

بهبود سیستم جداسازی منبع مبتنی بر آنالیز ترکیب شنیداری در زبان فارسی

مسعود گراوانچی زاده^۱، دانشیار، پریا دادور^۲، دانشجوی دکتری، بابک بهادرنیا^۳، کارشناسی ارشد

۱، ۲ و ۳- دانشکده مهندسی برق و کامپیوتر- دانشگاه تبریز- تبریز- ایران

Email: 1-geravanchizadeh@tabrizu.ac.ir, 2-pariadadvar@tabrizu.ac.ir, 3-b_bahadornia89@tabrizu.ac.ir

چکیده: در این مقاله، سیستم‌های جدیدی به منظور بهبود عملکرد سیستم جداکننده گفتار دوگوشی با نام MESSL ارائه می‌شود. در سیستم جداسازی سیگنال، ابتدا، با استفاده از الگوریتم EM، مدل‌های گوسی پارامترهای اختلاف فاز درون گوش (IPD) و اختلاف شدت درون گوش (ILD) به دست می‌آیند. سپس، با استفاده از مدل به دست آمده برای هر منبع، ماسک نرمی استخراج شده که با ضرب آن در تبدیل فوریه زمان کوتاه (STFT) سیگنال مخلوط، سیگنال هدف جدا می‌شود. به علت عملکرد ناقص سیستم در امر جداسازی، دو سیستم پس پردازش به منظور حذف سیگنال‌های ناخواسته از سیگنال هدف، پیشنهاد می‌شود. روش پیشنهادی اول حذف وافی نویز با استفاده از بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (LPSO) است. سیستم پس پردازش پیشنهادی دوم شامل دو مرحله است. در مرحله اول این سیستم، از روش حذف نویز تبدیل موجک به منظور حذف بخش اعظم سیگنال تداخل استفاده می‌شود. در مرحله دوم، روش حداقل میانگین مربعات خطا (MMSE) جهت ارتقاء هرچه بیشتر کیفیت سیگنال هدف جدا شده به کار می‌رود. ارزیابی و مقایسه سیستم‌های پیشنهادی برای دادگان فارسی نشان می‌دهد که سیستم پیشنهادی دوم در بهبود کیفیت سیگنال هدف جدا شده خوب عمل می‌کند و از نظر محاسباتی نیز کارآمد است. واژه‌های کلیدی: بهبود کیفیت گفتار، جداسازی منبع دوگوشی، تبدیل موجک، حداقل میانگین مربعات خطا (MMSE)، بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (LPSO)

Enhancement of CASA-Based Source Separation System in Farsi

M. Geravanchizadeh¹ Associate professor, P. Dadvar², PhD. Student, B. Bahadornia³, Msc.

Faculty of Electrical & Computer Engineering, University of Tabriz, Tabriz 5166615813, Iran

Email: 1-geravanchizadeh@tabrizu.ac.ir, 2-pariadadvar@tabrizu.ac.ir, 3-b_bahadornia89@tabrizu.ac.ir

Abstract: In this paper, new systems to enhance the performance of binaural source separation system, called MESSL, are proposed. In the source separation system, first, the Gaussian models for the interaural phase difference (IPD) and interaural level difference (ILD) parameters are obtained by using the EM algorithm. Then, by using the generated model for each source, a soft mask is extracted and multiplied with the short-time Fourier transform (STFT) of the mixture signal to separate the target signal. Because of incomplete performance of the separation system, two post-processing systems are proposed to remove the unwanted signals from the target signal. The first proposed method is the adaptive noise cancellation using learning-based particle swarm optimization (LPSO). The second proposed post-processing system includes two stages. In the first stage of this system, the denoising technique of the Wavelet transform is employed to remove the main part of the distracter signal. In the second step, the minimum mean-squares-error (MMSE) approach is used to enhance further the quality of the separated target signal. Evaluation and comparison of the proposed systems for Farsi database shows that the second proposed system performs well in the enhancement of the separated target speech and is also computationally efficient.

KeyWords: Speech enhancement, binaural source separation, wavelet transform, minimum mean-squares-error (MMSE), learning-based particle swarm optimization (LPSO)

تاریخ ارسال مقاله: ۱۳۹۳/۰۲/۱۶

تاریخ اصلاح مقاله: ۹۳/۵/۶ و ۹۴/۱/۲۵

تاریخ پذیرش مقاله: ۱۳۹۴/۶/۳

نام نویسنده مسئول: دکتر مسعود گراوانچی زاده

نشانی نویسنده مسئول: ایران- تبریز- بلوار ۲۹ بهمن- دانشگاه تبریز- دانشکده مهندسی برق و کامپیوتر

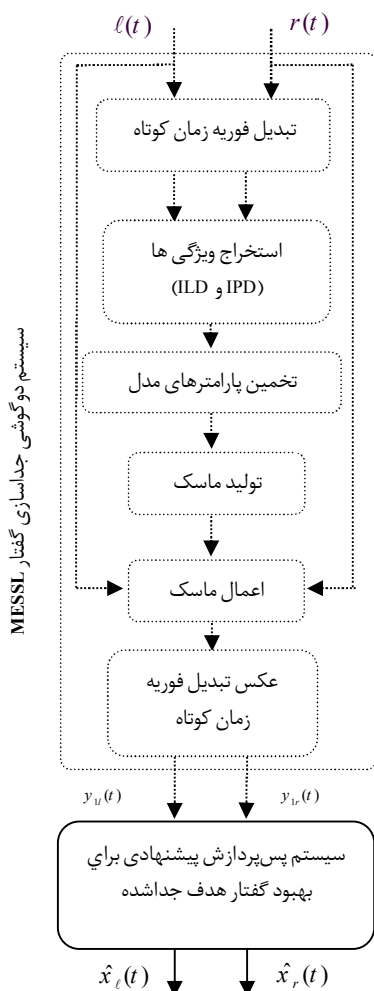
۱- مقدمه

یکی از مهم‌ترین چالش‌های پیش روی سیستم‌های پردازش صوت، حل مسئله مهمانی^۱ است. هدف اصلی در این چالش، ساخت پردازنده‌ای است که بتواند عمل بهبود کیفیت گفتار، جداسازی و مکان‌یابی منابع گفتار را در محیط‌های شلوغ به درستی انجام دهد. ما در این مقاله، جداسازی گفتار را مورد بحث قرار می‌دهیم. روش‌های متعددی برای حل مسئله جداسازی ارائه شده است که از آن جمله می‌توان به روش Beamforming اشاره کرد [۱]. ساده‌ترین نوع Beamformerها، delay-and-sum نام دارد. در این نوع سیستم‌ها، آرایه میکروفونی به صورت خطی قرار دارد و سیگنال رسیده از منبع گفتار، نسبت به زاویه آن با آرایه میکروفونی و فاصله بین میکروفون‌ها، تأخیر می‌یابد. با اعمال تأخیرهای مناسب، همه سیگنال‌های دریافتی در حوزه زمان هم‌تراز شده که این امر سبب حذف سیگنال‌های ناخواسته می‌شود [۲-۵]. از جمله روش‌های دیگر جداسازی منابع گفتار، روش‌های مبتنی بر جداسازی کور منابع (BSS) هستند. خود Beamformerها نیز زیرمجموعه‌ای از روش‌های BSS تلقی می‌شوند. واژه کور به این نکته اشاره می‌کند که نه خود سیگنال‌ها و نه نحوه ترکیب آن‌ها برای سیستم BSS مشخص نیست. در حالی که سیستم‌های مبتنی بر Beamforming روی ویژگی‌های مکانی منابع متمرکز هستند، روش آنالیز مؤلفه مستقل (ICA) که خود زیرمجموعه دیگری از روش‌های BSS است، از ویژگی آماری سیگنال‌ها برای امر جداسازی بهره می‌برد [۶-۷]. ICA برای اولین بار برای جداسازی ترکیب‌های آنی مورد استفاده قرار گرفت که در آن سیگنال مخلوط، ترکیبی بدون نویز و خطی از ورودی‌ها بود [۸]. اما از آنجایی که ترکیب‌های آکوستیکی در واقعیت شامل تأخیر و کانولوشن با پاسخ ضربه‌های فضایی هستند، به علت تبدیل کانولوشن به ضرب در حوزه فرکانس، ICA مستقیماً به باندهای فرکانسی اعمال شده است [۹]. یکی از مشکلات روش‌های مبتنی بر ICA، بحث ترتیب است، به این صورت که بایستی مشخص شود، ردیف‌های ماتریس جداکننده به کدام منبع تعلق دارد. علت اصلی این امر آن است که ردیف‌های ماتریس جداکننده حالت ثابتی ندارند و ترتیب آن‌ها از یک نقطه فرکانسی به نقطه‌ای دیگر تغییر می‌کند. برای حل این مسئله، روش‌های متعددی ارائه شده که از جمله آن‌ها می‌توان به روش مقایسه پوش سیگنال‌های جدا شده [۱۰-۱۲] و روش مبتنی بر معیار نسبت توان [۱۳] اشاره کرد. سیستم‌های جداساز کور مبتنی بر کانولوشن از هر دو ویژگی مکانی و آماری سیگنال بهره می‌برند. یکی از سیستم‌های جداکننده برای سیگنال‌های مخلوط کانولوشنی، TRINICON نام دارد که از هر سه خاصیت غیرسفید بودن، غیرایستادن بودن و غیرگوسی بودن برای جداسازی استفاده می‌کند [۱۴]. سه محدودیت عمده روش‌های ذکر شده در بالا عبارت‌اند از: ۱- وضعیت شخص گوینده بایستی در طول فرآیند جداسازی ثابت بماند. به عبارتی، از مواردی مانند چرخش سر که بیانگر شرایط غیرایستادن گوینده است، اجتناب شود. ۲- تعداد میکروفون‌ها بیش‌تر از

