Advances in Cognitive Sciences
Vol 22, No. 2, Summer 2020
Research Article
Pages 71-81

Evaluation of implicit emotion in the message through emotional speech processing based on Mel-Frequency Cepstral Coefficient and Short-Time Fourier Transform features

Mahsa Ravanbakhsh[1], Saeed Setayeshi[2*], Mir Mohsen Pedram[3], Azadeh Mirzaei[4]

1. PhD Student of Cognitive Linguistics, Institute for Cognitive Science Studies (ICSS), Tehran, Iran
2. Associate Professor of Department of Physics and Energy Engineering, Amirkabir University of Technology, Tehran, Iran
3. Associate Professor, Department of Electrical and Computer Engineering, Kharazmi University, Tehran, Iran
4. Assistant Professor of Linguistics, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran

Recieved: 28 Apr. 2019
Revised: 1 Dec. 2019
Accepted: 10 Dec. 2019

Keywords
Emotional speech
Emotion recognition
Short time Fourier transform
Mel-frequency Cepstral coefficients
Emotional speech processing

Corresponding author
Saeed Setayeshi, Associate Professor of Department of Physics and Energy Engineering, Amirkabir University of Technology, Tehran, Iran

Email: Setayesh@aut.ac.ir

doi.org/10.30699/icss.22.2.71

Abstract

Introduction: Speech is the most effective way to exchange information. In a speech, a speaker's voice carries additional information other than the words and grammar content of the speech, i.e., age, gender, and emotional state. Many studies have been conducted with various approaches to the emotional content of speech. These studies show that emotion content in speech has a dynamic nature. The dynamics of speech make it difficult to extract the emotion hidden in a speech. This study aimed to evaluate the implicit emotion in a message through emotional speech processing by applying the Mel-Frequency Cepstral Coefficient (MFCC) and Short-Time Fourier Transform (STFT) features.

Methods: The input data is the Berlin Emotional Speech Database consisting of seven emotional states, anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutral version. MATLAB software is used to input audio files of the database. Next, the MFCC and STFT features are extracted. Feature vectors for each method are calculated based on seven statistical values, i.e. minimum, maximum, mean, standard deviation, median, skewness, and kurtosis. Then, they are used as an input to an Artificial Neural Network. Finally, the recognition of emotional states is done by training functions based on different algorithms.

Results: The results revealed that the average and accuracy of emotional states recognized using STFT features are better and more robust than MFCC features. Also, emotional states of anger and sadness have a higher rate of recognition, among other emotions.

Conclusion: STFT features showed to be better than MFCC features to extract implicit emotion in speech.

Citation: Ravanbakhsh M, Setayeshi S, Pedram M, Mirzaei A. Evaluation of implicit emotion in the message through emotional speech processing based on Mel-Frequency Cepstral Coefficient and Short-Time Fourier Transform features. Advances in Cognitive Sciences. 2020;22(2):71-81.

مجله تازه‌های علوم شناختی
دوره ۲۲، شماره ۲، تابستان ۱۳۹۹
صفحات ۸۱-۷۱
مقاله پژوهشی

ارزیابی هیجان ضمن پیام از طریق پردازش گفتار هیجانی مبتنی بر استفاده از ویژگی‌های MFCC و STFT

مهسا روانبخش[۱]، سعید ستایشی[۲*]، میر محسن پدرام[۳]، آزاده میرزائی[۴]

. دانشجوی دکتری زبان‌شناسی شناختی، موسسه آموزش عالی علومشناختی، تهران، ایران
۲. دانشیار گروه مهندسی هسته‌ای، دانشکده فیزیک و انرژی، دانشگاه صنعتی امیرکبیر، تهران، ایران
۳. دانشیار گروه مهندسی الکترونیک و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه خوارزمی، تهران، ایران
۴. استادیار گروه زبان‌شناسی، دانشکده ادبیات و زبان‌های خارجی، دانشگاه علامه طباطبائی، تهران، ایران

دریافت: ۱۳۹۸/۰۲/۰۸
اصلاح نهایی: ۱۳۹۸/۰۹/۱۰
پذیرش: ۱۳۹۸/۰۹/۱۹

واژه‌های کلیدی
گفتار هیجانی
بازشناسی هیجان
تبدیل فوریه کوتاه مدت
ضرایب کپسترال فرکانس مل
پیردازش گفتار هیجانی

نویسنده مسئول
سعید ستایشی، دانشیار گروه مهندسی هسته‌ای، دانشکده فیزیک و انرژی، دانشگاه صنعتی امیرکبیر، تهران، ایران
ایمیل: Setayesh@aut.ac.ir

doi.org/10.30699/icss.22.2.71

چکیده

مقدمه: گفتار مؤثرترین ابزاری است که انسان‌ها برای انتقال اطلاعات از آن استفاده می‌کنند. گوینده در خلال گفتار خویش علاوه بر واژگان و دستور زبان اطلاعاتی همچون سن، جنسیت و حالت هیجانی خود را منتقل می‌کند. پژوهش‌های فراوانی با رویکردهای گوناگون پیرامون هیجان در گفتار هیجانی انجام شده است. این پژوهش‌ها نشان می‌دهند که هیجان ضمن پیام در گفتار هیجانی از طبیعتی پویا برخودار می‌باشد. این پویایی، مطالعه کمّی هیجان در گفتار هیجانی را با دشواری همراه می‌سازد. این پژوهش به ارزیابی هیجان ضمن پیام از طریق پردازش گفتار هیجانی با استفاده از ویژگی‌های ضرایب کپسترال فرکانس مِل (MFCC) و تبدیل فوریه زمان کوتاه (STFT) پرداخت.

روش کار: داده‌های ورودی، پایگاه‌داده استاندارد گفتار هیجانی Berlin شامل هفت حالت هیجانی خشم، کسلی، انزجار، ترس، شادی، غم و حالت خنثی می‌باشد. با استفاده از نرم افزار MATLAB ابتدا فایل‌های صوتی خوانده شدند. در مرحله بعد نخست ویژگی‌های MFCC و سپس ویژگی‌های STFT استخراج شدند. بردارهای ویژگی برای هر کدام از ویژگی‌ها بر اساس هفت مقدار آماری کمینه، بیشینه، میانگین، انحراف معیار، میانه، چولگی و کشیدگی محاسبه شدند و به عنوان ورودی شبکه عصبی مصنوعی مورد استفاده قرار گرفتند. در انتها، بازشناسی حالت‌های هیجانی با استفاده از توابع آموزشی مبتنی بر الگوریتم‌های مختلف انجام شد.

یافته‌ها: نتایج بدست آمده نشان داد میانگین و صحت بازشناسی حالت‌های هیجانی با استفاده از ویژگی‌های STFT نسبت به ویژگی‌های MFCC بهتر است. همچنین، حالت‌های هیجانی خشم و غم از نرخ بازشناسی بهتری برخوردار بودند.

