌ها باید ارتفاع

مناسبی را با توجه به پیک ماکزیمم داشته باشند.

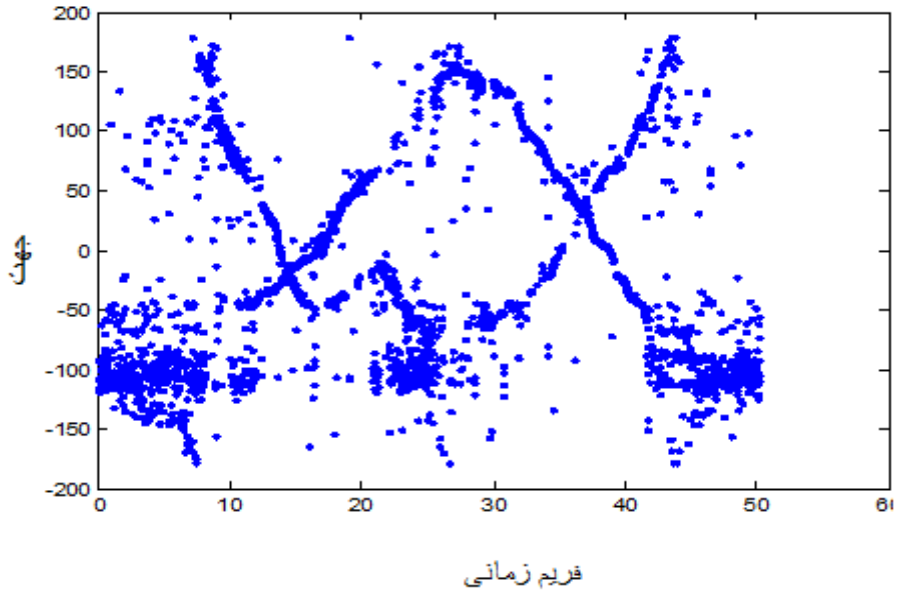
در ادامه ضرایب استفاده شده در الگوریتم فوق را که صورت تجربی محاسبه شده، بیان می‌شود ضریب مشارکت فریم‌های بی‌صدا را می‌توان معادل $\beta=$(۰.۴۵-۰.۷) و تعداد فریم‌های قبلی را $N =$ (۳-۵) در نظر گرفت. ضرایب استفاده شده در مرحله دوم شامل آستانه ثابت با توجه به نتایج بدست آمده و به طور تجربی برای تابع power GCC-PHAT برابر (۰.۰۶) در نظر می‌گیریم.و فاکتور مقدار مجاز پیک برای شرکت در الگوریتم تعیین عمل را نیز برابر (۰.۵۵-۰.۷۵) به صورت تجربی در نظر می‌گیریم.

۴-۳-۳ پیاده سازی شبکه‌ای تعیین محل چندین صحبت‌کننده به‌طورهمزمان

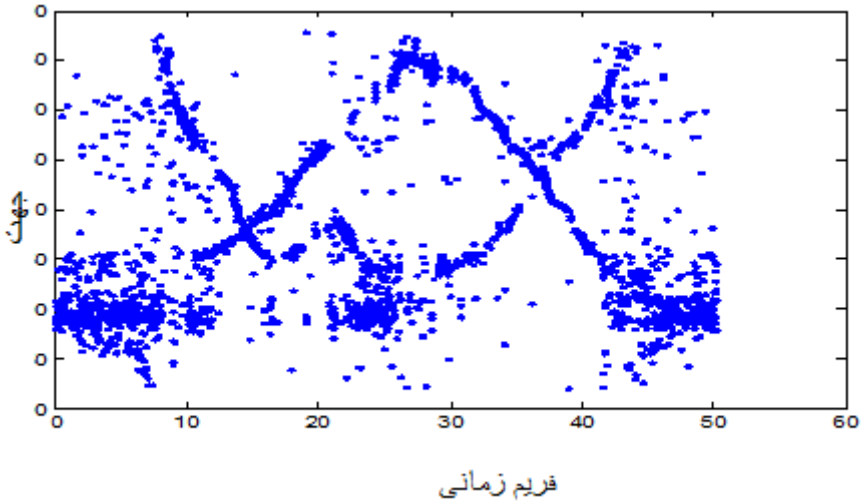
در این بخش توسط توابع SRP-PHAT,PROD-PHAT و تابع پیشنهادی OPROD-PHAT, الگوریتم پیشنهادی فوق پیاده‌سازی شده و نتایج را بررسی خواهیم کرد.

در اولین حالت، دو صحبت کننده در حال صحبت، در بخش‌های مختلف حرکت می‌کنند و کاملا همپوشانی صحبتی دارند. در هر فریم زمانی خروجی الگوریتم با یک نقطه در جهات مختلف نشان داده شده است. همان طور که نتایج درشکل ۴-۱۰ و شکل ۴-۱۱ دیده می‌شود افراد از روی هم تقاطع حرکتی دارند. این حالت از دیتابیس دایره‌ای تهیه شده است. نتایج هر دو تابع، همان‌طور که از شکل‌ها هم دیده می‌شود، تقریباً یکسان است. محور عمودی جهت قرارگیری افراد و محور افقی فریم های زمانی را نشان می‌دهند.

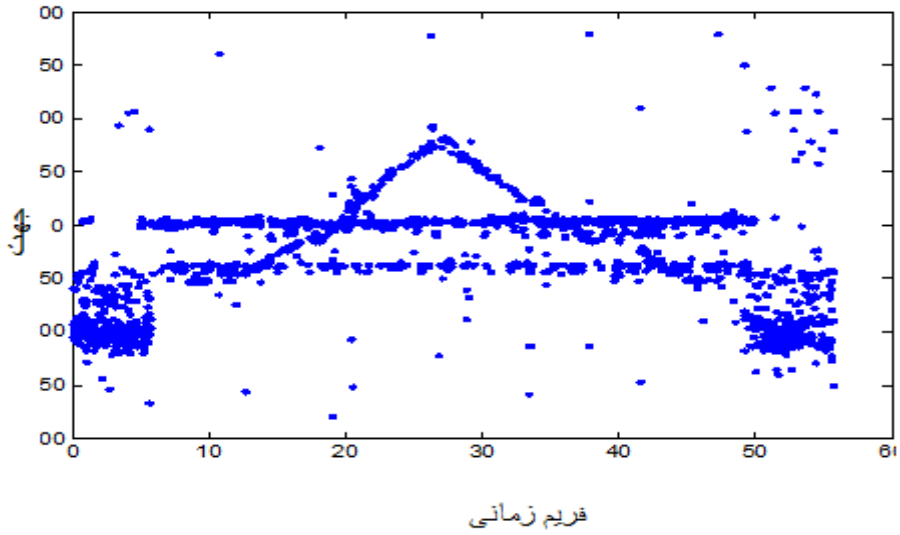
در ادامه حالتی را در نظر می‌گیریم که دو فرد در کنار یکدیگر نشسته و در حال صحبت می‌باشند و به طور همزمان فرد دیگری در حال حرکت به صحبت‌کردن مشغول است. کاملاً همپوشانی صوتی بین افراد وجود دارد. در شکل ۴-۱۲ و شکل ۴-۱۳ اجرای الگوریتم پیشنهادی را با توابع SRP-PHAT,PROD-PHAT مشاهده می‌نماییم.



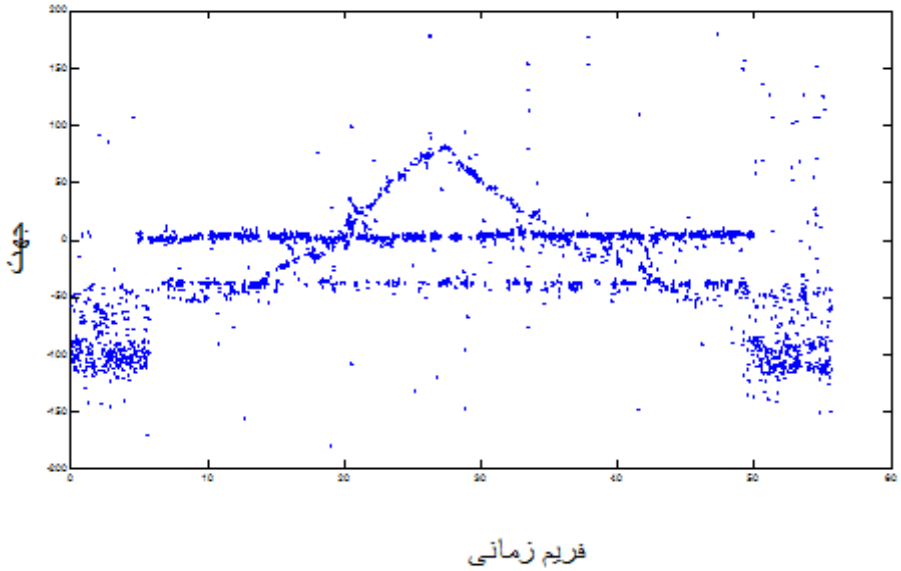
شکل ۴-۱۰: تخمین تعیین محل لحظه‌ای. دو صحبت کننده، در حال صحبت، در بخش های مختلف حرکت می کنند. از تابع SRP-PHAT استفاده شده است.



شکل ۴-۱۱: تخمین تعیین محل لحظه‌ای. دو صحبت کننده، در حال صحبت، در بخش های مختلف حرکت می کنند. از تابع PROD-PHAT استفاده شده است.



شکل ۴-۱۲: تخمین تعیین محل لحظه‌ای. حالتی که دو صحبت کننده نشسته و یک صحبت کننده در حال حرکت می‌باشد از تابع SRP-PHAT استفاده شده است.



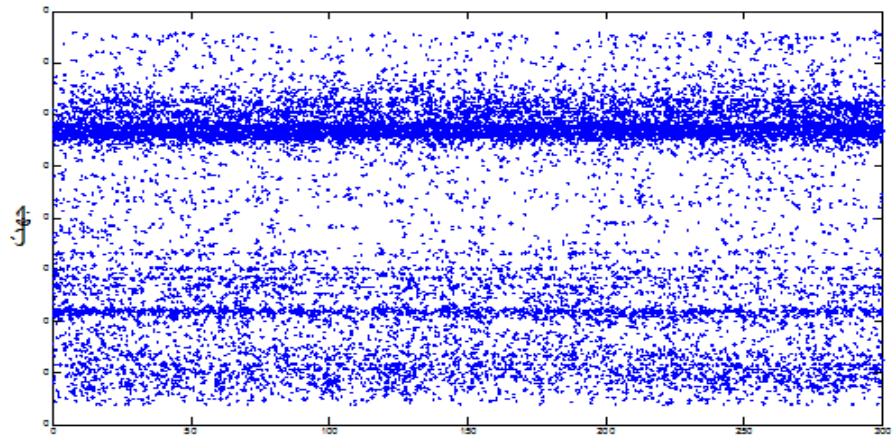
شکل ۴-۱۳: تخمین تعیین محل لحظه‌ای. حالتی که دو صحبت کننده نشسته و یک صحبت کننده در حال حرکت می‌باشد از تابع PROD-PHAT استفاده شده است

مواردی را که بررسی کردیم، مربوط دیتابیس دایره‌ای بود، برای تابع پیشنهادی OPROD-PHAT که از اطلاعات جهت سر برای تقویت سیگنال‌های مستقیم و حذف انعکاسات به‌کار می‌رفت، نمی‌توان از این دیتابیس کمک گرفت، چراکه اگر جهت سر صحبت‌کننده به سمت دیوارها باشد دیگر تابع پیشنهادی ما برای این دیتابیس کاربرد نخواهد داشت.

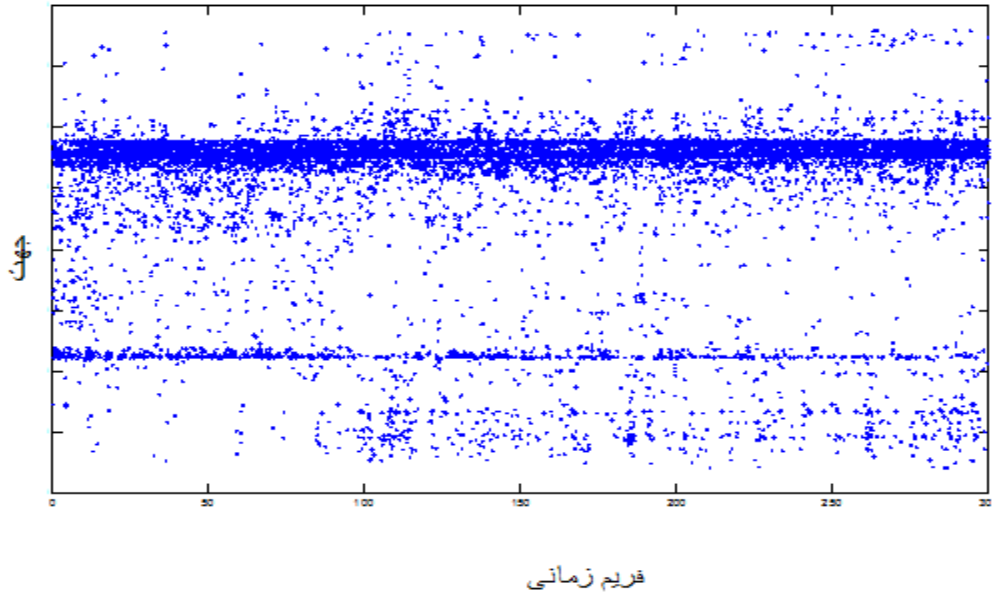
در ادامه قصد داریم از اطلاعات جهت منبع صوتی برای بهبود عملکرد الگوریتم پیشنهادی، همان‌طور که در بخش‌های قبلی در مورد آن صحبت شد، استفاده نماییم. اگر آرایه‌های میکروفنی مانند دیتابیس CHIL ، بروی دیواره‌ها قرار داشته باشند، برای استفاده از اطلاعات جهت سر صحبت کننده در تصحیح الگوریتم‌های ASL مناسب‌تر بوده و می‌توان در هر نقطه فرضی برای منبع، توان را در تمامی جهات بدست آورد و سپس تاکید بروی ماکزیمم توان را که مسلماً در جهت مستقیم انتشار است، داشته باشیم.

با استفاده از تابع PROD-PHAT و توسط دیتابیس CLEAR-CHIL، عملیات ASL را اجرا می‌کنیم. در ادامه در هر فریم زمانی تابع PROD-PHAT با توجه به جهت سر صحبت کننده وزن‌دهی شده و تاکید بیشتر روی جفت میکروفن‌هایی خواهد بود که در جهت سر صحبت کننده قرار می‌گیرند. برای افزایش سرعت ما تنها از ۱۶ جهت و سیگما ۲ برای تابع گوسین استفاده کردیم.

