مقاله

تشخیص احساس از سیگنال گفتار با استفاده از موجک بیونیک

رسول یوسفی‌نژاد*
کارشناس ارشد مهندسی صدا
دانشگاه صدا و سیما
ryousefinezhad@yahoo.com

بابک حاجی باقر نایینی
استادیار دانشکده مهندسی رسانه
دانشگاه صدا و سیما
bnaeeni@yahoo.com

معصومه شفیعیان
استادیار دانشکده مهندسی رسانه
دانشگاه صدا و سیما
shafieian@iust.ac.ir

تاریخ دریافت: ۱۳۹۴/۰۸/۰۳ تاریخ پذیرش: ۱۳۹۵/۰۱/۲۵

چکیده

تشخیص احساس برای رایانه امری چالش‌برانگیز است. دلیل اصلی این موضوع نیز عدم توانایی رایانه در درک احساس کاربر است. هدف از این مقاله، طراحی یک سیستم تشخیص احساس از گفتار و ارائهٔ روشی نوین جهت بهبود این سیستم است. تاکنون در این زمینه از ویژگی‌های متفاوتی استفاده شده است، اما هیچ یک عملاً به ارتباط بین دامنهٔ صوت و حالت‌های احساسی نپرداخته‌اند. چون موجک بیونیک به این ارتباط بیشتر پرداخته است، به‌نظر می‌رسد بتواند در جداسازی حالت‌های مختلف احساسی کمک کند. برای این منظور، در این پژوهش از موجک بیونیک برای استخراج ویژگی از سیگنال‌های صوتی، در تشخیص خودکار احساسات از گفتار استفاده شده است. ساختارموجک بیونیک منطبق بر ساختار گوش انسان است و چون انسان درک خوبی از احساسات گفتار دارد، می توان انتظار داشت که استفاده از موجک بیونیک برای تشخیص خودکار احساسات از گفتار مفید باشد. ساختار پیشنهادی روی پایگاه داده‌های برلین و دادگان گفتار احساسی فارسی ارزیابی شده است، که این دادگان شامل عبارات و جملاتی کوتاه و بیانگر احساسات منفی ترس، خشم، ناراحتی و حالت طبیعی و جز این‌ها هستند. نتایج آزمایش‌ها نشان می‌دهد الگوریتم پیشنهادی عملکرد قابل قبولی در مقایسه با سیستم‌های تشخیص خودکار احساسات از گفتار موجود ارائه می‌دهد.

واژگان کلیدی: تشخیص احساسات از گفتار، استخراج ویژگی، موجک بیونیک، پایگاه داده، ماشین بردار پشتیبان

۱. مقدمه

سیستم‌های بازشناسی احساس در سیگنال‌های گفتار انسان به‌دلیل افزایش تراکنش میان انسان و ماشین نقش مهمی در زندگی روزمره پیدا کرده‌اند و نیاز به این سیستم‌ها افزایش روزافزون داشته است. از طرفی، مشکلاتی نیز در این سیستم‌ها وجود دارد. نخستین مشکل در تحلیل احساسات، دشواربودن 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 به‌شدت وابسته است. دومین مشکل پیچیدگی احساسات

است که اغلب به‌هنگام برقرای ارتباط بین اشخاص، احساسات کامل، خالص و پایه بروز نمی‌کنند، بلکه معمولاً ترکیبی از احساسات در یک لحظه ممکن است بروز نماید [۱]. بنابراین احساس پدیده‌ای مبهم، پیچیده و مرکب است و جداسازی، تشریح و تشخیص آن بسیار دشوار می‌باشد. نحوهٔ 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 زمینهٔ تشخیص احساس از روی گفتار است. اطلاعات جنسیتی بر نتایج دسته‌بندی احساسات مختلف تأثیر می‌گذارد [۲]. در برخی از تحقیقات در این زمینه فقط از گفتار بانوان استفاده شده است [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. نحوهٔ طراحی سیستم‌های تشخیص احساسات در حالت مستقل از گوینده؛ یعنی وقتی جملهٔ واحد توسط گویندگان مختلف ادا می‌شود و حالت مستقل از متن؛ یعنی زمانی‌که جملات متفاوت توسط گوینده واحد ادا می‌شود متفاوت است. در این سیستم‌ها باید اطلاعاتی از سیگنال گفتار استخراج شود که حداکثر همبستگی را با احساسات داشته باشند و در عین‌حال به سایر عوامل از جمله محتوای گفتار و تغییر گوینده وابسته نباشند. تمامی این مسائل روند تشخیص احساس از گفتار را پیچیده‌تر می‌کنند. ضمن اینکه اغلب پژوهش‌ها در این حوزه روی دادگانی انجام شده است که در یک زبان خاص، به یک روش خاص و برای کاربردی خاص تهیه شده‌اند [۴]. نبود دادگان استاندارد و قابل اطمینان برای انجام تحقیقات در این زمینه مخصوصاً برای زبان فارسی کار را مشکل‌تر می‌کند. دادگانی که در کارهای معدود قبلی مورد استفاده قرار گرفته‌اند نیز کاستی‌هایی دارند [۵]. مثلاً در برخی موارد یکسان‌بودن متن دادگان سبب شده است که آثار تغییرات محتوایی حذف شود. مسائل بیان‌شده اهمیت شناسایی احساسات از گفتار را نشان می‌دهد. به‌طور کلی اطلاعاتی از سیگنال گفتار باید استخراج شود که حداکثر همبستگی را با احساسات داشته باشند و در عین‌حال به سایر عوامل از جمله محتوای گفتار وابسته نباشند. بنابراین انتخاب بردار ویژگی بسیار حائز اهمیت است و می‌تواند کارآیی سیستم را به‌شدت تحت تأثیر قرار دهد. این پژوهش سعی در انتخاب ویژگی مناسب برای بهبود تشخیص احساس و طبقه‌بندی درست آنها بر اساس سیگنال گفتار دارد. بر این اساس ویژگی جدید موجک بیونیک براساس تلفیق رفتاری دامنه و فرکانس سیگنال مورد بررسی قرار گرفته است و نشان داده می‌شود که ویژگی‌های استخراج‌شده از این روش خود به‌تنهایی می‌تواند در تشخیص احساسات از گفتار بسیار کارآمد باشد. بر این اساس ابتدا تبدیل موجک بیونیک تشریح و پس از آن روش پیشنهادی معرفی می‌شود.

