شناسایی احساس سیگنال گفتار فارسی با استفاده از تحلیل ویژگی‌های طیفی–فرکانسی

مریم مومنی[۱]

[۱]استادیار گروه مهندسی برق، دانشکده فنی و مهندسی، دانشگاه اراک، m-momeni@araku.ac.ir

چکیده

امروزه تشخیص احساس از گفتار در مواردی که ارتباط متقابل انسان و ماشین وجود دارد مورد توجه قرار گرفته است. با وجود تلاش‌های زیاد در این زمینه همچنان فاصله زیادی بین احساسات طبیعی انسان و درک کامپیوتر نسبت به آن وجود دارد. دلیل اصلی این موضوع نیز عدم توانایی رایانه در درک احساس کاربر است. هدف از این مقاله، طراحی یک سیستم تشخیص احساس از گفتار بر روی پایگاه داده گفتار احساسی فارسی که شامل ۵ احساس خوشحالی، تنفر، ترس، ناراحتی و عصبانیت است. در این مقاله، پس از استخراج داده‌های چهار بعدی مقیاس، نرخ (سرعت)، زمان و فرکانس گفتار به کمک سیستم مدل شنوایی گوش انسان، داده دو بعدی مقیاس و فرکانس حاصل شد که بیشینه مقدار این داده‌ها به‌عنوان بردار ویژگی استفاده شد. در نهایت با استفاده از طبقه‌بند ماشین بردار پشتیبان احساس این پایگاه داده طبقه‌بندی شدند. نتایج آزمایش‌ها نشان می‌دهد الگوریتم پیشنهادی عملکرد قابل قبولی در مقایسه با سیستم‌های تشخیص خودکار احساسات از گفتار در زبان فارسی ارائه می‌دهد.

کلیدواژه

شناسایی خودکار احساس، گفتار زبان فارسی، ویژگی‌های طیفی–فرکانسی.

مقدمه

سیگنال گفتار سریع‌ترین و طبیعی‌ترین روش ارتباط بین انسان-ها می‌باشد. بر این اساس گفتار به‌عنوان یک روش سریع و کارآمد برای تعامل انسان و کامپیوتر بکار گرفته می‌شود. اکنون تلاش-های زیادی در زمینه تشخیص گفتار انجام شده است. با وجود پیشرفت‌های زیاد در این زمینه، فاصله زیادی بین تعامل طبیعی انسان و کامپیوتر وجود دارد [۱]. از طرفی، مشکلاتی نیز در این سیستم‌ها وجود دارد. نخستین مشکل در تحلیل احساسات، دشوار بودن تفکیک ویژگی‌های احساس می‌باشد زیرا ویژگی‌های مختلف در احساسات، به افراد مختلف و حالت فعلی گوینده در زمان بیان جملات احساسی مانند حوصله، وضعیت درونی، نگرش و خصوصیات شخصیتی فرد به شدت وابسته است. دومین مشکل پیچیدگی احساسات است که اغلب به‌هنگام برقراری ارتباط بین اشخاص، احساسات کامل، خالص و پایه بروز نمی‌کنند بلکه معمولا ترکیبی از احساسات در یک لحظه ممکن است بروز نماید [۲]. بنابراین احساس پدیده‌ای مبهم، پیچیده و مرکب است و جداسازی، تشریح و تشخیص آن بسیار دشوار می‌باشد. نحوه بروز احساسات در گفتار به فرهنگ و زبان، محتوای گفتار، جنسیت و سن گوینده و بسیاری از عوامل دیگر وابسته است. اثر جنسیت و سن بر احساس مسئله‌ای مهم برای پژوهش در زمینه تشخیص احساس از روی گفتار است. اطلاعات جنسیتی بر نتایج دسته-بندی احساسات مختلف تاثیر می‌گذارد [۳]. اثر افزایش سن و جنسیت بر تشخیص احساس از روی گفتار و تفاوت جنسیت در رفتارهای احساسی، موضوعاتی است که نظر تعدادی از پژوهشگران را به خود جلب کرده است. بانوان احساس خود را غلیظ‌تر بیان می‌کنند و نسبت به آقایان درک بیشتری از احساسات دارند. همچنین بانوان احساسات را با شدت بیشتر و فرکانس بالاتر در گفتار مدوله می‌کنند. از طرفی، آقایان مهارت بیشتری در کنترل احساسات دارند. تمامی این مسائل روند تشخیص احساس از گفتار را پیچیده‌تر می‌کنند [۴]. سیستم تشخیص احساس از روی گفتار از دیدگاه تشخیص الگو شامل سه بخش می‌باشد: ۱-استخراج ویژگی، ۲-کاهش ویژگی و ۳-طبقه‌بندی. مهمترین چالش‌های تشخیص احساس از روی گفتار عمدتاً به مرحله استخراج ویژگی مرتبط می‌باشند. دلیل اصلی، نامعلوم بودن ویژگی‌های مؤثر در تشخیص احساس و تنوع صوتی می‌باشد که خود ناشی از وجود کلمات متنوع، گوینده‌های مختلف، سبک صحبت کردن و نرخ صحبت کردن متفاوت است [۵].

تاریخ دریافت: ۱۳۹۷/۰۷/۱۶ تاریخ پذیرش: ۱۳۹۸/۰۴/۲۷

پیشینه تحقیق

همانطور که قبلا هم بیان شد، یکی از پارامترهای مهم در شناسایی خودکار زبان و احساس، پایگاه داده و زبان آن داده‌ها می‌باشد. پایگاه‌های داده متفاوتی با زبان‌های متفاوت مثل آلمانی، اسلواکی، ترکی و غیره وجود دارد و تحقیقات زیادی در زبان‌های متفاوت و با روش‌ها و الگوریتم‌های گوناگون در راستای استخراج ویژگی و طبقه‌بندی انجام شده است [۱۲-۶،۴] که در جدول ۱، به این تحقیق‌ها اشاره شده است.

تحقیقاتی هم برای شناسایی احساس در زبان فارسی و بر روی پایگاه داده فارسی PESD[۱] انجام شده است که در مقاله [۱۳] خلاصه شده است. از آنجاییکه در مقاله حاضر هم از همین پایگاه داده استفاده می‌شود، در جدول ۱، برخی از تحقیقات انجام شده بر روی این پایگاه داده ارائه شده است.

