تشخیص حالت احساسی از سیگنال گفتار در حالت مستقل از گوینده با استفاده از آنتروپی بسته موجک

مینا کدخدایی الیادرانی[۱] - سید حمید محمودیان[۲] - غزال شیخی[۳]

(۱) کارشناس ارشد – گروه برق، موسسه آموزش عالی بنیان، شاهین شهر، اصفهان، ایران

(۲) استادیار – گروه برق، دانشکده مهندسی برق، واحد نجف‌آباد، دانشگاه آزاد اسلامی، نجف‌آباد، اصفهان، ایران

(۳) دانشجوی دکتری- دانشکده مهندسی کامپیوتر، دانشگاه مدیترانه شرقی، ترکیه

تاریخ دریافت: ۱۳۹۳/۳/۵ تاریخ پذیرش: ۱۳۹۳/۹/۱۲

خلاصه: در این مقاله آنتروپی بسته موجک برای بازشناسی احساسات از گفتار در حالت مستقل از گوینده پیشنهاد شده است. پس از پیش‌پردازش، بسته موجکِ db3 سطح ۴ در هر فریم محاسبه شده است و آنتروپی شانون در گره‌های آن به عنوان ویژگی در نظر گرفته شده است. ضمناً ویژگی‌های نواییِ گفتار شامل فرکانس چهار فرمنت اول، جیتر یا دامنه تغییرات فرکانس گام و شیمر یا دامنه تغییرات انرژی به عنوان ویژگی‌های پرکاربرد در حوزه تشخیص احساسات در کنار ضرایب فرکانسی کپسترال مل (MFCC) برای تکمیل بردار ویژگی مورد استفاده قرار گرفته‌اند. طبقه‌بندی با استفاده از ماشین بردار پشتیبان (SVM) انجام شده است و ترکیب‌های مختلفی از بردار ویژگی در حالت چند دسته‌ای برای همه احساسات و دودسته‌ای نسبت به حالت طبیعی مورد بررسی قرار گرفته‌اند. ۴۶ بیانِ مختلف از جمله واحد در دادگان احساسی دانشگاه برلین به زبان آلمانی انتخاب شده که توسط ۱۰ گوینده مختلف با حالت‌های احساسی ناراحتی، خوشحالی، ترس، ملالت، خشم و حالت طبیعی بیان شده‌اند. نتایج نشان می‌دهند استفاده از ضرایب آنتروپی به عنوان بردار ویژگی نرخ بازشناسی را در حالت چند دسته‌ای بهبود می‌بخشد. علاوه بر آن ویژگی‌های پیشنهادی در ترکیب با سایر ویژگی‌ها باعث بهبود نرخ تشخیص احساس خشم، ترس و خوشحالی نسبت به حالت طبیعی می‌شوند.

کلمات کلیدی: تشخیص احساسات از گفتار، بسته موجک، ضرایب آنتروپی شانون، ماشین بردار پشتیبان.

Wavelet Packet Entropy in Speaker-Independent Emotional State Detection from Speech Signal

Mina Kadkhodaei Elyaderani[1] - Seyed Hamid Mahmoodian[2] - Hossein Pourghassem[2] - Ghazaal Sheikhi[3]

(1) MSc. - Bonyan Institute of Higher Education Shahinshahr, Isfahan, Iran
minakadkhodaee@gmail.com

(2) Assistant Professor - Electrtical Engineering Department, Najafabad Branch, Islamic Azad University, Najafabad, Esfahan, Iran
mahmoodian_hamid@yahoo.com
h_pourghasem@iaun.ac.ir

(3) Phd Student, Computer Engineering Department, Eastern Mediterranean University, Turkey
ghazaal.sheikhi@gmail.com

Abstract:In this paper, wavelet packet entropy is proposed for speaker-independent emotion detection from speech. After pre-processing, wavelet packet decomposition using wavelet type db3 at level 4 is calculated and Shannon entropy in its nodes is calculated to be used as feature. In addition, prosodic features such as first four formants, jitter or pitch deviation amplitude, and shimmer or energy variation amplitude besides MFCC features are applied to complete the feature vector. Then, Support Vector Machine (SVM) is used to classify the vectors in multi-class (all emotions) or two-class (each emotion versus normal state) format. 46 different utterances of a single sentence from Berlin Emotional Speech Dataset are selected. These are uttered by 10 speakers in sadness, happiness, fear, boredom, anger, and normal emotional state. Experimental results show that proposed features can improve emotional state detection accuracy in multi-class situation. Furthermore, adding to other features wavelet entropy coefficients increase the accuracy of two-class detection for anger, fear, and happiness.

Index Terms: Speech emotion recognition, wavelet Packet, shannon entropy coefficients, support vector machine.

نویسنده مسئول: مینا کدخدایی الیادرانی، دانشگاه آزاد اسلامی، واحد نجف‌آباد، minakadkhodaee@gmail.com