همان‌طور که از شکل ۴-۱۴ و شکل ۴-۱۵دیده می‌شود، استفاده از اطلاعات مربوط به جهت سر صحبت کننده در این دیتا بیس، با توجه به اینکه آرایه‌های میکروفنی بروی دیواره‌ها نصب شده‌اند به خوبی تاثیر انعکاسات را از بین برده و خروجی مناسبی را خواهیم داشت. در هر ms ۱۰ یک خروجی برای تعیین محل افراد صحبت‌کننده داریم. در این دیتابیس افراد اکثراً در محل خود نشسته و در حال صحبت کردن می‌باشند. که در تصاویر هم خروجی ها اکثراً بروی دو جهت تمرکز دارند، البته برخی از نقاط ناشی از نویز زمینه می باشند.



شکل ۴-۱۴: تخمین تعیین محل لحظه‌ای. اجرای الگوریتم PROD-PHAR با استفاده از دیتابیس CLEAR-CHIL . محور عمودی جهت و محور افقی فریم‌های زمانی را نشان می‌دهد.



شکل ۴-۱۵: تخمین تعیین محل لحظه‌ای. اجرای الگوریتم OPROD-PHAT با استفاده از دیتابیس CLEAR-CHIL. محور عمودی جهت و محور افقی فریم‌های زمانی را نشان می‌دهد.

در بخش فوق با استفاده از شبکه‌ای از نقاط که کل فضای اتاق سمینار را پوشش می‌داد، به تعیین محل چندین صحبت‌کننده به‌طورهمزمان پرداختیم. یک الگوریتم دومرحله‌ای برای حذف نویز و استفاده از توابع همبستگی برای تعیین محل هم‌زمان چندین صحبت کننده پیشنهاد شد.
یکی از مشکلات الگوریتم پیشنهادی فوق، عدم تعیین صحیح تعداد افراد صحبت‌کننده به طور همزمان، در شرایطی که سه یا بیشتر از سه نفر در حال صحبت‌کردن هستند، می‌باشد.

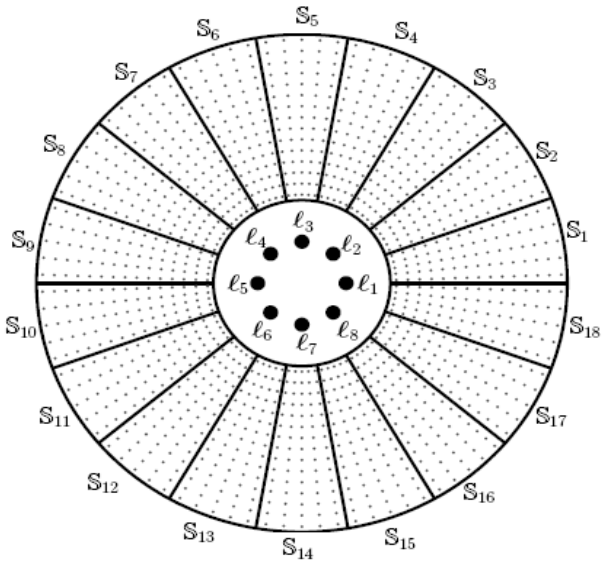
در بخش بعدی برای افزایش سرعت الگوریتم دربرابر پویش کل فضای اتاق سمینار، و همچنین بهبود عملکرد سیستم تعیین تعداد افراد فعال در هر فریم زمانی، از بخش‌بندی فضای اتاق استفاده شده و یک معیار فعالیت صوتی از تابع SRP-PHAT برای هر بخش تهیه می‌گردد، با متوسط‌گیری بروی کل فضای یک بخش تا حدودی ازحساسیت الگوریتم در برابر نویز می‌کاهیم؛ چراکه وجود منبع صوتی در یک نقطه از فضای اتاق تاحدودی در نقاط اطراف آن هم تاثیرگذار می‌باشد.

۴-۴- تعیین فعالیت صوتی در بخش‌های مختلف اتاق سمینار

در کل مشکل اصلی هر دو روش مستقیم CSSP, SRP-PHAT اتلاف زمان زیادی برای جستجو در کل فضای اتاق سمینار می‌باشد که می تواند فضای بسیار بزرگی باشد، که برای رفع این مشکل در [۴۱] پیشنهاد شده که فضای اطراف آرایه میکروفنی به Ns حجم متصل به هم تقسیم گردد و با این کار فضای جستجو را کاهش می‌دهیم. در این کار از آرایه میکروفنی دایره‌ای استفاده شده است، که نسبت به زاویه در تمامی جهات کاملاً یکنواخت می‌باشد. و فضای اطراف آرایه را همان‌طور که در شکل ۴-۱۶ دیده می‌شود به بخش‌های مختلف و یکسانی تقسیم می‌کنند.
رابطه مربوط به تقسیم فضای اطراف یک آرایه دایره‌ای، در (۴-۹) نشان داده شده است:

$$\forall S \in \{1...N_S\} \qquad \mathbb{S}_S = \{(\theta,\varphi,\rho) \mid 2\pi\frac{S-1}{N_S} \le \theta < 2\pi\frac{S}{N_S}, 0 \le \varphi \le \frac{\pi}{2}, \rho \ge \rho_0\} \qquad (۹-۴)$$

پس از تقسیم کردن فضای اتاق به چندین بخش، به دنبال تخمین میزان فعالیت صوتی در هر قسمت می‌باشیم. با این کار بجای جستجو در تمام نقاط اتاق سمینار، تنها خود را محدود به بدست آوردن میزان فعالیت صوتی در بخش‌های از پیش تعریف شده‌ای می‌کنیم، که صرفه‌جویی در زمان جستجو می‌نماید. علاوه بر آن همان طور که در ادامه می‌بینیم، چون وجود منبع صوتی را، نه در یک نقطه، بلکه در یک حجم فضایی، مورد بررسی قرار می‌دهیم، عملکرد بسیار بهتری در برابر نویز و انعکاسات از خود نشان خواهد داد.



شکل ۴-۱۶: تقسیم فضای اطراف آرایه دایره‌ای به بخش‌های مختلف

برای تعیین فعال بودن هر بخش در[۴۲] توان beamformer را در نقطه وسط هر بخش از فضا بدست آورده اند، ولی این روش، زمانی که منبع صوتی دقیقاً در مرز بین دو بخش باشد با مشکل مواجه می شود. برای حل این مشکل می توان میانگین فعالیت صوتی را در فضای حجمی مربوط به

هر بخش انجام دهیم. در ادامه با استفاده از توابع مختلف پیشنهادی در مرحله قبل، روابط مربوطه را بدست می‌آوریم.

۴-۴-۱ محاسبه روابط تعیین میزان فعالیت صوتی در هر بخش با استفاده از توابع مختلف

در این بخش قصد داریم که برای تعیین محل منابع صوتی، از الگوریتم‌های SRP-PHAT PROD-PHAT,OSRP-PHAT,OPROD-PHAT استفاده نماییم که در آن محل منابع صوتی معادل نقاطی است که ماکزیمم توان beamforming را داشته باشیم. در ابتدا کلیه روابط را برای تابع SRP-PHAT محاسبه کرده و در ادامه برای سایر توابع، از بسط روابط بدست آمده استفاده خواهیم کرد.

در فصل قبلی نشان دادیم که ماکزیمم کردن توان SRP-PHAT معادل مینیمم کردن یک ترم معرفی شده در رابطه(۳-۳۳) می‌باشد که برای ساده شدن محاسبات از آن استفاده می‌کنیم.

در ادامه در هر فریم زمانی و برای هر بخش یک معیار فعالیت باند پهن صوتی را بر پایه متوسط‌گیری از معیار $d(a,b)$ آورده شده در فصل قبل برای روش SPR-PHAT و مقایسه میان مقادیر مختلف بدست آمده برای هر بخش در هر فرکانس K، معرفی می‌کنیم.

محاسبه متوسط مقدار فعالیت صوتی از SRP-PHAT معادل با RMS گرفتن از معیار فوق بروی فضای هر بخش می‌باشد. به عبارت دیگر برای تعیین فعالیت صوتی برای یک بخش، یک فریم زمانی داده شده و در یک فرکانس گسسته ما RMS متریک را برای بردار فازی مشاهده شده و تمام بردارهای فازی تئوری در تمام نقاط موجود در آن بخش محاسبه می‌کنیم.[۵]

معیار مورد نظر را به فرم زیر معرفی کردیم:

$$d(a,b)=\sqrt{\frac{1}{N_q}\sum_{q=1}^{N_q}\sin^2(\frac{a-b}{2})} \quad (۱۰-۴)$$

در ادامه برای RMS گرفتن از این معیار خواهیم داشت:

$$D_s^{(t)}(k)=\{\int_{\mathbb{S}_s}[d(u^{(t)}(k)-u^{th}(k,\lambda))]^2 P_s(\lambda)d\lambda\}^{\frac{1}{2}} \quad (۱۱-۴)$$

در رابطه فوق $P_s(\lambda)$ بیانگر توزیع اولیه از محل احتمالی منبع صوتی فعال در هر بخش است؛ که برای مثال می‌تواند تواند توزیع یکنواخت و یا گوسی داشته باشد.

برای رسیدن به یک راه حل تحلیلی مناسب می‌بایست از فضای پیوسته به گسسته برویم، بنابراین بجای انتگرال از سیگما استفاده می‌کنیم:

$$D_s^{(t)}(k)=\sqrt{\frac{1}{N_v}\sum_{n=1}^{N_v}[d(u^{(t)}(k)-u^{th}(k,v_{s,n}))]^2} \quad (۱۲-۴)$$

$v_{s,n}$ مجموعه ای از N_v نقطه در فضای سه بعدی هر بخش است که برای تخمین توزیع پیوسته مرحله قبل بکار می رود.

در ادامه به ساده سازی روابط فوق به منظور پیاده سازی عملی می‌پردازیم:

$$(D_s^{(t)}(k))^2=\frac{1}{N_v}\sum_{n=1}^{N_v}\frac{1}{N_q}\sum_{q=1}^{N_q}\sin^2(\frac{u_q^{(t)}(k)-u_q^{(th)}(k,v_{s,n})}{2}) \quad (۱۳-۴)$$

با استفاده از $\sin^2(x)=\frac{1}{2}(1-\cos\ 2(x))$ داریم:

$$(D_s^{(t)}(k))^2=\frac{1}{2.N_q}\sum_{q=1}^{N_q}\{1-\frac{1}{N_v}\sum_{n=1}^{N_v}\cos(u_q^{(t)}(k)-u_q^{(th)}(k,v_{s,n}))\} \quad (۱۴-۴)$$

$$= \frac{1}{2.N_q} \sum_{q=1}^{N_q} \{1 - \Re[\frac{1}{N_v} \sum_{n=1}^{N_v} e^{j(u_q^{(t)}(k) - u_q^{(th)}(k, v_{s,n}))}]\} \quad (۱۵-۴)$$

$$= \frac{1}{2.N_q} \sum_{q=1}^{N_q} \{1 - \Re[e^{j(u_q^{(t)}(k)} \frac{1}{N_v} \sum_{n=1}^{N_v} e^{-ju_q^{(th)}(k, v_{s,n})}]\} \quad (۱۶-۴)$$

$$= \frac{1}{2.N_q} \sum_{q=1}^{N_q} \{1 - \Re[e^{j(u_q^{(t)}(k)} Z_s^*(k, \{v\})]\}$$
$$Z_s(K, \{V\}) = \frac{1}{N_v} \sum_{n=1}^{N_v} e^{ju_q^{(th)}(k, v_{s,n})} \quad (۱۷-۴)$$

و رابطه نهایی را به فرم زیر خواهیم داشت:

$$(D_s^{(t)}(k))^2 = \frac{1}{2.N_q} \sum_{q=1}^{N_q} \{1 - |Z_s^*(k, \{v\})| \cos[u_q^{(t)}(k) - \angle Z_s(k, \{v\})]\} \quad (۱۸-۴)$$

بنابراین تخمین میزان فعالیت صوتی در هر فریم، در هر بخش و در هر فرکانس کلاً شامل یافتن پارامتر Z است، که تنها یک بار نیاز به محاسبه دارد. برای نتیجه واقعی‌تر می‌توان فضای گسسته را با نقاط بیشتری توسط $v_{s,n}$ تخمین بزنیم. در ادامه برای تابع PROD-PHAT روابط مورد نظر را محاسبه می‌کنیم.

همان‌طور که در فصل قبل دیدیم، استفاده از تابع PROD-PHAT بجای میانگین‌گیری، از اشتراک اطلاعات مربوط به TDOA جفت میکروفن‌ها استفاده کرده و این مساله باعث شده که پیک‌هایی با ارتفاع بیشتری داشته باشیم و تاثیر نویز و اطلاعات اضافی را حذف خواهد کرد. در این بخش ما پیشنهاد می‌کنیم از این تابع برای سنجش میزان فعالیت صوتی در هر بخش استفاده گردد.

در این صورت تمام روابط آورده شده برای SRP-PHAT در اینجا هم صادق خواهد بود و تنها بجای

استفاده از متوسط‌گیری از ضرب جفت میکروفن‌ها استفاده می‌کنیم. متریک آورده شده به صورت زیر تعریف می‌شود:

$$d'(a,b)=\sqrt{\frac{1}{N_q}\prod_{q=1}^{N_q}\sin^2(\frac{a-b}{2})} \qquad (۱۹-۴)$$

و بنابراین رابطه نهایی معیار فعالیت صوتی هم برابر خواهد با:

$$(D_s'^{(t)}(k))^2=\frac{1}{2.N_q}\prod_{q=1}^{N_q}\{1-\left|Z_s^*(k,\{v\})\right|\cos[u_q^{(t)}(k)-\angle Z_s(k,\{v\})]\} \qquad (۲۰-۴)$$

در ادامه مانند بخش‌های قبلی پیشنهاد می‌کنیم که از اطلاعات مربوط به جهت منبع صوتی در معیار میزان فعالیت صوتی در هر بخش استفاده کنیم. برای این منظور ما برای تعیین میزان فعالیت صوتی در هر بخش و در فرکانس‌های مختلف، از متوسط گیری وزن‌دار جفت میکروفن‌ها که تمامی جهات را پوشش داده‌اند، استفاده خواهیم کرد. و برای هر بخش تنها جهتی را که ماکزیمم مقدار را به ما بدهد، تقویت می‌کنیم و اطلاعات مربوط به جفت میکروفن‌هایی که در سایر جهات قرار گرفته‌اند را تضعیف می‌نماییم.

بنابراین با بکارگیری این روش ما به طور خودکار امواج انعکاسی را تضعیف کرده و تنها از اطلاعات مربوط به جفت میکروفن‌هایی که در مسیر مستقیم امواج رسیده از منابع صوتی بوده‌اند، استفاده می‌کنیم.

