مقاله پژوهشی

تفکیک کور منابع گفتار دوکاناله بر اساس مکان‌یابی

حسن علی‌صوفی، مرتضی خادمی و عباس ابراهیمی‌مقدم

چکیده: در این مقاله یک روش جدید برای تفکیک کور منابع گفتار دوکاناله، بدون نیاز به دانش قبلی در مورد منابع گفتار آمده است. در روش پیشنهادی، با وزن‌دادن به طیف سیگنال ترکیب‌شده بر اساس فاصله منابع گفتار با میکروفون، تفکیک منابع گفتار انجام می‌شود. بنابراین ابتدا با تشکیل اسپکتوگرام زاویه‌ای توسط تابع همبستگی متقابل تعمیم‌یافته، منابع گفتار موجود در سیگنال ترکیب‌شده مکان‌یابی می‌شوند. سپس با توجه به موقعیت مکانی منابع از نظر فاصله با میکروفون‌ها، اندازه طیف سیگنال ترکیب‌شده، وزن‌دهی می‌شود. با ضرب اندازه طیف وزن داده شده در مقادیر حاصل از اسپکتوگرام زاویه‌ای و مقایسه آنها با هم، برای هر منبع یک نقاب باینری ساخته می‌شود. با اعمال نقاب باینری به اندازه طیف سیگنال ترکیب‌شده، منابع گفتار موجود در آن از هم جدا می‌شوند. این روش روی داده‌های پایگاه داده SiSEC آزمایش و از ابزار سنجش و معیارهای موجود در این پایگاه، برای ارزیابی استفاده شده است. نتایج نشان می‌دهد که روش پیشنهادی، از جهت معیارهای موجود در پایگاه مذکور با روش‌های رقیب قابل مقایسه بوده و پیچیدگی محاسباتی کمتری دارد.

***کلیدواژه:* اسپکتوگرام زاویه‌ای، تابع همبستگی متقابل تعمیم‌یافته، تفکیک کور منابع گفتار.**

۱- مقدمه

تفکیک صدا در مهمانی[1] یک مسئله معروف در حوزه تفکیک منابع گفتار است [۱]. صداهای ترکیب‌شده با هم، باید به نحوی از هم جدا شوند که تا حد امکان از اعوجاج و مصنوعی‌شدن صدا جلوگیری شود و تداخل صداهای مزاحم از بین برود.

وجود صداهای مزاحم در ارتباط تلفنی، امری آزاردهنده است که با حذف آنها می‌توان ارتباط بهتری ایجاد کرد. با پیشرفت الگوریتم‌های تفکیک صدا و جداسازی بهتر منابع گفتار موجود در سیگنال ترکیب‌شده از هم، عملکرد ارتباطات تلفنی نیز قابل بهبود است. همچنین با تجهیزکردن سمعک یا هدفون به جداکننده صدا، می‌توان صدای اصلی را از تداخل‌های مزاحم و یا نویز محیط جدا کرد و درک شنیداری افراد کم‌شنوا را افزایش داد. ابزارهای تشخیص گوینده و تشخیص گفتار نیز می‌توانند صدای اصلی را از بین صداهای دیگر جدا کنند و این امر باعث افزایش عملکرد آنها می‌شود.

این مقاله در تاریخ ۲۲ تیر ماه ۱۳۹۸ دریافت و در تاریخ ۲۸ بهمن ماه ۱۳۹۹ بازنگری شد.

حسن علی‌صوفی، گروه برق، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: hassan_alisoofi@um.ac.ir).

مرتضی خادمی (نویسنده مسئول)، گروه برق، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: khademi@um.ac.ir).

عباس ابراهیمی‌مقدم، گروه برق، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: a.ebrahimi@um.ac.ir).

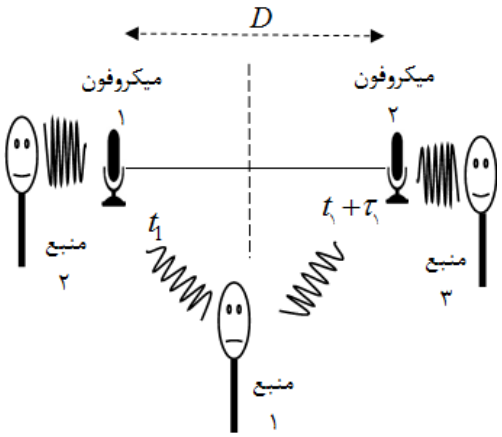
1. Cocktail Party Problem

روش‌های متعددی در حوزه تفکیک منابع گفتار وجود دارد. در برخی از روش‌های تفکیک منابع گفتار، بایستی تعداد منابع موجود در سیگنال ترکیب‌شده کوچک‌تر یا مساوی تعداد میکروفون باشد [۲]. روش‌های دیگری مانند [۳] محدودیت قبل را ندارند ولی برای گفتار ترکیب‌شده بدون انعکاس مناسب هستند. در این تحقیق، دو چالش مهم در تفکیک منابع گفتار بررسی می‌شود. یک چالش، مربوط به زمانی است که تعداد منابع موجود در سیگنال ترکیب‌شده از تعداد میکروفون‌ها بیشتر باشد و چالش دیگر وقتی است که محیط ضبط صدا، مشابه با محیط‌های واقعی، انعکاس صدا داشته باشد [۴]. اگر سیگنال ترکیب‌شده دریافتی دوکاناله (دومیکروفونه) باشد، از اختلاف شدت و اختلاف زمانی بین دو کانال، می‌توان برای غلبه بر چالش‌های مذکور استفاده کرد [۵] و [۶]. در بسیاری از تحقیقات اخیر از ترکیب مکان‌یابی [۷] با یکی از ابزارهای شبکه عصبی [۸] یا تجزیه نامنفی ماتریس [۹]، تفکیک منابع گفتار انجام شده است.

مراجع [۸] و [۱۰] از ترکیب مکان‌یابی منابع و شبکه عصبی عمیق برای تفکیک منابع گفتار استفاده کرده‌اند. در گفتار دوکاناله، اختلاف فاز و اختلاف شدت بین دو کانال، ورودی‌های شبکه عصبی عمیق هستند. هرچه شبکه عصبی با داده‌های بیشتری آموزش ببیند، تفکیک منابع گفتار به صورت بهتری انجام می‌شود. وابستگی به آموزش، مهم‌ترین ضعف این روش است. همچنین اگر اختلاف بین داده‌های آموزش و آزمایش زیاد باشد، عملکرد شبکه عصبی پایین می‌آید. این روش برای محیط‌های واقعی که در آن انعکاس صدا وجود دارد مناسب است.

