

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 مدل های دوگوشی، از معیارهای طیفی مانند STI و SII یا دیگر روابط تحلیلی برای تعیین میزان قابلیت فهم دوگوشی استفاده شده است. در مدل پیشنهادی، بر خلاف تمام مدل های پیشگویی قابلیت فهم دوگوشی، از بازناسانگر خودکار گفتار در قسمت پایانی به عنوان واحد تصمیمگیری استفاده نمیشود. یک مزیت استفاده از این روش، امکان تحلیل میزان بازناسانی قسمت های کوچک گفتار مانند واج و سیلا ب است. مزیت دیگر این مدل استفاده از پیش پردازش هایی است که وجود آنها در دستگاه شنوایی انسان به اثبات رسیده است. با استفاده از ماتریس ویژگی پیشنهادی در بازناسانگر گفتار، این مدل دارای پیشگویی های خوبی در حضور یک منبع نویه ایستگاه شبکه گفتار است. مقایسه نتایج مدل با نتایج حاصل از آزمایش های شنوایی، مقادیر همبستگی بالا و میانگین قدر مطلق خطای پایین را نشان می دهد. همچنین، ماتریس های ابهام برای همخوان ها همبستگی بالابی را بین پیشگویی ها و اندازه گیری ها نشان می دهد. آستانه ادراک گفتار پیشگویی شده 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، مدل های دوگوشی، فیلتربانک مدولاسیون، مدل های میکروسکوپی، مدل های 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

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.

اهمیت هر باند در قابلیت فهم وزن دهی شده و با هم جمع می شوند. عدد به دست آمده، اندیس قابلیت فهم گفتار نامیده می شود. روش شناخته شده دیگر، استفاده از اندیس انتقال گفتار^۸ (STI) است که در آن با استفاده ازتابع تبدیل مدولاسیون^۹، میزان تخریب مدل اسیون های گفتار به وسیله سامانه انتقال و در پی آن مقدار قابلیت فهم، پیش گویی می شود [18]. روش های SII و STI دو روش شناخته شده پیش گویی قابلیت فهم گفتار مبتنی بر اندیس هستند که در آنها ویژگی های سیگنال های ورودی مانند نسبت سیگنال به نویه (در) SII و مدولاسیون گفتار (در) STI به یک اندیس (یا عدد) نگاشته می شوند.

در روش های بالا، جزئیات پردازش های زمانی سیگنال که نقش اساسی در ادراک دستگاه شنوایی دارند، نادیده گرفته می شوند؛ مثلاً در SII، طیف بلندمدت^{۱۰} گفتار و نویه به عنوان ورودی های مدل در نظر گرفته می شوند و پیش پردازش دیگری روی سیگنال صورت نمی گیرد. در مدل تک گوشی پیش گویی قابلیت فهم ارائه شده توسط کریستینسن [11]، از پیش پردازش های دقیق تری مشابه دستگاه گوش انسان استفاده شده است. در این روش، ضرایب همبستگی متقابل بین الگوهای پیش پردازش شده سیگنال نویه ای و سیگنال تمیز، در باندهای فرکانسی مختلف و در بازه های زمانی بیست میلی ثانیه، محاسبه می شوند. براساس ضرایب تابع همبستگی و توان سیگنال تمیز و با وزن دهی این مقادیر، مقداری برای قابلیت فهم گفتار محاسبه می شود. در پایان نیز، با استفاده از نگاشت تابع منطقی^{۱۱} غیر خطی، درصد قابلیت فهم گفتار به دست می آید. مدل تک گوشی دیگری مبتنی بر پیش پردازش های گوش توسط هلوپ و همکارانش [20] ارائه شده است. در در قسمت پایانی این مدل، برخلاف مدل ارائه شده در [11] و دیگر روش های مبتنی بر اندیس، از بازناساگر گفتار استفاده می شود که نرخ تشخیص را به طور مستقیم و بدون نیاز به تابعی غیر خطی ارائه می کند. در این مدل، علاوه بر متوسط نرخ تشخیص، نرخ تشخیص و ابهام برای هر کدام از

۱- مقدمه

دستگاه شنوایی انسان با استفاده از دو گوش، توانایی جداسازی اصوات مورد نظر در محیط های نویه ای را دارد. «مسئله مهمانی^۱» که در علم شنود سنجی^۲ به این قابلیت می پردازد، در سال های دور و توسط چری [9] مطرح شده است. دانش آنالیز صحنه ای شنوایی^۳ شامل مطالعه پدیده های آکوستیکی، مبحث پوشیدگی صوت^۴، پردازش دو گوشی^۵ و دیگر عوامل مؤثر در جداسازی و تشخیص صدای های مختلف است [7]. بررسی میزان قابلیت فهم گفتار^۶ زیر مجموعه ای از دانش تحلیل ترکیب شنیداری است که به بررسی ادراک گفتار در موقعیت های آکوستیکی مختلف می پردازد [8]. پیش گویی قابلیت فهم گفتار دو گوشی با آگاهی از موقعیت و نوع منابع صوت کاربردهایی مانند طراحی اتاق با ویژگی های آکوستیکی مورد نظر، شنود سنجی، تولید سمعک و فهم دقیق تر مکانیزم های درون گوش دارد. توانایی انسان در جداسازی اصوات مختلف تا حدود زیادی به استفاده از دو گوش، که شنوایی دو گوشی نامیده می شود، بستگی دارد. تفاوت موقعیت فضایی صوت هدف و تداخل، آستانه ادراک گفتار را تا دوازده دسیبل در محیط های بدون پژواک، کاهش می دهد [8].

اهمیت پیش گویی قابلیت فهم در ابتدا، برای شنود تلفنی مطرح شد. مطالعات زیادی که در این زمینه انجام گرفته است، منجر به تهیه استانداردهای پیش گویی قابلیت فهم تک گوشی شد. دو استاندارد [3] و [4] برای پیش گویی قابلیت فهم گفتار تک گوشی در شرایط نویه ای و غیر نویه ای استفاده می شوند. پارامترهای دیگری نیز مانند ضعف شنوایی افراد در این استانداردها در نظر گرفته شده است. در استاندارد [4]، از روش اندیس قابلیت فهم^۷ (SII) استفاده می شود. در این روش، مقادیر سیگنال به نویه در باندهای فرکانسی مختلف محاسبه می شود و این مقادیر بر مبنای

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

دوگوشی دیگری، ون [32] از پردازش برابرسازی و حذف بر روی سیگنال دوگوشی دارای خطای سیگنال ناپایدار^۹، به جای استفاده از اعمال مستقیم خطاهای مصنوعی، مشابه کار بویتلمن و همکاران، در مدل‌سازی پردازش غیرکامل دوگوشی^{۱۰} بهره برده است. لاواندیر و همکاران [22] مدلی برای پیش‌گویی قابلیت فهم گفتار دوگوشی در اتفاق‌های 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 نووفه جدا از هم، ارائه کرده‌اند. لکلر و همکاران [25] مدل پیشنهادی توسط لاواندیر و همکاران [22] را برای پیش‌گویی ستاریوهای بیشتر و شرایطی که سیگنال گفتار هدف در فواصل دورتری از شنوونده حضور دارد، بهبود داده‌اند.

