پیاده‌سازی سیستم‌های ترکیبی حذف انعکاس از گفتار و ارائه دادگان دوکاناله گفتار فارسی برای سنجش کارآیی آنها

فرهاد فغانی[۱] – حمیدرضا ابوطالبی[۲]

(۱) استادیار - دانشکده برق، دانشگاه آزاد اسلامی، واحد نجف‌آباد

(۲) دانشیار - دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد

تاریخ دریافت: پاییز۱۳۹۱ تاریخ پذیرش: زمستان ۱۳۹۱

خلاصه: در این مقاله جوانب مختلف مسأله حذف انعکاس از سیگنال گفتارفارسی بررسی می‌شود. به عنوان مصالحه‌ای بین میزان بهسازی گفتار از یکسو و پیچیدگی و بار محاسباتی روش از سوی دیگر، تعداد دو میکروفون در سیستم در نظر گرفته شده است. به منظور امکان سنجش و مقایسه کارآیی سیستم‌های مورد نظر، ابتدا مجموعه لغات فارسی برای آزمون قابلیت فهم گفتار تهیه و سپس با استفاده از آن، دادگان دوکاناله گفتار فارسیِ دارای انعکاس طراحی و ضبط گردید. در این تحقیق، روش‌های پایه (تک‌کاناله و دوکاناله) حذف انعکاس از گفتار مطالعه و پیاده سازی و با استفاده از دادگان تهیه شده مقایسه گردیده‌اند. بر پایه این مطالعه، یک سری از روش‌های ترکیبی که قابلیت فهم و کیفیت سیگنال گفتار آغشته به انعکاس را به نحو مطلوبی افزایش می‌دهد، ارائه شده‌است و نشان داده شده است که ترکیب سه روش delay-and-sum beamforing، فیلتر معکوس، و روش مبتنی بر خطای فاز نسبت به سایر سیستم‌های ترکیبی از لحاظ میزان کیفیت و قابلیت فهم گفتار خروجی منجر به نتایج بهتری می‌شود.

کلمات کلیدی: بهبود قابلیت فهم، حذف انعکاس، بهسازی سیگنال آغشته به انعکاس، دادگان گفتارفارسی، آزمون DRT.

۱- مقدمه

هرچند به طور معمول و منطقی، نویز به عنوان اصلی‌ترین عامل مخرب سیگنال گفتار و کاهش کارآیی سیستم‌های ارتباط گفتاری مطرح می‌شود ولی در بسیاری موارد اثر تخریبی انعکاس را نمی‌توان در مقابل نویز نادیده گرفت. در دهه‌های اخیر، برای بهسازی سیگنال گفتار نویزی روش‌های بسیار متعددی ارائه شده ولی توجه صِرف به بحث انعکاس و بهسازی (و بهبود قابلیت فهم) سیگنال گفتار آغشته به انعکاس هنوز جای کار فراوان دارد.

تکنیک‌های متعددی به صورت تک‌کاناله (تک‌میکروفونه) به بهسازی سیگنال گفتار دارای انعکاس می‌پردازند؛ از طرف دیگر، الگوریتم‌های زیادی نیز برای بهسازی سیگنال گفتار با استفاده از چند کانال ورودی (چند میکروفون) ارائه شده است. علیرغم بهره قابل توجهی که با استفاده از چند کانال ورودی در بهسازی و بهبود قابلیت فهم حاصل می‌گردد، باید به مصالحه‌ای بین این مزایا از یکسو و پیچیدگی و حجم محاسبات از سوی دیگر دست یافت: هر چه تعداد میکروفون‌ها بیشتر باشد، کیفیت کار بهتر ولی پیچیدگی سیستم و حجم محاسبات نیز افزایش می‌یابد. با در نظر گرفتن این مسأله، استفاده از سیستم‌های دومیکروفونه به عنوان یک حد وسط منطقی به نظر می‌رسد. روشهای متنوعی برای حذف انعکاس موجود است. به عنوان مثال می‌توان به استفاده از روش فیلتر معکوس [۱] و یا استفاده از روش‌های آماری [۲] اشاره نمود. ولی این روش‌ها عمدتاً در کانالهای زیاد (بیش از دو کانال) بازدهی خوبی نشان می‌دهند که این امر منجر به پیچیدگی روند پیاده‌سازی فیزیکی و همچنین پیچیدگی محاسباتی می‌شود. امروزه بسیاری از سیستم‌های مخابراتی از دو کانال صوتی استفاده می‌نمایند.

هدف از این تحقیق، بررسی مسأله بهسازی سیگنال گفتار آغشته به انعکاس با استفاده از سیستم‌های دومیکروفونه و با تأکید بر روی افزایش میزان قابلیت فهم می‌باشد. از این رو، مجموعه کلمات لازم برای انجام آزمون DRT (Diagnostic Rhyme Test) در زبان فارسی استخراج و دادگانی دوکاناله از سیگنال گفتار دارای انعکاس تهیه گردید. در سراسر این تحقیق، ارزیابی روش‌های موجود (قبلی) و پیشنهادی (جدید) بر روی این دادگان صورت پذیرفته است.

در مقاله حاضر، جوانب مختلف مسأله بهسازی و بهبود قابلیت فهم گفتار آغشته به انعکاس به طور جامع مورد بررسی و مطالعه قرار گرفته است. در بخش دوم به صورت اجمالی مفهوم انعکاس و نحوه اندازه‌گیری آن

بررسی شده است. بخش سوم، به بحث در مورد نحوه تعیین تعداد و چیدمان میکروفونها اختصاص دارد. روش‌های مهم و پایه حذف انعکاس از گفتار در بخش چهارم معرفی شده است. در بخش پنجم، پس از معرفی آزمون‌های استاندارد سنجش قابلیت فهم گفتار، به تشریح نحوه تهیه مجموعه لغات و دادگان سیگنال گفتار دوکاناله مورد نیاز در این پژوهش پرداخته می‌شود. ارزیابی و مقایسه روش‌های پایه حذف انعکاس در بخش ششم گزارش شده و در همانجا سیستم‌های ترکیبی معرفی و ارزیابی می‌شود. بخش هفتم نیز به جمع‌بندی و نتیجه‌گیری تحقیق اختصاص دارد.

۲- نگاهی اجمالی بر مفهوم انعکاس

۲-۱- تعریف انعکاس

انعکاس، حضور صوت است پس از آن که صوت اصلی از سیگنال دریافت شده حذف گردد. در واقع انعکاس، بازتاب‌های همان سیگنال صوتی است که با تأخیرهای متفاوت و به صورت تضعیف شده، دریافت می‌شود. این پدیده هنگامی بهتر نمایان می‌شود که منبع صوتی سکوت می‌کند ولی هنوز بازتاب‌های آن با تضعیف دریافت می‌شوند؛ به زبان علمی‌تر می‌توان سیگنال دریافت شده در میکروفون(ها) را به صورت زیر بیان نمود [۳]:

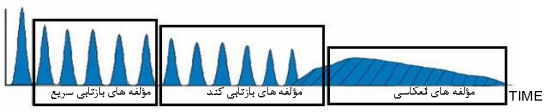
$$y(n)=\sum_{k=0}^{\infty}a_k x(n-k)+w(n) \qquad (۱)$$

که در آن a_kها ضرایب تضعیف و $w(n)$ نویز دریافتی است. با صرف نظر از نویز دریافتی، رابطه فوق را می‌توان به صورت زیر بازنویسی نمود:

$$y(n)=\sum_{k=0}^{\infty}a_k x(n-k) \qquad (۲)$$

اگرچه در برخی موارد (مانند فضاهای داخلی مساجد، کلیساها، سالن‌های نمایش و ...) از انعکاس به عنوان یک پدیده هنری مطلوب برای جلب نظر شنونده بهره گرفته می‌شود، لیکن در بسیاری از موارد این بازتاب‌های صوتی باعث ناراحتی و آزردگی خاطر شنونده شده و کاهش کیفیت و قابلیت فهم گفتار را به دنبال دارد. هنگامی که سخن از حذف انعکاس می‌شود، معمولاً هدف بالا بردن کیفیت یا قابلیت فهم گفتار، و یا بهبود بهره سیستم‌های بازشناسی خودکار گفتار می‌باشد. در پژوهش حاضر، مسأله حذف انعکاس به عنوان پیش‌پردازشی برای بازشناسی گفتار مطرح نبوده و هدف نهایی، همان بهبود کیفیت و قابلیت فهم گفتار خروجی است.

