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

مسعود گراوانچی‌زاده[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 به‌منظور بهبود عملکرد سیستم جداکننده گفتار دوگوشی با نام MESSL ارائه می‌شود. در سیستم جداساز سیگنال، ابتدا، با استفاده از الگوریتم EM، مدل‌های گوسی پارامترهای اختلاف فاز درون‌گوشی (IPD) و اختلاف شدت درون‌گوشی (ILD) به دست می‌آیند. سپس، با استفاده از مدل به‌دست‌آمده برای هر منبع، ماسک نرمی استخراج شده که با ضرب آن در تبدیل فوریه زمان‌کوتاه (STFT) سیگنال مخلوط، سیگنال هدف جدا می‌شود. به‌علت عملکرد ناقص سیستم در امر جداسازی، دو سیستم پس‌پردازش به‌منظور حذف سیگنال‌های ناخواسته از سیگنال هدف، پیشنهاد می‌شود. روش پیشنهادی اول حذف وفقی نویز با استفاده از بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (LPSO) است. سیستم پس‌پردازش پیشنهادی دوم شامل دو مرحله است. در مرحله اول این سیستم، از روش حذف نویز تبدیل موجک به‌منظور حذف بخش اعظم سیگنال تداخل استفاده می‌شود. در مرحله دوم، روش حداقل میانگین مربعات خطا (MMSE) جهت ارتقاء هرچه بیش‌تر کیفیت سیگنال هدف جداشده به کار می‌رود. ارزیابی و مقایسه سیستم‌های پیشنهادی برای دادگان فارسی نشان می‌دهد که سیستم پیشنهادی دوم در بهبود کیفیت سیگنال هدف جداشده 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. Geravanchizadeh[1] Associate professor, P. Dadvar[2], PhD. Student, B. Bahadornia[3], 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 گراوانچی‌زاده
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 به‌صورت 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 [۵-۲]. ازجمله 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. درحالی‌که 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 [۷-۶]. 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 ازآنجایی‌که 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 ازجمله آن‌ها می‌توان به روش مقایسه پوش سیگنال‌های جداشده [۱۲-۱۰] 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، غیرایستان بودن و غیرگوسی بودن برای جداسازی استفاده می‌کند [۱۴]. سه محدودیت عمده روش‌های ذکرشده در بالا عبارت‌اند از: ۱- وضعیت شخص گوینده بایستی در طول فرآیند جداسازی ثابت بماند. به‌عبارتی، از مواردی مانند چرخش سر که بیانگر شرایط غیرایستان 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 ذکرشده 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 غیرایستان 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 بیشینه‌سازی 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 به‌دست‌آمده و با بررسی تعلق هر واحد زمان- فرکانسی (T-F) به هر یک از منابع، ماسک مربوط به هر منبع صوتی استخراج می‌شود. درنهایت، با اعمال ماسک استخراج‌شده به سیگنال مخلوط، منابع گفتار از یکدیگر جدا می‌شوند. عمل جداسازی به‌طور کامل در سیستم موردبحث اتفاق نمی‌افتد و اندکی از گفتار منابع غیرهدف نیز در سیگنال جداشده باقی می‌ماند. در [۱۷] یک سیستم پس‌پردازش دومرحله‌ای با استفاده از روش حذف نویز موجک[۶] و روش حداقل میانگین مربعات خطا (MMSE) برای بهبود کیفیت سیگنال‌های جداشده توسط روش MESSL ارائه شده است. بهبود کیفیت گزارش‌شده از ارزیابی‌های انجام‌شده برای دادگان TIMIT در این مرجع رضایت‌بخش است. هرچند روش‌های متعددی، ازجمله روش تفریق 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 ارائه‌شده، ولی از میان این روش‌ها حذف نویز بر اساس تبدیل موجک عملکرد بهتری در بهبود سیگنال‌های جداشده از خود نشان می‌دهد [۲۱]. در این مقاله، دو سیستم بهبود کیفیت گفتار به‌منظور حذف سیگنال‌های ناخواسته از سیگنال هدف جداشده توسط سیستم دوگوشی MESSL پیشنهاد و ارزیابی شده است. یکی از این سیستم‌ها روش بهبود گفتار دو کانالی ارائه‌شده 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 ارائه‌شده دوم بنام

