استخراج نقشه‌های برجستگی وزن‌دار در مدل‌سازی توجه پایین به بالای شنوایی

مسعود گراوانچی زاده[۱]، دانشیار، سپیده اختری خسروشاهی[۲]، دانشجوی کارشناسی ارشد، سحر ذاکری[۳]، دانشجوی دکترا

۱- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران – geravanchizadeh@tabrizu.ac.ir
۲- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران – s.akhtari96@ms.tabrizu.ac.ir
۳- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران – s.zakeri@tabrizu.ac.ir

چکیده: شنوایی یکی از قسمت‌های مهم زندگی روزانه انسان‌ها است. با وجود اینکه انسان‌ها در معرض صداهای مختلف از منابع متفاوت قرار دارند و تعداد گیرنده‌های سیستم عصبی برای تجزیه و تحلیل این صداها محدود هستند، انسان‌ها می‌توانند مخلوط‌های شنیداری پیچیده را به خوبی پردازش کنند. یکی از دلایل این توانایی انسان، پدیده توجه است. توجه شنوایی را می‌توان به دو دسته توجه پایین به بالا و توجه بالا به پایین تقسیم‌بندی کرد. در این مقاله، مدلی برای شبیه‌سازی توجه پایین به بالا با استفاده از نقشه‌های برجستگی وزن‌دار، در سیستم شنوایی ارائه شده است. دادگان به کار رفته در این پژوهش از ترکیب نویزهای پس زمینه مختلف با صوت‌های موجود در پایگاه دادگان ESC به عنوان قسمت‌های برجسته، در SNR های متفاوت بدست آمده است. برای ارزیابی مدل، از معیار میانگین خطا استفاده شده است که بصورت اختلاف زمانی بین نقطه برجسته واقعی و نقطه برجسته تشخیص داده شده توسط مدل تعریف می‌شود. ترکیب وزن‌دار نقشه‌های آشکار حاصل از ویژگی‌ها، با استفاده از الگوریتم ژنتیک، سبب شده است که مدل پیشنهادی با خطای متوسط ۰/۹۲ ثانیه عملکرد بهتری را نسبت به مدل پایه با خطای متوسط ۱/۹۱ ثانیه داشته باشد.

واژه‌های کلیدی: مدل‌سازی شنوایی توجه، توجه پایین به بالا، نقشه برجستگی، الگوریتم ژنتیک.

Extraction of Weighted Saliency Maps in Modelling Bottom-Up Auditory Attention

Masoud Geravanchizadeh, Associate Professor[1], Sepideh Akhtari Khosroshahi, Master Student[2], Sahar Zakeri, PhD Student[3]

1- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: geravanchizadeh@tabrizu.ac.ir
2- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: s.akhtari96@ms.tabrizu.ac.ir
3- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: s.zakeri@tabrizu.ac.ir

Abstract: Hearing is an important part of human daily life. Although humans are exposed to various sounds from different sources and the numbers of receptors of the neural system are limited, they can process complex auditory scenes well. One of the reasons for this human ability is the phenomenon of attention. Auditory attention can be divided into two categories: bottom-up attention and top-down attention. In this paper, a model for simulating the bottom-up attention using weighted saliency maps in the auditory system is proposed. The dataset in this research work is obtained by combining different background noises with the sounds in the ESC database as salient regions, at different SNRs. To evaluate the model, the mean-error criterion was used, which is defined as the time difference between the actual salient point and the salient point detected by the model. The weighted combination of the conspicuity maps of the features using the Genetic algorithm makes the proposed model with an average error of 0.92 seconds to perform better than the baseline model having an average error of 1.91 seconds.

Keywords: Auditory Attention Modelling, Bottom-Up Attention, Saliency Map, Genetic Algorithm.

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

۱- مقدمه

مغز عضوی بسیار پیشرفته برای پردازش اطلاعات است و بخش بزرگی از توانایی تجزیه و تحلیل آن به پردازش ورودی‌های حسّی مربوط به سیستم بینایی و شنوایی اختصاص دارد. فرآیندهای عصبی پیچیده‌ای باعث می‌شوند که مغز اطلاعات موجود در محیط را پردازش کند و وقایع مهم محیط اطراف شناخته شوند. در میان این فرآیندها، پدیده توجه[۱]، با تمرکز بر روی منابع حسّی و شناختی نقش مهمی را در ادراک محیط اطراف بازی می‌کند. توجه را می‌توان نوعی پردازش اطلاعات دانست که به طور مداوم از ورودی‌های حسّی نمونه‌برداری می‌کند و زیرمجموعه‌ای از اطلاعات حسّی ورودی را جهت پردازش انتخاب می‌کند [۱]. این انتخاب ترکیبی از توجه پایین به بالا[۲] و توجه بالا به پایین[۳] است. ابتدا، فرآیند سریع پایین به بالا سیگنال ورودی را پردازش می‌کند تا بر اساس ویژگی‌های بدست آمده بصورت ناخودآگاه توجه را به سمت نقاط برجسته[۴] جلب کند. سپس، فرآیند بالا به پایین بصورت ارادی باعث جلب توجه به قسمت‌هایی می‌شود که از نظر شناختی آشنا هستند [۲].

اوّلین مدل محاسباتی توجه پایین به بالا در زمینه بینایی در سال ۱۹۸۷ [۳] معرفی شد و توسط ایتی و همکاران [۴] بهبود یافت. اساس این مدل بر پایه تئوری ادغام ویژگی‌ها است که در آن ویژگی‌های مختلفی از تصویر ورودی استخراج می‌شوند و از ادغام آن‌ها نقشه برجستگی[۵] بدست می‌آید. این نقشه، مشخص‌کننده قسمت‌های برجسته و مهم تصویر است که باعث جلب توجه می شوند. در زمینه شنوایی، مدل‌های توجه پایین به بالا بسیار کم مورد بررسی قرار گرفته‌اند. پژوهش‌های محدود در این زمینه از نمونه‌های مشابه در زمینه بینایی اقتباس شده‌اند. نخستین مدل توجه پایین به بالا در زمینه شنوایی توسط کایزر و همکاران [۵] معرفی شده است. در این پژوهش، فرآیند پیدا کردن نقطه برجسته مشابه مدل [۴]، روی نمایش زمانی-فرکانسی، به عنوان تصویر ورودی، اعمال شده است. این مدل قادر به مطابقت با نتایج تجربی در محرک‌های برجسته ساده، مانند تشخیص صدای تُن در پس‌زمینه نویزی، است. کالینلی و نارایانان [۶] با اضافه کردن دو ویژگی دیگر بر روی مدل قبلی و تغییر روش نرمال‌سازی مدل جدیدی را معرفی کرده‌اند. دانگودم و اندرسون [۷] برای استخراج ویژگی‌ها از صوت ورودی، ساز و کارهای زیست‌شناختی قابل قبول‌تری که اصول پردازش در سیستم شنوایی محیطی و مرکزی را تقلید می‌کنند، به کار برده‌اند. آن‌ها از ویژگی‌های مدولاسیون طیفی-زمانی مشابه با شبیه‌سازی پاسخ عصبی در قشر شنوایی پستانداران، استفاده کرده‌اند. هرچند این مدل از نظر شبیه‌سازی عملکرد بیولوژیکی سیستم شنوایی مهم است، اما از نظر نتایج ارائه شده نسبت به سایر مدل‌ها دارای برتری نیست.

با توجه به اهمیت سیستم شنوایی در شناخت محیط اطراف و نقش مهم فرآیند توجه در پردازش ورودی‌های حسی، هدف این مقاله، ارائه مدل کارآمدتری برای شبیه سازی توجه پایین به بالا در سیستم شنوایی با استفاده از ویژگی‌های بدست آمده از صوت ورودی و ترکیب وزن‌دار نمودهای حاصل از این ویژگی‌ها، جهت استخراج نقشه برجستگی است.

۲- روش پیشنهادی برای بهبود نقشه برجستگی شنوایی

شکل (۱) بلوک‌دیاگرام روش پیشنهادی جهت محاسبه نقشه‌های برجستگی وزن‌دار در مدل‌سازی توجه پایین به بالای سیستم شنوایی را نشان می‌دهد. در این مقاله، از مدل پایه ارائه شده توسط کایزر و همکاران [۵] برای بدست آوردن نقشه برجستگی شنوایی استفاده شده است. طی زیر بخش‌های بعدی جزئیات سیستم پیشنهادی توصیف می‌شود.