مدل دوگوشی پیشنهادشده در این مقاله، تعمیم مدل تک‌گوشی قابلیت فهم گفتار مبتنی بر بازناساگر 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، برخلاف مدل‌های قبلی، از محاسبه نرخ تشخیص بازناساگر گفتار برای پیش‌گویی آستانه ادراگ 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^{۱۱} و اثر حذف ماسک دوگوشی^{۱۲}، 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

⁹ Jitter

¹⁰ Incomplete Binaural Processing

¹¹ Better Ear Effect

¹² Binaural Unmasking Effect

واج^۱‌ها به صورت 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 هلوپ و همکارانش، تعییر 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 دهنده؛ اما ارزیابی آنها نیاز به زمان طولانی‌تری در مقایسه با روش‌های مبتنی بر اندیسی 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 به عنوان پیش‌پردازش در مدل‌های تک‌گوشی قابلیت فهم مورد استفاده قرار می‌گیرند. مزیت ناشی از تعامل دوگوشی^۲ در کاهش سطح نووفه پوشانده^۳ بعد از پردازش دوگوشی مدل می‌شود. بویتلمن و همکاران در [5] از تئوری برابرسازی و حذف^۴ و معیار SII، ولی با نادیده‌گرفتن جزئیات پردازش‌های گوش انسان، استفاده کرده است. در این مدل، از پردازش مستقیم بر روی سیگنال گفتار عبور داده شده از فیلتربانک گاماتون^۵، با هدف تخمین آستانه ادراک گفتار^۶ در زوایای مختلف سیگنال نووفه، برای شرایط آکوستیکی گوناگون استفاده شده است. با اعمال پردازش برابرسازی و حذف در هر باند فرکانسی و اعمال خطاهای مصنوعی^۷، 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 به منظور کاهش بار محاسباتی تعییر داده است. این مدل به اختصار BSIM^۸ نامیده شده است. در مدل

¹ Phoneme

² Binaural Interaction

³ Masker Noise

⁴ Equalization and Cancellation (EC)

⁵ Gammatone Filterbank

⁶ speech Reception Threshold (SRT)

⁷ Artificial Errors

⁸ Binaural Speech Intelligibility Model

از یکسوساز نیموج عبور داده می‌شود و توسط فیلتر پایین‌گذر مرتبه نخست، با فرکانس قطع یک کیلو هرتز که مدل ساده‌شده‌ای برای سلول‌های مویی^۵ 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.

خروچی مدل سلول‌های مویی توسط پنج حلقه انتطبق^۶ فشرده‌سازی می‌شود. ثوابت زمانی مدل سلول حلقه‌های انتطبق برابر با $\tau_1 = 5 \text{ ms}$, $\tau_2 = 129 \text{ ms}$, $\tau_3 = 50 \text{ ms}$, $\tau_4 = 253 \text{ ms}$ و $\tau_5 = 500 \text{ ms}$ است که این مقادیر با توجه به مقاله [14] انتخاب شده‌اند. حلقه‌های انتطبق، سیگنال‌های زمانی ایستان را به طور تقریبی به صورت 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 انتطبق به صورت $O = C^{1/32}$ 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^۷ و پایان^۸‌های سیگنال در خروجی حلقه‌ها، به طور برجسته‌تر تقویت می‌شوند. اگر نوسانات سیگنال ورودی در مقایسه با ثابت‌های زمانی حلقه‌های انتطبق سریع‌تر باشد، سیگنال به طور تقریبی بدون تغییر از حلقه‌ها عبور می‌کند و در غیر این صورت، سیگنال فشرده می‌شود. حتی هنگامی که سیگنال تحریک ورودی حلقه‌ها قطع می‌شود، به دلیل پردازش غیرخطی، تا مدت زمانی در خروجی حلقه‌ها تحریک وجود دارد [14]. در پایان، سیگنال از فیلتربانک مدولاسیون^۹ [23] عبور داده می‌شود. این فیلتربانک، شامل چهار کanal مدولاسیون به‌ازای هر کanal فرکانسی گاماتون است به‌نحوی که در آن از یک فیلتر پایین‌گذر با فرکانس قطع ۲/۵ هرتز و سه فیلتر

⁵ Hair-Cells (HCs)

⁶ Adaptation Loops (ALs)

⁷ Onset

⁸ Offset

⁹ Modulation Filterbank (MFB)

آزمایش‌های شناوی‌یی، شامل آستانه‌های شناوی‌یی اندازه‌گیری-شده و نرخ تشخیص همخوان‌ها، و همچنین، شبیه‌سازی‌های مدل پیشنهادی گنجانده شده است. به منظور مقایسه با روش‌های پیشین پیش‌گویی قابلیت فهم گفتار، نتایج پیش-گویی‌های مدل BSIM ارائه شده 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 بازناسان‌ساز گفتار، برای مقایسه دادگان آزمون^۱ با دادگان آموزش^۲ 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.

۱-۱- استخراج ماتریس‌های ویژگی تک‌گوشی و دوگوشی

گفتار و نویه با استفاده از تابع تبدیل سر مربوطه^۳ فیلتر می‌شوند تا سیگنال دارای موقعیت فضایی شبیه‌سازی شود. از این سیگنال‌ها به عنوان ورودی مدل استفاده می‌شود. ماتریس ویژگی پیشنهادی برای بازناسان‌ساز 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 ارائه شده در [20] است. پیش‌پردازش در ابتدا، شامل عبور از فیلتربانک گاماتون برای مدل‌سازی پردازش حلزون گوش انسان است [19].