در ادامه معیار فعالیت صوتی که نسبت به جهات مختلف وزن‌دار شده است برای تابع SRP-PHAT به فرم زیر تعریف می‌کنیم:

$$(D_{s,d}^{(t)}(k))^2=\frac{1}{2.N_q}\sum_{q=1}^{N_q}\{1-\left|Z_s^*(k,\{v\})\right|\cos[u_q^{(t)}(k)-\angle Z_s(k,\{v\})]\bullet\omega(\theta_{qd})\} \qquad (۲۱-۴)$$

و به همین صورت برای تابع PROD-PHAT خواهیم داشت:

$$(D'^{(t)}_{s,d}(k))^2 = \frac{1}{2.N_q}\prod_{q=1}^{N_q}\{1-\left|Z^*_s(k,\{v\})\right|\cos[u^{(t)}_q(k)-\angle Z_s(k,\{v\})]\bullet\omega(\theta_{qd})\} \qquad (۲۲-۴)$$

۲-۴-۴ پیاده‌سازی میزان فعالیت صوتی در بخش‌های مختلف اتاق سمینار

در این بخش در ابتدا با استفاده از دیتابیس دایره‌ای عملیات مربوط به تعیین میزان معیار فعالیت صوتی در هر بخش از فضای اتاق سمینار را با توابع تعیین شده در بخش قبلی و در حالت‌های مختلف بررسی خواهیم کرد. و در ادامه بااستفاده از دیتابیس UKA CHIL پیاده سازی‌ها را اجرا خواهیم نمود.

همان‌طور که گفته‌شد در این قسمت قصد داریم فعالیت صوتی را در هر فریم زمانی، در هر بخش با استفاده از روابط بحث شده در بخش‌های قبلی، تخمین بزنیم. علت استفاده از دیتابیس دایره‌ای وجود آرایه میکروفنی دایره‌ای بروی میز، تقریبا در وسط اتاق سمینار می‌باشد که تمامی جهات را بخوبی پوشش داده و برای عملیات بخش کردن و تعیین فعالیت صوتی مناسب می‌باشد.

ابتدا می‌بایست فضای اطراف آرایه میکروفنی را به چندین بخش تقسیم کنیم. برای پوشش دادن کل فضای اتاق سمینار ما از مختصات کروی استفاده می‌کنیم. با فرض اینکه آرایه میکروفنی در مبدا مختصات قرار گرفته باشد، تعداد بخش‌ها باید به نحوی باشد که زمان زیادی برای تعداد بالای بخش‌ها صرف نکرده باشیم و هم از طرفی فضای جستجو را برای عملیات تعیین محل نقطه‌ای منبع به اندازه کافی کوچک کند.

ما در مختصات فضای اطراف آرایه دایره ای را به ۱۸ بخش ۲۰ درجه ای تقسیم می کنیم، تا کل ۳۶۰ درجه را پوشش داده باشیم. یکی از دلایل انتخاب آرایه دایره ای هم این مطلب است که مشخصات آن نسبت به جهت تغییر ناپذیر است.

مرکز آرایه میکروفنی را به عنوان مبدا مختصات کروی در نظر می گیریم. محل دقیق قرار گیری میکروفن‌ها (۸ عدد) را با توجه به اینکه شعاع آرایه میکروفنی ۱۰ سانتی متر است، در مختصات کروی بدست می آوریم. آرایه میکروفنی هم سطح با مبدا مختصات فرض می شوند.

همان طور که در قسمت‌های قبلی داشتیم ، فضای پیوسته را توسط N_v نقطه در هر بخش، به فضای گسسته تخمین زدیم. ما $N_v = 80^3$ در نظر می‌گیریم به طوری که در هر بخش، برای هر یک از پارامترهای (r, φ, θ)، ۸۰ نقطه در نظر گرفته می‌شود.بنابراین در این شرایط هر بخش را توسط ۵۱۲۰۰۰ نقطه،که هریک میتواند محل تخمینی برای منابع صوتی باشد، به فضای گسسته تخمین می زنیم.

در نهایت مختصات تمام نقاط را برای همه ۱۸ بخش در مختصات کارتزین بدست می آوریم. مختصات این نقاط برای بدست آوردن فاز تئوری معرفی شده در بخش‌های قبلی استفاده می‌شود. در ادامه قصد داریم پارامتر Z_s، بدست آوریم.

ما ۸ عدد میکروفن داریم که روی هم ۲۸ جفت میکروفن خواهیم داشت . طول هر فریم زمانی را ۵۱۲ نمونه در نظر می‌گیریم بنابراین به همین تعداد فرکانس دیجیتال هم خواهیم داشت. همان طور که در روابط قبلی دیدیم ، محاسبه پارامتر Z_s وابسته به فاز تئوری می‌باشد. در کل ما برای هر بخش، ۱۸ بخش، و برای هر جفت، ۲۸ جفت، با توجه به رابطه فاز تئوری زیر:

$$u_q^{(th)}(k, L) = \pi \frac{k-1}{N_F} (TOF(L, L_m) - TOF(L, L_{m'})) \qquad (۲۳-۴)$$

که در آن L نشانگر کل نقاط موجود در هر بخش ، ۵۱۲۰۰۰ نقطه، اختلاف زمان پرواز جفت میکروفن‌ها را نسبت به هم بدست می آوریم. فرکانس نمونه برداری Khz ۱۶ می‌باشد. و با ضرب‌کردن حاصل در هر فرکانس، ۵۱۲ فرکانس دیجیتال، فاز مشاهداتی را محاسبه می‌کنیم.

بنابراین در نهایت برای پارامتر Z_s یک ماتریس با ابعاد (۵۱۲،۲۸،۱۸) خواهیم داشت. که به ترتیب با توجه به تعداد فرکانس‌های دیجیتال، تعداد جفت میکروفن‌ها و تعداد بخش‌های اتاق سمینار، ایجاد شده اند.

در ادامه می خواهیم مقدار فعالیت صوتی در هر بخش را برای تمام فریم‌های زمانی بدست آوریم. داده‌های جمع آوری شده توسط میکروفن‌های مختلف را، فریم به فریم دریافت می‌کنیم و در ادامه عملیات را برای هر فریم از کل ۸ کانال انجام می‌دهیم. از همپوشانی ٪۵۰ بین فریم‌ها استفاده می‌نماییم. اجرا عملیات برای یک فریم توضیح می دهیم و سپس عملیات را برای کل فریم‌ها به همین ترتیب اجرا می‌نماییم.

همان طور که گفته شد طول فریم ما ۵۱۲ است بنابراین در هر مرحله داده‌های ما در یک ماتریس به ابعاد (۵۱۲،۸) ذخیره می‌شود. سپس از آنها FFT می‌گیریم. که در این میان از پنجره همینگ استفاده می‌شود. با توجه به اینکه پارامتر Z_s در مرحله قبل محاسبه شده است، در نهایت به یک ماتریس (۵۱۲،۱۸) می‌رسیم.

مشکلی که وجود دارد این است که زمانی که یک منبع صوتی در یک بخش در فرکانس گسسته k وجود دارد، ما افزایش متوسط توان خروجی را در بخش‌های همسایه نیز خواهیم داشت. بنابراین مستقیما نمی توانیم از Ds(k) برای معیار فعالیت صوتی در یک بخش در فرکانس k استفاده کنیم. در[۴۳] نشان داده شده است که در یک فرکانس تنها یک منبع در اندازه غالب فرض شده و سایر منابع می‌توانند نادیده گرفته شوند. بنابراین مقادیر بخش‌های مختلف در هر فرکانس با هم مقایسه شده و تنها به بخشی که بیشترین مقدار فعالیت صوتی را داشته یک امتیاز داده می‌شود. اگر چند بخش در یک فرکانس مقدار یکسانی داشته باشند کلاً آن فرکانس را در نظر نمی‌گیریم.

$$s_{\min}(k) = \arg\min D_s^{(k)}(k) \qquad (۲۴-۴)$$

حال برای بدست آوردن فعالیت صوتی در کل پهنای باند فرکانسی برای یک بخش در فضا و برای یک فریم کافی است تعداد فرکانس‌های گسسته مثبتی را که آن بخش غالب بر سایر بخش‌ها بوده است شمارش شود.

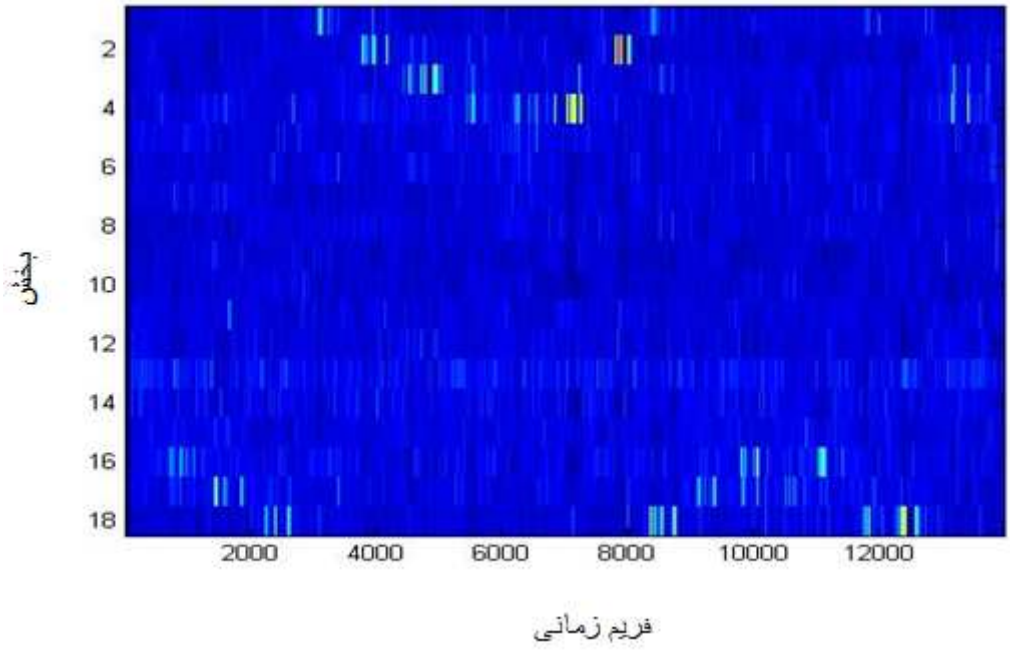
$$\zeta_s^{(t)} = \sum_{k=2}^{N_F} \delta_{kr}(s - s_{\min}(k)) \qquad (۲۵-۴)$$

و اگر یک بخش شامل یک منبع صوتی با پهنای باند بالا مانند گفتار باشد، انتظار داریم که در خیلی از فرکانس‌های گسسته نسبت به سایر بخش‌ها غالب باشد . و مقادیر بزرگتری در معیار فعالیت داشته باشد. به این نکته توجه داشته باشیم که در هر فریم مجموع فعالیت برای تمام بخش‌ها ثابت و برابر N_F می‌باشد.

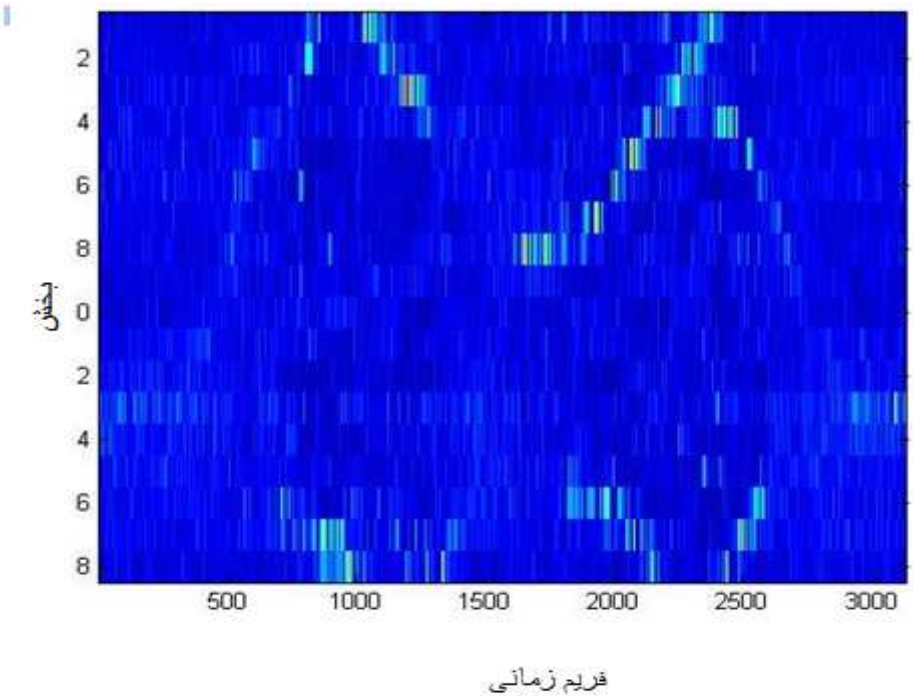
$$\forall t \qquad \sum_{s=1}^{N_s} \zeta_{s,t} = N_F \qquad (۲۶-۴)$$

در ادامه بحث نتایج حاصل از فعالیت صوتی را در دو حالت برای تابع SRP-PHAT مورد بررسی قرار می‌دهیم:

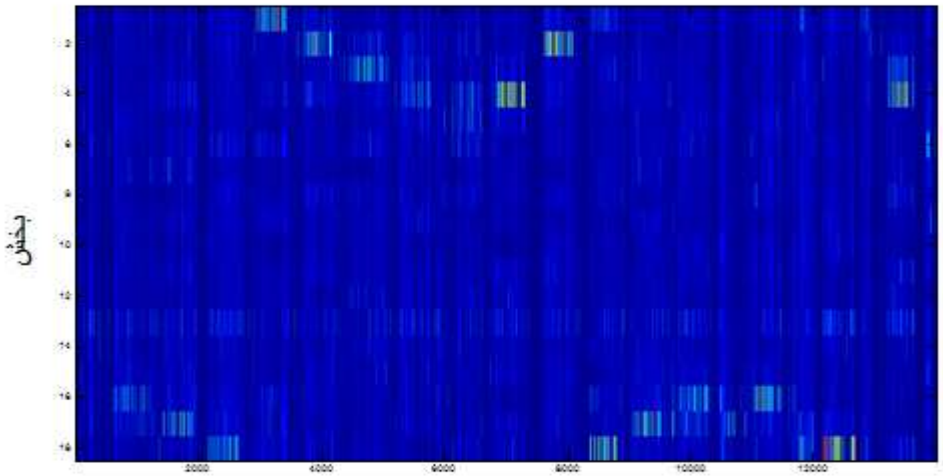
- در ابتدا ساده ترین حالت ممکن را بررسی می‌کنیم. تنها یک صحبت کننده به طور ساکن ایستاده و صحبت می‌کند و پس از پایان صحبت محل خود را تغییر داده و این عمل را تکرار می‌کند. محور عمودی در شکل ۴-۱۷نشان دهنده ۱۸ بخش معرفی شده در قسمتهای قبلی می‌باشد و محور افقی تعداد کل فریم‌ها را نشان می‌دهد. فعالیت صوتی به صورت قله‌های بلندتر،پررنگ‌تر، دیده می‌شوند و در هر فریم زمانی میزان نویز کمتری داشته باشیم، میزان نمره‌هایی که یک بخش می‌گیرد بیشتر است، که این حالت می‌تواند در مواقعی که صحبت کننده به آرایه میکروفنی نزدیک تر است اتفاق بیفتد.