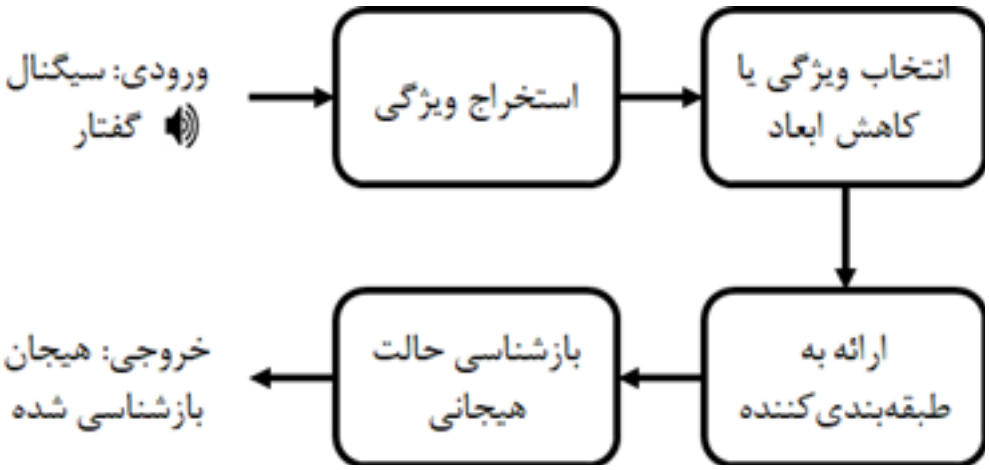
نتیجه‌گیری: ویژگی‌های STFT نسبت به ویژگی‌های MFCC هیجان ضمن پیام در گفتار هیجانی را بهتر بازنمایی می‌کنند.

مقدمه

گفتار ابزاری است که انسان از آن برای ارتباط با دیگران و انتقال اطلاعات استفاده می‌کند. هر گفتار علاوه بر واژگان و دستور زبان خاص هر زبان حاوی اطلاعات دیگری از جمله سن، جنسیت و حالت هیجانی گوینده آن نیز می‌باشد. مطالعه کمّی حالت‌های هیجانی ضمن پیام در گفتار هیجانی بسیار دشوار است. برای غلبه به این دشواری در گام نخست لازم است ویژگی‌های مناسبی از سیگنال گفتار هیجانی استخراج شود که بتوانند به نحو شایانی حالت‌های هیجانی ضمن پیام در گفتار را بازنمایی کنند. البته در گامی‌های بعدی استفاده از طبقه‌بندی‌کننده‌ای که بتواند به درستی حالت‌های هیجانی را بازشناسی کند از اهمیت ویژه‌ای برخوردار است. بازشناسی حالت‌های هیجانی ضمن پیام در گفتار هیجانی با عنوان

بازشناسی هیجان گفتار (Emotion Speech Recognition) شناخته می‌شود. با بررسی پژوهش‌های انجام گرفته در زمینه بازشناسی هیجان گفتار، شاهد انجام یک فرآیند مشابه در تمامی این پژوهش‌ها هستیم. شکل ۱ مراحل مشترک انجام بازشناسی حالت‌های هیجانی در گفتار هیجانی را نشان می‌دهد.

همان‌طور که در تصویر مشاهده می‌شود، ابتدا از نمونه‌های سیگنال گفتار هیجانی ویژگی‌هایی استخراج می‌گردد. در مرحله بعد با توجه به این که ویژگی‌های استخراج شده از ابعاد بالایی برخوردار هستند، با استفاده از روش‌های محاسباتی، ویژگی‌هایی از میان تمامی ویژگی‌های استخراج شده انتخاب می‌شوند تا هر نمونه در قالب یک بردار بازنمایی شود. این بردارها به عنوان بردارهای ورودی برای طبقه‌بندی‌کننده استفاده می‌شوند. در گام آخر طبقه‌بندی‌کننده، بردارهای ورودی را دسته‌بندی می‌کند و به بازشناسی حالت‌های هیجانی می‌پردازد. خروجی طبقه‌بندی‌کننده، حالت‌های هیجانی بازشناسی شده هستند.



شکل ۱. مراحل بازشناسی هیجان گفتار

دو رویکرد قالب در بازشناسی هیجان گفتار وجود دارد. در رویکرد نخست موضوع مورد بررسی استخراج و ارائه ویژگی‌هایی از سیگنال صوتی گفتار است که به بهترین وجه بتواند حالت‌های هیجانی را بازنمایی کند. در پژوهش Hasrul و همکاران نشان داده شد که برای بررسی حالت‌های هیجانی، ویژگی‌های گام (Pitch)، انرژی، شدت، فرمانت، دیرش، ضرایب کپسترال فرکانس مل (Mel-Frequency Cepstral Coefficients (MFCC)) و مشتقات آن بیش از سایر ویژگی‌ها مورد استفاده قرار گرفته‌اند (۱). در این میان MFCC بیش از سایر ویژگی‌های برای بازشناسی هیجان گفتار مورد استفاده قرار گرفته‌اند (۱). اما در رویکرد دوم موضوع مورد بررسی انواع طبقه‌بندی‌کننده‌ها، طراحی و بهینه‌سازی آنها می‌باشد. در ادامه چند نوع از طبقه‌بندی‌کننده‌های به کار رفته در چند پژوهش را معرفی می‌کنیم. Demircan و Kahramanli با استفاده از الگوریتم نزدیکترین همسایگی k بازشناسی حالت هیجانی در گفتار هیجانی را انجام دادند (۲). در پژوهشی دیگر Nwe و همکاران یک شیوه مستقل از متن برای دسته‌بندی هیجان گفتار بر اساس Hidden Markov Model به عنوان دسته‌بندی‌کننده ارائه کردند (۳). Lalitha و همکاران (۴) و همچنین Nicholson و همکاران (۵) از شبکه‌های عصبی مصنوعی برای بازشناسی و طبقه‌بندی حالت‌های هیجانی گفتار استفاده کردند. Franti و همکاران شبکه‌های عصبی کانولوشن را به عنوان طبقه‌بندی‌کننده مورد استفاده قرار دادند (۶). علاوه بر این روش‌ها، از شیوه‌های دیگری همچون مدل ترکیبی Gaussian (۷)، الگوریتم درخت تصمیم‌گیری (۸، ۹)، ماشین بردار پشتیبان (۱۰، ۱۱)، شبکه‌های بیزی (۱۲) و مدارهای سلسه مراتبی (۱۳، ۱۴) برای بازشناسی و طبقه‌بندی حالت‌های هیجانی گفتار استفاده شده است. همچنین از روش‌های نوینی چون مدل‌های شناختی از جمله مدل (Brain Emotional Learning (BEL)) برای بازشناسی حالات هیجانی گفتار استفاده می‌شود (۱۷-۱۵). همچنین پژوهش‌های فراوان دیگری برای درک پیچیدگی‌های زیستی سیستم شنوایی انجام گرفته‌اند؛ Golipour و Gazor یک مدل برای سیستم شنوایی ارائه کردند که پیچیدگی‌های زیستی حرکت موج در حلزون گوش برای تمامی سیگنال‌های ورودی به گوش از جمله گفتار را در نظر می‌گرفت (۱۸). آنها برای بررسی تغییرات فرکانس‌ها در طی زمان و اثر آنها بر رفتار بخش‌های مختلف گوش از ویژگی‌های تبدیل فوریه زمان کوتاه (Short Time Fourier Transform (STFT)) استفاده کردند (۱۸).