مراجع [۲]، [۹] و [۱۱] از ترکیب مکان‌یابی منابع و تجزیه نامنفی ماتریس برای تفکیک منابع گفتار استفاده کرده‌اند. با استفاده از اختلاف زمانی بین دو کانال، می‌توان منابع موجود در سیگنال ترکیب‌شده را از طریق یافتن تأخیر زمانی بین دو کانال برای هر منبع، مکان‌یابی کرد. در برخی از روش‌های مکان‌یابی، فقط از اختلاف زمانی بین دو میکروفون برای مکان‌یابی منابع استفاده می‌شود و چون اساس عملکرد روش‌های مذکور محاسبه اختلاف فاز است، نیازی به محاسبه اختلاف شدت بین دو میکروفون نیست [۲]. استفاده از اطلاعات مکانی و تجزیه نامنفی ماتریس سیگنال ترکیب‌شده، تفکیک منابع گفتار را به خوبی انجام می‌دهد. تجزیه نامنفی ماتریس، ابزاری مناسب برای تفکیک منابع گفتار هم‌زمان است و تداخل را به مقدار قابل قبولی حذف می‌کند ولیکن پیچیدگی محاسباتی این رویکرد زیاد است [۹].

در این تحقیق یک روش جدید برای تفکیک منابع گفتار موجود در یک سیگنال ترکیب‌شده دوکاناله ارائه شده که تفکیک منابع گفتار موجود در سیگنال ترکیب‌شده را بر اساس فاصله آنها از میکروفون‌های ۱ و ۲ انجام می‌دهد. در روش پیشنهادی برای مکان‌یابی منابع، از تابع همبستگی



شکل ۱: منبع ۱ بین دو میکروفون است و صدای آن در لحظه t_1 به میکروفون ۱ و در لحظه $t_1+\tau_1$ به میکروفون ۲ می‌رسد. منابع ۲ و ۳ به ترتیب دارای تأخیر D/V و $-D/V$ هستند.

متقابل تعمیم‌یافته (GCC)[1] استفاده شده است [۱۲]. همچنین از فیلتر میانگین به عنوان یک پیش‌پردازش برای هموارسازی طیف سیگنال ترکیب‌شده جهت بهبود مکان‌یابی منابع استفاده می‌شود [۱۳] و [۱۴]. این روش دارای پیچیدگی محاسباتی کمی بوده و به اطلاعات قبلی از منابع گفتار نیازمند نیست.

در بخش بعدی، مبانی مکان‌یابی با تابع همبستگی متقابل تعمیم‌یافته که روش پیشنهادی بر مبنای آن است، بررسی می‌شود. روش پیشنهادی در بخش ۳ شرح داده می‌شود. ارائه نتایج شبیه‌سازی و مقایسه با روش‌های دیگر در بخش ۴ انجام می‌گردد و نتیجه‌گیری کلی در بخش ۵ بیان می‌شود.

۲- مکان‌یابی منابع گفتار با روش GCC-PHAT

با توجه به این که در روش پیشنهادی این مقاله، مکان‌یابی منابع بر مبنای روش مشهور و پرکاربرد GCC-PHAT[2] می‌باشد، لازم است ابتدا روش مذکور مختصراً معرفی گردد. اگر سیگنال ترکیب‌شده ورودی دوکاناله باشد، از اختلاف زمانی بین کانال‌ها برای مکان‌یابی منابع گفتار موجود در سیگنال ترکیب‌شده استفاده می‌شود. تأخیر در حوزه زمان معادل جابه‌جایی فاز در حوزه فرکانس است و GCC-PHAT از این موضوع و با تشکیل اسپکتوگرام زاویه‌ای، برای یافتن تأخیر زمانی بین دو میکروفون استفاده می‌کند [۱۲]. به عنوان مثال شکل ۱ وضعیت قرارگرفتن سه منبع گفتار و دو میکروفون را نشان می‌دهد. همان طور که در این شکل دیده می‌شود، منبع شماره ۱ بین دو میکروفون قرار گرفته و صدای آن در لحظه t_1 به میکروفون شماره ۱ و با تأخیر τ_1 به میکروفون ۲ می‌رسد. اگر فاصله بین دو میکروفون D و سرعت صدا V باشد آن گاه با توجه به شکل ۱، بیشترین تأخیر زمانی (از نظر جبری) بین دو میکروفون D/V و کمترین تأخیر $-D/V$ است.

شکل ۲ بلوک دیاگرام روش GCC-PHAT [۱۲] را برای مکان‌یابی (یافتن تأخیر زمانی بین دو میکروفون) نشان می‌دهد. در این روش تبدیل فوریه هر دو سیگنال ترکیب‌شده ورودی محاسبه می‌شود. از آنجا که سیگنال گفتار، غیر ایستان است، از تبدیل فوریه زمان‌کوتاه (STFT) برای محاسبه طیف آن استفاده شده است. مطابق شکل ۲، $x_1(n)$ و $x_2(n)$ دو سیگنال ترکیب‌شده ورودی در حوزه زمان و $X_1(m,f)$ و $X_2(m,f)$ به ترتیب تبدیل فوریه زمان‌کوتاه (متغیرهای m و f به ترتیب بیانگر زمان و فرکانس هستند) کانال‌های ۱ و ۲ می‌باشند. در این روش یکی از میکروفون‌ها (مثلاً میکروفون ۱) به عنوان مبنا در نظر گرفته می‌شود. سپس مزدوج مختلط $X_2(m,f)$ ($X_2^*(m,f)$) را محاسبه کرده و به ازای مقادیر مختلف تأخیر (δ) بین D/V و $-D/V$، اسپکتوگرام زاویه‌ای سه‌بعدی ($G(m,f,\delta)$) طبق (۱) محاسبه می‌شود [۱۲]

$$G(m,f,\delta)=\mathrm{Re}\{\frac{X_1(m,f)X_2^*(m,f)}{|X_1(m,f)||X_2(m,f)|}e^{j2\pi f\delta}\} \qquad (۱)$$