۲. بررسی چند پایگاه داده احساسی متداول

در زمینهٔ تشخیص احساس از روی گفتار پایگاه‌های دادهٔ مختلفی وجود دارد، که بسیاری از آنها برای استفادهٔ عموم در دسترس نمی‌باشند و همین مسئله، پژوهش در زمینهٔ تشخیص احساس از روی گفتار را با چالش مواجه نموده است. به‌طور کلی، پایگاه‌های داده در دو دستهٔ مصنوعی و طبیعی می‌باشند. در پایگاه دادهٔ مصنوعی از بازیگران خواسته می‌شود که جملات مختلفی را با احساسات گوناگون بیان کنند و این کار سبب می‌شود نتایج با آنچه در واقعیت رخ می‌دهد فاصله داشته باشد. پایگاه دادهٔ طبیعی اما از گفتگوهای روزمره گرفته می‌شود. واضح است که تجزیه و تحلیل در سطح عامیانه نتیجهٔ بهتری می‌دهد، اما متأسفانه دسترسی به گفتگوهای عامیانه مردم بسیار مشکل

و شاید هم غیرممکن می‌باشد. همچنین در اکثر پایگاه‌های داده احساسات خشم، شادی، غم، تعجب، خستگی، انزجار و خنثی وجود دارد. در نهایت بسیاری از پایگاه‌های داده احساسی، احساسات بزرگسالان را در نظر گرفته‌اند، به‌جز دو مورد پایگاه‌های دادهٔ کیسمت[1] و بی‌بی ایبرز[2] که احساس‌های نوزادان را در نظر گرفته‌اند. برخی از محققان بر این باورند که شناخت احساسات نوزادان در تعامل انسان و روبات بسیار مفید است [۶]. تقریباً تمامی پایگاه‌های دادهٔ احساسی از روی گفتار، دارای محدودیت‌هایی برای ارزیابی عملکرد سیستم‌های تشخیص احساس می‌باشند. برخی از این محدودیت‌ها به اجمال ذکر می‌شود:

۱. در اکثر پایگاه‌های داده احساسی به‌صورت طبیعی و واضح نتوانسته‌اند احساس نهفته در گفتار را شبیه‌سازی کنند و بر این اساس به نرخ تشخیص پایینی از احساسات انسان دست یافته‌اند. در برخی از پایگاه‌های داده [۷] نیز نرخ تشخیص انسان حدود ۶۵ درصد است.
۲. در برخی از پایگاه‌های داده مانند کیسمت، کیفیت ضبط جملات چندان خوب نیست، علاوه بر این فرکانس نمونه‌برداری برخی از پایگاه‌های داده خیلی پایین است (۸ کیلوهرتز).
۳. آواشناسی با بعضی از پایگاه‌های داده، مانند بی‌بی ایبرز[۸] ارائه نشده است. بنابراین استخراج محتوای زبانی جملات این‌گونه پایگاه‌ها بسیار مشکل است.

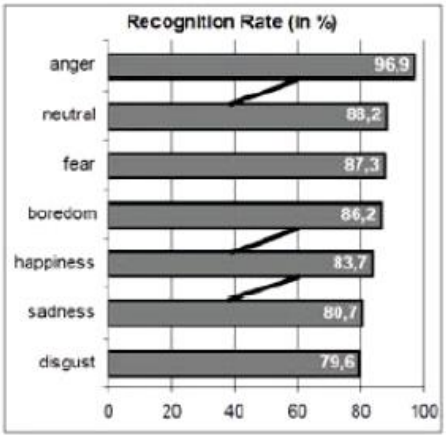
۲-۱. پایگاه داده احساسی برلین

پایگاه دادهٔ برلین [۹] برای آزمایش روی طبقه‌بندی احساسات در فضای گسسته طراحی شده است. این پایگاه داده در دسترس عموم است و از محبوب‌ترین پایگاه‌های دادهٔ احساسی در زمینهٔ تشخیص احساس از روی گفتار است، در نتیجه به‌راحتی می‌توان نتایج حاصل از این پایگاه داده را با پایگاه‌های دیگر مقایسه کرد. این پایگاه مشتمل بر ۵ گویندهٔ مرد و ۵ گویندهٔ زن است که هر ۱۰ گوینده، جملاتی را (۵ جملهٔ کوتاه و ۵ جملهٔ بلند؛ طول جملات نیز بین ۱/۵ تا ۵ ثانیه) به زبان آلمانی بیان نموده‌اند. در مجموع ۲۲/۸۰ دقیقه از گفتار ضبط‌شده در دسترس است. این پایگاه داده شامل ۷ احساس عصبانیت، خستگی، نفرت، ترس، خوشحالی، ناراحتی، طبیعی است. پایگاه دادهٔ اولیه، شامل ۸۰۰ جمله بوده است. این جملات را در اختیار ۲۰ شنونده قرار داده‌اند، تا از نظر نوع احساس ارزیابی شوند و از این طریق به نرخ تشخیص بالای ۸۰ درصد رسیدند. نرخ تشخیص هر احساس در شکل ۱ نمایش داده شده است. سپس برای تشکیل پایگاه دادهٔ نهایی، برچسب‌گذاری احساسات بر این اساس بوده است که هر احساس حداقل آرای ۶۰ درصد شنوندگان را دریافت نماید. در شکل ۲ نیز نسبت جملات پایگاه دادهٔ برلین، به تفکیک هفت کلاس نمایش داده شده است.

۳-۱. پایگاه داده گفتار احساسی زبان فارسی

پایگاه دادهٔ گفتار احساسی زبان فارسی[3]، مجموعه‌ای جامع و معتبر از گفتار احساسی برای زبان فارسی است که در دانشگاه فرای برلین[4] ساخته و سپس رواسازی شده است [۱۰]. برای ساخت این مجموعه دو بازیگر فارسی‌زبان (یک زن و یک مرد) ۹۰ جمله را در پنج آهنگ عاطفی[5] خشم، شادی، غم، ترس، چندش و نیز خنثی، طی شرایط خاصی در سه دستهٔ همگون[6]، ناهمگون[7] و پایه[8]، 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 مطالعهٔ رفتاری جداگانه رواسازی شده بودند. حاصل این کار ۴۷۲ جمله صوتی با آهنگ‌های عاطفی متفاوت است. رواسازی محتوایی این جملات صوتی در یک مطالعهٔ رفتاری توسط ۳۴ فارسی‌زبان مورد ارزیابی قرار گرفته و ۴۶۸ جملهٔ 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 جملهٔ صوتی نمایانگر تفاوت معناداری در زمینهٔ شدت، زیر و بمی صدا و کشش (دیرند) بیان جملات در پنج آهنگ عاطفی مورد مطالعه است. در مجموع حدود ۱۳ دقیقه و ۲۰ ثانیه از گفتار ضبط‌شده در دسترس است.



