Advances in Cognitive Sciences
Vol 21, No. 4, Winter 2020
Research Article
Pages 113-127

Emotion recognition based on multimodal fusion using mixture of brain emotional learning

Zeinab Farhoudi[1], Saeed Setayeshi[2*], Farbod Razazi[3], Azam Rabiee[4]

1. PhD Student of Artificial Intelligence, Department of Computer Engineering, Science and Reserach Branch, Islamic Azad University, Tehran, Iran
2. Professor of Department of Energy Engineering and Physics, Amirkabir University of Technology, Tehran, Iran
3. Professor of Department of Electrical and Computer Engineering, Science and Reserach Branch, Islamic Azad University, Tehran, Iran
4. Professor of Department of Computer Science, Dolatabad Branch, Islamic Azad University, Isfahan, Iran

Recieved: 6 Mar. 2019
Revised: 6 Sep. 2019
Accepted: 16 Sep. 2019

Keywords
Multimodal emotion recognition
Brain emotional learning
Mixture of neural networks
Fusion
Deep learning

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

Email: Setayesh@aut.ac.ir

doi.org/10.30699/icss.21.4.113

Abstract

Introduction: Multimodal emotion recognition due to receiving information from different sensory resources (modalities) from a video has a lot of challenges and has attracted many researchers as a new method of human computer interaction. The purpose of this paper was to automatically recognize emotion from emotional speech and facial expression based on the neural mechanisms of the brain. Therefore, based on studies on brain-inspired models, a general framework for bimodal emotion recognition inspired by the functionality of the auditory and visual cortics and brain limbic system is presented.

Methods: The hybrid and hierarchical proposed model consisted of two learning phases. The first step: the deep learning models for the representation of visual and auditory features, and the second step: a Mixture of Brain Emotional Learning (MoBEL) model, obtained from the previous stage, for fusion of audio-visual information. For visual feature representation, 3D-convolutional neural network (3D-CNN) was used to learn the spatial relationship between pixels and the temporal relationship between the video frames. Also, for audio feature representation, the speech signal was first converted to the log Mel-spectrogram image and then fed to the CNN. Finally, the information obtained from the two above streams was given to the MoBEL neural network model to improve the efficiency of the emotional recognition system by considering the correlation between visual and auditory and fusion of information at the feature level.

Results: The accuracy rate of emotion recognition in video in the eNterface'05 database using the proposed method was on average of 82%.

Conclusion: The experimental results in the database show that the performance of the proposed method is better than the hand-crafted feature extraction methods and other fusion models in the emotion recognition.

Citation: Farhoudi Z, Setayeshi S, Razazi F, Rabiee A. Emotion recognition based on multimodal fusion using mixture of brain emotional learning. Advances in Cognitive Sciences. 2020;21(4):113-127.

مجله تازه‌های علوم شناختی
دوره ۲۱، شماره ۴، زمستان ۱۳۹۸
صفحات ۱۲۷-۱۱۳
مقاله پژوهشی

بازشناسی هیجان مبتنی بر همجوشی اطلاعات چندوجهی با استفاده از مدل ترکیبی یادگیری هیجانی مغز

زینب فرهودی[۱]، سعید ستایشی[۲*]، فربد رزازی[۳]، اعظم ربیعی[۴]

۱. دانشجوی دکتری هوش مصنوعی، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران
۲. استاد گروه مهندسی پرتوپزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران
۳. استاد گروه الکترونیک و مهندسی رایانه، دانشگاه علوم و تحقیقات آزاد اسلامی، تهران، ایران
۴. استاد دانشکده مهندسی رایانه، دانشگاه آزاد اسلامی واحد دولت آباد، اصفهان، ایران

دریافت: ۱۳۹۷/۱۲/۱۵
اصلاح نهایی: ۱۳۹۸/۰۶/۱۵
پذیرش: ۱۳۹۸/۰۶/۲۵

واژه‌های کلیدی
بازشناسی هیجان چندوجهی
یادگیری هیجانی مغز
مدل ترکیب شبکه‌های عصبی همجوشی
یادگیری عمیق

نویسنده مسئول
سعید ستایشی، استاد گروه مهندسی پرتوپزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران

ایمیل: Setayesh@aut.ac.ir

doi.org/10.30699/icss.21.4.113

چکیده

مقدمه: بازشناسی هیجان چندوجهی به واسطه دریافت اطلاعات از منابع حسی (وجه‌های) مختلف از یک ویدیو دارای چالش‌های فراوانی است و به عنوان روش جدیدی برای تعامل طبیعی انسان با رایانه مورد توجه محققان زیادی قرار گرفته است. هدف از این پژوهش، بازشناسی هیجان به طور خودکار از روی گفتار هیجانی و حالات چهره، مبتنی بر ساز و کارهای عصبی مغز بود. بنابراین، با توجه به مطالعات صورت گرفته در زمینه مدل‌های الهام گرفته از مغز، یک چارچوب کلی برای بازشناسی هیجان دومدالیتی با الهام از عملکرد کورتکس شنوایی و بینایی و سیستم لیمبیک مغز ارائه شود.