۱- مقدمه

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. 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. نتایج این پژوهش‌ها در برنامه‌های کاربردی کامپیوتر، ابزار تشخیص برای درمانگران، مراکز پاسخگویی خودکار، ارتباطات تلفن همراه و غیره کاربرد فراوانی دارد. سیستم‌های تشخیص احساسات از گفتار، بخش مهمی از تحقیقات رو به رشد در این حوزه را به خود اختصاص داده‌اند [۱-۳]. با این حال علیرغم تحقیقات گسترده، مشکلات فراوانی نیز در این سیستم‌ها وجود دارد. احساس انسان، پدیده‌ای پیچیده، مبهم و مرکب است. در اغلب اوقات در هنگام برقراری ارتباط بین افراد، احساسات کامل، خالص و پایه بروز نمی‌کنند بلکه معمولاً ترکیبی از احساسات مختلف در یک لحظه ممکن است بروز کنند [۴]. بنابراین جداسازی، تشخیص و تشریح محتوای احساسیِ گفتار حتی توسط عوامل انسانی، بسیار دشوار است. علاوه بر آن نحوه بروز احساسات در گفتار به فرهنگ و زبان، محتوای گفتار، جنسیت و سن گوینده و بسیاری از عوامل دیگر وابسته است [۵]. تمامی این مسائل روند تشخیص حالت احساس از گفتار را پیچیده‌تر می‌کنند. به طور کلی سیستم تشخیص احساسات از گفتار شامل دو مرحله است: استخراج ویژگی و طبقه‌بندی. در مرحله اول، باید اطلاعاتی از سیگنال گفتار استخراج شود که حداکثر همبستگی را با احساسات داشته باشند و در عین حال به سایر عوامل از جمله محتوای گفتار و تغییر گوینده وابسته نباشد. نشان داده شده است که ویژگی‌های نوایی[1] مثل فرکانس پایه، فرمنت‌ها و انرژی گفتار اطلاعات احساسی زیادی را منتقل می‌کنند [۶]، [۷]. ویژگی‌های زمان-فرکانس مثل ضرایب کپسترال مِل نیز در این حوزه بسیار پرکاربرد هستند [۸]. به طور کلی پرکاربردترین ویژگی‌ها در زمینه تحقیقات بازشناسی احساسات از گفتار عبارتند از ضرایب کپسترال فرکانسی مل (MFCC) و مشتقات آنها [۹،۱۰]، ضرایب پیشگویی خطی (LPC) [۱۱]، فرکانس فرمنت‌ها، جیتر یا دامنه تغییرات فرکانس گام و شیمر یا دامنه تغییرات انرژی [۲]، [۱۲-۱۳].

کارآیی این ویژگی‌ها و ترکیب‌های مختلف آنها برای تشخیص احساسات از گفتار در زبان‌هایی مثل انگلیسی [۱۴]، آلمانی [۱۵]، هلندی [۱۶] و غیره مورد بررسی قرار گرفته است. با این حال این ویژگی‌های متداول با تغییر گوینده به شدت تغییر می‌کنند و انتخاب نحوه ترکیب مؤثر ویژگی‌ها برای ساخت بردار ویژگی بسیار حائز اهمیت بوده و می‌تواند کارآیی سیستم را به شدت تحت تأثیر قرار دهد. بنابراین نیاز به استفاده از ابزارهای پردازشی نوین برای استخراج ویژگی در این حوزه به شدت وجود دارد. تبدیل موجک به عنوان ابزاری غیرایستا می‌تواند تغییرات طیفی را بسیار بهتر از روش‌های متداول آشکار کند و به همین دلیل کاربرد فراوانی در حوزه پردازش گفتار دارد. به عنوان مثال در [۱۷] نشان داده شده است که آنتروپی بسته موجک می‌تواند کارآیی سیستم تشخیص گوینده را بهبود بخشد. در این مقاله بازشناسی احساسات از گفتار در حالت مستقل از گوینده با استفاده از آنتروپی بسته موجک مورد بررسی قرار می‌گیرد. بسته موجک، تعمیم یافته تبدیل موجک است و سیگنال را در هر مرحله از درخت موجک به دو زیرباند تفکیک می‌کند. این سیگنال‌های تفکیک شده به دلیل حجم بالای اطلاعات، مستقیماً به صورت بردار ویژگی قابل استفاده نیستند. به همین دلیل آنتروپی در گره‌های درخت موجک به عنوان معیاری از محتوای اطلاعاتی محاسبه شده و به عنوان ویژگی مورد استفاده قرار گرفته است. در اینجا، ویژگی‌های نواییِ گفتار شامل فرکانس چهار فرمنت اول، جیتر یا دامنه تغییرات فرکانس گام و شیمر یا دامنه تغییرات انرژی به عنوان ویژگی‌های پرکاربرد در حوزه تشخیص احساسات، در کنار ضرایب فرکانسی کپسترال مل (MFCC) برای ساخت بردار ویژگی مورد استفاده قرار گرفته‌اند. برای تکمیل بردار ویژگی استفاده از آنتروپی در گره‌های بسته موجک پیشنهاد شده است. بسته موجکِ db3 سطح ۴ در هر فریم محاسبه شده است و آنتروپی شانون در گره‌های آن به عنوان ویژگی در نظر گرفته شده است. پس از استخراج ویژگی‌ها، از ماشین بردار پشتیبان (SVM) [۱۸ و ۱۹] به عنوان طبقه‌بندی کننده استفاده شده است. آزمایشات بر روی بخشی از دادگان گفتار احساسی دانشگاه برلین (Emo-DB) به زبان آلمانی انجام شده است. در این مقاله از یک جمله واحد که توسط ۱۰ گوینده مختلف با حالت‌های احساسی ناراحتی، خوشحالی، ترس، ملالت، خشم و حالت طبیعی بیان شده است، استفاده شده تا حالت مستقل از گوینده مورد بررسی قرار گیرد. به منظور مقایسه ترکیب‌های مختلفی از بردار ویژگی برای دسته‌بندی حالات احساسی در حالت دو کلاسه و چندکلاسه مورد آزمایش قرار گرفته است. در ادامه مقاله، در بخش دوم ابتدا استخراج ویژگی توضیح داده شده است. بخش سوم به طبقه‌بندی کننده SVM می‌پردازد. در بخش چهارم به نتایج آزمایشات پرداخته شده است و نهایتاً در بخش پنج نتایج مورد بحث و بررسی قرار گرفته‌اند.