WT_MMSE شامل دو مرحله است. ابتدا، با استفاده از روش حذف نویز موجک بر روی سیگنال هدف جداشده، سیگنال‌های ناخواسته باقی‌مانده از مرحله جداسازی دوگوشی حذف می‌شود. در مرحله بعد، برای افزایش کیفیت سیگنال جداشده هدف، از روش حداقل میانگین مربعات خطا (MMSE) [۲۲] استفاده می‌شود. سیستم‌های پس‌پردازش ارائه‌شده 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

بلوک‌دیاگرام پیشنهادی به‌منظور 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 نقطه‌چین مشخص شده است.

۲-۱- تعریف مدل

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

$$\begin{aligned} \ell(t) &= s(t-\tau_\ell) * h_\ell(t) \\ r(t) &= s(t-\tau_r) * h_r(t), \end{aligned} \quad (۱)$$

که در آن $h_\ell(t)$ و $h_r(t)$، به ترتیب، پاسخ ضربه محیط برای سیگنال‌های دریافتی از منبع توسط گوش‌های چپ و راست می‌باشند. نسبت تبدیل فوریه زمان‌کوتاه (STFT) دو سیگنال رسیده به گوش‌های چپ و راست، به‌عنوان طیف‌نگاره درون‌گوشی[۹]، با روابط زیر بیان می‌شود:

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \quad (۲)$$

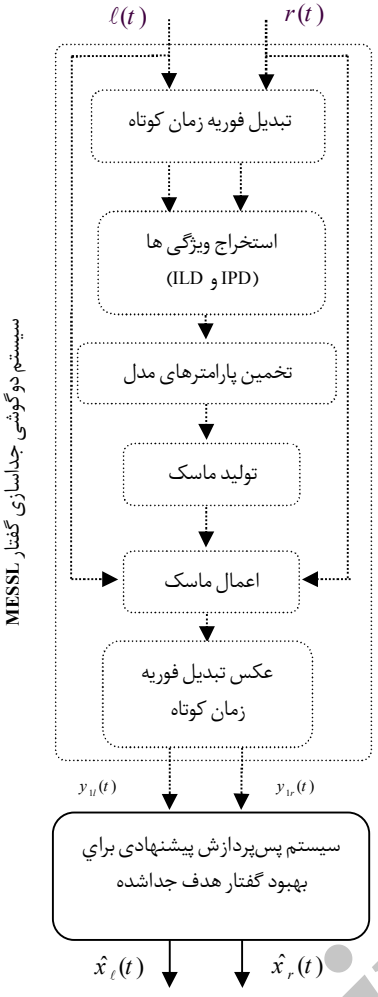
$$\approx 10^{\alpha(\omega)/20} e^{-j\omega\tau(\omega)} \quad (۳)$$

به‌طوری‌که:

$$H(\omega) = \mathcal{F}[h_\ell(t)] / \mathcal{F}[h_r(t)] \quad (۴)$$

$$\tau(\omega) = \tau_\ell - \tau_r + \angle H(\omega) \quad (۵)$$

$$\alpha(\omega) = 20\log_{10}|H(\omega)|. \quad (۶)$$



شکل ۱: بلوک‌دیاگرام روش پیشنهادی جهت ارتقای عملکرد سیستم دوگوشی جداسازی گفتار **MESSL**. سیستم جداسازی دوگوشی گفتار **MESSL** با نقطه‌چین و سیستم پس‌پردازش پیشنهادی با خط پررنگ مشخص شده است.