روش کار: مدل ترکیبی و سلسله مراتبی پیشنهادی از دو مرحله یادگیری تشکیل شده بود. مرحله اول: مدل‌های یادگیری عمیق برای بازنمایی ویژگی‌های بینایی و شنوایی و مرحله دوم: مدل ترکیبی یادگیری هیجانی مغز (MoBEL) بدست آمده از مرحله قبل برای همجوشی اطلاعات شنیداری_دیداری. برای بازنمایی ویژگی‌های بینایی به منظور یادگیری ارتباط مکانی بین پیکسل‌ها و ارتباط زمانی بین فریم‌های ویدئو از مدل شبکه عصبی یادگیری عمیق 3D-CNN استفاده شد. همچنین به منظور بازنمایی ویژگی‌های شنوایی، ابتدا سیگنال گفتار به تصویر لگاریتم مل_اسپکتروگرام تبدیل شده سپس به مدل یادگیری عمیق CNN برای استخراج ویژگی‌های مکانی_زمانی داده شد. در نهایت، اطلاعات به دست آمده از دو جریان فوق به شبکه عصبی ترکیبی MoBEL داده شد تا با در نظر گرفتن همبستگی بین وجه‌های بینایی و شنوایی و همجوشی اطلاعات در سطح ویژگی، کارایی سیستم بازشناسی هیجان را بهبود بخشد.

یافته‌ها: نرخ بازشناسی هیجان در ویدیو با استفاده از مدل ارائه شده بر روی پایگاه داده eNterface'05 بطور میانگین ۸۲ درصد شد.

نتیجه‌گیری: نتایج تجربی در پایگاه داده مذکور نشان می‌دهد که کارکرد روش پیشنهادی بهتر از روش‌های استخراج ویژگی‌های دستی و سایر مدل‌های همجوشی در بازشناسی هیجان است.

مقدمه

انسان یک سیستم هوشمند نهایی مجهز به سنسورهای چند وجهی است که قادر به پردازش، یادگیری و پاسخ به محرک‌های چندوجهی می‌باشد. به نظر می‌رسد انسان، تناظرات نماهای مختلف را یاد می‌گیرد و از آن به همراه سایر روش‌های ترکیب اطلاعات چندوجهی در سطوح مختلف انتزاع، استفاده می‌کند. این یک رویکرد ایده‌آل برای همجوشی اطلاعات چندوجهی به عنوان نمونه‌ای موفق از طرح‌های مدل‌سازی سلسله مراتبی می‌باشد. با این وجود، پیشرفت مهمی لازم است تا رایانه بتواند در سطح انسان اطلاعات چندوجهی را پردازش کند. در سال‌های اخیر، بازشناسی هیجان چندوجهی به ویژه بازشناسی حالات چهره و هیجان گفتار در بسیاری از کاربردها از جمله: تعامل انسان و رایانه (۱)،

بازی‌های رایانه‌ای (۲)، یادگیری الکترونیکی (۳)، نظارت بر سلامت انسان (۴) مود توجه محققان زیادی قرار گرفته است. با این وجود بازشناسی هیجان صوتی_تصویری هنوز یک چالش در زمینه بینایی ماشین و یادگیری ماشین است. اول اینکه، بازنمایی ویژگی‌های حالات چهره و هیجان گفتار به دلیل تنوع ظهور هیجان در هر فرد، زاویه و شدت روشنایی چهره مشکل است. دوم اینکه، نحوه همجوشی وجه‌های مختلف صوتی و تصویری با در نظر گرفتن همزمانی داده‌های ورودی و همبستگی موجود در اطلاعات مکان_زمان در ویدیو در محاسبه دقت نهایی تأثیر می‌گذارد. بنابراین، بازنمایی ویژگی‌ها و همجوشی اطلاعات دو مرحله مهم در بازشناسی هیجان فرآیند دیداری_شنیداری می‌باشند. برای استخراج ویژگی‌های حالات چهره در تصاویر ثابت، مطالعات زیادی با استفاده از روش‌های کلاسیک و استخراج ویژگی دستی انجام شده است که به طور کلی به دو دسته «مبتنی بر هندسه» و «مبتنی بر ظاهر» تقسیم می‌شوند (۵، ۶). ویژگی‌های هندسی، مؤلفه‌های شکل و مکانی چهره مانند چشم‌ها، ابروها، لب و غیره را بازنمایی می‌کند و ویژگی ظاهری، بافت چهره مانند چین و چروک، برآمدگی و گودی را نمایش می‌دهند. ویژگی‌های هندسی بر اساس نتایج ترازبندی مؤلفه‌های چهره از طریق روش مدل ظاهر فعال (Active Appearance Model (AAM)) بدست می‌آیند (۵). Chang و همکاران از مدل شکلی که به وسیله ۵۸ نقطه برجسته در چهره تعریف شده است استفاده کردند (۷). ویژگی‌های ظاهری نیز از طریق روش‌هایی مانند موجک گابور (Gabor Wavelet) (۶)، ویژگی Haar (۸) و غیره بدست می‌آید. در بسیاری از مطالعات استفاده از ویژگی‌های ظاهری و هندسی، بهترین گزینه برای طراحی بازشناسی حالات چهره می‌باشد (۹). در این میان، بازنمایی (Local Binary Patterns (LBP)) (۱۰) و (Local Phase Quantization (LPQ)) (۱۱) دو بازنمایی از روش‌های استخراج ویژگی مبتنی بر ظاهر می‌باشند. در سال‌های اخیر، استخراج ویژگی با استفاده از شبکه‌های عصبی کانولوشن CNN برای بازنمایی حالات چهره در تصاویر ثابت و پویا مورد استفاده قرار گرفته است (۱۲). برای دنباله‌ای از تصاویر پویا، بازنمایی ویژگی‌های حرکات ماهیچه در طول زمان استفاده می‌شود (۱۳). به طور مثال در مراجع برای استخراج ویژگی‌های چهره از مدل‌های یادگیری عمیق 3D-CNN (سه بعدی) و CNN-RNN با در نظر گرفتن توالی مکان_زمان بروز هیجان در چهره در ویدیو استفاده می‌شود (۱۴، ۱۵).