شکل ۱. نرخ تشخیص هر احساس در پایگاه داده برلین

شکل ۲. مقدار جملات پایگاه داده برلین در هر کلاس

۳. پیشینهٔ تحقیق

در این بخش فعالیت‌های پژوهشی انجام‌شده در زمینهٔ تشخیص احساس از گفتار با استفاده از پایگاه دادهٔ برلین مرور می‌شود. پژوهشگران در مرجع [۱۲]، با استفاده از ویژگی‌های طیفی و MLS و طبقه‌بندی‌های متفاوت و SOM به تشخیص احساسات از گفتار پرداختند و نشان دادند که ترکیب ویژگی‌های پروزودیک و طیفی سبب بهبود نرخ تشخیص می‌شود؛ آنها به نرخ تشخیص ۶۶/۸۳ درصد دست یافتند. یانگ و همکاران (۲۰۱۰) با استفاده از ویژگی‌هایی چون انرژی، فرمنت و نرخ عبور از صفر و طبقه‌بند GMM در حالت شادی نرخ تشخیص ۵۲/۷ درصد را گزارش کردند [۱۳]. نویسندگان تحقیق [۱۴] نشان دادند که با استخراج ویژگی در سطح واج دقت دسته‌بندی افزایش می‌یابد. آنها با استفاده از طبقه‌بندی‌کنندهٔ ماشین بردار پشتیبان[۹] به نرخ تشخیص ۶۷/۷ درصد رسیدند. در مرجع [۱۵] از ویژگی‌های متداول و ماشین بردار پشتیبان برای دسته‌بندی احساسات مختلف از پایگاه داده به زبان‌های دانمارکی، آلمانی و صربستانی استفاده شده است. تأکید مقاله بر توسعهٔ روش‌های گوناگون برای ماشین بردار پشتیبان دودویی است که آن را برای طبقه‌بندی احساسات مناسب می‌سازد. بهترین نرخ تشخیص، مربوط به دسته‌بندی 3DEC گزارش شده است که برابر با ۶۳/۲ درصد بوده است. مرجع [۱۶] به بررسی پارامتر نرخ صحبت، در تشخیص احساسات از گفتار می‌پردازد. نتایج نشان می‌دهد این پارامتر طبقه‌بندی را آسان می‌کند. این گروه از پژوهشگران با استفاده از طبقه‌بندی‌کنندهٔ MLP نرخ تشخیص ۶۱/۴۲ درصد را گزارش کردند. گاروف[۱۰] در [۱۷] با ترکیب دو طبقه‌بندی‌کنندهٔ ماشین بردار پشتیبان و مدل‌های مخفی مارکوف[۱۱]، نرخ تشخیص بالاتری نسبت به ماشین بردار پشتیبان به‌تنهایی به‌دست آورد و نرخ تشخیص ۶۵ درصد را گزارش نمود.

۴. تبدیل موجک بیونیک[۱۲]

تبدیل موجک بیونیک، در سال ۲۰۰۱ م، توسط یاآو و ژانگ به‌عنوان یک روش تحلیل زمان - فرکانسی براساس مدل شنوایی مطرح شد [۱۷–۱۸]. تفاوت این تبدیل با تبدیل موجک متداول این است که رزولوشن در حوزهٔ زمان و فرکانس نه‌تنها با تغییر فرکانس که با تغییرات لحظه‌ای دامنه و مشتق اول آن به‌صورت وقفی تغییر می‌کند. این وقفی‌بودن تابع موجک مادر است که تبدیل موجک را وقفی می‌کند و در تبدیل موجک بیونیک سازوکار فعال کنترلی منطبق بر مدل شنوایی انسان، موجک مادر را براساس سیگنال مورد نظر تطبیق می‌دهد. به‌طور کلی ایدهٔ

تبدیل موجک بیونیک این است که پوش موجک مادر براساس مشخصات سیگنال ورودی با زمان تغییر کند. این کار با معرفی تابع متغیر با زمان T انجام می‌شود که روابط آن به‌صورت زیر بررسی می‌شود [۱۷–۱۸]. رابطهٔ ۱ موجک مادر متداول را در تبدیل موجک پیوسته نشان می دهد:

$$\varphi(t)=\frac{1}{\sqrt{a}}\tilde{\varphi}(t)\exp(j\omega_0 t) \qquad (۱)$$

به‌طوری‌که در این رابطه $\tilde{\varphi}(t)$ پوش $\varphi(t)$ است. برای محاسبهٔ تبدیل موجک سیگنال $x(t)$ از نسخهٔ شیفت‌یافته و منبسط‌شدهٔ موجک مادر، از رابطهٔ ۲ استفاده می‌شود:

$$WT_f(a,\tau)=$$
$$\frac{1}{\sqrt{a}}\int x(t)\tilde{\varphi}^*\left(\frac{t-\tau}{a}\right)\exp\left(-j\omega_0\left(\frac{t-\tau}{a}\right)\right)dt \qquad (۲)$$

به‌طوری‌که در آن a ضریب مقیاس و τ شیفت زمانی است. در تبدیل موجک بیونیک، تابع T برای اعمال تابع کنترلی سلول‌های مویی سیستم شنوایی، موجک مادر را به صورت رابطهٔ ۳ تنظیم می‌کند.

$$\varphi(t)=\frac{1}{T\sqrt{a}}\tilde{\varphi}\left(\frac{t}{T}\right)\exp(j\omega_0 t) \qquad (۳)$$

وجود T در قسمت اول رابطه بیانگر فاکتور مقیاس است (برای اطمینان از یکسان‌بودن انرژی در هر موجک مادر) و در قسمت دوم رابطه، پوش موجک مادرِ $\varphi(t)$ را بدون تنظیم فرکانس مرکزی آن تنظیم می‌کند. با تعریف موجک مادر جدید، تبدیل موجک منطبق بر سیستم شنوایی به صورت رابطهٔ ۴ تعریف می‌شود.

$$BWT_f(a,\tau)=$$
$$\frac{1}{T\sqrt{a}}\int x(t)\tilde{\varphi}^*\left(\frac{t-\tau}{Ta}\right)\exp\left(-j\omega_0\left(\frac{t-\tau}{a}\right)\right)dt \qquad (۴)$$

که تابع T مرحله به مرحله و براساس مدل فعال غشای پایه[۱۳] تعیین می‌شود و در کل مقدار آن به دامنهٔ لحظه‌ای سیگنال و مشتق اول آن وابسته است.