فاز عبارت $(X_1(m,f)X_2^*(m,f)/|X_1(m,f)||X_2(m,f)|)e^{j2\pi f\delta}$ (برای δ ایده‌آل) صفر و قسمت حقیقی آن یک می‌شود ولی در عمل به علت وجود تداخل و انعکاس این مقدار کمتر از یک است. بنابراین برای محاسبه تأخیر بین دو کانال، فقط قسمت حقیقی عبارت فوق در نظر گرفته می‌شود. برای تعیین اختلاف زمانی منابع، بایستی حاصل‌جمع $G(m,f,\delta)$ را روی ابعاد فرکانس و زمان محاسبه کرد. شکل ۳ نمایش دوبعدی $\sum_f G(m,f,\delta)$ (یعنی حاصل‌جمع روی بعد فرکانس) و شکل ۴ نمایش یک‌بعدی $\sum_m\sum_f G(m,f,\delta)$ (یعنی حاصل‌جمع روی ابعاد فرکانس و زمان) را نشان می‌دهند. سه خط تیره‌رنگ در شکل ۳ نشان‌دهنده سه منبع گفتار در این شکل است. رابطه (۲) نحوه یافتن تأخیر زمانی برای هر منبع را نشان می‌دهد [۱۲]

$$\tau_i=\arg\max_{\delta}\Big(\sum_m\sum_f G(m,f,\delta)\Big)\ ,\ i=1,2,\ldots,I \qquad (۲)$$

در این رابطه I تعداد منابع و τ_i تأخیر زمانی منبع iام است. این رابطه به تعداد منابع، نقاط ماکسیمم دارد. سه نقطه ماکسیمم در شکل ۴، بیانگر وجود سه منبع گفتار و محل حداکثرشدن آنها، نشان‌دهنده تأخیر زمانی بین دو میکروفون برای هر منبع است. وقتی فاصله بین دو میکروفون زیاد است، GCC-PHAT یک ابزار مناسب برای مکان‌یابی منابع است. با افزایش تعداد منابع، اگر دو میکروفون به هم نزدیک باشند و از روش GCC-PHAT برای مکان‌یابی منابع استفاده شود، آن گاه تعداد نقاط حداکثر از تعداد منابع بیشتر و یافتن تأخیر منابع دچار ابهام می‌شود [۱۵]. برای حذف نقاط حداکثر خطا از توابع سیگموئید استفاده می‌شود که مهم‌ترین آنها tanh است [۱۶]. بنابراین وقتی دو میکروفون نزدیک هستند، برای جلوگیری از بروز خطا در یافتن تأخیر زمانی از مدل غیر خطی ($\tilde{G}(m,f,\delta)$) طبق (۳) استفاده می‌شود [۱۵] و [۱۷]

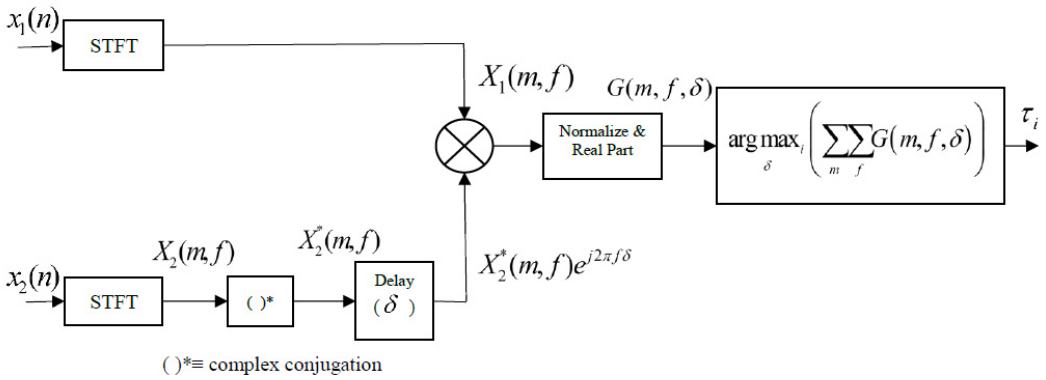
$$\tilde{G}(m,f,\delta)=1-\tanh(\gamma\sqrt{1-G(m,f,\delta)}) \qquad (۳)$$

در این رابطه γ یک ضریب وزنی و عددی مثبت است که با تغییر آن می‌توان دقت مکان‌یابی را تغییر داد. در این تحقیق برای داده‌های ناشی از دو میکروفون نزدیک به هم، از این مدل غیر خطی استفاده شده است.

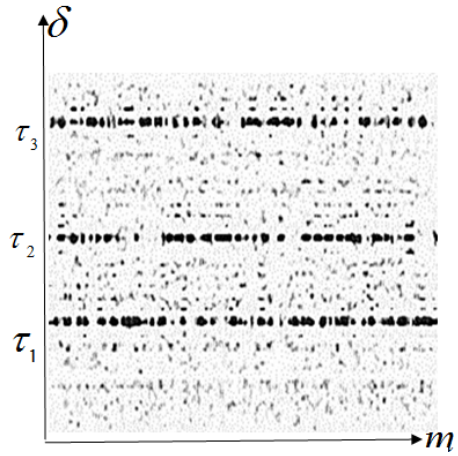
۳- روش پیشنهادی

شکل ۵ بلوک دیاگرام روش پیشنهادی را نشان می‌دهد. در روش پیشنهادی از فیلتر میانگین برای کاهش اثر تداخل و بهبود مکان‌یابی برای تفکیک منابع گفتار موجود در سیگنال ترکیب‌شده دوکاناله استفاده می‌شود. این فیلتر در بخش ۳-۱ معرفی می‌شود. مکان‌یابی منابع گفتار نیز توسط روش GCC-PHAT انجام می‌شود. مؤثرترین بخش در این روش، دادن وزن مناسب به طیف سیگنال ترکیب‌شده بر اساس مکان منابع نسبت به دو میکروفون است. این موضوع در بخش ۳-۲ توضیح