بازشناسی هیجان از روی گفتار یکی از موضوعات چالش‌برانگیز در زمینه پردازش گفتار به شمار می‌آید. به طور کلی استخراج ویژگی در سیستم بازشناسی هیجان از روی گفتار به دو دسته ویژگی‌های عروضی و طیفی تقسیم می‌شوند. ویژگی‌های عروضی (Prosodic) (یا به عبارتی مشخصه‌های گفتار پیوسته) شامل فرکانس گام (پیچ)، انرژی و نرخ عبور از صفر (۱۶، ۱۷) و ویژگی‌های طیفی (Spectral) شامل فرمنت‌ها، ضرایب (Mel-frequency Cepstral Coefficient (MFCC)) و (Perceptual Linear Prediction (PLP)) می‌باشند (۱۸). در بسیاری از روش‌ها از ترکیب ویژگی‌های عروضی و طیفی به همراه طبقه‌بندهای سنتی مانند (Hidden Markov Model (HMM)) و SVM استفاده می‌شود (۱۹، ۲۰). Zeng و همکارانش از روش استخراج ویژگی‌های عروضی و طبقه‌بند (Multi-stream HMM (MS-HMM)) برای بازشناسی هیجان گفتار در ویدیو استفاده کرده است (۲۱). Nta-lampiras و همکاران، ویژگی‌ها فرکانس گام، دامنه موجک و ویژگی MFCC را استخراج کرده و روش همجوشی به نام F-HMM را بر روی ویژگی‌ها اعمال کرده‌اند (۲۲). اخیراً در بازشناسی هیجان گفتار روش‌های یادگیری عمیق بکار رفته‌اند که به نتایج خوبی دست یافته‌اند. به طور مثال، Zhang و همکاران، ابتدا ویدیو را به کلیپ‌هایی با مقدار همپوشانی مشخص تقسیم کرده و سپس در هر کلیپ برای استخراج ویژگی‌ها روش CNN را بر روی تصویر مل_اسپکتروگرام سه کاناله اعمال کردند (۲۳).

پس از مرحله بازنمایی ویژگی‌ها، همجوشی چندمدالیتی برای ادغام ویژگی‌ها یا اطلاعات وجه‌های مختلف بینایی و شنوایی برای بازشناسی هیجان در ویدیو استفاده می‌شود. علی‌رغم مزیت‌های همجوشی چندوجهی، چالش‌های زیر در بازشناسی هیجان وجود دارد: ۱) سیگنال‌های بینایی و شنوایی از نظر زمانی هم‌زمان نیستند ۲) ساخت مدلی که از اطلاعات مکمل و نه از اطلاعات تکمیلی استفاده کند مشکل است ۳) هر مدالیتی انواع مختلف و سطوح مختلفی از نویز را در زمان‌های مختلف نشان می‌دهد. همجوشی اطلاعات دارای طیف گسترده‌ای از برنامه‌های کاربردی می‌باشد که عبارتند از: بازشناسی گفتار صوتی_تصویری (۲۴)، بازشناسی هیجان چندمدالیتی (۲۵)، تحلیل تصاویر پزشکی (۲۶) و بازشناسی رویدادهای چندمدالیتی (۲۷). بررسی‌های زیادی نیز در زمینه تحلیل چندمدالیتی، بازیابی اطلاعات و بازشناسی هیجان انجام شده است (۲۷). مطالعات انجام شده در زمینه همجوشی وجه‌های مختلف برای حل چالش‌های مذکور به ۴ راهبرد همجوشی: همجوشی سطح ویژگی (سطح اول) (۲۸)، همجوشی سطح تصمیم‌گیری (سطح آخر) (۲۹) و همجوشی سطح طبقه‌بند (۳۰) و همجوشی ترکیبی سطح اول و آخر (۳۱) تقسیم شده‌اند. در مطالعات اخیر، خط مرز بین بازنمایی چندمدالیتی و همجوشی آنها برای مدل‌هایی مانند شبکه‌های عصبی عمیق نامشخص می‌باشد. در این شبکه‌ها، یادگیری بازنمای ویژگی‌ها با طبقه‌بندی یا رگرسیون اشیاء

همخوانی دارد (۳۲). اگرچه این روش‌ها کارایی بالایی را در همجوشی اطلاعات چندوجهی برای بازشناسی هیجان نشان داده‌اند اما نتوانسته‌اند یک ارتباط غیرخطی و مکمل موجود بین ویژگی‌های بینایی و شنوایی در سطوح بالا بدست آورند. بنابراین نیازمند طراحی سیستمی هستیم که بتواند با درنظر گرفتن همبستگی غیرخطی بین وجه‌های مختلف، همجوشی را در سطح بالای ویژگی انجام دهد. از آنجایی که در مغز انسان، سیستم لیمبیک مسئول پاسخ به محرک‌های هیجانی است، طراحی ماشینی که با الهام از ساز و کارهای عصبی مغز بتواند کار بازشناسی هیجان را با دقت بالا انجام دهد ایده خوبی خواهد بود. از این رو، ما برای انجام این کار یک مدل نو به نام اختلاط خبره‌ها مبتنی بر یادگیری هیجانی مغز ((MoBEL) Mixture of Brain Emotional Learning) را پیشنهاد کرده‌ایم. جامع‌ترین مدلی که الهام گرفته از سیستم لیمبیک مغز پستانداران است مدل BEL ارائه شده توسط مورن وبالکینیوس می‌باشد (۳۳). اخیراً، مدل‌های توسعه‌یافته BEL در برنامه‌های کاربردی کنترلی به نام BELBIC (۳۴)، بازشناسی الگوها (۳۵) و بازشناسی هیجان گفتار (۳۶) به کار رفته است. این مدل‌ها مبتنی بر دو مؤلفه آمیگدالا و اوربیتوفرانتال می‌باشند. آمیگدال و اوربیتوفرانتال (OFC) دو قسمت اصلی از سیستم لیمبیک مغز هستند. در مدل BEL، آمیگدال یاد می‌گیرد که به محرک هیجانی پاسخ دهد و OFC تجربه‌های نامناسب و اتصالات یادگیری را مهار می‌کند. مدل‌های یادگیری می‌بایست ویژگی سرعت بالا و پیچیدگی محاسباتی پایین را دارا باشند که مدل BEL دارای این مزایا می‌باشد. اگر چه مدل اصلی BEL می‌تواند بسیاری از مسائل بازشناسی الگو را حل کند اما نمی‌تواند مسائل غیرخطی و تقریباً درجه n بیتی را حل کند. برای غلبه بر این مشکل، لطفی و همکاران، مدل رقابتی BEL، الهام گرفته شده از مغز را پیشنهاد کرده‌اند (۳۷).