جدول ۱. برخی از مطالعات انجام شده در زمینه شناسایی احساس با استفاده از سیگنال گفتار

مراجع	ضرایب ویژگی مورد استفاده	طبقه‌بند	دقت (بر حسب درصد)		
[۴]	موجک بیونیک[۳]	SVM[۲]	۷۸		
[۶]	ویژگی‌های طیفی و MLS[۵]	SOM[۴]	۶۳/۸۳		
[۷]	ویژگی‌هایی چون انرژی، فرمنت[۷] و نرخ عبور از صفر[۸]	GMM[۶]	۵۲/۷		
[۸]	استخراج ویژگی در سطح واج	SVM	۶۷/۷		
[۹]	ویژگی‌های متداول در احساسات مختلف از پایگاه داده به زبان‌های دانمارکی، آلمانی و صربستانی	SVM	۶۳/۲		
[۱۰]	بررسی پارامتر نرخ صحبت، در تشخیص احساسات از گفتار	MLP[۹]	۶۱/۴۲		
[۱۱]	پیچ[۱۱]، انرژی، ضریب اغتشاشات فرکانسی محلی[۱۲]، ضریب اغتشاشات محلی دامنه[۱۳]، نرخ عبور از صفر، ضرایب کپستروم[۱۴]	ترکیب دو طبقه‌بندی کننده SVM و HMM[۱۰]	۶۵		
[۱۲]	در این مقاله، ابتدا از ویژگی‌های عروضی و طیفی بر مبنای سطح برانگیختگی طبقه‌بندی می‌شود؛ سپس احساس‌های با سطح برانگیختگی یکسان با استفاده از ویژگی‌های پیشنهادی دینامیکی غیرخطی از یکدیگر جدا می‌شوند. ویژگی‌های دینامیکی غیرخطی از روی مشخصات هندسی فضای بازسازی شده سیگنال گفتار استخراج می‌شوند.	طبقه‌بند مبتنی بر مدل احساسی برانگیختگی- جاذبه	۹۲/۳۴		
[۱۴]	فرکانس گام، انرژی، تعداد عبور از صفر، ضرایب پیشگویی خطی، ضرایب کپستروم و مشتقات آن، ضریب اغتشاشات فرکانسی محلی، ضریب اغتشاشات محلی دامنه، فرمنت‌ها، ضرایب فوریه، مینیمم، ماکزیمم، انحراف معیار	SVM	۷۸		
		KNN[۱۵]	۹۱		
[۱۵]	ویژگی‌های پروزودیک[۱۶] مثل پیچ، شدت صدا و مشخصه‌های عمومی سیگنال گفتار	شبکه عصبی چند لایه	۷۸		
[۱]	ویژگی‌های گوناگونی مثل پیچ، نرخ عبور از صفر، ضرایب کپسترال مل، انرژی، فرمنت و PLP[۱۸]	LDA[۱۷]	برلین	مرد	زن
				۷۴/۳۲	۷۸/۶۴
			پایگاه داده فارسی پیشنهادی مقاله	۴۷/۲۸	۵۵/۷۴
[۱۶]	ویژگی‌های پایه مثل پیچ، شدت، ضرایب کپسترال مل، انرژی، فرمنت و غیره	SVM	فارسی	۹۹/۴۴	
		NN[۱۹]	برلین	۸۷/۲۱	

۱ Persian Emotional Speech Database
۲ Support Vector Machine
۳ Bionic Wavelet Transform
۴ Self-Organizing Map
۵ Mean of Log-Spectrum
۶ Gaussian Mixture Model
۷ Formants
۸ Zero Crossing
۹ Multi-Layer Perceptron
۱۰ Hidden Markov Model
۱۱ Pitch
۱۲ Jitter
۱۳ Shimmer
۱۴ Mel Frequency Cepstral Coefficients
۱۵ K-Nearest Neighbors
۱۶ Prosodic
۱۷ Linear Discriminant Analysis
۱۸ Perceptual Linear Predictive
۱۹ Neural Network

[۱۷]	ویژگی‌های مبتنی بر طیف یعنی MFCC و LPC[۲۱]	MLP, SVM, RBF[۲۰], GMM, Bays	بالای ۹۰ درصد
[۱۸]	۲۸ ویژگی صوتی مانند سه فرمنت اول، ویژگی‌های طیفی و دامنه بیان احساسی	SVM	۹۷
		NN	۹۳

از طرفی، برخی مراجع به بررسی تاثیر عوامل مختلف مثل جنسیت [۱۹]، اثر فشرده‌سازی سیگنال در تشخیص احساس [۲۰]، تاثیر نوع احساس بر روی ویژگی‌های سیگنال گفتار فارسی [۲۳-۲۱]، نویز محیط [۲۵،۲۴] را بررسی کرده‌اند. از آنجاییکه سیستم مدل شنوایی انسان [۲۶] بعنوان بهترین تشخیص‌دهنده سیگنال گفتار عمل می‌کند لذا می‌تواند بعنوان بهترین مدل برای برای شناسایی سیگنال گفتار استفاده شود. سیستم مدل شنوایی انسان در شناسایی و تحلیل سیگنال گفتار، شبیه‌سازی، کد کردن و اندازه‌گیری کیفیت صدا کاربرد دارد [۲۷]. در مراجع [۳۰-۲۸] به کمک سیستم مدل شنوایی واج و سیگنال‌های گفتار مورد بررسی و تحقیق قرار گرفته‌اند. در [۳۱] با توجه به ساختار شنوایی انسان بررسی‌های روانشناسی و نوروفیزیولوژیک یک مدل محاسباتی برای تجزیه و تحلیل صوت ارائه شده است. این مدل ساختاری یکپارچه از نمایش ویژگی‌های طیفی و زمانی سیگنال را ارائه می‌دهد. مدل ریاضی کاملی ارائه می‌شود که چگونگی تغییر سیگنال‌های پیچیده در سطوح مختلف مدولاسیون طیفی-زمانی را نشان می‌دهد و آن را با مدل‌های موجود در پردازش شنوایی مقایسه می‌کند. در [۳۲] با مقایسه مدل مدولاسیون طیفی-زمانی با دقت بالا نشان می‌دهد که تنظیم دقیق مدولاسیون توانایی جداسازی صدای‌های طبیعی را افزایش دهد. در [۳۳] مدل محاسباتی صوتی معرفی شده است که نمایش طیفی-زمانی (MTF)[۲۲] بر اساس خصوصیات برجسته داده ارائه می‌دهد تا ارتباط بین MTF را در راستای ارزیابی توانایی شناسایی صوت در نویز و شرایط مختلف نشان دهد. در[۳۴] مدولاسیون‌های طیفی-زمانی که برای درک توسط انسان لازم است را مشخص می‌کند. در [۳۵] مدولاسیون طیفی-زمانی را به عنوان روشی قدرتمند برای استخراج ویژگی از سیگنال گفتار معرفی کرده و نتایج استخراج ویژگی توسط این روش را با ویژگی‌های استخراج شده از ضرایب کپسترال فرکانس مل مقایسه کرده است.

با توجه به اینکه سایر مقالات با استفاده از مدولاسیون طیفی-زمانی نتایج قابل توجهی در پردازش گفتار کسب کرده‌اند، این مقاله بر آن شده است که از این روش در پردازش سیگنال گفتار بیماران آلزایمری و افراد سالم استفاده کند. هدف اصلی از این مقاله پردازش سیگنال گفتار بیمارن آلزایمری توسط مدولاسیون‌های طیفی، زمانی و فرکانسی حاصل از مدل سیستم مدل شنوایی است. این مدولاسیون‌ها در مقیاس‌های مختلفی انجام می‌گیرند و تعیین کننده رفتار پردازش گفتار مانند قابلیت فهم گفتار است، تحقیقات نشان می‌دهد که سیستم مدل شنوایی انسان قادر است مدولاسیون یک بعدی را به خوبی مدولاسیون دو بعدی تشخیص دهد و شناسایی کند [۳۱].