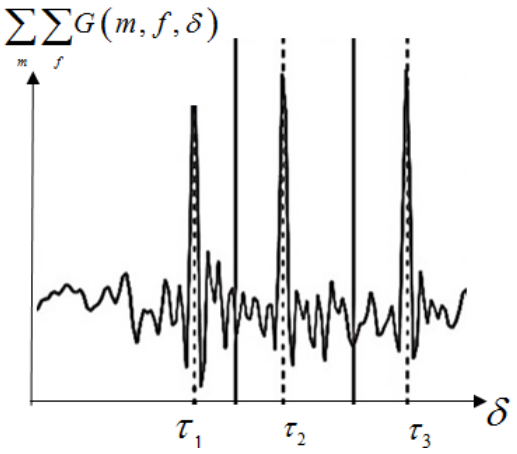
1. Generalized Cross Correlation
2. GCC-Phase Transform



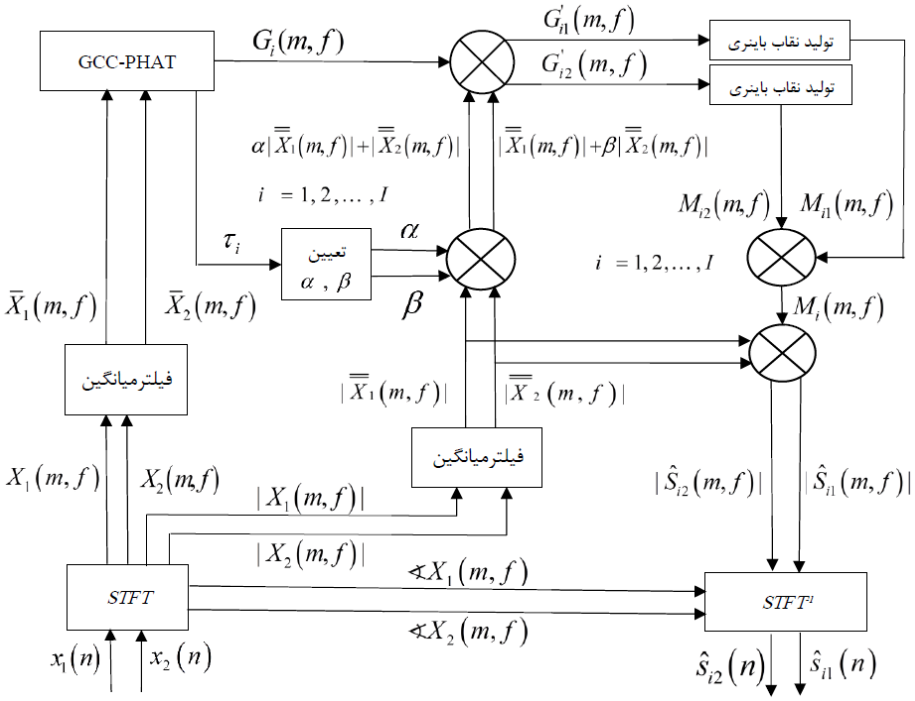
شکل ۲: بلوک دیاگرام روش GCC-PHAT [۱۰].



شکل ۳: نمایش دوبعدی اسپکتوگرام زاویه‌ای ($\sum_f G(m,f,\delta)$) روی ابعاد m و δ که سه خط تیره‌رنگ بیانگر تأخیر زمانی سه منبع گفتار هستند.



شکل ۴: نمایش یک‌بعدی اسپکتوگرام زاویه‌ای ($\sum_m \sum_f G(m,f,\delta)$) روی بعد δ که سه نقطه مشخص‌شده، تأخیر زمانی سه منبع گفتار را نشان می‌دهند.



شکل ۵: بلوک دیاگرام روش پیشنهادی.

داده می‌شود. در بخش ۳-۳ تولید نقاب باینری و در بخش ۳-۴ بازسازی منابع گفتار تفکیک‌شده بررسی می‌شوند.

۳-۱ فیلتر میانگین

تداخل امواج صوتی باعث ایجاد تغییرات شدید در طیف سیگنال گفتار می‌شود. بنابراین در این مقاله از فیلتر میانگین به عنوان یک پیش‌پردازش برای هموارکردن طیف سیگنال ترکیب‌شده جهت بهبود مکان‌یابی منابع استفاده می‌شود. همچنین استفاده از فیلتر میانگین باعث کاهش اثر تداخل در منابع بازسازی‌شده می‌گردد [۱۳] و [۱۸]. مطابق شکل ۵ فیلتر میانگین در دو جا استفاده شده است. در مورد اول، ورودی فیلتر، طیف مختلط

سیگنال ترکیب‌شده ورودی است و چون تغییرات قسمت حقیقی و موهومی طیف گفتار شبیه هم است لذا فیلتر میانگین مطابق (۴) به هر فریم زمانی قسمت حقیقی و موهومی به صورت جداگانه اعمال می‌شود

$$\bar{X}_C(m,f) = \{\tfrac{1}{4}X_c(m,f-1) + \tfrac{1}{2}X_c(m,f) + \tfrac{1}{4}X_c(m,f+1)\}\ ,\ c=1,2 \quad (4)$$

در این رابطه c اندیس کانال و $\bar{X}_C(m,f)$ خروجی فیلتر میانگین است. این فیلتر به صورت تجربی و برای حصول تفکیک بهتر کانال‌ها به دست آمده و در این فیلتر هر نمونه، از نمونه قبلی و بعدی آن تأثیر می‌بینند. تأثیر نمونه‌های بیشتر باعث افزایش محاسبات می‌شود و در برخی موارد، باعث کاهش کیفیت صدای خروجی می‌گردد. ضمناً اوزان فیلتر چنان انتخاب می‌شود که باعث سادگی محاسبات می‌شود. یعنی ضرایب وزنی، توان‌های منفی از عدد دو انتخاب شده‌اند که این موضوع باعث سادگی در پیاده‌سازی سخت‌افزاری می‌گردد. در مورد دوم و طبق (۵) فیلتر میانگین به اندازه تبدیل فوریه ($|X_C(m,f)|$) اعمال می‌شود

$$\left|\bar{\bar{X}}_C(m,f)\right| = \{\tfrac{1}{4}|X_c(m,f-1)| + \tfrac{1}{2}|X_c(m,f)| + \tfrac{1}{4}|X_c(m,f+1)|\}\ ,\ c=1,2 \quad (5)$$