در روابط بالا، $\mathcal{F}$ بیانگر عملگر تبدیل فوریه، $\alpha(\omega)$ اختلاف شدت درون‌گوشی وابسته به فرکانس، τ_ℓ، τ_r، تأخیرهای مستقل از فرکانس گوش‌های چپ و راست و $\tau(\omega)$ اختلاف زمانی درون‌گوشی وابسته به فرکانس هستند. روابط (۲) و (۳)، به ترتیب، به‌عنوان مدل واقعی و مدل تقریبی مشاهدات طیف‌نگاره درون‌گوشی محسوب می‌شوند. طیف‌نگاره درون‌گوشی 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 درون‌گوشی (IPD) به اختلاف زمان درون‌گوشی (ITD)[۱۰]، مقادیر نماینده ITD (یا τ) به فضای IPD نگاشت می‌شوند تا τای را که با مشاهدات بیش‌ترین تطابق را دارد، پیدا شود. بنابراین، به جای مدل کردن فاز واقعی، فاز باقی‌مانده[۱۱] $\hat{\phi}(\tau)$، یعنی اختلاف بین فاز واقعی و فاز تخمین‌زده‌شده توسط τ، به‌صورت زیر مدل می‌شود:

$$\hat{\phi}(\omega,t;\tau)=\arg(e^{j\phi(\omega,t)}e^{-j\omega\tau(\omega)}), \quad (۷)$$

به‌طوری‌که همواره $\hat{\phi}(\omega,t;\tau)$ در بازه $[-\pi,\pi]$ قرار دارد.

برای مدل کردن IPD و ILD، به ترتیب، از توزیع گوسی به‌صورت زیر استفاده می‌شود [۱۶]:

$$\begin{aligned} p\left(\phi(\omega,t)|\tau(\omega),\sigma(\omega)\right)&=N\left(\hat{\phi}(\omega,t;\tau(\omega))\middle|0,\sigma^2(\omega)\right)\\ &\approx N\left(\phi(\omega,t)\middle|\omega\tau(\omega),\sigma^2(\omega)\right), \end{aligned} \quad (۸)$$

$$p\left(\alpha(\omega,t)\middle|\mu(\omega),\eta^2(\omega)\right)=N\left(\alpha(\omega,t)\middle|\mu(\omega),\eta^2(\omega)\right), \quad (۹)$$

که در آن $\mu(\omega)$ 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 به‌صورت زیر بیان می‌شود:

$$\begin{aligned} p\left(\phi(\omega,t)\cdot\alpha(\omega,t)|\Theta\right)=&N\left(\hat{\phi}(\omega,t)|\xi(\omega),\sigma^2(\omega)\right)\\ &\cdot N\left(\alpha(\omega,t)\middle|\mu(\omega),\eta^2(\omega)\right), \end{aligned} \quad (۱۰)$$

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 داده‌شده در بخش ۲-۱ را نمی‌توان به‌صورت مستقیم از سیگنال مخلوط به دست آورد، چون به دلیل توزیع متفاوت منابع بر روی IPD و ILD، نواحی مختلف طیف‌نگاره درون‌گوشی توسط منابع مختلف اشغال شده است. در اینجا فرض می‌شود که تنها واحدهای T-F مربوط به یک منبع که دارای تأخیرهای برابری هستند، به‌طور یکسان توزیع شده‌اند. بنابراین پارامترهای مدل، تنها زمانی می‌توانند تخمین زده شوند که معلوم شود واحدهای T-F طیف‌نگاره درون‌گوشی متعلق به کدامین تأخیر و کدامین منبع هستند. این نوع روش جداسازی، در حقیقت مسئله داده‌های گم‌شده[۱۲] است که پارامترهای احتمال بیشینه (ML)[۱۳] آن توسط الگوریتم EM[۱۴] استخراج می‌شوند. الگوریتم EM برای هر یک از منابع در سیگنال مخلوط، محل‌هایی از طیف‌نگاره درون‌گوشی 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 به‌صورت زیر مدل می‌شود [۱۶]:

$$\tau(\omega)=\tau+\omega^{-1}\xi(\omega). \quad (۱۱)$$

در رابطه (۱۱)، τ تأخیر مستقل از فرکانس است که در مکان‌یابی مورد استفاده قرار می‌گیرد. برای آن‌که سیگنال‌های تداخل نیز قابل مسیریابی باشند، متغیر پنهان τ را به‌صورت متغیر تصادفی گسسته و به‌صورت مجموعه‌ای از تأخیرهای مجاز پیش‌فرض[۱۷] در نظر می‌گیریم. پارامتر $\xi(\omega)$، مقدار افست 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، به‌نحوی‌که مقدار این متغیر، در صورتی‌که نقطه (ω,t) از منبع i و تأخیر τ باشد، یک و در غیر این‌صورت، صفر است. هر واحد T-F از طیف‌نگاره باید به منبع و تأخیری تعلق داشته باشد که درنتیجه 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 بیشینه‌سازی الگوریتم EM محاسبه می‌شوند. ازآنجایی‌که τها 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، درصورتی‌که 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 $\Theta\equiv\{\xi_{i\tau}(\omega),\sigma_{i\tau}(\omega),\mu_i(\omega),\eta_i(\omega),\psi_{i\tau}\}$ به‌عنوان بردار مشاهدات و با تعریف $N_{\hat{\phi}}=N\left(\hat{\phi}(\omega,t;\tau)\middle|\xi_{i\tau}(\omega),\sigma^2_{i\tau}(\omega)\right)$ و $N_\alpha=N\left(\alpha(\omega,t)\middle|\mu_i(\omega),\eta_i^2(\omega)\right)$، می‌توان احتمال لگاریتمی کل را به‌صورت رابطه (۱۲) تعریف کرد:

$$\begin{aligned} \Gamma(\Theta)&=\sum_{\omega,t}\log p\left(\phi(\omega,t),\alpha(\omega,t)|\Theta\right)\\ &=\sum_{\omega,t}\log\sum_{i,\tau}\left[N_{\hat{\phi}}.N_\alpha.\psi_{i\tau}\right]. \end{aligned} \quad (۱۲)$$

در این رابطه، به‌تعداد $N_i\times N_\tau$ 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 $\psi_{i\tau}$ به‌عنوان ثابت وزن‌دهی ترکیب‌کننده 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 بیشینه‌کردن Θ تعریف کرد:

$$\begin{aligned} Q\left(\Theta|\Theta_s\right)=&\\ k+\sum_{\omega,t}\sum_{i,\tau}&\Big[p\left(z_{i\tau}(\omega,t)|\phi(\omega,t),\alpha(\omega,t),\Theta_s\right)\\ &\cdot\log p\left(z_{i\tau}(\omega,t),|\phi(\omega,t),\alpha(\omega,t)|\Theta\right)\Big], \end{aligned} \quad (۱۳)$$

به‌طوری‌که Θ_s تخمین پارامترهای Θ بعد از s تکرار و ثابت k نیز مستقل از Θ است.

در مرحله امید ریاضی (E)، متغیر پنهان $z_{i\tau}(\omega,t)$ به‌شرط مشاهدات (ϕ و α) و پارامتر Θ_s محاسبه می‌شود و در مرحله بیشینه‌سازی (M)، مقدار بیشینه Q نسبت به Θ و امید ریاضی $z_{i\tau}(\omega,t)$ به دست می‌آید. در مرحله E، احتمال حضور منبع i و تأخیر τ در واحد T-F را به‌صورت زیر خواهیم داشت:

$$\begin{aligned}\upsilon_{i\tau}(\omega,t) &\equiv p\left(z_{i\tau}(\omega,t)\middle|\phi(\omega,t),\alpha(\omega,t),\Theta_s\right)\\ &\propto p\left(z_{i\tau}(\omega,t),\phi(\omega,t),\alpha(\omega,t)\middle|\Theta_s\right)\\ &=\psi_{i\tau}\cdot N_{\phi}\cdot N_{\alpha}.\end{aligned} \quad (۱۴)$$

با استفاده از تعریف عملگر زیر به‌عنوان میانگین وزن‌دار:

$$\langle x\rangle_{t,\tau}\equiv\frac{\sum_{t,\tau}x\upsilon_{i\tau}(\omega,t)}{\sum_{t,\tau}\upsilon_{i\tau}(\omega,t)} \quad (۱۵)$$

در مرحله M، پارامترهای مدل به‌صورت زیر محاسبه می‌شوند:

$$\mu_i(\omega)=\langle\alpha(\omega,t)\rangle_{t,\tau} \quad (۱۶)$$

$$\eta_i^2(\omega)=\left\langle\left(\alpha(\omega,t)-\mu_i(\omega)\right)^2\right\rangle_{t,\tau} \quad (۱۷)$$

$$\xi_{i\tau}(\omega)=\left\langle\hat{\phi}(\omega,t;\tau)\right\rangle_t \quad (۱۸)$$

$$\sigma_{i\tau}^2(\omega)=\left\langle\left(\hat{\phi}(\omega,t;\tau)-\xi_{i\tau}(\omega)\right)^2\right\rangle_t \quad (۱۹)$$

$$\psi_{i\tau}=\frac{1}{\Omega T}\sum_{\omega,t}\upsilon_{i\tau}(\omega,t). \quad (۲۰)$$