روش

روش کلی مقاله به این صورت است که ابتدا داده‌های موجود در پایگاه داده برای استفاده در مراحل بعدی پیش پردازش و سپس با استفاده از سیستم مدل شنوایی و مدولاسیون طیفی-زمانی حاصل از آن، ویژگی‌های مورد نظر استخراج شده و توسط ماشین بردار پشتیبان طبقه‌بندی می‌شوند.

پیش‌پردازش

نویز موجود در داده های جمع آوری شده توسط فیلتر حذف و سیگنال گفتار نرمالیزه شده است. با توجه به ماهیت غیرایستان سیگنال گفتار توسط پنجره همینگ به قسمت های ۳۰-۵۰ میلی ثانیه با همپوشانی ۵۰٪ تقسیم شده‌اند و سکوت در سیگنال گفتار توسط عبور از صفر و انرژی سیگنال استخراج و حذف شده‌اند. پس از پیش‌پردازش در دو مرحله اخذ طیف سیگنال صوت و آنالیز طیفی-زمانی پردازش سیگنال گفتار ادامه می‌یابد.

سیستم مدل شنوایی

ابتدا تبدیل موجک متعارف از سیگنال حاصل می‌شود ($s(t)$) و سپس تجزیه و تحلیل طیفی توسط فیلتر کوچلر[۲۳] ۱۲۸ فیلتر میان‌گذر با همپوشانی و Q (Q10db =3) ثابت با فرکانس‌های مرکزی که به صورت یکنواخت در محدوده فرکانس لگاریتمی توزیع شده‌اند انجام می‌شود که پاسخ فرکانسی هر فیلتر با $h(t;x)$ مشخص شده است. خروجی این فیلتر کوچلر با $y_{coch}(t,x)$ نمایش داده می‌شود [۳۶]،

[۲۰] Radial Basis Function Kernel
[۲۱] Linear predictive coding
[۲۲] Modulation Transform Function
[۲۳] Cochlear

$$y_{coch}(t,x) = s(t) \otimes_t h(t;x), \quad (۱)$$

که $\otimes_t$ عملگر کانولوشن نسبت به زمان است. خروجی فیلتر کوچلر $y_{coch}(t,x)$ در یک الگوی عصبی شنوایی $y_{AN}(t,x)$ که شامل یک فیلتر بالاگذر و فشرده ساز غیرخطی $g(.)$ و فیلتر پایین گذر $w(t)$ است، قرار می‌گیرد [۳۶]،

$$y_{AN}(t,x) = g\big(\partial_t y_{coch}(t,x)\big) \otimes_t w(t), \quad (۲)$$

مرحله آخر عملکرد یک شبکه مهارکننده جانبی[۲۴]است که اساس کوچلر است. این شبکه به سادگی توسط مشتق مرتبه اول با توجه به محور تونوتپیک[۲۵]و یکسوساز نیم‌موج اعمال می‌شود [۳۶]،

$$y_{LIN}(t,x) = max(\partial_x y_{AN}(t,x), 0), \quad (۳)$$

$$y_{final}(t,x) = y_{LIN}(t,x) \otimes_t \mu(t;x), \quad (۴)$$

خروجی نهایی توسط انتگرال‌گیری در طول یک پنجره کوتاه به دست می‌آید [۳۶]،

$$\mu(t;x) = e^{-t/\tau} u(t), \quad (۵)$$

در این مرحله طیف زمانی و فرکانسی سیگنال گفتار از طریق یک بانک فیلتر که دارای پارامترهای طیفی-زمانی مدولاسیون است و تغییرات زمانی آن‌ها از نرخ آهسته به سریع و تغییرات طیفی آن‌ها از مقیاس‌های باریک به گسترده است تخمین زده می‌شود. حوزه دریافت فرکانس-زمان (STRF[۲۶]) این فیلترها در فرکانس‌های مختلفی در امتداد محور تونوتپیک قرار دارند. پاسخ طیفی-زمانی[۲۷] r_c پایین رونده و بالا رونده به شکل زیر خواهد بود [۳۶]،

$$\begin{aligned} r_{c\Downarrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) &= y(t,x) \otimes_{tx} \big[\big(h_t h_s \\ &- \widehat{h_t}\widehat{h_s}\big) cos(\theta_c + \phi_c) \\ &+ \big(\widehat{h_t} h_s + h_t \widehat{h_s}\big) sin(\theta_c + \phi_c)\big] \end{aligned} \quad (۶)$$

$$\begin{aligned} r_{c\Uparrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) &= y(t,x) \otimes_{tx} \big[\big(h_t h_s \\ &+ \widehat{h_t}\widehat{h_s}\big) cos(\theta_c - \phi_c) \\ &+ \big(\widehat{h_t} h_s - h_t \widehat{h_s}\big) sin(\theta_c - \phi_c)\big] \end{aligned} \quad (۷)$$

که در آن $\otimes_{tx}$ نشان دهنده کانولوشن زمان و مکان است و h_t و h_s به ترتیب نشان دهنده مدولاسیون زمانی و فرکانسی STRF است. θ و ϕ نشان دهنده ویژگی‌های فاز هستند.

فرمول مناسبی از r_c در صورتیکه خروجی دامنه و فاز، تابعی از تبدیل موجک باشند را می‌توان به شکل زیر نیز نشان داد [۳۶]،

$$\begin{aligned} r_{c\Downarrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) &= R\{z_{\Downarrow}\} cos(\theta_c + \phi_c) \\ &+ I\{z_{\Downarrow}\} sin(\theta_c + \phi_c) \\ &= |z_{\Downarrow}| cos(\psi_{\Downarrow} - \theta_c - \phi_c) \end{aligned} \quad (۸)$$

$$\begin{aligned} r_{c\Uparrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) &= R\{z_{\Uparrow}\} cos(\theta_c - \phi_c) \\ &- I\{z_{\Uparrow}\} sin(\theta_c - \phi_c) \\ &= |z_{\Uparrow}| cos(\psi_{\Uparrow} + \theta_c - \phi_c) \end{aligned} \quad (۹)$$

$R\{.\}$ و $I\{.\}$ به ترتیب نشان دهنده قسمت حقیقی و مختلط تابع است. توابع $z \Uparrow$ و $z \Downarrow$ که به صورت تابعی از تبدیل ویولت زمانی h_{TW} و ویولت فرکانسی h_{SW} است در زیر نوشته شده است [۳۶]،