۲-۱- پردازش شنوایی

به منظور مدل‌سازی سیستم شنوایی انسان، از نمایش زمانی-فرکانسی صوت ورودی استفاده می‌شود. غشاء پایه[۶] بصورت فیلتربانکی مدل می‌شود که در آن هر فیلتر، پاسخ فرکانسی یک نقطه مشخص از غشاء را بیان می‌کند. فیلتر گاماتون به عنوان مدلی برای پاسخ ضربه فیبرهای عصب شنوایی معرفی شده است. گاماتون فیلتری میان‌گذر است که پاسخ ضربه آن بصورت رابطه (۱) بیان می‌شود [۸].

$$g_{f_c}(t) = t^{N-1}\exp\left[-2\pi t b(f_c)\right]\cos(2\pi f_c t + \theta)u(t). \qquad (۱)$$

در رابطه (۱)، N مرتبه فیلتر، f_c فرکانس مرکزی فیلتر بر حسب هرتز، θ فاز، $u(t)$ پاسخ پله واحد و $b(f_c)$ پهنای باند فرکانس مرکزی داده شده است. پژوهش‌ها نشان داده است که فیلتر گاماتون به ازای مرتبه ۴ تناسب بسیار خوبی با تخمین‌های تجربی بدست آمده از شکل فیلتر شنیداری انسان دارد [۹].

ابتدا، سیگنال ورودی در حوزه فرکانس توسط بانکی متشکل از ۶۴ فیلتر گاماتون تجزیه می شود. فرکانس مرکزی این فیلترها روی مقیاس نرخ پهنای باند مستطیلی[۷] از ۸۰ تا ۵۰۰۰ هرتز توزیع شده است. در حوزه زمان، خروجی هر کانال به بازه‌های زمانی ۲۰ میلی‌ثانیه‌ای با جابجایی فریم به میزان ۱۰ میلی‌ثانیه تجزیه می شود. سپس، خروجی هر کانال از واحدی که سلول‌های مویی داخلی را شبیه سازی می‌کند [۱۰]، عبور داده می‌شود تا یک نمایش زمانی-فرکانسی از صوت ورودی بدست آید.

۲-۲- نقشه برجستگی شنوایی

سیستم شنوایی صداها را در یک مخلوط شنیداری پیچیده بر اساس ویژگی‌هایی مانند مدولاسیون فرکانس یا مدولاسیون زمان جدا می‌کند. این نوع ویژگی‌ها انسان را قادر می‌سازد تا در یک محیط شلوغ و پر سر و صدا، صداهای مورد علاقه را کشف کند.

نقشه برجستگی شنوایی بطور موازی این ویژگی‌ها را از مخلوط صوتی استخراج می‌کند که این کار بیانگر تحلیل ویژگی‌های صوت توسط نورون‌های قشر شنوایی است. در روش پیشنهادی، نقشه‌های

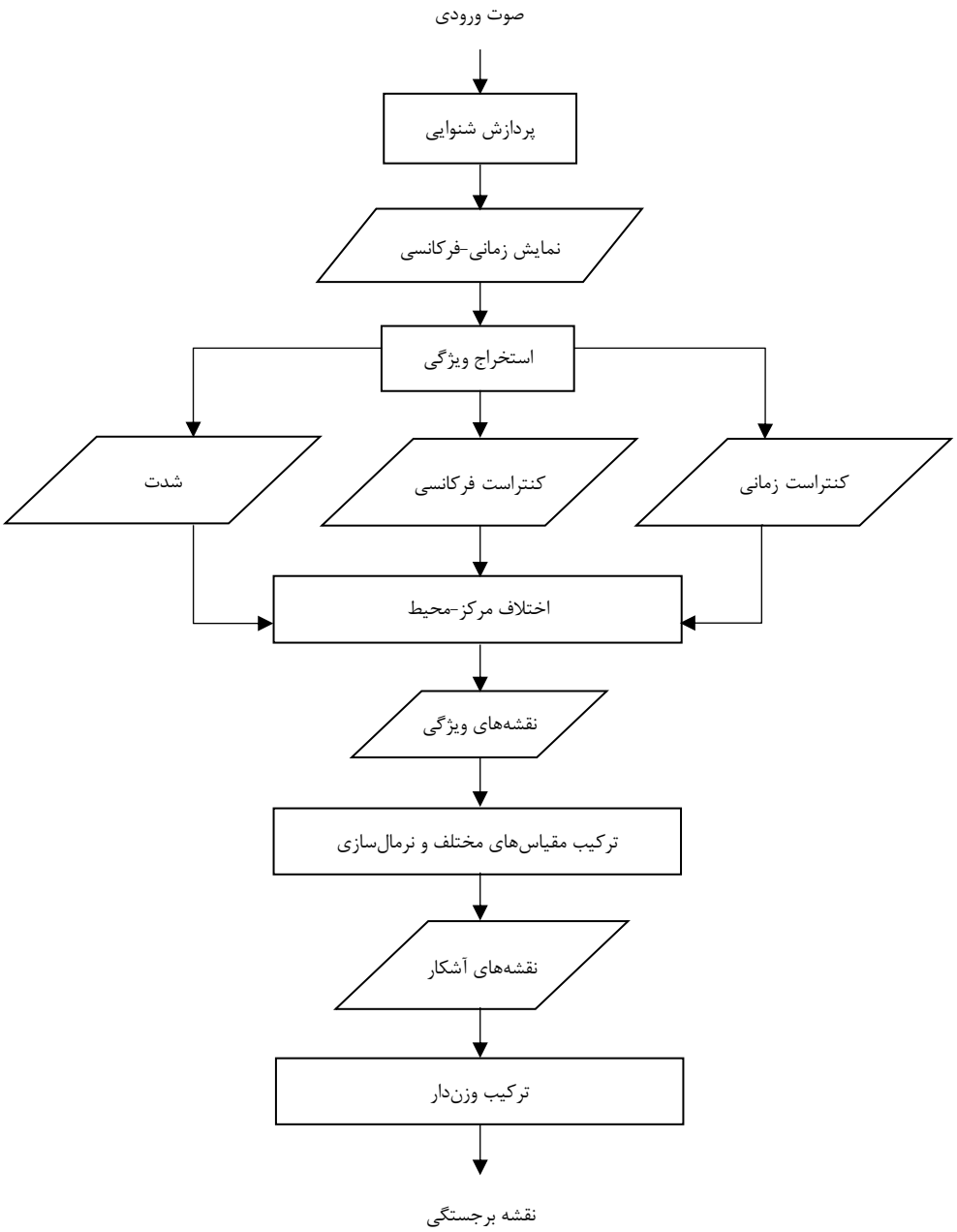
آشکار بدست آمده از ویژگی‌ها، بصورت وزن‌دار باهم جمع می شوند تا ویژگی‌های مهم‌تر که در شناسایی نقطه برجسته اهمیت بیشتری دارند، دارای وزن بالاتری باشند. پارامترهای وزنی بهینه برای هر ویژگی، توسط الگوریتم ژنتیک بدست می‌آیند.

۲-۲-۱- نحوه استخراج نقشه برجستگی شنوایی

در مرحله نخست، مجموعه‌ای از نقشه‌های خام ویژگی‌ها در مقیاس‌های زمانی و فرکانسی مختلف با عبور دادن اطلاعات زمانی-فرکانسی صوت ورودی از فیلترهایی که نواحی پذیرنده در قشر شنوایی را مدل می‌کنند، استخراج می شوند. این عمل در رابطه (۲) نمایش داده شده است.

$$r_{\alpha,\theta}^{\delta,\gamma}(t,f)=\left(s*h_{\alpha,\theta}^{\delta,\gamma}\right)(t,f). \qquad (۲)$$

در این رابطه، $r_{\alpha,\theta}^{\delta,\gamma}(t,f)$ بیانگر نقشه ویژگی مورد نظر، s نمایش زمانی-فرکانسی صوت ورودی و $h_{\alpha,\theta}^{\delta,\gamma}$ نشان‌دهنده پاسخ ضربه فیلتر گابور دوبُعدی [۱۱]، مطابق رابطه (۳) است.