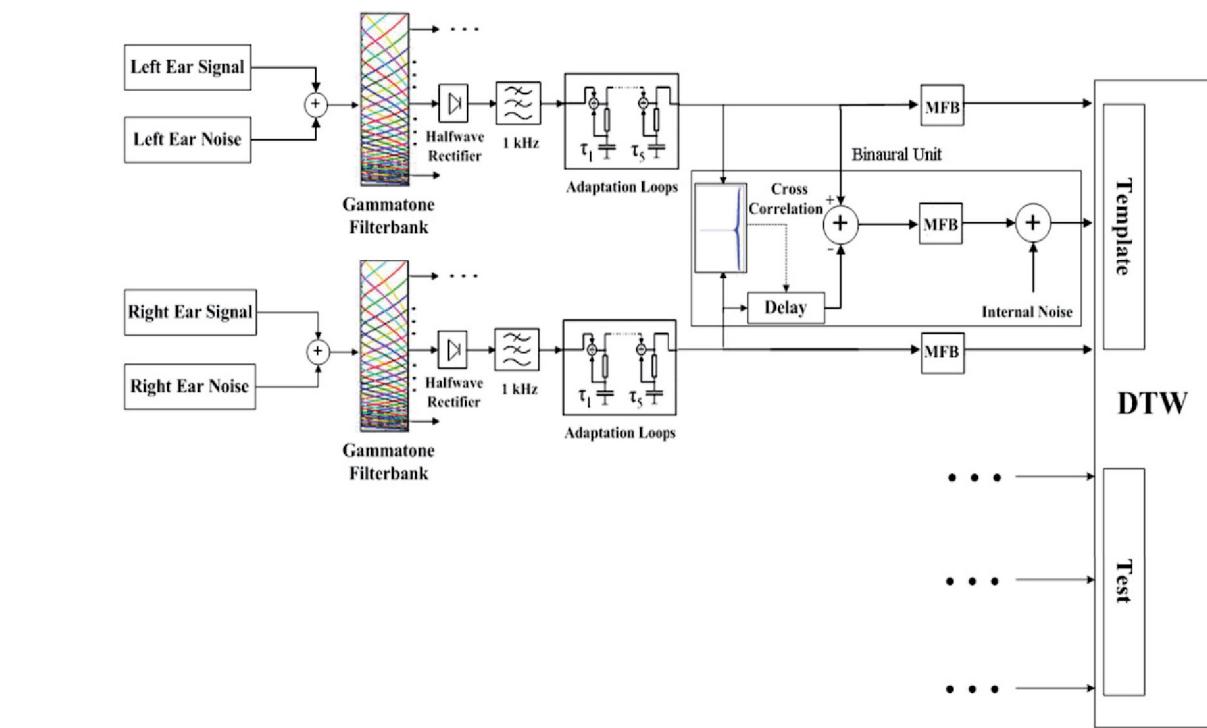
برای پیاده‌سازی فیلتربانک از ۲۷ فیلتر گاماتون که فرکانس مرکزی آنها به طور مساوی در مقیاس ERB^۴ روی محور فرکانس توزیع شده است، استفاده می‌شود. این فیلتربانک، بازه فرکانسی ۲۳۶ تا ۸۰۰۰ هرتز را مشابه مدل ارائه شده در [23] پوشش می‌دهد. خروجی هر فیلتر گاماتون،

¹ Test Data

² Train Data

³ Head-Related Transfer Function (HRTF)

⁴ Equivalent Rectangular Bandwidth (ERB)

⁴ Signal Detection⁵ Interaural Cross-Correlation Function (ICF)⁶ Interaural Time Difference (ITD)

(شکل-۱): مدل پیش‌گویی قابلیت فهم گفتار دوگوشی میکروسکوپی شامل دو مسیر تک گوشی و یک مسیر دوگوشی که در آن از پردازش برابرسازی و حذف بدون تنظیم بهره استفاده شده است. ماتریس ویژگی استفاده شده در DTW سه زیرماتریس دارد: زیرماتریس سیگنال گوش چپ، زیرماتریس گوش راست و زیرماتریس خروجی دوگوشی. قسمت بالای شکل مربوط به پردازش به منظور استخراج سیگنال‌های آموزش برای DTW و قسمت پایین (نشان داده شده با خطوط نقطه‌چین) دارای پردازش مشابه 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 processes 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 section 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.

کشف سیگنال^۴ دوگوشی ارائه شده توسط زریز [34] نیز استفاده شده است. در هر باند فرکانسی،تابع همبستگی مقابله دوگوشی^۵ بین سیگنال گوش‌های راست و چپ، پس از عبور از حلقه‌های انطباق، محاسبه می‌شود. مقدار تأخیر زمانی بین گوشی^۶ از روی تأخیر زمانی در بیشینه مقدار تابع همبستگی مقابله تعیین می‌شود. از آنجا که نرخ تشخیص گفتار در مقدار سیگنال به نوشه مثبت به طور تقریبی صد درصد است، و قابلیت فهم گفتار تنها برای مقادیر منفی سیگنال به نوشه مورد توجه است. در نتیجه، توان سیگنال نوشه همواره در مبحث قابلیت فهم بیشتر از گفتار است. با توجه به اندازه‌گیری‌های انجام شده توسط یورگنز و همکاران [23]، مقدار نرخ تشخیص گفتار اندازه‌گیری شده در نسبت

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

پهنهای باند برای فیلتر پایین گذر 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^۱ سیگنال نام دارد که برای کاهش تعداد نمونه‌ها در بازناسایگر گفتار، نرخ آن تا صد هرتز کاهش داده می‌شود. داو و همکاران برای خروجی مدل پیشنهادی، واحدی به نام «واحد مدل»^۲ 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^۳ است.

در قسمت دوگوشی مدل، از پردازش برابرسازی و حذف استفاده شده است. از پردازش برابرسازی و حذف در مدل

¹ Internal Representation (IR)² Model Unit (MU)³ Sound Pressure Level (SPL)

داخلی گوش راست (IR_{right}) و نمایش داخلی سیگنال دوگوشی ($\text{IR}_{\text{binaural}}$) است. با توجه به وجود ۲۷ باند، فرکانسی و ۴ فیلتر مدولاسیون و ۳ مسیر پردازش سیگنال، ابعاد IR برابر 27×12 خواهد بود. با استفاده از این ماتریس ویژگی پیشنهادی، مدل دوگوشی برای شرایطی که سیگنال نویه همراستا با سیگنال گفتار و از زاویه صفر درجه پخش می‌شود، نتایج مشابه مدل تک‌گوشی [23] را دارد. چون در حالت پخش سیگنال و نویه از زاویه صفر درجه، نمایش‌های داخلی گوش‌های راست و چپ به طور تقریبی مشابه هستند، بعد از عملیات برابرسازی و حذف، انرژی سیگنال دوگوشی به‌دلیل حذف همزمان و کامل گفتار و نویه، مقدار ناچیزی خواهد داشت. اگر $\text{IR}_{\text{train}} = [\mathbf{x}_{\text{train},\text{left}}, \mathbf{x}_{\text{train},\text{right}}, \mathbf{x}_{\text{train},\text{binaural}}]$ نمایش داخلی سیگنال آموزش و $\text{IR}_{\text{test}} = [\mathbf{x}_{\text{test},\text{left}}, \mathbf{x}_{\text{test},\text{right}}, \mathbf{x}_{\text{test},\text{binaural}}]$ نمایش داخلی سیگنال آزمون برای یک زمان خاص باشند، فاصله لورنتزی بین آنها برابر خواهد بود با:

$$\begin{aligned} D_{\text{Lor}} &= \\ &\sum_j \log \left(1 + \frac{1}{2} \left(x_{\text{train},\text{left},j} - x_{\text{test},\text{left},j} \right)^2 \right) \\ &+ \sum_j \log \left(1 + \frac{1}{2} \left(x_{\text{train},\text{right},j} - x_{\text{test},\text{right},j} \right)^2 \right) \\ &+ \sum_j \log \left(1 + \frac{1}{2} \left(x_{\text{train},\text{binaural},j} - x_{\text{test},\text{binaural},j} \right)^2 \right) \end{aligned} \quad (2)$$