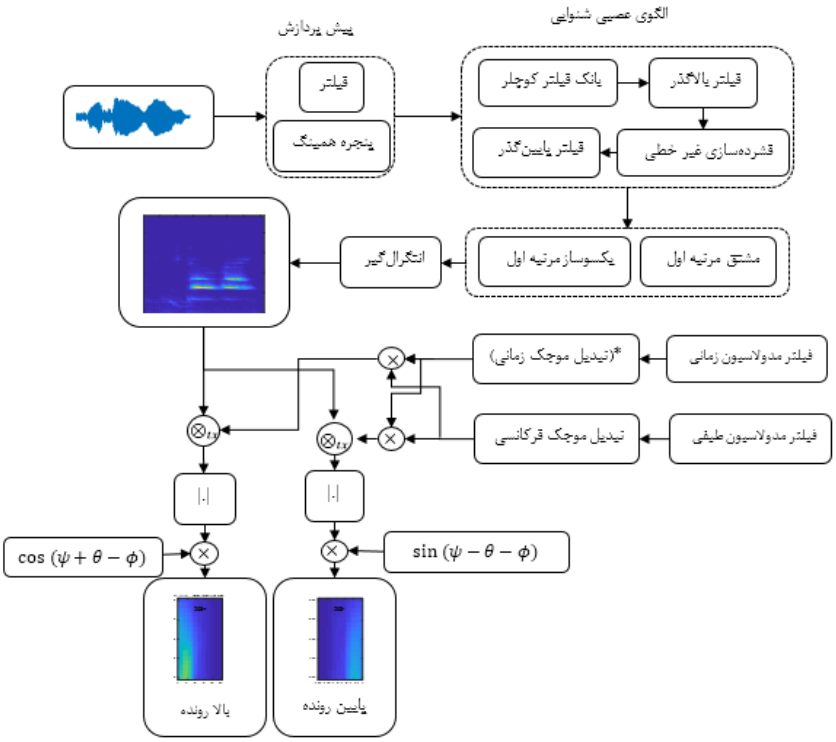
$$\begin{aligned} z_{\Downarrow}(t,x;\omega_c,\Omega_c) &= y(t,x) \otimes_{tx} [h_{TW}(t;\omega_c) h_{SW}(x;\Omega_c)] \\ &= |z_{\Downarrow}(t,x;\omega_c,\Omega_c)| e^{j\psi_{\Downarrow}(t,x;\omega_c,\Omega_c)}, \end{aligned} \quad (۱۰)$$

$$\begin{aligned} z_{\Uparrow}(t,x;\omega_c,\Omega_c) &= y(t,x) \otimes_{tx} [h_{TW}{}^*(t;\omega_c) h_{SW}(x;\Omega_c)] \\ &= |z_{\Uparrow}(t,x;\omega_c,\Omega_c)| e^{j\psi_{\Uparrow}(t,x;\omega_c,\Omega_c)}, \end{aligned} \quad (۱۱)$$

نمایش مدولاسیون زمانی و طیفی از فیلتر کردن اسپکتروم قابل حصول است؛ فیلترها برای مدولاسیون طیفی دارای مقیاس $\Omega = [0.5,1,2,4,8] cyc/oct$ و در مدولاسیون زمانی دارای فرکانس مرکزی $\omega = [1,2,4,8,16,32] Hz$ انتخاب و خروجی هر بانک فیلتر در طول زمان میانگین‌گیری شد.

بلوک دیاگرام مراحل پردازش سیگنال در سیستم مدل شنوایی مطرح شده در این مقاله در شکل (۱) نمایش داده شده است.

[۲۴] Lateral Inhibitory Network
[۲۵] Tonotopic
[۲۶] Spectro-Temporal Receptive Fields
[۲۷] Cortical Representation



شکل ۱. بلوک دیاگرام سیستم مدل شنوایی

پس از اخذ داده چهار بعدی r_c و پاسخ‌های بالا و پایین رونده تبدیل ویولت و پاسخ ضربه فرکانسی، بیشینه مقدار آن‌ها و انرژی اسپکتروم استخراج، تحلیل و با استفاده از SVM طبقه‌بندی شدند.

اساس کار SVM افزایش فاصله بین نمونه‌ها و مشخص نمودن مرز طبقه‌بندی است. این فاصله بعنوان حاشیه شناخته می‌شود و با افزایش آن قادر به تعمیم الگوهای ناشناخته است. راه حل افزایش حاشیه، به SVM این اجازه را می‌دهد تا بیشترین طبقه بندی‌های غیرخطی را در حضور نویز که یکی از مشکلات ASR است، انجام دهد. همچنین SVMها مشکلات همگرایی و پایداری معمول که اکثر شبکه‌های عصبی دارا هستند را ندارند. مفهوم اساسی که در SVM نهفته است کاهش خطای ساختاری است [۳۷]، یک دستگاه یادگیری به‌گونه‌ای انتخاب می‌شود که خطای آزمون را به حداقل برساند که باعث اندازه‌گیری مناسبی از تعمیم‌پذیری دستگاه است و تخمینی از نسبت بردارهای طبقه-بندی شده بر کل بردارهای آموزش است [۳۸].

نتایج

پایگاه داده

در زمینه تشخیص احساس از گفتار ۲ نوع پایگاه داده طبیعی و مصنوعی وجود دارد [۳۹] پایگاه داده طبیعی در واقع گفتگوی روزمره مردم است و تجزیه و تحلیل این پایگاه داده نتایج مطلوبی را نشان می‌دهد اما دسترسی به گفتگوهای روزمره ساده نیست. در پایگاه داده مصنوعی از بازیگران خواسته می‌شود که جملات مختلفی را با احساس‌های گوناگون بیان کنند. این کار باعث می‌شود نتایج بدست آمده با آنچه در واقعیت رخ می‌دهد فاصله داشته باشد [۱۴، ۴۰].

در این مقاله از پایگاه داده گفتار احساسی زبان فارسی PSED استفاده شده است که مجموعه‌ای جامع و معتبر از گفتار احساسی برای زبان فارسی است که در دانشگاه فرای برلین[۲۸] ساخته و سپس رواسازی شده است [۴۱]. برای ساخت این مجموعه دو بازیگر فارسی‌زبان (یک زن و یک مرد) ۹۰ جمله را در پنج آهنگ عاطفی خشم، شادی، غم، ترس، چندش و نیز خنثی طی شرایط خاصی در سه دسته همگون[۲۹](در بخش همگون پایگاه داده، جملات بیان شده در یک حالت با حالت عاطفی دیگر بیان نشده است)، ناهمگون[۳۰](در بخش ناهمگون پایگاه داده، جملات بیان شده در یک حالت با حالت عاطفی دیگر