مقدار T در بازه‌های زمانی کوتاه ثابت فرض می‌شود؛ یعنی $BWT_f(a,\tau)=K\times WT_f(a,\tau)$. البته باید توجه داشت که K به تابع T وابسته است. فرض کنید موجک مادر در موجک تبدیل بیونیک از نوع مورلت باشد، چنانچه در رابطهٔ ۵ نشان داده شده است.

$$\varphi(a,\tau)=\exp\left(-\frac{t}{T_0}\right)^2\exp(j\omega_0 t) \qquad (۵)$$

انتخاب موجک مادرِ مورلت برای دستیابی به کمترین حجم محاسبات در حوزهٔ زمان و فرکانس است. حال فاکتور K به صورت تقریبی از رابطهٔ ۶ به‌دست می‌آید:

$$K=e\frac{1.7725T_0}{\sqrt{T^2}+1} \qquad (۶)$$

مزایای تبدیل موجک بیونیک عبارت است از:

۱. تبدیل موجک بیونیک یک تبدیل غیر خطی است که دارای حساسیت بالا و گزینش فرکانسی می‌باشد.

۲. تبدیل موجک بیونیک سیگنال را با یک توزیع انرژی متمرکز نمایش می‌دهد.

۳. تبدیل موجک بیونیک معکوس می‌تواند سیگنال اصلی را از نمایش فرکانس - زمان آن استخراج کند.

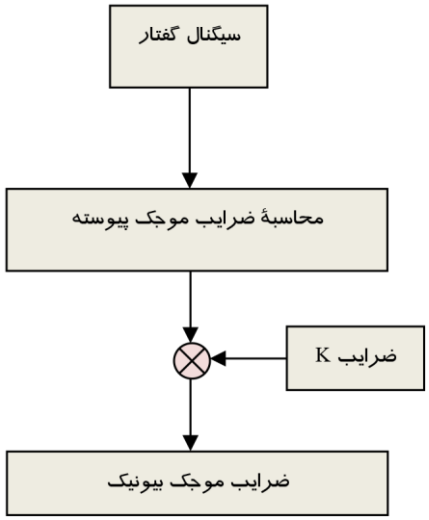
۵. روش کار

در پیاده‌سازی نرم‌افزاری با استفاده از نرم‌افزار متلب و استفاده از جعبه‌ابزار موجک‌ها در متلب، ۲۲ مقیاس منطبق با ۲۲ الکترود موجود در سیستم حلزون کاشتنی در نظر گرفته شده است. ایمپلنت حلزون گوش وسیله‌ای الکترونیک است که توسط جراحی کاشته شده و امکان حس اصوات را برای افراد با سطح شنوایی بسیار کم فراهم می‌کند. این ایمپلنت معمولاً گوش بیونیک نامیده می‌شود. ایمپلنت حلزون بر خلاف سمعک صدا را تقویت نمی‌کند، بلکه به‌طور مستقیم اعصاب شنوایی داخل حلزون را توسط پالس‌های الکتریکی تحریک می‌کند. سپس فرکانس‌های مرکزی از ۲۲۱ تا ۵۲۹۱ هرتز طبق مدل شنوایی تعریف شده‌اند. روند کار در شکل ۳ نشان داده شده است.

۵-۱. ساختار روش پیشنهادی

بلوک دیاگرام نمایش داده‌شده در شکل ۴ ساختار الگوریتم پیشنهادی با استفاده از آنتروپی بیونیک را نشان می‌دهد. در

این روش ابتدا برای استخراج ویژگی‌های ایستا[14]، باید سیگنال گفتارکه یک سیگنال غیرایستا[15] است را به بازه هایی در حدود ۲۰ تا ۱۰۰ میلی‌ثانیه تقسیم کرد و از چون در دادگان فارسی، گوینده مورد نظر مرد است، حداقل فرکانس گام ۵۰ هرتز درنظر گرفته شده و حداقل طول پنجره ۴۰ میلی‌ثانیه باشد و میزان همپوشانی[16] فریم‌ها ۵۰ درصد یا ۲۰ میلی‌ثانیه در نظر گرفته شود. این همپوشانی برای تغییرات هموار ویژگی‌ها از یک فریم به فریم بعدی لازم است.



شکل ۳. روند محاسبهٔ ضرایب بیونیک

بلوک دیاگرام نمایش داده‌شده در شکل ۴ ساختار الگوریتم پیشنهادی با استفاده از آنتروپی بیونیک را نشان می‌دهد. در این روش ابتدا برای استخراج ویژگی‌های ایستا، باید سیگنال گفتارکه یک سیگنال غیرایستا است را به بازه‌های در حدود ۲۰ تا ۱۰۰ میلی‌ثانیه تقسیم کرد، و از آنجایی که در دادگان فارسی، گوینده موردنظر مرد است حداقل فرکانس گام ۵۰ هرتز درنظر گرفته شده و بنابراین حداقل طول پنجره ۴۰ میلی‌ثانیه است و میزان همپوشانی فریم‌ها ٪۵۰ یا ۲۰ میلی‌ثانیه در نظر گرفته شده است. این همپوشانی برای تغییرات هموار ویژگی‌ها از یک فریم به فریم بعدی لازم است. در اکثر الگوریتم‌های پردازش گفتار، برای کمتر کردن تأثیر لبه‌ها در طیف، از پنجره‌های همینگ استفاده می‌شود که باعث تمرکز اطلاعات در وسط پنجره و کاهش آنها در لبه‌ها می‌باشد. بنابراین ویژگی‌ها از یک فریم به فریم بعدی، به صورت جهشی تغییر نمی‌کند و تغییرات نسبتاً آرامی دارد. برای کمینه کردن اعوجاج طیفی در هنگام فریم بندی سیگنال گفتار، هر فریم در یک پنجره همینگ ضرب می‌شود چنانکه در رابطه (۷)، نشان داده شده است.

$$w(n)=0.54-0.64\cos\left(\frac{2n\pi}{N-1}\right) \qquad (۷)$$

$$0\le n\le N-1$$

که در آن N طول فریم‌ها (برحسب تعداد نمونه) است. عمل پنجره‌بندی مانند یک فیلتر پایین‌گذر است که در آن سیگنال‌های موجود در مرکز پنجره بهبود یافته و سیگنال‌های موجود در لبه‌های پنجره هموار می‌گردند. پنجره‌بندی همینگ سبب هموارترشدن طیف در انتقال بین پنجره‌ها می‌شود. سپس با استفاده از مجموعهٔ فیلتر بانک بیونیک که از ۲۲ فیلتر با فرکانس‌های، مرکزی که در جدول ۱ آورده شده است، هر فریم گفتار به ۲۲ باند تجزیه و آنتروپی شانون[17] (آنتروپی اطلاعات) برای هر باند با استفاده از رابطهٔ ۸ محاسبه می‌شود.

$$e_m(i)=-\sum_j s_m{}^2(j)\log_{10}^{s_m{}^2(j)} \qquad (۸)$$

که در این رابطه $e_m(i)$ آنتروپی باند i ام در گره m می‌باشد (m=1,2, …, 31) و $s_m(j)$ نمونه‌های سیگنال تجزیه‌شده در این گره هستند و در نهایت با بررسی طبقه‌بندی کننده‌های مختلف از ماشین بردار پشتیبان به‌عنوان بهترین طبقه‌بندی‌کننده در این زمینه استفاده شده است.