شکل (۱): بلوک‌دیاگرام روش پیشنهادی برای استخراج نقشه برجستگی وزن‌دار در مدل‌سازی شنوایی توجه پایین به بالا.

$$h_{\alpha,\theta}^{\delta,\gamma}(t,f)=\exp\left[-\frac{t'^2+f'^2}{2\delta^2}\right]\cos^{\alpha}\left(\frac{2\pi t'}{\gamma}\right),$$
$$t'=\frac{t}{\Delta t}\cos\vartheta+\frac{f}{\Delta f}\sin\vartheta, \qquad (۳)$$
$$f'=\frac{t}{\Delta t}\sin\vartheta+\frac{f}{\Delta f}\cos\vartheta.$$

که در آن α، θ، δ، γ و ϑ پارامترهای فیلتر گابور هستند.

برای بدست آوردن نقشه‌های ویژگی کنتراست زمانی[۸]، کنتراست فرکانسی[۹] و شدت[۱۰] که در مقاله پایه هم از آن‌ها استفاده شده، پارامترهای فیلتر گابور مطابق مقاله پایه تنظیم شده است. این پارامترها از مشاهده عملکرد فیزیولوژیکی سیستم شنوایی بدست آمده‌اند. فیلتر مربوط به استخراج ویژگی شدت، مطابق با نواحی پذیرنده قشر شنوایی دارای فاز تحریک است. فیلتر استخراج ویژگی کنتراست فرکانسی دارای یک فاز تحریک و مهار باند کناری، بصورت همزمان با آن است. فیلتر طراحی‌شده برای استخراج ویژگی کنتراست زمانی دارای یک فاز تحریک و یک فاز بازدارنده بعدی است. در کوچکترین مقیاس، این فیلترها دارای پهنای باند ۲۰۰ هرتز و طول ۲۰ میلی‌ثانیه، برای هر دو فاز تحریک و بازدارنده هستند. این فیلترها در چهار مقیاس مختلف به نمایش زمانی-فرکانسی صوت ورودی اعمال می‌شوند. هر فیلتر از نمونه‌برداری مجدد فیلتر قبلی با مقیاس ۲ بوجود می‌آید [۵]. با این کار چهار مقیاس مختلف از هرکدام از ویژگی‌ها حاصل می‌شود.

در گام بعدی، مطابق رابطه (۴) اختلاف های مرکز و محیط[۱۱] از نقشه‌های خام محاسبه می‌شوند تا مقیاس‌های مختلف زمانی-فرکانسی از نقشه‌های ویژگی خام بدست آیند.

$$d_{\mu}^{\sigma_c,\Delta\sigma}(t,f)=\left|r_{\mu}^{\sigma_c}(t,f)-r_{\mu}^{\sigma_c+\Delta\sigma}(t,f)\right|,$$
$$\sigma_c=\{2,3,4\}, \qquad (۴)$$
$$\Delta\sigma_c=\{3,4\}.$$

در رابطه (۴)، μ نوع نقشه ویژگی، اعداد، نشان‌دهنده مقیاس‌های نقشه‌های ویژگی و $d_{\mu}^{\sigma_c,\Delta\sigma}(t,f)$ بیانگر مقیاس‌های مختلف محاسبه شده از هر کدام از نقشه‌های ویژگی خام هستند.

در مرحله بعدی، نقشه‌های ویژگی خام، با هدف مدل کردن رقابت بین نقاط برجسته همسایه، نرمالیزه می‌شوند. نقشه‌های ویژگی نرمالیزه شده در مقیاس‌های مختلف، با هم ترکیب می‌شوند تا نقشه‌های آشکار[۱۲] برای هر ویژگی بصورت مجزا مطابق رابطه (۵) بدست آید.

$$q_{\mu}(t,f)=\sum_{\sigma_c=2}^{4}\sum_{\Delta\sigma=3}^{4}N\left(d_{\mu}^{\sigma_c,\Delta c}(t,f)\right). \qquad (۵)$$

در این رابطه، $q_{\mu}(t,f)$ نشان‌دهنده نقشه آشکار برای هر ویژگی و $N\left(d_{\mu}^{\sigma_c,\Delta c}(t,f)\right)$ بیانگر نقشه‌های ویژگی نرمالیزه شده در ابعاد مختلف است. برای به‌دست آوردن نقشه‌های آشکار نرمالیزه شده، ابتدا نقشه‌های حاصل از رابطه (۵) به محدوده $[0,1]$ نگاشت می‌شوند؛ سپس، به عدد یک منهای میانگین ماکزیمم‌های محلی نقشه مربوطه ضرب می‌شوند [۵].

در نهایت، نقشه برجستگی شنوایی با ترکیب نقشه‌های آشکار وزن‌دار شده هر ویژگی، مطابق رابطه (۶) بدست می‌آید. نحوه وزن‌دهی به هر نقشه ویژگی در ادامه توضیح داده خواهد شد.

$$s(t,f)=\sum_{\mu=1}^{3}w_{\mu}q_{\mu}(t,f). \qquad (۶)$$

در رابطه (۶)، w_{μ} نشان‌دهنده وزن تخصیص داده شده به نقشه آشکار ویژگی μ و $s(t,f)$ نشان‌دهنده نقشه برجستگی است.

۲-۲-۲- نحوه وزن‌دهی نقشه‌های آشکار ویژگی‌ها

در این مقاله، جهت تعیین وزن مناسب برای هر نقشه آشکار بدست آمده از ویژگی‌ها، از الگوریتم ژنتیک استفاده شده است. الگوریتم‌های ژنتیک از جمله روش‌های بهینه‌سازی بر اساس جست‌وجو بوده که مبتنی بر اصول انتخاب طبیعی و ژنتیک هستند [۱۲].

الگوریتم ژنتیک براساس تکامل تدریجی نسل‌ها پایه‌ریزی شده است و هدف، آن است که یک‌سری پارامتر بهینه شود. در این الگوریتم ابتدا تعدادی راه‌حل منتخب بصورت تصادفی انتخاب می‌شوند و بعنوان جمعیت[۱۳] درنظر گرفته می‌شوند. در این پژوهش، جمعیت اولیه، به کمک تابع مولد اعداد تصادفی بزرگتر از صفر، با توزیع یکنواخت تولید شده است. هر راه‌حل که کروموزوم نیز نامیده می‌شود، با یک لیست از پارامترها نشان داده می‌شود. . طول کروموزوم‌های استفاده‌شده در این پژوهش برابر با ۳ و نحوه بیان آن‌ها بصورت اعداد حقیقی اعشاری است. در طول هر نسل هر راه‌حل ارزیابی شده و شایستگی[۱۴] آن توسط تابع شایستگی[۱۵] اندازه‌گیری می شود. بر اساس این تابع بعضی از راه‌حل‌ها باقی می‌مانند و بعضی دیگر حذف می‌شوند. گام بعدی الگوریتم ژنتیک، ایجاد دومین نسل از جامعه است. برای این منظور، الگوریتم، از عملگرهای جهش[۱۶] و تقاطع[۱۷] استفاده می‌کند. هریک از این عملگرها با یک احتمال بر روی جمعیت اعمال می‌شوند. در عملگر تقاطع، دو والد باهم ترکیب شده و فرزندی را بوجود می‌آورند. برای این منظور، از عملگر تقاطع میانی[۱۸] استفاده شده است. در این روش، فرزندان، مطابق رابطه (۷) به وجود می‌آیند:

$$O_1=P_1\times\alpha(P_2-P_1). \qquad (۷)$$

در رابطه (۷)، α ضریب مقیاس‌گذاری است که بصورت تصادفی توسط الگوریتم انتخاب می‌شود. P_1 و P_2 نیز، کروموزوم‌های والدین هستند.

برای ایجاد جهش در نسل فعلی، بر اساس احتمال تخصیص داده شده به جهش و با استفاده از عملگر جهش عملی سازگار[۱۹]، مقدار یک ژن بطور تصادفی تغییر می‌یابد. این عملگر، بطور تصادفی روش‌هایی را تولید می‌کند که با آخرین نسل موفق یا ناموفق سازگار هستند. طول گام در این روش‌ها به گونه‌ای انتخاب می شود که محدودیت‌های خطی و مرزهای درنظر گرفته شده برای مسئله نیز برآورده شوند. کل این فرآیندها برای نسل بعدی هم تکرار می شود و با این کار جمعیت سایر نسل‌ها بوجود می‌آیند. در طی این تکامل اجزایی که تابع شایستگی بزرگتری دارند، به نسل بعد منتقل می‌شوند و اجزایی با تابع شایستگی

کوچکتر حذف می‌شوند. این فرآیند آن قدر تکرار می‌شود که پس از تعداد محدودی تکرار متوالی، دیگر بهبودی در پاسخ حاصل نشود.

