مدل میکروسکوپی دوگوشی مبتنی بر فیلتر بانک مدولاسیون برای پیش‌گویی قابلیت فهم گفتار در افراد دارای شنوایی عادی

علی فلاح و مسعود گراوانچی‌زاده*

دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

چکیده

در این مطالعه، مدل پیش‌گویی قابلیت فهم دوگوشی میکروسکوپی بر مبنای فیلتربانک مدولاسیون ارائه می‌شود. تاکنون در مدل‌های دوگوشی، از معیارهای طیفی مانند STI و SII و یا دیگر روابط تحلیلی برای تعیین میزان قابلیت فهم دوگوشی استفاده شده است. در مدل پیشنهادی، بر خلاف تمام مدل‌های پیش‌گویی قابلیت فهم دوگوشی، از بازشناساگر خودکار گفتار در قسمت پایانی به‌عنوان واحد تصمیم‌گیری استفاده می‌شود. یک مزیت استفاده از این روش، امکان تحلیل میزان بازشناسی قسمت‌های کوچک گفتار مانند واج و سیلاب است. مزیت دیگر این مدل استفاده از پیش‌پردازش‌هایی است که وجود آنها در دستگاه شنوایی انسان به اثبات رسیده است. با استفاده از ماتریس ویژگی پیشنهادی در بازشناساگر گفتار، این مدل دارای پیش‌گویی‌های خوبی در حضور یک منبع نوفه ایستان شبه‌گفتار است. مقایسه نتایج مدل با نتایج حاصل از آزمایش‌های شنوایی، مقادیر همبستگی بالا و میانگین قدر مطلق خطای پایین را نشان می‌دهد. همچنین، ماتریس‌های ابهام برای همخوان‌ها همبستگی بالایی را بین پیش‌گویی‌ها و اندازه‌گیری‌ها نشان می‌دهد. آستانه ادراک گفتار پیش‌گویی‌شده توسط مدل پیشنهادی دارای میانگین قدر مطلق خطای کمتری (۰/۶ دسیبل) در مقایسه با مدل مبنای BSIM است.

واژگان کلیدی: پیش‌گویی قابلیت فهم گفتار، مدل‌های دوگوشی، فیلتربانک مدولاسیون، مدل‌های میکروسکوپی، مدل‌های ماکروسکوپی.

Binaural Microscopic Model Based on Modulation Filterbank for the Prediction of Speech Intelligibility in Normal-Hearing Listeners

Ali Fallah and Masoud Geravanchizadeh

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

Abstract

In this study, a binaural microscopic model for the prediction of speech intelligibility based on the modulation filter bank is introduced. So far, the spectral criteria such as the STI and SII or other analytical methods have been used in the binaural models to determine the binaural intelligibility. In the proposed model, unlike all models of binaural intelligibility prediction, an automatic speech recognizer (ASR) is used in the back-end as the decision unit. One advantage of using this approach is the possibility of analyzing the recognition rate of small parts of speech such as phonemes and syllables. Another advantage of this model lies in the use of pre-processing that their existence in the human auditory system has been verified. Using the proposed feature matrix in the speech recognizer, this model has good predictions in the presence of one source of stationary speech-shaped noise. Comparing the results of the proposed model with those of listening tests show high correlations and low mean absolute error values. Also, the confusion matrices of the consonants represent high correlation between predictions and measurements. The predicted speech

* نویسندهٔ عهده‌دار مکاتبات

[DOI: 10.18869/acadpub.jsdp.14.1.135]

reception threshold by the proposed model has a smaller mean absolute error (0.6 dB) than the baseline model of BSIM.

Keywords: Prediction of Speech Intelligibility, Binaural Models, Modulation Filter bank, Microscopic Models, Macroscopic Models.

۱-مقدمه

دستگاه شنوایی انسان با استفاده از دو گوش، توانایی جداسازی اصوات مورد نظر در محیط‌های نوفه‌ای را دارد. «مسئله مهمانی[1]» که در علم شنودسنجی[2] به این قابلیت می‌پردازد، در سال‌های دور و توسط چری [9] مطرح شده است. دانش آنالیز صحنه‌ای شنوایی[3] شامل مطالعه پدیده‌های آکوستیکی، مبحث پوشیدگی صوت[4]، پردازش دوگوشی[5] و دیگر عوامل مؤثر در جداسازی و تشخیص صداهای مختلف است [7]. بررسی میزان قابلیت فهم گفتار[6] زیرمجموعه‌ای از دانش تحلیل ترکیب شنیداری است که به بررسی ادراک گفتار در موقعیت‌های آکوستیکی مختلف می‌پردازد [8]. پیش‌گویی قابلیت فهم گفتار دوگوشی با آگاهی از موقعیت و نوع منابع صوت کاربردهایی مانند طراحی اتاق با ویژگی‌های آکوستیکی مورد نظر، شنودسنجی، تولید سمعک و فهم دقیق‌تر مکانیزم‌های درون گوش دارد. توانایی انسان در جداسازی اصوات مختلف تا حدود زیادی به استفاده از دو گوش، که شنوایی دوگوشی نامیده می‌شود، بستگی دارد. تفاوت موقعیت فضایی صوت هدف و تداخل، آستانه ادراک گفتار را تا دوازده دسیبل در محیط‌های بدون پژواک، کاهش می‌دهد [8].

اهمیت پیش‌گویی قابلیت فهم در ابتدا، برای شنود تلفنی مطرح شد. مطالعات زیادی که در این زمینه انجام گرفته است، منجر به تهیه استانداردهای پیش‌گویی قابلیت فهم تک‌گوشی شد. دو استاندارد [3] و [4] برای پیش‌گویی قابلیت فهم گفتار تک‌گوشی در شرایط نوفه‌ای و غیرنوفه‌ای استفاده می‌شوند. پارامترهای دیگری نیز مانند ضعف شنوایی افراد در این استانداردها در نظر گرفته شده است. در استاندارد [4]، از روش اندیس قابلیت فهم[7] (SII) استفاده می‌شود. در این روش، مقادیر سیگنال به نوفه در باندهای فرکانسی مختلف محاسبه می‌شود و این مقادیر بر مبنای اهمیت هر باند در قابلیت فهم وزن‌دهی شده و با هم جمع می‌شوند. عدد به‌دست آمده، اندیس قابلیت فهم گفتار نامیده می‌شود. روش شناخته‌شدهٔ دیگر، استفاده از اندیس انتقال گفتار[8] (STI) است که در آن با استفاده از تابع تبدیل مدولاسیون[9]، میزان تخریب مدلاسیون‌های گفتار به‌وسیله سامانه انتقال و در پی آن مقدار قابلیت فهم، پیش‌گویی می‌شود [18]. روش‌های SII و STI دو روش شناخته‌شده پیش‌گویی قابلیت فهم گفتار مبتنی براندیس هستند که در آنها ویژگی‌های سیگنال‌های ورودی مانند نسبت سیگنال به نوفه (در SII) و مدولاسیون گفتار (در STI) به یک اندیس (یا عدد) نگاشته می‌شوند.

در روش‌های بالا، جزئیات پردازش‌های زمانی سیگنال که نقش اساسی در ادراک دستگاه شنوایی دارند، نادیده گرفته می‌شوند؛ مثلاً در SII، طیف بلندمدت[10] گفتار و نوفه به‌عنوان ورودی‌های مدل در نظر گرفته می‌شوند و پیش‌پردازش دیگری روی سیگنال صورت نمی‌گیرد. در مدل تک‌گوشی پیش‌گویی قابلیت فهم ارائه‌شده توسط *کریستینسن* [11]، از پیش‌پردازش‌های دقیق‌تری مشابه دستگاه گوش انسان استفاده شده است. در این روش، ضرایب همبستگی متقابل بین الگوهای پیش‌پردازش‌شده سیگنال نوفه‌ای و سیگنال تمیز، در باندهای فرکانسی مختلف و در بازه‌های زمانی بیست میلی‌ثانیه، محاسبه می‌شوند. براساس ضرایب تابع همبستگی و توان سیگنال تمیز و با وزن‌دهی این مقادیر، مقداری برای قابلیت فهم گفتار محاسبه می‌شود. در پایان نیز، با استفاده از نگاشت تابع منطقی[11] غیرخطی، درصد قابلیت فهم گفتار به‌دست می‌آید. مدل تک‌گوشی دیگری مبتنی بر پیش‌پردازش‌های گوش توسط *هلوب و همکارانش* [20] ارائه شده است. درقسمت پایانی این مدل، برخلاف مدل ارائه‌شده در [11] و دیگر روش‌های مبتنی بر اندیس، از بازشناساگر گفتار استفاده می‌شود که نرخ تشخیص را به‌طورمستقیم و بدون نیاز به تابعی غیر خطی ارائه می‌کند. در این مدل، علاوه‌بر متوسط نرخ تشخیص، نرخ تشخیص و ابهام برای هر کدام از

[1] Cocktail Party Problem
[2] Audiology
[3] Auditory Scene Analysis (ASA)
[4] Masking
[5] Binaural Processing
[6] Speech Intelligibility
[7] Speech Intelligibility Index (SII)
[8] Speech Transmission Index (STI)
[9] Modulation Transfer Function (MTF)
[10] Long-term Spectrum
[11] Logistic Function

[DOI: 10.18869/acadpub.jsdp.14.1.135]

واج[1]ها به‌صورت جداگانه نیز محاسبه می‌شود. این مدل به‌دلیل استفاده از پیش‌پردازش‌های دقیق‌تر و امکان بررسی نرخ تشخیص برای اجزای کوچک‌تر زبان، میکروسکوپی نامیده می‌شود. در مدل ارائه‌شده توسط *یورگنز و همکاران* [23]، قسمت پایانی پیش‌پردازش مدل هلوب و همکارانش ، تغییر داده شده است. در این مدل میکروسکوپی جدید از فیلتربانک مدولاسیون در مرحله آخر استخراج ویژگی استفاده شده است. از ساختار کلی این مدل در [24] برای پیش‌گویی قابلیت فهم واج‌ها در افراد دارای ضعف شنوایی استفاده شده است. مدل‌های میکروسکوپی می‌توانند تخمین دقیقی از قابلیت فهم گفتار ارائه دهند؛ اما ارزیابی آنها نیاز به زمان طولانی‌تری در مقایسه با روش‌های مبتنی بر اندیس دارد.

پس از مدل‌های تک‌گوشی پیش‌گویی قابلیت فهم گفتار، به مدل‌هایی که مبتنی بر شنوایی دوگوشی در محیط‌های واقعی هستند، پرداخته شد. نخستین مدل‌های پیش‌گویی قابلیت فهم گفتار دوگوشی در مطالعات انجام‌شده در [21]، [31] و [35] مطرح شده‌اند. تمامی این مدل‌ها دارای واحدهای دوگوشی هستند که به‌عنوان پیش‌پردازش در مدل‌های تک‌گوشی قابلیت فهم مورد استفاده قرار می‌گیرند. مزیت ناشی از تعامل دوگوشی[2] در کاهش سطح نوفه پوشاننده[3] بعد از پردازش دوگوشی مدل می‌شود. *بویتلمن و همکاران* در [5] از تئوری برابرسازی و حذف[4] و معیار SII، ولی با نادیده‌گرفتن جزئیات پردازش‌های گوش انسان، استفاده کرده است. در این مدل، از پردازش مستقیم بر روی سیگنال گفتار عبور داده‌شده از فیلتربانک گاماتون[5]، با هدف تخمین آستانه ادراک گفتار[6] در زوایای مختلف سیگنال نوفه، برای شرایط آکوستیکی گوناگون استفاده شده است. با اعمال پردازش برابرسازی و حذف در هر باند فرکانسی و اعمال خطاهای مصنوعی[7]، که عدم دقت گوش انسان در پردازش دوگوشی را مدل می‌کند، بویتلمن و همکاران در نسخه دوم از کار خود [6] قسمت‌هایی از کار قبلی خود را به‌منظور کاهش بار محاسباتی تغییر داده است. این مدل به‌اختصار BSIM[8] نامیده شده است. در مدل دوگوشی دیگری، *ون* [32] از پردازش برابرسازی و حذف بر روی سیگنال دوگوشی دارای خطای سیگنال ناپایدار[9]، به جای استفاده از اعمال مستقیم خطاهای مصنوعی، مشابه کار بویتلمن و همکاران، در مدل‌سازی پردازش غیرکامل دوگوشی[10] بهره برده است. *لاواندیر و همکاران* [22] مدلی برای پیش‌گویی قابلیت فهم گفتار دوگوشی در اتاق‌های واقعی دارای پژواک ارائه کرده‌اند که مزیت‌های ناشی از تعامل دوگوشی با استفاده از یک رابطه تحلیلی محاسبه می‌شود. *کازنتینو و همکاران* [13] بر مبنای مدل بالا، روشی برای تخمین پارامترهای دوگوشی مورد نیاز در این رابطه تحلیلی، بدون نیاز به سیگنال‌های گفتار و نوفه جدا از هم، ارائه کرده‌اند. *لکلر و همکاران* [25] مدل پیشنهادی توسط لاواندیر و همکاران [22] را برای پیش‌گویی سناریوهای بیشتر و شرایطی که سیگنال گفتار هدف در فواصل دورتری از شنونده حضور دارد، بهبود داده‌اند.