البته باید متذکر شد مفهومی که در این تحقیق از خود انعکاس مدّنظر می‌باشد، حالتی است که در اثر بازتاب سیگنال از برخورد با سطوح اجسام مختلف موجود در محیط به وجود آمده و در برخی موارد به آن طنین یا پژواک نیز می‌گویند.



شکل (۱): بخش‌های مختلف سیگنال آغشته به انعکاس. مؤلفه‌های بازتابی کُند و مؤلفه‌های انعکاسی سبب کاهش قابلیت فهم گفتار می‌گردند.

Fig (1): Several parts and reflections. Late speech reverberant of speech reduce reverberation intelligibility

حالت دیگری از انعکاس که در سیستم‌های مخابراتی در اثر نشت الکتریکی بین فرستنده-گیرنده در سیگنال ظاهر می‌شود، نیز وجود دارد که خارج از بحث این مقاله است.

با یک نگاه دقیق‌تر به سیگنال دریافتی در میکروفون(ها)، می‌توان آن(ها) را مطابق با شکل (۱) به‌گونه دیگری نیز بخش‌بندی نمود: مؤلفه اصلی، مؤلفه‌های بازتابی سریع (early reflections)، مؤلفه‌های بازتابی کُند (late reflections) و مؤلفه‌های انعکاسی [۴].

هنگامی که سیگنال گفتار در فضای اتاق پخش می‌شود، از چند طریق به گیرنده می‌رسد: مسیر اول، مسیر مستقیم است. سیگنال رسیده از این مسیر دارای بیشترین دامنه و همراه با تأخیری ثابت، بر حسب فاصله بین منبع و گیرنده است. مسیرهای دسته دوم، مسیرهایی هستند که معمولاً از بازتاب سیگنال اصلی پس از برخورد با سطوح بزرگ مثل پنجره و دیوار به گیرنده رسیده و از آنها با نام مؤلفه بازتابیِ سریع یاد می‌شود و دارای دامنه قابل توجهی هستند. این دسته از مؤلفه‌ها که در رنگی‌کردن (Coloration) طیف تأثیر دارد، حداکثر حدود 60ms نسبت به مؤلفه اصلی تأخیر دارند. رنگی کردن را می‌توان به عنوان انحراف استاندارد پاسخ فرکانسی محیط تعبیر کرد. دسته سوم و چهارم مؤلفه‌ها توسط بازتاب‌های مکرر از اشیاء درون محیط به گیرنده رسیده و عامل اصلی تخریب و کاهش قابلیت فهم سیگنال گفتار است. در برخی مراجع مرتبط با علم آکوستیک، در یک بیان دقیق‌تر، از این دو دسته آخر با عنوان پژواک (Reverberation) یاد می‌شود.

انعکاس در محیط پدیده‌ای ایستان نیست. علت آن نیز جابجایی گوینده و یا اشیاء درون اتاق است. حتی زمانی‌که گوینده سر خود را نیز تکان می‌دهد این پاسخ ضربه تغییر می‌کند. پاسخ ضربه محیط با در دست داشتن سیگنال آغشته به انعکاس و سیگنال تمیز (بدون انعکاس) قابل محاسبه است؛ ولی از آنجائی‌که سیگنال تمیز معمولاً در دست نیست و در اکثر موارد علاوه بر انعکاس، نویز نیز وجود دارد، محاسبه پاسخ ضربه محیط، کار دشواری است [۵-۶].

۲-۲- اندازه‌گیری میزان انعکاس

میزان انعکاس موجود در محیط توسط معیاری به نام RT_{60} و یا زمان انعکاس سنجیده می‌شود. زمان انعکاس، مدت زمانی است که طول می‌کشد تا دامنه سیگنال صوتی در محیط، 60dB تضعیف شود. هرچه RT_{60} بیشتر باشد، قابلیت فهم سیگنال کمتر خواهد بود. مقدار قابل قبول برای زمان انعکاس به فضای پخش سیگنال و نوع سیگنال پخش شده بستگی دارد.

یکی دیگر از راه‌های معمول برای بیان میزان انعکاس موجود در یک سیگنال استفاده از پارامتر SRR (Signal-to-Reverberation Ratio) است. SRR عبارتست از نسبت توان مؤلفه مستقیم سیگنال به توان سایر مؤلفه‌ها (مؤلفه‌های انعکاسی). در واقع این معیار، مشابه همان SNR ولی برای نویز همبسته (انعکاس) می‌باشد. SRR از روی پاسخ ضربه آکوستیکی محیط، به روش زیر قابل تخمین است [۴]:

$$SRR(dB) = 10\log_{10}\frac{h^2(\delta)}{\sum_{l=0(l\neq\delta)}^{M-1} h^2(l)} \qquad (۳)$$

که در آن، h(n)، پاسخ ضربه آکوستیکی محیط بین گوینده-شنونده، M تعداد جملات و δ، زمان تأخیر (مسیر مستقیم) است.

به دست آوردن RT_{60} و SRR در کاربردهای واقعی (و نه در حالت شبیه‌سازی) پیچیده و شاید غیرممکن بوده و از این‌رو، در اکثر موارد در سنجش کارآیی سیستم‌های حذف انعکاس از معیارهای کیفی (که در ادامه مقاله به معرفی آنها می‌پردازیم)، استفاده می‌شود.

۳- انتخاب تعداد و نحوه چیدمان میکروفون‌ها

الگوریتم‌های موجود برای حل مسأله بهبود قابلیت فهم گفتار آغشته به انعکاس، عمدتاً به دو دسته تک‌کاناله (تک‌میکروفونه) و یا چندکاناله (چندمیکروفونه) تقسیم می‌شود. روش‌های تک‌میکروفونه، در بسیاری از سیستم‌های بازشناسی گفتار و یا سمعک‌ها کاربرد دارد. ولی در روش‌های تک‌کاناله، کیفیت سیگنال بازسازی شده از حدّ مشخصی فراتر نمی‌رود و از روش‌های چندمیکروفونه نتایج بهتری حاصل می‌شود [۵-۱۱].

معمولاً هرچه تعداد کانال‌ها بیشتر باشد، کیفیت سیگنال بهسازی شده بهتر خواهد بود. در عوض افزایش تعداد کانال‌ها، افزایش حجم محاسبات و پیچیدگی سیستم را در پی خواهد داشت. در نتیجه برای انتخاب تعداد کانال‌ها باید بین کیفیت سیگنال بهسازی شده و پیچیدگی سیستم به مصالحه‌ای دست یافت. در اکثر موارد، انتخاب دو کانال، انتخاب مناسبی خواهد بود.

برای آنکه حالت کلی از ضبط سیگنال را در نظر گرفته شده باشد، نمی‌توانیم مکان ثابتی برای گوینده نسبت به میکروفون‌ها در نظر بگیریم. لذا انتخاب میکروفون‌های همه‌جهته (Omni-directional) منطقی می‌باشد.

یک مسأله مهم، چگونگی قرار گرفتن این میکروفون‌هاست. اگرچه انواع مختلفی برای نحوه چیدمان میکروفون‌ها ارائه شده، ولی عمومی‌ترین نوع، و در بحث سیستم دومیکروفونه تنها گزینه ممکن، چیدمان خطی (و یکنواخت) میکروفون‌هاست.

با فرض دور بودن فاصله گوینده از محور میکروفون‌ها (فرض پردازش Far Field)، امواج صوتی رسیده به آرایه میکروفونی را می‌توان امواجی تخت در نظر گرفت. در چنین شرایطی، جهت جلوگیری از پدیده Spatial Aliasing، فاصله دو میکروفون باید در نامساوی زیر صدق کند:

$$d < \frac{\lambda_{min}}{2} \qquad (۴)$$

چنانچه محدوده فرکانسی مهم و قابل توجه سیگنال گفتار را (200Hz-2kHz) در نظر بگیریم، آنگاه از رابطه (۴)، نتیجه زیر قابل استخراج است:

$$d < \frac{\lambda_{min}}{2} = \frac{c/f_{max}}{2} \qquad (۵)$$

که در آن، c سرعت صوت و برابر 340m/s و f فرکانس صوت است. لذا خواهیم داشت:

$$d < 8.6cm \qquad (۶)$$

از این‌رو، دراین تحقیق، فاصله مناسب برای میکروفون‌ها، بین 7-8cm، پیشنهاد می‌شود. فاصله گوینده از میکروفون‌ها، به هدف ما از ضبط سیگنال و تعداد مورد نظر برای انعکاس موجود در سیگنال بستگی پیدا خواهدکرد. معمولاً مقادیر بیشتر از 1m برای فاصله گویندگان تا محور میکروفون‌ها مطلوب خواهد بود.