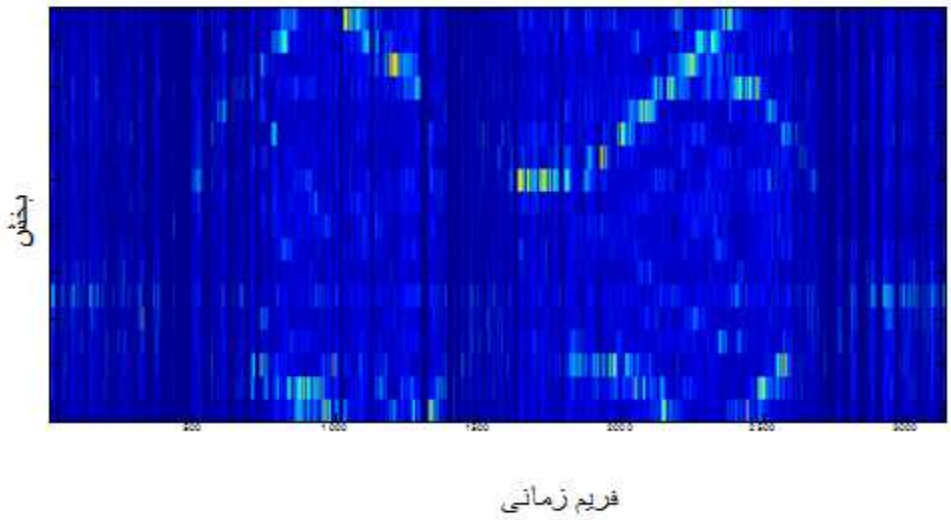
شکل ۴-۱۷: فعالیت صوتی در فریم‌هایی‌زمانی. یک صحبت کننده به طور ساکن در بخش‌های مختلف ایستاده و صحبت می‌کند.از تابع SRP-PHAT استفاده شده است.



شکل ۴-۱۸: دو صحبت کننده، در حال صحبت، در بخش‌های مختلف حرکت می‌کنند. از تابع SRP-PHAT استفاده شده است.



شکل ۴-۱۹: یک صحبت کننده به طور ساکن در بخش‌های مختلف ایستاده و صحبت می‌کند.از تابع MULTI-PHAT استفاده شده است.



شکل ۴-۲۰: دو صحبت کننده، در حال صحبت، در بخش‌های مختلف حرکت می‌کنند. از

- حالت بعدی از پیچیدگی بیشتری برخوردار است، به طور که دو صحبت کننده در حال صحبت، در بخش‌های مختلف حرکت می‌کنند و کاملا همپوشانی صحبتی دارند. همان طور که نتایج درشکل ۴-۱۸ دیده می‌شود افراد از روی هم تقاطع حرکتی دارند. حال همان دو حالت فوق را برای تابع PROD-PHAT انجام خواهیم داد.

در تمام شکل‌های آورده شده محور افقی،فریم‌های زمانی و محور عمودی نشاندهنده بخش‌های مختلف می‌باشد. و بخش‌های با روشنایی بیشتر، نشان‌دهنده احتمال بیشتر برای حظور صحبت‌کننده می‌باشد.

همان‌طور که در شکل ۴-۱۹ و شکل ۴-۲۰ مشاهده می‌گردد، خروجی تابع MULTI-PHAT کمتر تحت تاثیر نویز محیط قرار می‌گیرد.

در ادامه قصد داریم از اطلاعات جهت منبع صوتی برای بهبود عملکرد معیار فعالیت صوتی،همان‌طور که در بخش‌های قبلی در مورد آن صحبت شد، استفاده نماییم.

در استفاده از این دیتابیس چون آرایه میکروفنی در مرکز اتاق سمینار قرار دارد، پیشنهاد می‌کنیم که برای برای بدست آوردن میزان فعالیت صوتی در هر بخش تاکید بیشتری روی جفت میکروفن‌هایی که در مسیر مستقیم انتشار امواج در آن بخش قرار گرفته اند داشته باشیم، با فرض این مطلب که جهت سر صحبت کننده به سمت دیواره نباشد. ما ۲۸ جفت میکروفن و ۱۸ بخش داریم، مرکز هر بخش را به عنوان نقطه نماینده آن بخش فرض می‌کنیم. برای این منظور ما ۱۸ جهت با توجه به مرکز بخش‌ها در نظر می‌گیریم و می‌بایست زاویه بین مرکز جفت میکروفن‌ها و مرکز هر بخش را بدست آوریم. که یک ماتریس(۱۸،۲۸) بدست می‌آید.

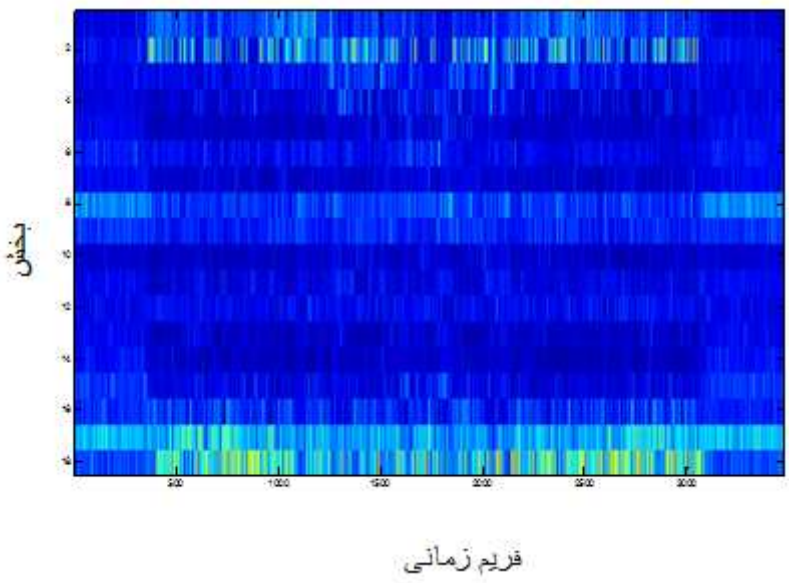
سپس معیار فعالیت صوتی را برای ۱۸ بخش مختلف، با تاکید بروی جفت میکروفن‌هایی که در مسیر مستقیم انتشار هستند بدست می‌آوریم.سایر مباحث مانند بخش‌های قبلی دقیقاً تکرار می‌گردد.

چون در این دیتابیس آرایه در مرکز اتاق سمینار قرار گرفته و فرض بر این است که افراد دور میز سمینار نشسته و صحبت می‌کنند و یا اینکه جهت سر صحبت کننده‌ها در کل به سمت مرکز اتاق است، نه به سمت دیواره‌ها، که البته در اکثر مواقع فرض درستی می‌باشد. ما تنها بروی جفت میکروفن‌هایی که در راستای مستقیم آن بخش قراردارند، تاکید می‌کنیم؛ چراکه اگر بیشترین توان از

جفت میکروفنی که در راستای مستقیم یک بخش قرار ندارد را داشته باشیم، حتماً ناشی از انعکاسات می‌باشد و بهتر است که نادیده گرفته شود.

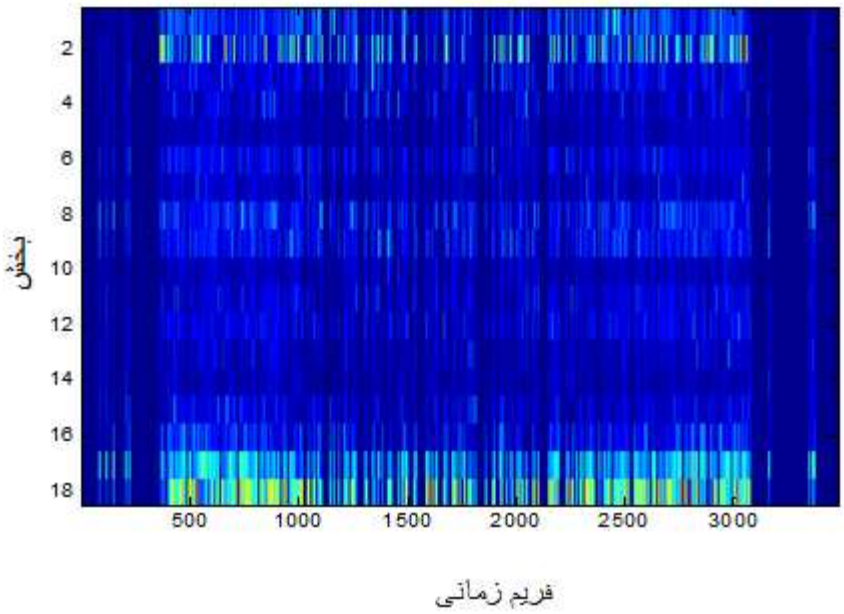
حال اگر آرایه‌های میکروفنی بروی دیواره‌ها قرار داشته باشند، مانند دیتابیس CHIL، برای استفاده از اطلاعات جهت سر صحبت کننده در تصحیح الگوریتم‌های ASL مناسب‌تر بوده و می‌توان برای هر نقطه فرضی برای منبع، توان را در تمامی جهات بدست آورد و سپس تاکید بروی ماکزیمم توان را که مسلماً در جهت مستقیم انتشار است، داشته باشیم.

در شکل فوق حالتی را که دو صحبت‌کننده نشسته و با هم در حال صحبت می‌باشند و یک صحبت کننده در حال حرکت، صحبت می‌کند را در نظر می گیریم. هر سه صحبت کننده با هم همپوشانی صحبتی دارند. ما این حالت را با استفاده از تابع SRP-PHAT و نیز با استفاده از تاکید بروی جفت میکروفن‌های راستای مستقیم هر بخش پیاده سازی می‌کنیم:



شکل ۴-۲۱: فعالیت صوتی برای حالتی که دو صحبت کننده نشسته و یک صحبت کننده در حال حرکت می‌باشد همپوشانی صوتی کامل داریم.از الگوریم SRP-PHAT استفاده شده است.

همان‌طور که دیده می‌شود با تاکید بروی جفت میکروفن‌هایی که در راستای مستقیم هر بخش قرار گرفته‌اند، تا حدودی از میزان نویز و انعکاسات کاسته شده است، و پیک‌هایی با ارتفاع بیشتری را خواهیم داشت.

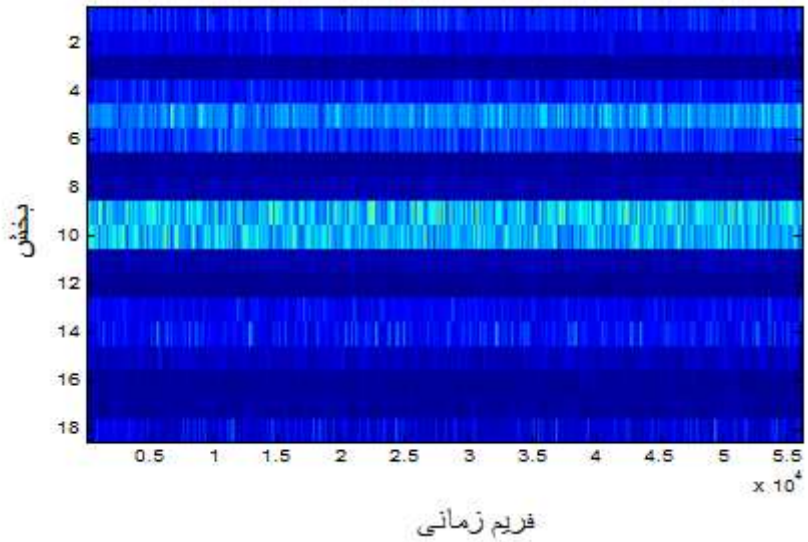


شکل ۴-۲۲: فعالیت صوتی برای حالتی که دو صحبت کننده نشسته و یک صحبت کننده در حال حرکت می‌باشد همچوشانی صوتی کامل داریم.از الگوریم SRP-PHAT با تاکید بروی میکروفن‌هایی که در راستای مستقیم هر بخش هستند، استفاده شده است.

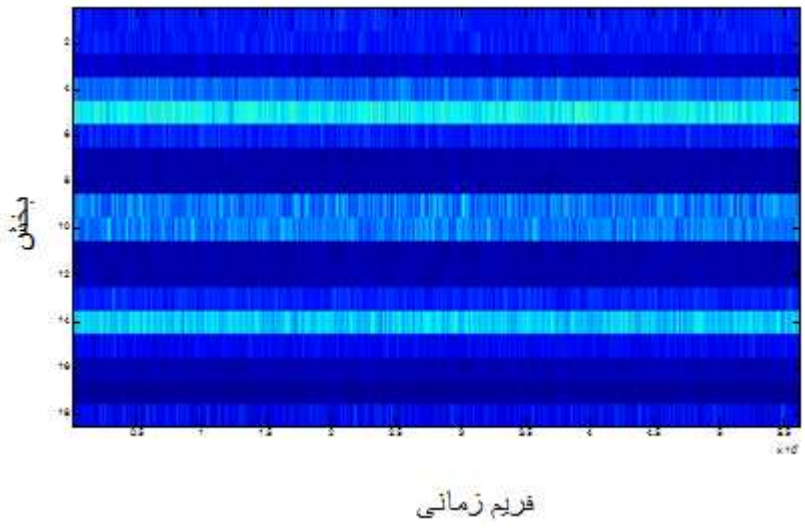
در ادامه قصد داریم با استفاده از دیتابیس UKA CHIL ، عملیات انجام شده را تکرار کنیم. افراد مختلفی به صورت ساکن دور یک میز کنفرانس نشسته و با یکدیگر صحبت می‌کنند. نویز محیط شامل صدای فن کامپوتر، صدای جابجا کردن صندلی و خندیدن افراد و زمزمه کردن دو نفر با همدیگر و صدای گذاردن کلید بروی میز، کار با دستگاه‌های مختلف که داخل اتاق قرار دارند و در طول سمینار این دستگاه چندین بار روشن و خاموش می‌شود تا تغییرات شرایط نویزی را به‌طور کامل داشته باشیم.

در ادامه ما با استفاده از تابع SRP-PHAT و MULTI-PHAT میزان فعالیت صوتی را مطابق شکل‌های زیر بدست آوردیم:

عملیات در این بخش دقیقاً مشابه حالت‌های قبلی تکرار می‌گردد.