به منظور انتخاب مقدار بهینه عملگرهای الگوریتم ژنتیک، برای هر کدام از آن‌ها سه سطح مختلف در نظر گرفته شده است. سطح‌های انتخاب شده برای اندازه جمعیت اولیه، ۴، ۱۶ و ۶۴، برای درصد جهش، ۰/۰۱، ۰/۰۵ و ۰/۱ و برای درصد تقاطع، ۰/۷، ۰/۸ و ۰/۹ هستند. با ترکیب این مقادیر، ۲۷ حالت مختلف بوجود می‌آید. نتیجه وزن‌دهی و خطای مدل با حالت‌های مختلف مورد ارزیابی قرار می‌گیرد. بهترین حالت مربوط به زمانی است که مقدار جمعیت اولیه برابر ۶۴، درصد جهش برابر ۰/۱ و درصد تقاطع برابر ۰/۸ فرض شود و از این مقادیر به عنوان عملگرهای الگوریتم ژنتیک در محاسبات استفاده می‌شود.

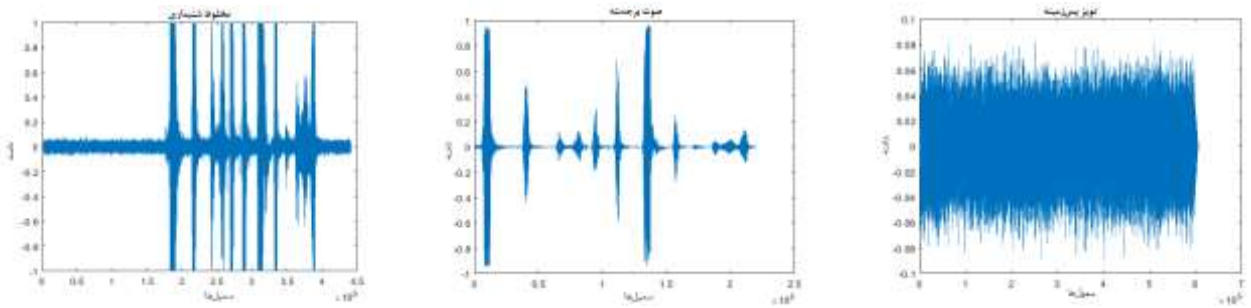
وزن‌های w_μ ذکر شده در رابطه (۶)، برای هر نقشه آشکار بدست آمده از ویژگی‌ها توسط الگوریتم بهینه‌سازی ژنتیک انتخاب می‌شوند، بطوریکه با ضرب‌کردن هر وزن به نقشه آشکار مربوطه و محاسبه نقشه برجستگی، امتیاز برجستگی در محل افزوده شدن صوت برجسته دارای مقدار بیشینه[۲۰] خود باشد. تابع شایستگی الگوریتم ژنتیک به گونه‌ای طراحی شده است که بیشینه نقشه برجستگی در محل افزوده شدن صوت برجسته به هر داده باشد. به عبارتی دیگر، این تابع سعی در کمینه ساختن اختلاف بین محل بیشینه نقشه برجستگی هر ترکیب شنیداری و محل افزوده شدن صوت برجسته به پس‌زمینه نویزی در آن داده را دارد. این‌کار با اعمال وزن‌های مناسب به نقشه‌های آشکار ویژگی‌ها صورت می‌گیرد. الگوریتم ژنتیک به ازای هر صوت موجود در پایگاه دادگان اجرا می‌شود و وزن‌های بدست آمده برای نقشه‌های آشکار ویژگی‌های هر صوت منحصربه‌فرد هستند؛ بطوریکه این وزن‌ها به ازای هر صوت ورودی تغییر می‌یابند.

۳- یافته‌ها و بحث

در این قسمت، ابتدا، دادگان مورد استفاده در شبیه‌سازی‌ها معرفی می‌شود. سپس، نقشه برجستگی وزن‌دار پیشنهادشده مورد تحلیل قرار گرفته و با نقشه برجستگی کایزر که به عنوان مدل پایه در نظر گرفته شده است، مقایسه می‌شود. در پایان، وزن‌های اختصاص یافته توسط الگوریتم ژنتیک به هر یک از ویژگی‌ها مورد بررسی قرار می‌گیرند.

۳-۱- دادگان مورد استفاده

برای تهیه دادگان، جهت ارزیابی مدل پیشنهادی، از نویزهای کارخانه[۲۱]، شبه‌گفتار[۲۲] و نویز سفید[۲۳] [۱۳] به طول ۱۰ ثانیه به عنوان پس‌زمینه استفاده شده است که یک نقطه برجسته دارند. نقطه برجسته از صوت‌های موجود در پایگاه دادگان ESC [۱۴] انتخاب شده است. این پایگاه دادگان حاوی ۲۰۰۰ صوت ۵ ثانیه‌ای در ۵۰ دسته متفاوت است. به تعداد ۱۰ دسته از این پایگاه دادگان، شامل دسته‌های صدای پارس سگ، آژیر، چکه قطرات آب، دست زدن، سرفه کردن، در زدن، کلیک ماوس، زنگ ساعت، شکستن لیوان و صدای قدم زدن، به عنوان فایل‌های صوتی دارای نقاط برجسته انتخاب شده است. دو صوت از هر دسته (مجموعاً ۲۰ صوت) در نقاط زمانی تصادفی[۲۴] و در نسبت‌های سیگنال به نویز[۲۵] ۲۰، ۱۵، ۱۰، ۵، ۰ و ۵- دسیبل به پس‌زمینه اضافه شده که حاصل کار ۳۶۰ مخلوط شنیداری است که به عنوان دادگان مورد استفاده برای ارزیابی مدل پیشنهادی بکار می‌روند. هدف مدل، پیدا کردن زمان نقطه صوت برجسته در مخلوط شنیداری است. در شکل (۲)، نمونه‌ای از داده‌های مورد استفاده شامل صوت پس‌زمینه، نمونه‌ای از صوت برجسته و مخلوط شنیداری حاصل از ترکیب این دو نشان داده شده‌اند.



(الف) (ب) (ج)

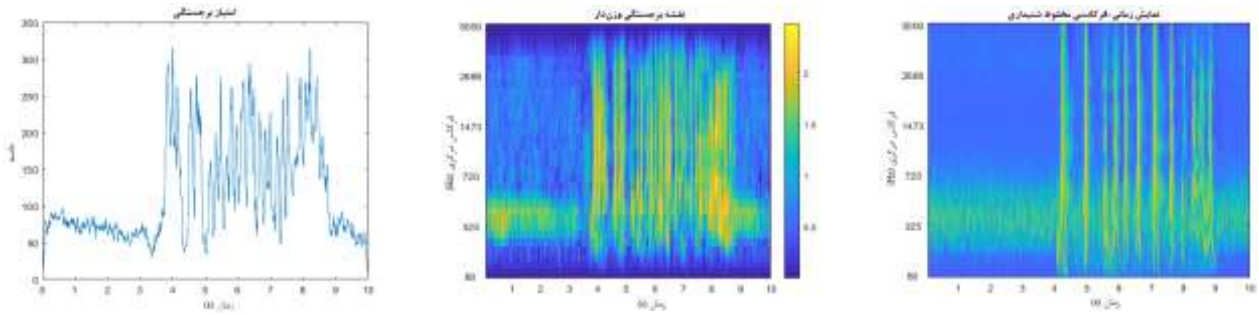
شکل (۲): نمونه‌ای از داده‌های بکار رفته در شبیه‌سازی‌ها: (الف) نویز کارخانه، (ب) صدای سگ که به عنوان صوت برجسته در نظر گرفته شده است، و (ج) مخلوط شنیداری که از اضافه شدن صدای سگ به صدای پس‌زمینه در ثانیه چهارم با SNR = 0 dB بدست آمده است.