در این مطالعه قصد داریم تا با استفاده از یک شبکه عصبی مصنوعی پیشخور با توابع آموزشی مبتنی بر الگوریتم‌های مختلف بررسی کنیم که کدامیک از ویژگی‌های MFCC یا STFT قابلیت بازنمایی حالت‌های هیجانی را بهتر انجام می‌دهند و کمک می‌کنند که حالت‌های هیجانی با نرخ بهتری بازشناسی شوند.

در بخش دوم روش ارزیابی هیجان ضمن پیام در گفتار هیجانی را با استفاده از شبکه عصبی مصنوعی مبتنی بر ویژگی‌های MFCC و STFT را شرح می‌دهیم. ابتدا نحوه محاسبه ویژگی‌های MFCC و STFT را توضیح می‌دهیم. سپس معماری شبکه عصبی مصنوعی مورد استفاده به عنوان طبقه‌بندی‌کننده به همراه تابع آموزش در شبکه‌های عصبی مصنوعی شرح داده خواهند شد. در ادامه به معرفی پایگاه‌داده گفتار هیجانی برلین می‌پردازیم. پس از آن، در مورد بردار ویژگی بازنمایی‌کننده گفتارهای هیجانی صحبت می‌کنیم. در ادامه نتایج بدست آمده از شبیه‌سازی نرم‌افزاری ارائه می‌شود و در انتها به نتیجه‌گیری می‌پردازیم.

روش کار

محاسبه و استخراج ویژگی‌های آکوستیکی گفتار

ویژگی‌های گوناگونی را می‌توان از گفتار هیجانی استخراج کرد. در این مقاله قابلیت ویژگی‌های MFCC و STFT را در بازنمایی حالت هیجانی ضمن پیام در گفتار هیجانی بررسی می‌کنیم. بنابراین در این بخش به ترتیب شیوه محاسبه ویژگی‌های MFCC و STFT را شرح خواهیم داد.

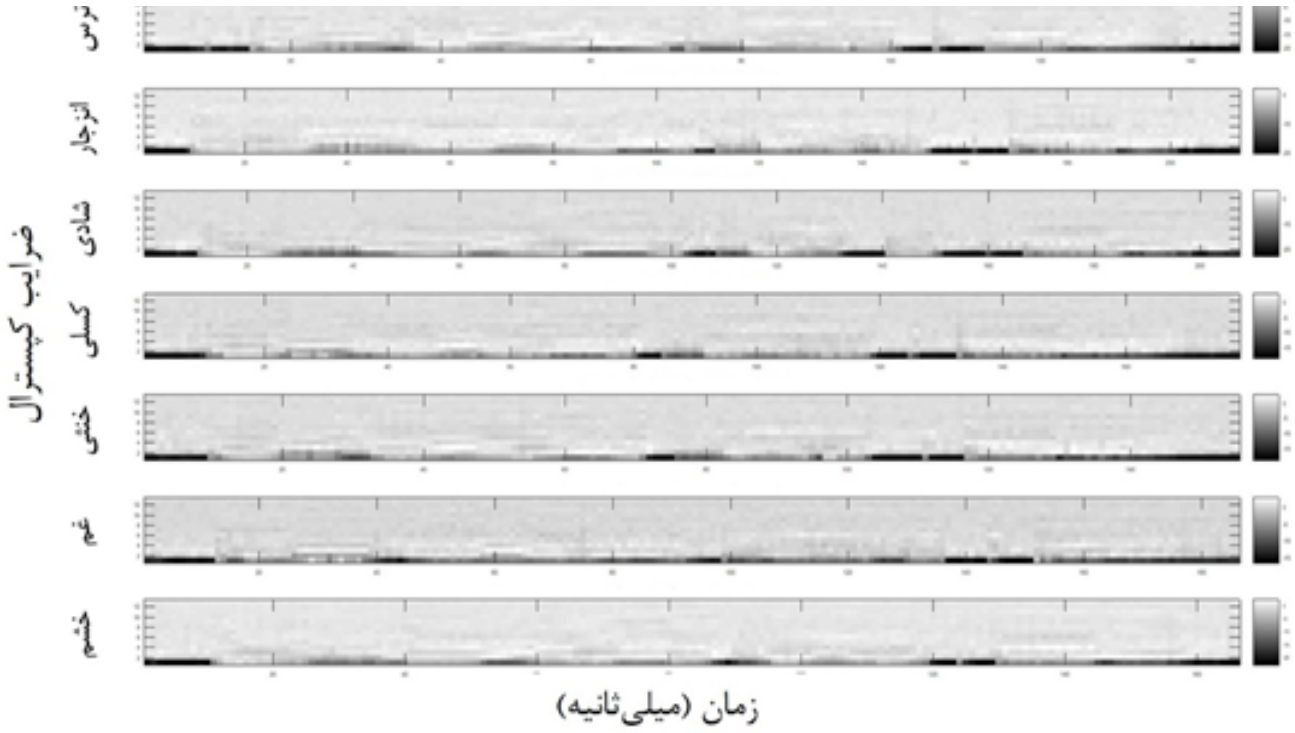
ضرایب کپسترال فرکانس مل (MFCC)

MFCC ویژگی‌هایی هستند که از خواص شنیداری گوش انسان در دریافت و فهم گفتار الهام گرفته شده‌اند (۱۹). برای به دست آوردن ویژگی‌های MFCC نخست باید طیف فوریه پنجره (فریم) با استفاده از تبدیل فوریه سریع بدست آید و دامنه آن محاسبه شود. سپس روی طیف بدست آمده به صورت لگاریتمی و بر اساس معادله (۱) بانک فیلتر اعمال می‌شود و خروجی فیلتر محاسبه می‌شود.

$$F_{mel} = ۲۵۹۳log_{۱۰}\left[۱ + \frac{F_{Hz}}{۷۰۰}\right] \quad (۱)$$

در مرحله بعد، با استفاده از این مقادیر خروجی و نیز استفاده از معادله (۲) ضرایب MFCC بدست می‌آیند. در این معادله F تعداد فیلترها، x_j خروجی به دست آمده از فیلتر jام، c_i ضرایب MFCC بدست آمده و N تعداد ضرایب MFCC می‌باشند. شکل ۲ ضرایب MFCC محاسبه شده برای یک پاره گفتار که توسط یک گوینده با هفت حالت هیجانی بیان شده است را ارائه می‌کند. در این مقاله ضرایب MFCC با استفاده از پنجره‌گذاری همینگ با پنجره‌هایی به اندازه ۲۰ میلی‌ثانیه بدست آمده‌اند (۲۰).