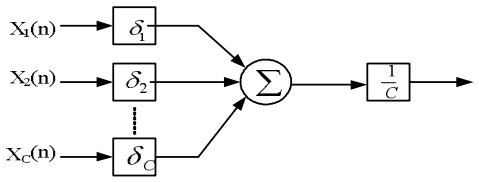
۴- روش‌های پایه بهسازی سیگنال گفتار آغشته به انعکاس

سیگنال جمع‌آوری‌شده توسط میکروفون‌ها در محیط معمولی، آغشته به نویز و انعکاس است. معمولاً میزان نویز اضافه شده مستقل از روابط بین فواصل میکروفون(ها) و گوینده است. در مقابل، انعکاس به محل قرار گرفتن میکروفون(ها) و گوینده بسیار وابسته است. با افزایش فاصله بین میکروفون و گوینده، مؤلفه‌های اصلی سیگنال، نسبت به مؤلفه‌های فرعی تضعیف شده و در نتیجه نسبت سیگنال به انعکاس (SRR) کاهش می‌یابد [۱۱]. همچنین با افزایش این فاصله، به دلیل تضعیف دامنه سیگنال اصلی (در مقایسه با نویز)، SNR نیز کاهش می‌یابد. در نتیجه باید به دنبال تکنیکی بود که دامنه مؤلفه اصلی (مستقیم) سیگنال را در مقایسه با نویز پس‌زمینه افزایش دهد و در عین حال باعث افزایش مؤلفه‌های انعکاسی هم نشود.

یکی از اصلی‌ترین مشکلات پیش‌رو، ماهیت غیرایستان سیگنال گفتار است. به دلیل تغییرات زمانی و فرکانسی سیگنال گفتار، SNR تابعی از زمان و همچنین تابعی از فرکانس است. این مسأله، کار تخمین سیگنال گفتار (تمیز) از روی نمونه تخریب‌شده آن (توسط نویز و انعکاس) را به شدت پیچیده نموده و در نهایت منجر به اُفت کارایی الگوریتمهای استاندارد حذف نویز مانند تفریق طیفی می‌شود [۲]. از سوی دیگر، بر اساس آنچه توسط Allen نشان داده شده است [۷]، اکثر محیط‌های واقعی، دارای پاسخ ضربه مینیمم‌فاز نبوده و بنابراین الگوریتم‌های خانواده deconvolution در آنها چندان خوب عمل نمی‌نماید.

در برخی از روش‌های موفق بهسازی گفتار دارای انعکاس از آنالیز پیش‌بینی خطی (LP: Linear Prediction) و کار بر روی سیگنال باقیمانده پیش‌بینی خطی (LP-Residual) استفاده شده است. سیگنال LP-Residual، مدلی از سیگنال تحریک می‌باشد. قسمت‌های واکدار گفتار، بر اثر تحریک تارهای صوتی تولید شده و ساختار در

زمان‌هایی که حنجره بسته می‌شود، این سیگنال تحریک به وضوح خود را نشان می‌دهد. در اکثر روش‌ها هم از همین وضوح برای بازسازی استفاده شده است. با مشخص کردن مکان‌هایی که حنجره بسته می‌شود، می‌توان سیگنال را در محل وقوع این نمونه‌ها، بازسازی نمود. به دلیل استحکام سیگنال در این نمونه‌ها SNR در اطرافشان، نسبت به سایر نقاط بالاست.



شکل (۲): بلوک دیاگرام سیستم Delay-and-Sum Beamforming.
Fig. (2): Block diagram of Delay-and-Sum Beamforming system

همان طور که قبلاً نیز عنوان شد، در شرایط عدم حضور نویز، کیفیت سیگنال آغشته به انعکاس به دو پارامتر عمده بستگی پیدا می‌کند: Coloration و Long-Term Reverberation. روش‌های تک‌کاناله معمولاً در دو مرحله انجام می‌شوند؛ مرحله اول -که دراکثر موارد مشترک است- شامل اعمال یک فیلتر معکوس از تابع تخمینی آکوستیکی محیط است که به حذف پدیده Coloration می‌پردازد [۲]. آنچه که متفاوت است، مرحله دوم این روش‌هاست. باید توجه داشت اکثر این روش‌ها، اگرچه به عنوان تک کاناله توصیف می‌شوند، ولی قابلیت تعمیم به چند کانال را نیز دارند.

در ادامه، چندین دسته از روش‌های پایه تک‌کاناله و چندکاناله برای بهسازی سیگنال گفتار آغشته به انعکاس مورد بررسی، قرار می‌گیرد. روش‌های پایه مورد بررسی عبارتند از:

- بهبود قابلیت فهم گفتار با استفاده از Delay-and-Sum beamforming.
- بهبود قابلیت فهم سیگنال گفتار با استفاده از LP-Residual.
- بهبود قابلیت فهم گفتار با استفاده از فیلتر معکوس.
- بهبود قابلیت فهم سیگنال گفتار با استفاده از ترکیب فیلتر معکوس و تفریق طیفی.
- بهبود قابلیت فهم گفتار با استفاده از فاز سیگنال دریافتی از دو میکروفون.

این تکنیک‌های پایه در پژوهش حاضر پیاده‌سازی و بررسی شده و در ادامه، سیستم‌هایی جدید (حاصل از ترکیب دو یا چند روش پایه) معرفی و ارزیابی می‌شود.

۴-۱- بهبود قابلیت فهم گفتار با استفاده از روش Delay-and-Sum Beam forming

روش Delay-and-Sum Beam forming، یکی از روش‌های پایه و رایج در حذف انعکاس به شمار می‌رود. ایده اصلی این روش، آن است که هنگامی که از یک آرایه میکروفون برای دریافت سیگنال استفاده می‌شود، سیگنال‌های دریافت شده با تأخیر نسبت به یکدیگر دریافت خواهند شد. میزان این تأخیرها به ساختار فیزیکی و چیدمان میکروفون‌ها بستگی پیدا می‌کند. در این روش سعی بر آن است که - تا حد امکان- این تأخیرها حذف گردند.

در این روش، یک کانال به عنوان مرجع در نظر گرفته می‌شود و TDOA (Time Delay Of Arrival) برای مابقی کانال‌ها محاسبه می‌گردد. سپس سیگنال‌های تأخیر داده شده با یکدیگر جمع می‌شوند:

$$\hat{s}(n)=\frac{1}{C}\sum_{c=1}^{C}x_c(n-\delta_c) \qquad (۷)$$

که در آن C تعداد کانال‌ها، x_c معرف سیگنال در کانال cام و $\hat{s}(n)$ سیگنال بهسازی شده است. بلوک دیاگرم یک Delay-and-Sum Beam former در شکل (۲) نمایش داده شده است.

این عمل سبب می‌شود که سیگنال‌های ناخواسته که از سایر مسیرها دریافت می‌شوند تا حدودی حذف شوند. کیفیت سیگنال نهایی بهتر از (در بدترین حالت برابر با) سیگنال دریافتی هریک از میکروفون‌هاست.

برای محاسبه تأخیرهای زمانی، همبستگی متقابل بین هر کانال با کانال مرجع توسط رابطه زیر تخمین زده می‌شود:

$$\hat{r}_{x_1x_2}(m)=\sum_{n=0}^{N-1}x_1(n)x_2(n-m) \qquad (۸)$$

که در آن $x_1(n)$ و $x_2(n)$ سیگنال‌های دریافتی از دو کانال بوده و N تعداد نمونه‌های $x_1(n)$ است. تأخیر بین دو کانال توسط رابطه زیر محاسبه می‌شود:

$$\delta_{x_1x_2}=\arg\max_m \hat{r}_{x_1x_2}(m) \qquad (۹)$$

که در آن $\delta_{x_1x_2}$ تأخیر بین دو کانال می‌باشد. این روش را Delay-and-Sum Beam forming بدون استفاده از تابع وزن‌دهی (No Weighted) و یا باختصار DS-NW می‌نامیم.