به طور خلاصه، یادگیری ویژگی‌های صوتی_تصویری برای بازشناسی هیجان یکی از گام‌های اصلی در پیدا کردن وابستگی بین ویژگی وجه‌های مختلف می‌باشد. مطالعات قبلی بیشتر بر روی ویژگی‌های دستی تمرکز شده بود که ثابت شده برای بازشناسی هیجان به طور کافی جدایی‌پذیر نیستند. در حالی که، هدف از این پژوهش، یادگیری توأم بازنمایی ویژگی‌های بینایی و شنوایی به طور خودکار با استفاده از مدل جدید برای همجوشی اطلاعات و طبقه‌بندی الگوها به نام MoBEL می‌باشد. این پژوهش، به این صورت سازماندهی شده است: در روش پیشنهادی، بازنمایی ویژگی‌های مکان_زمان گفتار هیجانی و حالات چهره و همچنین یادگیری مدل MoBEL ارائه شد. سپس در قسمت یافته‌ها، مدل پیشنهادی با سایر مدل‌ها بر روی پایگاه داده صوتی_تصویری eN-terface مورد ارزیابی قرار گرفت.

روش کار

از آنجا که ساختار مدل پیشنهادی الهام گرفته شده از مدل فیزیولوژی مغز بود؛ روند بازشناسی هیجان از روی حالات چهره و گفتار هیجانی بر اساس مسیر استخراج ویژگی در قشر بینایی و شنوایی شبیه‌سازی شد. با پذیرش مدل‌های قدرتمند یادگیری عمیق در بازنمایی ویژگی، یک مدل همجوشی MoBEL از طریق همجوشی ویژگی‌های یاد گرفته شده بینایی و شنوایی با استفاده از روش‌های یادگیری عمیق ارائه کرده‌ایم. شکل ۱، ساختار مدل پیشنهادی را نمایش می‌دهد. مدل پیشنهادی از دو مرحله تشکیل شده است: مرحله اول: روش‌های یادگیری عمیق بویژه CNN و شبکه‌های عصبی بازگشتی RNN و شبکه کانولوشن سه بعدی 3D-CNN برای بازنمایی ویژگی‌ها در سطح بالاتر اعمال می‌شوند. مرحله دوم: مدل همجوشی و طبقه‌بندی MoBEL که برای یادگیری مشترک ویژگی‌های یادگرفته شده صوتی_تصویری استفاده می‌شود. مطابق شکل این ساختار به طور جزئی‌تر از مراحل زیر تشکیل شده است:

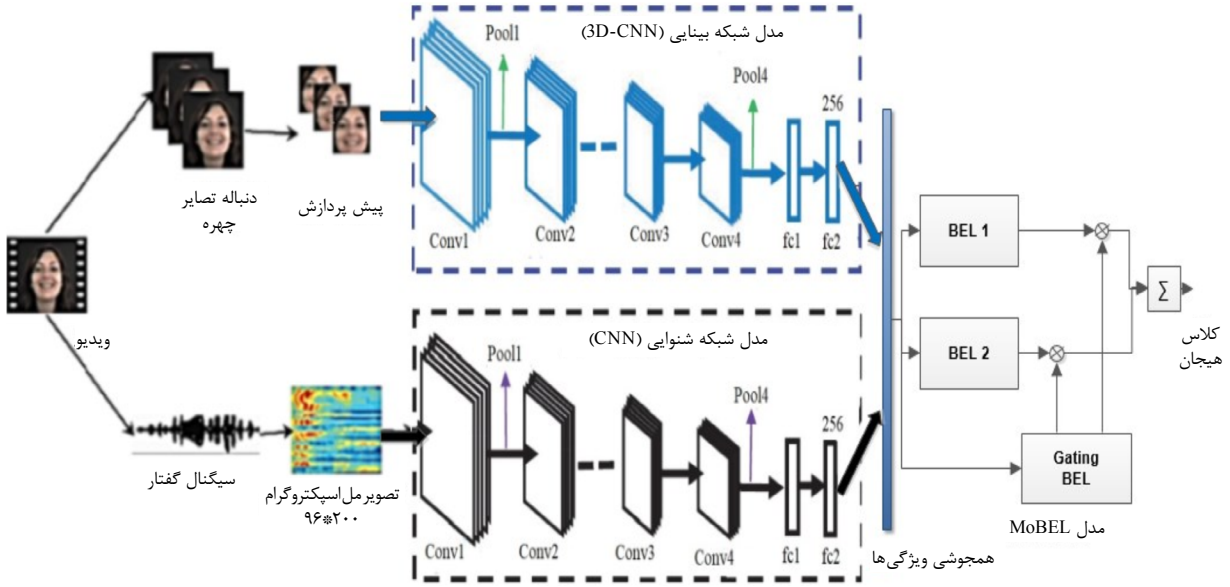
۱. ابتدا هر ویدیو به دو جریان بینایی و شنوایی تقسیم شد و سپس در جریان بینایی با یک نرخ فریم بر ثانیه مشخص، فریم‌های ویدیو استخراج گردید.