[۲۸] Freie Universität Berlin
[۲۹] Congruent
[۳۰] Incongruent

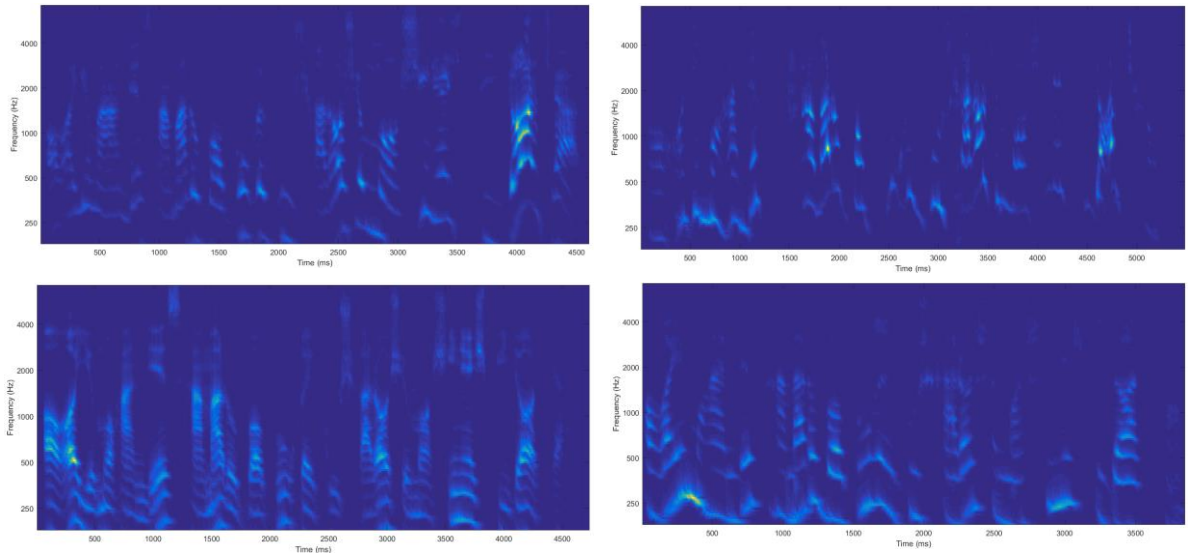
بیان شده است) و پایه در یک استودیو تخصصی ضبط صدا، زیر نظر یک زبانشناس و یک متخصص اکوستیک در شهر برلین آلمان اجرا کردند. متن این ۹۰ جمله پیشتر توسط ۱۱۲۶ فارسی‌زبان در دو مطالعه رفتاری جداگانه روانسازی شده بودند. حاصل این کار ۴۷۲ جمله صوتی با آهنگ‌های عاطفی متفاوت است. رواسازی محتوایی این جملات صوتی در یک مطالعه رفتاری توسط ۳۴ فارسی‌زبان مورد ارزیابی قرار گرفته و ۴۶۸ جمله صوتی که درصد تشخیص‌شان بالای ۷۱/۴۲ درصد بود (پنج بار بالاتر از سطح شانس) به‌عنوان جملات معتبر (رواسازی شده) در نظر گرفته شده‌اند. همچنین تجزیه و تحلیل اکوستیکی این ۴۶۸ جمله صوتی نمایانگر تفاوت معناداری در زمینه شدت، زیر و بمی صدا و کشش بیان جملات در پنج آهنگ عاطفی مورد مطالعه است. در مجموع حدود ۱۳ دقیقه و ۲۰ ثانیه از گفتار ضبط شده در دسترس است [۴].

در شکل (۲) اسپکتروگرام سیگنال گفتار نشان داده شده است که نشان دهنده شدت فرکانس در زمان معلوم است و تغییرات فرکانسی بر حسب زمان است. شکل (۲) ردیف بالا، اسپکتروگرام سیگنال گفتار حالت عصبانیت مرد (چپ)، زن (راست) و ردیف پایین سیگنال گفتار خنثی مرد(چپ)، زن(راست) را نمایش می-دهد.

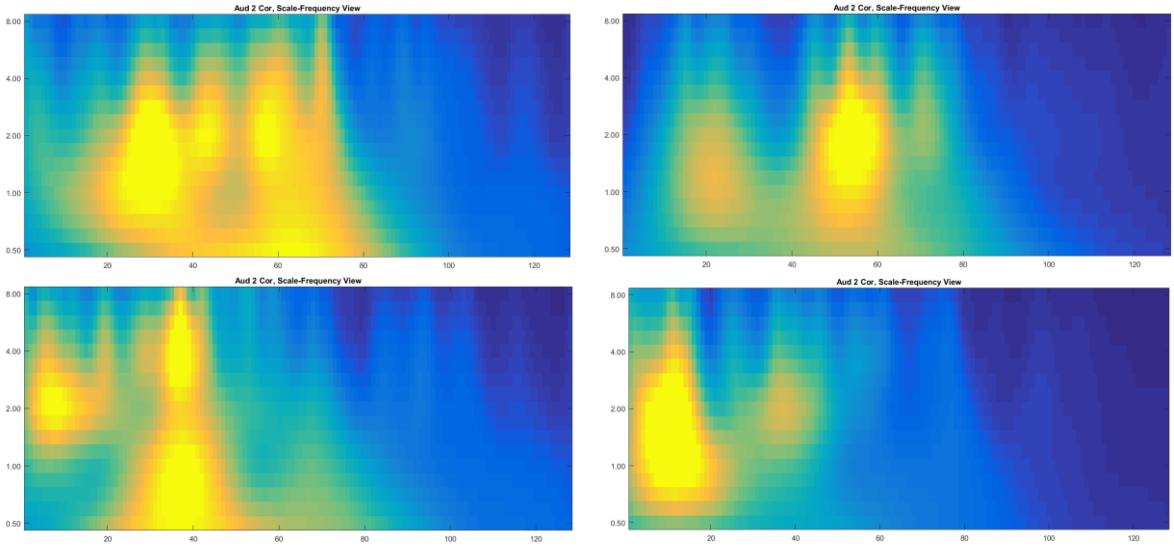
بخاطر چندبعدی بودن پاسخ r_c، نمایش آن علاوه بر تفسیر مشکل آسان نمی‌باشد از اینرو در راستای نمایش بهینه‌تر، انتگرال‌گیری در راستای تغییرات زمان انجام می‌شود که منجر به نمایش مقیاس-فرکانس می‌شود. شکل (۳) ردیف بالا، تغییرات مقیاس-فرکانس سیگنال گفتار حالت عصبانیت مرد (چپ)، زن (راست) است. ردیف پایین سیگنال گفتار حالت خنثی مرد (چپ)، زن (راست) را نمایش می‌دهد.

در این مرحله نیز بخاطر چندبعدی بودن پاسخ r_c و در راستای نمایش بهینه‌تر، انتگرال‌گیری در راستای تغییرات فرکانس انجام می‌شود که منجر به نمایش مقیاس-زمان می‌شود. شکل (۴) ردیف بالا، تغییرات مقیاس-زمان سیگنال گفتار حالت عصبانیت مرد (چپ)، زن (راست) و ردیف پایین سیگنال گفتار حالت خنثی مرد(چپ)، زن(راست) را نمایش می‌دهد.

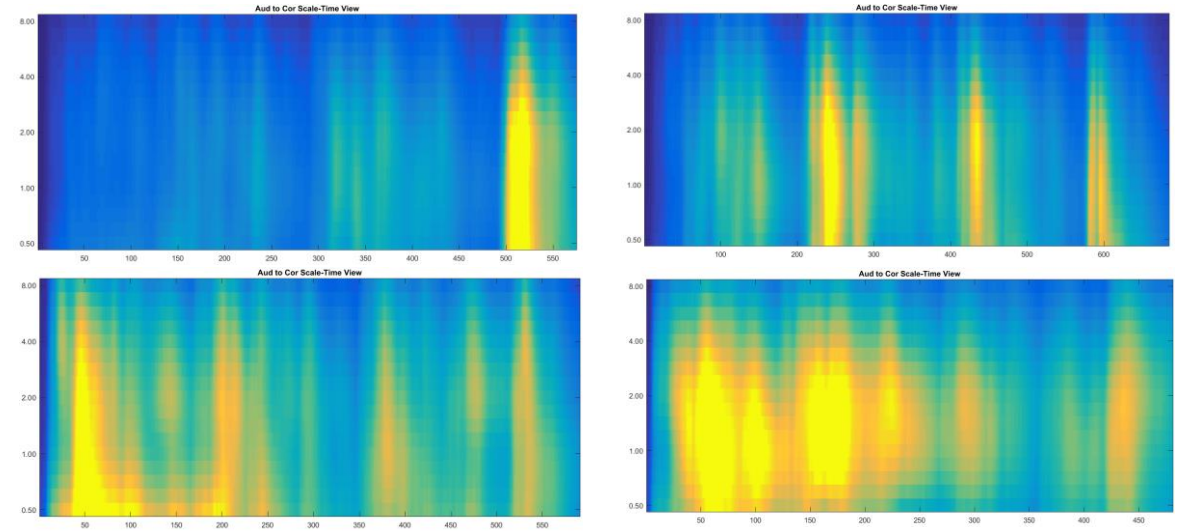
نتایج تجربی



شکل ۲. ردیف بالا، نمایش اسپکتروگرام سیگنال گفتار حالت عصبانیت مرد (چپ)، زن (راست) و ردیف پایین سیگنال گفتار خنثی مرد (چپ)، زن (راست)