که در آن، اندیس ز نشان‌دهنده زیرماتریس‌ها در نمایش داخلی سیگنال در یک زمان خاص است. با درنظر گرفتن این واقعیت که در سناریوی $S_x N_y S_0 N_0$ (یک زمان خاص در زوایای x و y است)، $x_{\text{train},\text{left},j} = x_{\text{train},\text{right},j} = x_{\text{train},j}$ و $x_{\text{test},\text{left},j} = x_{\text{test},\text{right},j} = x_{\text{test},j}$ است و با حذف کامل سیگنال‌ها در برابر سازی و حذف یعنی $x_{\text{temp},\text{binaural},j} - x_{\text{test},\text{binaural},j} \approx 0$ برای رابطه (۲) خواهیم داشت:

$$D_{\text{Lor}} = 2 \times \sum_j \log \left(1 + \frac{1}{2} \left(x_{\text{temp},j} - x_{\text{test},j} \right)^2 \right) \quad (3)$$

به عبارتی، تنها مقدار نهایی فاصله لورنتزی در مقدار ثابت ۲ ضرب می‌شود و مسیر بهینه در شبکه DTW^۴ تغییر نخواهد کرد و از این رو انتظار داریم که نرخ تشخیص مدل در این حالت مشابه مدل تک‌گوشی باشد؛ اما با جداشدن منبع نویه

⁴ DTW Grid

سیگنال به نویه 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 مقدار نرخ تشخیص در سناریوهای دوگوشی بسیار بیشتر از این مقدار باشد؛ لذا، در این مرحله، انتظار می‌رود که تأخیر زمانی بین گوشی مربوط به سیگنال نویه‌ای به تأخیر زمانی بین گوشی سیگنال نزدیک باشد. در مرحله برابرسازی، سیگنال یکی از گوش‌ها با توجه به مقدار تأخیر زمانی بین گوشی جایه‌جا می‌شود و از سیگنال گوش دیگر کسر می‌شود. بنابراین، عمل برابرسازی تنها در بعد زمانی انجام و از برابرسازی بهره^۱ صرف‌نظر می‌شود. در برخی از مدل‌های دوگوشی، از جمله مدل ارائه شده 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^۲ 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» نویه داخلی به سیگنال‌های خروجی تمامی باندهای مدولاسیون افزوده می‌شود و کاهش نرخ نمونه‌برداری، مشابه سیگنال‌های تک‌گوشی، تا مقدار صد هرتز صورت می‌پذیرد. از نویه داخلی و عدم برابرسازی بهره، به‌منظور اعمال ابهام به مدل و محدود کردن پردازش کامل دوگوشی، استفاده شده است.

۲-۲-ماتریس ویژگی پیشنهادی و بازناساگر گفتار

از بازناساگر DTW [30]، با معیار فاصله لورنتزی^۳ به عنوان واحد تصمیم‌گیری مدل استفاده شده است. ماتریس ویژگی مورد استفاده در DTW، ماتریسی ترکیبی و دارای سه زیرماتریس است:

$$\text{IR} = [[\text{IR}_{\text{left}}], [\text{IR}_{\text{right}}], [\text{IR}_{\text{binaural}}]], \quad (1)$$

که در آن IR ماتریس ویژگی مورد استفاده در DTW در یک زمان خاص را نشان می‌دهد. این ماتریس شامل زیرماتریس‌های نمایش داخلی گوش چپ (IR_{left} ، نمایش

¹ Gain Equalization

² Uncertainty

³ Lorentzian Distance

۳-۱-دادگان صوتی

دادگان گفتاری، دارای ساختار «واکه-همخوان-واکه»^۴ 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 نوچهای، بلامانع است. دادگان «واکه-همخوان-واکه»⁵، دارای واکه‌های یکسان در ابتدا و انتهای 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 /ا/، /ت/، /آ/، /ک/، /ف/، /س/، /ب/، /پ/، /و/، /ام/، /لن/، /اش/ و /ال/ می‌باشد که در بین یک جفت یکسان از ۵ واکه /آ، آ، آی، آ/ 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 صدای مابین صدای دو همخوان /ت/ و /س/ می‌باشد به عنوان 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 شنوایی‌سنجدی استاندارد در بازه ۲۵۰ هرتز تا ۸ کیلوهرتز، بیشتر از بیست دسیبل در مقیاس سطح شنوایی^۶ نیست.

۳-۳-شرایط آکوستیکی

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

از منبع سیگنال گفتار در فضای خروجی واحد پردازش دوگوشی دارای انرژی قبل 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 بازناساگر گفتار با استفاده از مفهوم «کشف‌کننده بهینه»^۱ مرتفع می‌شود [14]. در اینجا فرض می‌شود که در مرحله کشف سیگنال، آگاهی کامل از خود سیگنال هدف وجود دارد. از این رویکرد در بازناساگر گفتار مدل میکروسکوپی هلوب [20] نیز استفاده شده است. در این مدل، شکل‌موج‌های آموزش در مرحله آموزش به استثنای نوفه افزوده شده به آنها، مشابه شکل‌موج‌های سیگنال آزمون است.

در این مقاله، برای پیاده‌سازی فیلتربانک گاماتون، از کدهای موجود در پایگاه اینترنتی دانشگاه الدنبورگ^۲ استفاده شده که در آن از روش [19] برای پیاده‌سازی استفاده شده است. مدل سلول‌های مویی (HCs) و حلقه‌های انطباق (Al)s، با استفاده از فایل‌های MEX موجود در تارنمای همین دانشگاه شبیه‌سازی شده است. کدهای مربوط به فیلتربانک مدولاسیون، بخش پردازش دوگوشی و DTW توسط نویسنده‌گان مقاله تهیه شده است.

۳-دادگان و شرایط انجام آزمایش‌های

شنوایی و آزمون مدل

در این پژوهش، عملکرد افراد دارای شنوایی عادی در تشخیص صحیح همخوان‌ها^۳ 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. به علاوه، به عنوان مرجع مقایسه، نتایج پیش‌گویی‌های مدل BSIM ارائه شده توسط بویتلمن و همکاران [5] مورد استفاده قرار خواهد گرفت.

¹ Optimal Detector

² Oldenburg

³ Consonants

⁴ Vowel-Consonant-Vowel (VCV)

⁵ Hearing Level (HL)

چهارصد میلی ثانیه به منظور 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 نیم پنجره‌های هن صد میلی ثانیه‌ای برای پنجره کردن بخش‌های ابتدایی و انتهایی سیگنال‌های گفتار و نوفه استفاده می‌شود [23]. در پایان، هر کدام از سیگنال‌های گفتار و نوفه توسطتابع سر مربوط به خود، فیلتر می‌شود.