با استفاده از این مطلب که تابع همبستگی متقابل تبدیل فوریه معکوس چگالی طیف توان متقابل یا CSD (Cross Spectral Density) است، و با وارد کردن یک تابع وزن‌دهی، می‌توان رابطه زیر را برای محاسبه تابع همبستگی متقابل ارائه داد:

$$r_{x_1x_2}(m)=\frac{1}{2\pi}\int_{-\pi}^{\pi}W(\omega)G_{x_1x_2}(\omega)e^{j2\pi\omega m}d\omega \qquad (۱۰)$$

که در آن، $W(\omega)$ تابع وزن‌دهی و $G_{x_1x_2}(\omega)$، CSD است. در این تحقیق تابع وزن‌دهی را به صورت زیر در نظر گرفته و روش حاصل را Delay-and-Sum Beam forming با استفاده از تابع وزن‌دهی فاز (Phase Weighted) یا به اختصار DS-PHW می‌نامیم:

$$W(\omega)=\frac{1}{\left|G_{x_1x_2}(\omega)\right|} \qquad (۱۱)$$

۴-۲- بهبود قابلیت فهم گفتار با استفاده از LP Residual

این روش که توسط Yegnarayana و همکاران وی ارائه شده [۱۲]، بر پایه پردازش بر روی سیگنال LP-Residual می‌باشد. در این تکنیک، پس از اعمال یک فیلتر معکوس، ضرایب gross از روی سیگنال LP-Residual و ضرایب fine از روی مقدار LP-Error

استخراج می‌گردد. ضرایب وزن‌دهی نهایی از حاصلضرب ضرایب gross و fine به دست آمده و بر روی سیگنال اعمال می‌شود. در مقاله حاضر، از این روش با نام «LP-Res» یاد می‌شود.

نواحی شامل سطوح gross با استفاده از خواص آماری سیگنال LP-Residual به دست می‌آید. در سیگنال LP-Residual، آنتروپی توزیع نمونه‌ها در نواحی با SRR بالا، در مقایسه با آنتروپی نواحی با SRR پایین، کم است. دلیل این مسأله آن است که نمونه‌ها در نواحی با SRR پایین، تقریباً توزیع گوسی شکل دارند و بنابراین آنتروپی در این نواحی زیاد است. برای محاسبه آنتروپی، تابع چگالی احتمال از روی نمونه‌ها در هر قطعه 20ms از LP-Residualتخمین زده می‌شود. آنتروپی H_k به صورت زیر تعریف می‌گردد:

$$H_k = -\sum_{i=1}^{M} p_i \log(p_i) \qquad (۱۲)$$

که در آن، p_i، احتمال تخمینی binِ iام هیستوگرام و M تعداد binها در هیستوگرام است. در نظر گرفتن تعداد bin در حدود (۵-۲۰)، این اطمینان را می‌دهد که به تعداد کافی نمونه از LP-Residual وجود دارد. معمولاً M=7 انتخاب می‌شود.

جهت استخراج تابع وزن‌دهی fine، تابع خطای نرمالیزه (η) برای هر نمونه و در فریمهای 2ms و با آنالیز LP مرتبه ۵ محاسبه می‌گردد. تابع وزن‌دهی fine عبارتست از:

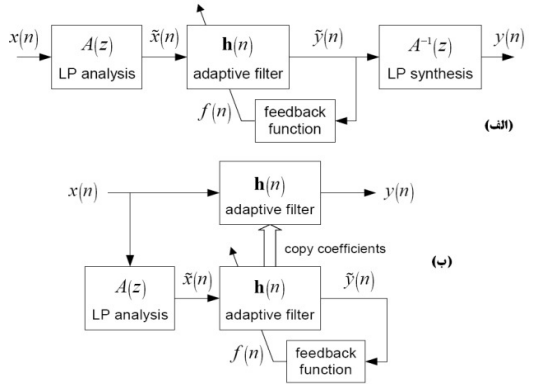
$$\omega_n^{fine} = \left(\frac{\omega_{max}^{fine} - \omega_{min}^{fine}}{2}\right)\tanh\left(\alpha_f \pi \eta_n\right) + \left(\frac{\omega_{max}^{fine} + \omega_{min}^{fine}}{2}\right)$$

$$\left\{\begin{array}{ll} \omega_n^{fine} & \text{weight value at the sampling instant n;} \\ \omega_{max}^{fine}(=1) & \text{maximum weight value;} \\ \omega_{min}^{fine} & \text{minimum wight value;} \\ \alpha_f(=1.5) & \text{positive constant which decides the slope of the weight function;} \\ \eta_n & \text{deternded normalized error value at the sampling instant n.} \end{array}\right\} \qquad (۱۳)$$

تابع وزن‌دهی نهایی از ضرب دو تابع وزن‌دهی gross و fine به دست می‌آید. سپس تابع وزن‌دهی نهایی و سیگنال LP-Residual در یکدیگر ضرب می‌شوند تا سیگنال LP-Residual اصلاح‌شده به دست آید. سیگنال LP-Residual اصلاح‌شده به عنوان سیگنال تحریک یک فیلتر تمام قطب مرتبه ۵، جهت بازسازی سیگنال به کار می‌رود.

۴-۳- بهبود قابلیت فهم گفتار با استفاده از فیلتر معکوس

بررسی‌های انجام‌شده [۴] حاکی از آن است که افزایش میزان انعکاس در سیگنال دریافتی، کاهش مقدار kurtosis سیگنال LP-Residual در پی خواهد داشت. با توجه به این نکته، ایده اصلی موجود در روش مورد بحث این قسمت، استفاده از فیلترهای وفقی است که نه بر مبنای معیار MMSE (Minimum Mean Square Error)، بلکه بر اساس ماکزیمم‌سازی آمارگان مرتبه ۴ (kurtosis) سیگنال LP-Residual کار می‌کند. بلوک دیاگرام حالت پایه این روش در شکل (۳) نشان داده شده است.



شکل (۳): بلوک دیاگرام سیستم فیلتر معکوس. (الف) دیاگرام سیستم تک کاناله وفقی در حوزه زمان برای ماکزیمم کردن kurtosis سیگنال LP-Residual. (ب) سیستم بهبود یافته.

Fig. (3): Block diagram of Inverse filter. (a) Diagram of adaptive single channel in time domain in order to maximizing kurtosis of LP-Residual Signal. (b) Enhanced system

از این روش -که توسط Gillespie و همکاران در مرجع [۱۳] ارائه شده- با نام تکنیک IF (Inverse Filtering) یاد می‌کنیم.

لازم به ذکر است تکنیک ارائه شده توسط Gillespie در حوزه زیرباند عمل نموده و قابلیت پیاده‌سازی به صورت هم تک‌کاناله و هم چندکاناله را دارا می‌باشد.

ضرایب LP در حالت کلی، شامل اطلاعات زیادی در مورد لوله صوتی انسان (vocal tract) می‌باشد؛ در حالی که سیگنال LP-Residual اطلاعات مربوط به انعکاس را در خود دارد. با استفاده از سیگنال LP-Residual می‌توان تا حد زیادی، همبستگی طولانی‌مدت (long-term correlation) را کاهش داد. همبستگی (طولانی‌مدت) موجود در سیگنال انعکاسی به نحو قابل توجهی به پاسخ ضربه آکوستیکی بین منبع و گیرنده بستگی دارد. از این رو با اعمال یک فیلتر معکوس می‌توان همبستگی را تا حد زیادی کاهش داد.

۴-۴- بهبود قابلیت فهم سیگنال گفتار با استفاده از ترکیب فیلتر معکوس و تفریق طیفی

روش دیگری که در مرجع [۶] معرفی شده و در اینجا نام اختصاری IF+SS (Inverse Filtering+Spectral Subtraction) را برای آن بکار می‌بریم، شامل دو مرحله است: در مرحله اول با اعمال یک فیلتر معکوس اثر coloration را کاهش می‌دهیم و در مرحله دوم، با استفاده ار تفریق طیفی، به کاهش اثر long-term reverberation می‌پردازیم.