۵-۲. جدول مربوط به تعداد فریم‌های احساسی مختلف در حالت آموزش و آزمایش

دو حالت مستقل ازمتن وم ستقل از گوینده را می‌توان بررسی نمود. در حالت مستقل از متن یک گویندهٔ واحد جملات گوناگونی را ادا می‌کند و در حالت مستقل از گوینده، جملهٔ واحدی توسط گوینده‌های مختلف ادا

می‌شود. در این پژوهش به‌دلیل برخی محدودیت‌ها در پایگاه‌های داده و انطباق بیشتر دو پایگاه داده، حالت مسقل از متن مورد تحلیل و بررسی قرار گرفته است. در هر دو پایگاه داده از صدای گوینده مرد استفاده شده است. در حالت گویندهٔ واحد (مستقل از متن)، پایگاه دادهٔ فارسی، شامل ۴۵ بیان احساسی مختلف، که شامل ۱۲ حالت احساسی خشم، ۴ حالت ترس، ۵ حالت غم، ۶ حالت چندش و ۱۸ حالت طبیعی می‌باشد و دادگان آلمانی نیز شامل ۴۵ بیان احساسی مختلف، که شامل ۱۲ حالت احساسی خشم، ۴ حالت ترس، ۷ حالت خوشحالی، ۷ حالت غم، ۵ حالت کسالت و ۱۰ حالت طبیعی می‌باشد. نسبت داده‌های آموزشی به داده‌های تست۶۵ به ۳۵ درصد است. سعی شده است برای حفظ تنوع جملاتی که با یک حالت احساسی مشخص، چند بار بیان شده‌اند همگی در داده‌های آموزشی قرار بگیرند. به این ترتیب داده‌های تست از بین جملاتی انتخاب شده‌اند که توسط گویندهٔ مشخص تنها یک‌بار در حالت احساسی مشخص بیان شده‌اند. جدول ۲ تعداد فریم‌های مربوط به هر حالت احساسی در حالت آموزش و آزمایش را نشان می‌دهد.

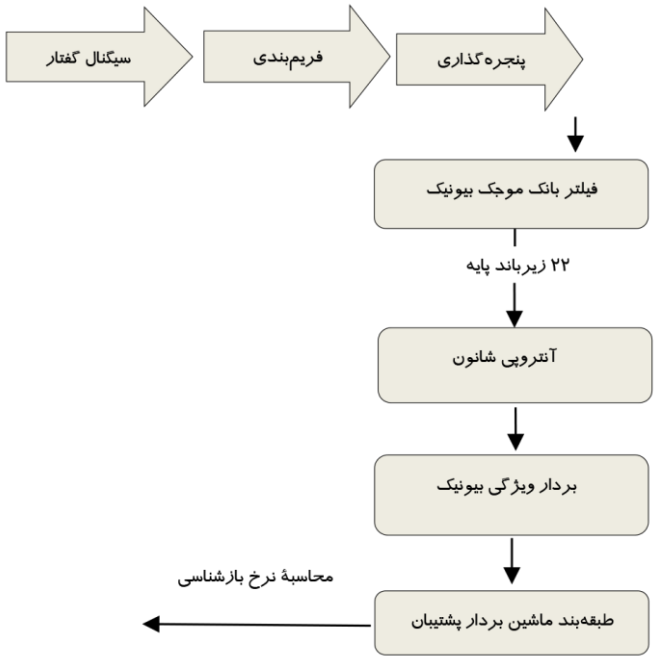
جدول ۱. فرکانس‌های مرکزی فیلتر بانک بیونیک

فیلتر	فرکانس مرکزی	فیلتر	فرکانس مرکزی
۷	۵۲۹۱	۱۲	۱۰۱۱
۲	۴۵۵۱	۱۳	۸۷۰
۳	۳۹۱۶	۱۴	۷۴۸
۴	۳۳۶۹	۱۵	۶۴۴
۵	۲۸۹۹	۱۶	۵۵۴
۶	۲۴۹۴	۱۷	۴۷۶
۷	۲۱۴۵	۱۸	۴۱۰
۸	۱۸۴۶	۱۹	۳۵۲
۹	۱۵۸۸	۲۰	۳۰۳
۱۰	۱۳۶۶	۲۱	۲۶۱
۱۱	۱۱۷۵	۲۲	۲۲۱

۵-۳. آنالیز نتایج جهت انتخاب بهترین کرنل ماشین بردار پشتیبان

روش ماشین بردار پشتیبان الگوریتمی ساده و کارآ با حجم محاسبات کم در یادگیری ماشین است و کاربرد گسترده‌ای در طبقه‌بندی الگوها دارد. در شرایطی که حجم داده‌ها محدود باشد، این روش نرخ بازشناسی بهتری نسبت به سایر روش‌های طبقه‌بندی دارد. در این مقاله برای تشخیص حالات احساسی در هر فریم گفتار از این روش استفاده شده است. ماشین بردار پشتیبان از الگوریتم‌های دسته‌بندی‌کننده با سرپرست است. توابع کرنل پرکاربرد در ماشین بردار پشتیبان عبارت‌اند از تابع چندجمله‌ای و تابع پایه شعاعی[۱۸]. پارامترهای توابع کرنل معمولاً به‌صورت تجربی یا توسط الگوریتم‌های جستجو تعیین می‌شوند. نتایج مربوط به کرنل چندجمله‌ای در مقایسه با کرنل تابع

پایه شعاعی نتایج مطلوبی ندارد. کرنل چندجمله‌ای در هیچ‌یک از حالات انتخاب مناسبی نیست. آزمایش‌ها نشان می‌دهد که کرنل تابع پایه شعاعی نتایج بهتری در زمینهٔ بازشناسی احساسات از گفتار دارد. در این مقاله هدف بررسی کرنل‌ها نبوده و بهبود بردار ویژگی و معرفی ویژگی جدید مبتنی بر بیونیک برای بهبود نرخ بازشناسی احساسات از گفتار مد نظر بوده است، لذا از ماشین بردار پشتیبان با کرنل بهینه برای بردار ویژگی جدید مبتنی بر موجک بیونیک استفاده شده است. با بررسی اولیه روی کرنل و تنظیمات ماشین بردار پشتیبان یک نتیجهٔ نسبی به‌دست آورده و از آن برای محاسبات بعدی استفاده شده است. برای دادگان فارسی کرنلِ تابع پایه شعاعی با c=15 و sigma=1 در نظر گرفته شده است و برای دادگان آلمانی کرنل تابع پایه شعاعی با c=1 و sigma=1 لحاظ شده است.