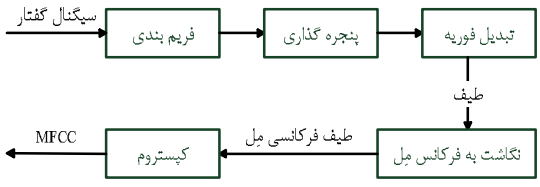
۲- استخراج ویژگی

در بازشناسی احساسات، محققان ویژگی‌های مختلف زمان-فرکانس و نوایی را استفاده کرده‌اند. در این مقاله از ویژگی‌های پرکاربرد صوتی یعنی ضرایب MFCC، فرمنت‌ها، جیتر و شیمر در ترکیب با ویژگی‌های پیشنهادی مبتنی بر آنتروپی بسته موجک استفاده شده است. در مرحله استخراج ویژگی با توجه به آنکه سیگنال گفتار، یک سیگنال غیر ایستا است، برای استخراج ویژگی‌های ایستا باید آن را به بازه‌های در حدود ۲۰ تا ۱۰۰ میلی‌ثانیه تقسیم کرد. در عین حال برای استخراج صحیح ویژگی‌های مبتنی بر فرکانس گام مثل جیتر، باید هر پنجره حاوی حداقل دو تناوب پایه باشد. از آنجایی که گویندگان

موجود شامل زن و مرد هستند، حداقل فرکانس گام با توجه به گویندگان مرد ۵۰ هرتز در نظر گرفته شده است. بنابراین طول فریم ۴۰ میلی‌ثانیه 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. برای کاهش اثر دامنه سیگنال در مرزهای فریم و تمرکز بر اطلاعات بخش مرکزی، فریم‌ها به وسیله پنجره همینگ پنجره‌گذاری شده‌اند. در ادامه جزئیات استخراج ویژگی‌های متداول و ویژگی‌های پیشنهادی آمده است.

۲-۱- استخراج ویژگی‌های متداول

پس از فریم‌بندی و پنجره‌گذاری، سیگنال برای استخراج ویژگی‌ها آماده است. ویژگی پایه در این تحقیق، ضرایب MFCC هستند. ایده اصلی در استخراج ضرایب MFCC، برگرفته از خواص گوش انسان در دریافت و فهم گفتار است و همین مسئله این ضرایب را به ابزاری قدرتمند در تمامی حوزه‌های پردازش و بازشناخت گفتار تبدیل کرده است. تعداد ضرایب مورد استفاده در بازشناسی گفتار معمولاً بین ۹ تا ۱۳ ضریب تغییر می‌کند. ضریب صفرم نشان‌دهنده انرژی است و از آنجایی که اطلاعات انرژی در ویژگی شیمر لحاظ شده است، تعداد ضرایب مورد استفاده در این مقاله ۱۲ عدد است. ویژگی شیمر و نحوه استخراج آن در ادامه آمده است. برای استخراج این ضرایب در مرحله اول، تبدیل فوریه بر سیگنال اعمال می‌شود. سپس توان طیف به دست آمده به مقیاس مِل نگاشته می‌شود و از توان در هر فرکانس مِل، لگاریتم گرفته می‌شود. در مرحله آخر طیف لگاریتمیِ مِل به حوزه زمان برگردانده می‌شود. نتیجه این تبدیلات نمایش کپسترال طیف سیگنال گفتار است که ویژگی‌های طیفی یک فریم از سیگنال گفتار را به خوبی نشان می‌دهد. شکل (۱) روند استخراج MFCC را نمایش می‌دهد. فرکانس گام و فرکانس چهار فرمنت اول با استفاده از الگوریتم چگالی توان طیفی (PSD) استخراج شده‌اند [۲۰ و ۲۱]. این روش شامل دو مرحله است. ابتدا چگالی توان طیفی سیگنال با استفاده از روش Yule-Walker خودبازگشتی (AR) تخمین زده می‌شود. این روش سیگنال را با فیلتر خودبازگشتی خطی مدل می‌کند به نحوی که خطای پیش‌بینی با معیار میانگین مربعات حداقل شود.



شکل (۱): روند استخراج MFCC

Fig. (1): MFCC extraction process

این فرمول معادله Yule-Walker را نتیجه می‌دهد که با روش بازگشتیِ Levinson-Durbin حل می‌شود. مجذور دامنه پاسخ فرکانسی مدل خودبازگشتی که از این روش به دست می‌آید، همان تخمین طیف سیگنال است. سپس چگالی توان طیفی در مقیاس لگاریتمی از روش زیر به دست می‌آید:

$$F_{P_{XX}} = \sum 10\log_{10}(P_{XX}) \qquad (۱)$$

که در آن P_{XX} تخمین طیف سیگنال و F_{PXX} چگالی توان طیفی است. پس از تخمینِ PSD، تغییرات طیف در (D) به صورت زیر محاسبه می‌شود.

$$D = F_{P_{XX}}(i+1) - F_{P_{XX}}(i) \qquad (۲)$$

سپس بیشینه‌های محلی به عنوان اندیس نقاطی که D از مثبت به منفی تغییر علامت می‌دهد، شناسایی می‌شوند. محل قرارگیری اولین نقطه بیشینه معادل فرکانس گام و چهار بیشینه بعدی نشان دهنده چهار فرمنت اول هستند. مرتبه مدل خودبازگشتی مورد استفاده ۱۲ است که برای سیگنال گفتار مناسب است. تعداد نقاط برای محاسبه طیفی نیز ۲۵۶ در نظر گرفته شده است که در نتیجه طیف تخمینی ۱۲۸ نقطه‌ای خواهد بود و برای نرمال‌سازی مقدار اندیس‌ها به ۱۲۸ تقسیم شده است. چهار فرمنت اول پس از نرمال‌سازی مستقیماً به عنوان ویژگی مورد استفاده قرار می‌گیرند و فرکانس گام برای استخراج ویژگی جیتر مورد استفاده قرار می‌گیرد.