$$c_i = \sum_{j=1}^{N} \log(x_j)\, cos\left[\frac{\pi_i(j-۰/۵۵)}{F}\right], \quad ۱ \le i \le F \quad (۲)$$



شکل ۲. ضرایب MFCC محاسبه شده برای یک پاره گفتار که توسط یک گوینده با هفت حالت هیجانی بیان شده است

تبدیل فوریه کوتاه مدت (STFT)

تجزیه طیفی ابزاری است که امکان مطالعه و تحلیل دقیق‌تر سیگنال از جمله سیگنال گفتار را فراهم می‌آورد. تبدیل فوریه امکان استخراج طیف فرکانسی سیگنال و بررسی کلی محتوای فرکانسی سیگنال را فراهم می‌آورد. در معادله (۳)، معادله تبدیل فوریه ارائه می‌شود. اما در بررسی سیگنال‌های غیر ایستا همچون گفتار که محتوای فرکانسی آنها با زمان تغییر می‌کند تنها بررسی فضای فرکانسی کافی نیست و داشتن نمایش دو بعدی در فضای زمان و فرکانس ضروری است. روش STFT امکان داشتن چنین نگاشتی را فراهم می‌کند.

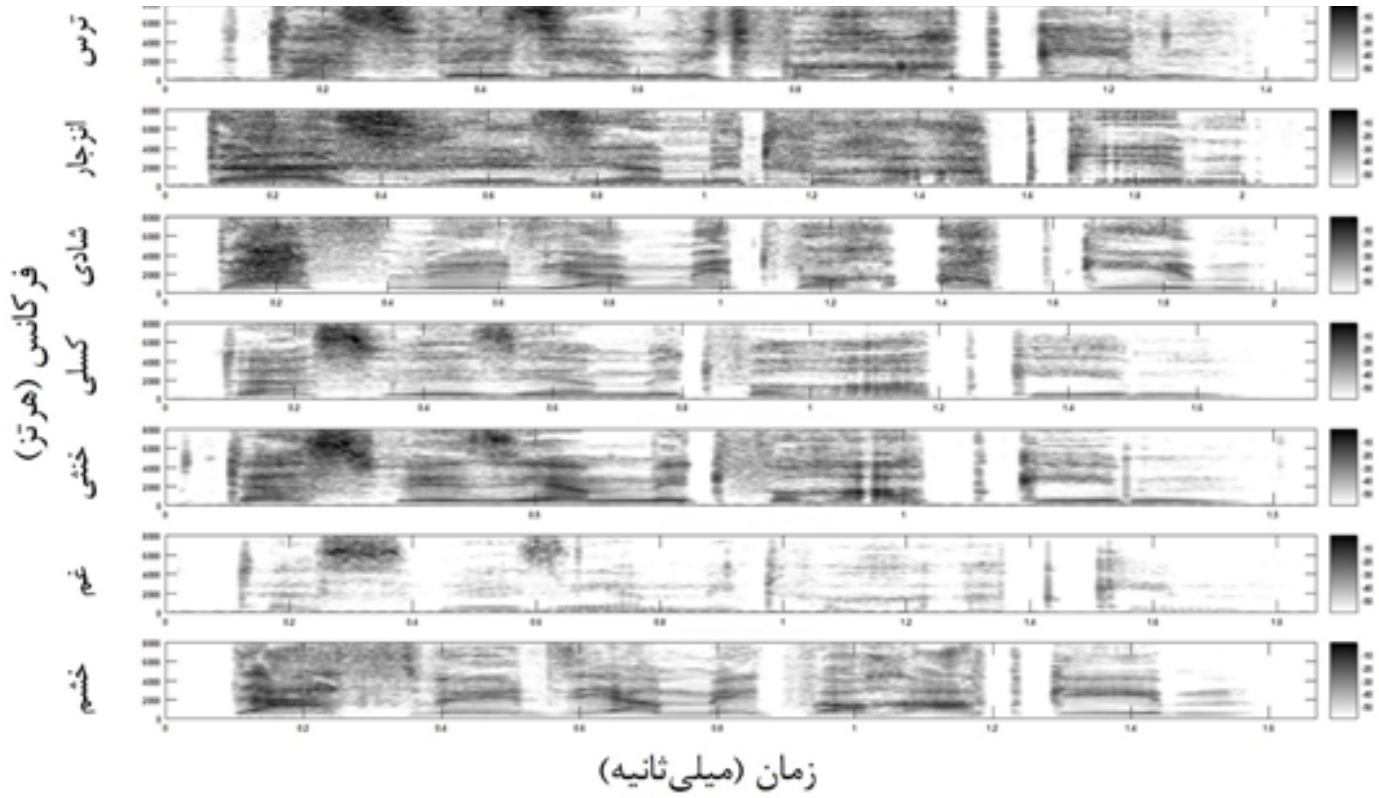
$$F(w) = \int_{-\infty}^{\infty} x(t)e^{-iwt}dt,\ w = 2\pi f \qquad (۳)$$

با استفاده از STFT نگاشت زمان_فرکانس سیگنال بدست می‌آید. برای بدست آوردن STFT یک سیگنال ابتدا سیگنال به پنجره‌های زمانی با طول محدود و معین تقسیم می‌شود و سپس تبدیل فوریه برای هر پنجره به طور مجزا محاسبه می‌شود. به این صورت STFT امکان تحلیل سیگنال روی پنجره‌های زمانی خاصی را فراهم می‌آورند (۲۱، ۲۲). معادله (۴)، معادله STFT را ارائه می‌کند.

$$STFT(\tau, f) = \int x(t)g(t-\tau)e^{-iwt}dt,\ w = 2\pi f \qquad (۴)$$

انتخاب اندازه پنجره مناسب برای محاسبه‌ی STFT یک سیگنال از اهمیت ویژه‌ای برخوردار است. با توجه به اصل عدم قطعیت، انتخاب پنجره‌های باریک باعث داشتن رزلوشن زمانی خوب و رزلوشن فرکانسی پایین می‌شود، در حالی که انتخاب پنجره‌های پهن باعث داشتن رزلوشن زمانی پایین و رزلوشن فرکانسی خوب می‌شود. همچنین با توجه به نوع سیگنال از پنجره‌های متفاوتی استفاده می‌شود.