اندازه‌گیری قابلیت فهم گفتار با استفاده از نرم‌افزاری که در محیط GUI-MATLAB تهیه شده، صورت گرفته است. شنوondگان در اتفاقی 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 شنوondگان با هدف آشنایی آنها پخش می‌شود. تابع روان-سنجدی^۴ رابطه زیر را برای با مجموعه بسته‌ای از دادگان در نظر می‌گیریم:

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

که در آن s شب تابع روان-سنجدی، SRT آستانه ادراک گفتار، L توان گفتار به نوفه و g احتمال تشخیص صحیح یک واج به صورت تصادفی است. تابع روان-سنجدی 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 به صورت مجموعه بسته‌ای از دادگان انجام نشود و شنوondگان توانایی حدس و انتخاب تصادفی صحیح را نداشته باشند، مقدار $g=0$ خواهد بود. در این صورت به ازای توان گفتار به نوفه برابر با آستانه ادراک گفتار، یعنی $L=SRT$ ، $P(L, SRT, s) = 50\%$. در $L=SRT$ بود. با توجه به اینکه آزمایش‌ها با مجموعه بسته‌ای از دادگان انجام شده‌اند، با درنظر گرفتن $\frac{1}{13}=g$ ، نرخ تشخیص در انجام اینکه آزمایش‌ها با مجموعه بسته‌ای از دادگان $L=SRT$ در حدود ۵۴٪ است. چون در DTW، یکی از گزینه‌ها به‌حتم انتخاب می‌شود، به منظور مقایسه بهتر مدل با اندازه‌گیری‌ها از روش افزایش-کاهش وزن دهنده شده^۵ [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 شنوond است. اندازه‌گیری از مقدار

شنوندگان پخش می‌شوند. منبع تداخلی (ICRA-noise) یک نویز ایستان شبه‌گفتار^۱ است [15] که از زوایای مختلف برای شنوondگان پخش می‌شود. این منبع در شدت صوت ثابت ۶۵ دسیبل توسط دستگاه شنوودسنج کالیبره شده برای شنوondگان پخش می‌شود. در این پژوهش از هفت سناریوی پخش نوفه در زوایای صفر، ۳۰، ۴۵، ۶۰، ۹۰، ۱۰۵ و ۱۳۵ درجه 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، به لحاظ به کارگیری آن در پژوهش انجام شده توسط پیسیگ و همکاران [28]، و نیز مقدار کم آستانه ادراک گزارش شده 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 به منظور ایجاد اکوستیک جهتی، فیلتر شده‌اند. از توابع تبدیل سر بدون پژواک تهیه شده توسط سر مصنوعی KEMAR به این منظور استفاده شده است [2]. سیگنال‌های گفتار نوفه‌ای دوگوشی از طریق دستگاه شنوایی‌سنجدی AC40 [10] که به رایانه وصل است، برای شنوondگان پخش شده را در شدت این دستگاه می‌توان شدت سیگنال پخش شده را در شدت صوت^۲ مورد نظر تنظیم نمود.

۴-۳- سیگنال تحریک و روش اندازه‌گیری

آستانه ادراک گفتار

دادگان «واکه-همخوان-واکه» به گروه‌های تقسیم شده‌اند که تفاوت آنها تنها در همخوان وسط است. با استفاده از این تقسیم‌بندی که شنوondه امکان انتخاب یکی از همخوان‌ها را از میان سیزده همخوان دارد، آزمایشی با مجموعه بسته‌ای از دادگان^۳ انجام می‌پذیرد. بسته‌بودن آزمایش به معنی امکان انتخاب فرد آزمایش شونده 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 نوف، چهارصد میلی ثانیه قبل از سیگنال گفتار صورت می‌پذیرد. بنابراین، چهارصد میلی ثانیه سکوت به ابتدای دادگان گفتاری افزوده می‌شود. زمان

⁴ Psychometric Function

⁵ Weighted Up and Down (WUD)

¹ Stationary Speech-Shaped Noise (SSN)

² Speech Pressure Level (SPL)

³ Closed Test



۴- شبیه‌سازی و نتایج آزمایش‌های شنوازی

در آزمایش شناوری نخست، آستانه‌های ادراک گفتار، با استفاده از یک روش استاندارد، برای افراد مختلف اندازه گیری شده و از مقادیر بهدست آمده از تمامی افراد در هر سناریوی دوگوشی میانگین‌گیری شده است. برای ارزیابی شبیه‌سازی‌ها، از معیارهای مریع ضریب همبستگی و میانگین قدر مطلق خطا بین مقادیر آستانه اندازه گیری شده و پیش‌گویی شده در تمام سناریوها استفاده شده است. حالات ایده‌آل برای پیش‌گویی‌های مدل، مقدار مریع ضریب

تشخیص داده شود، ذخیره می‌شود. برای مرحله بازناسی در مدل، نویه‌ای متفاوت از نویه اضافه شده به سیگنال‌های آموزش در همان نسبت سیگنال به نویه، این‌بار به هر کدام از سیگنال‌های گفتار آزمون افزوده می‌شود که یکی از این سیگنال‌های آزمون همان سیگنال استفاده شده در الگوی آموزش ذخیره شده است. در هر دو مرحله ذخیره الگوی آموزش و مرحله بازناسی، بعد از محاسبه نمایش داخلی برای سیگنال تولید شده، ماتریس‌های ویژگی مربوط به چهارصد میلی ثانیه پخش نویه، که قبل از شروع گفتار افزوده شده است، حذف می‌شود. این عمل به منظور در نظر گرفتن تنها اطلاعات مربوط به دادگان گفتار در بازناساگر گفتار صورت می‌پذیرد. با محاسبه فاصله هر الگوی سیگنال آزمون با الگوهای آموزشی ذخیره شده و تعیین الگویی با کمترین فاصله در DTW، نرخ تشخیص صحیح واج در نسبت سیگنال به نویه‌ای مختلف محاسبه می‌شود.

در این پژوهش، آستانه ادراک گفتار در شبیه‌سازی‌ها مشابه اندازه گیری‌ها، مقدار سیگنال به نویه‌ای در نظر گرفته شده است که در آن نرخ تشخیص برابر 54% می‌شود. روش جستجوی مقدار آستانه به این صورت است که نسبت سیگنال به نویه در مدل با گام‌های $1/0$ دسیبل تغییر داده می‌شود تا این که نزدیکترین نرخ تشخیص به 54% حاصل گردد. با انتخاب این گام جستجو، با توجه به نتایج شبیه‌سازی‌ها، حداقل انحراف از مقدار 54% برابر $1/5$ به دست آمده است. برای هر سناریوی دوگوشی، شبیه‌سازی‌ها با نمونه‌های انتخابی متفاوت از پوشش نویه SSN بیست‌بار تکرار می‌شوند تا میانگین خوبی از نرخ‌های تشخیص به دست آید.

سیگنال به نویه پنج- دسیبل و با گام کاهشی چهار دسیبل برای پاسخ‌های صحیح شروع می‌شود. این گام کاهشی بعد از دومین و چهارمین برگشت¹، 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 ذکر شده در مقدار $1/17$ ضرب می‌شود. از این‌رو مقدار نهایی گام افزایشی برای پاسخ‌های ناصحیح برابر $1/17$ دسیبل است. گام‌های یادشده 1 و $1/17$ دسیبل تا پخش تمام 65 داده صوتی ادامه می‌یابند. مقدار متوسط سیگنال به نویه پخش شده در تمامی اجراهای² با گام 1 و $1/17$ دسیبل به عنوان آستانه ادراک گفتار در نظر گرفته می‌شود. برای هر زاویه پخش نویه، ترتیب پخش تمامی ادراک‌ها به صورت تصادفی با نرم‌افزار تغییر داده می‌شود. از افراد خواسته شده است که با استفاده از موشواره بر روی گزینه مورد نظر کلیک نمایند و در صورتی که اصلاً قادر به تشخیص نباشند، یکی از گزینه‌ها را حدس بزنند.