شکل ۳. ردیف بالا، نمایش مقیاس-فرکانس سیگنال گفتار حالت عصبانیت مرد (چپ)، زن (راست) و ردیف پایین سیگنال گفتار حالت خنثی مرد(چپ)، زن(راست)



شکل ۴. ردیف بالا، نمایش مقیاس-زمان سیگنال گفتار حالت عصبانیت مرد (چپ)، زن (راست) و ردیف پایین سیگنال گفتار حالت خنثی مرد(چپ)، زن(راست)

ارزیابی نتایج

به کمک تحلیل چندرزولوشنی مدل شنوایی انسان می‌توان تحلیل سیگنال گفتار با حالات احساسی متفاوت داده‌های پایگاه داده فارسی را داشته باشیم. از آنجاییکه به دنبال تشکیل بردار ویژگی برای شناسایی جنسیت و احساس در زبان فارسی و با ویژگی‌های زمان و فرکانس بوده‌ایم، پس از استخراج داده‌های دو بعدی مقیاس و فرکانس به کمک انتگرال‌گیری نسبت به نرخ و زمان حاصل شد که بیشینه مقدار این داده‌ها بعنوان بردار ویژگی استفاده شد.

در نهایت با استفاده از طبقه‌بند SVM احساس این پایگاه داده طبقه‌بندی شدند. نسبت داده‌های آموزش و تست ۶۰ به ۴۰ درصد انتخاب و با استفاده از طبقه‌بند SVM با کرنل تابع پایه شعاعی (RBF) داده‌ها طبقه‌بندی شده و به کمک رابطه (۱۲) دقت عملکرد طبقه‌بندی ارزیابی شده است:

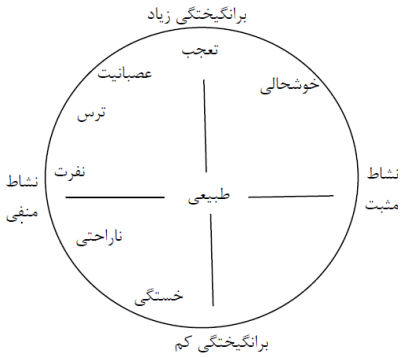
$$\text{دقت} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (۱۲)$$

که در آن، TP مثبت درست، TN منفی درست، FP مثبت نادرست و FN منفی نادرست است. جدول ۲ نتایج طبقه‌بندی الگوریتم پیشنهادی را نشان می‌دهد.

جدول ۲. نرخ شناسایی (بر حسب درصد) احساس گفتار در پایگاه داده فارسی بخش همگون و ناهمگون، زن (الف) و (د)، مرد (ب)، (ه) و زن و مرد (ج)، (و)

		ناراحتی	شادی	ترس	نفرت	عصبانیت	میانگین
همگون	(الف)	۸۰	۸۳/۳۳۳۳	۶۶/۶۶۶۷	۸۳/۳۳۳۳	۵۸/۳۳۳۳	۷۴/۳۳۳۳۲
	(ب)	۸۰	۱۰۰	۳۳/۳۳۳۳	۸۳/۳۳۳۳	۷۵	۷۴/۳۳۳۳۲
	(ج)	۹۰	۱۰۰	۸۳/۳۳۳۳	۷۵	۱۰۰	۹۱/۴
ناهمگون	(د)	۹۰	۸۰	۸۰	۷۰	۵۰	۷۴
	(ه)	۹۰	۱۰۰	۹۰	۱۰۰	۷۰	۹۰
	(و)	۶۰	۸۰	۱۰۰	۷۰	۶۰	۷۴

به‌طور کلی احساس‌ها دارای دو بعد برانگیختگی و میزان نشاط [۱] است (شکل ۵).



شکل ۵. مدل دوبعدی احساس [۱]

برانگیختگی به میزان انرژی لازم برای ادای یک احساس خاص اشاره دارد. براساس برخی از مطالعات فیزیولوژیکی از مکانیزم تولید احساسات مشخص شده است که سیستم عصبی برای احساس‌های خوشحالی، عصبانیت و ترس برانگیخته می‌شود اما بر اساس برانگیختگی نمی‌توان احساس‌ها را از هم تفکیک نمود. به‌عنوان مثال دو احساس عصبانیت و خوشحالی هر دو برانگیختگی بالایی دارند اما آنها از نظر احساسی کاملا متفاوت هستند. این تفاوت از نظر بعد میزان نشاط می‌باشد [۱]. بنابراین طبقه‌بندی بین احساس‌ها با برانگیختگی بالا و کم چالش برانگیز است. در این مقاله، با الگوریتم پیشنهادی نرخ شناسایی بین احساس‌های با برانگیختگی بالا مثل شادی-عصبانیت، با نشاط کم مثل ترس-نفرت و ناراحتی-نفرت بر روی بخش ناهمگون پایگاه داده مرد به ترتیب ۸۰، ۱۰۰ و ۱۰۰ درصد شده است.

با توجه به توضیحات ارائه شده، شکل ۵ و همچنین نتایج حاصل مشاهده می‌شود که دقت در ربع اول (برانگیختگی بالا و نشاط بالا) و سوم (برانگیختگی کم و نشاط کم) نسبت به ربع دوم (برانگیختگی بالا و نشاط کم) بالاتر است. بطور مثال در جدول ۲ (الف) و (ب)، نرخ شناسایی احساس شادی و ترس به ترتیب بالاترین و پایین‌ترین دقت را دارند.

از آنجا که یکی از عوامل تاثیر گذار در چگونگی بروز احساس، جنسیت گوینده است، نرخ شناسایی احساس در جدول ۲ (الف) و (ب) و همچنین در جدول ۲ (د) و (ه) قابل مقایسه است چراکه مقایسه همزمان جنسیت و احساس در جدول ۲ (ج) و (و) انجام می‌شود.

از طرفی برانگیختگی بالا معادل انرژی بیشتر و نشاط بالا معادل فرکانس بالاتر است. بطور مثال در شکل ۳، ردیف بالا، چپ، انرژی و فرکانس نسبت به شکل ۳، ردیف پایین، چپ بیشتر است که حاکی از برانگیختگی بیشتر عصبانیت نسبت به حالت خنثی است.