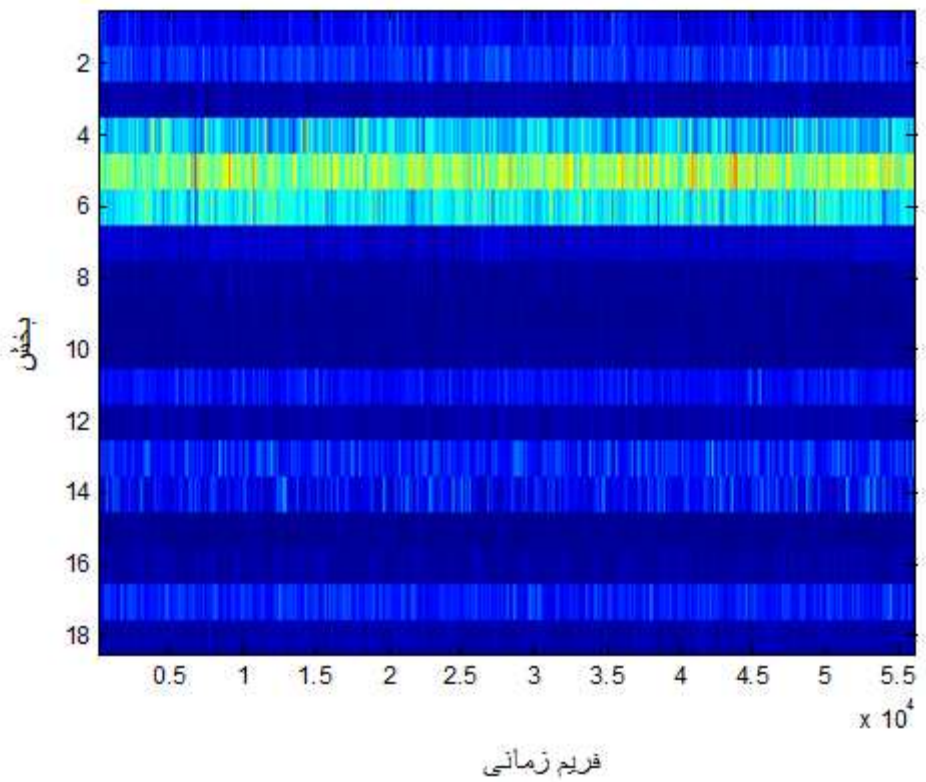


شکل ۴-۲۳: فعالیت صوتی با استفاده از دیتابیس UKA-CHILL و در آن از تابع SRP-PHAT استفاده شده است.



شکل ۴-۲۴: فعالیت صوتی با استفاده از دیتابیس UKA-CHILL و در آن از تابع MULTI-PHAT استفاده شده است.

در هر فریم زمانی تابع مربوط به میزان فعالیت صوتی با توجه به جهت سر صحبت کننده وزن‌دهی شده و تاکید بیشتر روی چفت میکروفن‌هایی خواهد بود که در جهت سر صحبت کننده قرار می‌گیرند. برای افزایش سرعت ما تنها از ۱۶ جهت و سیگما ۲ برای تابع گوسین استفاده کردیم.



شکل ۴-۲۵: فعالیت صوتی با استفاده از دیتابیس UKA-CHILL و در آن از تابع وزنی، با توجه به جهت سر صحبت کننده، استفاده شده است.

همان‌طور که در شکل هم دیده می‌شود، استفاده از اطلاعات مربوط به جهت سر صحبت کننده در این دیتا بیس، با توجه به اینکه آرایه‌های میکروفنی بروی دیواره‌ها نصب شده‌اند به خوبی تاثیر انعکاسات را از بین برده و خروجی مناسبی را خواهیم داشت.

در این بخش برای با استفاده از دیتابیس‌های مختلف، و توابع مختلف، میزان فعالیت صوتی را برای هر بخش از اتاق سمینار بدست‌آوردیم. در ادامه با استفاده از یک آستانه تطبیقی در هر فریم مشخص می‌نماییم که کدام بخش‌ها فعال می‌باشند.

۴-۵- تعیین آستانه تطبیقی

در بخش قبلی میزان فعالیت صوتی موجود در بخشهای مختلف را در کلیه محدوده فرکانسی موجود بررسی کردیم. این عملیات برای کلیه فریم‌های زمانی اجرا می‌گردد. اما قبل از تعیین محل دقیق، باید یک تصمیم باینری برای هر بخش گرفته شود. مثلا با مقایسه میزان فعالیت هر بخش با یک آستانه تصمیم می‌گیریم که آیا فعالیت صوتی واقعی بوده یا ناشی از نویز زمینه می باشد.
در ادامه مقدماتی را در مورد سیستم‌های SNS گفته شده، سیستمی برای انتخاب وفقی آستانه معرفی خواهیم کرد و در نهایت با استفاده از یکی از روش‌های SNS، به حذف محل‌های تخمینی منابع صوتی را که تحت تاثیر نویز زمینه هستند، می‌پردازیم.

انتخاب آستانه تطبیقی در یک روش بدون سرپرست، که به هیچ دیتای آموزشی نیاز ندارد، می‌بایست انجام شود. اگر دانش کافی از نوع توزیع چگالی میزان فعالیت صوتی در فریم‌های مختلف داشته باشیم می‌توان یک مدل احتمالی دو مولفه‌ای برای صحبت و غیر صحبت را بروی مشاهدات در دیتاهای آزمایشی تطبیق داده، و یک آستانه تطبیقی بر حسب این مدل با توجه به شرایط نویزی مختلف زمینه، و برای رسیدن به یک معیار هدف که توسط کاربر قابل تنظیم است، بدست آورد.[۴۴]
مدل دو مولفه‌ای با استفاده از رابطه زیر برای فعالیت و عدم فعالیت بدست می‌آید:

$$f(x) = \omega_0 f_0(x) + \omega_1 f_1(x) \qquad (۲۷-۴)$$

معیار هدف می‌تواند FRR[1],FAR[2] باشد که در شرایط مختلف (نویزی و یک و یا چندین صحبت‌کننده به‌طورهمزمان) ثابت می‌باشد و مانند یک نقطه‌کاری که در ابتدا توسط کاربر تعریف می‌شود، عمل می‌کند و به صورت زیر نشان داده می‌شود:

$$FAR = \frac{N_{FP}}{N_{FP} + N_{TN}} \quad (۲۸-۴)$$

که در آن، صورت کسر تعداد فریم‌هایی است که به اشتباه فعالیت صوتی در نظر گرفته شده اند و مخرج تعداد فریم‌های بی‌صدا را نشان می دهد. یک false alarm زمانی اتفاق می‌افتد که یک نمونه در ground truth غیر فعال، و در نتایج بدست‌آمده فعال باشد.

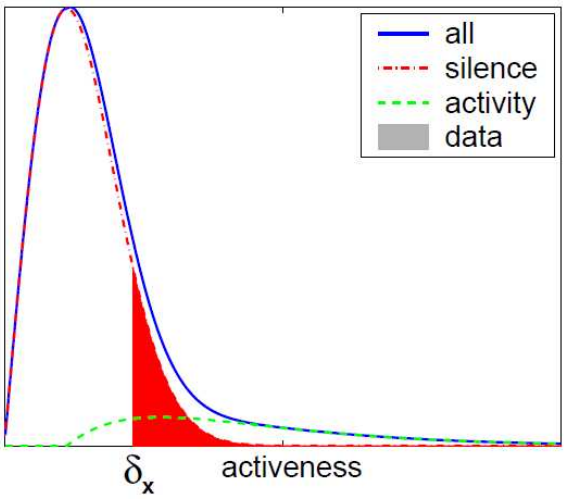
$$FRR = \frac{N_{FN}}{N_{FN} + N_{TP}} \quad (۲۹-۴)$$

که در آن صورت کسر تعداد فریم‌هایی است که به اشتباه بی‌صدا در نظر گرفته شده اند و مخرج تعداد فریم‌های فعال را نشان می دهد. یک false rejection زمانی اتفاق می‌افتد که یک نمونه در ground truth فعال، و در نتایج بدست‌آمده غیر فعال باشد.

تطبیق مدل دو مولفه‌ای از طریق روش expectation maximization (EM)، [۴۵] بروی مشاهدات در هر قسمت از دیتاهای آزمایشی، قابل اجرا می‌باشد

در ادامه مدل‌هایی که برای فعالیت و عدم فعالیت با توجه به نوع توزیع و نحوه تطبیق بروی داده‌های هر بخش تعریف می‌شود، بدست می‌آوریم. بنابراین با استفاده از EM یک مدل بروی دیتاها تطبیق داده شده و سپس مقدار آستانه تطبیقی تعیین می‌گردد.

[1] False rejection rate
[2] False alarm rate



شکل ۴-۲۶: تخمین FAR برا ی بدست آوردن آستانه تطبیقی با استفاده از روشEM . [۴۴]

در شکل ۴-۲۶ کل دیتاها صرف‌نظر از زمان و مکان در هر بخش نشان داده شده است، با استفاده از الگوریتم EM مدل دو مولفه‌ای بروی دیتاها تطبیق داده شده است.

بدست‌آوردن آستانه با توجه به تعریف FAR با توجه به مقدار تعریف شده توسط کاربر تعیین می‌گردد.

برای بخش‌های که فعالیت صوتی ندارند، در یک فرکانس مشخص انتخاب بخش فعال به فرم تصادفی است که معادل احتمال $1/S$ برای هر بخش می‌باشد. بنابراین در یک فریم زمانی یک توزیع یکنواخت برای بخش غیر فعال می‌تواند در نظر گرفته شود؛ اما در عمل وجود نویز زمینه و انعکاسات تاثیر گذار بوده و استفاده از توزیع RICE مناسب تر خواهد بود. البته برخی مقادیر از نمونه‌ها برابر صفر هستند؛ که بدین جهت برای عدم فعالیت از ترکیب توزیع RICE و تابع دلتا استفاده می‌گردد.

و برای بخش‌های فعال، به دلیل پیچیده بودن توزیع صحبت از RICE به‌خاطر انعطاف پذیری بالای آن استفاده می‌شود. البته محدوده کوچکی از فعالیت‌ها را به عنوان نویز باید در نظر داشت، و منطقی خواهد بود که فرض کنیم رنج کوچکی از مقادیر دیتاها، اطلاعاتی برای جداسازی میان فعالیت

و عدم فعالیت به ما نمی‌دهند.که بدین خاطر از RICE شیفت داده شده استفاده می‌شود توزیع RICE در ادامه معرفی می‌شوند:

$$R_{\sigma,V}(x)=\begin{cases}\frac{x}{\sigma^2}.e^{-\frac{x^2+|V|^2}{2\sigma^2}}.I_0(\frac{x|V|}{\sigma^2}) & if \quad x>0\\ 0 & if \quad x\le 0\end{cases} \qquad (۳۰-۴)$$

که در آن I_0 تابع بسل از نوع صفر می‌باشد.

بنابراین تابع احتمال برای دو حالت، برای عدم فعالیت و فعالیت به‌ترتیب، به صورت زیر خواهد بود:

$$f_0(\zeta)=\omega_0^D\delta(\zeta_S)+\omega_0^G R_{\sigma_0,V_0}(\zeta_S)$$
$$f_1(\zeta)=R_{\sigma_1,V_1}(\zeta_S-\sqrt{2.\sigma^2{}_0+V^2{}_0}) \qquad (۳۱-۴)$$

توضیحات کامل و مرحله‌به‌مرحله در[۴۴] مورد بررسی قرار گرفته‌است.

۴-۶- تعیین محل نقطه‌ای منابع صوتی با روش‌های غیر مستقیم

برای تعیین محل نقطه‌ای منابع صوتی معمولاً توسط یکی از روش‌های گرادیانی برای ماکزیمم کردن تابع استفاده می‌گردد که بر پایه یک کلاس از تکنیک‌های بهینه سازی هستند. بسیاری از شبکه‌های عصبی بر پایه تکنیک‌های شیب نزولی[1] استوار می‌باشند. این روشها معمولاً همگرایی ضعیفی دارند و نیازمند مشخص کردن پارامترها توسط کاربر می‌باشند و مقادیر این پارامترها اغلب در موفقیت الگوریتم تعیین کننده می‌باشد . برای نمونه می توان به روش انتشار به عقب[2] اشاره کرد. از نظر بهینه سازی شبکه عصبی در مرحله یادگیری به دنبال مینیمم کردن یک تابع خطااست که خود یک تابع چند متغیره می‌باشد و به تمامی وزن‌های شبکه مربوط است. شبکه عصبی در مرحله

[1] gradient descent
[2] standard back propagation

یادگیری نیازمند تنظیم وزن‌های زیادی است.در آنالیز عددی، تنها یک کلاس از روشهای بهینه‌سازی وجود دارد که قادر به تنظیم این تعداد بالای وزن‌ها می باشد که به روش‌های Conjugate Gradient معروف هستند.

در یک شبکه روبه جلو[1]، N تعداد وزن موجود می‌باشد که واحدهای مختلف از لایه‌های مختلفی را به هم وصل می‌کنند. بردار وزن‌ها را با W نشان می‌دهیم. در روش‌های بهینه‌سازی به دنبال مینیمم کردن تابع خطا E(W) می‌باشیم، که شامل یک فرایند تکراری محلی است بطوریکه تقریبی از تابع را در یک همسایگی از فضای وزن، مینیمم می‌گردد. تقریب مورد نظر معمولا بسط درجه یک از سری تیلر می‌باشد. با انتخاب مقادیر اولیه از وزن‌ها و نیز گام و جهت حرکت در یک فرآیند تکراری وزن‌ها را به طریقی آپدیت می نماییم که خطا در هر مرحله کاهش یافته و در نهایت به نقطه مینیمم برسیم.

برای مثال در الگوریتم شیب نزولی مقدار ثابتی برای گام حرکت در نظر گرفته شده و جهت حرکت را منفی گرادیان تابع خطا در نظر می‌گیریم. که برحسب تابع درجه یک از بسط تیلر می باشد؛ که این مساله و ثابت در نظر گرفتن گام حرکت از نقاط ضعف این الگوریتم می‌باشد.

در روش SCG از تقریب درجه دوم بسط تیلر و تخمین موثری از اطلاعات درجه دوم (هسین) توسط گرادیان، برای تابع خطا استفاده می شود. توضیحات کامل ونحوه عملکرد این الگوریتم در [۴۶] آورده شده است.این روش از سرعت بسیار خوبی برخوردار است به طوری که نهایتاً با تعداد کمی تکرار به جواب می‌رسیم.

در فصل قبل روش‌های دو مرحله‌ای TDOA را مورد بررسی قرار دادیم، که در مرحله اول تاخیر ناشی از جفت میکروفن‌های مختلف را با روش‌هایی مانند GCC-PHAT محاسبه می‌کردیم و در

[1] feed-forward

مرحله دوم محل دقیق منبع صوتی را تخمین می‌زدیم. یکی از نقاط ضعف روش‌های دومرحله‌ای این است که نمی‌توان از آنها برای تعیین محل چندین منبع صوتی به طور همزمان استفاده کرد، ما در این بخش با توجه به مباحثی که در تعیین میزان فعالیت صوتی در هر بخش از فضای اتاق سمینار معرفی کردیم، می‌توانیم روش‌های دومرحله‌ای را تنها در بخش‌های فعال بکار بگیریم که با استفاده از این ایده روش‌های دو مرحله‌ای را برای تعیین چندین منبع به طور همزمان استفاده می‌شود.