در این مطالعه مقادیر STFT گفتارهای هیجانی با استفاده از توابع ارائه شده در (۲۳) با پنجره‌بندی همینگ و با طول ۱۰۲۴ محاسبه شده است. در شکل ۳ مقادیر STFT برای هفت گفتار هیجانی با هفت حالت هیجانی متفاوت ارائه شده است. در این گفتارها گوینده و پاره‌گفتار بیان شده یکسان هستند و تنها نوع هیجان تغییر می‌کند.



شکل ۳. مقادیر STFT برای هفت گفتار هیجانی با هفت حالت هیجانی متفاوت که گوینده و پاره‌گفتار بیان شده در آنها یکسان

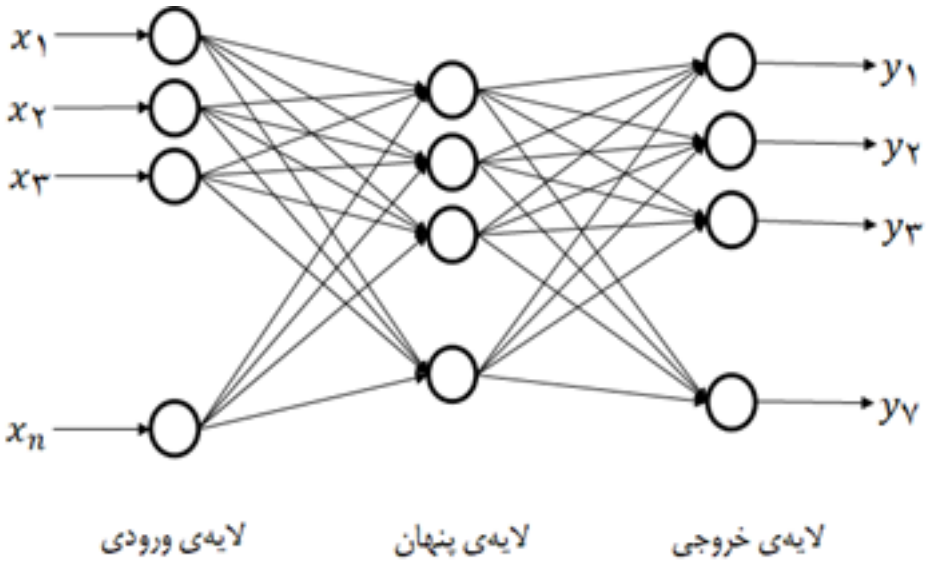
ساختار شبکه عصبی مصنوعی طبقه‌بندی‌کننده

شبکه‌های عصبی مصنوعی از ساختار و سیستم عصبی در موجودات زنده الهام گرفته شده‌اند. شبکه‌های عصبی مصنوعی از اتصال واحدهای پردازشگری به نام نورون در پیکربندی‌های گوناگون ساخته می‌شوند که

در حقیقت مدلی محاسباتی از سلول‌های عصبی یا نورون‌ها در سیستم‌های عصبی موجودات زنده می‌باشند. شبکه عصبی استفاده شده در این مطالعه به عنوان طبقه‌بندی‌کننده، یک شبکه عصبی پیش‌خور (Feedforward) با یک لایه پنهان است. برای تعیین تعداد نورون‌های لایه پنهان هفت حالت پایه هیجانی خشم، کسلی، انزجار، ترس، شادی، غم و حالت خنثی را با استفاده از ویژگی‌های STFT و تعداد مختلف نورون در لایه پنهان مورد بازشناسی قرار دادیم. در تمامی بازشناسی‌های انجام شده ۷۰ درصد داده‌ها برای فاز آموزش، ۱۵ درصد برای فاز تأیید و ۱۵ درصد برای فاز آزمون مورد استفاده قرار گرفتند. همچنین در تمام بازشناسی‌ها از تابع آموزش مبتنی بر الگوریتم پس‌انتشار گرادیان مزدوج مدرج (Scaled conjugate gradient backpropagation) در نرم‌افزار MATLAB به کار گرفته شد. نتایج حاصل از بازشناسی‌ها در جدول ۱ نشان داده شده‌اند. با توجه به نتایج ارائه شده در جدول ۱، تعداد ۲۵ نورون در لایه پنهان قرار داده شد. شکل ۴ نمایی کلی از شبکه‌ی عصبی مورد استفاده ارائه می‌دهد. تعداد نورون‌های لایه خروجی برابر هستند با تعداد حالت‌های پایه هیجانی مورد بازشناسی که در اینجا این تعداد ۷ عدد می‌باشد.

جدول ۱. نتایج حاصل از بازشناسی هفت حالت پایه هیجانی توسط شبکه عصبی پیشخور با تعداد نورون‌های مختلف در لایه پنهان (درصد)

مورد	تعداد نورون‌ها	آموزش	تأیید	آزمون	در مجموع
۱	۱۰	۹۴/۴	۷۰	۶۸/۸	۸۶/۹
۲	۱۵	۹۵/۵	۷۱/۳	۷۵	۸۸/۸
۳	۲۰	۹۸/۴	۸۰	۷۲/۵	۹۱/۸
۴	۲۵	۹۳/۶	۷۳/۸	۷۶/۳	۸۸
۵	۳۰	۱۰۰	۷۵	۷۶/۳	۹۲/۷
۶	۳۵	۹۸/۱	۷۳/۸	۷۵	۹۱
۷	۴۰	۸۸/۸	۷۱/۳	۷۱/۳	۸۳/۶



شکل ۴. نمایی کلی از پیکربندی شبکه عصبی طبقه‌بندی‌کننده

آموزش شبکه‌های عصبی

نورون‌ها یا واحدهای سازنده شبکه‌های عصبی مصنوعی به صورت موازی عمل می‌کنند. نورون‌ها به صورت گسترده‌ای با یکدیگر دارای

اتصالاتی هستند. این اتصالات عملکرد شبکه عصبی را مشخص می‌کنند. همچنین با تنظیم مقادیر (ارزش‌های) این اتصالات می‌توان شبکه عصبی را آموزش داد تا عملکرد ویژه‌ای را انجام دهد. آموزش به طور خلاصه عبارت است از فرآیندی که وزن‌های بهینه شبکه عصبی تعیین می‌شود (۲۴). از توابع آموزشی مبتنی بر الگوریتم‌های مختلفی همچون گرادیان کاهشی (Gradient Descent) برای تنظیم وزن‌های شبکه عصبی استفاده می‌شود. این توابع آموزشی عبارتند از: پس‌انتشار ارتجاعی (Resilient Backpropagation)، گرادیان مزدوج مدرج (Scaled Conjugate Gradient)، گرادیان مزدوج با بازآغازی‌های پاول/بیل (Conjugate Gradient with Powell/Beale Restarts)، گرادیان مزدوج فلچر_پاول (Fletcher-Powell Conjugate Gradient)، گرادیان مزدوج پلاک_ریبیر (Polak-Ribiére Conjugate Gradient)، تک مرحله‌ای متقاطع (One Step Secant) و گرادیان کاهشی (Gradient Descent).