مدل دوگوشی پیشنهادشده در این مقاله، تعمیم مدل تک‌گوشی قابلیت فهم گفتار مبتنی بر بازشناساگر گفتار ارائه‌شده توسط یورگنز و همکاران [23] است. بلوک‌های پردازشی مدل پیشنهادی، برگرفته از نتایج پژوهش‌های پیشین در مورد پردازش‌های تک‌گوشی و دوگوشی است که در بخش‌های بعدی به جزئیات آنها پرداخته خواهد شد. در مدل پیشنهادی، بر خلاف مدل‌های قبلی، از محاسبه نرخ تشخیص بازشناساگر گفتار برای پیش‌گویی آستانه ادراک گفتار استفاده می‌شود. از این رو امکان مقایسه مستقیم پیش‌گویی‌های مدل با اندازه‌گیری‌ها برای واحدهای کوچکتر زبان که در این پژوهش واج است، فراهم شده است. ماتریس ویژگی مناسبی در مدل جدید پیشنهاد شده است که پارامترهای شنوایی دوگوشی، شامل اثر گوش بهتر[11] و اثر حذف ماسک دوگوشی[12]، مدل می‌شوند. کنترل نرخ تشخیص مدل با استفاده از دو پارامتر معیار فاصله و خطاهای پردازش دوگوشی صورت می‌پذیرد.

بخش‌بندی ادامه مقاله به این ترتیب است که در بخش دوم، ساختار مدل پیشنهادی، شامل پردازش‌های تک‌گوشی، دوگوشی و ماتریس ویژگی، ارائه می‌شود. در بخش سوم، شرایط آزمایشگاهی، داده‌گان مورد استفاده، و همچنین، شرایط و روش انجام آزمایش‌های شنوایی و روش استخراج نتایج مدل شرح داده می‌شوند. در بخش چهارم، نتایج

[1] Phoneme
[2] Binaural Interaction
[3] Masker Noise
[4] Equalization and Cancellation (EC)
[5] Gammatone Filterbank
[6] peech Reception Threshold (SRT)
[7] Artificial Errors
[8] Binaural Speech Intelligibility Model
[9] Jitter
[10] Incomplete Binaural Processing
[11] Better Ear Effect
[12] Binaural Unmasking Effect

[DOI: 10.18869/acadpub.jsdp.14.1.135]

آزمایش‌های شنوایی، شامل آستانه‌های شنوایی اندازه‌گیری‌شده و نرخ تشخیص همخوان‌ها، و همچنین، شبیه‌سازی‌های مدل پیشنهادی گنجانده شده است. به‌منظور مقایسه با روش‌های پیشین پیش‌گویی قابلیت فهم گفتار، نتایج پیش‌گویی‌های مدل BSIM ارائه‌شده توسط بویتلمن و همکاران نیز لحاظ شده است. در پایان و در بخش پنجم، بحث و نتیجه‌گیری‌ها از شبیه‌سازی‌ها گنجانده شده است.

۲-ساختار مدل پیشنهادی پیش‌گویی قابلیت فهم گفتار

ساختار مدل پیشنهادی در شکل (۱) نشان داده شده است. این مدل شامل دو بخش پیش‌پردازش داده‌ها، برای استخراج ماتریس ویژگی، و بخش بازشناساگر گفتار، برای مقایسه دادگان آزمون[1] با دادگان آموزش[2] است. قسمت بالای شکل مربوط به بخش آموزش و قسمت پایین شکل که با نقطه‌چین نشان داده شده، مربوط به بخش آزمون مدل است. بخش آزمون مدل، دارای پیش‌پردازش‌های مشابه با بخش آموزش است.

۲-۱- استخراج ماتریس‌های ویژگی تک‌گوشی و دوگوشی

گفتار و نوفه با استفاده از تابع تبدیل سر مربوطه[3] فیلتر می‌شوند تا سیگنال دارای موقعیت فضایی شبیه‌سازی شود. از این سیگنال‌ها به‌عنوان ورودی مدل استفاده می‌شود.

ماتریس ویژگی پیشنهادی برای بازشناساگر گفتار دارای سه زیر ماتریس است: دو ماتریس ویژگی مربوط به هر کدام از گوش‌های راست و چپ و یک ماتریس که مربوط به خروجی واحد پردازش دوگوشی است.

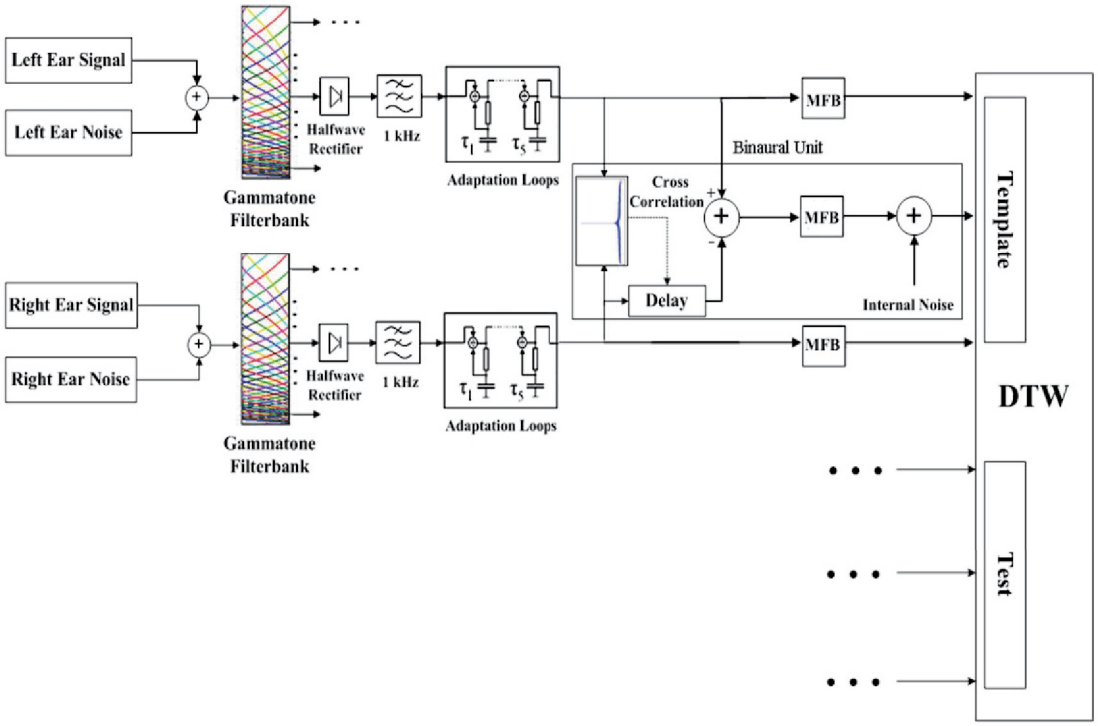
واحد پردازش تک‌گوشی، برگرفته از مدل تک‌گوشی ارائه‌شده در [20] است. پیش‌پردازش در ابتدا، شامل عبور از فیلتربانک گاماتون برای مدل‌سازی پردازش حلزون گوش انسان است [19].

برای پیاده‌سازی فیلتربانک از ۲۷ فیلتر گاماتون که فرکانس مرکزی آنها به‌طور مساوی در مقیاس ERB[4]روی محور فرکانس توزیع شده است، استفاده می‌شود. این فیلتربانک، بازه فرکانسی ۲۳۶ تا ۸۰۰۰ هرتز را مشابه مدل ارائه‌شده در [23] پوشش می‌دهد. خروجی هر فیلتر گاماتون، از یک‌سوساز نیم‌موج عبور داده می‌شود و توسط فیلتر پایین‌گذر مرتبه نخست، با فرکانس قطع یک کیلو هرتز که مدل ساده‌شده‌ای برای سلول‌های مویی[5] گوش داخلی است، فیلتر می‌شود. استفاده از این فیلتر پایین‌گذر سبب عبور بدون تغییر خروجی فیلترهای گاماتون در فرکانس‌های پایین، و استخراج پوش خروجی فیلترهای گاماتون در فرکانس‌های بالا می‌شود.

خروجی مدل سلول‌های مویی توسط پنج حلقه انطباق[6] فشرده‌سازی می‌شود. ثوابت زمانی مدل سلول حلقه‌های انطباق برابر با $\tau_1 = 5\,\text{ms}$، $\tau_2 = 50\,\text{ms}$، $\tau_3 = 129\,\text{ms}$، $\tau_4 = 253\,\text{ms}$ و $\tau_5 = 500\,\text{ms}$ است که این مقادیر با توجه به مقاله [14] انتخاب شده‌اند. حلقه‌های انطباق، سیگنال‌های زمانی ایستان را به‌طور تقریبی به‌صورت لگاریتمی فشرده می‌سازند. هر مسیر بازگشتی در این حلقه‌ها شامل یک مقسم و یک فیلتر پایین‌گذر با ثابت زمانی τ_i است. سیگنال ورودی به حلقه بر خروجی فیلتر پایین‌گذر تقسیم می‌شود. برای سیگنال با مقدار ثابت C، سیگنال خروجی برای هر حلقه انطباق به‌صورت $O = \sqrt{C}$ است. بنابراین، خروجی پنج حلقه برابر $O = C^{1/32}$ خواهد بود که به این صورت عملیات فشرده‌سازی لگاریتمی سیگنال ورودی تخمین زده می‌شود. با تغییر ثابت‌های زمانی با استفاده از ظرفیت خازن‌ها، می‌توان پاسخ سامانه به نوسانات سیگنال ورودی را تغییر داد. در مورد سیگنال‌های غیرایستان، مانند سیگنال گفتار، شروع[7] و پایان[8]های سیگنال در خروجی حلقه‌ها، به‌طور برجسته‌تر تقویت می‌شوند. اگر نوسانات سیگنال ورودی در مقایسه با ثابت‌های زمانی حلقه‌های انطباق سریع‌تر باشد، سیگنال به‌طور تقریبی بدون تغییر از حلقه‌ها عبور می‌کند و در غیر این صورت، سیگنال فشرده می‌شود. حتی هنگامی که سیگنال تحریک ورودی حلقه‌ها قطع می‌شود، به‌دلیل پردازش غیرخطی، تا مدت زمانی در خروجی حلقه‌ها تحریک وجود دارد [14]. در پایان، سیگنال از فیلتربانک مدولاسیون[9] [23] عبور داده می‌شود. این فیلتربانک، شامل چهار کانال مدولاسیون به‌ازای هر کانال فرکانسی گاماتون است به‌نحوی‌که در آن از یک فیلتر پایین‌گذر با فرکانس قطع ۲/۵ هرتز و سه فیلتر

[1] Test Data
[2] Train Data
[3] Head-Related Transfer Function (HRTF)
[4] Equivalent Rectangular Bandwidth (ERB)
[5] Hair-Cells (HCs)
[6] Adaptation Loops (ALs)
[7] Onset
[8] Offset
[9] Modulation Filterbank (MFB)

[DOI: 10.18869/acadpub.jsdp.14.1.135]



(شکل-۱): مدل پیش‌گویی قابلیت فهم گفتار دوگوشی میکروسکوپی شامل دو مسیر تک‌گوشی و یک مسیر دوگوشی که در آن از پردازش برابرسازی و حذف بدون تنظیم بهره استفاده شده است. ماتریس ویژگی استفاده‌شده در DTW سه زیرماتریس دارد: زیرماتریس سیگنال گوش چپ، زیرماتریس گوش راست و زیرماتریس خروجی دوگوشی. قسمت بالای شکل مربوط به پردازش به‌منظور استخراج سیگنال‌های آموزش برای DTW و قسمت پایین (نشان‌داده شده با خطوط نقطه‌چین) دارای پردازش‌های مشابه‌ی با قسمت بالایی شکل بوده که در آن ماتریس ویژگی سیگنال آزمون استخراج می‌شود.