بنابراین برای تعیین محل نقطه‌ای در مرحله دوم از تکنیک LI که در فصل قبل معرفی شد استفاده می‌کنیم. همان‌طور که در فصل قبل بحث شد، برای هر بخش فعال جهت منبع صوتی را تخمین خواهیم زد؛ از جفت میکروفن‌های مختلف که درروش LI استفاده می‌گردد، ممکن است جهت‌هایی مختلفی را برای صحبت‌کننده تخمین بزنیم که در نهایت با میانگین‌گیری از جهت‌های بدست آمده محل منابع صوتی در هر بخش را مشخص می‌کنیم.

در بخش بعدی با استفاده از دیتابیس‌های مختلف با جزییات بیشتری عملیات تعیین محل نقطه‌ای را توضیح خواهیم داد.

۴-۶-۱ پیاده سازی تعیین محل نقطه‌ای منابع صوتی

در بخش‌های قبلی ما میزان فعالیت صوتی را در هر بخش از اتاق کنفرانس با استفاده از دیتابیس‌های مختلفی مشخص کردیم. در ادامه با استفاده از یک مقدار آستانه وفقی، بخش‌هایی را که منبع صوتی واقعی در آنها وجود داشت را تعیین کرده و در نهایت با استفاده از یکی از تکنیک‌های closed-form محل دقیق منابع صوتی را در هر لحظه تخمین می‌زنیم.

در این بخش به توضیح عملکرد برنامه‌ها و تعیین برخی از پارامترها در برنامه می‌پردازیم و در نهایت نتایج را با استفاده از دیتاهای مختلف مورد بررسی قرار می دهیم.

برای تعیین فعال بودن هر بخش توسط معیار EM مقدار FAR=۰.۰۰۵ انتخاب شده برای تعداد کل افرادی که در هر لحظه در حال صحبت کردن هستند، محدودیت $N_{MAX}=4$ را انتخاب کرده و دیتاها را در بلوک‌های ۱۰ ثانیه‌ای در نظر گرفته و در پایان مدل‌های مربوط به آستانه تطبیقی آپدیت می‌گردد.

در ابتدا تعیین محل با دیتابیس دایره‌ای را در نظر می‌گیریم، با استفاده از هشت میکروفن دایره‌ای دو عدد آرایه چهارتایی از میکروفن‌ها که در مرکز اتاق سمینار قرار دارند. در هر فریم زمانی عمل تعیین محل نقطه‌ای تنها در داخل بخش‌هایی که فعال در نظر گرفته‌شده‌اند با استفاده از هر دو آرایه انجام می‌گیرد. برای جفت میکروفن‌ها از هر آرایه تابع GCC-PHAT در حوزه زمان را بدست‌آورده و تاخیر زمانی با استفاده از ماکزیمم کردن تابع GCC-PHAT از طریق محدوده مربوط به همان بخش بدست‌می‌آید. سپس با استفاده از روشLI معرفی شده در بخش ۲-۲-۱ جهت صحبت‌کننده فعال در هر بخش با استفاده دو تاخیر زمانی تعیین شده برای هر جفت میکروفن در هر آرایه‌میکروفنی را تعیین می‌کنیم.

با استفاده از چندین آرایه ممکن است چندین جهت برای هر یک از صحبت‌کننده‌های فعال در هر بخش ایجاد گردد؛ ما جهت‌های خارج ناحیه فعال را حذف کرده و برای تعیین جهت نهایی از میانگین چندین جهت بدست‌آمده برای یک صحبت‌کننده استفاده می‌گردد.

در دیتابیس CLEAR UKA هم از چهار آرایه T-shaped که در چهار طرف اتاق سمینار قرار گرفته اند استفاده می‌گردد و جهت هر صحبت‌کننده دقیقاً مانند حالت قبلی برای هر بخش فعال در هر فریم زمانی با روش LI محاسبه می‌گردد. در استفاده از این دیتابیس برای هر یک از آرایه‌ها ابتدا یک جهت با استفاده از مختصات محلی مربوط به آن آرایه تعیین جهت صورت گرفته و در نهایت تمامی جهات تخمین زده به مختصات کلی تبدیل می‌گردد و عملیات نهایی تعیین محل مانند حالت فوق با متوسط‌گیری بدست می‌آید. در ادامه نمودارهای حاصل از تعیین محل را مورد بررسی قرار

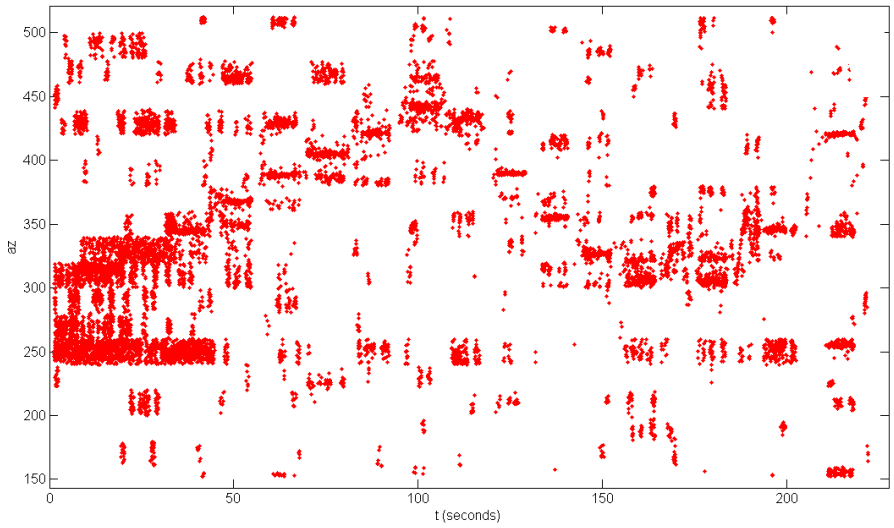
می‌دهیم.

همان‌طور که در ابتدای فصل برای پیاده‌سازی شبکه‌ای شرایط مختلفی را بررسی کردیم، اینجا هم برخی از معروف‌ترین حالات را بررسی خواهیم نمود:

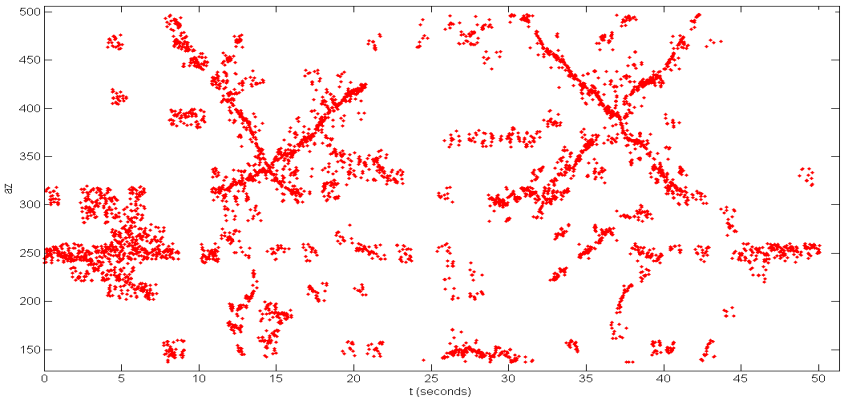
- اولین آزمایش را با استفاده از دیتابیس دایره‌ای برای حالتی که تنها یک فرد در فضای اتاق ایستاده و پس از صحبت‌کردن، عمل خود را در نقاط مختلف اتاق تکرار می‌کند، اجرا می‌کنیم.
- دومین حالت دو صحبت کننده در حال صحبت، در بخش‌های مختلف حرکت می‌کنند و کاملا همپوشانی صحبتی دارند و افراد از روی هم تقاطع حرکتی دارند.
- در ادامه حالتی را در نظر می‌گیریم که دو فرد در کنار یکدیگر نشسته و در حال صحبت می‌باشند و به طور همزمان فرد دیگری در حال حرکت به صحبت کردن مشغول است. کاملاً همپوشانی صوتی بین افراد وجود دارد و یکی از مشکل‌ترین شرایطی است که برای مساله ASL قابل بررسی است.
- در مرحله بعدی با استفاده از دیتابیس CHIL الگوریتم OPROD-PHAT را پیاده‌سازی می‌نماییم.

در نمودارهای آورده شده خروجی مرحله تعیین محل نقطه‌ای را با استفاده از دیتابیس دایره‌ای و CLEAR-CHIL می‌بینیم.

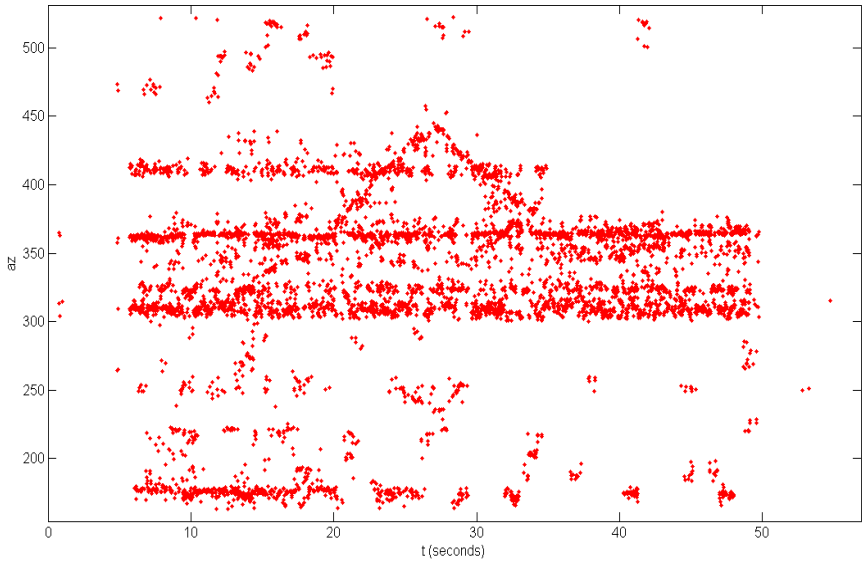
در این بخش پیاده‌سازی حالت‌های مختلفی را انجام دادیم. در بخش بعدی می‌خواهیم با استفاده از متریک‌هایی که برای تعییم محل منبع صوتی کاربرد دارند، به ارزیابی کلیه الگوریتم‌های معرفی شده در این تحقیق بپردازیم.



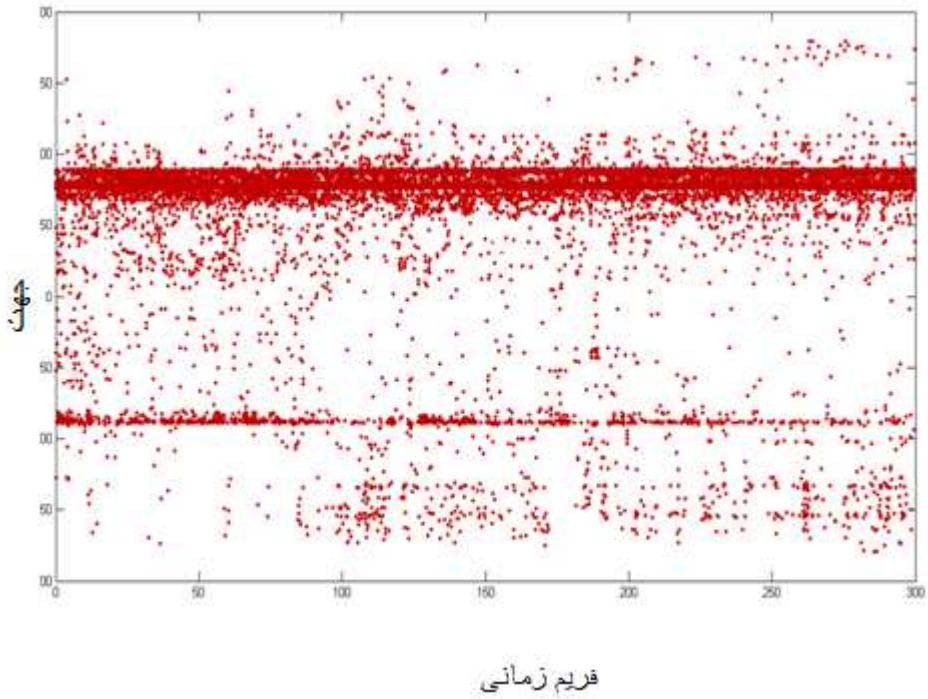
شکل ۴-۲۷: تعیین محل لحظه‌ای یک صحبت کننده به طور ساکن در بخش‌های مختلف ایستاده و صحبت می‌کند.از تابع SRP-PHAT استفاده شده است.



شکل ۴-۲۸:تعیین محل لحظه‌ای. دو صحبت کننده، در حال صحبت، در بخش‌های مختلف حرکت می‌کنند. از تابع PROD-PHAT استفاده شده است.



شکل ۴-۲۹ : تعیین محل لحظه‌ای حالتی که دو صحبت کننده نشسته و یک صحبت کننده در حال حرکت می‌باشد از تابع SRP-PHAT استفاده شده است.



شکل ۴-۳۰: تعیین محل لحظه‌ای الگوریتم OPROD-PHAT با استفاده از دیتابیس CLEAR-CHIL

۴-۷- نتایج ارزیابی الگوریتم‌های پیشنهادی

در ادامه این فصل برای مقایسه بین روش‌های پیشنهادی و روش‌های رایج در ASL متریک‌هایی را معرفی کرده و نتایج کلیه روش‌ها را با استفاده از این متریک‌ها با توجه به فایل ground-truth برا ی هر یک از دیتابیس‌ها، نشان خواهیم داد. خطای مربوط به ground-truth تقریباً برابر ۱۰cm می‌باشد.

۴-۷-۱ معرفی معیارهای ارزیابی تعیین محل صحبت‌کننده

معیار اولیه برای ارزیابی دقت الگوریتم ASL خطای تعیین محل E_r می‌باشد. یک آستانه E_r که به فرم شعاع یک کره می‌تواند در نظر گرفته شود، برای تقسیم‌بندی خطای تعیین محل به دو فرم خطای مناسب[1] و خطای نامناسب[2] به‌کار گرفته می‌شود.

با استفاده از فاصله اقلیدوسی بین محل تخمین‌زده شده برای هر صحبت‌کننده و محل واقعی آن‌ها این تقسیم بندی انجام می‌گیرد. به‌طوری که اگر $d(p_l, p_r) > E_r$ خطای نامناسب خواهیم داشت و در غیر این صورت خطا مناسب خواهد بود. این مقدار آستانه با توجه دقت مورد نظر برای کار ASL می‌تواند در نظر گرفته شود، برای مثال در کار تعیین محل داخل اتاق کنفرانس می‌تواند برابر ۵۰ cm در نظر گرفته شود. بر پایه این خطای تعیین محل چندین متریک برای ارزیابی الگوریتم‌های ASL ایجاد می‌گردد:

- خطای RMS (RMSE) : خطای ریشه میانگین مربعات[3] که برای تمام خروجی‌های الگوریتم تعیین محل محاسبه می‌گردد.