بحث و نتیجه‌گیری

با افزایش روزافزون تراکنش میان انسان و ماشین در بسیاری از زمینه‌ها، تحقیقات زیادی برای ایجاد ارتباط بهتر و آسان‌تر بین این دو در حال انجام است. از جمله می‌توان به برقراری ارتباط کلامی بین انسان و ماشین، درک احساسات انسانی از سوی ماشین و ارائه واکنش مناسب به آن اشاره کرد. سیستم‌های تشخیص احساسات از گفتار، بخش مهمی از تحقیقات رو به رشد در حوزه پردازش گفتار را به خود اختصاص داده‌اند. در این مقاله، به تشخیص احساس از گفتار پایگاه داده احساسی فارسی که شامل ۵ احساس خوشحالی، تنفر، ترس، ناراحتی و عصبانیت پرداخته شد. در این مقاله از جعبه‌ابزار NSL [۴۲] بهره گرفته و سیگنال گفتار احساسی پایگاه داده فارسی به آن داده شد. نتیجه چهاربعدی مقیاس، نرخ، زمان و فرکانس حاصل شد و با انتگرال-گیری نسبت به هر کدام از ابعاد داده خروجی می‌توان به شکل-های جدیدی دست یافت. همانطور که در شکل (۲) قابل مشاهده است، بطور مثال، در فرکانس ۱۰۰۰ هرتز، تغییرات فرکانس در طول زمان برای دو شکل ردیف بالا و همچنین دو شکل ردیف

پایین کاملا متفاوت است که حاکی از تفاوت جنسیت مرد و زن است. همچنین تغییرات فرکانسی در این فرکانس خاص برای دو شکل ستون چپ و دو شکل ستون راست نیز کاملا متفاوت و متمایز است که حاکی از تاثیر حالت عصبانیت در سیگنال گفتار مرد و زن است. پاسخ قوی در شکل (۳) در گستره فرکانسی ۲۰ تا ۸۰ و ۴۰ تا ۶۰ هرتز به ترتیب برای مرد(ردیف بالا، چپ) و زن (ردیف بالا، راست) است. ردیف پایین سیگنال گفتار حالت خنثی مرد(چپ)، زن(راست) را نمایش می‌دهد که پاسخ قوی این شکل در گستره فرکانسی ۲۰ تا ۶۰ و ۰ تا ۲۰ هرتز به ترتیب برای مرد و زن است. همچنین اگر گستره فرکانسی دو شکل ستون چپ و دو شکل ستون راست را مقایسه کنیم، بخاطر حالت عصبانیت مرد و زن گستره فرکانسی آنها متفاوت می‌باشد. همانطور که در شکل (۴) قابل مشاهده است، پاسخ قوی در این حالت در گستره زمانی دو شکل ردیف بالا و پایین و همچنین در مقایسه دو شکل ستون چپ و راست کاملا متفاوت است.

پس از استخراج داده‌های چهار بعدی مقیاس، نرخ (سرعت)، زمان و فرکانس گفتار به کمک مدل شنوایی گوش انسان [۲۶]، داده دو بعدی مقیاس و فرکانس حاصل شد که بیشینه مقدار این داده‌ها به‌عنوان بردار ویژگی استفاده شد. در نهایت با استفاده از طبقه‌بند SVM احساس این پایگاه داده طبقه‌بندی شدند. روش پیشنهادی بر روی دادگان گفتار احساسی فارسی ارزیابی شد. همانطور که قبلا هم اشاره شد نتایج آزمایش‌ها نشان می‌دهد که الگوریتم پیشنهادی عملکرد قابل قبولی در مقایسه با سیستم‌های تشخیص احساس موجود پایگاه داده فارسی بیان شده در پیشینه تحقیق ارائه می‌دهد. همانطور که از بررسی منابع مرتبط برمی‌آید در پایگاه داده استفاده شده در این مقاله، انتخاب ویژگی و طبقه-بندی کننده [۱۸-۱۴، ۲۳-۲۰] در نرخ شناسایی احساس نقش دارند که در اکثر مقالات ویژگی‌های آکوستیک یکسانی استفاده شده است و لذا روش استخراج ویژگی حاضر نسبت به ویژگی‌های متداول جدید است و شامل اطلاعات مدلاسیون فرکانس، زمان، طیف و نرخ است. در بررسی با نتایج مطالعات فوق، در مقالات با نتایج بالای ۹۰ درصد، حدود بالای ۴۰ ویژگی استفاده شده است که همان بردار ویژگی بر روی دیگر پایگاه‌های داده نتایج کمتری را داشته است [۱۶، ۱۷]. در مقاله حاضر تنها از بیشینه نقاط ویژگی‌های حاصل از سیستم مدل شنوایی استفاده شده است که برای بهتر شدن کارایی سیستم پیشنهاد می‌شود ویژگی‌های دیگری از سیگنال گفتار (همچون ضرایب کپسترال مل، فرمنت-ها، جیتر، نرخ عبور از صفر، انرژی و غیره) به سیستم اضافه گردد و با ترکیب آن با ویژگی‌های مبتنی با روش پیشنهادی به نرخ بازشناسی بهتری رسید. در این مقاله، تنها از دو بعد داده حاصل از مدل شنوایی انسان استفاده شد و در کارهای آتی سعی بر آن است که از دیگر ابعاد حاصل از این مدل بهره گرفته و اطلاعات دقیق‌تری از احساس دادگان استخراج و تفسیر کرد. همچنین استفاده از شبکه‌های عصبی و سیستم‌های فازی به‌عنوان طبقه-بند توصیه می‌شود.

تشکر و قدردانی

این مطالعه تحت قرارداد پژوهشی شماره ۹۴/۱۱۹۹۵ دانشگاه اراک از پشتیبانی مادی و معنوی بهره مند شده است.

مراجع

[۱] ح. مروی، ز. اسماعیلیان، "معرفی پایگاه داده فارسی جهت تشخیص احساس از روی گفتار،" بیست و یکمین کنفرانس مهندسی برق ایران، مشهد، دانشگاه فردوسی مشهد، ۱۳۹۲.

[2] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, D.M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," Proc. IEEE, Trans. Biomedical Eng., vol. 47(7), pp. 829-837, 2007.

[3] T. Pao, C. Wang. "A study on the search of the most discriminative speech features in the speaker dependent speech emotion recognition," Proc. IEEE Fifth Int. Sym. Parallel Architectures, Algorithm and Programming, 2012, pp. 157-162.

[۴] ر. یوسفی‌نژاد، ب. حاجی باقر نایینی، م. شفیعیان، " تشخیص احساس از سیگنال گفتار با استفاده از موجک بیونیک،" نشریه علمی ترویجی صوت و ارتعاش، سال پنجم، شماره نهم، ۱۳۹۵، ۷۱-۸۳.

[5] M.El. Ayadi, M.S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, pp. 572–587, 2011.

[6] E.M. Albornoz, D.H. Milone, H.L. Rufiner, "Spoken Emotion recognition using hierarchical classifier," Computer Speech and Language, vol. 25(3), pp. 556–570, 2011.

[7] B. Yang, M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Processing, vol. 90, pp.1415-1423, 2010.

[8] D. Bitouk, R. Verma, A. Nenkova, "class level spectral features for emotion recognition," Speech Communication, vol. 52, pp.613-625, 2010.