که در آن، $\left|\bar{\bar{X}}_C(m,f)\right|$ اندازه طیف هموارشده است. بعد از هموارسازی طیف سیگنال ترکیب‌شده، مکان‌یابی منابع گفتار انجام می‌شود، با این تفاوت که به جای $X_1(m,f)$ و $X_2(m,f)$ از $\bar{X}_1(m,f)$ و $\bar{X}_2(m,f)$ استفاده می‌گردد. با توجه به تأخیر زمانی پیداشده برای هر منبع (τ_i)، مطابق روش [۱۲]، $G_i(m,f)$ از اسپکتوگرام زاویه‌ای $(G(m,f,\delta))$ به ازای $\delta=\tau_i$ به دست می‌آید یعنی

$$G_i(m,f) = G(m,f,\delta)\big|_{\delta=\tau_i} \quad (6)$$

۳-۲ وزن‌دهی کانال‌ها بر اساس مکان منابع

بعد از مکان‌یابی منابع گفتار و یافتن τ_iها، می‌توان با وزن‌دادن مناسب به $\left|\bar{\bar{X}}_C(m,f)\right|$ باعث بهبود کیفیت صدای خروجی شد. روش کار این است: منبعی که به میکروفون ۱ نزدیک‌تر باشد، ضرایب تقریب طیف آن منبع برای کانال ۱ تقویت می‌گردند و بالعکس ضرایب تقریب طیف آن برای کانال ۲ تضعیف می‌شوند و عکس همین موضوع برای منبعی که به میکروفون ۲ نزدیک است نیز برقرار است. اگر منبعی در محدوده میانی بین دو میکروفون باشد، ضرایب تقریب طیف آن منبع برای هر دو کانال بدون تغییر است. رابطه زیر نحوه یافتن وزن‌های مناسب را نشان می‌دهد

$$\left\{\begin{array}{ll} \alpha>1\ ,\ \beta<1 & \text{if}\quad -\dfrac{D}{V_\cdot}\le\tau_i<-\dfrac{D}{4V_\cdot} \\ \alpha=\beta=1 & \text{if}\quad -\dfrac{D}{4V_\cdot}\le\tau_i\le+\dfrac{D}{4V_\cdot} \\ \alpha<1\ ,\ \beta>1 & \text{if}\quad +\dfrac{D}{4V_\cdot}<\tau_i\le+\dfrac{D}{V_\cdot} \end{array}\right\} \quad (7)$$

در این رابطه α و β به ترتیب ضرایب وزنی کانال‌های ۱ و ۲ هستند. همان طور که مشاهده می‌شود محدوده میانی بین دو زمان $(-D/4V_\cdot)$ و $(D/4V_\cdot)$ در نظر گرفته شده است. ذکر این نکته لازم است که برای ساده‌ترشدن الگوریتم، وقتی $-D/V_\cdot\le\tau_i<-(D/4V_\cdot)$ دو عدد برای α و β در نظر گرفته می‌شود و وقتی $(D/4V_\cdot)<\tau_i\le D/V_\cdot$ فقط جای α و β عوض می‌شود. بعد از اختصاص α و β، $\left|\bar{\bar{X}}_1(m,f)\right|$ و $\left|\bar{\bar{X}}_2(m,f)\right|$ وزن‌دهی می‌شوند و در فاز منابع $(G_i(m,f))$ ضرب می‌شوند. $G'_{i1}(m,f)$ و $G'_{i2}(m,f)$ فاز وزن داده شده کانال‌های ۱ و ۲ هستند و از (۸) به دست می‌آیند

$$\left\{\begin{array}{l} G'_{i1}(m,f) = G_i(m,f)[\alpha\left|\bar{\bar{X}}_1(m,f)\right| + \left|\bar{\bar{X}}_2(m,f)\right|] \\ G'_{i2}(m,f) = G_i(m,f)[\left|\bar{\bar{X}}_1(m,f)\right| + \beta\left|\bar{\bar{X}}_2(m,f)\right|] \end{array}\right\} \quad (8)$$

وزن‌دادن به سیگنال ترکیب‌شده بر اساس مکان منابع و ضرب آن در $(G_i(m,f))$ برای ساختن نقاب باینری است. هرچه α و β بهتر انتخاب شوند، نقاب باینری ساخته‌شده به سمت نقاب ایده‌آل نزدیک‌تر است و تفکیک منابع بهتر انجام می‌شود. انتخاب α و β بهینه، توسط آزمایش شنیداری منطبق بر استاندارد ITU [۱۹] انجام می‌شود.

۳-۳ نقاب باینری

هدف از محاسباتی که تا کنون انجام گردید، اختصاص نقاط زمان- فرکانس طیف سیگنال ترکیب‌شده به منابع گفتار بر اساس فاصله آنها نسبت به دو میکروفون است. این بدان معنی است که برای هر کانال از هر منبع، یک نقاب باینری ساخته می‌شود تا بتواند مقادیر زمان فرکانس منبع iام را جدا کرده و تداخل بقیه منابع را حذف کند. با مقایسه مقادیر $G'_{i1}(m,f)$ و $G'_{i2}(m,f)$ برای منابع مختلف، می‌توان نقاب‌های باینری $M_{i1}(m,f)$ و $M_{i2}(m,f)$ را به ترتیب برای کانال اول و دوم منبع iام ایجاد کرد. این مقایسه برای تمام نقاط زمان- فرکانس انجام می‌شود. اگر یک نقطه زمان- فرکانس مربوط به منبع iام و کانال cام باشد، مقدار آن یک و در غیر این صورت مقدار آن صفر است. رابطه (۹) نحوه ساختن نقاب باینری $M_{ic}(m,f)$ (نقاب باینری منبع iام و کانال cام) را نشان می‌دهد. بقیه نقاب‌ها به طور مشابه ساخته می‌شوند

$$M_{ic}(m,f) = \left\{\begin{array}{l} 1 \text{ if } G'_{ic}(m,f) \ge G'_{jc}(m,f) \\ \quad ,\ i\ne j\ ,\ i,j=1,2,\ldots,I\ ,\ c=1,2 \\ 0 \text{ if } G'_{ic}(m,f) < G'_{jc}(m,f) \\ \quad ,\ i\ne j\ ,\ i,j=1,2,\ldots,I\ ,\ c=1,2 \end{array}\right\} \quad (9)$$