۳-۲- نتایج ارزیابی مدل پیشنهادی و مقایسه با مدل پایه

برای ارزیابی مدل پیشنهادی جهت استخراج زمان وقوع قسمت‌های برجسته صوتی، هر کدام از ۲۰ صوت انتخابی، به عنوان بخش‌های برجسته صوتی، در نقطه زمانی تصادفی به صوت‌های پس‌زمینه (نویز کارخانه، نویز شبه‌گفتار و نویز سفید) در شش SNR ۲۰، ۱۵، ۱۰، ۵، ۰ و ۵- دسیبل اضافه شده و سپس، زمان وقوع صوت برجسته توسط مدل پیشنهادی برای همه ۳۶۰ مخلوط شنیداری محاسبه می‌شود. اختلاف زمانی شروع واقعی صوت برجسته و زمان بدست آمده توسط مدل، برای همه داده‌ها تحت عنوان خطا محاسبه شده و میانگین این خطاها به عنوان خطای مدل بر حسب ثانیه بیان می‌شود.

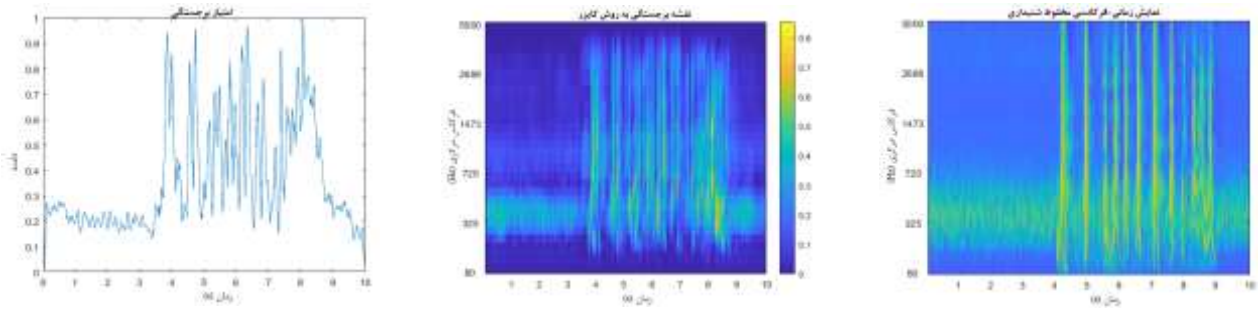
خروجی مدل پیشنهادی بصورت یک تصویر دو بعدی شامل بعدهای زمان و فرکانس تحت عنوان نقشه برجستگی است. برای بدست آوردن زمان نقطه برجسته از این تصویر مطابق روش ارائه شده در مرجع [۱۵] عمل می شود. بعد از محاسبه نقشه برجستگی، اطلاعات کانال‌های فرکانسی با هم جمع می شوند تا بُعد فرکانسی حذف شود و تنها اطلاعات محور زمان باقی بماند. سیگنال بدست آمده، امتیاز برجستگی[۲۶] نام دارد. زمان رخ دادن بیشینه امتیاز برجستگی به عنوان زمان برجسته در صوت در نظر گرفته می‌شود. اختلاف زمانی وقوع صوت برجستهٔ محاسبه شده توسط مدل و زمان برجسته واقعی به عنوان خطای تخمین زمان برجسته در هر داده شناخته می‌شود که میانگین این خطاها برای هر ۳۶۰ داده، خطای کلی مدل را معین می‌کند.

نمایش زمانی-فرکانسی صوت ورودی، نقشه برجستگی و امتیاز برجستگی برای یک مخلوط شنیداری (ترکیب صدای نویز کارخانه و صدای پارس کردن سگ)، در شکل (۳) و (۴)، به ترتیب، برای مدل‌های پیشنهادی و مدل پایه نشان داده شده‌اند.



(الف) (ب) (ج)

شکل (۳): محاسبه نقشه برجستگی وزن‌دار شده و امتیاز برجستگی توسط روش پیشنهادی: (الف) نمایش زمانی-فرکانسی صوت ورودی، (ب) نقشه برجستگی، و (ج) امتیاز برجستگی برای مخلوط نویز کارخانه با صدای پارس سگ در SNR = 0 dB.



(الف) (ب) (ج)

شکل (۴): محاسبه نقشه برجستگی و امتیاز برجستگی توسط روش پایه [۵]: (الف) نمایش زمانی-فرکانسی صوت ورودی، (ب) نقشه برجستگی و (ج) امتیاز برجستگی برای مخلوط نویز کارخانه با صدای پارس سگ در SNR = 0 dB.

جدول (۱): مقایسه خطای تخمین زمان برجسته (بر حسب ثانیه) در مدل پیشنهادی و مدل پایه برای گروه‌های نویز پس‌زمینه در SNR های مختلف.

نویزهای پس‌زمینه		مدل پیشنهادی		مدل پایه
		میانگین	واریانس	میانگین
نویز کارخانه	-۵ dB	۱/۲۵	۰/۰	۲/۰۹
	۰ dB	۰/۹۶	۰/۰۷	۲/۰۵
	۵ dB	۰/۹۵	۰/۰۶	۱/۹۴
	۱۰ dB	۰/۹۴	۰/۱۱	۱/۶۳
	۱۵ dB	۰/۹۳	۰/۱۵	۱/۶۱
	۲۰ dB	۰/۸۷	۰/۰۶	۱/۵۵
نویز شبه‌گفتار	-۵ dB	۰/۸۸	۰/۰۳	۲/۲۲
	۰ dB	۰/۸۲	۰/۰۱	۲/۱۳
	۵ dB	۰/۷۷	۰/۱۸	۲/۱۱
	۱۰ dB	۰/۶۷	۰/۰۷	۱/۷۷
	۱۵ dB	۰/۶۴	۰/۰۹	۱/۷۲
	۲۰ dB	۰/۵۹	۰/۰۶	۱/۷۰
نویز سفید	-۵ dB	۱/۲۹	۰/۰۳	۲/۰۴
	۰ dB	۱/۲۰	۰/۰۳	۲/۰۲
	۵ dB	۱/۰۵	۰/۰۴	۲/۰۱
	۱۰ dB	۰/۹۵	۰/۰۶	۲/۰۱
	۱۵ dB	۰/۹۳	۰/۰۷	۱/۹۷
	۲۰ dB	۰/۸۷	۰/۰۹	۱/۹۶
میانگین کل داده‌ها		**۰/۹۲**	**۰/۰۶**	**۱/۹۱**

در جدول (۱)، خطای تخمین زمان برجسته هر دسته داده ۲۰ تایی در نویزهای مختلف پس‌زمینه و SNR های گوناگون و نیز خطای میانگین کل مدل برای روش‌های پیشنهادی و پایه با هم مقایسه شده‌اند. لازم به ذکر است که به دلیل ماهیت تصادفی الگوریتم ژنتیک، در ارزیابی مربوط به مدل پیشنهادی، این الگوریتم ۱۰ مرتبه بر روی هر کدام از داده‌ها اجرا شده است و با استفاده از وزن‌های حاصل در هر مرتبه تکرار، نقشه برجستگی هر داده محاسبه و خطای آن بدست آمده است. سپس، میانگین و واریانس خطا برای هر داده محاسبه شده و نتایج گزارش شده در جدول (۱) میانگین و واریانس این خطاها برای هر مجموعه از دادگان (که از ۲۰ داده تشکیل شده‌اند) است.

همانگونه که از جدول مشخص است، مدل پیشنهادی عملکرد بهتری را در همه نویزهای پس‌زمینه و SNR ها نسبت به مدل پایه نشان می‌دهد. خطای متوسط مدل‌های پیشنهادی و پایه، به ترتیب، برابر ۰/۹۲ ثانیه و ۱/۹۱ ثانیه است که نشان دهنده عملکرد بهتر مدل پیشنهادی است.