دو ویژگیِ جیتر و شیمر نیز از جمله ویژگی‌های مورد استفاده در تشخیص احساسات هستند. ویژگی جیتر دامنه تغییرات در فرکانس گام را از یک فریم به فریم بعدی نشان می‌دهد. در اینجا از فرکانس گام نرمال شده، که در مرحله قبلی به دست آمده استفاده می‌شود. اختلاف مقدار فرکانس پایه در یک فریم با مقدار آن در فریم قبلی، به میانگین فرکانس گام در یک جمله تقسیم شده و به عنوان ویژگی جیتر مورد استفاده قرار می‌گیرد. فرمول (۳) نحوه محاسبه جیتر را نشان می‌دهد که در آن $F_0(i)$ نشان دهنده مقدار فرکانس پایه در فریم iام از یک جمله است که به n فریم تقسیم شده است. علامت | | نشان دهنده قدرمطلق است.

$$\text{Jitter}(i) = \left|F_0(i) - F_0(i-1)\right| \Big/ \text{mean}\{F_0(i), i = 2,3,...,n\} \qquad (۳)$$

ویژگی شیمر تغییرات انرژی زمان کوتاه[۲] بین فریم‌های متوالی است. برای محاسبه شیمر از انرژی جذر میانگین مربعات (RMS) استفاده شده است. انرژی RMS از معادله (۴) به دست می‌آید که در آن E(i) انرژی فریم iام و S(k) مقدار نمونه‌های سیگنال گفتار در این فریم پس از پنجره‌گذاری هستند.

$$E(i) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} S^2(k)} \qquad (۴)$$

پس از محاسبه مقادیر انرژی در یک جمله شامل n فریم، ویژگی شیمر برای فریمِ iام به صورت زیر به دست می‌آید:

$$\text{Shimmer}(i) = \left|E(i) - E(i-1)\right| \Big/ \text{mean}\{E(i), i = 2,3,...,n\} \qquad (۵)$$

۲-۲- استخراج ویژگی‌های پیشنهادی

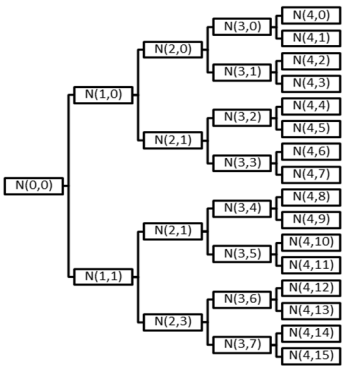
بسته موجک، تعمیم یافته تجزیه موجک است که امکان بیشتری برای تحلیل سیگنال ایجاد می‌کند. در تبدیل موجک، سیگنال به دو شاخه تقریب[۳] و جزییات[۴] تقسیم می‌شود و این روند بر روی شاخه تقریب تکرار می‌شود. برای n سطح یعنی درختی به عمق n، تعداد n+1 مسیر ممکن برای تجزیه یا کدگذاری سیگنال وجود دارد. در تحلیل بسته موجک، جزییات را نیز می‌توان همانند تقریبات تجزیه کرد. حاصل این کار دسترسی به بیش از 2^n روش متفاوت برای کدگذاری سیگنال است. در حقیقت تجزیه موجک گوشه‌ای از تجزیه به روش بسته موجک است. بنابراین بسته موجک نمایش بهتری از سیگنال نسبت به تبدیل موجک ارائه می‌دهد. به همین دلیل در این مقاله پیشنهاد شده است برای استخراج ویژگی‌های کمکی به منظور بهبود نرخ تشخیص به کار رود.

با افزایش عمق درخت موجک، سیگنال به صورت جزیی‌تری تجزیه می‌شود اما زمان مورد نیاز برای استخراج ویژگی و طبقه‌بندی داده‌ها به شدت افزایش می‌یابد. بنابراین اولاً انتخاب مناسب عمق پیشروی در درخت تجزیه حائز اهمیت است و ثانیاً این ضرایب به صورت مستقیم به عنوان ویژگی قابل استفاده نبوده و کاهش بعد ضرایب به دست آمده امری ضروری است. برای کاهش بعد، ضرایب آنتروپی شانون در گره‌های درخت تجزیه موجک محاسبه می‌شوند. آنتروپی معیاری از محتوای اطلاعاتی موجود در گره‌ها بوده و به نحو مؤثری اطلاعات طیفی سیگنال را بازنگاری می‌کند. عمق مناسب پیشروی در درخت نیز با انجام آزمایشات طبقه‌بندی به دست می‌آید که جزییات مربوط به آن در بخش چهار مقاله آمده است. شکل (۲) درخت موجک را با عمق پیشروی ۴ نشان می‌دهد. همان طور که مشاهده می‌شود در این حالتِ تجزیه ۳۱ گره مختلف وجود دارد.

پس از تجزیه سیگنال، آنتروپی شانون در تمامی ۳۱ گره به صورت زیر محاسبه می‌شود.

$$e_m(i) = -\sum_j s_m^2(j)\log_{10} s_m^2(j) \qquad (۶)$$

که در آن $e_m(i)$ آنتروپی فریم iام در گره m $(m = 1, 2, \ldots, 31)$ است و $s_m(j)$ نمونه‌های سیگنال تجزیه شده در این گره هستند. بردار آنتروپی نهایی برای هر فریم یک بردار ۳۱ بعدی خواهد بود.