۲-۳- تخمین ماسک جداسازی

می‌توان با جمع بر روی مقادیر τ، ماسک موردنظر برای هر منبع را به‌صورت زیر به دست آورد:

$$M_i(\omega,t)\equiv\sum_{\tau}\upsilon_{i\tau}(\omega,t). \quad (۲۱)$$

ماسک به‌دست‌آمده، احتمال تعلق هر یک از واحدهای T-F در طیف‌نگاره را به منبع iام نشان می‌دهد [۱۶].

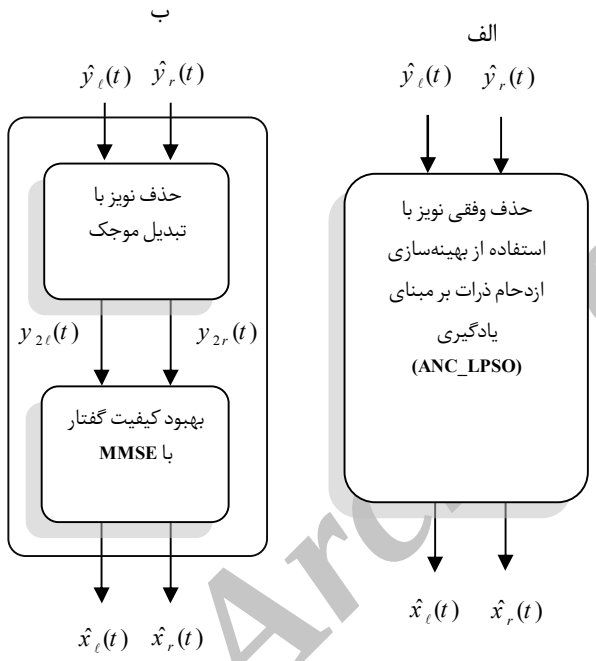
با اعمال ماسک به‌دست‌آمده بر روی تبدیل فوریه سیگنال‌های مخلوط گوش چپ و راست و سپس، انجام عمل عکس تبدیل فوریه، سیگنال هدف از سیگنال مخلوط ورودی استخراج می‌شود.

۳- سیستم‌های پیشنهادی جهت بهبود کیفیت سیگنال هدف جداشده توسط سیستم دوگوشی MESSL

به علت باقی ماندن بخشی از سیگنال گفتار منابع غیرهدف در سیگنال گفتار جداشده منبع هدف، دو سیستم پس‌پردازش بهبود کیفیت گفتار برای ارتقاء عملکرد سیستم جداسازی دوگوشی MESSL پیشنهاد می‌شود. بلوک‌دیاگرام 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 می‌شود. این مرحله از بهبود کیفیت گفتار، با درصد بالایی منابع گفتار ناخواسته را حذف می‌کند. در مرحله دوم، روش حداقل میانگین مربعات خطا (MMSE) جهت ارتقاء هرچه بیش‌تر کیفیت سیگنال جداشده هدف به کار می‌رود. بلوک‌دیاگرام سیستم‌های پس‌پردازش پیشنهادی در شکل ۲ آمده است.

۳-۱- حذف وفقی نویز با استفاده از بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (ANC_LPSO)

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



شکل ۲: بلوک‌دیاگرام سیستم‌های پس‌پردازش پیشنهادی برای ارتقاء عملکرد سیستم دوگوشی جداسازی گفتار MESSL؛ الف) روش حذف وفقی نویز با استفاده از بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (ANC_LPSO) [۲۰]، ب) روش دومرحله‌ای حذف نویز بر اساس تبدیل موجک و MMSE (WT_MMSE).

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

سیستم پس‌پردازش 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 می‌شود.