۳-۳- بررسی وزن‌های اختصاص یافته توسط الگوریتم ژنتیک

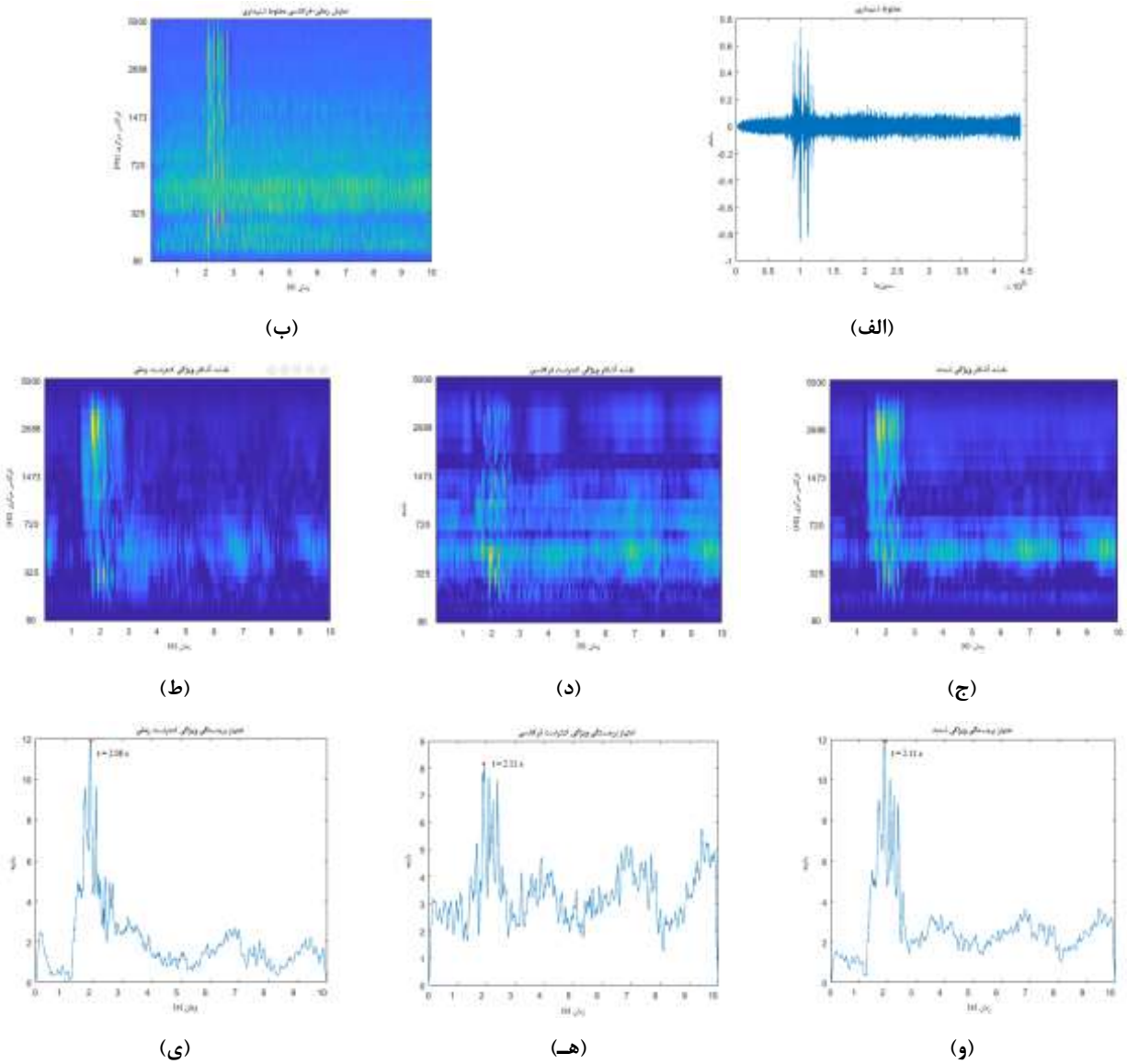
وزن‌های اختصاص یافته برای هر نقشه آشکار، توسط الگوریتم ژنتیک به دست آمده است. این وزن‌ها باعث پر رنگ شدن اهمیت ویژگی‌هایی می‌شود که در تشخیص بهتر قسمت برجسته مخلوط شنیداری ورودی موثرتر هستند. در شکل (۵)، نقشه‌های آشکار ویژگی‌ها و امتیاز برجستگی نقشه‌های آشکار، برای داده‌ای که از ترکیب نویز کارخانه به عنوان پس‌زمینه و صدای سرفه کردن به عنوان صوت برجسته، در نسبت سیگنال به نویز SNR = 0 dB بدست آمده است، ملاحظه می‌شوند. زمان افزوده شدن صوت برجسته به نویز پس‌زمینه ثانیه ۲-ام است. اختلاف این زمان و زمان رخ دادن بیشینه امتیاز برجستگی هر نقشه آشکار (مشابه روش ارائه شده در محاسبه خطای مدل)، برای نقشه‌های ویژگی شدت، کنتراست فرکانسی، و کنتراست زمانی به ترتیب ۰/۱۱، ۰/۰۸ و ۰/۱۱ ثانیه محاسبه شده است. مطابق این اعداد می‌توان گفت، ویژگی کنتراست فرکانسی عملکرد بهتری در محاسبه نقشه برجستگی این داده دارد.

جدول (۲) نشان دهنده وزن‌های اختصاص یافته برای هر نقشه آشکار مربوط به این داده، توسط الگوریتم ژنتیک است. همانگونه که در جدول (۲) مشاهده می‌شود، وزن اختصاص یافته به ویژگی بهتر، یعنی کنتراست فرکانسی، بالاتر از سایر وزن‌ها است. این مسئله باعث می‌شود نقشه برجستگی وزن‌دار (روش پیشنهادی)، با تمرکز و اهمیت دادن به ویژگی‌های مناسب‌تر عملکرد بهتری نسبت به روش پایه داشته باشد، بطوریکه برای داده مذکور خطا از ۰/۱۱ ثانیه برای مدل پایه، به ۰/۰۸ ثانیه برای مدل پیشنهادی کاهش پیدا کرده است.

بطور مشابه، در شکل (۶) همین محاسبات برای داده‌ای دیگر که از ترکیب نویز سفید به عنوان پس‌زمینه و صدای پارس سگ در ثانیه ۴-ام به عنوان صوت برجسته، در نسبت سیگنال به نویز SNR = 10 dB

بدست آمده، مشاهده می‌شود. اختلاف زمان واقعی وقوع رخداد برجسته و بیشینه نقشه‌های آشکار شدت، کنتراست فرکانسی، و کنتراست زمانی، به ترتیب، ۳/۳، ۲/۷ و ۰/۰۵ ثانیه است که نشان‌دهنده عملکرد بهتر ویژگی کنتراست زمانی در تعیین نقشه برجستگی است. مطابق جدول (۳) نیز، وزن بیشتری توسط الگوریتم ژنتیک به این ویژگی اختصاص داده شده است. بکار بردن وزن‌های بهینه در این داده، سبب کاهش خطا از ۰/۱ ثانیه برای مدل پایه به ۰/۰۵ ثانیه برای مدل پیشنهادی (نقشه برجستگی وزن‌دار) شده است.

در شکل (۷)، عملیاتی مشابه برای داده حاصل از ترکیب نویز شبه‌گفتار به عنوان پس‌زمینه و صدای آژیر در ثانیه ۲-ام به عنوان صوت برجسته که در نسبت سیگنال به نویز ۵ dB به پس‌زمینه افزوده شده است، مشاهده می‌شود. اختلاف زمان واقعی وقوع رخداد برجسته و بیشینه نقشه‌های آشکار شدت، کنتراست فرکانسی و کنتراست زمانی به‌ترتیب برابر با ۰/۲، ۰/۲۱ و ۰/۴۹ ثانیه است. این خطاها، نشان‌دهنده عملکرد بهتر ویژگی شدت در محاسبه نقشه برجستگی است. مطابق جدول (۴) نیز، وزن بیشتری توسط الگوریتم ژنتیک به این ویژگی اختصاص داده شده است. بکار بردن وزن‌های بهینه در این داده، سبب کاهش خطا از ۰/۵۰ ثانیه برای مدل پایه به ۰/۲ ثانیه برای مدل پیشنهادی شده است.