پایگاه‌داده گفتار هیجانی

برای دستیابی به اهداف پژوهشی خود در این مطالعه نیازمند استفاده از مجموعه دادگانی هستیم که متشکل از گفتارهای هیجانی باشد. همچنین هر نمونه با نام حالت هیجانی مشخص برچسب‌گذاری شده باشد. به این منظور در اینجا از پایگاه‌داده گفتار هیجانی Berlin استفاده شد (۲۵). این مجموعه دادگان شامل ۵۳۵ نمونه گفتار هیجانی است و دارای نمونه‌هایی برای هفت حالت پایه هیجانی خشم، کسلی، انزجار، ترس، شادی، غم و حالت خنثی می‌باشد. ۱۰ گوینده شامل ۵ زن و ۵ مرد، ۱۰ پاره گفتار مختلف را با توجه به این هفت حالت پایه هیجانی ذکر شده بیان کردند و نمونه‌های گفتار هیجانی این پایگاه‌داده را تولید کردند. تولیدکنندگان پایگاه‌داده گفتار هیجانی Berlin از میان تمامی نمونه‌های تولید شده آنهایی را که دارای نرخ بازشناسی بیش از ۸۰ درصد و طبیعی ۶۰ درصد بودند را انتخاب کردند و در قالب این مجموعه دادگان ارائه دادند.

هفت حالت پایه هیجانی خشم، کسلی، انزجار، ترس، شادی، غم و حالت خنثی نشان‌دهنده طبقه‌هایی هستند که می‌خواهیم گفتارهای هیجانی بازشناسی شده را بر اساس آنها طبقه‌بندی کنیم. بردار خروجی شبکه عصبی مورد استفاده بر اساس این هفت حالت پایه هیجانی مشخص می‌شود. جدول ۲ بردار خروجی برای هر حالت پایه هیجانی را نشان می‌دهد.

جدول ۲. بردارهای خروجی نشان‌دهنده هر یک از حالت‌های پایه هیجانی

مورد	حالت هیجانی	برچسب
۱	خشم	۱۰۰۰۰۰۰
۲	کسلی	۰۱۰۰۰۰۰
۳	انزجار	۰۰۱۰۰۰۰
۴	ترس	۰۰۰۱۰۰۰
۵	شادی	۰۰۰۰۱۰۰
۶	غم	۰۰۰۰۰۱۰
۷	حالت خنثی	۰۰۰۰۰۰۱

انتخاب ویژگی‌ها و تشکیل بردار ویژگی

داده‌های بدست آمده MFCC و STFT برای ارائه به شبکه عصبی مناسب نیستند و دارای ابعاد بزرگی هستند. برای ساختن بردار ویژگی‌ها از میان ویژگی‌های استخراج شده برخی مقادیر آماری شامل کمینه، بیشینه، میانگین، انحراف معیار، میانه، چولگی و کشیدگی (۲) محاسبه شدند. در اینجا ۱۳ ضریب نخست MFCC را داریم و هفت مقدار آماری ذکر شده را برای هر ضریب محاسبه شد و به این ترتیب بردار ویژگی برای هر نمونه گفتار هیجانی را تولید شد. شکل ۵ چگونگی کاهش اندازه برای ۱۳ ضریب نخست MFCC را نشان می‌دهد.

روند مشابهی برای ویژگی‌های STFT استخراج شده از نمونه‌های گفتار هیجانی تکرار شد. نخست ویژگی‌های STFT برای نمونه‌های گفتار هیجانی محاسبه شد. سپس دنباله با نرخ یک چهارم زیر نمونه‌برداری (Downsampling یا Decimation) شد. در ادامه برای تولید بردار

ویژگی هفت مقدار آماری شامل کمینه، بیشینه، میانگین، انحراف معیار، میانه، چولگی و کشیدگی را برای هر کدام از زیرنمونه‌ها محاسبه گردید. شکل ۶ روند کاهش اندازه بردار ویژگی برای مقادیر STFT را نشان می‌دهد.

شکل ۵. روند کاهش اندازه بردار ویژگی برای ۱۳ ضریب نخست MFC

شکل ۶. روند کاهش اندازه بردار ویژگی برای مقادیر STFT

یافته‌ها

نتایج بدست آمده از پیاده‌سازی نرم‌افزاری با استفاده از نرم‌افزار MATLAB R2016a در جدول‌های ۳ و ۴ ارائه شده است. بردار ویژگی برای هر یک از نمونه‌های گفتار هیجانی در پایگاه‌داده گفتار هیجانی Berlin محاسبه می‌شود. در مرحله نخست، بر اساس ویژگی‌های MFCC، هر نمونه در قالب یک بردار ویژگی شامل ۹۱ عضو بدست می‌آید. از مجموع ۵۳۵ نمونه گفتار هیجانی موجود ۷۰ درصد برای آموزش، ۱۵ درصد برای تأیید و ۱۵ درصد برای آزمون شبکه عصبی مصنوعی پیشخور استفاده شد. نتایج حاصل از پیاده‌سازی نرم‌افزاری مبتنی بر ویژگی‌های MFCC در جدول ۳ قابل مشاهده است. هر سطر بیان‌کننده نتایج بازشناسی حالت‌های هیجانی ضمن پیام در گفتار هیجانی می‌باشد. این نتایج از بازشناسی توسط شبکه عصبی با توابع آموزشی بر اساس الگوریتم‌های مختلف بدست آمده‌اند. همان‌گونه که در جدول ۳ مشاهده می‌شود به ترتیب حالت‌های هیجانی خشم و غم نسبت به سایر حالت‌های هیجانی از نرخ بازشناسی بهتری برخوردار هستند و حالت‌های هیجانی انزجار و شادی با نرخ پایین‌تری بازشناسی می‌شوند.

جدول ۳. نتایج بدست آمده از بازشناسی حالت‌های هیجانی بر اساس ویژگی‌های MFCC (درصد)

الگوریتم تابع آموزش	خشم	کسلی	انزجار	ترس	شادی	غم	حالت خنثی	در مجموع
پس‌انتشار ارتجاعی	۸۳/۵	۶۹/۱	۳۰/۴	۶۲/۳	۲۶/۸	۷۵/۸	۴۶/۸	۶۰/۲
گرادیان مزدوج مدرج	۸۱/۱	۵۹/۳	۲۶/۱	۴۹/۳	۳۸	۶۷/۷	۴۸/۱	۵۶/۸
گرادیان مزدوج با بازآغازی‌های پاول/بیل	۸۹	۴۹/۴	۳۴/۸	۴۷/۸	۳۳/۸	۷۷/۴	۵۷	۵۹/۶
گرادیان مزدوج فلچر_ پاول	۸۵	۵۶/۸	۲/۲	۴۶/۴	۱۲/۴	۶۷/۷	۵۷	۵۲/۹