۴-۵- بهبود قابلیت فهم گفتار با استفاده از فاز سیگنال دو میکروفون

این الگوریتم از الگوریتم‌های نوین در زمینه بهسازی سیگنال گفتار آغشته به انعکاس به شمار می‌رود و توسط اعرابی و همکاران[۵] ارائه شده است. این روش را می‌توان نوع خاصی از Delay-and-Sum

Beam forming وزن‌یافته با خطای فاز سیگنال‌های میکروفون‌ها در نظر گرفت. سیگنال دریافتی از دو کانال به صورت زیر است:

$$\begin{cases} x_1(t) = s(t) * h_{s1}(t) + n_1(t) \\ x_2(t) = s(t) * h_{s2}(t) + n_2(t) \end{cases} \quad (۱۴)$$

در حالت خاص می‌توان گفت:

$$\begin{cases} x_1(t) = h_1 \cdot s(t) + n_1(t) \\ x_2(t) = h_2 \cdot s(t-\tau) + n_2(t) \end{cases} \quad (۱۵)$$

در این روش سعی بر تخمین τ است. چنانچه $X_1(\omega)$ و $X_2(\omega)$ تبدیل‌های فوریه متناظر باشند، تخمین τ عبارتست از:

$$\tilde{\tau} = \arg\max_{\beta} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) \overline{X_2(\omega)} e^{-i\omega\beta} d\omega \quad (۱۶)$$

برای اطمینان از ایستان بودن سیگنال از قطعه‌های 10ms-20ms استفاده می‌کنیم. خواهیم داشت:

$$\tilde{\tau} = \arg\max_{\beta} \sum_{k=1}^{N} \int_{-\infty}^{\infty} W(k,\omega) X_{1,k}(\omega) \overline{X_{2,k}(\omega)} e^{-i\omega\beta} d\omega \quad (۱۷)$$

با توجه به اینکه با فرکانس‌های گسسته سروکار داریم، رابطه (۱۷) به صورت زیر بازنویسی می‌شود:

$$\tilde{\tau} = \arg\max_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} W(k,\omega) X_{1,k}(\omega) \overline{X_{2,k}(\omega)} e^{-i\omega\beta} \quad (۱۸)$$

تابع وزن‌دهی W بر اساس معیارهای مختلفی اتخاذ می‌شود که در ادامه به بعضی از آن‌ها اشاره می‌شود:

- Maximum Likelihood:

$$W_{ML}(k,\omega) = \frac{|X_{1,k}(\omega)||X_{2,k}(\omega)|}{|N_{2,k}(\omega) \cdot X_{1,k}(\omega)|^2 + |N_{1,k}(\omega) \cdot X_{2,k}(\omega)|^2} \quad (۱۹)$$

- Phase Transform (PHAT):

$$W_{PHAT}(k,\omega) = \frac{1}{|X_{1,k}(\omega) \cdot \overline{X_{2,k}(\omega)}|} \quad (۲۰)$$

- Un-filtered Cross Correlation:

$$W_{UCC}(k,\omega) = 1 \quad (۲۱)$$

با استفاده از تابع رابطه۲۰ و جایگذاری آن در رابطه (۱۸)، برای تخمین PHAT خواهیم داشت:

$$\tilde{\tau}_{PHAT} = \arg\max_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} \cos\left(\theta_{\beta,k}(\omega)\right) \quad (۲۲)$$

که در آن:

$$\theta_{\beta,k}(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) - \omega\beta \quad (۲۳)$$

در رابطه (۲۲) مقداربیشینه زمانی به دست می‌آید که β برابر τ باشد. می‌توان نشان داد که رابطه (۲۲) قابل بازنویسی به شکل زیر است:

$$\tilde{\tau}_{PHAT} = \arg\min_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} \theta_{\beta,k}^2(\omega) \quad (۲۴)$$

با توجه به رابطه فوق، MPV (Minimum Phase Variance) به صورت زیر تعریف می‌شود:

$$\psi_\tau = \min_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} \theta_{\beta,k}^2(\omega) \quad (۲۵)$$

در حالت ایده‌آل (هنگامی که τ صحیح تخمین زده شده باشد)، MPV، می‌بایست برابر صفر باشد. ولی به علت حضور نویز، انعکاس و پنجره‌کردن سیگنال، MPV در عمل برابر صفر نخواهد بود. هر چه انعکاس افزایش پیدا می‌کند، MPV نیز افزایش پیدا می‌کند.

در روش ارائه شده توسط Aarabi، نخست خطای فاز از روی دو فاز تبدیل فوریه سیگنال‌های دریافتی محاسبه می‌گردد. سپس برای هر زوج سیگنال یک تابع ماسک استخراج می‌شود. تابع ماسک عبارت است از:

$$M(\omega) = \frac{1}{1 + \gamma\theta_{\beta,k}^2(\omega)} \quad (۲۶)$$

که γ یک ثابت است.

برای حالت کلی سیستم چندمیکروفونه رابطه (۲۶) برای هر زوج میکروفون (i,j) به رابطه زیر تبدیل می‌شود:

$$M_{ij}(\omega) = \frac{1}{1 + \gamma\theta_{ij}^2(\omega)} \quad (۲۷)$$

در ادامه، میانگین هندسی توابع ماسک مرتبط با هر کانال (میکروفون) از رابطه زیر به دست می‌آید:

$$\Phi_i(\omega) = \left(\prod_{j=1, j\neq i}^{C} M_{ij}(\omega) \right)^{\frac{1}{k}} \quad (۲۸)$$

که در آن C تعداد کانال‌ها و k تعداد میکروفون‌ها می‌باشد. در نهایت تبدیل فوریه سیگنال بهسازی شده از رابطه زیر به دست آمده و از آنجا سیگنال بهسازی شده مشخص می‌شود:

$$\hat{S}(\omega) = \sum_{i=1}^{C} \Phi_i(\omega) X_i(\omega) \quad (۲۹)$$

در این مقاله، از این تکنیک با نام روش PEB (Phase Error Based) یاد می‌شود.

۵- ارزیابی قابلیت فهم سیگنال گفتار

معیارهای سنجش کیفیت (Quality) سیگنال گفتار، به ارزیابی «چگونگی» سیگنال می‌پردازد که بنوبه خود، به میزان عوامل مخرب سیگنال بستگی پیدا می‌کند. اما مسأله سنجش قابلیت فهم (Intelligibility) سیگنال گفتار، به ارزیابی «چیستی» سیگنال می‌پردازد که به تعداد قطعات گفتار که صحیح تشخیص داده می‌شوند بستگی پیدا می‌کند. مفهوم بودن لغات و یا اطلاعات جانبی دریافتی، از جمله مسائل تحت پوشش مسأله قابلیت فهم گفتار است.

آزمون‌ها و معیارهای عمومی سنجش کیفیت گفتار را می‌توان در دو دسته عمده قرار داد:

۱) معیارهای کمّی (Objective)، که بر پایه پارامترهای فیزیکی سیگنال یا کانال انتقال می‌باشد.

۲) آزمون‌های کیفی (Subjective)، که بر اساس نظرات شنوندگان است.

معیارهای کمّی نوعاً قادر به سنجش میزان قابلیت فهم سیگنال گفتار نبوده و تنها یک‌سری پارامترهای فیزیکی محیط یا سیگنال را ارزیابی می‌نمایند. اگرچه در برخی موارد، با یک مدل از پیش مشخص شده امکان پیش‌بینی قابلیت فهم گفتار از روی این معیارهای کمّی وجود دارد، لیکن ارزیابی یک سیستم بهسازی با چنین پارامترهایی با مشکلات و محدودیت زیادی روبروست.

آزمون‌های کیفی بر پایه نظر تعدادی شنونده سیگنال گفتار است که با شنیدن سیگنال مورد سنجش، کیفیت و یا قابلیت فهم آن را ارزیابی می‌کنند. از آنجائی که مسأله قابلیت فهم گفتار به نوعی زیرمجموعه‌ای از کیفیت گفتار به حساب می‌آید، ارزیابی قابلیت فهم سیگنال گفتار هم به نوبه خود ارزیابی کیفیت سیگنال گفتار محسوب می‌شود.

۵-۱- آزمون‌های کیفی سنجش قابلیت فهم گفتار

آزمون‌های کیفی سنجش قابلیت فهم سیگنال گفتار عموماً بر پایه سه ساختار همخوان‌ها، کلمات (با معنی یا بی‌معنی) و جملات طراحی می‌شود. در بسیاری از موارد کار با لغات معنی‌دار (مانند: کلمات، اعداد و ...) ساده‌تر و منطقی‌تر از کار با لغات بی‌معنایی است که از ترکیب تصادفی /CVC/ (و یا به عبارت دیگر، /Consonant-Vowel-Consonant/) با یکدیگر به وجود می‌آیند.

روش پاسخگوئی می‌تواند به صورت باز و یا بسته باشد. در روش باز، شنونده شخصاً در مورد آنچه که گمان می‌کند شنیده است اظهار نظر می‌کند. در روش بسته سعی می‌شود شنونده کلمه شنیده شده را از بین یک سری گزینه خاص انتخاب نماید. مزیت روش بسته نسبت به روش باز در این است که آموزش شنوندگان سریع‌تر و ساده‌تر بوده و به شنوندگان خبره نیاز چندانی نیست.

۵-۲- آزمون‌های Rhyme

در بسیاری از آزمون‌های ارزیابی قابلیت فهم گفتار از فهرست لغات Rhyme استفاده می‌کنند. در این آزمونها کلمات فهرست Rhyme به شنوندگان ارائه می‌شود که این کلمات فقط در واجهای آغازین باهم تفاوت دارند.