نقاب باینری هر کانال، وظیفه حذف تداخل همان کانال را دارد یعنی $M_{i1}(m,f)$ تداخل‌های مزاحم را برای کانال ۱ و منبع iام حذف می‌کند. به طور مشابه $M_{i2}(m,f)$ تداخل‌های مزاحم را برای کانال ۲ و منبع iام حذف می‌کند. چون نقاب‌ها ایده‌آل نیستند بنابراین هنگام بازسازی گفتار خروجی، برخی تداخل‌ها فقط در یک سمت شنیده می‌شوند. برای بهترشدن عملکرد نقاب‌گذاری در حذف تداخل، نقاب کانال‌های ۱ و ۲ به دست آمده از (۹) به ازای هر منبع در هم ضرب می‌شوند

$$M_i(m,f) = M_{i1}(m,f)M_{i2}(m,f) \quad (10)$$

که $M_i(m,f)$ نقاب باینری منبع iام است. این نقاب در هر دو کانال اندازه طیف سیگنال ترکیب‌شده ضرب می‌شود و صدای منبع iام را جدا و تداخل‌های مزاحم را حذف می‌کند.

۳-۴ بازسازی منابع

بعد از آن که برای هر منبع، نقاب باینری ساخته شد، مطابق شکل ۵ نقاب مذکور در $\left|\bar{\bar{X}}_C(m,f)\right|$ ضرب می‌شود و تخمینی از اندازه طیف منابع $(\left|\hat{S}_{ic}(m,f)\right|)$ را می‌دهد

جدول ۱: تأثیر تغییرات ضرایب وزنی α و β بر روی معیارهای ارزیابی SIR، SDR و SAR به صورت میانگین وقتی فاصله بین دو میکروفون ۵ سانتی‌متر است.

ضرایب وزنی		معیار		
α	β	SAR	SDR	SIR
۱	۱	۵٫۰۳	۲٫۲۳	۷٫۲۴
۱٫۲	۰٫۸	۴٫۰۴	۱٫۵۷	۱۰٫۳۲
۱٫۴	۰٫۶	۳٫۹۸	۱٫۰۱	۱۱٫۳۸
۱٫۸	۰٫۴	۳٫۸۱	۰٫۸۵	۱۱٫۹۱
۲	۰٫۲	۳٫۶۵	۰٫۶۴	۱۲٫۵۰

جدول ۳: نتایج مقایسه روش پیشنهادی با روش رقیب و نقاب باینری ایده‌آل با معیارهای SIR، SDR و SAR به صورت میانگین که روی داده‌های SiSEC ۲۰۱۶ DEV۱ اجرا شده است. نمرات بالاتر نشان‌دهنده تفکیک بهتر منابع گفتار است.

معیار زمان		معیارهای تفکیک			
۴ گوینده	۳ گوینده	SIR	SDR	SAR	روش
۵٫۷۴	۵٫۰۸	۸٫۵۸	۱٫۴۶	۳٫۹۹	روش پیشنهادی
۸۵۱٫۶۰	۶۴۴٫۸۲	۵٫۹۰	۳٫۰۰	۶٫۱۸	GCC-NMF [۱۱]
–	–	۹٫۳۳	۸٫۹۹	۹٫۳۱	IBM [۱۹]

$$\left|\hat{S}_{ic}(m,f)\right| = M_i(m,f)\left|\overline{\overline{X}}_C(m,f)\right| \quad , \ i = 1,2,\ldots,I \ , \ c = 1,2 \tag{11}$$

در این رابطه $M_i(m,f)$ در هر دو کانال $\left|\overline{\overline{X}}_C(m,f)\right|$ ضرب می‌شود. نقاب ساخته‌شده برای هر منبع، در اندازه طیف هموارشده ضرب گردیده است. در این تحقیق، فاز هر نمونه زمان- فرکانس، همان فاز سیگنال ترکیب‌شده در نظر گرفته می‌شود. در انتها با افزودن فاز سیگنال ترکیب‌شده ($\measuredangle X_c(m,f)$) به مقادیر اندازه و عکس تبدیل فوریه زمان‌کوتاه ($STFT^{-1}$) شکل زمانی منابع ($\hat{s}_{ic}(n)$) موجود در سیگنال ترکیب‌شده به دست می‌آید

$$\hat{s}_{ic}(n) = STFT^{-1}\left[\left|\hat{S}_{ic}(m,f)\right| \measuredangle X_c(m,f)\right] \quad , \ i = 1,2,\ldots,I \ , \ c = 1,2 \tag{12}$$

بدیهی است که سیگنال ترکیب‌شده ورودی دوکاناله بود و سیگنال خروجی منابع تفکیک‌شده نیز دوکاناله است.

٤- نتایج شبیه‌سازی و بحث

برای ارزیابی روش پیشنهادی در این مقاله از داده‌ها و معیارهای ارزیابی پایگاه داده SiSEC[1] استفاده شده است [۲۰]. این پایگاه شامل انواع مختلفی از سیگنال ترکیب‌شده است. در این مقاله سیگنال ترکیب‌شده از چند گوینده هم‌زمان و سیگنال ترکیب‌شده از یک گوینده در محیط نویزی بررسی می‌شود. SiSEC ۲۰۱۶ dev۱، شامل ۱۶ سیگنال گفتار ترکیب‌شده بوده که به صورت زنده ضبط شده‌اند و تعداد منابع موجود در سیگنال‌های ترکیب‌شده ۳ یا ۴ است. در این داده‌ها، سیگنال گفتار، ترکیب‌شده از گفتار سه زن، سه مرد، چهار زن و چهار مرد وجود دارد. همچنین انواع مختلفی از صداهای فوق از نظر فاصله بین دو میکروفون و زمان انعکاس اتاق موجود است. سیگنال ترکیب‌شده، با فاصله بین دو میکروفون ۵ سانتی‌متر و ۱۰۰ سانتی‌متر و همچنین با زمان

1. Signal Separation Evaluation Campaign

جدول ۲: تأثیر تغییرات ضرایب وزنی α و β بر روی معیارهای ارزیابی SIR، SDR و SAR به صورت میانگین وقتی فاصله بین دو میکروفون ۱۰۰ سانتی‌متر است.