تعداد منابع گفتار یا حداقل برابر با آن‌ها باشد. ۳- بحث ترتیب که تأثیر به‌سزایی روی خروجی سیستم جداکننده گفتار مبتنی بر BSS دارد، منظور شود.

برای برطرف نمودن محدودیت‌های ذکر شده در بالا، می‌توان از قابلیت‌های سیستم شنوایی انسان در امر جداسازی بهره برد. توانایی انسان در تمرکز روی صدای یک فرد خاص در محیط‌های شلوغ با در نظر گرفتن شرایط غیرایستادن گوینده، آن هم تنها با دو گوش، عامل اصلی در تلاش برای شبیه‌سازی این توانایی برای ماشین در بحث پردازش صوت است. شبیه‌سازی سیستم شنوایی انسان در شکل محاسباتی آن، آنالیز محاسباتی ترکیب شنیداری (CASA)^۲ نام دارد. هدف در CASA، طراحی پردازنده‌ای است که تا حد امکان بتواند عملکرد گوش انسان را شبیه‌سازی کند [۱۵].

در [۱۶] روش دوگوشی مکان‌یابی و جداسازی منابع مبتنی بر مدل با استفاده از پیشینه‌سازی انتظار (MESSL)^۳ ارائه شده است که به جداسازی و مکان‌یابی چندین منبع صدا از یک سیگنال ترکیب دو کانالی می‌پردازد. در این روش، ابتدا، پارامترهای درون‌گوشی سیگنال گفتار استخراج می‌شوند. این پارامترها عبارت‌اند از: اختلاف فاز درون‌گوشی (IPD)^۴ و اختلاف شدت درون‌گوشی (ILD)^۵. سپس، با در دست داشتن پارامترهای درون‌گوشی، مدل گوسی آن‌ها محاسبه می‌گردد. در مرحله بعد، به کمک مدل گوسی به دست آمده و با بررسی تعلق هر واحد زمان-فرکانسی (T-F) به هر یک از منابع، ماسک مربوط به هر منبع صوتی استخراج می‌شود. در نهایت، با اعمال ماسک استخراج شده به سیگنال مخلوط، منابع گفتار از یکدیگر جدا می‌شوند. عمل جداسازی به‌طور کامل در سیستم مورد بحث اتفاق نمی‌افتد و اندکی از گفتار منابع غیرهدف نیز در سیگنال جدا شده باقی می‌ماند. در [۱۷] یک سیستم پس‌پردازش دومرحله‌ای با استفاده از روش حذف نویز موجک^۱ و روش حداقل میانگین مربعات خطا (MMSE) برای بهبود کیفیت سیگنال‌های جدا شده توسط روش MESSL ارائه شده است. بهبود کیفیت گزارش شده از ارزیابی‌های انجام شده برای دادگان TIMIT در این مرجع رضایت‌بخش است. هرچند روش‌های متعددی، از جمله روش تفريق طیفی [۱۸] و یا روش فیلتر وینر [۱۹] و نیز حذف وفقی نویز^۷ با استفاده از الگوریتم‌های بهینه‌سازی اتفاقی مانند بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (LPSO)^۸ [۲۰]، برای بهبود کیفیت گفتار ارائه شده، ولی از میان این روش‌ها حذف نویز بر اساس تبدیل موجک عملکرد بهتری در بهبود سیگنال‌های جدا شده از خود نشان می‌دهد [۲۱]. در این مقاله، دو سیستم بهبود کیفیت گفتار به‌منظور حذف سیگنال‌های ناخواسته از سیگنال هدف جدا شده توسط سیستم دوگوشی MESSL پیشنهاد و ارزیابی شده است. یکی از این سیستم‌ها روش بهبود گفتار دو کانالی ارائه شده در مرجع [۲۰] است که در این مقاله به‌صورت سیستم ANC_LPSO نام‌گذاری شده است، به حذف وفقی نویز با استفاده از بهینه‌سازی ازدحام ذرات مبتنی بر یادگیری می‌پردازد. سیستم بهبود کیفیت گفتار ارائه شده دوم بنام



شکل ۱: بلوک‌دیگرام روش پیشنهادی جهت ارتقای عملکرد سیستم دوگوشی جداسازی گفتار MESSL. سیستم جداسازی دوگوشی گفتار MESSL با نقطه چین و سیستم پس پردازش پیشنهادی با خط پررنگ مشخص شده است.

در روابط بالا، $\mathcal{F}$ بیانگر عملگر تبدیل فوریه، $\alpha(\omega)$ اختلاف شدت درون گوش و وابسته به فرکانس، τ_r ، τ_l ، تأخیرهای مستقل از فرکانس گوش‌های چپ و راست و $\tau(\omega)$ اختلاف زمانی درون گوش و وابسته به فرکانس هستند. روابط (۲) و (۳)، به ترتیب، به عنوان مدل واقعی و مدل تقریبی مشاهدات طیف‌نگاره درون گوش محسوب می‌شوند. طیف‌نگاره درون گوش با دو پارامتر IPD $(\phi(\omega, t))$ و ILD $(\alpha(\omega, t))$ برحسب dB نمایش داده می‌شود.

برای رفع ابهام از فاز، از یک فرآیند بالا به پایین استفاده می‌شود، به این صورت که به جای نگاشت اختلاف فاز درون گوش (IPD) به اختلاف زمان درون گوش (ITD)، مقادیر نماینده (ITD) (یا τ) به فضای IPD نگاشت می‌شوند تا τ ای را که با مشاهدات بیش‌ترین تطابق را دارد، پیدا شود. بنابراین، به جای مدل کردن فاز واقعی، فاز باقی‌مانده $\hat{\phi}(\tau)$ ، یعنی اختلاف بین فاز واقعی و فاز تخمین زده شده توسط τ ، به صورت زیر مدل می‌شود:

WT_MMSE شامل دو مرحله است. ابتدا، با استفاده از روش حذف نویز مویک بر روی سیگنال هدف جدا شده، سیگنال‌های ناخواسته باقی‌مانده از مرحله جداسازی دوگوشی حذف می‌شود. در مرحله بعد، برای افزایش کیفیت سیگنال جدا شده هدف، از روش حداقل میانگین مربعات خطا (MMSE) [۲۲] استفاده می‌شود. سیستم‌های پس‌پردازش ارائه شده برای دادگان فارسی برحسب معیارهای مختلف مورد ارزیابی جامع قرار گرفته و عملکرد آن‌ها برای حالت دومنبعی مقایسه شده است.

ساختار مقاله به ترتیب زیر است. ابتدا، در بخش ۲ جزئیات سیستم جداکننده دوگوشی MESSL ارائه می‌شود. در بخش ۳، روش‌های پیشنهادی بهبود سیستم جداسازی دوگوشی منبع مطرح شده است. در قسمت اول این بخش، روش ANC_LPSO و در قسمت دوم روش WT_MMSE توصیف شده است. بخش ۴ به گزارش نتایج تجربی و شبیه‌سازی‌های حاصل از ارزیابی و مقایسه روش‌های پیشنهادی می‌پردازد. نتیجه‌گیری مقاله در بخش ۵ آمده است.

۲- جداسازی دوگوشی سیگنال مبتنی بر سیستم MESSL

بلوک‌دیگرام پیشنهادی به منظور جداسازی دوگوشی گفتار مبتنی بر سیستم MESSL در شکل ۱ نشان داده شده است. در این بخش، به توصیف جزئیات سیستم جداسازی دوگوشی MESSL می‌پردازیم. این سیستم در شکل ۱ با نقطه چین مشخص شده است.