(Figure-1): Binaural microscopic model for the prediction of speech intelligibility including two monaural paths and one binaural path, in which the processings of equalization and cancellation have been employed without adjusting the gain. The feature matrix used in DTW has three sub-matrices: sub-matrix of the left-ear signal, sub-matrix of the right-ear signal, and sub-matrix of the binaural output signal. The upper sectin of the diagram concerns the processing for the extraction of training signals for DTW and the lower section (shown as dotted lines) has a similar processing as the upper one where the feature matrix of the test signal is extracted.

میان‌گذر با فرکانس‌های مرکزی ۵، ۱۰ و ۱۶/۷ هرتز استفاده می‌شود.

پهنای باند برای فیلتر پایین‌گذر و فیلترهای با فرکانس‌های مرکزی ۵ و ۱۰ هرتز، برابر ۵ هرتز و برای فیلتر ۱۶/۷ برابر ۸/۳ هرتز است. خروجی این مدل، نمایش داخلی[1] سیگنال نام دارد که برای کاهش تعداد نمونه‌ها در بازشناساگر گفتار، نرخ آن تا صد هرتز کاهش داده می‌شود. داو و همکاران برای خروجی مدل پیشنهادی، واحدی به نام «واحد مدل»[2] در نظر گرفته‌اند [14]. در این مقاله آمده است که هر واحد مدل معادل با شدت سیگنال صوتی با مقدار یک دسیبل سطح فشار صوت[3] است.

در قسمت دوگوشی مدل، از پردازش برابرسازی و حذف استفاده شده است. از پردازش برابرسازی و حذف در مدل کشف سیگنال[4] دوگوشی ارائه‌شده توسط *زربز* [34] نیز استفاده شده است. در هر باند فرکانسی، تابع همبستگی متقابل دوگوشی[5] بین سیگنال گوش‌های راست و چپ، پس از عبور از حلقه‌های انطباق، محاسبه می‌شود. مقدار تأخیر زمانی بین‌گوشی[6] از روی تأخیر زمانی در بیشینهٔ مقدار تابع همبستگی متقابل تعیین می‌شود. از آنجا که نرخ تشخیص گفتار در مقدار سیگنال-به-نوفه مثبت به‌طور تقریبی صد درصد است، و قابلیت فهم گفتار تنها برای مقادیر منفی سیگنال به نوفه مورد توجه است، درنتیجه، توان سیگنال نوفه همواره در مبحث قابلیت فهم بیشتر از گفتار است. با توجه به اندازه‌گیری‌های انجام‌شده توسط یورگنز و همکاران [23]، مقدار نرخ تشخیص گفتار اندازه‌گیری‌شده در نسبت

[1] Internal Representation (IR)
[2] Model Unit (MU)
[3] Sound Pressure Level (SPL)
[4] Signal Detection
[5] Interaural Cross-Correlation Function (ICF)
[6] Interaural Time Difference (ITD)

[DOI: 10.18869/acadpub.jsdp.14.1.135]

سیگنال به نوفه صفر، بیش از ۹۰٪ برای دادگان گفتاری بی‌معنی در سناریوهای با تحریک یکسان، است. در سناریوهای دوگوشی با توجه به بهبود ناشی از پردازش دوگوشی، انتظار داریم مقادیر نرخ تشخیص در سناریوهای دوگوشی بسیار بیشتر از این مقدار باشد؛ لذا، در این مرحله، انتظار می‌رود که تأخیر زمانی بین‌گوشی مربوط به سیگنال نوفه‌ای به تأخیر زمانی بین‌گوشی سیگنال نوفه بسیار نزدیک باشد. در مرحله برابرسازی، سیگنال یکی از گوش‌ها با توجه به مقدار تأخیر زمانی بین‌گوشی جابه‌جا می‌شود و از سیگنال گوش دیگر کسر می‌شود. بنابراین، عمل برابرسازی تنها در بعد زمانی انجام و از برابرسازی بهره[1] صرف‌نظر می‌شود. در برخی از مدل‌های دوگوشی، از جمله مدل ارائه‌شده در [17] نیز، به‌دلیل تأثیر بیشتر برابرسازی زمانی در مقایسه با برابرسازی بهره، از برابرسازی بهره یا دامنه صرف‌نظر شده است. با صرف‌نظر از برابرسازی بهره، مقداری ابهام[2] در واحد پردازش دوگوشی اعمال می‌شود. اعمال چنین ابهاماتی برای پیش‌گویی‌های صحیح مدل برابرسازی و حذف ضروری است [16]. در قسمت پایانی واحد پردازش دوگوشی، سیگنال دوگوشی، مشابه سیگنال‌های تک‌گوشی، از فیلتربانک مدولاسیون عبور داده می‌شود. در پایان، مقدار «۴ واحد مدل» نوفه داخلی به سیگنال‌های خروجی تمامی باندهای مدولاسیون افزوده می‌شود و کاهش نرخ نمونه‌برداری، مشابه سیگنال‌های تک‌گوشی، تا مقدار صد هرتز صورت می‌پذیرد. از نوفه داخلی و عدم برابرسازی بهره، به‌منظور اعمال ابهام به مدل و محدودکردن پردازش کامل دوگوشی، استفاده شده است.

۲-۲-ماتریس ویژگی پیشنهادی و بازشناساگر گفتار

از بازشناساگر DTW [30]، با معیار فاصله لورنتزی[3] به‌عنوان واحد تصمیم‌گیری مدل استفاده شده است. ماتریس ویژگی مورد استفاده در DTW، ماتریسی ترکیبی و دارای سه زیرماتریس است:

$$\mathbf{IR}=\left[[\mathbf{IR}_{left}],[\mathbf{IR}_{right}],[\mathbf{IR}_{binaural}]\right], \qquad (۱)$$

که در آن $\mathbf{IR}$ ماتریس ویژگی مورد استفاده در DTW در یک زمان خاص را نشان می‌دهد. این ماتریس شامل زیرماتریس‌های نمایش داخلی گوش چپ ($\mathbf{IR}_{left}$)، نمایش داخلی گوش راست ($\mathbf{IR}_{right}$) و نمایش داخلی سیگنال دوگوشی ($\mathbf{IR}_{binaural}$) است. با توجه به وجود ۲۷ باند فرکانسی و ۴ فیلتر مدولاسیون و ۳ مسیر پردازش سیگنال، ابعاد $\mathbf{IR}$ برابر ۱۲×۲۷ خواهد بود. با استفاده از این ماتریس ویژگی پیشنهادی، مدل دوگوشی برای شرایطی که سیگنال نوفه همراستا با سیگنال گفتار و از زاویه صفر درجه پخش می‌شود، نتایجی مشابه مدل تک‌گوشی [23] را دارد. چون در حالت پخش سیگنال و نوفه از زاویه صفر درجه، نمایش‌های داخلی گوش‌های راست و چپ به‌طور تقریبی مشابه هستند، بعد از عملیات برابرسازی و حذف، انرژی سیگنال دوگوشی به‌دلیل حذف همزمان و کامل گفتار و نوفه، مقدار ناچیزی خواهد داشت. اگر $\mathbf{IR}_{train}=\left[\mathbf{x}_{train,lef},\mathbf{x}_{train,right},\mathbf{x}_{train,binaural}\right]$ نمایش داخلی سیگنال آموزش و $\mathbf{IR}_{test}=\left[\mathbf{x}_{test,left},\mathbf{x}_{test,right},\mathbf{x}_{test,binaural}\right]$ نمایش داخلی سیگنال آزمون برای یک زمان خاص باشند، فاصله لورنتزی بین آنها برابر خواهد بود با:

$$D_{Lor}=\sum_j\log\left(1+\frac{1}{2}\left(x_{train,left,j}-x_{test,left,j}\right)^2\right)+\sum_j\log\left(1+\frac{1}{2}\left(x_{train,right,j}-x_{test,right,j}\right)^2\right)+\sum_j\log\left(1+\frac{1}{2}\left(x_{train,binaural,j}-x_{test,binaural,j}\right)^2\right), \qquad (۲)$$

که در آن، اندیس j نشان‌دهندهٔ j-امین عنصر زیرماتریس‌ها در نمایش داخلی سیگنال در یک زمان خاص است. با درنظرگرفتن این واقعیت که در سناریوی S_0N_0 (S_xN_y بیان‌گر وجود سیگنال هدف و نوفه، به‌ترتیب، در زوایای x و y است)، $x_{train,left,j}=x_{train,right,j}=x_{train,j}$ و $x_{test,left,j}=x_{test,right,j}=x_{test,j}$ است و با حذف کامل سیگنال‌ها در برابرسازی و حذف یعنی $x_{temp,binaural,j}-x_{test,binaural,j}\approx 0$ برای رابطه (۲) خواهیم داشت:

$$D_{Lor}=2\times\sum_j\log\left(1+\frac{1}{2}\left(x_{temp,j}-x_{test,j}\right)^2\right) \qquad (۳)$$

به‌عبارتی، تنها مقدار نهایی فاصله لورنتزی در مقدار ثابت ۲ ضرب می‌شود و مسیر بهینه در شبکه DTW[4] تغییر نخواهد کرد و از این رو انتظار داریم که نرخ تشخیص مدل در این حالت مشابه مدل تک‌گوشی باشد؛ اما با جداشدن منبع نوفه

[1] Gain Equalization
[2] Uncertainty
[3] Lorentzian Distance
[4] DTW Grid

[DOI: 10.18869/acadpub.jsdp.14.1.135]

[DOI: 10.18869/acadpub.jsdp.14.1.135]

از منبع سیگنال گفتار در فضا، خروجی واحد پردازش دوگوشی دارای انرژی قابل ملاحظه‌ای خواهد بود. با استفاده از نمایش داخلی دوگوشی در کنار نمایش‌های داخلی سیگنال‌های تک‌گوشی، مدل پیشنهادی در سایر سناریوها به نرخ تشخیص بالاتری می‌رسد. مقایسه‌های مربوط به بررسی میزان تأثیر واحدهای پردازش تک‌گوشی و دوگوشی در نرخ تشخیص در بخش ۴ ارائه خواهد شد.