از آنجایی که مدل میکروسکوپی، قابلیت پیش‌گویی نرخ تشخیص و نیز ماتریس ابهام³ هر کدام از واج‌ها را به صورت جداگانه دارد، آزمایش شناوری دوم بعد از تخمین مقدار میانگین مقادیر آستانه ادراک گفتار در هر سناریو انجام شده است. پس از تخمین مقدار آستانه، تمامی دادگان این با برآورد نسبت سیگنال به نویه ثابت (برابر با آستانه ادراک گفتار تخمینی) برای شنوندگان در هر سناریو پخش می‌شوند. روش انجام این اندازه گیری مشابه رویکرد به کار گرفته شده در گروانچیزاده و همکاران^[1] است.

۳-۵- پیش‌گویی‌های مدل

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

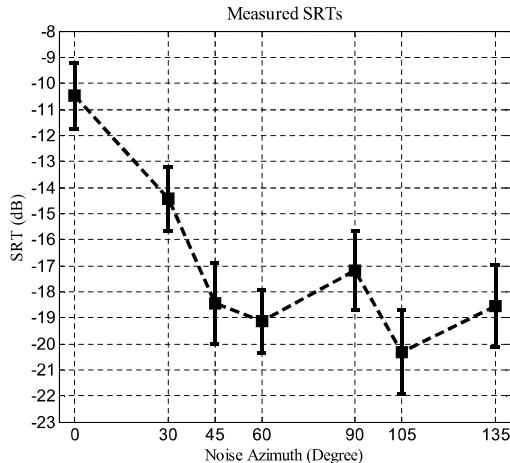
در مرحله آموزش مدل، نمونه‌هایی از نویه، در نسبت سیگنال به نویه مورد جستجو، به دادگانی که تفاوت آنها تنها در واج وسط است، افزوده می‌شود و نمایش داخلی این سیگنال‌های نویه‌ای به عنوان الگوهایی که می‌باشد

¹ Reversal

² Runs

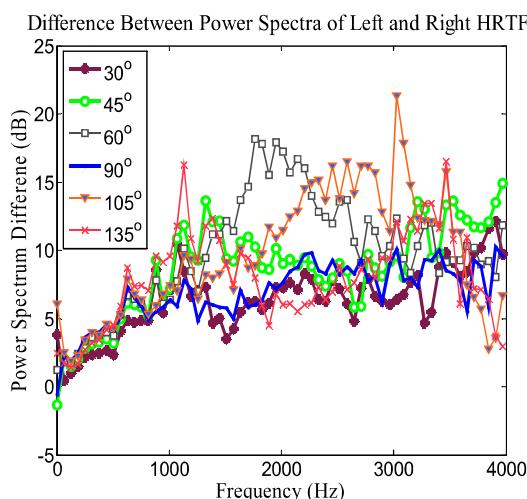
³ Confusion Matrix

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»^۵ است. این پدیده بهدلیل کاهش اثر سایه سر در زوایای نود و ۲۷۰ درجه است.



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

(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°.

⁵ Bright-Spot Phenomenon

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، به کار رفته‌اند.

۱-۴-نتایج آزمایش‌های شنوایی

میانگین و انحراف معیارهای آستانه‌های ادراک گفتار اندازه‌گیری شده ادراک گفتار برای دوازده فرد دارای شنوایی عادی در شکل (۲) نشان داده شده است. محور افقی، زوایای پخش سیگنال نوشه و محور عمودی، آستانه‌های ادراک گفتار در اندازه‌گیری‌ها را نشان می‌دهد. میله‌های خطأ^۱، 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^۲ تأثیر عامل زاویه بر اندازه‌گیری‌ها، توسط تحلیل واریانس^۳ (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 به دست‌آمده به‌طور تصادفی که بیانگر عدم وابستگی نتایج به زاویا است، با استفاده ازتابع 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 اندازه‌گیری شده ادراک گفتار است ($F(6,77)=58.78, p < 0.0001$). برای بررسی جامع‌تر، تحلیل واریانس یک‌طرفه^۴ 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، در حالی که 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

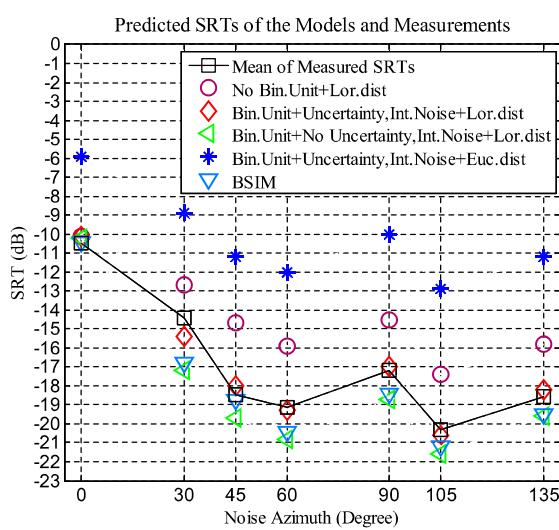
¹ Error Bars

² Statistical Significance

³ Analysis of Variance (ANOVA)

⁴ One-Way ANOVA

این مقدار مرجع برابر با میانگین SII های به دست آمده برای تمامی همخوانها و نوفه های جمع شونده، با نسبت سیگنال به نوفه مساوی $10/4$ - دسیبل در سناریوی S_0N_0 است. تمامی مقادیر آستانه های ادراک گفتار برای مدل پیشنهادی در شکل (۴) از میانگین گیری آستانه های پیش گویی شده برای ۶۵ همخوان و در ۲۰ تکرار اجرای برنامه، با استفاده از نمونه های متفاوت نوفه، به دست آمده است.



(شکل-۴): مقادیر آستانه ادراک گفتار اندازه گیری شده و پیش گویی شده در زوایای افقی مختلف منبع نوفه ای. برای مدل پیشنهادی تأثیر وجود 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.

همان گونه که در این شکل مشاهده می شود، برای حالت S_0N_0 نوع معیار فاصله، وجود یا عدم وجود واحد دوگوشی و نیز ابهام و نوفه داخلی تأثیر چندانی بر پیش گویی های مدل ندارند. این امر بیشتر با توجه به این واقعیت است که در S_0N_0 مزیت استفاده از گوش بهتر وجود ندارد؛ به علاوه، با توجه به حذف سیگنال هدف و نوفه به صورت همزمان در این سناریو، ماتریس ویژگی دوگوشی اهمیت کمتر و ماهیت نویزی است. برخلاف BSIM که در آن SII برابر مقدار مرجع $1/1$ تنظیم شده است، مدل پیشنهادی این مزیت را دارد که در آن نیازی به تعیین شرایط مرجع نیست و مدل به طور ذاتی دارای پیش گویی های نزدیک به اندازه گیری در حالت S_0N_0 است.