شکل (۲): درخت بسته موجک با عمق پیشروی ۴

Fig. (2): Wavelet packet tree at level 4

۲-۳- طبقه‌بندی

در سال‌های اخیر الگوریتم‌های طبقه‌بندی مختلفی در بازشناسی حالات احساسی پیشنهاد شده است از جمله شبکه‌های عصبی (NN)، مدل‌های مخلوط گاوسی (GMM)، مدل‌های مخفی مارکوف (HMM)، طبقه‌بندی کننده بیزینِ شباهت بیشینه (MLC)، طبقه‌بندی کننده بازگشتی کرنل[۵]، روش نزدیکترین همسایگی (KNN) و ماشین بردار پشتیبان (SVM). ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و محاسبات بازگشتی استفاده می‌شود. کاربردهای این روش در دسته‌بندی‌های دو و چند کلاسه در سال‌های اخیر افزایش یافته است. مبنای کار SVM دسته‌بندی خطی داده‌هاست. برای اینکه ماشین بتواند دادگانی با پیچیدگی بالا را دسته‌بندی کند باید دادگان توسط کرنل مناسب به فضای با ابعاد بالاتر نگاشت شوند. سپس اَبَرصفحه‌ای برای جداسازی دادگان انتخاب می‌شود که حاشیه اطمینان بیشتری داشته باشد.

این روش یکی از الگوریتم‌های ساده و کارآ با حجم محاسبات کم در یادگیری ماشین است و کاربرد گسترده‌ای در طبقه‌بندی الگوها دارد. علاوه بر آن در شرایطی که حجم داده‌ها محدود باشد، SVM نرخ بازشناسی بهتری نسبت به سایر روش‌های طبقه‌بندی خواهد داشت. بنابراین در این مقاله برای تشخیص حالات احساسی در هر فریم گفتار از این روش استفاده شده است. در آموزش SVM کرنل‌ها و پارامترهای آن نقش مهمی دارند. بنابراین باید به درستی انتخاب شوند تا دقت دسته‌بندی بهبود یابد. کرنل چندجمله‌ای با درجه پایین و کرنل با تابع پایه شعاعی اولین انتخاب‌ها در طبقه‌بندی با SVM هستند. بررسی‌ها نشان می‌دهند کرنل چندجمله‌ای در دسته‌بندی داده‌ها حالات احساسی انتخاب مناسبی نیست [۲۲]. اما استفاده از کرنل با تابع پایه شعاعی (RBF) می‌تواند به پاسخ‌های قابل قبولی منجر شود، لذا در این مقاله از کرنل پایه شعاعی استفاده شده است .

۲-۴- نتایج شبیه‌سازی

کارآیی روش پیشنهادی در این مقاله با استفاده از داده‌ها استاندارد گفتار احساسی زبان آلمانی مورد بررسی قرار گرفته است. داده‌ها گفتار احساسی برلین (Emo-DB) توسط دانشکده فناوری صوتی دانشگاه صنعتی برلین تهیه شده و دسترسی به آن برای همه محققین آزاد است. این داده‌ها در زمینه بازشناسی احساسات از گفتار بسیار پرکاربرد هستند و امکان دانلود بخش‌های مختلف آن به صورت انتخابی وجود دارد. هدف این مقاله بررسی کارآیی ویژگی‌های پیشنهادی به صورت مستقل از گوینده است و لازم است تنوع گویندگان در داده‌ها تست و تعلیم حفظ شود. به همین منظور ۴۶ بیان مختلف از جمله واحد که توسط ۱۰ گوینده مختلف با حالت‌های احساسی ناراحتی، خوشحالی، ترس، ملالت، خشم و حالت طبیعی بیان شده است به عنوان داده‌ها

انتخاب شده است. نسبت داده‌ها آموزش به تست ٪۷۰ به ٪۳۰ است. سعی شده است برای حفظ تنوع، جملاتی که با یک حالت احساسی مشخص، چندبار بیان شده‌اند همگی در داده‌ها آموزشی قرار بگیرند. به این ترتیب داده‌ها تست از بین جملاتی انتخاب شده‌اند که توسط گوینده مشخص تنها یک بار با یک حالت احساسی خاص بیان شده‌اند.

در اولین آزمایش عمق مناسب پیشروی در درخت موجک مورد بررسی قرار گرفته است. به این منظور سیگنال در هر فریم به وسیله بسته موجکِ db3 با عمق پیشروی مشخص آنالیز شده و آنتروپی شانون در گره‌های آن به عنوان بردار ویژگی درنظر گرفته شده است. انتخاب موجک db3، ترکیبی از روند سعی و خطا و تجربیات قبلی است. کاربرد موفقیت آمیز موجک daubechies برای استخراج ویژگی بسیار گزارش شده است [۲۳]. با مقایسه عملکرد تشخیص، با چندین نوع موجک مادر daubechies، موجک مادر db3، مناسب‌تر تشخیص داده شد. جدول (۱)، دلیلی بر این مدعاست.

ضرایب آنتروپی برای عمقِ ۲، ۳، ۴، ۵ و ۶ در درخت بسته موجک استخراج شده‌اند. سپس این بردارهای ویژگی برای طبقه‌بندی شش حالت احساسی ناراحتی، خوشحالی، ترس، ملالت، خشم و حالت طبیعی به صورت چنددسته‌ای به وسیله ماشین بردار پشتیبان مورد استفاده قرار گرفته‌اند.