۲-۱- تعریف مدل

با فرض نمایش منابع صوتی موجود در محیط به صورت $s(t)$ ، سیگنال‌های رسیده به گوش‌های چپ و راست، متناظر با، به ترتیب، $\ell(t)$ و $r(t)$ ، به شکل زیر تعریف می‌شوند [۱۶]:

$$\begin{aligned} \ell(t) &= s(t - \tau_l) * h_l(t) \\ r(t) &= s(t - \tau_r) * h_r(t), \end{aligned} \quad (1)$$

که در آن $h_l(t)$ و $h_r(t)$ ، به ترتیب، پاسخ ضربه محیط برای سیگنال‌های دریافتی از منبع توسط گوش‌های چپ و راست می‌باشند. نسبت تبدیل فوریه زمان کوتاه (STFT) دو سیگنال رسیده به گوش‌های چپ و راست، به عنوان طیف‌نگاره درون گوش^۱، با روابط زیر بیان می‌شود:

$$\frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega, t)/20} e^{j\phi(\omega, t)} \quad (2)$$

$$\approx 10^{\alpha(\omega)/20} e^{-j\omega\tau(\omega)} \quad (3)$$

به‌طوری‌که:

$$H(\omega) = \mathcal{F}[h_l(t)] / \mathcal{F}[h_r(t)] \quad (4)$$

$$\tau(\omega) = \tau_l - \tau_r + \angle H(\omega) \quad (5)$$

$$\alpha(\omega) = 20 \log_{10} |H(\omega)|. \quad (6)$$

مسیریابی باشند، متغیر پنهان τ را به صورت متغیر تصادفی گسسته و به صورت مجموعه‌ای از تأخیرهای مجاز پیش فرض^{۱۷} در نظر می‌گیریم. پارامتر $\xi(\omega)$ ، مقدار افسست در بازه $(-\pi, \pi)$ است که اجازه اندکی انحراف از مدل مستقل از فرکانس را می‌دهد. هر دو پارامتر i (اندیس منبع) و τ (متغیر پنهان) در متغیر پنهان $z_{i\tau}(\omega, t)$ قرار می‌گیرند، به نحوی که مقدار این متغیر، در صورتی که نقطه (ω, t) از منبع i و تأخیر τ باشد، یک و در غیر این صورت، صفر است. هر واحد T-F از طیف‌نگاره باید به منبع و تأخیری تعلق داشته باشد که در نتیجه آن $\sum_{i,\tau} z_{i\tau}(\omega, t) = 1$ خواهد شد.

پارامترهای مربوط به توزیع‌های گوسی موجود در مدل به همراه پارامتر عضویت کرانه‌ای^{۱۸} $\psi_{i\tau} \equiv p(i, \tau)$ ، که نشان‌دهنده احتمال حضور واحدهای T-F از منبع i و تأخیر τ است، در مرحله پیشینه‌سازی الگوریتم EM محاسبه می‌شوند. از آنجایی که τ ها مقادیر گسسته‌ای در این مدل هستند، $\psi_{i\tau}$ یک ماتریس دوبعدی از مقادیر احتمالی حضور در هر حالت گسسته است. پارامترهای IPD $\xi_{i\tau}(\omega)$ و $(\sigma_{i\tau}(\omega))$ ، به ω و τ وابسته هستند، در صورتی که پارامترهای ILD $(\mu_i(\omega))$ و $(\eta_i(\omega))$ ، فقط به ω وابسته‌اند. پارامترهای IPD و ILD بستگی به i ، یعنی منبعی که از آن نشأت می‌گیرند، دارند. با فرض $\Theta \equiv \{\xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\}$ به عنوان بردار مشاهدات و $N_\phi = N(\phi(\omega, t; \tau) | \xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega))$ و $N_\alpha = N(\alpha(\omega, t) | \mu_i(\omega), \eta_i^2(\omega))$ می‌توان احتمال لگاریتمی کل را به صورت رابطه (۱۲) تعریف کرد:

$$\Gamma(\Theta) = \sum_{\omega, t} \log p(\phi(\omega, t), \alpha(\omega, t) | \Theta) \\ = \sum_{\omega, t} \log \sum_{i, \tau} [N_\phi \cdot N_\alpha \cdot \psi_{i\tau}]. \quad (12)$$

در این رابطه، به تعداد $N_i \times N_\tau$ توزیع گوسی وجود دارد که با ضرایب $\psi_{i\tau}$ با یکدیگر ترکیب شده و مقداردهی اولیه آن‌ها نیز توسط هیستوگرام PHAT صورت می‌گیرد [۲۳]. رابطه (۱۲) در اصل یک مدل ترکیب گوسی (GMM)، با یک توزیع گوسی به ازای هر ترکیب از (i, τ) و $\psi_{i\tau}$ به عنوان ثابت وزن دهی ترکیب کننده است. در اینجا، تعداد منابع بایستی از قبل مشخص شده باشد. با استفاده از این احتمال لگاریتمی، می‌توان تابع کمکی زیر را بر اساس متغیر پنهان $z_{i\tau}(\omega, t)$ جهت پیشینه‌کردن Θ تعریف کرد:

$$Q(\Theta | \Theta_s) = k + \sum_{\omega, t} \sum_{i, \tau} [p(z_{i\tau}(\omega, t) | \phi(\omega, t), \alpha(\omega, t), \Theta_s) \cdot \log p(z_{i\tau}(\omega, t) | \phi(\omega, t), \alpha(\omega, t) | \Theta)] \quad (13)$$

به طوری که Θ_s تخمین پارامترهای Θ بعد از s تکرار و ثابت k نیز مستقل از Θ است.

$$\hat{\phi}(\omega, t; \tau) = \arg(e^{j\hat{\phi}(\omega, t)} e^{-j\omega\tau(\omega)}), \quad (7)$$

به طوری که همواره $\hat{\phi}(\omega, t; \tau)$ در بازه $(-\pi, \pi]$ قرار دارد. برای مدل کردن IPD و ILD، به ترتیب، از توزیع گوسی به صورت زیر استفاده می‌شود [۱۶]:

$$p(\phi(\omega, t) | \tau(\omega), \sigma(\omega)) = N(\hat{\phi}(\omega, t; \tau(\omega)) | 0, \sigma^2(\omega)) \\ \approx N(\hat{\phi}(\omega, t) | \omega\tau(\omega), \sigma^2(\omega)), \quad (8)$$

$p(\alpha(\omega, t) | \mu(\omega), \eta^2(\omega)) = N(\alpha(\omega, t) | \mu(\omega), \eta^2(\omega))$ ، (۹) که در آن میانگین ILD و $\sigma(\omega)$ و $\eta(\omega)$ به ترتیب، بیانگر انحراف معیار توزیع‌های گوسی IPD و ILD هستند. توزیع توأم IPD و ILD، با فرض استقلال آماری این دو پارامتر در هر واحد T-F، در حالت کلی به صورت زیر بیان می‌شود:

$$p(\phi(\omega, t), \alpha(\omega, t) | \Theta) = N(\hat{\phi}(\omega, t) | \xi(\omega), \sigma^2(\omega)) \\ \cdot N(\alpha(\omega, t) | \mu(\omega), \eta^2(\omega)), \quad (10)$$

که در آن $\xi(\omega)$ میانگین توزیع گوسی متغیر تصادفی $\hat{\phi}$ و Θ شامل همه پارامترهای مدل است.

۲-۲- تخمین پارامترهای مدل

پارامترهای مدل شرح داده شده در بخش ۲-۱ را نمی‌توان به صورت مستقیم از سیگنال مخلوط به دست آورد، چون به دلیل توزیع متفاوت منابع بر روی IPD و ILD، نواحی مختلف طیف‌نگاره درون گوشه توسط منابع مختلف اشغال شده است. در اینجا فرض می‌شود که تنها واحدهای T-F مربوط به یک منبع که دارای تأخیرهای برابری هستند، به طور یکسان توزیع شده‌اند. بنابراین پارامترهای مدل، تنها زمانی می‌توانند تخمین زده شوند که معلوم شود واحدهای T-F طیف‌نگاره درون گوشه متعلق به کدامین تأخیر و کدامین منبع هستند. این نوع روش جداسازی، در حقیقت مسئله داده‌های گم‌شده^{۱۲} است که پارامترهای احتمال پیشینه (ML)^{۱۳} آن توسط الگوریتم EM^{۱۴} استخراج می‌شوند. الگوریتم EM برای هر یک از منابع در سیگنال مخلوط، محل‌هایی از طیف‌نگاره درون گوشه را انتخاب می‌کند که به بهترین شکل ممکن با پارامترهای منبع مطابقت داشته باشند و در نهایت پارامترهایش را از همان نواحی تخمین می‌زند. الگوریتم EM به جای استفاده از یک ماسک باینری سخت^{۱۵}، از یک ماسک نرم^{۱۶} مبتنی بر احتمال حضور هر یک از منابع در هر واحد T-F برای استخراج سیگنال‌ها استفاده می‌کند.