در سال ۱۹۵۸، Fairbanks با استفاده از کلمات با ساختار تک‌هجایی از میان لغات Rhyme، به سنجش سیگنال گفتار پرداخت [۹]. در این روش یک شنونده، یک برگه پاسخگوئی داشت که بر روی آن باید حروف ابتدائی سیگنال‌های پخش شده را تشخیص می‌داد. در سال ۱۹۶۵، House و همکاران روش Fairbanks را تغییر دادند. در این روش، شنوندگان باید کلمه شنیده شده را از میان تعداد محدودی (۶ لغت) از لغات Rhyme تشخیص می‌دادند. در نهایت، Voiers در سال ۱۹۷۷، توانست با استفاده از فهرست Rhyme، آزمون DRT را ارائه دهد [۷]. دادگان اصلی DRT شامل ۹۶ جفت لغت است. در این روش، لغات به دسته های دوتائی محدود می‌شوند. هر زوج از کلمات به سنجش ویژگی خاصی از سیگنال (شامل: voicing، nasality، sustention، sibilation، graveness، و compactness) می‌پردازد.

در هر نوبت اجرای آزمون، یکی از دو کلمه برای شنونده پخش می‌شود و شنونده باید تشخیص بدهد کدام کلمه را شنیده است. در نهایت نمره آزمون DRT از طریق رابطه زیر به دست می‌آید:

$$\mathrm{DRT}\% = \frac{\mathrm{N_{correct}} - \mathrm{N_{incorrect}}}{\mathrm{N_{tests}}} \times 100 \qquad (۳۰)$$

که در آن $\mathrm{N_{tests}}$، تعداد زوج‌های مورد آزمایش قرار گرفته، $\mathrm{N_{correct}}$، تعداد پاسخ‌های صحیح و $\mathrm{N_{incorrect}}$، تعداد پاسخ‌های نادرست می‌باشد.

آزمون MRT (Modified Rhyme Test)، به نوعی گسترش‌یافته آزمون DRT است که به سنجش قابلیت فهم همخوان‌های ابتدائی و انتهائی می‌پردازد. این آزمون دارای ۵۰ مجموعه است که هر مجموعه دارای ۶ کلمه تک هجایی است که در نهایت یک مجموعه ۳۰۰ کلمه‌ای را ارائه می‌دهد. مجموعه ۶ کلمه‌ای، یک بار پخش می‌شود و شنونده کلماتی را که گمان می‌کند شنیده است را به ترتیب از بین گزینه‌های موجود در برگه پاسخگوئی انتخاب می‌کند. نیمه ابتدائی جهت سنجش همخوان اول و نیمه دوم برای سنجش همخوان دیگر است.

۵-۳- مجموعه کلمات آزمون DRT برای زبان فارسی

جدول (۱)، مجموعه زوج کلمات آزمون DRT در زبان فارسی را معرفی می‌کند. این جدول،که از نوآوریهای این تحقیق به شمار می‌رود – مشابه با نمونه انگلیسی آن از ۹۶ جفت کلمه تشکیل شده که هر جفت، یکی از ۶ ویژگی مطرح در بخش (۵-۲) را مورد مقایسه قرار می‌دهد.

برای تهیه این مجموعه کلمات فارسی، با مقایسه کلمات جدول موجود در زبان انگلیسی و جایگزینی همخوان‌های مشابه در زبان فارسی، زوج کلمات مناسب با استفاده از فرهنگ لغات فارسی استخراج و انتخاب شده است. عمده کلمات استفاده شده در این دادگان نیز همانند مشابه انگلیسی آن تک‌هجایی و دارای ساختار /CVC/ هستند. ولی بنا به ساختار متفاوت زبان فارسی با زبان انگلیسی، در بعضی از کلمات این محدودیت لحاظ نشده است. مانند /چلتوک/ و شلتوک/، و یا در جایی دیگر /بو/ و /مو/. (برای تشریح بیشتر مراحل تهیه مجموعه کلمات DRT برای زبان فارسی به مرجع [۳] مراجعه شود.)

۵-۴- تهیه دادگان گفتار فارسی آغشته به انعکاس

از آنجائی‌که هدف اصلی ما بهسازی سیگنال گفتار فارسی آغشته به انعکاس است می‌بایست روش‌های موجود و پیشنهادی را با دادگان فارسی مورد سنجش قرار دهیم. با توجه به عدم وجود دادگانی از سیگنال گفتار فارسی که به دلیل انعکاس قابلیت فهم خود را از دست داده باشد، دو راه برای ما باقی می‌ماند:

۱. استفاده از نرم‌افزارهای شبیه‌ساز از قبیل: Cool Edit Pro و یا Adobe Audition.

۲. تهیه و ضبط سیگنال گفتار در شرائط حضور انعکاس شدید.

هرچند که راه اول ساده‌تر به نظر می‌رسد ولی این نرم‌افزارها به طور مصنوعی سیگنال را تخریب می‌کنند و در نتیجه، سیگنال حاصل با

سیگنال ضبط شده در شرایط واقعی -که در آن وجود نویز محیط کاملاً طبیعی و قطعی است- تفاوت بسیار دارد.

ضبط سیگنال در شرایط طبیعی، خود به تنهایی موضوعی است که ملاحظات زیادی را می‌طلبد و بررسی تمامی جوانب آن، می‌تواند خود مبنای یک کار تحقیقاتی مستقل باشد. با این وجود در این پژوهش، ما این راه را به راه حل نخست برتری داده و با در نظر گرفتن یک سری شرایط به ضبط داده‌های گفتاری اقدام نمودیم. این دادگان- که از دستاوردهای تحقیق حاضر است- می‌تواند در سایر پژوهش‌های مرتبط با سیگنال دومیکروفونه گفتار فارسی مورد استفاده واقع شود.

۵-۴-۱- ضبط سیگنال گفتار در شرایط حضور انعکاس

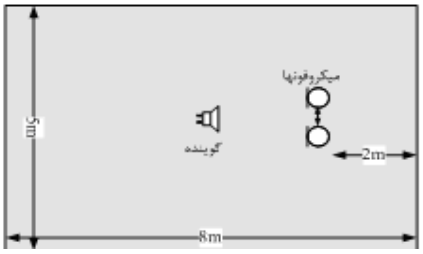
در طرح کلی پژوهش حاضر، ارزیابی عملکرد روش‌ها از لحاظ میزان بهبود قابلیت فهم و نیز بهبود کیفیت گفتار مد نظر بوده است. از این‌رو، به ترتیب آزمونهای کیفی DRT و Mean Opinion Score: MOS برای نیل به مقصود مورد استفاده قرار گرفته است.

برای امکان انجام آزمون DRT لیست کلمات فارسی این آزمون (به شرح آمده در بخش (۵-۳)) در شرائط محیطی با انعکاس، توسط گویندگان قرائت و از طریق سیستم دو میکروفونه ضبط گردید.

به منظور تکمیل دادگان گفتار فارسی آغشته به انعکاس و نیز برای فراهم‌سازی امکان سنجش کیفیت سیگنال، علاوه بر کلمات جدول DRT، یک سری جمله نیز طراحی و در هنگام تهیه دادگان توسط سیستم دومیکروفونه ضبط گردید. این جملات در آزمون MOS برای سنجش کیفیت گفتار مورد استفاده قرار گرفت.

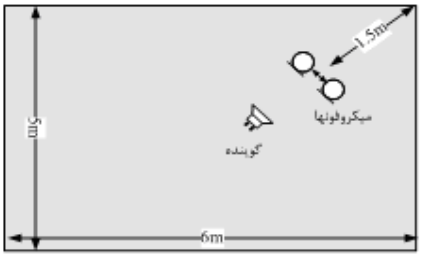
همان‌گونه که قبلاً نیز عنوان شد، گزینه مناسب برای فاصله بین میکروفون‌ها، 7-8 cm است. در این تحقیق، از میکروفون‌های همه‌جهته استفاده شده است. برای انتخاب فاصله گوینده از محور یا خط واصل بین میکروفون‌ها، پس از انجام یک سری آزمایش اولیه بر روی فواصل 50cm، 100cm، 150cm، 200cm، 300cm، 400cm و 450cm، به این نتیجه رسیدیم که برای فواصل 50cm و 100cm قابلیت فهم سیگنال گفتار خوب بوده و انجام آزمایش و مقایسه معنی‌دار بر روی آن میسر نمی‌باشد. با توجه به این ملاحظات، از میان فواصل ذکر شده، در تهیه دادگان تنها به ضبط نمونه‌ها در فواصل 150cm، 300cm و 450cm اکتفا کردیم. برای پوشش بهتر بر روی حالت‌های ناایستان طبیعی، چندین دسته نمونه نیز در وضعیت راه رفتن گوینده در فاصله بین 100cm تا 300cm از میکروفونها ضبط گردید.