شکل ۴. بلوک‌دیاگرام ساختار پیشنهادی سیستم تشخیص احساسات از گفتار

۶. آنالیز نتایج نهایی

آنالیز نتایج حاصل از این پژوهش در جدول ۳ آمده است. همهٔ آزمایش‌های فوق پنج بار تکرار و میانگین نتایج حاصل در جدول ارائه شده است. در حالت طبقه‌بندی دو دسته‌ای برای دادگان آلمانی، حالت‌های خشم، کسالت، ترس، شادی و غم، با حالت طبیعی دسته‌بندی شده است و نتایج به‌صورت خشم، کسالت، ترس، شادی و غم در جدول بیان شده است. در حالت دو دسته‌ای بین حالت چندش و طبیعی، استفاده ازروش‌های استخراج ویژگی مبتنی بر موجک بیونیک با نرخ تشخیص ۸۴/۲ درصد بهترین نرخ تشخیص را نشان می‌دهد، که می‌تواند دلیل ارتباط بین این ویژگی‌ها با دامنهٔ صوت باشد. همچنین در حالت چند دسته ای، همه حالت‌های خشم و کسالت، ترس و شادی و غم طبقه‌بندی شده است (چند دسته‌ای)؛ که در نتایج طبقه‌بندی در جدول بیان شده است. در حالت چند دسته‌ای (جملهٔ واحد)، با استفاده از روش استخراج ویژگی آنتروپی موجک بیونیک نرخ بازشناسی ۷۸ درصد برای دادگان فارسی به‌دست می‌آید که در مقایسه با روش‌های متداول بهبود قابل توجهی دیده می‌شود. دلیل آن نیز استفاده از آنتروپی بیونیک است. می‌توان از انرژی بیونیک هم استفاده

کرد؛ اما استفاده از آنتروپی بیونیک نتایج بهتری دارد، چون انرژی فقط شدت سیگنال هر باند را نشان می‌دهد، اما آنتروپی محتوای اطلاعاتی هر باند را نشان می‌دهد. هرچند نتایج در حالت دو دسته‌ای بهتر عمل کرده‌اند، اما هدف یک سیستم طبقه‌بندی‌کنندهٔ احساسات شناسایی انواع احساسات در حالت چند دسته‌ای است. استفاده از روش‌های استخراج ویژگی مبتنی بر بیونیک در حالت کلی سبب بهبود نرخ بازشناسی در حالت چند دسته‌ای می‌شود.

جدول ۲.تعداد فریم‌های آموزش و آزمایش در دو دسته دادگان مورد استفاده به تفکیک حالت احساسی

دادگان	حالت احساسی	تعداد فریم‌ها
آلمانی مستقل از متن	خشم	Train=3627, Test=1718
	کسالت	Train=2363, Test=1287
	ترس	Train=2055, Test=1270
	شادی	Train=2768, Test=1067
	غم	Train=3252, Test=1515
	طبیعی	Train=4900, Test=3062
فارسی مستقل از متن	خشم	Train=2927, Test=1541
	چندش	Train=2236, Test=1281
	ترس	Train=2028, Test=1330
	غم	Train=2213, Test=1372
	طبیعی	Train=3404, Test=1925

جدول ۳. تست مستقل از متن؛ نرخ بازشناسی با استفاده از ویژگی‌های مبتنی بر تبدیل موجک بیونیک (گویندهٔ واحد)

دادگان پایگاه داده برلین						
حالت احساسی	خشم	کسالت	ترس	شادی	غم	چ د[19]
نتایج آنتروپی بیونیک	۷۴/۲	۶۸/۵	۶۶/۲	۷۱/۴	۶۸/۵	۶۶
دادگان پایگاه داده احساسی زبان فار						
حالت احساسی	خشم	چندش	ترس	غم	چند دسته‌ای	
نتایج آنتروپی بیونیک	۷۰	۸۴/۲	۷۸/۵	۷۸/۵	۷۸	

۷. مقایسهٔ ساختار پیشنهادی با کارهای دیگران

در این بخش نتایج حاصل از الگوریتم پیشنهادی و مقایسهٔ آن با دیگر روش‌های بازشناسی ارائه‌شده در کارهای دیگران ارزیابی می‌شود. برای این منظور مقالاتی که در آن از پایگاه داده برلین استفاده شده، بررسی شده است. فاکوتاکیس و تالامپیراس (۲۰۱۲) مدل‌کردن تغییرات زمانی