تأخیر $\tau(\omega)$ منبعی که در هر واحد T-F طیف‌نگاره غالب است، شامل یک متغیر پنهان τ است که به صورت زیر مدل می‌شود [۱۶]:

$$\tau(\omega) = \tau + \omega^{-1} \xi(\omega). \quad (11)$$

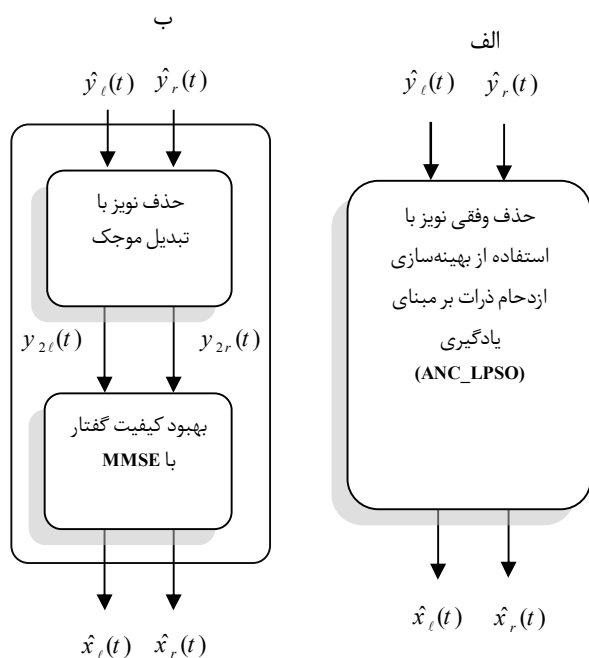
در رابطه (۱۱)، τ تأخیر مستقل از فرکانس است که در مکان‌یابی مورد استفاده قرار می‌گیرد. برای آن که سیگنال‌های تداخل نیز قابل

در مرحله دوم، روش حداقل میانگین مربعات خطا (MMSE) جهت ارتقاء هرچه بیش تر کیفیت سیگنال جدا شده هدف به کار می رود. بلوک دیاگرام سیستم های پس پردازش پیشنهادی در شکل ۲ آمده است.

۳-۱- حذف و فقی نویز با استفاده از بهینه سازی ازدحام ذرات بر

مبنای یادگیری (ANC_LPSO)

سیستم پس پردازش پیشنهادی اول، الگوریتم حذف و فقی نویز با استفاده از بهینه سازی ازدحام ذرات بر مبنای یادگیری (ANC_LPSO) است. ANC_LPSO که یک روش دوکانالی برای بهبود گفتار می باشد، عملکرد بسیار مناسبی داشته و برتری چشمگیر آن نسبت به روش هایی مانند الگوریتم ژنتیک (GA)، حذف و فقی نویز با استفاده از بهینه سازی ازدحام ذرات استاندارد (SPSO) و حداقل مربعات میانگین نرمالیزه (NLMS) در مسئله بهبود کیفیت سیگنال نویزی نشان داده شده است [۲۰].



شکل ۲: بلوک دیاگرام سیستم های پس پردازش پیشنهادی برای ارتقاء عملکرد سیستم دوگوشی جداسازی گفتار (الف)؛ روش حذف و فقی نویز با استفاده از بهینه سازی ازدحام ذرات بر مبنای یادگیری (ANC_LPSO) [۲۰]. (ب) روش دومرحله ای حذف نویز بر اساس تبدیل موجک و MMSE (WT_MMSE).

۳-۲- حذف نویز دومرحله ای بر اساس تبدیل موجک و MMSE (WT_MMSE)

سیستم پس پردازش پیشنهادی دوم شامل دو مرحله حذف نویز بر اساس تبدیل موجک و روش MMSE است. در این بخش به توصیف مراحل پردازشی این روش پرداخته می شود.

در مرحله امید ریاضی (E)، متغیر پنهان $z_{i\tau}(\omega, t)$ به شرط مشاهدات $(\phi$ و $\alpha)$ و پارامتر Θ_s محاسبه می شود و در مرحله بیشینه سازی (M)، مقدار بیشینه Q نسبت به Θ و امید ریاضی $z_{i\tau}(\omega, t)$ به دست می آید. در مرحله E، احتمال حضور منبع i و تأخیر τ در واحد T-F را به صورت زیر خواهیم داشت:

$$v_{i\tau}(\omega, t) \equiv p(z_{i\tau}(\omega, t) | \phi(\omega, t), \alpha(\omega, t), \Theta_s) \propto p(z_{i\tau}(\omega, t), \phi(\omega, t), \alpha(\omega, t) | \Theta_s) = \psi_{i\tau} \cdot N_{\phi} \cdot N_{\alpha} \quad (14)$$

با استفاده از تعریف عملگر زیر به عنوان میانگین وزن دار:

$$\langle x \rangle_{i,\tau} \equiv \frac{\sum_{i,\tau} x v_{i\tau}(\omega, t)}{\sum_{i,\tau} v_{i\tau}(\omega, t)} \quad (15)$$

در مرحله M، پارامترهای مدل به صورت زیر محاسبه می شوند:

$$\mu_i(\omega) = \langle \alpha(\omega, t) \rangle_{i,\tau} \quad (16)$$

$$\eta_i^2(\omega) = \langle \alpha(\omega, t) - \mu_i(\omega) \rangle_{i,\tau}^2 \quad (17)$$

$$\xi_{i\tau}(\omega) = \langle \hat{\phi}(\omega, t; \tau) \rangle_i \quad (18)$$

$$\sigma_{i\tau}^2(\omega) = \langle \hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega) \rangle_i^2 \quad (19)$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega, t} v_{i\tau}(\omega, t). \quad (20)$$

۳-۲- تخمین ماسک جداسازی

می توان با جمع بر روی مقادیر τ ، ماسک مورد نظر برای هر منبع را به صورت زیر به دست آورد:

$$M_i(\omega, t) \equiv \sum_{\tau} v_{i\tau}(\omega, t). \quad (21)$$

ماسک به دست آمده، احتمال تعلق هر یک از واحدهای T-F در طیف نگاره را به منبع نام نشان می دهد [۱۶].

با اعمال ماسک به دست آمده بر روی تبدیل فوریه سیگنال های مخلوط گوش چپ و راست و سپس، انجام عمل عکس تبدیل فوریه، سیگنال هدف از سیگنال مخلوط ورودی استخراج می شود.

۳- سیستم های پیشنهادی جهت بهبود کیفیت سیگنال هدف

جدا شده توسط سیستم دوگوشی MESSL

به علت باقی ماندن بخشی از سیگنال گفتار منابع غیرهدف در سیگنال گفتار جدا شده منبع هدف، دو سیستم پس پردازش بهبود کیفیت گفتار برای ارتقاء عملکرد سیستم جداسازی دوگوشی MESSL پیشنهاد می شود. بلوک دیاگرام این روش در شکل ۲ نشان داده شده است. در قسمت اول، روش ANC_LPSO مرجع [۲۰] و در قسمت دوم روش WT_MMSE توصیف شده است. در روش اخیر، ابتدا، از روش حذف نویز به وسیله تبدیل موجک استفاده می شود. این مرحله از بهبود کیفیت گفتار، با درصد بالایی منابع گفتار ناخواسته را حذف می کند.

$$\int_{-\infty}^{\text{median}(|\hat{N}|)} \text{pdf}(x) dx = \frac{1}{2}. \quad (26)$$

با جایگذاری رابطه (۲۵) در رابطه (۲۶)، خواهیم داشت:

$$\int_0^{\text{median}(|\hat{N}|)} \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{\hat{N}^2}{2\sigma^2}} d\hat{N} = \frac{1}{2}. \quad (27)$$

از رابطه (۲۷) می توان $\text{median}(|\hat{N}|)$ را به صورت زیر به دست آورد:

$$\text{median}(|\hat{N}|) = \sigma\sqrt{2} \text{erf}^{-1}\left(\frac{1}{2}\right) \approx 0.6745 \sigma, \quad (28)$$