ضبط داده‌ها در چند مرحله و در دو اتاق به ابعاد 8m×5m×3m و 6m×5m×3m صورت پذیرفت. کف‌پوش این دو اتاق موزائیک، دیوارها تا حد خوبی صیقلی، پنجره‌ها فاقد پرده پارچه‌ای و هر اتاق دارای حدود ۶۰ صندلی چوبی صیقلی بود. شکل‌های ۴ و ۵ نحوه قرار گرفتن گوینده و مجموعه میکروفونها را در هر یک از دو اتاق نشان می‌دهد. لازم به ذکر است از ۴ نفر (دو مرد و دو زن) به عنوان گوینده در تهیه دادگان استفاده شد. دادگان در فرکانس نمونه‌برداری 16kHz و با اختصاص ۱۶ بیت برای هر نمونه ضبط شده است. برای تشریح بیشتر مراحل تهیه دادگان دوکاناله گفتار فارسیِ دارای انعکاس به مرجع [۲] مراجعه شود.



شکل (۴): نحوه قرار گرفتن گوینده و میکروفونها ، اتاق اول

Fig. (4): Speaker and microphones arrangement in the first room



شکل (۵): نحوه قرار گرفتن گوینده و میکروفونها، اتاق دوم

Fig. (5): Speaker and microphones arrangement in the second room

۶- ارزیابی عملکرد روش‌های موجود و ارائه سیستم‌های ترکیبی برای حذف انعکاس

در این تحقیق، برای ارزیابی میزان بهبود قابلیت فهم گفتار، از آزمون DRT بهره گرفته شد. برای اجرای آزمون DRT، از هر جفت کلمات جدول DRT یکی به طور تصادفی برای شنونده پخش می‌گردد؛ شنونده باید حدس بزند که کدام کلمه را شنیده است. در پایان نمره آزمون DRT از روی نسبت تعداد کلمات درست تشخیص داده شده به دست می‌آید.

برای ارزیابی کیفیت سیگنال خروجی از آزمون MOS استفاده نمودیم. برای این منظور جملات (فارسی) برای شنونده پخش شده و از شنونده خواسته می‌شود که به کیفیت سیگنال شنیده شده امتیازی بین ۱ تا ۵ بدهد.

مجموعه نتایج حاصل از آزمون‌های DRT و MOS در جدول (۲) خلاصه شده است. از آنجا که تعداد کلمات جدول DRT زیاد بود، در عمل برای هر شنونده، یک سری از لغات به عنوان سیگنال نمونه انتخاب و در اجرای آزمون مورد استفاده قرار گرفت. سیگنال نمونه شامل ۲۴ کلمه بود که توسط ۳ گوینده (یک مرد و دو زن) و در فواصل 150cm، 300cm و 450cm از محور میکروفون‌ها قرائت شده بود.

جدول (۱): مجموعه جفت کلمات فارسی برای آزمون DRT

Voicing		Nasality		Sustention		Sibilation		Graveness		Compactness	
فهم	وهم	بین	مین	باک	واک	چی	کی	ری	وی	هک	فک
فاو	واو	دخیل	نخیل	چَرت	شَرط	جُرم	گُرم	پیشه	تیشه	روز	یوز
بیش	پیش	بار	مار	بخت	وقت	قران	گران	مالید	نالید	حربه	فربه
بیست	پیست	دور	نور	چرک	شرک	جاری	گاری	رهن	وهن	سُکر	شُکر
بند	پند	بطن	متن	بعید	وعید	جثه	غصه	مادی	نادی	تک	کک
جار	چار	دام	نام	چرم	شرم	جبر	گبر	باد	داد	در	گر
جین	چین	بور	مور	بال	وال	چاه	کاه	رحم	وهم	تل	کل
جیره	چیره	دشتی	نشتی	تیران	ثیران	جوش	گوش	پاس	تاس	سیر	شیر
دشت	تشت	بو	مو	چست	شست	چینه	کینه	بید	دید	پود	کود
بور	پور	دوش	نوش	بل	ول	جوهر	گوهر	رعد	وعد	تر	کر
دار	تار	بست	مست	چانه	شانه	خشت	کشت	پیپ	تیپ	سیب	شیب
دور	تور	دوان	نوان	بادی	وادی	قور	گور	میش	نیش	پوک	کوک
زال	سال	بال	مال	چلتوک	شلتوک	جن	ژن	رداء	وداع	تُرد	کُرد
زود	سود	دهر	نهر	بَلَد	وَلَد	جور	گور	پیر	تیر	صمد	شمد
گام	کام	بیوه	میوه	چال	شال	جیوه	ژیوه	بام	دام	پور	کور
گر	کر	دما	نما	بَنگ	وَنگ	زور	ژور	پیروز	دیروز	حنا	فنا

لازم به ذکر است در انجام آزمون‌های شنیداری DRT و MOS از ۷ نفر شنونده (که آشنایی قبلی با سیگنال‌ها نداشتند) استفاده گردیده و نتایج ارائه شده در جدول (۲) متوسط نظرات این شنوندگان می‌باشد. به عنوان مبنای مقایسه، نتایج حاصل از آزمون DRT و MOS بر روی سیگنال اولیه (سیگنال ورودی آغشته به انعکاس که هیچ پردازشی برای بهبود بر روی آن انجام نشده) در ردیف اول جدول آورده شده است. در ادامه، برای بررسی عملکرد روش‌های پایه حذف انعکاس، خروجی این روش‌ها مورد ارزیابی قرار گرفت که نتایج آن در ردیف‌های دوم تا هفتم جدول (۲) ارائه شده است. به وضوح پیداست که:

۱- الگوریتم‌های «IF+SS» و «LP-Res» در کاربرد مورد نظر نتایج مطلوبی ارائه نمی‌دهد. این الگوریتم‌ها بر طبق اظهارات مبتکران آن‌ها برای سیستم‌های بازشناسی گفتار پاسخ خوبی داشته است. این می‌تواند دلیلی بر این مدعا باشد که معیارهایی که ذهن انسان برای فهم گفتار در نظر می‌گیرد، با معیارهایی که مبنای کار سیستم بازشناسی گفتار است، برابر نیستند.

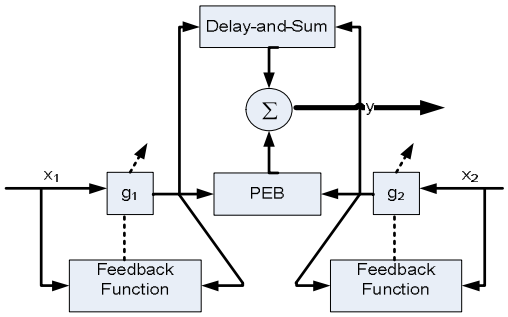
۲- نتایج حاصل از شبیه‌سازی الگوریتم‌های PEB، DS-NW و DS-PHW بیانگر بهبود هم در قابلیت فهم سیگنال و هم در کیفیت سیگنال، نسبت به سیگنال اصلی است.

پس از انجام این آزمایش‌های اولیه، در تحقیق حاضر، اقدام به طراحی یک سری روش‌های ترکیبی نموده و قابلیت فهم و کیفیت سیگنال‌های بهسازی شده را ارزیابی و مقایسه نموده‌ایم. در ارائه این روش‌ها، از فرض وجود دو میکروفون استفاده شده است. در ادامه، به معرفی قدم به قدم روش‌های ترکیبی پیاده‌سازی‌شده می‌پردازیم. ارائه و ارزیابی این سیستم‌های ترکیبی را می‌توان از دیگر دستاوردهای این پژوهش برشمرد.

همان‌گونه که قبلاً نیز عنوان شد، می‌توان پدیده long-term correlation موجود در سیگنال آغشته به انعکاس را با استفاده از یک فیلتر معکوس حذف نمود. اگرچه نتایج مربوط به الگوریتم IF چندان رضایتبخش نیست، ولی این فرضیه محتمل می‌باشد که اعمال روش IF به عنوان پیش‌پردازشی برای سایر تکنیکها می‌تواند منجر به بهبود نتایج حاصله گردد. برای این منظور، فیلتر معکوس را به سیگنال آغشته به انعکاس اعمال نموده و سپس الگوریتم‌های PEB، DS-NW و DS-PHW را به صورت جداگانه بر روی آن اعمال نمودیم. نتایج ارزیابی عملکرد این تکنیک‌های ترکیبی در بهبود کیفیت و قابلیت فهم گفتار در ردیفهای هشتم تا دهم جدول (۲) ارائه شده است. نتایج به دست آمده حاکی از آن است که این روش ترکیبی برای ترکیب فیلتر معکوس (IF) و Delay and Sum Beamforming (DS-NW و DS-PHW)، از دید آزمون DRT موفق‌تر از ترکیب فیلترمعکوس با روش استفاده از خطای فاز(PEB) بوده است. (هرچند که نتایج آزمون MOS چنین برتری را نشان نمی‌دهد!)