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، آستانه‌گذاری ضرایب و دوباره‌سازی سیگنال با استفاده از ضرایب فیلترشده. در ابتدا، تحلیل موجک را تا مرحله $N = ۱۲$ با موجک Daubechies 12 انجام می‌دهیم. سپس، ضرایب مربوط به موجک را آستانه‌گذاری می‌کنیم. درنهایت، دوباره‌سازی سیگنال را با استفاده از ضرایب آستانه‌گذاری‌شده خواهیم داشت.

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

$$Y_{1\,j,k} = X_{j,k} + N_{j,k}, \tag{۲۲}$$

به‌طوری‌که $Y_{1\,j,k}$، $X_{j,k}$ و $N_{j,k}$، به ترتیب، kامین ضرایب موجک در مقیاس jام مربوط به سیگنال‌های $y_1(t)$، $x(t)$ و $n(t)$ می‌باشند [۲۱].

با فرض گوسی‌بودن نویز و توزیع یکنواخت آن روی ضرایب موجک، می‌توان مشخصات نویز را از ضرایب موجک آن استخراج کرد. مرحله آستانه‌گذاری به دو صورت انجام می‌شود: آستانه‌گذاری متفاوت در هر مرحله و آستانه‌گذاری یکسان برای کل مراحل. در این مقاله، از یک مقدار ثابت برای آستانه‌گذاری در تمام مراحل استفاده می‌کنیم. برای تعیین مقدار آستانه δ از روش Donoho [۲۵] بهره می‌بریم که به‌صورت زیر تعریف می‌شود:

$$\delta = \sigma\sqrt{2\log L}. \tag{۲۳}$$

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

برای تخمین σ، به‌صورت زیر عمل می‌کنیم. با فرض گوسی‌بودن نویز، تابع چگالی احتمالی (pdf) نویز را به‌صورت زیر خواهیم داشت [۲۱]:

$$\text{pdf}(\hat{N}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\hat{N}^2}{2\sigma^2}}, \tag{۲۴}$$

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 به‌صورت زیر خواهد بود:

$$\text{pdf}(|\hat{N}|) = u(\hat{N})\,\frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{\hat{N}^2}{2\sigma^2}}, \tag{۲۵}$$

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

$$\int_{-\infty}^{\text{median}(|\hat{N}|)} \text{pdf}(x)\,dx = \frac{1}{2}. \tag{۲۶}$$

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

$$\int_{0}^{\text{median}(|\hat{N}|)} \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{\hat{N}^2}{2\sigma^2}} d\hat{N} = \frac{1}{2}. \tag{۲۷}$$

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

$$\text{median}(|\hat{N}|) = \sigma\sqrt{2}\,\text{erf}^{-1}\left(\frac{1}{2}\right) \approx 0.6745\,\sigma, \tag{۲۸}$$

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

$$\sigma = \frac{\text{median}(|Y_1|)}{0.6745}. \tag{۲۹}$$

مرحله آستانه‌گذاری خود به دو روش انجام می‌گیرد: آستانه‌گذاری سخت و آستانه‌گذاری نرم. در روش آستانه‌گذاری سخت تمام ضرایب کوچک‌تر از آستانه δ مقدار صفر می‌گیرند.

$$Y_{1\,j,k}^{Hard} = \begin{cases} Y_{1\,j,k}, & |Y_{1\,j,k}| > \delta \\ 0, & \text{otherwise.} \end{cases} \tag{۳۰}$$

آستانه‌گذاری نرم یک مرحله نیز بیش‌تر پیش می‌رود و دامنه ضرایب را به اندازه δ کاهش می‌دهد.

$$Y_{1\,j,k}^{Soft} = \text{sign}(Y_{1\,j,k})\,\max\left(|Y_{1\,j,k}| - \delta, 0\right). \tag{۳۱}$$

روش آستانه‌گذاری سخت اگرچه دامنه سیگنال را حفظ می‌کند، ولی به علت عدم پیوستگی در ضرایب موجک بعد از دوباره‌سازی سیگنال، اندکی نویز را وارد سیگنال می‌کند، درحالی‌که روش نرم با حذف این گسستگی در ضرایب موجک، سیگنال نرم‌تر، ولی با دامنه کاهش‌یافته، را به دست می‌دهد. در این مقاله، از روش آستانه‌گذاری نرم برای ضرایب تبدیل موجک سیگنال نویزی استفاده می‌شود.