به طوری که $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx$. در نهایت، با فرض یکسان بودن توزیع ضرایب موجک سیگنال نویزی با نویز، $\text{pdf}(Y_1) \approx \text{pdf}(\hat{N})$ ، انحراف معیار مربوط به ضرایب تبدیل موجک سیگنال نویزی به صورت زیر بیان می شود:

$$\sigma = \frac{\text{median}(|Y_1|)}{0.6745}. \quad (29)$$

مرحله آستانه گذاری خود به دو روش انجام می گیرد: آستانه گذاری سخت و آستانه گذاری نرم. در روش آستانه گذاری سخت تمام ضرایب کوچکتر از آستانه δ مقدار صفر می گیرند.

$$Y_{1,j,k}^{\text{Hard}} = \begin{cases} Y_{1,j,k}, & |Y_{1,j,k}| > \delta \\ 0, & \text{otherwise.} \end{cases} \quad (30)$$

آستانه گذاری نرم یک مرحله نیز بیش تر پیش می رود و دامنه ضرایب را به اندازه δ کاهش می دهد.

$$Y_{1,j,k}^{\text{Soft}} = \text{sign}(Y_{1,j,k}) \max(|Y_{1,j,k}| - \delta, 0). \quad (31)$$

روش آستانه گذاری سخت اگر چه دامنه سیگنال را حفظ می کند، ولی به علت عدم پیوستگی در ضرایب موجک بعد از دوباره سازی سیگنال، اندکی نویز را وارد سیگنال می کند، در حالی که روش نرم با حذف این گسستگی در ضرایب موجک، سیگنال نرم تر، ولی با دامنه کاهش یافته، را به دست می دهد. در این مقاله، از روش آستانه گذاری نرم برای ضرایب تبدیل موجک سیگنال نویزی استفاده می شود.

۳-۲-۲- حذف نویز به روش MMSE

با وجود اینکه روش حذف نویز با موجک تا حد قابل قبولی سیگنال های مربوط به منابع تداخل را حذف می کند، اندکی سیگنال ناخواسته در خروجی باقی می ماند. برای ارتقاء کیفیت سیگنال هدف جدا شده توسط سیستم جداسازی دوگوشی، می توان از روش بهبود کیفیت گفتار MMSE در مرحله دوم سیستم پس پردازش پیشنهادی بهره برد [۲۵]. سیگنال رسیده به مرحله بهبود کیفیت گفتار با MMSE در حوزه فوری به صورت $Y_{2k} e^{j\theta_{y_2}(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)}$ نشان داده می شود که در آن $\{Y_{2k}, X_k, D_k\}$ اندازه و $\{\theta_{y_2}(k), \theta_x(k), \theta_d(k)\}$ فاز در نقطه فرکانسی k ام، به ترتیب، برای سیگنال گفتار نویزی، سیگنال تمیز (سیگنال هدف) و سیگنال نویز می باشد. در مرحله بعد، یک تخمین گر

۳-۲-۱- حذف نویز بر اساس تبدیل موجک

حذف نویز بر اساس تبدیل موجک شامل سه مرحله است: تجزیه سیگنال به ضرایب تبدیل موجک، آستانه گذاری ضرایب و دوباره سازی سیگنال با استفاده از ضرایب فیلتر شده. در ابتدا، تحلیل موجک را تا مرحله $N = 12$ با موجک Daubechies انجام می دهیم. سپس، ضرایب مربوط به موجک را آستانه گذاری می کنیم. در نهایت، دوباره سازی سیگنال را با استفاده از ضرایب آستانه گذاری شده خواهیم داشت.

اگر سیگنال رسیده به مرحله حذف نویز با موجک را به صورت $y_1(t) = x(t) + n(t)$ نشان دهیم که در آن $x(t)$ سیگنال مربوط به منبع هدف و $n(t)$ سیگنال های ناخواسته باشند، در این صورت پس از اعمال تبدیل موجک بر روی سیگنال ورودی، رابطه (۲۲) را برای ضرایب موجک خواهیم داشت [۲۴]:

$$Y_{1,j,k} = X_{j,k} + N_{j,k}, \quad (22)$$

به طوری که $Y_{1,j,k}$ ، $X_{j,k}$ و $N_{j,k}$ ، به ترتیب، k امین ضرایب موجک در مقیاس زام مربوط به سیگنال های $y_1(t)$ ، $x(t)$ و $n(t)$ می باشند [۲۱].

با فرض گوسی بودن نویز و توزیع یکنواخت آن روی ضرایب موجک، می توان مشخصات نویز را از ضرایب موجک آن استخراج کرد. مرحله آستانه گذاری به دو صورت انجام می شود: آستانه گذاری متفاوت در هر مرحله و آستانه گذاری یکسان برای کل مراحل. در این مقاله، از یک مقدار ثابت برای آستانه گذاری در تمام مراحل استفاده می کنیم. برای تعیین مقدار آستانه δ از روش Donoho [۲۵] بهره می بریم که به صورت زیر تعریف می شود:

$$\delta = \sigma\sqrt{2 \log L}. \quad (23)$$

در رابطه (۲۳)، L تعداد نمونه های سیگنال مورد پردازش و σ انحراف معیار مربوط به ضرایب تبدیل موجک نویز است.

برای تخمین σ ، به صورت زیر عمل می کنیم. با فرض گوسی بودن نویز، تابع چگالی احتمالی (pdf) نویز را به صورت زیر خواهیم داشت [۲۱]:

$$\text{pdf}(\hat{N}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\hat{N}^2}{2\sigma^2}}, \quad (24)$$

که در آن، $\hat{N}$ ضرایب تبدیل موجک نویز و σ انحراف معیار ضرایب تبدیل موجک نویز است. تابع چگالی احتمالی قدرمطلق $\hat{N}$ نیز به صورت زیر خواهد بود:

$$\text{pdf}(|\hat{N}|) = u(\hat{N}) \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{\hat{N}^2}{2\sigma^2}}, \quad (25)$$

که در آن تابع $u(\cdot)$ بیانگر تابع پله است. برای یافتن مقدار میانه $|\hat{N}|$ ، رابطه زیر بایستی برقرار باشد:

به طوری که $0 < a < 1$ ، ثابت وزن دهی و $\hat{X}_k^2(m-1)$ تخمین گر دامنه در فریم قبلی است. بهترین نتیجه ممکن برای $a=0.98$ به دست می آید.

۴- نتایج آزمایش ها

برای سنجش عملکرد سیستم های بهبود گفتار پیشنهادی، از معیارهای نسبت سیگنال به اغتشاش (SDR)، نسبت سیگنال به تداخل (SIR) و نسبت سیگنال به مصنوعات (SAR)، به ترتیب، با روابط زیر استفاده می شود [۲۶]:

$$\text{SDR} = 10 \log_{10} \frac{\|x_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2}, \quad (40)$$

$$\text{SIR} = 10 \log_{10} \frac{\|x_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}, \quad (41)$$

$$\text{SAR} = 10 \log_{10} \frac{\|x_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2}, \quad (42)$$

که در آن x_{target} سیگنال هدف، e_{interf} سیگنال تداخل و e_{artif} سیگنال های مصنوعی ایجاد شده هستند. از میان معیارهای ذکر شده، معیار SDR مطابقت بالایی با نتایج تشخیص خودکار گفتار (ASR) دارد و می تواند به عنوان معیار اندازه گیری قابلیت فهم نیز به کار رود [۲۷]. همچنین، برای سنجش میزان عملکرد سیستم پیشنهادی، از معیار دیگری نیز بنام معیار PESQ که همبستگی بالایی با معیار ذهنی دارد، استفاده شده است [۲۸].

فایل های صوتی مورد استفاده در آزمایش ها، اعم از سیگنال های هدف و تداخل، از پایگاه داده فارس دات (FARSDAT) انتخاب شده اند [۲۹]. FARSDAT یک پایگاه داده گفتار فارسی پیوسته است که از ۶۰۰۰ گفته و از ۳۰۰ گوینده با لهجه های مختلف فارسی موجود در ایران تشکیل شده است. این جملات از یک مجموعه ۳۹۰ جمله ای توسط گوینده های مختلف خوانده شده است. پانزده جمله از جملات فوق به شکل تصادفی انتخاب شده و در ترکیب با سیگنال های تداخل به کار رفته است. برای پاسخ ضربه اتاق از داده های موجود در مرجع [۳۰] استفاده شده است. فرکانس کاری مراجع [۲۹-۳۰] ۴۴۱۰۰ هرتز است که با نمونه برداری مجدد مقدار آن به ۱۶ کیلوهرتز کاهش یافته است. طول فریم ها در تحلیل STFT، ۶۴ میلی ثانیه و مقدار همپوشانی آن ها ۱۶ میلی ثانیه در نظر گرفته شده است. برای پنجره بندی سیگنال از پنجره hamming استفاده شده است. در شبیه سازی های انجام گرفته، عملکرد الگوریتم ها در حالت های متفاوتی بررسی شده است. در جدول ۱، شرایط مختلف شبیه سازی ها با پارامتر حالت θ_{ij} نشان داده شده است. در اینجا، اندیس i و j ، به ترتیب، میزان پیچیدگی پارامترهای ILD و IPD را، از نقطه نظر اینکه این پارامترها دارای مقدار ثابت بوده و یا وابسته به فرکانس هستند، بیان می کنند. اندیس های i و j مقادیر ۰، ۱ و یا Ω را به خود می گیرند که به ترتیب، معرف ساده ترین حالت، حالت پیچیده تر و حالت با بیش ترین پیچیدگی هستند.