در این زوایا، منبع نوفه به طور تقریبی در راستای خط اتصال دو گوش قرار می گیرد و موجب کاهش اختلاف سطح توان دوگوشی^۱، و درنتیجه، کاهش مزیت دوگوشی در این زوایا نسبت به زوایای 60° و 105° درجه می شود. به منظور مقایسه میزان عدم شباهت بین توابع تبدیل گوش های راست و چپ، اختلاف دامنه پاسخ فرکانسی دو گوش در مقیاس دسیبل برای هر زاویه در شکل (۳) رسم شده است. فاصله فرکانسی 1500 تا 3000 هرتز، اهمیت زیادی در تشخیص همخوانها دارد. با توجه به شکل، اختلاف بین طیف توان های مربوط به توابع تبدیل سر گوش های راست و چپ در زوایای 60° و 105° درجه بیشتر از سایر زوایا است. مقادیر اختلاف برای زاویه 90° درجه کمتر از 60° و 105° درجه و نزدیک به 30° و 45° درجه است. در مطالعات قبلی، کمترین مقدار آستانه ادراک گفتار در زوایای افقی -100 - 120 درجه گزارش شده است [8]. کمترین مقدار آستانه ادراک گفتار اندازه گیری شده که در 105° درجه است، این نتایج را تأیید می کند.

۴-۲-پیش گویی های مدل

پیش گویی های مدل پیشنهادی در شکل (۴) نشان داده شده است. محور عمودی، آستانه های ادراک گفتار پیش گویی شده توسط مدل را نشان می دهد. همچنین، میانگین آستانه های ادراک گفتار اندازه گیری شده 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 (به ترتیب، مثلث رو به چپ و لوزی) و ۳- تأثیر استفاده از معیار فاصله افکلیدسی به جای فاصله لورنتزی بر پیش گویی ها (ستاره) در این شکل نشان داده شده است. همچنین، به منظور مقایسه، پیش گویی های مدل شناخته شده 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 پژواک دار و بدون پژواک، به خوبی پیش گویی می کند [6]. برای BSIM، با توجه به مقدار آستانه ادراک گفتار اندازه گیری شده در سناریوی S_0N_0 که برابر با $-10/4$ - دسیبل است، مقدار SII مرجع، برابر با $1/1$ تنظیم می شود.

^۱ Interaural Level Difference (ILD)

است. اگر از معیار اقلیدسی به جای لورنتزی استفاده شود، خطای پیش‌گویی‌ها به $6/6$ دسیبل می‌رسد. بدون استفاده از واحد پردازش دوگوشی، مقدار خطای تخمین برابر $2/5$ دسیبل است. همچنین، اگر ابهام و نوفه داخلی در پردازش دوگوشی در نظر گرفته نشوند، خطابه $1/4$ دسیبل می‌رسد. پیش‌گویی مدل BSIM نیز دارای میانگین قدرمطلق خطای قابل قبول 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 اندازه‌گیری شده پیش‌گویی می‌کند. به دلیل 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 خطای (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 نوبه داخلی	1.4	0.95
BSIM	1	0.95

در شکل (۵)، مقایسه مزیت دوگوشی به دست آمده 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 اندازه‌گیری شده در سه پژوهش مشابه و مزیت‌های دوگوشی اندازه‌گیری و پیش‌گویی شده را در این پژوهش نشان می‌دهد. این نتایج از اندازه‌گیری‌های پلات [29]، کازنتینو [13] و پیسیگ [28] برگرفته شده‌اند. در این پژوهش‌ها، مشابه کار ارائه شده در این مقاله، از نویه شبکه‌گفتار به عنوان سیگنال تداخل استفاده شده اما از دادگان گفتاری متفاوتی در هر کدام استفاده شده است.

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 به دست آمده برای این حالت و اندازه‌گیری‌ها، مشابه یکدیگر است. از این‌رو، پیش‌گویی‌های مدل، تأثیر عامل گوش بهتر در شناوی دوگوشی را به خوبی 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، مدل سازی شده و مقادیر پیش‌گویی آستانه ادراک گفتار به مقادیر اندازه‌گیری شده نزدیک می‌شود. در شکل (۴)، تأثیر ابهام و نوفه داخلی نیز بر پیش‌گویی‌های مدل نشان داده است. هنگامی که ابهام و نوفه داخلی در پیاده‌سازی مدل نادیده گرفته شوند (مثلث رو به پایین)، پیش‌گویی‌های مدل با توجه به تغوری برابر سازی و حذف، مقادیر آستانه ادراک گفتار کمتری را نشان می‌دهند. با اعمال ابهام و نوفه داخلی (لوزی)، انطباق خوبی بین اندازه‌گیری و پیش‌گویی‌ها به دست می‌آید. نوع معیار فاصله استفاده شده در بازناساگر گفتار، تأثیر قابل توجهی 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، در حالی که معیار لورنتزی برای دادگان دارای توزیع لورنتز^۱ بهینه است [27]. در مدل تک‌گوشی یورگنز و همکاران نشان داده شده است که مقادیر فاصله که در DTW مورد استفاده قرار می‌گیرند، دارای هیستوگرامی هستند که با استفاده ازتابع 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 خطای^۲ 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 در نظر گرفتن واحد پردازش دوگوشی دارای ابهام و نوفه داخلی، میزان خطای تخمین بسیار کم و برابر $0/4$ دسیبل

¹ Lorentz Distribution

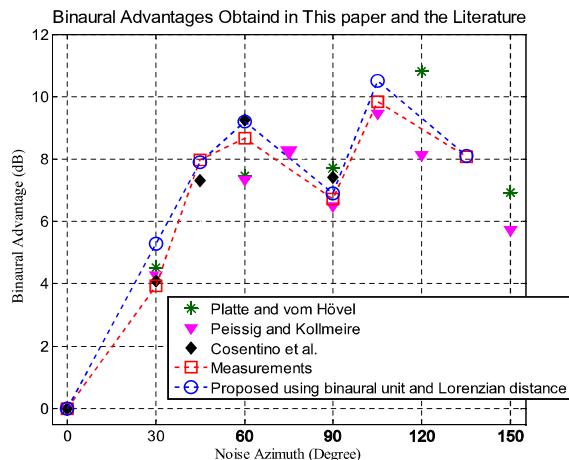
² Mean Absolute Error (MAE)



سمت چپ مربوط به ماتریس ابهام پیش‌گویی برای مدل دارای واحد دوگوشی به همراه ابهام و نویز داخلی با استفاده از معیار فاصله لورنتزی (لوژی در شکل (۴)) است. سطر و ستون‌های ماتریس ابهام به ترتیب، مربوط به همخوان‌های پخش‌شده و تشخیص‌داده شده هستند. عناصر قطری این ماتریس‌ها، میانگین نرخ تشخیص صحیح همخوان‌ها را نشان می‌دهند و عناصر غیرقطري، بیان گر 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 تعاظت زیادی با الگوهای تشخیص در اندازه‌گیری‌ها ندارد. به عنوان مثال، شنوندگان و پیش‌گویی‌های مدل، برای همخوان‌های /ت، ام، اس/ و /ش/ نرخ تشخیصی نزدیک به 100% را نشان می‌دهند. این نتایج، مطابق با پیش‌گویی‌های مدل بورگنز و همکاران [23] و نتایج گراوانچی‌زاده و همکاران [11] است. همچنین، نرخ تشخیص پایین برای همخوان‌های /ب، ام، ان، اف/ و /ال/، دامنه گسترده‌تری از انتخاب‌ها را در ماتریس ابهام نشان می‌دهد. میانگین نرخ‌های تشخیص‌داده شده در اندازه‌گیری‌ها برای هفت سناریو، برابر با 52% است. این مقدار، نتایج اندازه‌گیری‌های آستانه ادراک گفتار با استفاده از روش وفقی WUD را تأیید می‌کند.