۳-۲-۲- حذف نویز به‌روش MMSE

با وجود اینکه روش حذف نویز با موجک تا حد قابل‌قبولی سیگنال‌های مربوط به منابع تداخل را حذف می‌کند، اندکی سیگنال ناخواسته در خروجی باقی می‌ماند. برای ارتقاء کیفیت سیگنال هدف جداشده توسط سیستم جداسازی دوگوشی، می‌توان از روش بهبود کیفیت گفتار MMSE در مرحله دوم سیستم پس‌پردازش پیشنهادی بهره برد [۲۵]. سیگنال رسیده به مرحله بهبود کیفیت گفتار با MMSE در حوزه فوریه به‌صورت $Y_{2k}e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)}$ نشان داده می‌شود که در آن $\{Y_{2k}, X_k, D_k\}$ اندازه و $\{\theta_y(k), \theta_x(k), \theta_d(k)\}$ فاز در نقطه فرکانسی kام، به ترتیب، برای سیگنال گفتار نویزی، سیگنال تمیز (سیگنال هدف) و سیگنال نویز می‌باشد. در مرحله بعد، یک تخمین‌گر

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

$$E\left\{\left(\log X_k - \log \hat{X}_k\right)^2\right\}. \quad (۳۲)$$

تخمین‌گر بهینه اندازه log-MMSE را می‌توان به‌صورت زیر به دست آورد:

$$\hat{X}_k = \frac{\xi_k}{1+\xi_k}\exp\left\{\frac{1}{2}\int_{\upsilon_k}^{\infty}\frac{e^{-t}}{t}dt\right\}Y_k \quad (۳۳)$$
$$= G_{LSA}(\xi_k, \upsilon_k)Y_{2k},$$

که در آن $G_{LSA}(\xi_k, \upsilon_k)$ تابع بهره تخمین‌گر log-MMSE است. متغیر υ_k به‌صورت زیر تعریف می‌شود [۲۵]:

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \quad (۳۴)$$

که در آن متغیرهای ξ_k و γ_k، به ترتیب، SNR پیشین و پسین بوده و به‌صورت زیر تعریف می‌شوند:

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}, \quad (۳۵)$$

$$\gamma_k = \frac{Y_k^2}{\lambda_d(k)}, \quad (۳۶)$$

که در آن $\lambda_d(k)$ واریانس نویز است. برای تشکیل طیف سیگنال تمیز $\hat{X}(\omega_k) = \hat{X}_k \exp\left(j\hat{\theta}_x(k)\right)$، نیاز به تخمین فاز نیز است. تخمین‌گر MMSE فاز به‌صورت مسئله کمینه‌سازی مشروط زیر به دست می‌آید:

$$\min_{\exp(j\hat{\theta}_x)} = E\left\{\left|\exp(j\theta_x) - \exp(j\hat{\theta}_x)\right|^2\right\}, \quad (۳۷)$$
$$\text{به شرط:} \quad \left|\exp\left(j\hat{\theta}_x\right)\right| = 1.$$

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

$$\exp(j\hat{\theta}_x) = \exp(j\theta_y). \quad (۳۸)$$

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

تخمین‌گر اندازه log-MMSE در رابطه (۳۳)، با این فرض به‌دست آمد که SNR پیشین ξ_k و واریانس نویز $\lambda_d(k)$ معلوم هستند. ازآنجاکه در عمل، فقط به سیگنال آلوده دسترسی وجود دارد، می‌توان واریانس نویز را با فرض ایستان‌بودن نویز توسط روش آشکارساز فعالیت نویز (VAD) در بازه‌های عدم حضور گفتار به دست آورد. برای محاسبه SNR پیشین ξ_k، از روش همسو با تصمیم[۱۹] استفاده می‌شود [۲۲].

$$\hat{\xi}_k(m) = a\frac{\hat{X}_k^2(m-1)}{\lambda_d(k, m-1)} + (1-a)\max\left[\gamma_k(m) - 1, 0\right], \quad (۳۹)$$

به‌طوری‌که $0 < a < 1$، ثابت وزن‌دهی و $\hat{X}_k^2(m-1)$ تخمین‌گر دامنه در فریم قبلی است. بهترین نتیجه ممکن برای a=۰/۹۸ به دست می‌آید.