Table (1): Detection rate for different wavelets of daubechies family

جدول (۱): نرخ تشخیص به ازای موجکهای مختلف خانواده دابیچی

بردار ویژگی	نوع موجک	درصد تشخیص
Wavelet Entropy	db3	61.66
Wavelet Entropy	db2	57.13
Wavelet Entropy	db4	52.11
Wavelet Entropy	db5	49.8
Wavelet Entropy	db1	44.65

Table (2): Emotional state detection rates for different values of decomposition depth in wavelet tree

جدول (۲): نرخ تشخیص حالات احساسی به ازای مقادیر مختلفی از عمق پیشروی در درخت موجک

عمق پیشروی	تعداد ضرایب آنتروپی	نرخ تشخیص (درصد)
2	7	37.42
3	15	51.00
4	31	61.66
5	63	61.86
6	127	55.30

جدول (۲) نتایج نرخ تشخیص را به ازای مقادیر مختلفی از عمق پیشروی نشان می‌دهد. نرخ تشخیص از تقسیم تعداد فریم‌هایی که به درستی طبقه‌بندی شده‌اند، به تعداد کل فریم‌ها در دادگان تست به دست آمده و به صورت درصد بیان شده است. تعداد ضرایب آنتروپی یا همان ابعاد بردارهای ویژگی در هر حالت نیز در جدول نمایش داده شده است. همان طور که مشاهده می‌شود با افزایش عمق پیشروی در درخت موجک از ۴ به ۵، نرخ تشخیص افزایش چشم‌گیری پیدا نمی‌کند. این در حالی است که ابعاد بردار ویژگی تقریباً دو برابر شده و حجم محاسبات و زمان آن افزایش می‌یابد. افزایش بیشتر عمق پیشروی از ۵ به ۶ باعث افت نرخ تشخیص شده است. دلیل این مسئله می‌تواند بزرگ شدن بیش از حد ابعاد بردارهای ویژگی باشد که مانع از دسته‌بندی صحیح می‌شود. با توجه به نتایج این جدول در ادامه آزمایشات، عمق مناسب برابر ۴ با تعداد ۳۱ ضریب در نظر گرفته شده است. در آزمایش بعدی کارآیی روش پیشنهادی در مقایسه با روش‌های متداول بررسی شده است. به این منظور ترکیب‌های مختلفی از بردار ویژگی ساخته شده و به صورت چند دسته‌ای طبقه‌بندی شده‌اند. نرخ تشخیص و ابعاد بردار ویژگی در هر حالت، در جدول (۳) نشان داده شده است. همانطور که مشاهده می‌شود، استفاده از ویژگی‌های پیشنهادی بهترین نرخ تشخیص را منجر می‌شود. البته نتیجه مشابهی از ترکیب این ویژگی با MFCC به دست آمده است که می‌تواند به دلیل وجود اطلاعات زائد در بردار ویژگی باشد. استفاده از بردار ترکیبی شامل ضرایب آنتروپی، فرمنت‌ها، جیتر و شیمر نیز نتایج قابل قبولی (نرخ تشخیص ٪۶۰) در بر دارد. اما با وجودی که ابعاد بردار ویژگی در این حالت نسبت به ضرایب آنتروپی کمتر است، حجم محاسبات مورد نیاز برای استخراج این ویژگی‌ها بسیار بیشتر از روش پیشنهادی است. به منظور مقایسه، نتایج مقاله [۲۳] نیز در جدول آمده است.

در ۲۰۱۲، Ntalampiras و Fakotakis مدل کردن تغییرات زمانی پارامترهای صوتی را برای بازشناسی احساسات پیشنهاد کردند. در این مقاله بازشناسی احساسات در حالت مستقل از گوینده را بر روی ۶ حالت احساسی با ۱۰ گوینده مختلف مورد بررسی قرار داده‌اند. داده‌ها مورد استفاده در مقاله مذکور دادگان گفتار احساسی برلین است و از ویژگی‌هایی با پیچیدگی محاسباتی بالا استفاده شده است. ویژگی‌های مختلفی از جمله ویژگی‌های آماری زمان‌کوتاه، ممنتوم‌های طیفی و مدل‌های بازگشتی به همراه ویژگی‌های جدیدی مبتنی بر بسته موجک ادراکی[۶] در این تحقیق مورد بررسی قرار گرفته‌اند. طبقه‌بندی با استفاده از مدل‌های مخفی مارکوف (HMM) انجام شده است که از نظر زمان مورد نیاز و پیچیدگی محاسباتی بر ماشین بردار پشتیبان برتری ندارد.

علاوه بر آن در مقاله مذکور نتایج و بررسی‌های مختلفی گزارش شده است که از جمله آن تأثیر تغییر طول فریم‌های گفتار است. توضیح جزییات مقاله و بررسی تمامی نتایج آن از حوصله بحث خارج است. بنابراین در اینجا بخشی از نتایج که از نظر طول فریم و نوع ویژگی‌ها با روش پیشنهادی ما قابل مقایسه است، مورد مقایسه قرار گرفته است. مقدار درج شده در جدول (۳) نرخ بازشناسی به دست آمده در مقاله [۲۴] را نشان می‌دهد که با طول فریم ۴۰ میلی‌ثانیه با تکیه بر ۲۵ ویژگی آماری زمان کوتاه و ۱۷ ویژگی مبتنی بر تبدیل موجک ادراکی به دست آمده است. بنابراین ابعاد بردار ویژگی ۴۲ است که نسبت به روش پیشنهادی بردار بزرگتری است. با این حال مشخص است که

استفاده از ویژگی‌های مبتنی بر آنتروپی بسته موجک نرخ بازشناسی نسبتاً بهتری نسبت به این روش دارد.

به منظور بررسی بیشتر، در آزمایش بعدی طبقه‌بندی دودسته‌ای برای هر حالت احساسی نسبت به حالت طبیعی انجام شده است. داده‌ها آموزش و تست با حالت طبقه‌بندی چنددسته‌ای یکسان هستند. در اینجا نیز ترکیب‌های مختلفی از بردار ویژگی مورد آزمایش قرار گرفته‌اند. نتایج در جدول (۴) آمده است و بهترین نرخ تشخیص در هر ستون با فونت ضخیم مشخص شده است. برای آنکه امکان مقایسه نتایج در دو حالت چنددسته‌ای و دو دسته‌ای فراهم شود، تعداد فریم‌های آموزش و تست نیز به تفکیک حالت احساسی در جدول (۴) درج شده است.