الگوریتم تابع آموزش	خشم	کسلی	انزجار	ترس	شادی	غم	حالت خنثی	در مجموع
گرادیان مزدوج پلاک_ ربیبر	۸۵	۶۶/۷	۵۶/۵	۵۹/۴	۳۳/۸	۷۷/۴	۴۳	۶۲/۶
تک مرحله‌ای متقاطع	۷۸	۶۰/۵	۲۸/۳	۵۳/۶	۲۱/۱	۶۴/۵	۵۵/۷	۵۵/۵
گرادیان کاهشی	۳۸/۶	۱۶	۱۰/۹	۲۱/۷	۱۹/۷	۶۲/۹	۲۶/۶	۲۹/۲

در مرحله دوم، بردار ویژگی هر نمونه گفتار هیجانی مبتنی بر ویژگی‌های STFT بدست آمد. هر بردار ویژگی در این مرحله شامل ۹۰۳ عضو بود. در این مرحله نیز از مجموع ۵۳۵ نمونه گفتار هیجانی ۷۰ درصد برای آموزش، ۱۵ درصد برای تأیید و ۱۵ درصد برای آزمون شبکه عصبی مصنوعی پیشخور استفاده شد. در جدول ۴، نتایج حاصل از بازشناسی حالت‌های مختلف هیجانی توسط شبکه عصبی مصنوعی پیشخور با استفاده از توابع آموزشی بر اساس الگوریتم‌های مختلف را ارائه شده است. با توجه به نتایج ارائه شده در جدول ۴ حالت‌های هیجانی خشم، غم و کسلی با نرخ بهتری بازشناسی شده‌اند. هر چند در پاره‌ای از موارد حالت هیجانی غم به خوبی بازشناسی نشده است.

جدول ۴. نتایج بدست آمده از بازشناسی حالت‌های هیجانی بر اساس ویژگی‌های STFT (درصد)

الگوریتم تابع آموزش	خشم	کسلی	انزجار	ترس	شادی	غم	حالت خنثی	در مجموع
پس‌انتشار ارتجاعی	۹۸/۴	۴۵/۷	۰	۱/۴	۴/۲	۰	۶/۳	۳۲
گرادیان مزدوج مدرج	۹۲/۹	۵۸	۳۰/۴	۳۷/۷	۳۲/۴	۹۱/۹	۴۴/۳	۵۹/۸
گرادیان مزدوج با بازآغازی‌های پاول/بیل	۹۵/۳	۷۶/۵	۶۰/۹	۶۰/۹	۴۹/۳	۷۷/۴	۶۷/۱	۷۲/۷
گرادیان مزدوج فلچر_ پاول	۹۷/۶	۷۶/۵	۸۷	۷۵/۴	۷۴/۶	۹۳/۵	۸۸/۵	۸۵/۸
گرادیان مزدوج پلاک_ ربیبر	۹۴/۵	۷۹	۶۷/۴	۷۹/۷	۴۹/۳	۹۶/۸	۸۲/۳	۸۰/۴
تک مرحله‌ای متقاطع	۹۴/۵	۸۲/۷	۸۷	۷۹/۷	۶۷/۶	۹۵/۲	۷۵/۹	۸۳/۹
گرادیان کاهشی	۸۹/۸	۶۹/۱	۲۸/۳	۵۶/۵	۴۹/۳	۹۶/۸	۵۷	۶۷/۷

بحث

گفتار مؤثرترین ابزاری است که انسان‌ها از آن برای انتقال اطلاعات استفاده می‌کنند. هر گفتار علاوه بر واژگان و دستور زبان حاوی اطلاعات فراوان دیگری همچون سن، جنسیت و حالت هیجانی گوینده آن نیز می‌باشد. هیجان ضمن پیام که گوینده در خلال گفتار خود آن را انتقال می‌دهد از طبیعتی پویا برخوردار است. پژوهش‌های مختلفی پیرامون بازشناسی هیجان و حالت‌های هیجانی در گفتار هیجانی با رویکردهای گوناگون انجام شده است. نتایج بدست آمده از این پژوهش‌ها نشان می‌دهد که مطالعه کمّی هیجان ضمن پیام در گفتار هیجانی بسیار دشوار است. پژوهش‌های انجام شده پیرامون بازشناسی هیجان گفتار با دو رویکرد قالب انجام می‌گیرند (۱). رویکرد نخست به بررسی ویژگی‌هایی در سیگنال صوتی گفتار می‌پردازند که قادرند تا به بهترین وجه حالات هیجانی را بازنمایی کنند. در حالی که در رویکرد دوم موضوع مورد بررسی انواع طبقه‌بندی‌کننده‌ها، طراحی و بهینه‌سازی آنها می‌باشد. در اکثر پژوهش‌های انجام شده MFCC و مشتقات آن بیش از سایر ویژگی‌ها برای بازنمایی حالات هیجانی مورد استفاده قرار می‌گیرند (۱، ۲). این موضوع نشان‌دهنده توانمندی این ویژگی‌ها برای بازنمایی حالات هیجانی می‌باشد. همگام با این پژوهش‌ها، شاهد مطالعات دیگری هستیم که در تلاش هستند که پیچیدگی‌های زیستی سیستم شنوایی را مورد بررسی قرار دهند و این سیستم را مدل‌سازی کنند. در پژوهش انجام شده Golipour و Gazor سیستم شنوایی با استفاده از STFT مدل‌سازی شد (۱۸).

در این بررسی، بازشناسی حالت‌های هیجان ضمن پیام در گفتار هیجانی

مبتنی بر ویژگی‌های MFCC و STFT مورد ارزیابی قرار دادیم. برای انجام این ارزیابی از نمونه‌های گفتار هیجانی در پایگاه داده گفتار هیجانی Berlin استفاده کردیم که شامل نمونه‌های گفتار هیجانی برای هفت حالت پایه هیجانی خشم، کسلی، انزجار، ترس، شادی، غم و حالت خنثی می‌باشد (۲۵). در مرحله نخست بردار ویژگی گفتارهای هیجانی بر اساس ویژگی‌های MFCC بدست آمد. سپس توسط شبکه عصبی مصنوعی پیشخور و با استفاده از توابع آموزشی با الگوریتم‌های مختلف، بازشناسی حالت‌های هیجانی مختلف در گفتار هیجانی انجام شد. مرحله دوم مشابه با مرحله نخست بود، با این تفاوت که بردار ویژگی گفتارهای هیجانی بر اساس ویژگی‌های STFT می‌باشد. نتایج بدست آمده در مرحله نخست نشان داد که به ترتیب حالت‌های هیجانی خشم و غم نسبت به سایر حالت‌های هیجانی با استفاده از ویژگی‌های MFCC با درصد بهتری بازشناسی می‌شوند و حالت‌های هیجانی انزجار و شادی با درصد پایین‌تری بازشناسی شدند. از نتایج بدست آمده در مرحله دوم مشاهده می‌شود که بر اساس ویژگی‌های STFT، حالت‌های هیجانی خشم، غم و کسلی با نرخ بهتری بازشناسی می‌شوند. البته در پاره‌ای از موارد حالت هیجانی غم به خوبی بازشناسی نمی‌شود. در مجموع از مقایسه نتایج مرحله اول و دوم، بازشناسی درست هیجان و حالت‌های هیجانی ضمن پیام در گفتار هیجانی با استفاده از ویژگی‌های STFT از نرخ بهتری برخوردار است. هر چند بردار ویژگی بدست آمده از ویژگی‌های STFT به مراتب دارای اعضای بیشتری از بردار ویژگی بدست آمده از ویژگی‌های MFCC است، با توجه به طبیعت پویای هیجان ضمن پیام در گفتار هیجانی، ویژگی‌های STFT بهتر از ویژگی‌های MFCC الگوی‌های هیجانی را بازنمایی می‌کنند و برای بازشناسی تعداد بیشتری از هیجانات و حالت‌های هیجانی ضمن پیام در گفتار هیجانی کارآمدتر می‌باشند.