[1] Fine error
[2] Gross error
[3] Root mean square

- خطای RMS مناسب: همان خطای RMSE است که تنها بروی خروجی‌هایی که دارای خطای مناسب می‌باشند، محاسبه می‌گردد.
- نرخ تعیین محل: این نرخ با توجه به درصد خطای مناسب به کل خروجی سیستم، تعریف می‌گردد. اگر N_f تعداد خروجی با خطای مناسب باشد و N_T تعداد کل خروجی‌های سیستم باشد، نرخ تعیین محل به صورت زیر تعریف می‌شود:

$$P_{LOC} = \frac{N_f}{N_T}.100 \qquad (۳۲-۴)$$

سیستم‌های تعیین محل ممکن است با نرخ متفاوتی خروجی تولید کنند به‌طوری‌که دوربین‌هایی که برای تولید ground-truth به‌کار می‌روند معمولاً با نرخ پایین‌تری قادر به‌کار کردن می‌باشند، ازاین‌رو نیازمند یک متوسط‌گری بروی خروجی سیستم‌های ASL می‌باشیم.

- False alarm rate (FAR): درصد خروجی سیستم در زمانی که محلی را برای صحبت‌کننده تخمین می‌زنیم، درحالی‌که هیچ صحبت‌کننده‌ای نداریم، به تعداد فریم‌هایی که در ground-truth به صورت silence می‌باشند.

  برای به کارگیری FAR سیستم ASL ما می‌بایست شامل الگوریتمی برای جداسازی فریم‌های silence از سایرین باشد.
- Deletion rate: زمانی که شخصی در حال صحبت‌کردن می‌باشد و سیستم ASL هیچ تخمینی را فراهم نمی‌کند. بنابراین به نسبت تعداد فریم‌های عدم تعیین محل بر کل فریم‌های موجود شامل صحبت‌کننده در ground-truth می‌گویند.

در کل برای هر فریم زمانی با توجه به نرخ فایل مرجع و میانگین‌گیری از خروجی سیستم تعیین محل، یکی از چهار حالت، خطای مناسب، خطای نامناسب، false alarm، deletion را خواهیم داشت.

در مواقعی که چندین صحبت‌کننده به‌طور همزمان وجود دارد، برای ارزیابی شرایط با استفاده از فاصله اقلیدسی هر محل تخمینی را به نزدیکترین محل صحبت‌کننده در فایل مرجع نسبت می‌دهیم. اگر در یک فریم کسی صحبت نکند و یا در شرایط چند نفره تعداد افراد به درستی تخمین زده نشده باشد، آن فریم نادیده گرفته می‌شود. بنابراین معیارهای FAR, DEL rate تنها در شرایطی که یک صحبت‌کننده در حال صحبت‌کردن باشد، به‌کار گرفته می‌شود.

در ادامه بحث ارزیابی نهایی روش‌های پیشنهادی در این فصل را با استفاده از معیارهای معرفی شده، ارائه خواهیم نمود.

۴-۷-۲ ارزیابی الگوریتم‌های مختلف

در ابتدا ارزیابی را برای حالتی که تنها یک صحبت کننده داریم انجام می‌دهیم. تمامی الگوریتم‌های پیاده‌سازی شده در بخش‌های قبلی، با استفاده از معیارهای معرفی شده ارزیابی می‌گردند. تکنیک های زیر را با توجه به ترتیب آوردن در هر سطر با الگوریتم های پیشنهادی نام‌گذاری شده‌اند.

جدول ۴-۱ :نتایج ارزیابی مربوط به حالتی که یک صحبت کننده به طور ساکن در بخش‌های مختلف ایستاده و صحبت می‌کند.

Technique	RMSE [mm]	Fine RMSE [mm]	LOC. rate [%]	FA. Rate [%]	DEL. Rate [%]
SRP-PHAT شبکه‌ای[DIB00]	۶۲۴	۳۹۳	۷۳	۳۱	۱۹
PROD-PHAT شبکه‌ای (الگوریتم پیشنهادی ۱)	۵۶۸	۳۵۲	۸۱	۲۸	۱۳
SRP-PHAT بخش‌بندی	۷۴۱	۴۱۱	٦٤	۱٦	۸
PROD-PHAT بخش‌بندی (الگوریتم پیشنهادی ۲)	٦۹۲	۲۸٦	٦۹	۱۲	۶

برای شرایطی که چندین صحبت‌کننده داریم، FAR,DEL rate را در نظر نمی‌گیریم. چراکه این معیارها تنها در شرایطی که یک صحبت‌کننده وجود دارد، ایجاد می‌گردند.

در ادامه برای حالت های دو نفره و سه نفره با استفاده از دیتابیس دایره‌ای، ارزیابی‌ها انجام می‌شود و در ادامه این ارزیابی ها را برای دیتابیس CHIL اجرا خواهیم نمود.

جدول۴-۲ :دو صحبت کننده، در حال صحبت، در بخش های مختلف حرکت می کنند.

Technique	RMSE [mm]	Fine RMSE [mm]	LOC. rate [%]
SRP-PHAT شبکه‌ای	**۷۲۱**	**۴۰۱**	**۷۱**
PROD-PHAT شبکه‌ای (الگوریتم پیشنهادی ۱)	**۷۳۸**	**۳۸۳**	**۷۴**
SRP-PHAT بخش‌بندی	**۷۹۲**	**۴۰۹**	**۶۳**
PROD-PHAT بخش‌بندی (الگوریتم پیشنهادی۲)	**۷۸۷**	**۳۲۲**	**۶۹**

جدول ۴-۳ : حالتی که دو صحبت کننده نشسته و یک صحبت کننده در حال حرکت می باشد

Technique	RMSE [mm]	Fine RMSE [mm]	LOC. rate [%]
SRP-PHAT شبکه‌ای	**۹۷۴**	**۴۲۲**	**۵۹**
PROD-PHAT شبکه‌ای (الگوریتم پیشنهادی ۱)	**۹۴۰**	**۴۱۱**	**۵۴**
SRP-PHAT بخش‌بندی	**۸۲۱**	**۳۹۴**	**۷۱**
PROD-PHAT بخش‌بندی (الگوریتم پیشنهادی ۲)	**۷۹۸**	**۳۹۱**	**۶۳**

جدول ۴-۴ :ارزیابی با استفاده از دیتابیس CHIL UKA

Technique	RMSE [mm]	Fine RMSE [mm]	LOC. rate [%]
SRP-PHAT شبکه‌ای	۷۸۳	۳۸۲	۶۵
PROD-PHAT شبکه‌ای (الگوریتم پیشنهادی ۱)	۷۶۶	۳۴۰	۶۷
OPROD-PHAT شبکه‌ای (الگوریتم پیشنهادی ۳)	۶۰۸	**۲۳۲**	۸۱
SRP-PHAT بخش‌بندی	۶۷۹	۳۱۱	۷۲
PROD-PHAT بخش‌بندی (الگوریتم پیشنهادی ۲)	۶۳۸	۳۵۶	۷۵
OPROD-PHAT بخش‌بندی (الگوریتم پیشنهادی ۴)	**۵۸۰**	۲۸۹	**۸۴**

نتایج مربوط به ارزیابی در جداول فوق آورده شده است. در شرایطی که تنها یک صحبت‌کننده داریم استفاده از تکنیک‌های شبکه‌ای بسیار سریع‌تر و با دقت بالاتر، نسبت به روش‌های بخش‌بندی می‌باشد. همان‌طورکه نتایج RMSE نشان می‌دهند، استفاده از تکنیک شبکه‌ای PROD-PHAT بهترین عملکرد را از خود نشان می‌دهد. در مورد معیار LOC. Rate هم خروجی ۸۱% را برای این تکنیک داریم، که بسیار مناسب می‌باشد. در سایر معیارها، خروجی PROD-PHAT بخش‌بندی مناسب‌تر می‌باشد. البته سرعت و حجم محاسباتی روش‌های شبکه‌ای بسیار مناسب‌تر می‌باشند. خروجی معیار FAR در تمامی روش‌ها، چندان مناسب نمی‌باشد، چراکه معمولاً پس از سیستم‌های تعیین محل از SNS برای تقویت خروجی و حذف تخمین‌های اشتباه استفاده می‌شود و نتایج نهایی پس از حذف غیرصحبت‌ها از خروجی سیستم در مقالات ارایه می‌گردد.

در شرایطی که چند صحبت‌کننده داریم، مشکل اصلی روش‌های شبکه‌ای تعیین تعداد افراد در حال صحبت، به طورهمزمان، می‌باشد. بدین جهت برخی از فریم‌های خروجی به‌دلیل عدم تعیین صحیح تعداد افراد فعال، نادیده گرفته می‌شوند، بنابراین در کل در شرایط چند نفره استفاده از روش‌های بخش‌بندی مناسب‌تر می‌باشد.

البته همان‌طور که از نتایج در جدول ۳-۲ و جدول ۳-۳ دیده می‌شود، در شرایط دونفره هنوز خروجی مربوط به متریک خطای مناسب و LOC rate مربوط به روش‌های شبکه‌ای بهتر است، اما در شرایط سه‌نفره همان‌طور که توضیح داده شد، نتایج مناسب‌تر مربوط به روش‌های بخش‌بندی می‌باشد.

نتایج مناسب در حالت دو نفره تا حدودی می‌تواند تحت تاثیر این نکته باشد که برخی از فریم‌ها به دلیل عدم تعیین صحیح تعداد افراد صحبت‌کننده، کنار گذاشته‌شده‌اند و معمولاً چنین فریم‌هایی نتایج خروجی مناسبی را نیز به‌همراه نداشته‌اند.

در مورد دیتابیس CHIL اگرچه میزان نویز بالاتری نسبت به حالات قبلی دارد و همچنین تعداد افراد شرکت‌کننده در سمینار بیشتر است، اما افراد در طول زمان سمینار به‌صورت نشسته و به‌طور ساکن در حال صحبت می‌باشند و در کل میزان خطاها از حالات قبلی کمتر می‌باشد.

نکته قابل توجه خروجی الگوریتم OPROD-PHAT می‌باشد، که RMSE برابر ۵۸۰ mm دارد، در حالی که شرایط کامل نویزی در این حالت برقرار است. همچنین نرخ تعیین محل برابر ۸۴ % است که عملکرد بسیار مناسب این الگوریتم را نشان می‌دهد.

۴-۸- نتیجه‌گیری

در این فصل کلیه الگوریتم‌های پیشنهادی و سایر روش‌های معمول، با استفاده از دو دیتابیس در شرایط نویزی و تعداد شرکت‌کننده‌های مختلف مورد بررسی و پیاده‌سازی قرار گرفت.

قسمت اول پیاده‌سازی‌ها، شبکه‌ای از نقاط در کل اتاق سمینار به عنوان نقاط تخمینی برای صحبت کننده‌ها در نظر گرفته‌شد. در شرایط یک نفره خروجی سیستم بسیار سریع و مناسب می‌باشد.

برای شرایط چندین صحبت‌کننده الگوریتمی پیشنهاد شد، که از اطلاعات تابع همبستگی تعداد افراد صحبت‌کنندهد تخمین زده شده و سپس از همان الگوریتم یک‌نفره چندین بار برای تعیین محل سایر صحبت کننده‌ها استفاده می‌شود.

در حالت‌هایی که سه یا تعداد بیشتری صحبت‌کننده وجود داشت، الگوریتم پیشنهادی در تشخیص تعداد افراد، گاهاً دچار مشکل شده، بنابراین با در قسمت دوم پیاده‌سازی‌ها، با بخش‌بندی کردن فضای اتاق سمینار، یک معیار فعالیت صوتی برای هر بخش در نظر گرفته شد.
با استفاده از یک آستانه تطبیقی با روش EM مشخص می‌شود که در هر زمان کدام بخش‌ها شامل صحبت‌کننده فعال می‌باشد. سپس با استفاده از یکی از روش‌های تعیین محل به‌نام روش LI تنها در بخش‌های فعال و در بخش‌های فعال تعیین محل صحبت‌کننده ها انجام می‌شود.
در ادامه الگوریتمی برای استفاده از اطلاعات جهت سر در تقویت الگوریتم ها ئر برابر انعکاسات پیشنهاد شد، و با استفاده از تابعی به نام OPROD-PHAT تعیین محل منبع صوتی اجرا گردید. نتایج آورده شده در جداول پایانی عملکرد مناسب این تابع پیشنهادی را نشان می‌دهند.

فصل پنجم

نتیجه گیری و پیشنهادات

۵-۱- مقدمه

درابتدا جمع‌بندی کلی در مورد کارهای انجام شده در این تحقیق مورد بررسی قرار می‌گیرد و در ادامه مورد پیشنهادات برای کارهای آینده بحث خواهیم نمود.

۵-۲- نتیجه‌گیری

در این تحقیق، هدف تعیین محل لحظه‌ای چندین صحبت‌کننده به‌طورهمزمان در داخل یک اتاق کنفرانس بوده است. کاربردهای مورد نظر برای این تحقیق شامل: ایندکس‌گذاری کنفرانس‌ها، روباتیک، کاربردهای امنیتی و نظامی، هدایت خودکار دوربین به سمت گوینده در فعالیت‌های رسانه‌ای و... می‌باشد.

تعیین محل صحبت‌کننده‌ها در هر فریم‌زمانی و در شرایط مختلفی نظیر: یک یا چند صحبت کننده به‌طورهمزمان و اینکه صحبت کننده ها به هر دو فرم ساکن و در حال حرکت، باشند و در شرایط نویزی مختلف می‌تواند مورد بررسی قرار گرفته‌است.

در ابتدا با شبکه‌بندی فضای اتاق سمینار محل لحظه‌ای افراد صحبت‌کننده را مشخص نمودیم. از دو دیتابیس مختلف شامل چندین حالت، برای انجام عملیات استفاده شده است.با استفاده از اطلاعات جهت سر صحبت کننده‌ها، از آرایه‌های میکروفنی که در راستای مستقیم با جهت سر صحبت‌کننده قرار دارند استفاده کرده، و تاثیر انعکاسات گرفته‌شده از سایر میکروفن‌ها را تضعیف می‌نماییم. تابع استفاده شده در این الگوریتم را OPROD-PHAT نامیده‌ایم. با انجام این کار عملکرد بسیار مناسب‌تری در برابر انعکاسات خواهیم داشت.

در این بخش یک سیستم برای تعیین محل همزمان چندین صحبت‌کننده پیشنهاد شده است، که در آن، توسط یک الگوریتم دو مرحله‌ای ابتدا با استفاده از یک آستانه تطبیقی بر حسب انرژی هر فریم،

فریم‌های بی‌صدا را جدا کرده و در ادامه توسط مشخصات همبستگی متقابل بین سیگنال یک جفت میکروفن، در مورد تعیین محل لحظه‌ای صحبت‌کننده‌ها تصمیم‌گیری می‌گردد.

در شرایط یک نفره خروجی سیستم بسیار سریع و مناسب می‌باشد. برای شرایطی که چندین صحبت‌کننده داریم، الگوریتمی پیشنهاد شد، که از اطلاعات تابع همبستگی تعداد افراد صحبت‌کننده تخمین زده شده و سپس از همان الگوریتم یک‌نفره چندین بار برای تعیین محل سایر صحبت کننده‌ها استفاده می‌شود.