Table(3): Emotional state detection rates for various combinations feature vector

جدول (۳) نرخ تشخیص حالات احساسی برای ترکیب‌های مختلف بردار ویژگی

بردار ویژگی	ابعاد بردار ویژگی	نرخ تشخیص (درصد)
MFCC	12	58.29
Wavelet Entropy	31	61.66
MFCC + Formants + Jitter + Shimmer	18	60
Wavelet Entropy + Formants + Jitter + Shimmer	37	53.33
Wavelet Entropy + MFCC	43	61.66
Wavelet Entropy + MFCC + Formants + Jitter + Shimmer	49	55
Short-Term Statistics+PWP [16]	17+25	60.30

Table(4): Detection rates in two-class classification for each emotional state regarding normal

جدول (۴): نرخ تشخیص در طبقه‌بندی دو دسته‌ای برای هر حالت احساسی نسبت به حالت طبیعی

حالت احساسی	خشم	ملالت	ترس	خوشحالی	ناراحتی
تعداد فریمها	آموزش: ۲۴۰۳	آموزش: ۱۶۰۹	آموزش: ۱۱۲۱	آموزش: ۱۴۶۲	آموزش: ۱۷۲۳
بردار ویژگی	تست: ۱۰۳۳	تست: ۶۹۱	تست: ۴۷۸	تست: ۶۲۹	تست: ۷۴۰
MFCC	76.28	71.42	60.00	70.00	75.71
Wavelet Entropy	77.14	70.00	65.71	65.71	68.57
MFCC + Formants + Jitter + Shimmer	77.14	80.00	65.71	70.00	78.57
Wavelet Entropy + Formants + Jitter + Shimmer	71.42	67.14	61.35	61.42	72.85
Wavelet Entropy + MFCC	78.57	67.14	58.57	71.42	77.14
Wavelet Entropy + MFCC + Formants + Jitter + Shimmer	81.42	65.71	62.85	64.28	68.57

همان طور که نتایج نشان می‌دهند در حالت احساسی خشم با اضافه شدن ضرایب پیشنهادی نرخ تشخیص بهبود یافته است. این حالت احساسی دارای بیشترین تعداد فریم در داده‌ها است. با این حال بهترین نرخ تشخیص از ترکیب تمامی ویژگی‌ها به دست آمده است و تعداد زیاد داده‌ها برای این حالت احساسی نمی‌تواند نتایج جدول (۳) را تحت تأثیر قرار دهد. برای حالت ملالت، استفاده از ضرایب پیشنهادی منجر به بهبود نرخ تشخیص نمی‌شود و بهترین نتیجه از ترکیب MFCC با ویژگی‌های نوایی به دست می‌آید. برای حالت احساسی ترس، ضرایب آنتروپی نتیجه‌ای مشابه بردار ترکیبی از ضرایب MFCC و ویژگی‌های نوایی دارند. اما همان طور که پیشتر اشاره شد ضرایب آنتروپی از نظر حجم و زمان محاسبه نسبت به این بردار ترکیبی ارجحیت دارند. برای حالت خوشحالی، ترکیب ضرایب پیشنهادی با ضرایب MFCC منجر به بهبود نرخ تشخیص شده است. در نهایت برای حالت احساسی ناراحتی مشابه حالت ملالت، بهترین نرخ تشخیص از ترکیب MFCC با ویژگی‌های نوایی به دست می‌آید. در نهایت به نظر نمی‌رسد تعداد فریم‌های مربوط به هر حالت احساسی در دادگان، تأثیری بر نتایج طبقه‌بندی چند دسته‌ای داشته باشد. به طور کلی مقایسه نتایج جدول (۳) و (۴) دو نکته را مشخص می‌کند. اولاً نمی‌توان ارتباط مستقیمی بین نتایج در دو حالت دو دسته‌ای و چند دسته‌ای یافت. ثانیاً استفاده از ضرایب پیشنهادی در حالت چند دسته‌ای منجر به بهبود قطعی در نرخ تشخیص می‌شود. حالت چند دسته‌ای طبقه‌بندی پیچیده‌تری بوده و معمولاً در مقالات مورد نظر است. ضمناً برای حالات احساسی خاص، از جمله خشم، ترس و خوشحالی استفاده از ضرایب آنتروپی به بهبود نرخ تشخیص منجر می‌شود.

۳- بحث و نتیجه گیری

در این مقاله بازشناسی احساسات از گفتار در حالت مستقل از گوینده با استفاده از ویژگی‌های مبتنی بر آنتروپی بسته موجک مورد بررسی قرار گرفته است. به این منظور بسته موجکِ db3 در سطح ۴ در هر فریم محاسبه شده است و آنتروپی شانون در گره‌های آن به عنوان بردار ویژگی درنظر گرفته شده است. بازشناسی احساسات از گفتار، بسیار پیچیده است و انتخاب ویژگی‌های مناسب در این حوزه از اهمیت ویژه‌ای برخوردار است. علاوه بر آن، ویژگی‌های نواییِ گفتار شامل فرکانس چهار فرمنت اول، جیتر یا دامنه تغییرات فرکانس گام و

شیمر یا دامنه تغییرات انرژی به عنوان ویژگی‌های پرکاربرد در حوزه تشخیص احساسات در کنار ضرایب فرکانسی کپسترال مل (MFCC)برای تکمیل بردار ویژگی مورد استفاده قرار گرفته‌اند.

پس از استخراج بردار ویژگی طبقه‌بندی با استفاده از ماشین بردار پشتیبان (SVM) و با نسبت داده‌ها ٪۷۰ به ٪۳۰ برای آموزش و تست انجام شده است. طبقه‌بندی ابتدا به صورت طبقه‌بندی چنددسته‌ای بر روی تمامی حالت‌های احساسی انجام گرفته است و ترکیب‌های مختلفی از بردار ویژگی بررسی قرار گرفته‌اند.