پارامترهای صوتی را برای بازشناسی احساسات از گفتار پیشنهاد کردند [۱۹]. در این مقاله بازشناسی احساسات در حالت مستقل از گوینده را روی ۶ حالت احساسی مورد بررسی قرار دادند. داده‌های مورد استفاده در پژوهش مذکور دادگان گفتار احساسی برلین است و از ویژگی‌هایی با پیچیدگی محاسباتی بالا استفاده شده است. ویژگی‌های مختلفی از جمله ویژگی‌های آماری زمان کوتاه، ممنتوم‌های طیفی و مدل‌های بازگشتی به‌همراه ویژگی‌های جدیدی مبتنی بر بستهٔ موجک ادراکی[۲۰] در این پژوهش مورد بررسی قرار گرفته‌اند. طبقه‌بندی با استفاده از مدل‌های مخفی مارکوف انجام شده است که از نظر زمان مورد نیاز و پیچیدگی محاسباتی بر ماشین بردار پشتیبان برتری ندارد. علاوه بر آن، در مقالهٔ مذکور نتایج و بررسی‌های مختلفی گزارش شده است که از آن جمله تأثیر تغییر طول فریم گفتار است. توضیح جزئیات مقاله و بررسی تمام نتایج آن از حوصلهٔ این بحث خارج است. فقط به‌عنوان بخشی از این مقالهٔ جامع که از لحاظ طول فریم (طول فریم ۴۰ میلی ثانیه) و نوع ویژگی با روش پیشنهادی ما قابل مقایسه است اینکه در مقالهٔ فوق بازشناسی گفتار با استفاده از آنتروپی بستهٔ موجک ادراکی حدود ۶۱ درصد به‌دست آمده است که بعد با ترکیب آن با ضرایب کپسترال مل[۲۱] نهایت نرخ ۶۱/۶۶ درصد را گزارش نمودند که در مقایسه با روش آنتروپی موجک بیونیک و طبقه‌بندی با ماشین بردار پشتیبان برتری ندارد. گاروف (۲۰۰۸) با ترکیب ویژگی‌هایی نظیر ضرایب کپسترال مل، انرژی، جیتر، شیمر، نرخ عبور از صفر و طبقه‌بند SVM+GMM در حالت چند دسته‌ای به نرخ تشخیص ۶۵ درصد رسید که در مقایسه با نتایج به دست آمده با استفاده از آنتروپی موجک بیونیک، کمتر است [۱۶]. یانگ و همکاران (۲۰۱۰) با استفاده از ویژگی‌هایی چون انرژی، فرمنت و نرخ عبور از صفر و طبقه‌بند GMM در حالت شادی نرخ تشخیص ۵۲/۷ درصد را گزارش کردند که از مقایسهٔ این نتیجه با روش پیشنهادی مشاهده می‌شود که با استفاده از ویژگی آنتروپی بیونیک نتیجهٔ ۷۱/۴ درصد به‌دست آمده است [۱۲]. بیتوک و همکاران (۲۰۱۰) با استفاده از ویژگی‌های پروزودی و طبقه‌بند ماشین بردار پشتیبان، به نرخ تشخیص ۶۷/۷ درصد رسیدند که در مقایسه با روش پیشنهادی، که در آن فقط از آنتروپی موجک بیونیک استفاده شده، بسیار نزدیک است و نسبت به نتایج پایگاه داده فارسی کمتر می‌باشد [۱۳]. البرز و همکاران (۲۰۱۱) با استفاده از ویژگی‌های طیفی و طبقه بندی‌کننده‌های متفاوت در حالت چند دسته‌ای، نرخ تشخیص ۶۶/۸۳ درصد را گزارش کردند [۱۱] که در مقایسه با روش پیشنهادی در حالت چند دسته‌ای با ویژگی مبتنی بر موجک بیونیک، نرخ بازشناسی تقریباً برابری را نشان می‌دهد. هابنر و همکاران (۲۰۱۲) با استفاده از ویژگی نرخ عبور از صفر و طبقه‌بندی‌کنندهٔ شبکه عصبی MLP نرخ تشخیص ۶۱/۴۲ در حالت شادی گزارش کرده‌اند [۱۶]. مقایسهٔ این نتیجه با روش پیشنهادی بهبود قابل توجهی در بازشناسی را با استفاده از موجک بیونیک را نشان می‌دهد. حسان و همکاران (۲۰۱۲) با استفاده از ترکیب ویژگی‌های طیفی در حالت چند دسته‌ای با ماشین بردار پشتیبان به نرخ تشخیص ۶۳/۲ درصد رسیدند که در مقایسه با روش پیشنهادی نرخ بازشناسی پایین‌تری دارد [۱۴].

۸. جمع‌بندی

با افزایش روزافزون تراکنش میان انسان و ماشین در بسیاری زمینه‌ها، تحقیقات زیادی برای ایجاد ارتباط بهتر و آسان‌تر بین این دو در حال انجام است. ازجمله 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 روبه رشد در حوزهٔ پردازش گفتار را به‌خود اختصاص داده‌اند. در این مقاله از ویژگی‌های مبتنی بر موجک بیونیک برای تشخیص احساسات از گفتار استفاده گردید. علت استفاده از این ویژگی‌ها در تشخیص احساسات

از گفتار این است که موجک بیونیک منطبق بر ساختار گوش انسان است و چون انسان درک خوبی از احساسات گفتار دارد، نتایج آزمایش‌ها نشان می‌دهد که این موجک می‌تواند در تشخیص خودکار احساسات از گفتار مؤثر واقع شود. سیگنال موجود در هر فریم ابتدا توسط بیونیک به ۲۲ باند تجزیه و سپس آنتروپی شانون در هر باند محاسبه می‌شود و به‌عنوان بردار ویژگی مورد استفاده قرار می‌گیرد. استفاده از آنتروپی نتایج خوبی دربر دارد و دلیلش این است که آنتروپی محتوای اطلاعاتی هر باند را نشان می‌دهد. ساختار پیشنهادی روی پایگاه داده‌های برلین و دادگان گفتار احساسی فارسی ارزیابی شده است. نتایج آزمایش‌ها نشان می‌دهد که الگوریتم پیشنهادی عملکرد قابل قبولی در مقایسه با سیستم‌های تشخیص احساسات از گفتار موجود ارائه می‌دهد. هرچند ویژگی‌های متداول در حالت دو دسته‌ای خوب عمل می‌کنند، اما هدف یک سیستم طبقه‌بندی‌کنندهٔ احساسات، شناسایی انواع احساسات در حالت چند دسته‌ای است که کارایی سیستم پیشنهادی در حالت چند دسته‌ای نیز بسیار خوب بوده است. به‌طور کلی مقایسهٔ نتایج دو نکته را مشخص می‌کند: اولاً نمی‌توان ارتباط مستقیمی بین نتایج یافت، ثانیاً استفاده از روش پیشنهادی در حالت چند دسته‌ای منجر به بهبود قطعی در نرخ تشخیص می‌شود. معمولاً در مقالات هم، حالت چند دسته‌ای مد نظر است. کاربرد عملی این پژوهش در تشخیص احساسات از گفتار، در برنامه‌های کاربردی کامپیوتر، ابزار تشخیص برای درمانگران، مراکز پاسخگویی خودکار، ارتباطات تلفن همراه و جز این‌هاست.

۹.فعالیت‌های آتی

برای بهترشدن کارایی سیستم پیشنهاد می‌شود ویژگی‌های دیگری ازسیگنال گفتار (همچون ضرایب کپسترال مل[۲۲]، فرمنت‌ها[۲۳]،جیتر[۲۴] و جز این‌ها) به سیستم اضافه گردد و با ترکیب آن با ویژگی‌های مبتنی بر موجک به نرخ بازشناسی بهتری رسید. همچنین می‌توان پژوهش انجام‌شده را برای گفتار احساسی گویندهٔ زن هم بررسی و مقایسه کرد. استفاده از شبکه‌های عصبی و سیستم‌های فازی به‌عنوان طبقه‌بند با موجک بیونیک نیز توصیه می‌شود.

۱۰. مآخذ

[1] France, D. J., et al. “Acoustical properties of speech as indicators of depression and suicidal risk.” *Proc. IEEE, Trans.Biomedical Eng*, vol. 47(7), 2007, pp. 829-837.

[2] Pao, T., 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, Algoritm and Programming, 2012, pp. 157-162.

[3] Chandaka, S., A. Chatterjee, S. Munshi. “Support vector machines employing cross-correlation for emotional speech recognition.” *Measurment*, vol. 42, 2009, pp. 611-618.

[4] Hubner, D., B. Velasenko. “The performance of the speaking rate parameter in emotion recognition from speech”, *Proc. IEEE Int. Conf. Multimedia and Expo Workshops*, 2012, pp. 296-301.