۲. در جریان بینایی برای دنباله‌ای از تصاویر ابتدا با استفاده از عملیات شناسایی چهره، تصاویر چهره بدست آمد و پس از انجام پیش‌پردازش‌های لازم دنباله‌ای از تصاویر چهره برای بازنمایی ویژگی‌های مکان_زمان به مدل شبکه عصبی 3D-CNN داده شد.

۳. در جریان شنوایی، ابتدا سیگنال گفتار به تصویر لگاریتم مل_اسپکتروگرام تبدیل گردید سپس شبکه CNN بر روی تصویر بدست آمده اعمال شد.

۴. آخرین لایه تماماً اتصال هر کدام از جریان‌های فوق با هم الحاق شد و به عنوان یک بردار ویژگی جدید به شبکه عصبی ترکیبی MoBEL برای آموزش داده شد. وظیفه این شبکه، یادگیری همبستگی غیرخطی بین ویژگی‌های حالات چهره و هیجان گفتار است. در نهایت خروجی کلاس هیجان در بازشناسی هیجان ویدیو داده شد.

مدل پیشنهادی MoBEL از مدل اختلاط خبره‌های شبکه عصبی الهام گرفته شده است با این تفاوت که در اینجا اختلاط خبره‌ها (Mixture of Experts (MoE))دل BEL است. اختلاط خبره‌ها بر پایه راهبرد تقسیم و حل استوار است. به صورتی که مجموعه آموزش با توجه به میزان شباهت تقسیم‌بندی می‌شود و در طی آموزش، طبقه‌بندهای مختلف سعی می‌کنند تا قسمت‌های مختلفی از فضای ورودی را مدل کنند. یک مکانیزم رقابتی وجود دارد که با استفاده از یک شبکه میانجی، طبقه‌بندها را به صورت محلی ماهر می‌نماید.



شکل ۱. ساختار مدل پیشنهادی. بازشناسی هیجان از روی ویدیو با استفاده از ویژگی‌های بینایی و شنوایی و اعمال مدل MoBEL

شواهدی وجود دارد که نورون‌های رقابتی در مغز از مجموعه‌ای از اتصالات برانگیختگی و مهارکننده تشکیل شده است (۳۸) و همچنین مدل MoE از قشر حافظه انجمنی مغز الهام گرفته شده است که می‌تواند اطلاعات منابع حسی مختلف را ادغام کند.

بازنمایی ویژگی حالات چهره

در این مدل برای بازشناسی حالات چهره در تصاویر ویدئو از شبکه عصبی کانولوشن سه بعدی استفاده شده است. به این صورت که ابتدا ویدئو بر اساس نرخ فریم داده شده به فریم‌هایی تقسیم می‌شود. سپس عملیات پیش پردازش بر روی فریم‌های بدست آمده از ویدئو اعمال می‌شود. این عملیات شامل آشکارسازی چهره، ترازبندی چهره و تغییر اندازه تصویر می‌باشد.

در انتها هر نمونه ویدیو ورودی در قالب یک بردار چهار بعدی شامل دنباله‌ای از تصاویر چهره در فریم‌های هر ویدئو به اندازه ۱۴ فریم متوالی که اندازه هر تصویر ۹۶×۱۰۰ می‌باشد (۳×۱۰۰×۹۶×۱۴) به شبکه عصبی 3D-CNN داده می‌شود. برای آموزش بهتر شبکه سه بعدی 3D-CNN از آموزش اولیه این شبکه برای مقداردهی اولیه وزن‌ها که قبلاً بر روی پایگاه داده بزرگی از تصاویر و ویدیوها آموزش داده شده بود به نام C3D-Sports-1M استفاده شده است. اما برای جلوگیری از Overfitting شبکه و مواجه نشدن با خطای کمبود حافظه، لایه‌های آخر شبکه را حذف کرده‌ایم. به این ترتیب، شبکه 3D-CNN دارای ۸ لایه کانولوشن، ۵ لایه Max-pooling و ۲ لایه تماماً اتصال (FC) می‌باشد که آخرین لایه FC دارای ۲۵۶ نورون و لایه Softmax دارای ۶ نورون به تعداد کلاس‌های هیجان است.

بازنمایی ویژگی هیجان گفتار

همان‌طور که قبل‌تر توضیح داده شد؛ برای بازنمایی ویژگی هیجان گفتار، ابتدا سیگنال گفتار به تصویر لگاریتم مل_اسپکتروگرام تبدیل شد. اسپکتروگرام یک بازنمایی تصویری از شدت سیگنال صوت در طول زمان به ازای فرکانس‌های مختلف می‌باشد. یک تصویر دوبعدی است که در محور افقی زمان و در محور عمودی فرکانس و دامنه ضرایب فرکانس در یک زمان مشخص با شدت رنگ در تصویر مشخص می‌شود. از آنجایی که CNN یک روش قدرتمند و مقاوم در استخراج ویژگی‌های متمایز برای تصاویر و ویدیو می‌باشد، تبدیل سیگنال یک بعدی صوت به تصویر اسپکتروگرام دوبعدی و اعمال روش قدرتمند CNN بر روی آن ایده جالبی می‌تواند باشد. تصویر لگاریتم مل_اسپکتروگرام از طریق اعمال فیلتر بانک‌های MFCC و لگاریتم خروجی بدست می‌آید. در این پژوهش، ابتدا سیگنال‌های صوتی را به طول ۴ ثانیه در نظر گرفته شد. همچنین طول پنجره همینگ ۲۰ میلی‌ثانیه و طول پنجره همپوشانی ۱۰ میلی‌ثانیه در نظر گرفته شد.