برای مدل‌سازی میزان نرخ تشخیص صحیح در انسان، مشکل نرخ خطای بالا در بازشناساگر گفتار با استفاده از مفهوم «کشف‌کننده بهینه[1]» مرتفع می‌شود [14]. در اینجا فرض می‌شود که در مرحله کشف سیگنال، آگاهی کامل از خود سیگنال هدف وجود دارد. از این رویکرد در بازشناساگر گفتار مدل میکروسکوپی هلوب [20] نیز استفاده شده است. در این مدل، شکل‌موج‌های آموزش در مرحله آموزش به استثنای نوفه افزوده‌شده به آنها، مشابه شکل‌موج‌های سیگنال آزمون است.

در این مقاله، برای پیاده‌سازی فیلتربانک گاماتون، از کدهای موجود در پایگاه اینترنتی دانشگاه الدنبورگ[2] استفاده شده که در آن از روش [19] برای پیاده‌سازی استفاده شده است. مدل سلول‌های مویی (HCs) و حلقه‌های انطباق (Als)، با استفاده از فایل‌های MEX موجود در تارنمای همین دانشگاه شبیه‌سازی شده است. کدهای مربوط به فیلتربانک مدولاسیون، بخش پردازش دوگوشی و DTW توسط نویسندگان مقاله تهیه شده است.

۳-دادگان و شرایط انجام آزمایش‌های شنوایی و آزمون مدل

در این پژوهش، عملکرد افراد دارای شنوایی عادی درتشخیص صحیح همخوان‌ها[3] در زوایای مختلف پخش سیگنال نوفه، توسط اندازه‌گیری مقادیر آستانه ادراک گفتار مورد ارزیابی قرار می‌گیرد. مقادیر متوسط نرخ تشخیص و همچنین نرخ تشخیص هر کدام از همخوان‌ها برای پیش‌گویی‌های مدل پیشنهادی و نتایج آزمایش‌های شنوایی مقایسه خواهد شد. به‌علاوه، به‌عنوان مرجع مقایسه ،نتایج پیش‌گویی‌های مدل BSIM ارائه‌شده توسط بویتلمن و همکاران [5] مورد استفاده قرار خواهد گرفت.

[1] Optimal Detector
[2] Oldenburg
[3] Consonants

۳-۱-دادگان صوتی

دادگان گفتاری، دارای ساختار «واکه-همخوان-واکه[4]» هستند که از پایگاه داده OLLO [33] دانشگاه الدنبورگ انتخاب شده‌اند. این دادگان توسط گویندگان آلمانی ادا شده‌اند. نتایج آزمایش‌های شنوایی که در ابتدا برای سیگنال تمیز بدون نوفه انجام گرفت، نشان داده است که شنوندگان آذری‌زبان در تشخیص همخوان‌ها در این ساختار مشکلی ندارند و استفاده از این دادگان در آزمایش‌های شنوایی برای سیگنال‌های نوفه‌ای، بلامانع است. دادگان «واکه-همخوان-واکه» OLLO، دارای واکه‌های یکسان در ابتدا و انتها هستند. برای مقایسه مستقیم نتایج مدل با آزمایش‌های شنوایی، از سیگنال‌های صوتی مشابهی در مدل و آزمایش‌ها استفاده می‌شود. فایل‌های صوتی انتخاب‌شده، مربوط به گوینده مرد با شناسه S10M_NO و دارای سرعت بیان معمول واج است. همخوان‌های مورد استفاده یکی از ۱۳ آوای /د/، /ت/، /گ/، /ک/، /ف/، /س/، /ب/، /پ/، /و/، /م/، /ن/، /ش/ و /ل/ می‌باشد که در بین یک جفت یکسان از ۵ واکه /آ/، /اِ/، /ای/، /اُ/ و /او/ قرار می‌گیرد. در نتیجه ۶۵ فایل صوتی متفاوت برای آزمایش‌های شنوایی و آزمون مدل خواهیم داشت. از همخوان /ts/ پایگاه داده OLLO که دارای صدایی مابین صدای دو همخوان /ت/ و /س/ می‌باشد به‌علت ناآشنایی شنوندگان با آن استفاده نشده است.

۳-۲-افراد شرکت‌کننده در آزمایش‌های شنوایی

دوازده فرد آذری‌زبان دارای شنوایی عادی (۱۱ مرد و ۱ زن) که دارای سن ۲۳ تا ۳۰ سال بوده‌اند، در آزمایشات مربوط به اندازه‌گیری آستانه ادراک گفتار شرکت کرده‌اند. آستانه ادراک مطلق این افراد برای سیگنال‌های سینوسی خالص با استفاده از شنوایی‌سنجی استاندارد در بازه ۲۵۰ هرتز تا ۸ کیلوهرتز، بیشتر از بیست دسیبل در مقیاس سطح شنوایی[5] نیست.

۳-۳-شرایط آکوستیکی

اندازه‌گیری‌های این پژوهش در آزمایشگاه گفتار دانشکده توانبخشی دانشگاه علوم پزشکی تبریز انجام شده است. سیگنال گفتار همواره از روبه‌رو (زاویه افقی صفر درجه) برای

[4] Vowel-Consonant-Vowel (VCV)
[5] Hearing Level (HL)

شنوندگان پخش می‌شوند. منبع تداخلی (ICRA-noise) یک نویز ایستان شبه‌گفتار[1] است [15] که از زوایای مختلف برای شنودگان پخش می‌شود. این منبع در شدت صوت ثابت ۶۵ دسیبل توسط دستگاه شنودسنج کالیبره‌شده برای شنوندگان پخش می‌شود. در این پژوهش از هفت سناریوی پخش نوفه در زوایای صفر، ۳۰، ۴۵، ۶۰، ۹۰، ۱۰۵ و ۱۳۵درجه افقی استفاده شده است. در طراحی سناریوها، بحث تقارن در نظر گرفته نشده است. قطعاً بررسی جامع تر نیازمند اندازه‌گیری‌ها و شبیه‌سازی‌ها در زوایای ۷۵، ۱۲۰، ۱۵۰ و نیز ۱۸۰ درجه افقی است. استفاده از زاویه ۱۰۵ درجه افقی، به‌لحاظ به‌کارگیری آن در پژوهش انجام‌شده توسط *پیسیگ و همکاران* [28]، و نیز مقدار کم آستانه ادراک گزارش‌شده این زاویه نسبت به زاویه ۱۲۰ درجه در این پژوهش، از اهمیت خاصی برخوردار بوده و لذا، در آزمایش‌ها مورد استفاده قرار گرفته است. سیگنال‌های گفتار و نوفه توسط مجموعه‌ای از توابع تبدیل سر به‌منظور ایجاد اکوستیک جهتی، فیلتر شده‌اند. از توابع تبدیل سر بدون پژواک تهیه‌شده توسط سر مصنوعی KEMAR به این منظور استفاده شده است [2]. سیگنال‌های گفتار نوفه‌ای دوگوشی از طریق دستگاه شنوایی‌سنجی AC40 [10] که به رایانه وصل است، برای شنودگان پخش می‌شود. با استفاده از این دستگاه می‌توان شدت سیگنال پخش‌شده را در شدت صوت[2] مورد نظر تنظیم نمود.

۳-۴-سیگنال تحریک و روش اندازه‌گیری آستانه ادراک گفتار

دادگان «واکه-همخوان-واکه» به گروه‌هایی تقسیم شده‌اند که تفاوت آنها تنها در همخوان وسط است. با استفاده از این تقسیم‌بندی که شنونده امکان انتخاب یکی از همخوان‌ها را از میان سیزده همخوان دارد، آزمایشی با مجموعه بسته‌ای از دادگان[3] انجام می‌پذیرد. بسته‌بودن آزمایش به معنی امکان انتخاب فرد آزمایش‌شونده از تعدادی گزینه است. فرکانس نمونه‌برداری نوفه و توابع تبدیل سر به مقدار شانزده کیلوهرتز، که فرکانس نمونه‌برداری داده‌های گفتاری است، تغییر داده می‌شود. پخش نوفه، چهارصد میلی‌ثانیه قبل از سیگنال گفتار صورت می‌پذیرد. بنابراین، چهارصد میلی‌ثانیه سکوت به ابتدای دادگان گفتاری افزوده می‌شود. زمان چهارصد میلی‌ثانیه به‌منظور مقداردهی اولیه به خروجی حلقه‌های انطباق، درنظر گرفته شده است. سیگنال نوفه برابر ۶۸ دسیبل و سیگنال گفتار در توانی که نسبت سیگنال به نوفه مورد نظر را تولید کند، تنظیم می‌شود. برای جلوگیری از پخش و قطع لحظه‌ای سیگنال در آزمایش‌ها ادراک شنوایی از نیم‌پنجره‌های هن صد میلی‌ثانیه‌ای برای پنجره‌کردن بخش‌های ابتدایی و انتهایی سیگنال‌های گفتار و نوفه استفاده می‌شود [23]. در پایان، هر کدام از سیگنال‌های گفتار و نوفه توسط تابع سر مربوط به خود، فیلتر می‌شود.

اندازه‌گیری قابلیت فهم گفتار با استفاده از نرم‌افزاری که در محیط GUI-MATLAB تهیه شده، صورت گرفته است. شنوندگان در اتاقکی که تا حد زیادی عایق صوت است، با استفاده از یک رایانه شخصی در آزمون شرکت می‌کنند. قبل از شروع آزمون اصلی، تمامی داده‌گان بدون نوفه برای شنوندگان با هدف آشنایی آنها پخش می‌شود. تابع روان-سنجی[4] رابطه زیر را برای با مجموعه بسته‌ای از دادگان در نظر می‌گیریم:

$$P(L,SRT,s)=\frac{1-g}{1+\exp\left(4s.(SRT-L)\right)}+g, \qquad (۴)$$

که در آن s شیب تابع روان‌سنجی، SRT آستانه ادراک گفتار، L توان گفتار به نوفه و g احتمال تشخیص صحیح یک واج به‌صورت تصادفی است. تابع روانسنجی در حالت کلی تابعی غیر خطی از توان گفتار به نوفه است که در آن مقدار توان گفتار به نوفه به درصد تشخیص صحیح دادگان نگاشت می‌شود. اگر آزمایش به‌صورت مجموعه بسته‌ای از دادگان انجام نشود و شنوندگان توانایی حدس و انتخاب تصادفی صحیح را نداشته باشند، مقدار $g=0$ خواهد بود. در این صورت به‌ازای توان گفتار به نوفه برابر با آستانه ادراک گفتار، یعنی $L=SRT$ ، $P(L,SRT,s)=50\%$ خواهد بود. با توجه به اینکه آزمایش‌ها با مجموعه بسته‌ای از دادگان انجام شده‌اند، با درنظرگرفتن $g=\frac{1}{13}$، نرخ تشخیص در $L=SRT$ در حدود ۵۴٪ است. چون در DTW، یکی از گزینه‌ها به‌حتم انتخاب می‌شود، به‌منظور مقایسه بهتر مدل با اندازه‌گیری‌ها از روش افزایش-کاهش وزن‌دهی‌شده[5] [26] در اندازه‌گیری مقادیر آستانه استفاده شده است. تفاوتی که این روش با روش‌های وفقی پیشین اندازه‌گیری مقادیر آستانه دارد، در استفاده از گام‌های متفاوت برای پاسخ‌های صحیح و پاسخ‌های اشتباه شنونده است. اندازه‌گیری از مقدار

[1] Stationary Speech-Shaped Noise (SSN)
[2] Speech Pressure Level (SPL)
[3] Closed Test
[4] Psychometric Function
[5] Weighted Up and Down (WUD)

[DOI: 10.18869/acadpub.jsdp.14.1.135]

سیگنال به نوفه پنج- دسیبل و با گام کاهشی چهار دسیبل برای پاسخ‌های صحیح شروع می‌شود. این گام کاهشی بعد از دومین و چهارمین برگشت[1]، نصف می‌شود. برگشت در آزمایش به معنی پاسخ صحیح پس از پاسخ ناصحیح و یا پاسخ ناصحیح پس از پاسخ صحیح است. بنابراین، مقدار نهایی گام کاهشی برای پاسخ‌های صحیح برابر یک دسیبل خواهد بود. برای پاسخ‌های ناصحیح که مقدار گام باید افزایش یابد، تمامی گام‌های ذکرشده در مقدار ۱/۱۷ ضرب می‌شود. ازاین‌رو مقدار نهایی گام افزایشی برای پاسخ‌های ناصحیح برابر ۱/۱۷ دسیبل است. گام‌های یادشده ۱ و ۱/۱۷ دسیبل تا پخش تمام ۶۵ داده صوتی ادامه می‌یابند. مقدار متوسط سیگنال به نوفه پخش‌شده در تمامی اجراها[2] با گام ۱ و ۱/۱۷دسیبل به‌عنوان آستانه ادراک گفتار در نظر گرفته می‌شود. برای هر زاویه پخش نوفه، ترتیب پخش تمامی دادگان به‌صورت تصادفی با نرم‌افزار تغییر داده می‌شود. از افراد خواسته شده است که با استفاده از موشواره بر روی گزینه مورد نظر کلیک نمایند و درصورتی‌که اصلاً قادر به تشخیص نباشند، یکی از گزینه‌ها را حدس بزنند.