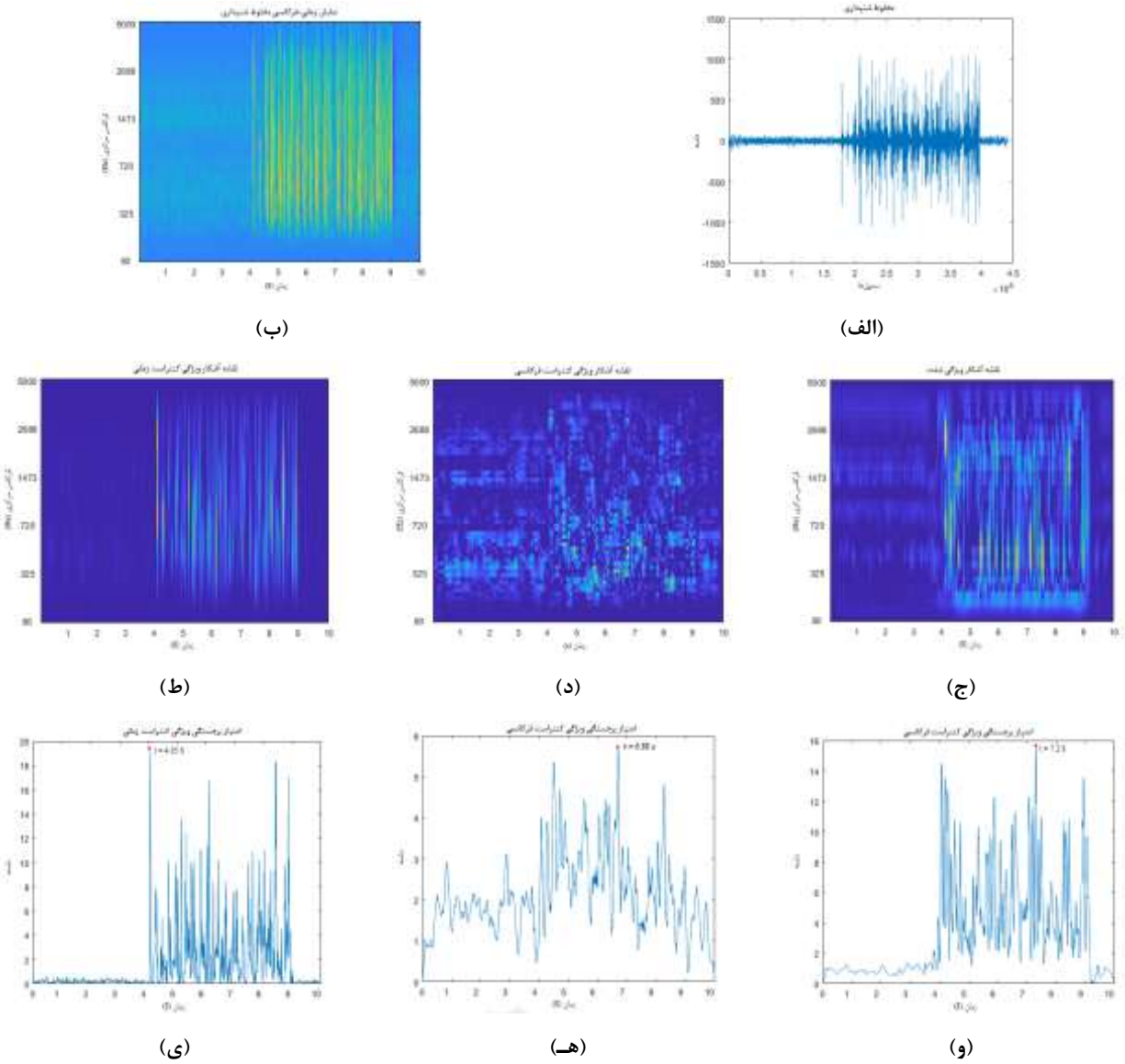
شکل (۵): نقشه‌های آشکار و امتیاز برجستگی هر یک از آن‌ها برای مخلوط نویز کارخانه با صدای سرفه کردن در SNR = 0 dB. (الف) مخلوط شنیداری، (ب) نمایش زمانی-فرکانسی مخلوط شنیداری، (ج) نقشه آشکار ویژگی شدت، (د) نقشه آشکار ویژگی کنتراست فرکانسی، (ط) نقشه آشکار ویژگی کنتراست زمانی، (و) امتیاز برجستگی ویژگی شدت، (هـ) امتیاز برجستگی ویژگی کنتراست فرکانسی و (ی) امتیاز برجستگی ویژگی کنتراست زمانی.

۴- نتیجه‌گیری

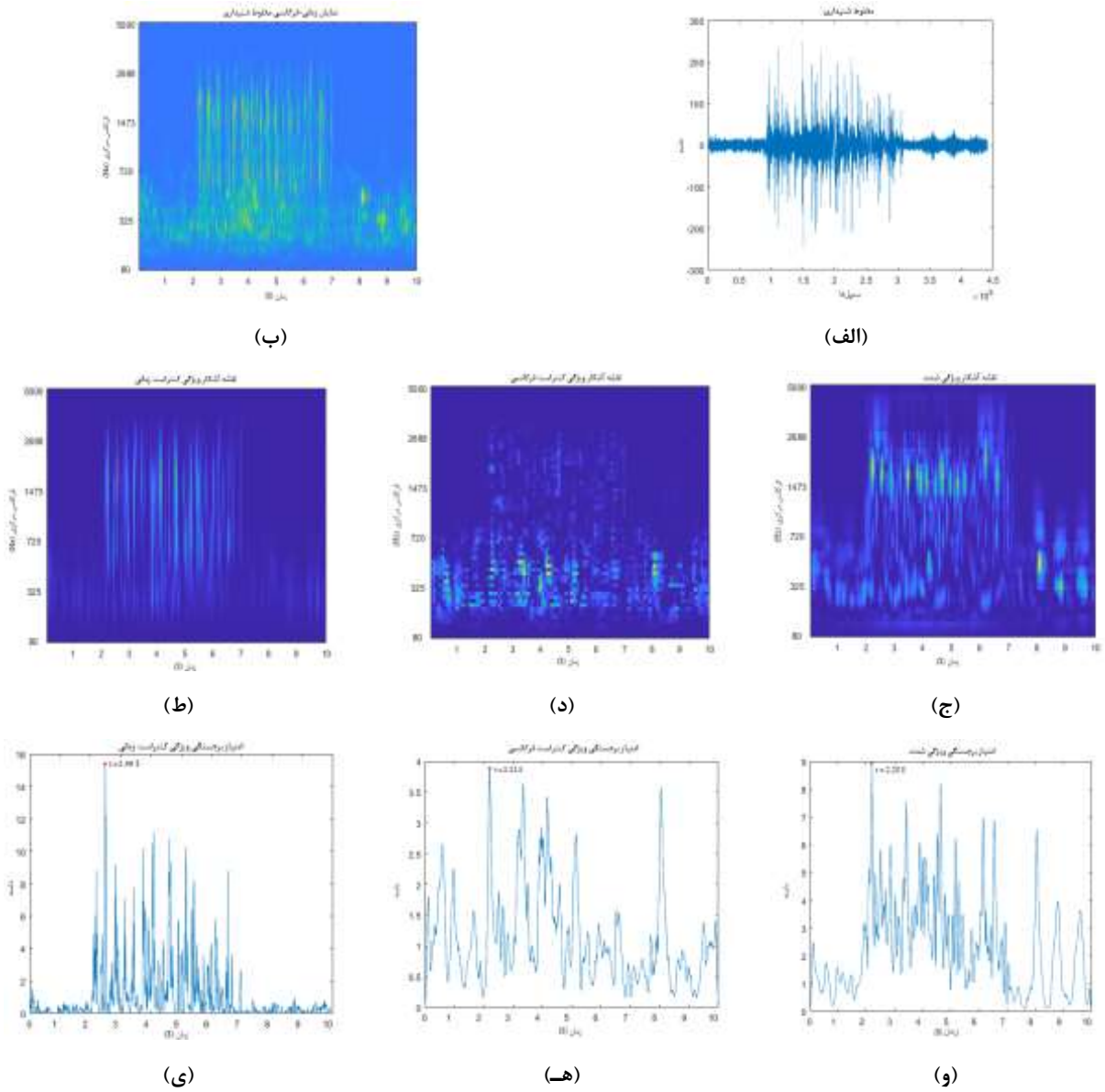
در این مقاله، روشی جهت محاسبه نقشه‌های برجستگی وزن‌دار در مدل‌سازی شنوایی توجه پایین به بالا ارائه شده است. مدل معرفی شده، یک مدل توسعه یافته از مدل پایه‌ی کایزر و همکاران [۵]، که شامل نوآوری در بخش ترکیب نقشه‌های آشکار بدست آمده از ویژگی‌ها است. نتایج حاصل از ارزیابی‌ها نشان دهنده بهبود عملکرد مدل پیشنهادی نسبت به مدل پایه است. میانگین خطای تخمین زمان برجسته از ۱/۹۱ ثانیه برای روش پایه به ۰/۹۲ ثانیه برای روش پیشنهادی کاهش یافته است.

افزودن ویژگی‌های مناسب دیگر مبتنی بر شنوایی به مدل پیشنهادی می‌تواند باعث بهبود نتایج تخمین زمان وقوع صوت برجسته شود.