توجه شود در پایگاه داده eNterface’05 مورد استفاده، حداکثر طول گفتار ۴ ثانیه است. اگر طول مدت زمان گفتار کمتر از ۴ ثانیه باشد با استفاده از روش Zero padding به اندازه اختلاف زمانی با طول ۴ ثانیه، تعدادی صفر به ابتدای ماتریس گفتار اضافه نمودیم و اگر هم

بیشتر از ۴ ثانیه بود فقط ۴ ثانیه ابتدای گفتار را انتخاب کردیم. سپس فیلتر مل_اسپکتروگرام بر روی هر کدام پنجره‌های همینگ از طریق محاسبه رابطه ۱ اعمال شد:

$$f_{mel}(f) = 2595\log_{10}(1+\frac{f_{HZ}}{700}) \quad (۱)$$

برای هر کدام از سیگنال‌های صوتی پس از اعمال رابطه فوق، ۹۶ ضریب اول فیلتر_مل انتخاب می‌شود. بنابراین اندازه تصویر سیگنال صوتی پس از اعمال لگاریتم مل_اسپکتروگرام برابر است با ۲۰۰×۹۶ که محور عمودی نشان‌دهنده ضرایب فیلتر مل و محور افقی نشان‌دهنده طول زمان که به صورت رابطه زیر محاسبه شد: طول پنجره همینگ/طول مدت زمان گفتار=محور افقی. در مرحله آخر، شبکه CNN برای استخراج ویژگی‌های گفتار بر روی تصاویر بدست آمده از سیگنال‌های صوتی اعمال شد.

مدل پیشنهادی MoBEL

ساختار اختلاط شبکه‌های عصبی که زیرمجموعه اختلاط خبره‌ها است از چند خبره و یک شبکه میانجی تشکیل شده است. شبکه میانجی دو وظیفه دارد یکی اینکه فضای ورودی را به صورت هوشمندانه بین طبقه‌بندها تقسیم کند. دوم اینکه با توجه به توانمندی طبقه‌بند برای طبقه‌بندی صحیح الگوی ورودی یک وزن به آن تخصیص دهد. همزمان با یادگیری طبقه‌بندها، شبکه میانجی یاد می‌گیرد که چگونه وزن مربوط به نظر هر طبقه‌بند را به صورت تابعی از الگوی ورودی محاسبه کند. به عبارتی با دو مدل از یادگیری مواجه هستیم: یکی یادگیری با ناظر در درون هر خبره و دیگری یادگیری بدون ناظر که در شبکه میانجی انجام می‌شود. همان‌طور که گفته شد در ساختار ارائه شده از مدل BEL در خبره‌ها و شبکه میانجی استفاده شد.

در شکل ۲، دو خبره و یک شبکه میانجی نشان داده شده است. فرض کنید که $p=\{p_1, p_2, \ldots, p_n\}$ بردار الگوی ورودی که مطابق شکل، n برابر با ۵۱۲ ویژگی می‌باشد و E_1 و E_2 بردارهای خروجی به ترتیب از خبره ۱ و خبره ۲ و g_1 و g_2 به ترتیب وزن تخصیص داده شده به شبکه BEL1 و شبکه BEL2 توسط شبکه میانجی می‌باشد. g_i تخمینی از احتمال پسین آن است که شبکه iام بتواند الگوی p را به درستی طبقه‌بندی کند. تعداد نورون‌های لایه خروجی شبکه میانجی با تعداد خبره‌ها برابر است. y نیز خروجی نهایی مدل است. خروجی نهایی سیستم y به صورت رابطه ۲ محاسبه شد:

$$y = \sum_i E_i\, g_i\,. \qquad i = 1.\ldots.N \quad (۲)$$

و g_i خروجی شبکه میانجی به صورت زیر تابعی از الگوی ورودی و وزن‌های یادگیری است. (رابطه ۳)

$$g_i = \frac{\exp(O_{gi})}{\sum_{j=1}^{N}\exp(O_{gi})} \quad (۳)$$

با تعریف رابطه فوق برای gi مجموع وزن‌های تخصیص داده شده به هر کدام از طبقه‌بندهای پایه برابر با یک است. اگر اختلاف خروجی هر کدام از طبقه‌بندهای پایه با خروجی مطلوب کمتر باشد یعنی خطای آن کمتر است و در نتیجه وزن g_i بیشتری به آن اختصاص داده می‌شود. در شبکه میانجی از یک مدل BEL (شامل ۲ آمیدلا و ۲ اوربیتوفرانتال به منظور ایجاد دو خروجی) با دو وزن خروجی تشکیل شده است. در این مدل BEL مقدار خروجی مطلوب h_i از طریق رابطه ۴ تعریف می‌شود. این عبارت را می‌توان عملیات Softmax در نظر گرفت.

$$h_i = \frac{g_i \exp(-\frac{1}{2}(t-E_i)^T(t-E_i))}{\sum_j g_i \exp(-\frac{1}{2}(t-E_i)^T(t-E_i)} \quad (۴)$$

که t مقدار خروجی مطلوب و E_i خروجی هر کدام از طبقه‌بندهای پایه است. در اختلاط خبره‌ها با تعریف تابع خطا یا هزینه و اعمال گرادیان نزولی روشی برای آموزش با ناظر در هر کدام از طبقه‌بندهای پایه ارائه شد. رابطه ۵ تابع خطای کل سیستم بود.