ضرایب وزنی		معیار		
α	β	SAR	SDR	SIR
۱	۱	۵٫۱۱	۲٫۲	۵٫۸۷
۱٫۵	۰٫۷	۴٫۵۱	۱٫۷۰	۶٫۰۳
۲	۰٫۵	۳٫۹۵	۱٫۳۶	۶٫۸۵
۲٫۵	۰٫۳	۳٫۷۸	۱٫۲۸	۷٫۰۹
۳	۰٫۱	۳٫۶۵	۱٫۲۰	۸٫۲۵

انعکاس اتاق ۱۳۰ میلی‌ثانیه و ۲۵۰ میلی‌ثانیه ساخته شده است. فرکانس نمونه‌برداری سیگنال‌های فوق ۱۶۰۰۰ هرتز و طول سیگنال ۱۰ ثانیه است. سیگنال‌های گفتار مرجع (ایزوله‌شده) نیز در پایگاه داده برای ارزیابی الگوریتم تفکیک موجود است.

پایگاه داده SiSEC، یک برنامه آماده با کد Matlab به نام BSS_EVAL برای ارزیابی الگوریتم تفکیک در اختیار محققان قرار داده است [۲۱]. سه معیار اصلی در این برنامه، نسبت سیگنال به اعوجاج (SDR)[2]، نسبت سیگنال به تداخل (SIR)[3] و نسبت سیگنال به مصنوعی‌شدن (SAR)[4] است.

در این مقاله برای محاسبه $STFT$، سیگنال گفتار در پنجره هنینگ به طول ۱۲۸۰ نمونه و مقدار پرش ۱۲۸ نمونه که به صورت تجربی به دست آمده‌اند، ضرب شده و از حاصل‌ضرب، تبدیل فوریه ۲۰۴۸ نقطه‌ای گرفته شده است. وقتی فاصله دو میکروفون از هم ۵ سانتی‌متر است، در (۳) ضریب $\gamma = 2$ برای مکان‌یابی منابع استفاده شده است. از آنجا که تغییر فاصله بین دو میکروفون باعث تغییر ضرایب وزنی می‌شود، روش پیشنهادی برای فاصله بین دو میکروفون، ۵ سانتی‌متر و ۱۰۰ سانتی‌متر، جدا ارزیابی شده است. جداول ۱ و ۲ معیارهای ارزیابی را به ازای وزن‌های مختلف α و β، برای فاصله بین دو میکروفون ۵ سانتی‌متر و ۱۰۰ سانتی‌متر روی داده‌های SiSEC ۲۰۱۶ dev۱ نشان می‌دهند. وقتی فاصله بین دو میکروفون ۵ سانتی‌متر است، $1 < \alpha < 2$ و $0.1 < \beta < 1$ بوده و وقتی فاصله بین دو میکروفون ۱۰۰ سانتی‌متر است $1 < \alpha < 10$ و $0.1 < \beta < 1$ می‌باشد. برای یافتن ضرایب وزنی α و β بهینه، ۱۰۰ حالت مختلف از تغییرات α و β آزمایش شده که در جداول ۱ و ۲ فقط ۵ حالت آن نمایش داده شده است. تعیین α و β بهینه توسط آزمایش شنیداری منطبق بر استاندارد ITU [۱۹] انجام شده است. یعنی به ازای مقادیر مختلف α و β، سیگنال منابع گفتار استخراج گردیده و توسط افراد شرکت‌کننده در آزمون شنیداری ارزیابی گردیده است. مقادیر α و β بهینه، وقتی فاصله بین دو میکروفون ۵ سانتی‌متر است، $\alpha = 1.2$ و $\beta = 0.8$ و برای فاصله ۱۰۰ سانتی‌متر میکروفون‌ها، $\alpha = 2$ و $\beta = 0.5$ است. البته می‌توان با افزایش α و کاهش β، تداخل بیشتری را حذف کرد (افزایش SIR) ولی این امر باعث افزایش اعوجاج (کاهش SDR) و مصنوعی‌شدن (کاهش SAR) صدای خروجی می‌شود. لذا باید یک مصالحه بین معیارهای ارزیابی صورت پذیرد.

نتایج ارزیابی و مقایسه زمان اجرای روش پیشنهادی با روش GCC-NMF [۱۳] و IBM [۲۲] در جدول ۳ آورده شده است. روش

2. Source Distortion Ratio
3. Source Interference Ratio
4. Source Artifact Ratio

(IBM) به معنای اعمال نقاب باینری ایده‌آل است. مقایسه جدول ۳ نشان می‌دهد روش پیشنهادی در مقایسه با روش GCC-NMF عملکرد بهتری در حذف تداخل دارد که مهم‌ترین دلیل آن، وزن‌دادن مناسب به سیگنال میکروفون‌ها بر اساس مکان منابع گفتار است. اگرچه صدای خروجی در روش پیشنهادی دارای اعوجاج و مصنوعی‌شدن بیشتری است (کاهش SDR و SAR)، اما این موضوع اثر کمتری نسبت به تداخل روی کیفیت شنیداری انسان دارد. در جدول ۳ مقایسه زمان اجرای هر دو الگوریتم با سخت‌افزار یکسان برای سیگنال ترکیب‌شده، شامل سه گوینده و چهار گوینده آورده شده است. زمان اجرای روش پیشنهادی بسیار کمتر است و افزایش تعداد منابع موجود در سیگنال ترکیب‌شده از سه به چهار باعث افزایش ۱۲ درصدی زمان اجرا در روش پیشنهادی و افزایش ۳۲ درصدی در روش رقیب است. همچنین روش GCC-NMF به مقادیر اولیه که به ماتریس‌های پایه و ضرایب داده می‌شود وابسته است و هر بار اجرای آن باعث خروجی متفاوت می‌شود.