در حالت‌هایی که سه یا تعداد بیشتری صحبت‌کننده وجود داشت، الگوریتم پیشنهادی در تشخیص تعداد افراد، گاهاً دچار مشکل شده، بنابراین با در قسمت دوم پیاده‌سازی‌ها، با بخش‌بندی کردن فضای اتاق سمینار، یک معیار فعالیت صوتی برای هر بخش در نظر گرفته شد.
در ادامه برای کاهش تاثیر نویز زمینه، و افزایش سرعت جستجوی فضای اتاق سمینار و همچنین تعیین محل مناسب‌تر چندین صحبت‌کننده به‌طورهمزمان، با چندین بخش کردن فضای اتاق، یک میزان فعالیت صوتی را در هر بخش تعیین نموده و در هر فریم زمانی به دنبال این هستیم که کدام بخش شامل صحبت کننده فعال می‌باشد و سپس تعیین محل تنها در بخش های فعال اجرا می‌گردد.

به دلیل وجود نویز متغیر، از یک آستانه وفقی با استفاده از الگوریتم EM برای تعیین فعال بودن هر بخش بهره می‌گیریم. سپس با استفاده از یکی از روش‌های غیرمستقیم TDE، به نام روش LI، در بخش‌های فعال محل صحبت‌کننده‌ها را تعیین نمودیم. توسط این ایده می‌توان روش‌های دو مرحله‌ای را برای تعیین چندین منبع صوتی به طور همزمان استفاده کرد.

۵-۳- پیشنهادات برای کارهای آینده

در این تحقیق محل لحظه‌ای افراد صحبت کننده در هر فریم زمانی با استفاده از الگوریتم‌های مختلف مورد بررسی قرار گرفته‌است. وجود انواع نویزهای محیطی مانند: صدای زنگ تلفن همراه، صدای بسته شدن درب اتاق، صداهای ناشی شده از افراد حاضر مثل خندیدن، سرفه کردن، نفس نفس زدن و... باعث می‌شود که سیستمی برای جداسازی صحبت از غیر صحبت، پس از تعیین محل افراد صحبت‌کننده هم استفاده گردد. بنابراین برای تقویت خروجی الگوریتم‌های تعیین محل پیشنهاد می‌شود، با استفاده از یک سیستم SNS مناسب، محل‌های تخمین‌زده شده که ناشی از منابع غیر صحبتی می‌باشند حذف گردد.

یکی از مباحثی همواره پس از تعیین محل منابع صوتی مطرح می‌گردد، ردیابی منابع صوتی می‌باشد. با استفاده از ردیابی صوتی، از اطلاعات لحظات گذشته برای تخمین محل منبع صوتی در زمان حال استفاده شده و تا حد زیادی از تاثیر نویز و انعکاسات و تخمین‌های اشتباه کاسته می‌شود. بنابراین برای بهبود عملکرد سیستم تعیین محل لحظه‌ای پیشنهاد می‌شود که از یکی از فیلترهای ردیابی استفاده گردد. همان‌طور که در فصل دوم به طور خلاصه در مورد برخی از فیلترهای ردیابی صحبت شد، اخیراً از پارتیکل فیلتر برای عملیات ردیابی در بسیاری از مقالات استفاده شده است.

عمل تعیین محل منبع صوتی به تنهایی با استفاده از اطلاعات صوتی، همواره تحت تاثیر انعکاسات و نویز زمینه بوده و در کل کار بسیار مشکلی می‌باشد. وجود گسستگی بین صحبت‌های افراد و همپوشانی صوتی هم مزید بر علت شده و اخیراً در مقالات از اطلاعات تصویری و صوتی به طور همزمان برای عمل تعیین محل منبع صوتی استفاده می‌گردد.

افراد همواره با استفاده از دوربین‌های مختلف در اتاق کنفرانس مورد ردیابی قرار گرفته و زمانی که صحبت‌کنند از اطلاعات تعیین محل صوتی برای تصحیح ردیابی تصویری استفاده می‌گردد.

بنابراین در نهایت پیشنهاد می‌گردد که به طور صوتی-تصویری عمل تعیین محل منبع صوتی را انجام دهیم.

در فصل چهارم، برای بدست آوردن محل نقطه‌ای افراد صحبت کننده، از یکی از روش‌های closed-form به نام روش LI استفاده شده است که حجم محاسباتی کمی دارد و روش بسیار ساده‌ای در پیاده سازی می‌باشد، همان‌طور که در فصل دوم در مورد این روش‌ها صحبت شد، روش chan یکی دیگر از روش‌های مناسب در این زمینه می‌باشد که در[۲۲] توضیحات کامل آن آمده است. ما پیشنهاد می کنیم برای افزایش دقت تعیین محل نقطه ای از این روش هم استفاده گردد. همچنین در بخش‌های فعال بدست آمده می‌توان الگوریتم SRP-PHAT را اجرا نمود.

خلاصه پیشنهادات در این بخش شامل موارد زیر می‌باشد:

- استفاده از یک سیستم SNS برای بهبود خروجی الگوریتم‌های تعیین محل منبع صوتی
- استفاده از فیلترهای ردیابی صوتی برای بهبود عملکرد سیستم ASL
- استفاده از اطلاعات ردیابی تصویری و صوتی به‌طور همزمان

مراجع

[١]. Shriberg, E., Stolcke, A., and Baron, D. Can prosody aid the automatic processing of multiparty meetings? Evidence from predicting punctuation and disfluencies, and overlapping speech. In Proceedings the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding. 2001.

[٢] Sony Corp. (2006). Sony AIBOTM. http://www.sony.net/Products/aibo/

[٣] Cerwin, S. Ears in the sky. Technology Today. 2004.

[٤] Spriet, A. Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids. PhD thesis, Faculty of Engineering, K.U.Leuven, Leuven, Belgium 2004.

[٥] Lathoud, G., Bourgeois, J., and Freudenberger, J. Sector-based detection for hands-free speech enhancement in cars. EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing. 2006.

[٦] T. Rossing, The Science of Sound. Addison-Wesley, 2nd edn., 1990.

[٧] Moon, T. K. and Stirling,W. C. Mathematical Methods and Algorithms for Signal Processing. 2000.

[٨] Oppenheim, A., Schafer, R., and Buck, J. Discrete-Time Signal Processing. 2, 1999.

[٩] Rabiner, L. and Schafer, R. Digital Processing of Speech Signals. Englewood Cliffs. 1978.

[١٠] S. Haykin, Adaptive filter theory, Prentice Hall, second ed., 1991.

[١١] Buchner, H, Aichner, R, Stenglein, J, Teutsch, H and Kellermann, W. Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, PA, USA. 2005.

[١٢] H. V. Trees, Detection, Estimation, and Modulation Theory. Part IV, Optimum Array Processing, John Wiley & Sons, 2002.

[١٣] J. DiBiase, H. Silverman, and M. Brandstein, Microphone Arrays. Springer-Verlag, 2001.

[١٤] Herbert Buchner, Robert Aichner, and Walter Kellermann. Trinicon-based blind system identification with application to multiple-source localization and separation. In S. Makino, T.W. Lee, and S. Sawada, editors, Blind Speech Separation. Springer-Verlag, Berlin, 2007.

[١٥] P. Stoica and J. Li, Source Localization from Range-Difference Measurements. IEEE Signal Processing Mag., 23:pp. 63 – 66, November 2006.

[١٦] M. Gillette and H. Silverman, A linear closed-form algorithm for source localization from time-differences of arrival. IEEE Signal Processing Letters, 15(1):pp. 1–4, 2008.

[١٧] B. Friedlander, A passive localization algorithm and its accuracy analysis. IEEE Journal of Oceanic Engineering, 12(1):pp. 234 – 245, 1987.

[۱۸] H. So and S. Hui, Constrained localization algorithm using tdoa measurements. IEICE Trans. Fundamentals, E86-A (12):pp. 3291 – 3293, December 2003.

[۱۹] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, Real-time passive source localization: A practical linear-correction least-squares approach. IEEE Transactions on Speech and Audio Processing, 9(8):pp. 943 – 956, 2001

[۲۰] Yiteng Huang, Jacob Benesty, Gary Elko, and Russell Mersereau. Real-time passive source localization: a practical linear-correction least-squares approach. IEEE Transactions on Speech and Audio Processing, 9(8):943-955, November 2001.

[۲۱] H. Schau and A. Robinson. Passive source localization employing intersecting spherical surfaces from time-of-arrival dierences. IEEE Transaction on Acoustics, Speech and Signal Processing, 35(12):1661-1669, December 1987.

[۲۲] Y. T. Chan and K. C. Ho. A simple and ecient estimator for hyperbolic location. IEEE transaction on signal processing, 42(8):1905-1915, August 1994.

[۲۳] Brandstein, M. A Framework for Speech Source Localization Using Sensor Arrays. PhD thesis, Brown University. 1995.

[۲۴] Norbert Strobel and Rudolf Rabenstein. Classication of time delay estimates for robust speaker localization. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 6, pages 3081-3084, Phoenix, AZ, USA, March 15-19 1999.

[۲۵] Paris Smaragdis and Petro Boufounos. Learning source trajectories using wrapped-phase hidden markov models. In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pages 114-117, New Paltz, NY, USA, October 16-19 2005.

[۲۶] S. Kay, Fundamentals of Statistical Signal Processing, Volume I: Esti- mation Theory. Prentice Hall Signal Processing Series, Prentice Hall, 1998

[۲۷] S. Gannot and G. Dvorkind, Microphone array speaker localizers using spatial-temporal information. EURASIP Journal on Applied Signal Processing, 2006:pp. 1–17, 2006, article ID 59625.

[۲۸] E. Lehmann and A. Johansson, Particle filter with integrated voice activity detection for acoustic source tracking. EURASIP Journal on Applied Signal Processing, 2007(1):pp. 1–11, 2007.

[۲۹] Thomas Funkhouser, Nicolas Tsingos, Ingrid Carlbom, Gary Elko, Mohan Sondhi, James E. West, Gopal Pingali, Patrick Min, and Addy Ngan. A beam tracing method for interactive architectural acoustics. Journal of Acoustical Society of America, 115(2):739-756, February 2003.

[۳۰] G.C, Knapp C. H and Carter. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing . August 1976, pp. 320-327.

[۳۱] Kevin D. Donohue, Jens Hannemann, Henry G. Dietz, Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments, lSignal Processing 87 (2007) 1677–1691

[۳۲] J. Chen, J. Benesty, and Y. Huang, Performance of GCC- and AMDF based time-delay estimation in practical reverberant environments. EURASIP J. Applied Signal Process. 2005(1):pp. 25 – 36, 2005.

[۳۳] J. Benesty, Adaptive Eigenvalue Decomposition Algorithm for Passive Acoustic Source Localization. J. Acoust. Soc. Am, 107(1):pp. 384 – 391, 2000.

[۳۴] DiBiase, J. A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments. PhD thesis. 2000.

[۳۵] Pasi Pertilä, Teemu Korhonen, and Ari Visa, Measurement Combination for Acoustic Source Localization in a Room Environment. EURASIP Journal on Audio, Speech, and Music Processing, vol. 2008, Article ID 278185, 14 pages, 2008.

[۳۶] Moore, D. (2002). The IDIAP Smart Meeting Room. IDIAP-COM 02-07, IDIAP

[۳۷] Chu, W. T., & Warnock, A. C. C. Detailed directivity of sound fields around humantalkers. Tech. rep. IRC-RR-104, National Research Council Canada. 2002.

[۳۸] Alberto Abad, Carlos Segura, Climent Nadeu and Javier Hernando, Audio-based approaches to head orientation estimation in a smart-room, *INTERSPEECH 2007,PP 390-393,* August 27-31, Antwerp, Belgium

[۳۹] E. Scheirer and M. Slaney, Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, ICASSP, 1997, vol. 1, pp. 1331-1334.

[۴۰] J. Ajmera, I. McCowan, and H. Bourlard, Speech/Music Segmentation using Entropy and Dynamism Features in a HMM Classification Framework, Speech Communication, vol. 40, pp. 351-363, May 2003.

[۴۱] Lathoud, G. and Magimai.-Doss, M. A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2005.

[۴۲] Zotkin, D. and Duraiswami, R. Accelerated speech source localization via a hierarchical search of steered response power. IEEE Transactions on Speech and Audio Process., 12(5). 2004.

[۴۳] Roweis, S Factorial Models and Refiltering for Speech Separation and Denoising. In Proceedings of Eurospeech. .2003.

[۴۴] G. Lathoud and M. Magimai.-Doss, Threshold Selection for Unsupervised Detection, with an Application to Microphone Arrays, IDIAP-RR-05-52, 2005.

[۴۵] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39:1–38. 1977.

[۴۶] Moller, M. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks, 6:525–533. 1993.

Abstract:

Recent advances in computer technology and speech and language processing have made possible that some new ways of person-machine communication and computer assistance to human activities start to appear feasible. Concretely, the interest on the development of new challenging applications in indoor environments equipped with multiple multimodal sensors, also known as smart-rooms, has considerably grown.

This thesis investigates the analysis of spontaneous multi-party speech; the goal is to estimate where the various speakers are talking.

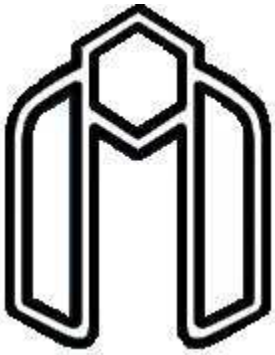
The speed, the versatility and the robustness of the proposed techniques are tested on a variety of real indoor recordings, including multiple moving speakers as well as seated speakers in meetings. Optimized implementations are provided in most cases.

At the first, by using the Combined of hyperbolae produced by time delay estimation (TDE) between several microphones pair and the head orientation information, a new acoustic multi-speaker localization function has been proposed that we call it OPROD-PHAT function. We implement a grid-based, multiple speaker localization method. On the multiple moving speaker location estimation, the new approach has been proposed, that to find number of active source in each time frame, the power of cross correlation function has been used. After find the loudest source present by maximizing the energy of a steered beamformer, in order to localize other source, the process is repeated by removing the contribution of the first source.

We used to discretize the physical space into a few sectors, to reduce the impact of background noise. And speed up, and for each time frame, an automatic threshold selection system by using the EM algorithm has been implemented to determine which sectors contain active acoustic sources

Then, using the LI method, the location of speakers in each active sector, has been determined.Finally, the proposed algorithms has been evaluated. The result of simulation show superior performance of proposed system.

Keywords- multiperson localization, time delay of arrival (TDOA) head oriantation, microphone array



Shahrood University of Technology

Facultyof.Electrical and Robotic Engineering.

Multiple Speaker localization in a smart room

Mohamad Hesam mahmodi nejad

Supervisor(s):

Dr. Hossein Marvi

September 2010