$$e = \sum_i g_i \|y - E_i\|^2 \quad (۵)$$

و خروجی هر کدام از طبقه‌بندهای پایه بر اساس رابطه ۶ و ۷ محاسبه شد:

$$E_i = f\left(f_{amg}\left(\sum_{j=1}^{n} v_j\, p_j + v_{n+1}p_{th}\right) - f_{ofc}\left(\sum_{j=1}^{n} w_j\, p_j\right)\right) \quad (۶)$$

$$p_{th} = \max(p_j)\,. \qquad j = 1.\ldots.n \quad (۷)$$

که f تابع فعال‌ساز، v_j وزن‌های AMG و w_j وزن‌های OFC، p_j الگوی ورودی و p_{th} بیشترین مقدار ورودی که به صورت بایاس عمل می‌کند، است. بر اساس روابط بالا و مشتق خطا نسبت به هر کدام از وزن‌های AMG و OFC، اصلاح وزن‌ها در مدل بهبود یافته BEL بر اساس روابط ۸ و ۹ محاسبه شد.

$$\Delta v_{ij} = -\gamma\, v_{ij} + \alpha\, h_i\, max(t - E_i.0)\, p_j\,. \quad for\ i = 1.2\ for\ j = 1.\ldots.n \quad (۸)$$

$$\Delta w_{ij} = \beta\, h_i\, (E_i - t)\, p_j \quad (۹)$$

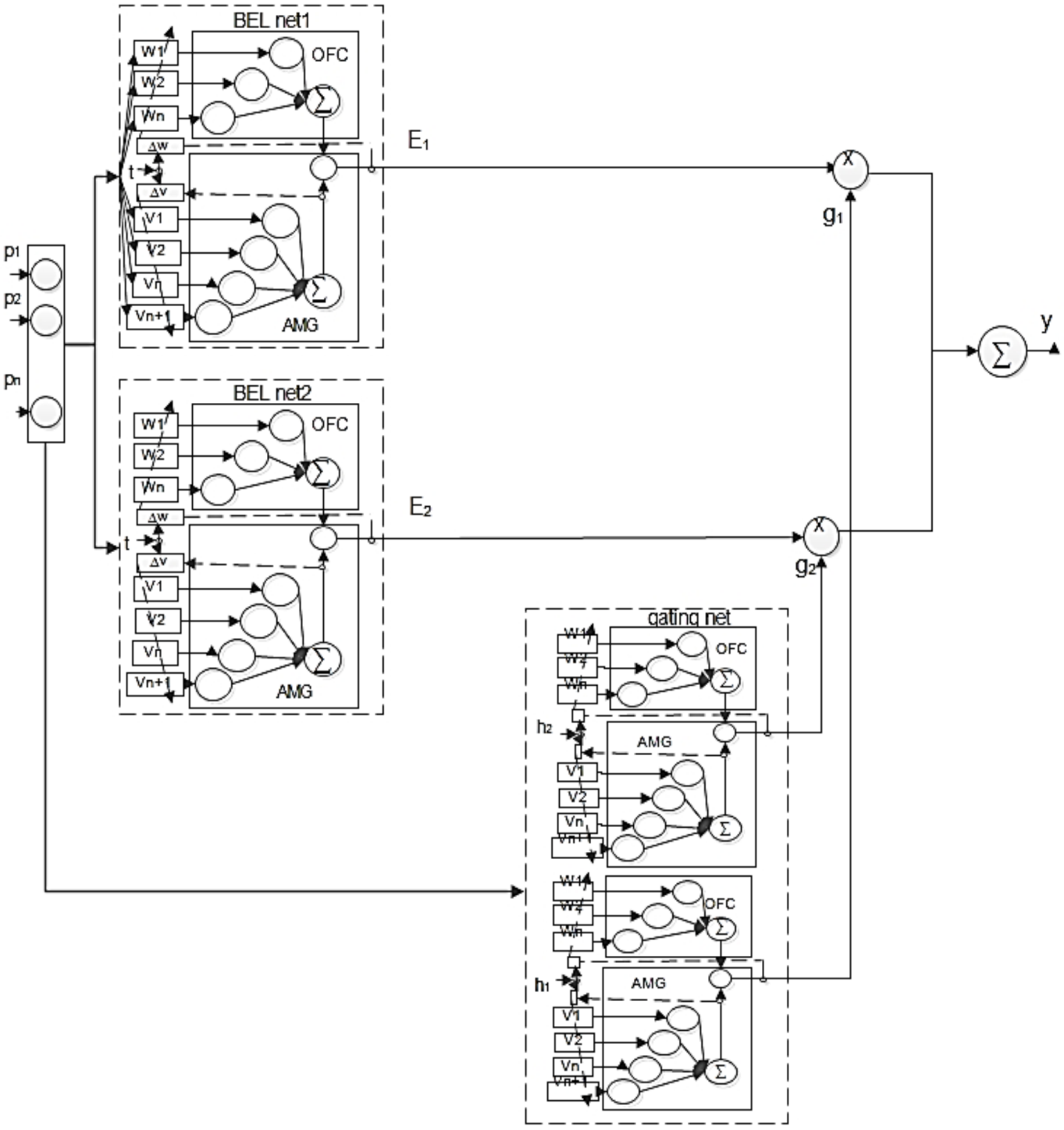
که α نرخ یادگیری AMG، β نرخ یادگیری OFC و γ نرخ تنزل برای مومنتوم بود. همچنین در شبکه میانجی اصلاح وزن‌های AMG و

OFC به صورت روابط ۱۰ و ۱۱ است.

$$\Delta v_{gj} = -\gamma\, v_{gj} + \alpha_g\, max(h_i - g_i.0)\, p_j\, .for\ g = gating\ \ for\ j = 1.\ldots.n \quad (۱۰)$$

$$\Delta w_{gj} = \beta\ \ (g_i - h_i)\, p_j \quad (۱۱)$$

که α_g نرخ یادگیری β، AMG نرخ یادگیری OFC در شبکه میانجی هستند. به این ترتیب در فرآیند آموزش، به ازای هر ورودی خبره‌ها با هم در رقابت هستند و شبکه میانجی بر اساس خطای هر خبره برنده را انتخاب می‌کند.