از آنجایی که مدل میکروسکوپی، قابلیت پیش‌گویی نرخ تشخیص و نیز ماتریس ابهام[3] هرکدام از واج‌ها را به‌صورت جداگانه دارد، آزمایش شنوایی دوم بعد از تخمین مقدار میانگین مقادیر آستانه ادراک گفتار در هر سناریو انجام شده است. پس از تخمین مقدار آستانه ، تمامی دادگان این بار با نسبت سیگنال به نوفه ثابت (برابر با آستانه ادراک گفتار تخمینی) برای شنوندگان در هر سناریو پخش می‌شوند. روش انجام این اندازه‌گیری مشابه رویکرد به‌کارگرفته‌شده در *گراوانچی‌زاده و همکاران* [۱] است.

۳-۵-پیش‌گویی‌های مدل

از سیگنال‌های پخش‌شده و انتخاب‌های موجود در آزمایش‌های شنوایی، برای استخراج نتایج مدل استفاده شده است. بنابراین، مدل و شنوندگان امکان انتخاب را از گروه یکسانی از دادگان دارند.

در مرحله آموزش مدل، نمونه‌هایی از نوفه، در نسبت سیگنال به نوفه مورد جستجو، به دادگانی که تفاوت آنها تنها در واج وسط است، افزوده می‌شود و نمایش داخلی این سیگنال‌های نوفه‌ای به‌عنوان الگوهایی که می‌بایست تشخیص داده شود، ذخیره می‌شود. برای مرحله بازشناسی در مدل، نوفه‌ای متفاوت از نوفهٔ اضافه‌شده به سیگنال‌های آموزش در همان نسبت سیگنال به نوفه، این‌بار به هر کدام از سیگنال‌های گفتار آزمون افزوده می‌شود که یکی از این سیگنال‌های آزمون همان سیگنال استفاده‌شده در الگوی آموزش ذخیره شده است. در هر دو مرحله ذخیره الگوی آموزش و مرحله بازشناسی، بعد از محاسبه نمایش داخلی برای سیگنال تولیدشده، ماتریس‌های ویژگی مربوط به چهارصد میلی‌ثانیه پخش نوفه، که قبل از شروع گفتار افزوده شده است، حذف می‌شود. این عمل به‌منظور در نظرگرفتن تنها اطلاعات مربوط به دادگان گفتار در بازشناساگر گفتار صورت می‌پذیرد. با محاسبه فاصله هر الگوی سیگنال آزمون با الگوهای آموزشی ذخیره‌شده و تعیین الگویی با کمترین فاصله در DTW، نرخ تشخیص صحیح واج در نسبت سیگنال به نوفه‌های مختلف محاسبه می‌شود.

در این پژوهش، آستانه ادراک گفتار در شبیه‌سازی‌ها مشابه اندازه‌گیری‌ها، مقدار سیگنال به نوفه‌ای در نظر گرفته شده است که در آن نرخ تشخیص برابر ۵۴٪ می‌شود. روش جستجوی مقدار آستانه به این صورت است که نسبت سیگنال به نوفه در مدل با گام‌های ۰/۱ دسیبل تغییر داده می‌شود تا این‌که نزدیکترین نرخ تشخیص به ۵۴٪ حاصل گردد. با انتخاب این گام جستجو، با توجه به نتایج شبیه‌سازی‌ها، حداکثر انحراف از مقدار ۵۴٪ برابر ۱/۵٪ به‌دست آمده است. برای هر سناریوی دوگوشی، شبیه‌سازی‌ها با نمونه‌های انتخابی متفاوت از پوشه نوفه SSN بیست‌بار تکرار می‌شوند تا میانگین خوبی از نرخ‌های تشخیص به‌دست آید.

۴-شبیه‌سازی و نتایج آزمایش‌های شنوایی

در آزمایش شنوایی نخست، آستانه‌های ادراک گفتار، با استفاده از یک روش استاندارد، برای افراد مختلف اندازه گیری شده و از مقادیر به‌دست‌آمده از تمامی افراد در هر سناریوی دوگوشی میانگین‌گیری شده است. برای ارزیابی شبیه‌سازی‌ها، از معیارهای مربع ضریب همبستگی و میانگین قدرمطلق خطا بین مقادیر آستانه اندازه‌گیری‌شده و پیش‌گویی‌شده در تمام سناریوها استفاده شده است. حالت ایده‌آل برای پیش‌گویی‌های مدل، مقدار مربع ضریب

[1] Reversal
[2] Runs
[3] Confusion Matrix

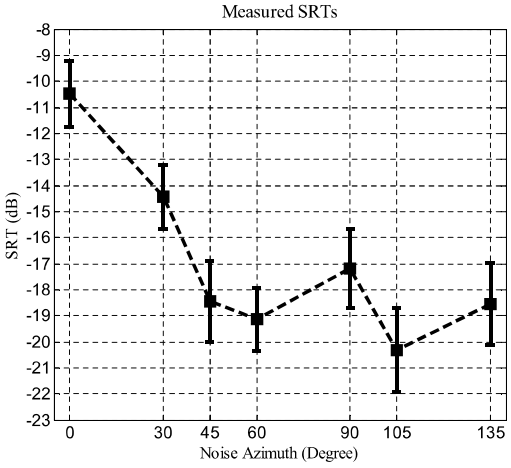
[DOI: 10.18869/acadpub.jsdp.14.1.135]

همبستگی برابر با یک و میانگین قدر مطلق خطای برابر با صفر است. پس از اندازه‌گیری مقادیر آستانه شنوایی، در آزمایش شنوایی دوم از پخش دادگان در توان ثابت به‌منظور محاسبه نرخ تشخیص تک‌تک واج‌ها استفاده شده است. در این مرحله نیز معیارهای ارزیابی یادشده در آزمایش شنوایی نخست، منتها این بار برای مقایسه پیش‌گویی‌های مدل و اندازه‌گیری‌ها در تشخیص واج‌ها، به‌کار رفته‌اند.

۴-۱-نتایج آزمایش‌های شنوایی

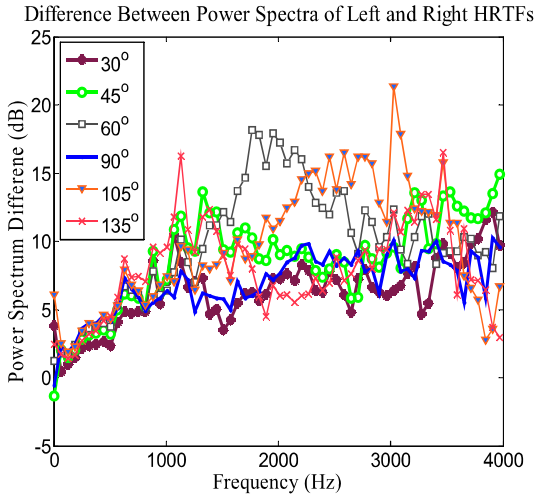
میانگین و انحراف معیارهای آستانه‌های ادراک گفتار اندازه‌گیری‌شده ادراک گفتار برای دوازده فرد دارای شنوایی عادی در شکل (۲) نشان داده شده است. محور افقی، زوایای پخش سیگنال نوفه و محور عمودی، آستانه‌های ادراک گفتار در اندازه‌گیری‌ها را نشان می‌دهد. میله‌های خطا[1]، انحراف معیار استاندارد اندازه‌گیری‌ها را نشان می‌دهد. برای تمامی زوایای افقی، دامنه انحراف معیار استاندارد، در بازه ۱/۲ تا ۱/۶ دسیبل قرار گرفته است. میزان معناداری آماری[2] تأثیر عامل زاویه بر اندازه‌گیری‌ها، توسط تحلیل واریانس[3] (ANOVA) برای هفت سناریو و دوازده فرد آزمایش‌شونده انجام شده و در آن وابستگی نتایج اندازه‌گیری افراد به زاویه مورد بررسی قرار گرفته است. این تحلیل نشان می‌دهد که احتمال وقوع نتایج به‌دست‌آمده به‌طور تصادفی که بیانگر عدم وابستگی نتایج به زوایا است، با استفاده از تابع چگالی احتمال F ، کمتر از یک سطح معنادار است. سطح معناداری برابر با ۰/۰۵ انتخاب شده است. مقدار این تحلیل، نشان‌دهنده تأثیر معنادار زاویه افقی و وجود مزیت دوگوشی در آستانه‌های اندازه‌گیری‌شده ادراک گفتار است ($F(6,77)=58.78,\ p<0.0001$). برای بررسی جامع‌تر، تحلیل واریانس یک‌طرفه[4] نیز برای هر زاویه به‌طور جداگانه انجام شده است. مقادیر آستانه ادراک گفتار ، برای زوایای افقی صفر، ۳۰، ۹۰ و ۱۰۵ درجه دارای تفاوتی معنادار با بقیه زوایا هستند، درحالی‌که بین زوایای افقی ۴۵، ۶۰ و ۱۳۵ درجه، تفاوت معناداری وجود ندارد.

به‌طور طبیعی انتظار داریم که مزیت دوگوشی با افزایش فاصله منبع تداخل از منبع گفتار به‌صورت تدریجی افزایش یابد. با این انتظار، مزیت دوگوشی تا زاویه ۶۰ درجه و تا مقدار نه دسیبل افزایش یافته است؛ اما بعد از آن به مقدار هفت دسیبل در زاویه نود درجه کاهش می‌یابد. افت مزیت دوگوشی در زاویه نود درجه به‌دلیل پدیده «مکان روشن»[5] است. این پدیده به‌دلیل کاهش اثر سایه سر در زوایای نود و ۲۷۰ درجه است.



(شکل-۲): میانگین آستانه‌های ادراک گفتار اندازه‌گیری‌شده در دوازده فرد دارای شنوایی عادی برای هفت زاویه افقی مختلف نوفه.

(Figure-2): The average of the measured SRTs for 12 normal-hearing subjects in 7 different noise azimuths.



(شکل-۳): اختلاف بین طیف توان‌های مربوط به توابع سر گوش‌های چپ و راست. طیف توان زوایای ۱۰۵ و ۶۰ درجه دارای اختلاف زیادی در بازه فرکانسی ۱۵۰۰ تا ۳۰۰۰ هرتز هستند که این بازه در تشخیص همخوان‌ها اهمیت زیادی دارد. مقادیر اختلاف برای زاویه ۹۰ درجه کمتر از ۶۰ و ۱۰۵ درجه و نزدیک به ۳۰ و ۴۵ درجه است.

(Figure-3): The difference between power spectra of left and right HRTFs. The power spectra for the azimuths of 105° and 60° have a large difference in the frequency range of 1500-3000 Hz and this range is very important in the recognition of consonants. The difference values for the 90° are smaller than 60° and 105° and are close to 30° and 45° .