بهینه اندازه طیف با روش MMSE، که میانگین مربعات خطای مربوط به لگاریتم اندازه طیف زیر را کمینه می کند، به دست می آید [۲۲]:

$$E \left\{ \left(\log X_k - \log \hat{X}_k \right)^2 \right\}. \quad (32)$$

تخمین گر بهینه اندازه log-MMSE را می توان به صورت زیر به دست آورد:

$$\begin{aligned} \hat{X}_k &= \frac{\xi_k}{1 + \xi_k} \exp \left\{ \frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t} dt \right\} Y_k \\ &= G_{LSA}(\xi_k, v_k) Y_{2k}, \end{aligned} \quad (33)$$

که در آن $G_{LSA}(\xi_k, v_k)$ تابع بهره تخمین گر log-MMSE است. متغیر v_k به صورت زیر تعریف می شود [۲۵]:

$$v_k = \frac{\xi_k}{1 + \xi_k} \gamma_k, \quad (34)$$

که در آن متغیرهای ξ_k و γ_k ، به ترتیب، SNR پیشین و پسین بوده و به صورت زیر تعریف می شوند:

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}, \quad (35)$$

$$\gamma_k = \frac{Y_k^2}{\lambda_d(k)}, \quad (36)$$

که در آن $\lambda_d(k)$ واریانس نویز است. برای تشکیل طیف سیگنال تمیز $\hat{X}(\omega_k) = \hat{X}_k \exp(j\hat{\theta}_x(k))$ ، نیاز به تخمین فاز نیز است. تخمین گر MMSE فاز به صورت مسئله کمینه سازی مشروط زیر به دست می آید:

$$\begin{aligned} \min_{\exp(j\hat{\theta}_x)} &= E \left\{ \left| \exp(j\theta_x) - \exp(j\hat{\theta}_x) \right|^2 \right\}, \\ &\text{به شرط: } \left| \exp(j\hat{\theta}_x) \right| = 1. \end{aligned} \quad (37)$$

با استفاده از روش ضرایب لاگرانژ، فاز بهینه را به صورت زیر خواهیم داشت:

$$\exp(j\hat{\theta}_x) = \exp(j\theta_y). \quad (38)$$

به عبارتی، فاز بهینه برای بازسازی سیگنال تمیز برابر با فاز سیگنال آلوده است.

تخمین گر اندازه log-MMSE در رابطه (۳۳)، با این فرض به دست آمد که SNR پیشین ξ_k و واریانس نویز $\lambda_d(k)$ معلوم هستند. از آنجاکه در عمل، فقط به سیگنال آلوده دسترسی وجود دارد، می توان واریانس نویز را با فرض ایستادن بودن نویز توسط روش آشکارساز فعالیت نویز (VAD) در بازه های عدم حضور گفتار به دست آورد. برای محاسبه SNR پیشین ξ_k ، از روش همسو با تصمیم^{۱۹} استفاده می شود [۲۲].

$$\hat{\xi}_k(m) = a \frac{\hat{X}_k^2(m-1)}{\lambda_d(k, m-1)} + (1-a) \max[\gamma_k(m) - 1, 0], \quad (39)$$

جدول ۱: شرایط مختلف شبیه‌سازی برای پارامتر حالت θ ، میانگین m و std انحراف معیار است [۲۰]

پارامتر حالت	ILD		IPD	
	m	std	m	std
θ_{00}	0	∞	0	σ_i
θ_{10}	μ_i	η_i	0	σ_i
θ_{01}	0	∞	$\xi_{i\tau}$	$\sigma_{i\tau}$
θ_{11}	μ_i	η_i	$\xi_{i\tau}$	$\sigma_{i\tau}$
$\theta_{0\Omega}$	0	∞	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$
$\theta_{\Omega 0}$	$\mu_i(\omega)$	$\eta_i(\omega)$	0	σ_i
$\theta_{\Omega \Omega}$	$\mu_i(\omega)$	$\eta_i(\omega)$	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$

عملکرد سیستم‌های پس‌پردازش پیشنهادی برای حالت دو منبعی مقایسه شده است. در جدول ۲، میانگین مقادیر SDR، SIR، SAR و PESQ مربوط به سیگنال‌های ترکیب ورودی و سیگنال‌های به‌دست‌آمده از سیستم جداسازی دوگوشی MESSL، سیستم‌های جداسازی دوگوشی با پس‌پردازش‌های پیشنهادی ANC_LPSO و WT_MMSE آورده شده است. همان‌گونه که مشاهده می‌شود، با استفاده از سیستم بهبود کیفیت گفتار پیشنهادی WT_MMSE، مقادیر SDR، SIR، SAR و PESQ به‌میزان قابل‌توجهی افزایش یافته است. بیش‌ترین مقدار افزایش SDR، SIR، SAR و PESQ نسبت به سیگنال ترکیب ورودی متعلق به حالت $\theta_{\Omega\Omega}$ بوده و، به ترتیب، به‌میزان ۱۳/۵۸، ۷/۱۲، ۱۳/۷۴، دسی‌بل و ۱/۷۵ (MOS) است. متوسط افزایش SDR، SIR، SAR و PESQ در مورد این سیستم پیشنهادی، در جدول ۲، به ترتیب، ۵/۹۴، ۴، ۵/۹۴، دسی‌بل و ۱/۰۸ (MOS) است. همچنین، متوسط مقدار افزایش SDR، SIR، SAR و PESQ نسبت به خروجی سیستم جداسازی دوگوشی MESSL، به ترتیب، به‌میزان ۱/۳۴، ۰/۲۹، ۱/۳۳، دسی‌بل و ۰/۷۸ (MOS) است. در مجموع، عملکرد سیستم پیشنهادی دوم (WT_MMSE) بسیار رضایت‌بخش ارزیابی می‌شود. همچنین، نتایج مندرج در جدول ۲ نشان می‌دهند که سیستم پس‌پردازش پیشنهادی WT_MMSE در مقایسه با سیستم ANC_LPSO بسیار بهتر عمل می‌کند. نتایج مربوط به حالتی که دو سیگنال تداخل در طرفین منبع هدف قرار گرفته است، در جدول ۳ و حالتی که سیگنال‌های تداخل هر دو در یک سمت قرار گرفته‌اند، در جدول ۴ آورده شده است. در حالتی که منابع تداخل نزدیک هم و

در یک سمت قرار گرفته‌اند، میزان افزایش SDR نسبت به سیگنال ترکیب ورودی به‌طور میانگین، ۱/۷۷ دسی‌بل بیش‌تر از حالتی است که سیگنال‌های تداخل در دو سمت منبع هدف قرار گرفته‌اند. این مسئله توانایی بهتر سیستم شنوایی دوگوشی را در جداسازی سیگنال هدف در حضور سیگنال‌های تداخلی که در سمت یک گوش متمرکز هستند، با استفاده از مزیت گوش برتر^۲، نشان می‌دهد. با اعمال سیستم پیشنهادی WT_MMSE به خروجی سیستم جداسازی دوگوشی، مقادیر SDR در جدول ۳ به‌طور متوسط به‌اندازه ۲/۱ دسی‌بل نسبت به خروجی MESSL افزایش می‌یابد که بیش‌ترین آن مربوط به حالت $\theta_{\Omega\Omega}$ (به‌میزان ۷/۲۹ دسی‌بل) است. میزان افزایش PESQ نیز به‌طور متوسط ۰/۶۶ (MOS) است که بیش‌ترین آن مربوط به حالت $\theta_{\Omega 0}$ است. در جدول ۴ نیز متوسط افزایش SDR و PESQ نسبت به خروجی MESSL، به ترتیب، به‌میزان ۰/۵۴، دسی‌بل و ۰/۳۷ (MOS) می‌باشند. کاهش میانگین بهره مرحله پس‌پردازش پیشنهادی WT_MMSE برحسب همه معیارها در جدول ۴ نسبت به جدول ۳ نشان می‌دهد که در حالتی که دو سیگنال تداخل در یک سمت قرار گرفته‌اند، مرحله پس‌پردازش پیشنهادی نسبت به سیستم جداکننده دوگوشی MESSL نقش کم‌تری در ارتقاء کیفیت سیگنال‌های جداشده دارد. ارتقاء کیفیت گفتار هدف جداشده توسط سیستم جداسازی دوگوشی با پارامتر حالت پیچیده‌تر ($\theta_{\Omega\Omega}$) در هر دو شرایط دو منبعی و سه منبعی برحسب تمام معیارهای به‌کاررفته به مراتب بیش‌تر است.