۵- نتیجه‌گیری

در این مقاله روشی جدید معرفی گردید که تفکیک منابع گفتار را بر اساس موقعیت مکانی منابع گفتار و فاصله آنها از دو میکروفون انجام می‌دهد. از فیلتر میانگین برای هموارکردن طیف و کاهش اثر تداخل استفاده می‌شود. وجود ضرایب وزنی باعث بهبود کیفیت تفکیک منابع گفتار می‌شود. این روش بدون نیاز به آموزش و اطلاعات قبلی از منابع گفتار، تفکیک منابع گفتار موجود در سیگنال ترکیب‌شده را انجام می‌دهد و همچنین دارای پیچیدگی محاسباتی کمتری نسبت به رقیبان است. این روش برای ۳ و ۴ گوینده هم‌زمان و با وجود انعکاس محیط آزمایش شده است. در داده‌هایی که دو میکروفون به هم نزدیک هستند (فاصله ۵ سانتی‌متر)، حذف تداخل به خوبی انجام می‌شود و این موضوع اهمیت روش پیشنهادی را در مکالمات تلفن همراه نشان می‌دهد.

مراجع

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875-1902, Sept. 2005.

[2] K. Itakura, *et al*., "Bayesian multichannel audio source separation based on integrated source and spatial models," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 4, pp. 831-846, Apr. 2018.

[3] Y. Xie, K. Xie, Z. Wu, and S. Xie, "Underdetermined blind source separation of speech mixtures based on K-means clustering," in *Proc. Chinese Control Conf., CCC'19*, pp. 42-46, Guangzhou, China, 27-30 Jul. 2019.

[4] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'97,* vol. 1, pp. 375-378, Munich, Germany, 21-24 Apr. 1997.

[5] Z. Ding, W. Li, and Q. Liao, "Dual-channel speech separation by sub-segmental directional statistics," in *Proc. Int. Conf. on Wireless Communications, Signal Processing and Networking, WiSPNET'16*, pp. 2287-2291, Chennai, India, 23-35 Mar. 2016.

[6] X. Li, Z. Ding, W. Li, and Q. Liao, "Dual-channel cosine function based ITD estimation for robust speech separation," *Sensors*, vol. 17, no. 6, Article No.: 1447, 13 pp. 2017.

[7] T. Maitheen and M. S. Lekshmi, "Enhancement of DUET blind source separation using wavelet," *International Research Journal of Engineering and Technology*, vol. 4, no. 5, pp. 3551-3553, May 2017.

[8] X. Zhang and D. Wang, "Binaural reverberant speech separation based on deep neural networks," in *Proc. Interspeech*, vol. pp. 2018-2022, Stockholm, Sweden, 20-24 Aug. 2017.

[9] S. U. N. Wood, *et al.*, "Blind speech separation and enhancement with GCC-NMF," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 745-755, Apr. 2017.

[10] Y. Yu, W. Wang, J. Luo, and P. Feng, "Localization based stereo speech separation using deep networks," in *Proc. IEEE Int. Conf. Digit. Signal Process,* pp. 153-157, Singapore, Singapore, 21-24 Jul. 2015.

[11] S. U. N. Wood and J. Rouat, "Unsupervised low latency speech enhancement with RT-GCC-NMF," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 332-346, May 2019.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust.*, vol. 24, no. 4, pp. 320-327, Aug. 1976.

[13] M. A. J. Sathya and S. P. Victor, Noise Reduction Techniques and Algorithms for Speech Signal Processing, .

[14] A. P. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'01,* vol. 5, pp. 3381-3384, Salt Lake City, UT, USA, 7-11 May 2001.

[15] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950-1960, Aug. 2012.

[16] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," in *Proc. IWAENC*, 4 pp., Seattle, WA, USA, 14-17 Sept. 2008.

[17] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Proc. 9th Int Conf. on Latent Variable Analysis and Signal Separation*, 8 pp., St. Malo, France, 27-30 Sept. 2010.

[18] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'08,* vol. 1, pp. 45-48, Las Vegas, NV, USA, 30 Mar- 4 Apr. 2008.

[19] [Online]. Available: www.itu.com

[20] [Online]. Available: https://sisec.wiki.irisa.fr.

[21] C. Fevotte, R. Gribonval, and E. Vincent, *BSS_EVAL Toolbox User Guide--Revision 2.0*, 2005.

[22] A. Liutkus, *et al.*, "The 2016 signal separation evaluation campaign," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, pp. 323-332, Grenoble, France, Feb. 2017.

حسن علی‌صوفی مدرک کارشناسی برق گرایش مخابرات خود را در سال ۱۳۸۱ از دانشگاه شهید باهنر کرمان اخذ نموده است. پس از آن مدرک کارشناسی ارشد مخابرات، گرایش سیستم را در سال ۱۳۹۸ از دانشکاه فردوسی دریافت کرد. زمینه‌ی علاقمندی ایشان پردازش سیگنال تصویر و صدا است.

مرتضی خادمی تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی برق به‌ترتیب در سال‌های ۱۳۶۴ و ۱۳۶۶ در دانشگاه صنعتی اصفهان به پایان رسانده است. ایشان از سال ۱۳۶۶ تا ۱۳۷۰ به عنوان عضو هیات علمی (مربی) در دانشگاه فردوسی مشهد به کار مشغول بود. پس از آن به دوره دکترای مهندسی برق در دانشگاه ولونگونگ (استرالیا) وارد گردیده و در سال ۱۳۷۴ موفق به اخذ درجه دکترا در مهندسی برق از دانشگاه مذکور گردید. دکتر خادمی از سال ۱۳۷۴ مجدداً در دانشکده مهندسی دانشگاه فردوسی مشهد مشغول به فعالیت گردید و اینک نیز استاد این دانشکده است. زمینه‌های علمی مورد علاقه نام‌برده شامل موضوعاتی مانند مخابرات ویدئویی، فشرده‌سازی ویدئو، پردازش تصویر، پردازش سیگنال‌های پزشکی و پنهان‌سازی اطلاعات در ویدئو می‌باشد.

عباس ابراهیمی مقدم مدرک کارشناسی و کارشناسی ارشد برق گرایش مخابرات خود را به ترتیب از دانشگاه‌های صنعتی شریف و صنعتی خواجه نصیر اخذ کرده است. ایشان مدرک دکتری خود را از دانشگاه مک‌مستر کانادا دریافت کرده و از سال ۱۳۹۰ در دانشگاه فردوسی مشهد مشغول تدریس و تحقیق می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان پردازش گفتار، پردازش تصویر و ویدیو، بینایی ماشین و پردازش سیگنال‌های حیاتی می‌باشد.