به منظور مقایسه الگوی پاسخ‌های صحیح، مقادیر^۲ 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 به خوبی پیش‌گویی می‌کند.

به منظور مقایسه الگوی ابهام هر کدام از همخوان‌ها، دویاره از ضرایب^۲، ولی این بار به عنوان معیار شباخت 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 به دلیل مقادیر بالای نرخ تشخیص صحیح هر همخوان در مقدار آستانه ادراک گفتار است. برای همخوان‌های /ت، ام، اس/ و /ش/ مقادیر^۲ نزدیک به مقدار واحد هستند؛ در حالی که همخوان‌های /ب، /او/ و /ال/ کمترین مقادیر همبستگی را نشان می‌دهند. به علاوه، برخلاف نرخ تشخیص پایین /ام، ان/ و /ال/، این همخوان‌ها همبستگی بالایی را به دلیل شباهت عناصر غیر- قطر اصلی دارند.



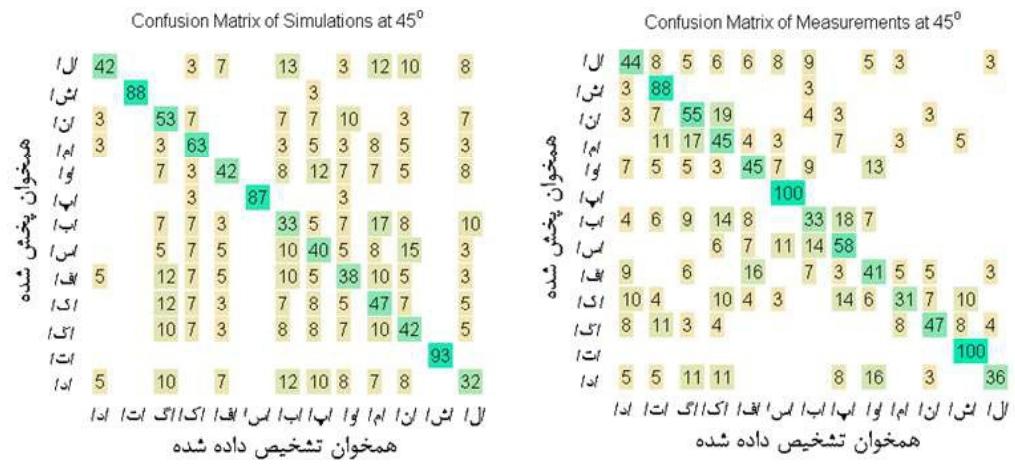
(شکل-۵): مزیت دوگوشی به دست آمده در اندازه‌گیری‌ها و پیش‌گویی‌های مدل پیشنهادی به همراه نتایج مطالعات قبلی.

(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 نقطه نظر 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 -90° انجام شده است که در شکل (۵)، تنها برای زوایای $+90^{\circ}$ مثبت، این مقادیر اندازه‌گیری ذکر شده است. اندازه‌گیری‌های ما بیشترین مزیت را در زوایای افقی 60° و 105° درجه نشان می‌دهد. مطابق این شکل روال و مزیت دوگوشی به دست آمده برای زوایای مختلف در نتایج این مقاله تا حد زیادی با نتایج مطالعات قبلی همخوانی دارد. افت در مقدار مزیت دوگوشی برای زاویه 90° 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 به دست آمده در آزمایش شنوایی نخست، مخلوط گفتار و نویه در سناریوهای دوگوشی و در سیگنال به نویه ثابتی در هر سناریو برای شنوندگان پخش شد. مقدار سیگنال به نویه در آزمایش شنوایی دوم برابر با آستانه ادراک گفتار برای همان سناریو در آزمایش شنوایی نخست تنظیم شده است. در شکل (۶) به عنوان نمونه ماتریس‌های ابهام برای اندازه‌گیری‌ها و پیش‌گویی‌های مدل برای سناریوی پخش نویه از زاویه 45° درجه نشان داده شده است. در اینجا، شکل سمت راست مربوط به ماتریس ابهام اندازه‌گیری و شکل



(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

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 بدون معنا استفاده شود.

مدل میکروسکوپی دیگری که پیش‌پردازش‌های متفاوتی از پژوهش انجام‌شده در این مقاله را دارد، مربوط به مدل «کشف دید کوتاه»^۲ 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

نویسنده‌گان 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.

6-References

۶-مراجع

[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 بازناساگر خودکار گفتار»، *فصلنامه علمی پژوهشی پردازش علائم و داده‌ها*، دوره ۱۲، شماره ۱، صفحات ۷۹-۹۰، ۱۳۹۴.

[۱] M. Geravanchizadeh, A. Fallah, and A. Eteraf Oskoie, "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.

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

² Glimpse Detection

سال ۱۳۹۶ شماره ۱ پیاپی ۳۱

۵-نتیجه‌گیری و پیشنهاد ادامه کار

در این مطالعه، مدل میکروسکوپی دوگوشی برای پیش‌گویی قابلیت فهم گفتار بر مبنای فیلتربانک مدولاسیون و بازناساگر 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^۱ 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 بازناساگر 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 بازناساگر گفتار، امکان تحلیل ماتریس‌های ابهام نیز فراهم شده است. با استفاده از این مدل، مقدار میانگین قدر مطلق خطای در پیش‌گویی مقادیر آستانه ادراک گفتار برابر 0.4° دسیبل به‌دست آمده است. مریع ضریب همبستگی بین آستانه‌های ادراک گفتار پیش‌گویی شده و اندازه‌گیری شده نیز برابر 0.97 است.

مدل پیشنهادی در شرایط بدون پژواک و با حضور یک منبع نویه شبکه گفتار، قابلیت پیش‌گویی خوبی دارد. از آنجا که دادگان استفاده شده در این پژوهش دارای طول محدود هستند، امکان ارزیابی پژواک برای آنها وجود ندارد؛ اما یک

¹ Diotic

- 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*. Massachusetts, 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.interacoustics.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.

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 فلاخ 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 شناوری است.
نشانی رایانمه ایشان عبارت است از:

ali.fallah@tabrizu.ac.ir



مسعود گروانچی‌زاده مدرس
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،
به ترتیب، در سال‌های ۱۳۷۴ و ۱۳۸۰ از دانشگاه-
Ruhr- 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
جدازی سیگنال گفتار، پردازش سیگنال‌های تصادفی و
پردازش سیگنال دوگوشی است.
نشانی رایانمه ایشان عبارت است از:

geravanchizadeh@tabrizu.ac.ir

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," *Perception 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 sprachverständlichkeitsemessungen 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",