از آنجائی‌که الگوریتم PEB به تنهایی موفق‌تر از هنگامی بوده از فیلتر معکوس استفاده شده است، بر آن شدیم تا اثر ترکیب الگوریتم PEB و Delay and Sum Beamforming (DS-NW و DS-PHW) را نیز بررسی نماییم. نتایج حاصل از شبیه‌سازی این ترکیب در ردیف‌های یازدهم و دوازدهم جدول (۲) ملاحظه می‌شود. نتایج مقایسه‌ای الگوریتم‌های PEB، DS-NW، DS-PHW، PEB+DS-NW و

PEB+DS-PHW با نتایج حاصل از مقایسه کارآیی همین روش‌ها برای سیستم بازشناسی خودکار گفتار -که توسط Aarabi و همکاران در مرجع [۵] ارائه شده است- همخوانی داردو بسیار به آنها نزدیک است؛ اما نتایج به دست آمده در مقایسه با نتایج روش‌های ترکیبی IF+PEB، IF+DS-NW وIF+DS-PHW چندان مطلوب به نظر نمی‌رسد.



شکل (۶): بلوک دیاگرام سیستم ترکیبی ((IF+(DS-NW+PEB)) یا ((IF+(DS-PHW+PEB)).

Fig. (6): Block diagram of hybrid system (IF+(DS-NW+PEB)) or (IF+(DS-PHW+PEB)).

الگوریتم‌های ترکیبی که تا به حال مورد بررسی قرار گرفت، به بهبود قابلیت فهم سیگنال گفتار آغشته به انعکاس و بالا بردن کیفیت آن کمک می‌کند، ولی از آنجائی که هرکدام از الگوریتم‌های PEB، DS-NW وDS-PHW به تنهایی قادر به بهسازی سیگنال گفتار هستند، در نهایت برآن شدیم تا از تمام ویژگی‌های مثبت موجود در این الگوریتم‌ها استفاده کرده و الگوریتم‌های ترکیبی IF+(DS-NW+PEB) و IF+(DS-PHW+PEB) را پیاده‌سازی و ارزیابی نماییم. شکل (۶) بلوک دیاگرام سیستم ترکیبی را در این حالت به نمایش گذارده است. نتایج حاصل از شبیه‌سازی‌ها نیز در ردیف‌های سیزدهم و چهاردهم جدول (۲) آورده شده است. نتایج بیانگر بهبود در قابلیت فهم و کیفیت سیگنال گفتار آغشته به انعکاس می‌باشد.

لازم به ذکر است که در تمامی الگوریتم‌های ترکیبی فوق در ابتدا یک فیلتر معکوس بر روی دو کانال اعمال گردیده و سپس پردازش (پردازش‌های) دیگر اعمال می‌شود (به جای جمع‌زدن دو سیگنال). همچنین به دلیل فاصله ثابت بین دو میکروفون، تأخیر ثابتی بین دو سیگنال دریافتی وجود دارد. از این‌رو، باید در ابتدا با استفاده از یک تکنیک تخمین تأخیر، سیگنال دریافتی از دو کانال را سنکرون کرد.

۷- جمع‌بندی و نتیجه‌گیری

در این تحقیق، موضوع بهسازی و بهبود قابلیت فهم گفتار آغشته به انعکاس مورد بررسی قرار گرفت. در ابتدا مفاهیم پایه‌ای بحث شامل تعریف انعکاس، نحوه اندازه‌گیری آن، و نیز اصول حاکم بر بحث تعداد و نحوه چیدمان میکروفونها مورد مطالعه قرار گرفت. در ادامه به بررسی چند روش پایه در زمینه حذف انعکاس پرداختیم. برای ارزیابی عملکرد این روش‌ها و نیز روش‌های ترکیبی پیشنهادی، مجموعه لغات فارسی برای آزمون DRT تهیه و سپس با استفاده از آن، دادگان دو کاناله گفتار دارای انعکاس طراحی و ضبط گردید.

روشن است که آزمون‌های DRT و MOS برای مقایسه بین الگوریتم‌های مختلف که نتایج آنها نزدیک به یکدیگر است و انتخاب الگوریتم برتر در این میان، مناسب نمی‌باشد؛ با در نظر گرفتن این مطلب و نتایج حاصل از جدول (۲)، به عنوان نتیجه کلی این تحقیق، بجای انتخاب الگوریتم بهینه، الگوریتم‌های برتر را معرفی می‌نماییم.

براین اساس، دو الگوریتم «IF+(DS-NW+PEB)» و «IF+(DS-PHW+PEB)» -که از نتایج بهتری نسبت به سایر الگوریتم‌ها برخوردارند- برای دستیابی به بهبود قابل توجه در میزان قابلیت فهم و کیفیت سیگنال بهسازی شده، توصیه می‌گردد.

جدول (۲): نتایج آزمون‌های DRT و MOS

Algorithm	DRT%	MOS
OS	76.6	3
IF	72.4	3.3
IF+SS	62.1	1.3
LP-Res	59.0	1.3
PEB	77.9	3.7
DS-NW	75.2	3.8
DS-PHW	81.6	3.5
IF+PEB	74.3	4.2
IF+DS-NW	80.7	3.9
IF+DS-PHW	80.2	3.9
PEB+DS-NW	79.3	3.4
PEB+DS-PHW	78.4	3.7
IF+(DS-NW+PEB)	79.3	4.5
IF+(DS-PHW+PEB)	81.6	4.6

سپاسگزاری

مراحل متعددی از این تحقیق با حمایت آزمایشگاه پردازش سیگنال دانشگاه یزد صورت گرفته است.

References

[1] I. Kodrasi, S. Doclo, "The effect of inverse filter length on the robustness of acoustic multi channel equalizagtion", EUSIPCO, pp. 2442-2446, Aug, 2012.

[2] X. Bao, j. Zhu, Z. Haung, "Blind speech dereverberation based on a statistical model", IEEE/ICME, pp. 467-472, 2012.

[3] H. Abutalebi, F. Faghani, "Proposing Farsi database for DRT test to evaluating Farsi speech Intelligibility", ICEE, May, 2006.

[4] F. Faghani, "Enhancement and intelligibility improvement of reverberated speech signals",Ms.C. Thesis, Yazd University, 2007.

[5] P. Aarabi, G. Shi, "Phase-based dual-microphone robust speech enhancement", IEEE Trans. on Sys., Man, and Cyb., Vol. 34, No. 4, Aug. 2004.

[6] H.R. Abutalebi, F. Faghani, "Dual microphone speech dereverberation", ISSPA, Jan. 2007.

[7] J.B. Allen, "Effects of small room reverberation on subjective preference", Jou. of Aco. Soc. of Ame., Vol. 71, 1982.

[8] D.A. Berkley, J.B. Allen, "Normal listening in typical rooms: The physical and psychophysical correlates of reverberation", In Acoustical factors affecting hearing aid performance, G. A. Studebaker and I. Hochberg, Eds., 2nd ed., Needham Heights, MA: Allyn and Bacon, pp. 3-14, 1993.

[9] J.R. Deller, J.G. Proakis, J.H.L. Hansen, "Discrete-time processing of speech signal", Second Edition, IEEE Press, 2000.

[10] J.L. Flanagan, J.D. Johnson, R. Zahn, G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", Jou. of Aco. Soc. of Am., Vol. 78, pp. 1508-1518, 1985.

[11] M. Ferras, "Multi microphone signal processing for automatic speech recording in meeting room", Ph.D. Thesis, Electrical Engineering Dept., Berkeley University, 2006.

[12] B. Yegnanarayana, C. Avendaño, H. Hermansky, P. Satyanarayana Murthy, "Processing linear prediction residual for speech enhancement", EUROSPEECH, Patras, Greece, pp. 1399–1402, Sep. 1997.

[13] B.W. Gillespie, H.S. Malvar, D.A.F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering", ICASSP, pp. 3701-3704, 2001.