[1] Error Bars
[2] Statistical Significance
[3] Analysis of Variance (ANOVA)
[4] One-Way ANOVA
[5] Bright-Spot Phenomenon

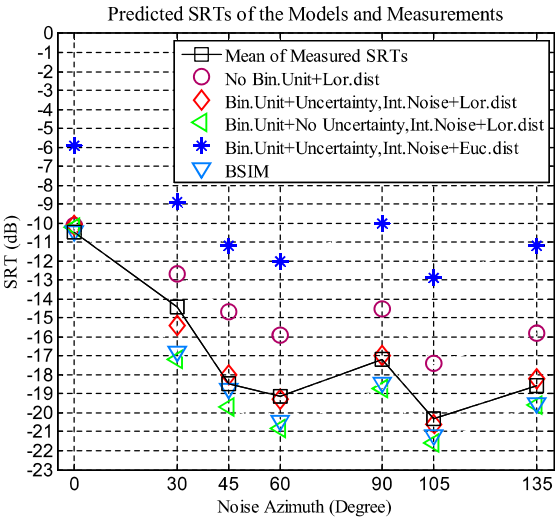
در این زوایا، منبع نوفه به‌طور تقریبی در راستای خط اتصال دو گوش قرار می‌گیرد و موجب کاهش اختلاف سطح توان دوگوشی[1]، و درنتیجه، کاهش مزیت دوگوشی در این زوایا نسبت به زوایای ۶۰ و ۱۰۵ درجه می‌شود. به‌منظور مقایسه میزان عدم شباهت بین توابع تبدیل گوش‌های راست و چپ، اختلاف دامنه پاسخ فرکانسی دو گوش در مقیاس دسیبل برای هر زاویه در شکل (۳) رسم شده است. فاصله فرکانسی ۱۵۰۰ تا ۳۰۰۰ هرتز، اهمیت زیادی در تشخیص همخوان‌ها دارد. با توجه به شکل، اختلاف بین طیف توان‌های مربوط به توابع تبدیل سر گوش‌های راست و چپ در زوایای ۶۰ و ۱۰۵ درجه بیشتر از سایر زوایا است. مقادیر اختلاف برای زاویه ۹۰ درجه کمتر از ۶۰ و ۱۰۵ درجه و نزدیک به ۳۰ و ۴۵ درجه است. در مطالعات قبلی، کمترین مقدار آستانه ادراک گفتار در زوایای افقی ۱۰۰- ۱۲۰ درجه گزارش شده است [8]. کمترین مقدار آستانه ادراک گفتار اندازه‌گیری شده که در ۱۰۵ درجه است، این نتایج را تأیید می‌کند.

۴-۲-پیش‌گویی‌های مدل

پیش‌گویی‌های مدل پیشنهادی در شکل (۴) نشان داده شده است. محور عمودی، آستانه‌های ادراک گفتار پیش‌گویی‌شده توسط مدل را نشان می‌دهد. همچنین، میانگین آستانه‌های ادراک گفتار اندازه‌گیری‌شده نیز در این شکل نشان داده شده است. تأثیر عواملی مانند: ۱- عدم وجود واحد پردازش دوگوشی (دایره)، ۲- وجود یا عدم وجود ابهام و نوفه داخلی در پردازش دوگوشی (به‌ترتیب، مثلث رو به چپ و لوزی) و ۳- تأثیر استفاده از معیار فاصله اقلیدسی به‌جای فاصله لورنتزی بر پیش‌گویی‌ها (ستاره) در این شکل نشان داده شده است. همچنین، به‌منظور مقایسه، پیش‌گویی‌های مدل شناخته‌شده BSIM (مثلث رو به پایین) نیز در این شکل لحاظ شده است. نشان داده شده که BSIM، آستانه ادراک گفتار را در حضور یک یا چند منبع نوفه در شرایط پژواک‌دار و بدون پژواک، به‌خوبی پیش‌گویی می‌کند [6]. برای BSIM، با توجه به مقدار آستانه ادراک گفتار اندازه‌گیری‌شده در سناریوی S_0N_0 که برابر با ۱۰/۴- دسیبل است، مقدار SII مرجع، برابر با ۰/۱ تنظیم می‌شود.

[1] Interaural Level Difference (ILD)

این مقدار مرجع برابر با میانگین SIIهای به‌دست‌آمده برای تمامی همخوان‌ها و نوفه‌های جمع‌شونده، با نسبت سیگنال به نوفه مساوی ۱۰/۴- دسیبل در سناریوی S_0N_0، است. تمامی مقادیر آستانه‌های ادراک گفتار برای مدل پیشنهادی در شکل (۴) از میانگین‌گیری آستانه‌های پیش‌گویی‌شده برای ۶۵ همخوان و در ۲۰ تکرار اجرای برنامه، با استفاده از نمونه‌های متفاوت نوفه، به‌دست آمده است.



(شکل-۴): مقادیر آستانه ادراک گفتار اندازه‌گیری‌شده و پیش‌گویی‌شده در زوایای افقی مختلف منبع نوفه‌ای. برای مدل پیشنهادی تأثیر وجود و یا عدم وجود واحد پردازش دوگوشی، ابهامات و نوفه داخلی بر پردازش دوگوشی و همچنین، نوع معیار فاصله نشان داده شده است.

(Figure-4): The measured and predicted SRTs for different azimuths of the noise source. The effects of presence or absence of binaural processing unit, uncertainties, and internal noise on the binaural processing and also the type of distance measure have been illustrated for the proposed model.

همان‌گونه‌که در این شکل مشاهده می‌شود، برای حالت S_0N_0 نوع معیار فاصله، وجود یا عدم وجود واحد دوگوشی و نیز ابهام و نوفه داخلی تأثیر چندانی بر پیش‌گویی‌های مدل ندارند. این امر بیشتر با توجه به این واقعیت است که در S_0N_0 مزیت استفاده از گوش بهتر وجود ندارد؛ به‌علاوه، با توجه به حذف سیگنال هدف و نوفه به‌صورت هم‌زمان در این سناریو، ماتریس ویژگی دوگوشی اهمیت کمتر و ماهیت نویزی است. برخلاف BSIM که در آن SII برابر مقدار مرجع ۰/۱ تنظیم شده است، مدل پیشنهادی این مزیت را دارد که در آن نیازی به تعیین شرایط مرجع نیست و مدل به‌طور ذاتی دارای پیش‌گویی‌های نزدیک به اندازه‌گیری در حالت S_0N_0 است.

[DOI: 10.18869/acadpub.jsdp.14.1.135]

برای سایر زوایای حضور سیگنال تداخل، پیش‌گویی‌های مدل بدون واحد پردازش دوگوشی و با استفاده از معیار فاصله لورنتزی (دایره) دارای مقادیر آستانه ادراک گفتار بالاتری از اندازه‌گیری‌ها است. با این وجود، روال پیش‌گویی‌ها، یعنی شکل دو منحنی به‌دست‌آمده برای این حالت و اندازه‌گیری‌ها، مشابه یکدیگر است. ازاین‌رو، پیش‌گویی‌های مدل، تأثیر عامل گوش بهتر در شنوایی دوگوشی را به‌خوبی منعکس می‌کند. با اعمال واحد پردازش دوگوشی در مدل (لوزی و مثلث رو به چپ)، اثر حذف ماسک دوگوشی نیز در کنار اثر گوش بهتر، مدل‌سازی شده و مقادیر پیش‌گویی آستانه ادراک گفتار به مقادیر اندازه‌گیری‌شده نزدیک می‌شود. در شکل (۴)، تأثیر ابهام و نوفه داخلی نیز بر پیش‌گویی‌های مدل نشان داده شده است. هنگامی که ابهام و نوفه داخلی در پیاده‌سازی مدل نادیده گرفته شوند (مثلث رو به پایین)، پیش‌گویی‌های مدل با توجه به تئوری برابرسازی و حذف، مقادیر آستانه ادراک گفتار کمتری را نشان می‌دهند. با اعمال ابهام و نوفه داخلی (لوزی)، انطباق خوبی بین اندازه‌گیری و پیش‌گویی‌ها به‌دست می‌آید. نوع معیار فاصله استفاده‌شده در بازشناساگر گفتار، تأثیر قابل‌توجهی بر پیش‌گویی‌های مدل دارد. پیش‌گویی‌های مدل با استفاده از واحد پردازش دوگوشی و استفاده از فاصله اقلیدسی (ستاره) دارای مقادیر آستانه ادراک گفتار بسیار بالاتری از اندازه‌گیری‌ها است. معیار فاصله اقلیدسی برای دادگان دارای توزیع گوسی بهینه است؛ درحالی‌که معیار لورنتزی برای دادگان دارای توزیع لورنتز[1] بهینه است [27]. در مدل تک‌گوشی یورگنز و همکاران نشان داده شده است که مقادیر فاصله که درDTW مورد استفاده قرار می‌گیرند، دارای هیستوگرامی هستند که با استفاده از تابع توزیع لورنتز بسیار بهتر از تابع چگالی گوسی تخمین زده می‌شود. در نتیجه، نرخ تشخیص با استفاده از فاصله لورنتزی بیشتر خواهد بود [23]. مربع مقادیر همبستگی (r^2) و میانگین قدر مطلق خطا[2] بین پیش‌گویی‌های مدل پیشنهادی و اندازه‌گیری‌ها و تأثیر عوامل مختلف بر پیش‌گویی‌های مدل در کنار نتایج پیش‌گویی‌های مدل BSIM در جدول (۱) نشان داده شده است. با استفاده از معیار فاصله لورنتزی و درنظرگرفتن واحد پردازش دوگوشی دارای ابهام و نوفه داخلی، میزان خطای تخمین بسیار کم و برابر ۰/۴ دسیبل است. اگر از معیار اقلیدسی به جای لورنتزی استفاده شود، خطای پیش‌گویی‌ها به ۶/۶ دسیبل می‌رسد.

بدون استفاده از واحد پردازش دوگوشی، مقدار خطای تخمین برابر ۲/۵ دسیبل است. همچنین، اگر ابهام و نوفه داخلی در پردازش دوگوشی در نظر گرفته نشوند، خطا به ۱/۴ دسیبل می‌رسد. پیش‌گویی مدل BSIM نیز دارای میانگین قدرمطلق خطای قابل‌قبول یک دسیبل است. در مقایسه با پیش‌گویی مدل پیشنهادی که در شکل (۴) با علامت لوزی نشان داده شده است، مدل BSIM مقادیر آستانه ادراک گفتار پایین‌تری را نسبت به آستانه‌های اندازه‌گیری‌شده پیش‌گویی می‌کند. به‌دلیل شباهت روند پیش‌گویی‌ها با روند اندازه‌گیری‌ها در زوایای مختلف، معیار مربع ضریب همبستگی، مقادیر بالایی را برای تمامی حالات مدل نشان می‌دهد.

(جدول-۱): مقادیر میانگین قدرمطلق خطا (MAE) و مربع ضریب همبستگی (r^2) بین پیش‌گویی‌های مدل پیشنهادی و مدل BSIM با اندازه‌گیری‌ها.

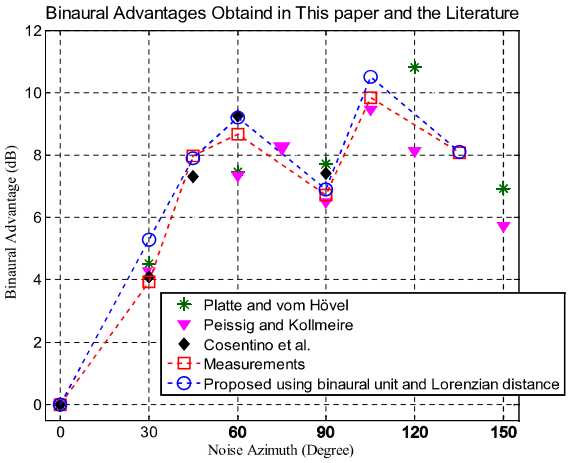
(Table-1): The values of the mean absolute error (MAE) and the square of correlation coefficient (r^2) between the predictions of the proposed model and BSIM model and the measurements.