[9] A. Hassan, R. Damper, "Classification of emotional speech using 3DEC hierarchical

classifier," Speech Communication, vol. 54, pp. 903-916, 2012.

[10] D. Philippou-Hübner, B. Vlasenko, R. Böck, A. Wendemuth, "The Performance of The Speaking Rate Parameter in emotion recognition from speech," IEEE, International conference on Multimedia and Expo Workshops, Melbourne, VIC, Australia, 9-13 July 2012.

[11] M. Gaurav, "Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech," 2008 IEEE Spoken Language Technology Workshop, Goa, India, 15-19 Dec. 2008.

[۱۲] ع. حریمی، ع. احمدی فرد، ع. شهزادی، خ. یغمایی، "تشخیص احساس از روی گفتار با استفاده از طبقه‌بند مبتنی بر مدل و ویژگی‌های دینامیکی غیر خطی،" نشریه، ب- مهندسی کامپیوتر، سال ۱۵، شماره ۲، تابستان ۱۳۹۶، ۱۴۵-۱۵۲.

[13] O.M. Nezami, P. Jamshid Lou, M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection, " Language Resources and Evaluation, March 2019, Vol. 53, Issue 1, pp 1–16, 2019.

[۱۴] ب. ابراهیم پور، ح. محمودیان، "تشخیص احساسات گفتار با استفاده از انتخاب ویژگی بر اساس مدل های بازگشتی،" هفتمین کنفرانس ملی مهندسی برق و الکترونیک ایران، دانشگاه آزاد اسلامی گناباد، ۲۸ و ۲۹ مرداد ماه ۹۴.

[15] M. Hamidi, M. Mansoorizade, "Emotion Recognition From Persian Speech With Neural Network," International Journal of Artificial Intelligence & Applications (IJAIA), vol.3(5), pp. 107-112, 2012.

[16] A. Shirani, A. R. N. Nilchi, "Speech Emotion Recognition based on SVM as Both Feature Selector and Classifier," I.J. Image, Graphics and Signal Processing, vol. 4, pp. 39-45, 2016.

[17] M. Shamsi, "Modeling of Emotion Recognition in Persian Speech by Machine Learning Method, "3rd. International Conference on Science and Engineering, 2 June 2016, Istanbul, Turkey.

[۱۸] م. کرمی، پ. جمشیدلو، ح. صامتی، "تشخیص حس وابسته به گوینده گفتار فارسی با استفاده از ویژگیهای آکوستیکی،" نشریه علمی ترویجی صوت و ارتعاش، سال دوم، شماره چهارم، ۱۳۹۲، صفحات ۳-۱۴، ۱۳۹۲.

[۱۹] ن. کشتیاری، "تاثیر جنسیت بر درک نوای عاطفی گفتار در زبان فارسی، "فصلنامه زبان-شناسی اجتماعی، سال اول، شماره اول، زمستان ۱۳۹۵، صفحات ۸۷-۹۹.

[20] H. Namvar Arefi, S.J. Sameni, H. Jalilvand, M. Kamali, "Effect of hearing aid amplitude compression on emotional speech recognition," Aud Vestib Res., vol. 26(4), pp. 223-230, 2017.

[۲۱] د. غرویان، س.م. احدی، "بازشناسی گفتار احساسی و شناسایی حالت گفتار در زبان فارسی،" مجله فنی و مهندسی مدرس، شماره ۳۴، صفحات ۱۳-۲۷، زمستان ۱۳۸۷.

[22] D. Gharavian, "Statistical Variation Analysis of Formant and Pitch Frequencies in Anger and Happiness Emotional Sentences in Farsi Language," Amirkabir International Journal of Science& Research (Electrical & Electronics Engineering), vol. 44(2), pp. 33- 45, Fall 2012.

[23] P. Jamshidlou, N. Keshtiari, M. Eslami, M. Bahrani, "Acoustic Representation of Intonational Elements in Persian Emotional Speech," Fifth International Conference on Iranian Linguistics (ICIL5), vol. 24, 2013.

[24] M.Bashirpour, M. Geravanchizadeh, "Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments," EURASIP Journal on Audio, Speech, and Music Processing, vol. 9, pp. 1-13, 2018.

[25] M. Bashirpour, M. Geravanchizadeh, "Speech Emotion Recognition Based on Power Normalized Cepstral Coefficients in Noisy Conditions," Iranian Journal of Electrical & Electronic Engineering, vol. 12(3), pp. 197-205, September 2016.

[26] T. Chi, Y. Gao, M. C. Guyton, P. Ru, S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility, " J. Acoust Soc Am, vol. 106, pp: 2719–2732, 1999.

[27] M. Karjalainen, "Auditory models for speech processing", Proc. of Int. Congr. of Phonetic Sciences, 1987.

[28] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations," IEEE Transactions on Audio, Speech and Language Processing, vol. 14(3), pp. 920–930, May 2006.

[29] P.K. Ghosh, L.M. Goldstein, and S.S. Narayanan, "Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures," The Journal of the Acoustical

Society of America, vol. 129(6), pp. 4014–4022, Jun. 2011.
[30] A. Klapuri, "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 255–266, Feb. 2008.
[31] T. Chi, P. Ru, and S.A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," The Journal of the Acoustical Society of America, vol.118(2), pp. 887-906, May 2005
[32] S.M.N. Woolley, T.E. Fremouw, A. hsu and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," Nature Neuroscience, vol. 8, pp. 1371–1379, 2005.
[33] T. Chi, Y. Gao, M.C. Guyton, P. Ru, S. Shamma, "Spectro-Temporal Modulation Transfer Functions and Speech Intelligibility," The Journal of the Acoustical Society of America, vol. 106(5), pp. 2719-32, 1999.
[34] T.M. Elliott, F.E. Theunissen , "The Modulation Transfer Function for Speech Intelligibility," PLoS Comput Biol, vol. 5(3): e1000302, pp. 1-14, 2009.
[35] M. R. Scha¨dler, B.T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," Acoustical Society of America, vol. 131(5), pp. 4134-51
[36] R. Santoro, M. Moerel, F. De Martino, R. Goebel, K. Ugurbil, E. Yacoub, E. Formisano, "Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex," PLoS Comput Biol 10(1): e1003412, pp. 1-14, 2014.
[37] Y. Li, L. Zhang, B. Li, Y. Xu, S. Wu, X. Wei, X. Liu, R. Lin, Q. Wang, "The Simulation Study of Three Typical Time Frequency Analysis Methods", BIO Web of Conferences, vol. 8, p. 02007, 2017.
[38] S. J. Chaudhari R. M. Kagalkar, "Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition", International Journal of Computer Applications, vol. 117(17), May 2015.
[39] B. Schuller, "Towards intuitive speech interaction by the integration of emotional aspects," Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Hammamet, Tunisia, October 6-9, 2002.
[40] N. Keshtiari, M. Kuhlmann, M. Eslami, G. Klann-Delius, "A database of Persian Emotional Speech," Paper presented at the 1st Basic and Clinical Neuroscience Congress, Tehran University of Medical Sciences, 2012.
[41] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klann-Delius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," Behavior Research Methods, vol. 47, pp. 275-294, 2015.
[42] https://isr.umd.edu/Labs/NSL/Software.htm