با توجه به مطالعات انجام شده توسط نویسندگان در زمینه استخراج نقشه‌های برجستگی شنوایی، مقاله‌ای که در آن به وزن‌دهی نقشه‌های آشکار ویژگی‌ها بپردازد، یافت نشد. لذا، به عنوان نقطه شروع، روش ژنتیک برای یافتن وزن‌های بهینه ارائه گردید و نتایج آن با مقاله پایه که در آن وزن‌ها مساوی و برابر یک هستند، مورد مقایسه قرار گرفت.



شکل (۶): نقشه‌های آشکار و امتیاز برجستگی هر یک از آن‌ها برای مخلوط نویز سفید با صدای پارس کردن سگ در SNR = 10 dB. (الف) مخلوط شنیداری، (ب) نمایش زمانی-فرکانسی مخلوط شنیداری، (ج) نقشه آشکار ویژگی شدت، (د) نقشه آشکار ویژگی کنتراست فرکانسی، (ط) نقشه آشکار ویژگی کنتراست زمانی، (و) امتیاز برجستگی ویژگی شدت، (هـ) امتیاز برجستگی ویژگی کنتراست فرکانسی و (ی) امتیاز برجستگی ویژگی کنتراست زمانی.



شکل (۷): نقشه‌های آشکار و امتیاز برجستگی هر یک از آن‌ها برای مخلوط نویز شبه‌گفتار با صدای آژیر در SNR = 5 dB. (الف) مخلوط شنیداری، (ب) نمایش زمانی-فرکانسی مخلوط شنیداری، (ج) نقشه آشکار ویژگی شدت، (د) نقشه آشکار ویژگی کنتراست فرکانسی، (ط) نقشه آشکار ویژگی کنتراست زمانی، (و) امتیاز برجستگی ویژگی شدت، (هـ) امتیاز برجستگی ویژگی کنتراست فرکانسی و (ی) امتیاز برجستگی ویژگی کنتراست زمانی.

نتایج بد ست آمده، حاکی از آن ا ست که الگوریتم ژنتیک قابلیت خوبی در مشخص کردن وزن‌های مناسب برای نقشه‌های آشکار و کاهش خطای محا سبه نق شه برج ستگی دارد؛ ولی با افت ن سبت سیگنال به نویز خطای نقشه برجستگی نیز افزایش یافته است. در الگوریتم ژنتیک اگر مقدار تابع شایستگی یک کروموزوم نسبت به سایر کروموزوم‌ها زیاد با شد (خیلی شایسته‌تر از بقیه با شد) ممکن ا ست دیگر کروموزوم‌ها را به سوی پا سخ بهینه محلی[۲۷] سوق دهد. در نتیجه راه‌حل‌های دیگر به جست‌وجوی خود ادامه نمی‌دهند و الگوریتم از پیداکردن پاسخ بهینه باز می‌ماند. می‌توان گفت در SNR های پایین، این اتفاق سبب شده ا ست تا خطای مدل بیشتر شود. برای حل این مشکل می‌توان از سایر روش‌های وزن‌دهی مانند وزن‌دهی باینری یا وزن‌دهی با الگوریتم‌های بهینه‌سازی جدید، مانند الگوریتم ازدحام ذرات[۲۸] [۱۶] یا الگوریتم جنگل[۲۹] [۱۷]، استفاده کرد و نتایج حاصل از آن‌ها را با نتایج حاصل از الگوریتم ژنتیک مقایسه نمود.

جدول (۲): وزن‌های اختصاص یافته به هریک از ویژگی‌ها توسط الگوریتم ژنتیک برای مخلوط نویز کارخانه با صدای سرفه کردن در SNR = 0 dB.

خطای ویژگی (ثانیه)	وزن اختصاص یافته	نوع ویژگی
۰/۱۱	۰/۱۱	شدت
۰/۰۸	۱۰/۹۹	کنتراست فرکانسی
۰/۱۱	۰/۰۱	کنتراست زمانی

جدول (۳): وزن‌های اختصاص یافته به هریک از ویژگی‌ها توسط الگوریتم ژنتیک برای مخلوط نویزسفید با صدای پارس کردن سگ در SNR = 10 dB.

خطای ویژگی (ثانیه)	وزن اختصاص یافته	نوع ویژگی
۳/۳۳	۱/۳۴	شدت
۲/۷	۸/۸۸	کنتراست فرکانسی
۰/۰۵	۱۷/۷۰	کنتراست زمانی

جدول (۴): وزن‌های اختصاص یافته به هریک از ویژگی‌ها توسط الگوریتم ژنتیک برای مخلوط نویز شبه‌گفتار با صدای آژیر در SNR = 5 dB.

خطای ویژگی (ثانیه)	وزن اختصاص یافته	نوع ویژگی
۰/۲۰	۲/۷۳	شدت
۰/۲۱	۲/۶۷	کنتراست فرکانسی
۰/۴۹	۱/۶۲	کنتراست زمانی

مراجع

[1] R. Desimone, J. Duncan, "Neural mechanisms of selective visual attention," Annual Review of Neuroscience, vol. 18, no. 1, pp. 193-222, 1995.

[2] E. M. Kaya, M. Elhilali, "Modelling auditory attention," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 372, no. 1714, p. 20160101, 2017.

[3] C. Koch, S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," Matters of Intelligence: Springer, pp. 115-141, 1987.

[4] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, 1998.

[5] C. Kayser, C. I. Petkov, M. Lippert, N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," Current Biology, vol. 15, no. 21, pp. 1943-1947, 2005.

[6] O. Kalinli, S. S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August, 2007.

[7] V. Duangudom, D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," 15th European Signal Processing Conference, Poznan, Poland, pp. 1206-1210, September, 2007.

[8] D. Wang, G. J. Brown, "Fundamentals of computational auditory scene analysis," John Wiley and Sons, 2006.

[9] M. Slaney, "Auditory toolbox," Interval Research Corporation, Tech. Rep, vol. 10, 1998.

[10] R. Meddis, M. J. Hewitt, T. M. Shackleton, "Implementation details of a computation model of the inner hair-cell auditory-nerve synapse," Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1813-1816, 199

[11] R. Mehrotra, K. R. Namuduri, N. Ranganathan, "Gabor filter-based edge detection," Pattern Recognition, vol. 25, no. 12, pp. 1479-1494, 1992.

[12] D. Whitley, "A genetic algorithm tutorial," Statistics and Computing, vol. 4, no. 2, pp. 65-85, 1994.

[13] F. Font, G. Roma, X. Serra, "Freesound technical demo," 21st ACM International Conference on Multimedia, Barcelona, Spain, pp. 411-412, October, 2013.

[14] K. J. Piczak, "ESC: Dataset for environmental sound classification," 23rd ACM International Conference on Multimedia, Brisbane, Australia, pp. 1015-1018, 2015.

[15] O. Kalinli, *Biologically inspired auditory attention models with applications in speech and audio processing*, PhD thesis, University of Southern California, 2009.

[16] B. S. G. de Almeida, V. C. Leite, *Particle Swarm Optimization: A Powerful Technique for Solving Engineering Problems*, Swarm Intelligence-Recent Advances, New Perspectives and Applications: IntechOpen, 2019.

[17] M. Ghaemi, M.-R. Feizi-Derakhshi, "Forest optimization algorithm," Expert Systems with Applications, vol. 41, no. 15, pp. 6676-6687, 2014.

زیرنویس‌ها

[۱] Attention

[۲] Bottom-Up Attention

[۳] Top-Down Attention

[۴] Salient

[۵] Saliency Map

[۶] Basilar Membrane

[۷] Equivalent Rectangular Bandwidth (ERB)

[۸] Temporal Contrast

[۹] Frequency Contrast

[۱۰] Intensity

[۱۱] Center-Surround Differences

[۱۲] Conspicuity Maps

[۱۳] Population

[۱۴] Fitness

[۱۵] Fitness Function

[۱۶] Mutation

[۱۷] Cross Over

[۱۸] Intermediate Cross Over

[۱۹] Adaptive Feasible Mutation

[۲۰] Maximum
[۲۱] Factory Noise
[۲۲] Speech-Shaped Noise (SSN)
[۲۳] White Noise
[۲۴] Random Time Points
[۲۵] Signal to Noise Ratio (SNR)
[۲۶] Saliency Score
[۲۷] Local Optimum
[۲۸] Particle Swarm Optimization (PSO)
[۲۹] Forest Optimization