مدل پیش‌گویی قابلیت فهم گفتار	MAE (dB)	r^2
مدل پیشنهادی با فاصله لورنتزی	0.4	0.97
مدل پیشنهادی با فاصله اقلیدسی	6.6	0.98
مدل پیشنهادی بدون واحد دوگوشی	2.5	0.96
مدل پیشنهادی بدون ابهام و نویز داخلی	1.4	0.95
BSIM	1	0.95

در شکل (۵)، مقایسه مزیت دوگوشی به‌دست‌آمده در اندازه‌گیری‌ها و نتایج بهترین پیش‌گویی مدل پیشنهادی (مدل دارای واحد پردازش دوگوشی و معیار فاصله لورنتزی) با نتایج تعدادی از پژوهش‌های مشابه نشان داده شده است. محور عمودی، مزیت‌های دوگوشی اندازه‌گیری‌شده در سه پژوهش مشابه و مزیت‌های دوگوشی اندازه‌گیری و پیش‌گویی‌شده را در این پژوهش نشان می‌دهد. این نتایج از اندازه‌گیری‌های *پلات* [29]، کازنتینو [13] و پیسیگ [28] برگرفته شده‌اند. در این پژوهش‌ها، مشابه کار ارائه‌شده در این مقاله، از نوفه شبه‌گفتار به‌عنوان سیگنال تداخل استفاده شده اما از دادگان گفتاری متفاوتی در هر کدام استفاده شده، است.

[1] Lorentz Distribution
[2] Mean Absolute Error (MAE)

[DOI: 10.18869/acadpub.jsdp.14.1.135]



(شکل-۵): مزیت دوگوشی به‌دست‌آمده در اندازه‌گیری‌ها و پیش‌گویی‌های مدل پیشنهادی به همراه نتایج مطالعات قبلی.

(Figure-5): The binaural advantage obtained in the measurements and predictions of the proposed model along with the results of previous studies.

این شکل نتایج پژوهش‌های دیگر را تنها از نقطه‌نظر مقایسه اندازه‌گیری‌ها و نه پیش‌گویی مدل‌ها شامل می‌شود. این شکل‌ها روند کلی تغییرات SRT را در زوایای مختلف نشان می‌دهند. اندازه‌گیری‌های [13] تنها برای زوایای ۹۰- تا ۹۰+ انجام شده است که در شکل (۵)، تنها برای زوایای مثبت، این مقادیر اندازه‌گیری ذکر شده است. اندازه‌گیری‌های ما بیشترین مزیت را در زوایای افقی ۶۰ و ۱۰۵ درجه نشان می‌دهد. مطابق این شکل روال و مزیت دوگوشی به‌دست‌آمده برای زوایای مختلف در نتایج این مقاله تا حد زیادی با نتایج مطالعات قبلی همخوانی دارد. افت در مقدار مزیت دوگوشی برای زاویه ۹۰ درجه در نتایج پژوهش کازنتینو [13] و پیسیگ [28] نیز مشاهده می‌شود.

۴-۳-نرخ تشخیص همخوان‌ها و ماتریس‌های ابهام

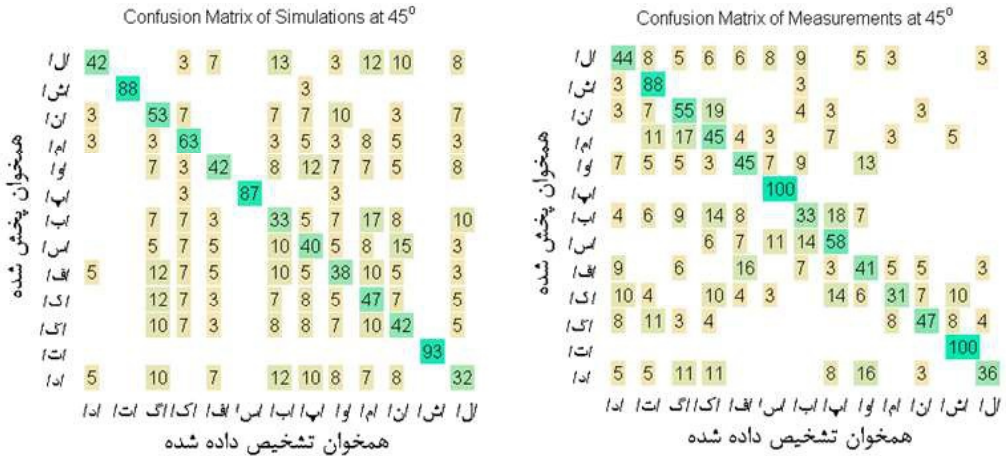
در آزمایش شنوایی دوم، بر اساس آستانه‌های ادراک گفتار به‌دست‌آمده در آزمایش شنوایی نخست، مخلوط گفتار و نوفه در سناریوهای دوگوشی و در سیگنال به نوفه ثابتی در هر سناریو برای شنوندگان پخش شد. مقدار سیگنال به نوفه در آزمایش شنوایی دوم برابر با آستانه ادراک گفتار برای همان سناریو در آزمایش شنوایی نخست تنظیم شده است.

در شکل (۶) به‌عنوان نمونه ماتریس‌های ابهام برای اندازه‌گیری‌ها و پیش‌گویی‌های مدل برای سناریوی پخش نوفه از زاویه ۴۵ درجه نشان داده شده است. در اینجا، شکل سمت راست مربوط به ماتریس ابهام اندازه‌گیری و شکل سمت چپ مربوط به ماتریس ابهام پیش‌گویی برای مدل دارای واحد دوگوشی به‌همراه ابهام و نویز داخلی با استفاده از معیار فاصله لورنتزی (لوزی در شکل (۴)) است. سطر و ستون‌های ماتریس ابهام به‌ترتیب، مربوط به همخوان‌های پخش‌شده و تشخیص‌داده‌شده هستند. عناصر قطری این ماتریس‌ها، میانگین نرخ تشخیص صحیح همخوان‌ها را نشان می‌دهند و عناصر غیرقطری، بیان‌گر ابهامات ایجادشده در تشخیص آنها است. تمامی مقادیر تشخیص در شکل‌ها، به درصد نشان داده شده‌اند. مقایسه کیفی ماتریس ابهام اندازه‌گیری‌ها و پیش‌گویی‌ها نشان می‌دهد که الگوی پیش‌گویی‌های مدل تفاوت زیادی با الگوهای تشخیص در اندازه‌گیری‌ها ندارد. به‌عنوان‌مثال، شنوندگان و پیش‌گویی‌های مدل، برای همخوان‌های /ت/، /س/ و /ش/ نرخ تشخیصی نزدیک به ۱۰۰٪ را نشان می‌دهند. این نتایج، مطابق با پیش‌گویی‌های مدل یورگنز و همکاران [23] و نتایج گراوانچی‌زاده و همکاران [۱] است. همچنین، نرخ تشخیص پایین برای همخوان‌های /ب/، /م/، /ن/، /ف/ و /ل/، دامنه گسترده‌تری از انتخاب‌ها را در ماتریس ابهام نشان می‌دهد. میانگین نرخ‌های تشخیص‌داده‌شده در اندازه‌گیری‌ها برای هفت سناریو، برابر با ۵۲٪ است. این مقدار، نتایج اندازه‌گیری‌های آستانه ادراک گفتار با استفاده از روش وقفی WUD را تأیید می‌کند.

به‌منظور مقایسه الگوی پاسخ‌های صحیح، مقادیر r^2 بین عناصر قطر اصلی ماتریس‌های ابهام اندازه‌گیری‌ها و پیش‌گویی‌ها در جدول (۲) نشان داده شده است. این مقادیر همبستگی بالا نشان می‌دهد که مدل، نرخ تشخیص صحیح همخوان‌ها توسط شنوندگان را به‌خوبی پیش‌گویی می‌کند.

به‌منظور مقایسه الگوی ابهام هر کدام از همخوان‌ها، دوباره از ضرایب r^2 ولی این‌بار به‌عنوان معیار شباهت بین هر سطر ماتریس ابهام اندازه‌گیری‌ها با سطر متناظر در ماتریس ابهام پیش‌گویی‌ها، در جدول (۳) استفاده شده است. مقادیر بالای همبستگی در این جدول تا حد زیادی به‌دلیل مقادیر بالای نرخ تشخیص صحیح هر همخوان در مقدار آستانه ادراک گفتار است. برای همخوان‌های /ت/، /س/ و /ش/ مقادیر r^2 نزدیک به مقدار واحد هستند؛ درحالی‌که همخوان‌های /ب/، /و/ و /م/ کمترین مقادیر همبستگی را نشان می‌دهند. به‌علاوه، برخلاف نرخ تشخیص پایین /م/، /ن/ و /ل/ ، این همخوان‌ها همبستگی بالایی را به‌دلیل شباهت عناصر غیر-قطر اصلی دارند.

[DOI: 10.18869/acadpub.jsdp.14.1.135]



(شکل-۶): ماتریس‌های ابهام همخوان‌ها در اندازه‌گیری‌ها (سمت راست) و شبیه‌سازی‌های مدل (سمت چپ) برای زاویه ۴۵ درجه.

(Figure-6): The confusion matrices of consonants in the measurements (right side) and model simulations (left side) for the 45°.

(جدول-۲): مربع ضرایب همبستگی (r^2) بین عناصر قطری ماتریس‌های ابهام اندازه‌گیری‌ها و شبیه‌سازی‌ها.

(Table-2): The square of correlation coefficients (r^2) between the diagonal elements of confusion matrices of the measurements and simulations.

زاویه نویز	0°	30°	45°	60°	90°	105°	135°	میانگین ۷ سناریو
r^2	0.91	0.91	0.83	0.85	0.88	0.87	0.82	0.86

(جدول-۳): مربع ضرایب همبستگی (r^2) بین سطرهای ماتریس‌های ابهام اندازه‌گیری‌ها و شبیه‌سازی‌ها.

(Table-3): The square of correlation coefficients (r^2) between the rows of confusion matrices of the measurements and simulations.

زاویه نویز / همخوان	0°	30°	45°	60°	90°	105°	135°	میانگین ۷ سناریو
/د/	0.69	0.84	0.78	0.87	0.79	0.83	0.77	0.79
/ت/	0.98	0.99	0.98	0.98	0.99	0.99	0.99	0.98
/گ/	0.90	0.83	0.87	0.91	0.90	0.77	0.68	0.83
/ک/	0.64	0.74	0.80	0.75	0.71	0.65	0.73	0.72
/ف/	0.62	0.85	0.78	0.90	0.75	0.68	0.67	0.75
/س/	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
/ب/	0.48	0.71	0.41	0.68	0.56	0.63	0.73	0.58
/پ/	0.83	0.86	0.77	0.79	0.74	0.81	0.85	0.80
/و/	0.68	0.58	0.83	0.62	0.61	0.73	0.24	0.61
/م/	0.55	0.59	0.62	0.46	0.54	0.60	0.62	0.54
/ن/	0.68	0.68	0.74	0.81	0.79	0.82	0.78	0.76
/ش/	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
/ل/	0.78	0.71	0.61	0.72	0.83	0.79	0.58	0.70

[DOI: 10.18869/acadpub.jsdp.14.1.135]

۵-نتیجه‌گیری و پیشنهاد ادامه کار

در این مطالعه، مدل میکروسکوپی دوگوشی برای پیش‌گویی قابلیت فهم گفتار بر مبنای فیلتربانک مدولاسیون و بازشناساگر گفتار ارائه شده است. ماتریس ویژگی پیشنهادی در مدل جدید به‌گونه‌ای انتخاب شده که با آن بتوان مقادیر آستانه ادراک گفتار را در سناریوهای با تحریک یکسان دوگوش[1] نیز پیش‌گویی کرد. ماتریس ویژگی دارای ۳ زیرماتریس است: دو ماتریس ویژگی مربوط به خروجی واحدهای پردازش تک‌گوشی و ماتریس سوم مربوط به خروجی واحد پردازش دوگوشی است. پیش‌پردازش‌های تک‌گوشی شامل عبور از فیلتربانک گاماتون، یک‌سوسازی نیم‌موج، حلقه‌های انطباق و فیلتربانک مدولاسیون است. تفاوت بخش پردازش دوگوشی مدل پیشنهادی با مدل دوگوشی ارائه‌شده توسط زربز [34]، در قسمت برابرسازی دوگوشی است. در مدل پیشنهادی، برابرسازی تنها برای تأخیر زمانی بین‌گوشی انجام گرفته و از برابرسازی اختلاف سطح توان دوگوشی صرف‌نظر شده است. مدل پیشنهادی در وهله نخست، تنها با استفاده از دو ماتریس ویژگی تک‌گوشی شبیه‌سازی شده و سپس، با افزودن ماتریس ویژگی دوگوشی تکمیل شده است.