نتایج نشان می‌دهند در حالت چند دسته‌ای استفاده از ضرایب آنتروپی بسته موجک به عنوان بردار ویژگی، نرخ بازشناسی را بهبود می‌بخشد. این نتایج مؤثر بودن ضرایب آنتروپی بسته موجک در نمایش محتوای اطلاعاتی طیف سیگنال را نشان می‌دهند. برای بررسی بیشتر طبقه‌بندی در هر حالت احساسی نسبت به حالت طبیعی به صورت دودسته‌ای انجام گرفته است. نتایج تایید می‌کنند که ویژگی‌های پیشنهادی می‌توانند در ترکیب با سایر ویژگی‌ها در طبقه‌بندی احساس خشم، ترس و خوشحالی نرخ تشخیص را بهبود می‌بخشد. این احساسات بر خلاف دو حالت احساسی ملالت و ناراحتی‌که بیشتر ویژگی‌های نوایی گفتار را تغییر می‌دهند، تأثیر پیچیده‌ای بر طیف سیگنال گفتار دارند که با آنالیز بسته موجک بهتر از ویژگی‌های متداول قابل آشکارسازی است. هرچند نمی‌توان نتایجِ دو حالت مختلف از دسته‌بندی را مستقیماً مورد مقایسه قرار داد، در حالت کلی می‌توان گفت ضرایب آنتروپی بسته موجک در بازنمایی اطلاعات احساسی گفتار کارآیی مناسبی دارند.

پی‌نوشت

1- Prosodic
2- Short-term
3- Approximation
4- Detail
5- Kernel Regression
6- Perceptual Wavelet Transform (PWP)

References

[1] M. Ayadi, M. Kamel, "Servey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, Vol. 44, pp. 72-587, 2011.

[2] D. Ververidis, C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods", Speech Communication, Vol. 48, pp. 1162–1181, 2006.

[3] B. Schuller, G. Rigoll, M. Long, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine –belief network architecture", Proceedings of the IEEE/ICASSP, Vol. 1, pp. 577–580, May 2004.

[4] France,et. al., "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Trans. on Biomedical Engineering, Vol. 47, No. 7, pp. 829-837, 2000.

[5] T. Pao, C. Wang, "A study on the search of the most discriminative speech features in the speaker dependent speech emotion recognition", Proceeding of the IEEE/PAAP, pp. 157-162, 2012.

[6] C. Busso, S. Lee, S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", IEEE Trans. on Audio Speech Language Process, Vol. 17, pp. 582–596, 2009.

[7] B. Schuller, et. al.,"The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals", Proc. Inter speech, pp. 2253–2256, 2007.

[8] B. Vlasenkoet. al, "Combiningframe and turn-level information for robust recognition of emotions within speech", Proc. Interspeech, pp. 2225–2228, 2007.

[9] X. Mao, L. Chenand L. Fu, "Multi-level speech emotion recognition based on HMM and ANN", Proceeding of the World Cong. on Computer Science and Information Engineering, 2009.

[10] T.Polzehl, et. al, "Anger recognition in speech using acoustic and linguistic cues", Speech Communication, Vol. 53, pp. 1198–1209, 2011

[11] H.Marvi, Z.Esmaileyan, A.Harimi, "Estimation of LPC coefficients using evolutionary algorithms", Journal of AI and Data mining, Vol. 1, pp. 111–118, 2013.

[12] L.S. Chen,et. al, "Emotion recognition from audiovisual information", Proceeding of the IEEE/MMSP, pp. 83–88, Redondo Beach, CA, Dec. 1998.

[13] X.Li, "Speech feature toolbox design and emotional speech feature extraction", Thesis Submitted to the Faculty of Graduate School, Marquette University, In Partial Fulfillment of the Requirements for the Degree of Master of Science

[14] Y.Pan, P. Shen, L.Shen, "Feature extraction and selection in speech emotion recognition", Proceeding of the onlinepresent.org, Vol. 2, pp. 64 -69, 2012.

[15] M. Gaurav, "Performance analyses of spectral and prosodic features and their fusion for emotion recognition in speech", Proceeding of the IEEE/SLT, pp. 313-316, Goa, Dec. 2008.

[16] T.Athanaselist, S.Bakamidis, "ASR for emotional speech: clarifying the issues and enhancing performance", Journal of Neural Network, Vol. 18, pp. 437-444, 2005.

[17] K.Daqrouq, "Wavelet entropy and neural network for text-independent speaker identification", Journal Engineering Applications of Artificial Intelligence, Vol. 24, No. 5, pp. 796–802, 2011.

[18] Y.Pan, P. Shen, L.Shen, "Speech Emotion Recognition using support vector machine", International Journal of Smart Home, Vol. 6, No. 2, pp. 101 -108, 2012.
[19] A.Statinkov, et. al, "A Gentle introduction to support vector machines in biomedicine", world scientific.2011
[20] A. Cherif, L. Bouafif, T. Dabbabi, "Pitch detection and formants analysis of Arabic speech processing", Applied Acoustics, Vol. 62, No. 10, pp. 1129–1140, 2001.
[21] J. Clark, C. Yallop, J. Fletcher, "An introduction to phonetics and phonology", 3rded.malden MA, USA: Blackwell publishers.
[22] M. Kadkhodaee, G.H. Sheikhi, H. Mahmoodian, "Survey on time–frequency features for speaker emotion recognition in persian", National Conference Shushtar, 2014.
[23] I. Elamvazuthi, G. Ling, K. Nurhanim, P. Vasant, S. Parasuraman, " Surface electromyography feature extraction based on daubechies wavelets", Proceeding of the ICIEA, pp. 1492–1495, 2013.
[24] S.Ntalampiras, N.Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition", IEEE Trans. on Affective Computing, Vol. 3, No. 1, pp. 116–125, 2012.