همچنین، با توجه به آنچه که در ابتدای این بخش در مورد همبستگی بالای معیارهای SDR و PESQ، به ترتیب، با ASR و تست‌های شنوایی افراد ذکر شد، می‌توان نتیجه گرفت که سیستم پیشنهادی WT_MMSE قابلیت فهم سیگنال‌های جداشده را نیز به-میزان چشمگیری افزایش می‌دهد.

جدول ۲: میانگین مقادیر SAR ، SIR ، SDR و PESQ برای سیگنال‌های ترکیب، در حالت دو منبعی با تداخل در زاویه $\theta = 45^\circ$ و سیگنال‌های به‌دست‌آمده توسط

سیستم جداسازی دوگوشی MESSL و سیستم‌های جداسازی دوگوشی با پس‌پردازش پیشنهادی WT_MMSE و ANC_LPSO

پارامتر حالت	Input Mixture $\theta = 45^\circ$				(MESSL) $\theta = 45^\circ$				(MESSL+ANC+LPSO) $\theta = 45^\circ$				(MESSL+WT+MMSE) $\theta = 45^\circ$			
	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)
θ_{00}	-۱/۱۷	۲۱/۸۲	-۱/۰۸	۱/۳۵	۰/۹	۲۴/۵۱	۱/۱۸	۱/۴۳	-۱/۳۱	۱۴/۴۵	-۰/۹	۱/۵۴	-۰/۸۴	۲۴/۱۱	-۰/۸۶	۲/۰۷
θ_{10}					۳/۰۳	۲۶/۹۵	۳/۰۷	۱/۷۴	-۰/۳۵	۱۷/۳۶	-۰/۶۴	۱/۳۶	۳/۴۱	۲۶/۹۳	۳/۴۶	۲/۳
θ_{01}					۰/۸۴	۲۳/۲۲	-۰/۹۱	۱/۳۴	-۱/۳۸	۱۴/۸	-۰/۹۶	۱/۲۹	۲/۳۱	۲۳/۹۹	۲/۳۹	۲/۲۵
θ_{11}					۳/۶۱	۲۶/۸۲	۳/۶۵	۱/۷۲	-۰/۹۳	۱۸/۶۱	۱/۱۱	۱/۳۷	۵/۳۳	۲۷/۵۷	۵/۴	۲/۴۹
$\theta_{0\Omega}$					۳/۳۳	۲۳/۶۸	۳/۴۲	۱/۴	۱/۲۳	۱۷/۷	۱/۴۸	۱/۴	۵/۶۶	۲۴/۲۸	۵/۸	۲/۵۴
$\theta_{\Omega 0}$					۳/۲۴	۲۵/۷۹	۳/۲۹	۱/۷۶	-۰/۵۴	۱۸/۹	-۰/۷	۱/۴۵	۳/۴۳	۲۴/۹۴	۳/۴۹	۲/۲۹
$\theta_{\Omega \Omega}$					۹/۰۶	۲۷/۷۹	۹/۲۵	۲/۱۸	۴/۶۵	۲۱/۲۲	۴/۸۸	۱/۶۵	۱۲/۴۱	۲۸/۹۴	۱۲/۶۶	۳/۱
میانگین	-۱/۱۷	۲۱/۸۲	-۱/۰۸	۱/۳۵	۳/۴۳	۲۵/۵۳	۳/۵۳	۱/۶۵	-۰/۷۱	۱۷/۵۷	-۰/۹۹	۱/۴۳	۴/۷۷	۲۵/۸۲	۴/۸۶	۲/۴۳
میانگین بهره نسبت به سیگنال مخلوط	-				۴/۶	۳/۷۱	۴/۶۱	-۰/۳	۱/۸۸	-۴/۲۵	۲/۰۷	-۰/۰۸	۵/۹۴	۴	۵/۹۴	۱/۰۸
میانگین بهره نسبت به سیگنال MESSL	-				-				-۲/۷۲	-۷/۹۶	-۲/۵۴	-۰/۲۲	۱/۳۴	-۰/۲۹	۱/۳۳	-۰/۷۸

جدول ۳: میانگین مقادیر SAR ، SIR ، SDR و PESQ برای سیگنال‌های ترکیب، در حالت سه منبعی با تداخل در زوایای $\theta = -65^\circ, 45^\circ$ و سیگنال‌های

به‌دست‌آمده توسط سیستم جداسازی دوگوشی MESSL و سیستم جداسازی دوگوشی با پس‌پردازش پیشنهادی WT_MMSE

پارامتر حالت	Mixture $\theta = -65^\circ, 45^\circ$				(MESSL) $\theta = -65^\circ, 45^\circ$				(MESSL+WT+MMSE) $\theta = -65^\circ, 45^\circ$			
	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)
θ_{00}	-۵/۷۵	۱۳/۴۱	-۵/۵	۱/۰۸	-۳/۳۴	۱۳/۹۸	-۲/۰۹	۱/۵۲	-۲/۴۷	۱۵/۴۴	-۲/۲۸	۲/۲
θ_{10}					-۳/۱۵	۱۳/۹۴	-۲/۹۹	۱/۵۸	-۱/۳۶	۱۶/۴۳	-۱/۱۹	۲/۲۸
θ_{01}					-۱/۵۴	۱۶/۰۶	-۱/۳۶	۱/۵۷	-۱/۱۶	۱۶/۷۵	۱-	۲/۲
θ_{11}					-۱/۷۸	۱۵/۰۹	-۱/۵۶	۱/۶۴	-۰/۱	۱۷/۲۷	-۰/۵۸	۲/۲۹
$\theta_{0\Omega}$					۱/۷۱	۱۷/۳۳	۱/۹۱	۱/۸۵	۲/۵۶	۱۸/۳۳	۲/۷۴	۲/۵۲
$\theta_{\Omega 0}$					-۲/۲۲	۱۵/۶۳	-۲/۰۴	۱/۶۴	-۰/۵۸	۱۷/۹۴	-۰/۴۵	۲/۳۷
$\theta_{\Omega \Omega}$					۴/۳۷	۱۷/۶	۴/۶۶	۲/۴۱	۱۱/۶۶	۱۹/۴۸	۱۲/۵	۲/۸۹
میانگین	-۵/۷۵	۱۳/۴۱	-۵/۵	۱/۰۸	-۰/۸۵	۱۵/۶۶	-۰/۶۴	۱/۷۴	۱/۲۵	۱۷/۳۸	۱/۵۶	۲/۴
میانگین بهره نسبت به سیگنال مخلوط	-				۴/۹	۲/۲۵	۴/۸۶	-۰/۶۶	۷	۳/۹۷	۷/۰۶	۱/۳۲
میانگین بهره نسبت به سیگنال خروجی MESSL	-				-				۲/۱	۱/۷۲	۲/۲	-۰/۶۶

جدول ۴: میانگین مقادیر SIR ، SDR ، SAR و PESQ برای سیگنال‌های ترکیب، در حالت سه منبعی با تداخل در زوایای $\theta = 45^\circ, 65^\circ$ و سیگنال‌های به‌دست‌آمده توسط سیستم جداسازی دوگوشی MESSL و سیستم جداسازی دوگوشی با پس‌پردازش پیشنهادی WT_MMSE