نتیجه‌گیری

گفتار مؤثرترین شیوه ارتباطی است. هر گفتار شامل اطلاعات فراوانی است همچون حالت هیجانی گوینده آن گفتار. این حالت هیجانی که در خلال گفتار بیان می‌شود و از طبیعتی پویا برخوردار است. در این مقاله قابلیت بازنمایی حالات هیجانی با استفاده از MFCC و STFT مورد ارزیابی قرار گرفتند. در پژوهش‌های بسیاری از MFCC برای بازنمایی حالات هیجانی استفاده شده است. اما نوآوری این مطالعه استفاده از STFT برای بازنمایی حالات هیجانی در گفتار هیجانی است. نتایج بدست آمده از بازشناسی هیجان گفتار با استفاده از MFCC و STFT نشان دادند که ویژگی‌های STFT همچون ضرایب MFCC ویژگی‌های توانمندی برای بازنمایی حالات هیجانی هستند و در برخی موارد STFT دارای عملکرد بهتری از MFCC است.

تشکر و قدردانی

این مقاله حاصل از رساله دکتری علوم‌شناختی-زبان‌شناسی در مؤسسه آموزش عالی علوم شناختی می‌باشد.

References

1. Hasrul MN, Hariharan M, Yaacob S. Human affective (Emotion) behaviour analysis using speech signals: A review. In International Conference on Biomedical Engineering 2012 (ICoBE). 2012 Feb 27-28; Penang, Malaysia. pp. 217–22.

2. Demircan S, Kahramanlı H. Feature extraction from speech data for emotion recognition. *Journal of Advances in Computer Networks*. 2014;2(1):28-30.

3. Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. *Speech Communication*. 2003;41(4):603-623.

4. Lalitha S, Geyasruti D, Narayanan R, Shravani M. Emotion detection using MFCC and cepstrum features. Procedia Computer Science. 2015;70:29-35.

5. Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks. *Neural Computing & Applications*. 2000;9(4):290-296.

6. Franti E, Ispas I, Dragomir V, Dascalu M, Zoltan E, Stoica IC. Voice based emotion recognition with convolutional neural networks for companion robots. *Science and Technology*. 2017;20(3):222-240.

7. Ververidis D, Kotropoulos C. Emotional speech classification using Gaussian mixture models. In 2005 IEEE International Symposium on Circuits and Systems. 2005 May 23-26; Kobe, Japan. pp. 2871-2874. IEEE.

8. Cichosz J, Slot K. Emotion recognition in speech signal using emotion-extracting binary decision trees. In Doctoral

Consortium. *Proceeding Affective Computer Intelligent Interaction*. Lisbon:Springer;2007.

9. Hua A, Litman DJ, Forbes-Riley K, Rotaru M, Tetreault J, Purandare A. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. InNinth International Conference on Spoken Language Processing. 2006 Sep 17-21; Pittsburgh, PA, USA. pp. 797–800.

10. Wu C, Chuang Z. Emotion recognition from speech using ig-based feature compensation. *International Journal of Computational Linguistics & Chinese Language Processing (Special Issue on Affective Speech Processing)*. 2007;12(1):65–78.

11. Hoch S, Althoff F, McGlaun G, Rigoll G. Bimodal fusion of emotional data in an automotive environment. In Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. 2005 Mar. 23; Philadelphia, PA, USA. pp. 1085–1088.

12. Fersini E, Messina E, Archetti F. Emotional states in judicial courtrooms: An experimental investigation. *Speech Communication*. 2012;54(1):11-22.

13. Albornoz EM, Milone DH, Rufiner HL. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*. 2011;25(3):556-570.

14. Lee CC, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*. 2011;53(9-10):1162-1171.

15. Motamed S, Setayeshi S, Rabiee A. Speech emotion recognition based on a modified brain emotional learning model. *Biologically Inspired Cognitive Architectures*. 2017;19:32-38.

16. Motamed S, Setayeshi S, Rabiee A. Speech emotion recognition based on brain and mind emotional learning model. *Journal of Integrative Neuroscience*. 2018;17(3-4):577-591.

17. Motamed S, Setayeshi S, Farhoudi Z, Ahmadi A. speech emotion recognition based on learning automata in Fuzzy Petri-net. *Journal of Mathematics and Computer Science*. 2014;12(3):173-185.

18. Golipour L, Gazor S. A biophysical model of the human cochlea for speech stimulus using STFT. In Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005. 2005 Dec 21; Athens, Greece. pp. 46–50.

19. Ayat S. Fundamentals of speech signal processing, Tehran:Payame Noor University Press;2008. (Persian)

20. Sahidullah M, Kinnunen T, Hanilçi C. A comparison of features for synthetic speech detection. Conference of the International Speech Communication Association (INTERSPEECH) 2015. 2015 Sep 6-10; Dresden, Germany. pp. 2087–2091.

21. Mallat S. A wavelet tour of signal processing the sparse way. 3rd ed. Burlington:Academic Press;2009.

22. Mertins A. Signal analysis: Wavelets, filter banks, time-frequency transforms and applications. West Sussex, England:John Wiley & Sons;1999.

23. Wojcicki K. Speech Spectrogram. MATLAB Central File Exchange; 2020 [updated 23 July 2020 cited 23 April 2019]. https://www.mathworks.com/matlabcentral/fileexchange/29596-speech-spectrogram.

24. Forouzanfar M, Dajani HR, Groza VZ, Bolic M, Rajan S. Comparison of Feed-Forward Neural Network training algorithms for oscillometric blood pressure estimation. In 4th International Workshop on Soft Computing Applications 2010 Jul. 15-17; Arad, Romania. pp. 119–123.

25. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In 6th Interspeech 2005 and 9th European Conference on Speech Communication and Technology 2005. 2005 Sep. 4-8; Lisbon, Portugal. pp. 1517–1520.