[5] Esmaileyan, Z., H. Marvi. “A Database for Automatic Persian Speech Emotion Recognition: Collection, Processing and Evaluation.” *IJE Trans. A: Basics*, Vol. 27, No. 1, 2014, pp. 79 -90.

[6] Breazeal, C., L. Aryananda. “Recognition of affective communicative intent in robot directed speech.” *Autonomous Robots*, vol. 2, 2002, pp. 83-104.

[7] New, T., S. Foo, D. Silva. “Speech Emotion Recognition using hidden Markov model.” *speech communication*, vol. 41, 2003, pp.603-623.

[8] Slaney, M., G. McRoberts. “Babyears: a recognition system for affective vocalizations.” *speech communication*, vol. 39, 2003, pp. 367–384.

[9] Burkhardt, F., et al, “A Database of German Emotional Speech.” *in Proceedings of the interspeech*, Lissabon, Portugal, 2005, pp.1517-1520.

[10] Keshtiari, N., M. Kuhlmann, M. Eslami, G. Klann-Delius. “Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD).” 2014.

[11] Albornoz, E., et al. “Spoken Emotion recognition using hierarchical classifier.” *Computer speech and Language*, 2011, pp. 556–570.

[12] Yang, B., M. Lugger. “Emotion recognition from speech signals using new harmony features.” *signal processing* 90, 2010, pp.1415-1423.

[13] Bitouk, D., et al. “class level spectral features for emotion recognition.” *speech communication*, vol. 52, 2010, pp.613-625.

[14] Hassan, A., R. Damper. “Classification of emotional speech using 3DEC hierarchical classifier.” *speech communication* 54, 2012, pp. 903-916.

[15] Hubner, D., et al. “The Performance of The Speaking Rate Parameter in emotion recognition from speech.” *IEEE, International conference on Multimedia and Expo Workshops*, 2012.

[16] Gaurav, M. “Performance Analysis of spectral and prosodic features and their fusion for emotion recognition in speech.” *IEEE. SLT*, 2008.

[17] Yao, J., Y. Zhang. “Bionic Wavelet Transform: A New Time-Frequency Method Based on an Auditory Model.” *Proc. IEEE, Trans.Biomedical Eng*, vol. 48(8), 2001, pp. 856-863.

[18] Chen, F., Zhang. “A new implementation of discrete bionic wavelet transform: Adaptive tiling.” *Digit. Signal Process.* vol. 16, issue 3, 2006, pp. 233-246.

[19] Ntalampiars, S., N. Fakotakis. “Modeling the temporal evolution of acoustic parameters for speech emotion recognition.” *IEEE Trans. on Affective Computing*, Vol. 3, No. 1, 2012, pp. 116-125.

پی‌نوشت

1. KISMET
2. BabyEars
3. Persian ESD
4. Freie Universität Berlin (Free University of Berlin, often abbreviated as FU Berlin or simply FU), http://www.fu-berlin.de (accessed August 30, 2016)

۵. پنج حس خشم، شادی، غم، ترس و چندش به "احساسات پایه" مشهورند. طبق نظر برخی از روانشناسان احساسات پایه بدون در نظر گرفتن تفاوت‌های فرهنگی و زبانی میان تمامی افراد بشر مشترک‌اند.

6. Congruent: emotional lexical content articulated in congruent emotional prosody

7. Incongruent: neutral lexical content articulated in various emotional prosody

8. Baseline: both neutral and emotional lexical content articulated in neutral prosody

9. Support Vector Mashine

10. Gaurav

11. HMM

12. Bionic Wavelet Transform

13. Basilar membrane Displacement

14. stationary

15. non-stationary

16. overlap

۱۷. در تئوری اطلاعات، آنتروپی (یا به‌صورت دقیق‌تر آنتروپی شانون) امید ریاضی اطلاعات موجود در هر پیام دریافت شده است. در اینجا پیام لزوماً متن نیست و هر جریانی از اطلاعات می‌تواند باشد. آنتروپی هر پیام مقدار عدم قطعیت آن است؛ هرچه پیام تصادفی‌تر باشد مقدار آن افزایش می‌یابد و هرچه کمتر تصادفی باشد مقدار آن کاهش می‌یابد. مقصود در اینجا این است که هرچه یک رویداد کمتر تصادفی باشد، در زمان رخداد اطلاعات بیشتری را فراهم می‌کند. در ابتدا این قضیه وارونه به‌نظر می‌رسد: این‌گونه تصور می‌شود که پیام‌هایی که ساختاریافته‌ترند، اطلاعات بیشتری دارند. اما این‌طور نیست. برای مثال، پیام " آآآآآآ" (که به‌نظر بسیار ساختار یافته و غیرتصادفی به‌نظر می‌رسد (هرچند در واقع می‌تواند از یک فرایند تصادفی ایجاد شده باشد) حاوی اطلاعات خیلی کمتری نسبت به پیام "الفبا" (که تا حدودی ساختار یافته است، اما تصادفی‌تر) یا پیام "آبستیخدم۲" (که بسیار تصادفی است) می‌باشد. در نظریهٔ اطلاعات، "اطلاعات" لزوماً به‌معنای اطلاعات مفید نیست؛ بلکه تنها مقدار تصادفی‌بودن پیام را مشخص می‌کند. بنابراین در مثال بالا، پیام اول دارای کمترین اطلاعات و پیام آخر دارای بیشترین اطلاعات است، هرچند در اصطلاح روزمره پیام وسطی؛ یعنی "الفبا" دارای اطلاعات بیشتری نسبت به رشته‌هایی از حروف تصادفی به نظر می‌رسد. بنابراین، در تئوری اطلاعات، پیام اول دارای آنتروپی کمی است، پیام دوم آنتروپی بالاتری از اولی دارد و پیام سوم بیشترین آنتروپی را داراست. به صورت فنی‌تر، دلائلی برای تعریف اطلاعات به‌صورت منفی لگاریتم توزیع احتمال وجود دارد. جعبه‌ابزار موجک‌ها در متلب، آنتروپی شانون به‌صورت تعریف‌شده وجود دارد. به‌صورت کلی، آنتروپی به عدم قطعیت یا بی‌نظمی

اشاره می‌کند. آنتروپی شانون در سال ۱۹۴۸ م در مقاله‌ای تحت عنوان یک *نظریه ریاضی برای ارتباطات* توسط کلاود شانون مطرح شد.

18. radial basis function

۱۹. چنددسته‌ای؛ که در این حالت بین تمام احساسات مورد بررسی سیستم، احساس موجود در گفتار را تشخیص می‌دهد.

20. oerceptual wavelet transform (PWP)

21 MFCC

22. Msel Frequency Cepstal Cofficients

23. Formants

24. Jitter