پارامتر حالت	Mixture $\theta = 45^\circ, 65^\circ$				(MESSL) $\theta = 45^\circ, 65^\circ$				(MESSL+WT+MMSE) $\theta = 45^\circ, 65^\circ$			
	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)	SDR (dB)	SIR (dB)	SAR (dB)	PESQ (MOS)
θ_{00}					۱/۵۱	۱۸/۰۲	۱/۶۷	۲/۰۳	۱/۶۶	۱۷/۵	۱/۸۶	۲/۵۸
θ_{10}					۲/۳	۱۸/۰۸	۲/۴۹	۲/۱۸	۲/۷۵	۱۸/۱۹	۲/۹۵	۲/۶۱
θ_{01}					۴	۲۰/۰۴	۴/۱۵	۲/۱۸	۳/۸۴	۱۹/۸	۴	۲/۶۷
θ_{11}	-۳/۰۲	۱۶/۵۸	-۲/۸۸	۱/۱۶	۵/۱۶	۲۰/۱	۵/۳۳	۲/۳۷	۵/۱۷	۲۰/۰۵	۵/۳۶	۲/۷۲
$\theta_{0\Omega}$					۷/۸۵	۲۱/۱۷	۸/۱	۲/۴۸	۸/۲۸	۲۲/۰۶	۸/۵	۲/۸۹
$\theta_{\Omega 0}$					۳/۷۶	۲۰/۲۸	۳/۹	۲/۲۲	۴/۲۱	۲۰/۴۸	۴/۳۶	۲/۶۱
$\theta_{\Omega\Omega}$					۱۱/۹	۲۲/۷۳	۱۲/۳	۲/۸۱	۱۴/۳۵	۲۵/۱۵	۱۴/۷۴	۲/۷۵
میانگین	-۳/۰۲	۱۶/۵۸	-۲/۸۸	۱/۱۶	۵/۲۱	۲۰/۰۶	۵/۴۲	۲/۳۲	۵/۷۵	۲۰/۴۶	۵/۹۷	۲/۶۹
میانگین بهره نسبت به سیگنال مخلوط	-				۸/۲۳	۳/۴۸	۸/۳	۱/۱۶	۸/۷۷	۳/۸۸	۵/۸۵	۱/۵۳
میانگین بهره نسبت به سیگنال خروجی MESSL	-				-				۰/۵۴	۰/۴	-۲/۴۵	۰/۳۷

مراجع

۵- نتیجه‌گیری

- [1] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Constant directivity beam forming," *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Ed. Springer, pp. 3-17, 2001.
- [2] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57 no. 8, pp. 1408-1418, 1969.
- [3] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926-935, 1972.
- [4] L. Griffiths, and C. Jim, "An alternative approach to linearly constrained adaptive beam forming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27-34, 1982.
- [5] A. S. Feng, and D. L. Jones, "Localization-based grouping," *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, pp. 187-207, 2006.
- [6] A. Hyvarinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94-128, 1999.
- [7] S. Choi, A. Cichocki, H. Park, and S. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing, Letters and Reviews*, vol. 6, no. 1, pp. 1-57, 2005.
- [8] A. J. Bell, and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129-1159, 1995.
- [9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21-34, 1998.
- [10] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beam forming," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, vol. 5, pp. 2733-2736, 2001.
- [11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation

سیستم جداساز دوگوشی MESSL با محاسبه پارامترهای مدل گوسی مربوط به ویژگی‌های دوگوشی IPD و ILD توسط الگوریتم EM . اقدام به استخراج ماسک نرمی می‌کند که با ضرب آن در STFT سیگنال مخلوط، سیگنال گفتار هدف جدا می‌شود. در این مقاله، به علت عملکرد ناقص سیستم MESSL در جداسازی سیگنال‌ها، دو سیستم پس‌پردازش برای بهبود کیفیت گفتار جدا شده توسط این سیستم ارائه شده است. یک سیستم، حذف وقتی نویز با استفاده از بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (ANC_LPSO) و سیستم پیاده‌سازی شده دیگر، حذف نویز با استفاده از تبدیل موجک و MMSE (WT_MMSE) است. مقایسه نتایج به‌دست‌آمده از ارزیابی‌های مختلف در حالت دومنبعی برای شرایط مختلف پارامتر حالت، نشان می‌دهد که سیستم پس‌پردازش پیشنهادی دوم (WT_MMSE) بهتر از روش مبتنی بر LPSO عمل می‌کند. همچنین، در هر دو حالت دومنبعی و سه‌منبعی، هرچه شرایط شبیه‌سازی (پارامتر حالت) پیچیده‌تر باشند (میانگین آماری و واریانس توزیع IPD و ILD تابع فرکانس باشند)، افزایش مقادیر SDR و PESQ بیش‌تر است. باید توجه داشت که روش‌های پیشنهادی بهبود کیفیت گفتار برای شرایط بدون پژواک ارائه شده است. همچنین، جداسازی تنها بر روی سیگنال‌های از پیش ضبط‌شده انجام می‌شود. علیرغم محدودیت‌های ذکرشده، سیستم پیشنهادی MESSL+WT+MMSE نسبت به سایر سیستم‌های مقایسه شده (MESSL و MESSL+ANC+LPSO) ضمن داشتن پیچیدگی محاسباتی اندک، عملکرد بسیار رضایت‌بخشی در جداسازی سیگنال گفتار هدف از سیگنال مخلوط ورودی دارد.

- of Farsi spoken language,” *Proc. 5th Australian Int. Conf. Speech Science and Technology (SST’94)*, pp. 826-831, 1994.
- [30] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF database,” *IEEE Workshop Applcat. Signal Process. Audio Acoust.*, pp. 99-102, 2001.
- زیرنویس‌ها
-
- ¹ Cocktail Party Problem
- ² Computational Auditory Scene Analysis (CASA)
- ³ Model-Based Expectation-Maximization Source Separation and Localization (MESSL)
- ⁴ Interaural Phase Difference (IPD)
- ⁵ Interaural Level Difference (ILD)
- ⁶ Wavelet Denoising Technique
- ⁷ Adaptive Noise Cancellation (ANC)
- ⁸ Learning-Based Particle Swarm Optimization (LPSO)
- ⁹ Interaural Spectrogram
- ¹⁰ Interaural Time Difference (ITD)
- ¹¹ Phase Residual
- ¹² Missing Data Problem
- ¹³ Maximum Likelihood (ML)
- ¹⁴ Expectation Maximization (EM)
- ¹⁵ Hard Binary Mask
- ¹⁶ Soft Mask
- ¹⁷ *A Priori*
- ¹⁸ Marginal Class Membership
- ¹⁹ Decision-Directed
- ²⁰ Better Ear
- problem of frequency-domain blind source separation,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530-538, 2004.
- [12] Sh. Ikeda, and N. Murata, “A method of ICA in time-frequency domain,” *Proceedings of the ICA*, pp. 365-371, 1999.
- [13] M. Geravanchizadeh, and M. Hesam, “Convolutional ICA for audio signals,” *Independent Component Analysis for Audio and Biosignal Applications*, G. R. Naik, Ed. Intech, pp 137-162, 2012.
- [14] H. Buchner, R. Aichner, and W. Kellermann, “Trinicon: a versatile framework for multichannel blind signal processing,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, vol. 3, pp. 89-92, 2004.
- [15] D. L. Wang, and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.
- [16] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, “Model-based expectation-maximization source separation and localization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382-394, 2010.
- [17] B. Bahadornia, and M. Geravanchizadeh, “Enhanced binaural source separation using wavelet denoising technique,” *Proceedings of 21th Iranian Conference on Electric Engineering, ICEE*, 2013.
- [18] H. Gustafsson, S. E. Nordholm, and I. Claesson, “Spectral subtraction using reduced delay convolution and adaptive averaging,” *IEEE Transaction on Speech and Audio Processing*, 2001.
- [19] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction,” *Speech Comm.*, vol. 49, no. 7-8, pp. 636-656, 2007.
- [20] L. B. Asl, and M. Geravanchizadeh, “Dual-channel speech enhancement based on stochastic optimization strategies,” *Inf. Sci. Signal Process. their Appl., ISSPA, 10th Int. Conf.*, pp. 229-232, 2010.
- [21] V. Balakrishnan, N. Borges, and L. Parchment, *Wavelet Denoising and Speech Enhancement*, Spring 2006.
- [22] Y. Ephraim, and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP, vol. 23, no. 2, pp. 443-445, 1985.
- [23] P. Aarabi, “Self-localizing dynamic microphone arrays,” *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, pp. 474-484, 2002.
- [24] T. Nguyen, and G. Strang, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, 1996.
- [25] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 32, pp. 1109-1121, 1984.
- [26] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, 2006.
- [27] M. Mandel, *Binaural Model-based Source Separation and Localization*, Ph.D. Dissertation, Columbia University, 2010.
- [28] L. Di Persia, D. Milone, H. Rufiner, and M. Yanagida, “Perceptual evaluation of blind source separation for robust speech recognition,” *Signal Process.*, vol. 88, no. 10, pp. 2578-2583, 2008.
- [29] M. Bijankhan, J. Sheikhzadegan, M. R. Roohani, Y. Samareh, C. Lucas, and M. Tebiani, “The speech database