مقایسه پیش‌گویی‌های مدل با نتایج آزمایشات شنوایی دوازده فرد دارای شنوایی عادی در هفت زاویه افقی مختلف سیگنال نویز، نشان می‌دهد که مدل جدید با استفاده از فاصله لورنتزی در بازشناساگر گفتار، دارای دقت خوبی در تخمین متوسط نرخ تشخیص تمامی همخوان‌ها و همچنین، نرخ تشخیص هر کدام از همخوان‌ها است. برخلاف مدل‌های پیشین پیش‌گویی قابلیت فهم گفتار، در مدل پیشنهادی به‌دلیل استفاده از بازشناساگر گفتار، امکان تحلیل ماتریس‌های ابهام نیز فراهم شده است. با استفاده از این مدل، مقدار میانگین قدر مطلق خطا در پیش‌گویی مقادیر آستانه ادراک گفتار برابر ۰/۴ دسیبل به‌دست آمده است. مربع ضریب همبستگی بین آستانه‌های ادراک گفتار پیش‌گویی‌شده و اندازه‌گیری‌شده نیز برابر ۹۷/ ۰است.

مدل پیشنهادی در شرایط بدون پژواک و با حضور یک منبع نوفه شبه‌گفتار، قابلیت پیش‌گویی خوبی دارد. از آنجا-که دادگان استفاده‌شده در این پژوهش دارای طول محدود هستند، امکان ارزیابی پژواک برای آنها وجود ندارد؛ اما یک مدل دوگوشی میکروسکوپی کامل می‌بایست قابلیت پیش‌گویی را در شرایط پژواک‌دار واقعی و در حضور منابع و تعداد مختلف سیگنال‌های تداخلی داشته باشد. برای بررسی تأثیر پژواک باید از جملات به‌جای دادگان بدون‌معنا استفاده شود.

مدل میکروسکوپی دیگری که پیش‌پردازش‌های متفاوتی از پژوهش انجام‌شده در این مقاله را دارد، مربوط به مدل «کشف دید کوتاه[2]» است که توسط کوک [12] مطرح شده است. در این مدل، از شناساگر مدل پنهان مارکوف استفاده شده است. این مدل از قسمت‌های تمیز سیگنال در شرایط نویزی برای پیش‌گویی قابلیت فهم استفاده می‌کند. نتایج کوک نشان می‌دهد که «دیدهای کوتاه» اطلاعات کافی جهت پیش‌گویی نرخ تشخیص انسانی را دارد. تعمیم این مدل تک‌گوشی پیش‌گویی قابلیت فهم به یک مدل دوگوشی می‌تواند پژوهش بعدی در راستای ارائه مدل دوگوشی میکروسکوپی جدید و کامل‌تری برای پیش‌گویی در محیط‌های واقعی باشد.

تقدیر و تشکر

نویسندگان این مقاله از هم‌فکری خانم نگین صالحی در دانشکده توانبخشی دانشگاه علوم پزشکی تبریز در استفاده از آزمایشگاه صوت و انجام آزمایش‌های شنوایی کمال تشکر را دارند.

۶-مراجع 6-Refrences

[۱]گراوانچی‌زاده، مسعود، فلاح، علی، اعتراف اسکویی، میرعلی، "پیش‌گویی قابلیت فهم همخوان‌ها در افراد دارای شنوایی عادی با استفاده از مدل‌های میکروسکوپی دارای معیار فاصله متفاوت در بازشناساگر خودکار گفتار"، *فصل‌نامه علمی پژوهشی پردازش علائم و داده‌ها*، دوره ۱۲، شماره ۱، صفحات ۷۹-۹۰، ۱۳۹۴.

[1] M. Geravanchizadeh, A. Fallah, and A. Eteraf Oskoiuei, "Prediction of consonants intelligibility for listeners with normal hearing using microscopic models of speech perception considering different distance measures in automatic speech recognizer," *Signal and Data Processing* (*JSDP*), vol. 12, no. 1, pp-79-90, 2015.

[2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF

[1] Diotic

[2] Glimpse Detection

[DOI: 10.18869/acadpub.jsdp.14.1.135]

131, pp. 796–807, 2014.

[14] T. Dau and A. Kohlrausch, "Modeling auditory processing of amplitude modulation I. Detection and masking with narrowband-carriers," *J. Acoust. Soc. Am.*, vol. 102, pp. 2893–2905, 1997.

[15] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, pp. 148–157, 2001.

[16] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, pp. 1206–1218, 1963.

[17] N. I. Durlach, "Binaural signal detection: equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, vol. 2, edited by J. V. Tobias, Academic Press, New York, 1972, Chap. 10, pp. 369–462.

[18] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, vol. 28, pp. 66–73, 1973.

[19] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica United with Acustica*, vol. 88, pp. 433–442, 2002.

[20] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.*, vol. 100, pp. 1703–1716, 1996.

[21] H. Levitt and L. R. Rabiner, "Predicting binaural gain in intelligibility and release from masking for speech," *J. Acoust. Soc. Am.*, vol. 42, pp. 820–828, 1967.

[22] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.*, vol. 131, pp. 218–231, 2012.

[23] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *J. Acoust. Soc. Am.*, vol. 126, pp. 2635–2648, 2009.

[24] T. Jürgens, S. D. Ewert, B. Kollmeier, and T. database," In Proc. *IEEE Work, Appl. Signal Process. to Audio Acoust Proc*, New Platz, NY, pp. 99–102,

[3] ANSI (1969), "Methods for the calculation of the articulation index, American National Standard S3.5–1969, Standards Secretariat," *Acoustical Society of America*, 1969.

[4] ANSI (1997), "Methods for calculation of the speech intelligibility index, American National Standard S3.5–1997, Standards Secretariat," *Acoustical Society of America*, 1997.

[5] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, pp. 331–342, 2006.

[6] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," *J. acoust. Soc. Am.*, vol. 127, pp. 2479–2497, 2010.

[7] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. Massach-usetts, Cambridge, The MIT Press, 1990.

[8] A. W. Bronkhorst, "The cocktail party phenomenon: a review of research on speech intelligibility in multiple talker conditions," *Acta Acustica United with Acustica*, vol. 86, pp. 117–128, 2000.

[9] E. C. Cherry, "Some experiments on the recognition of speech with one and with two ears," *J. acoust. Soc. Am.*, vol. 25, pp. 975–979, 1954.

[10] Clinical Audiometer AC40, "Instructions for Use," Interacoustics, DK-5610 Assens, Denmark. [Online]. Available: www.interac-oustics.com. [Accessed: June 7, 2017].

[11] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communications*, vol. 52, pp. 678–692, 2010.

[12] M. P. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 199 , pp. 1562–1573, 2006.

[13] S. Cosentino, T. Marquardt, D. McAlpine, J. F. Culling, and T. H. Falk, "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," *J. Acoust. Soc. Am.*, vol.

[DOI: 10.18869/acadpub.jsdp.14.1.135]

[DOI: 10.18869/acadpub.jsdp.14.1.135]

Brand, "Prediction of consonant recognition in quiet for listeners with normal and impaired hearing using an auditory model," *J. Acoust. Soc. Am.*, vol. 135, pp. 1506–1517, 2014.

[25] T. Leclère, M. Lavandier, and J. F. Culling, "Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation," *J. Acoust. Soc. Am.*, vol. 137, pp. 3335–3345, 2016.

[26] C. Kaernbach, "Adaptive threshold estimation with unforced-choice tasks," *Percepion and Psychophys*. vol. 63, 1377–1388, 2001.

[27] W. Press, S. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*. Massachusetts, Cambridge, The MIT Press, 1992.

[28] J. Peissig and B. Kollmeier, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," *J. Acoust. Soc. Am.*, vol. 101, pp. 1660–1670, 1997.

[29] H. J. Platte and H. vom Hövel, "Zur deutung der ergebnisse von sprachverständlichkeitsmessungen mit störschall im sreifeld," *Acta Acustica United with Acustica*, vol. 45, pp. 139–150, 1980.

[30] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust, Speech, Signal Process.*, vol. 26, pp. 43–49, 1978.

[31] H. vom Hövel, "On the importance of the transmission properties of the outer ear and the binaural auditory system in disturbed speech transmission", Ph.D. Thesis, University of Aachen, Germany, 1984.

[32] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, pp. 3678–3690, 2010.

[33] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier, "Oldenburg logatom speech corpus (OLLO) for speech recognition experiments with humans and machines," in *Proceedings of Interspeech*, pp. 1273–1276, Lisbon, Portugal, 2005.

[34] C. Zerbs, "Modeling the effective binaural signal processing in the auditory system",

Ph.D. Thesis, Carl-von-Ossietzky-Universitaet Oldenburg, Germany, 2000.

[35] P. M. Zurek, "Binaural advantages and directional effects in speech intelligibility," in *Acoustical factors affecting hearing aid performance*, edited by G. A. Studebaker, I. Hockberg, Allyn and Bacon, Boston, 2nd Ed., 1993, Chap. 15, pp. 255–276.

علی فلاح مدارک کارشناسی، کارشناسی ارشد و دکترا در رشته برق، گرایش مخابرات را در سال‌های ۱۳۸۵ و ۱۳۸۸ و ۱۳۹۵ به ترتیب، از دانشگاه‌های صنعتی امیرکبیر، دانشگاه شاهد و دانشگاه تبریز اخذ کرده است. زمینه‌های پژوهشی مورد علاقه ایشان پردازش سیگنال آرایه‌ای، پردازش سیگنال گفتار، پردازش سیگنال دوگوشی و مدل های ادراک شنوایی است.

نشانی رایانامه ایشان عبارت است از:

ali.fallah@tabrizu.ac.ir

مسعود گراوانچی‌زاده مدرک کارشناسی را در رشته مهندسی الکترونیک در سال ۱۳۶۵ از دانشگاه تبریز اخذ نموده است. وی سپس، مدارک کارشناسی ارشد و دکترا در رشته پردازش سیگنال را، به‌ترتیب، در سال‌های ۱۳۷۴ و ۱۳۸۰ از دانشگاه Ruhr-University Bochum آلمان کسب کرده است. ایشان هم اکنون عضو هیأت علمی دانشکده مهندسی برق و کامپیوتر دانشگاه تبریز با مرتبه علمی دانشیاری هستند. زمینه‌های پژوهشی مورد علاقه ایشان بهبود کیفیت گفتار، مکان‌یابی و جداسازی سیگنال گفتار، پردازش سیگنال‌های تصادفی و پردازش سیگنال دوگوشی است.

نشانی رایانامه ایشان عبارت است از:

geravanchizadeh@tabrizu.ac.ir