۴- نتایج آزمایش‌ها

برای سنجش عملکرد سیستم‌های 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 می‌شود [۲۶]:

$$\mathrm{SDR} = 10\log_{10}\frac{\|x_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2}, \quad (۴۰)$$

$$\mathrm{SIR} = 10\log_{10}\frac{\|x_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}, \quad (۴۱)$$

$$\mathrm{SAR} = 10\log_{10}\frac{\|x_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2}, \quad (۴۲)$$

که در آن x_{target} سیگنال هدف، e_{interf} سیگنال تداخل و e_{artif} سیگنال‌های مصنوعی ایجادشده هستند. از میان معیارهای ذکرشده، معیار SDR مطابقت بالایی با نتایج تشخیص خودکار گفتار (ASR) دارد و می‌تواند به‌عنوان معیار اندازه‌گیری 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].

فایل‌های صوتی مورداستفاده در آزمایش‌ها، اعم از سیگنال‌های هدف و تداخل، از پایگاه داده فارس‌دات (FARSDAT) انتخاب شده‌اند [۲۹]. FARSDAT یک پایگاه داده گفتار فارسی پیوسته است که از ۶۰۰ گفته و از ۳۰۰ گوینده با لهجه‌های مختلف فارسی موجود در ایران تشکیل شده است. این جملات از یک مجموعه ۳۹۰ جمله‌ای توسط گوینده‌های مختلف خوانده شده است. پانزده جمله از جملات فوق به شکل تصادفی انتخاب شده و در ترکیب با سیگنال‌های تداخل به کار رفته است. برای پاسخ ضربه اتاق از داده‌های موجود در مرجع [۳۰] استفاده شده است. فرکانس کاری مراجع [۳۰-۲۹] ۴۴۱۰۰ هرتز است که با نمونه‌برداری مجدد مقدار آن به ۱۶ کیلوهرتز کاهش یافته است. طول فریم‌ها در تحلیل STFT، ۶۴ میلی‌ثانیه و مقدار همپوشانی آن‌ها ۱۶ میلی‌ثانیه در نظر گرفته شده است. برای پنجره‌بندی سیگنال از پنجره hamming استفاده شده است. در شبیه‌سازی‌های انجام‌گرفته، عملکرد الگوریتم‌ها در حالت‌های متفاوتی بررسی شده است. در جدول ۱، شرایط مختلف شبیه‌سازی‌ها 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 می‌کنند. اندیس‌های i و j مقادیر 0، 1 و یا Ω را به خود می‌گیرند که به ترتیب، معرف ساده‌ترین حالت، حالت پیچیده‌تر و حالت با بیش‌ترین پیچیدگی هستند.

جدول ۱: شرایط مختلف شبیه‌سازی برای پارامتر حالت θ. **m** میانگین 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	$_\infty$	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$
$\theta_{\Omega 0}$	$\mu_i(\omega)$	$\eta_i(\omega)$	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، همه 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[۲۰]، 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 به‌میزان چشمگیری افزایش می‌دهد.

جدول ۲: میانگین مقادیر **SDR**، **SIR**، **SAR** 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 **ANC_LPSO** و **WT_MMSE**

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}					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}	-۱/۱۷	۲۱/۸۲	-۱/۰۸	۱/۳۵	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 **SDR**، **SIR**، **SAR** 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}					-۳/۳۴	۱۳/۹۸	-۳/۰۹	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}	-۵/۷۵	۱۳/۴۱	-۵/۵	۱/۰۸	-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 **SDR**، **SIR**، **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

۵- نتیجه‌گیری

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 جداشده توسط این سیستم ارائه شده است. یک سیستم، حذف وفقی نویز با استفاده از بهینه‌سازی ازدحام ذرات بر مبنای یادگیری (ANC_LPSO) و سیستم پیاده‌سازی‌شده 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.

مراجع

[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

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, "Convolutive 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 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 Applicat. 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