شکل ۲. ساختار مدل پیشنهادی MoBEL. از دو خبره BEL و یک شبکه میانجی BEL تشکیل شده است

یافته‌ها

به منظور بررسی و ارزیابی مدل پیشنهادی در بازشناسی هیجان صوتی_تصویری، ساختار مدل پیشنهادی را بر روی پایگاه داده صوتی_تصویری eNterface'05 اعمال کردیم. برای ارزیابی کارایی بازشناسی هیجان چندوجهی، ابتدا نتیجه بازشناسی هیجان در هر وجه (صوتی و تصویری) را به طور مستقل نشان دادیم سپس نتایج بازشناسی هیجان پس از اعمال انواع روش‌های همجوشی فرآیند شنوایی_بینایی و مقایسه آنها با مدل پیشنهادی همجوشی را نشان داده شد.

مطابق شکل ۱، مدل پیشنهادی از دو مرحله تشکیل شده بود. جزئیات پیاده‌سازی مدل‌های یادگیری عمیق CNN و 3D-CNN به ترتیب مدل شنوایی و بینایی به صورت زیر بود: اندازه هر دسته (Batch) برابر با ۳۲ بود. از بهینه‌سازی Adam در محاسبه تابع هزینه با نرخ آموزش ۰/۰۰۰۱ و مومنتوم ۰/۰۰۰۰۱ و در این روش از تکنیک توقف زودهنگام در زمان آموزش استفاده شد. تعداد دفعات تکرار در مرحله آموزش شبکه‌ها به ترتیب برای 3D-CNN برابر با ۵۰۰ و CNN برابر با ۴۰۰ تنظیم شده است. پیاده‌سازی مدل با چارچوب Tensorflow و بر روی ماشینی با مشخصات GPU NVIDIA GTX با ۸ گیگ حافظه GPU انجام شد. برای آموزش شبکه عصبی MoBEL، از دو شبکه عصبی خبره بر پایه مدل BEL و یم شبکه میانجی بر پایه BEL استفاده شد. هر شبکه BEL نیز شامل یک مؤلفه آمیگدال و اوربیتوفرانتال است که نرخ آموزش هم در آمیگدال و هم در اوربیتوفرانتال برابر با ۰/۰۰۰۹ می‌باشد. همچنین از گرادیان نزولی با مومنتوم ۰/۰۰۰۱ در الگوریتم پس انتشار خطا در آموزش BEL با تعداد دفعات تکرار ۲۰۰ استفاده شد. تابع فعال‌ساز در شبکه عصبی آمیگدال و اوربیتوفرانتال برابر با tansig و مقدار پاداش در مدل‌های BEL خبره برابر با مقدار خروجی مطلوب و در شبکه میانجی برابر با مقدار خروجی مطلوب h مطابق رابطه است (۱۱). بنابراین شبکه‌های خبره BEL به صورت یادگیری با نظارت و شبکه میانجی BEL به صورت بدون نظارت همزمان آموزش دیدند. آزمایش بازشناسی هیجان بر روی پایگاه داده eNterface'05 با در نظر گرفتن ۷۰ درصد مجموعه داده برای آموزش و ۱۰ درصد مجموعه داده برای اعتبارسنجی و ۲۰ درصد برای آزمایش به صورت اعتبارسنجی_متقاطع (Cross validation) انجام شد و میانگین ۵ بار اجرا گزارش شده است.

نتایج بازشناسی حالات چهره

در این قسمت، نتایج آزمایشات مختلف برای بازشناسی حالات چهره مورد بررسی قرار گرفت. همان‌طور که قبلاً گفته شد، پس از انجام عملیات پیش‌پردازش، دنباله‌ای از تصاویر چهره بدست آمد. به ازای تعدادی دنباله این آزمایش انجام شد و بهترین حالت ۱۴ فریم متوالی بود که هم حافظه کمتری مصرف می‌کرد و هم کارایی بالاتری داشت. نکته قابل توجه این بود که پس از انجام عملیات پیش‌پردازش، ممکن بود تعداد فریم‌های یک نمونه کمتر یا بیشتر از تعداد فریم‌های موردنیاز برای پردازش باشد. اگر کمتر بود ویدیو را با نرخ فریم کمتری تقسیم می‌کردیم تا فریم‌های بیشتری از ویدیو بدست آید و اگر زیاد بود، ضرایبی از فریم‌ها را انتخاب می‌کردیم. پس از اعمال مدل 3D-CNN بر روی دنباله‌ای از تصاویر چهره به منظور بازشناسی حالات چهره و انجام آزمایشات مختلف بر روی پایگاه داده eNterface'05 به طور متوسط دقت بازشناسی هیجان چهره به ازای تمام کلاس‌ها ۶۲ درصد بدست آمد. در جدول ۱، ماتریس درهم‌ریختگی بازشناسی حالات چهره با استفاده از مدل 3D-CNN بر روی پایگاه داده eNterface نشان داده شده است. مطابق جدول ۱ بر روی این پایگاه داده، هیجان «تنفر» و «خوشحالی» در بازشناسی حالات چهره بالاترین دقت بازشناسی و هیجان «عصبانیت» و «تعجب» کمترین دقت بازشناسی را دارند.

به منظور نمایش مزیت مدل پیشنهادی بازشناسی حالات چهره، کارایی این مدل با سایر روش‌هایی که از پایگاه داده eNterface'05 استفاده

جدول ۱. ماتریس درهم‌ریختگی بازشناسی حالات چهره با استفاده از 3D-CNN بر روی پایگاه داده eNterface-05 (درصد)

	عصبانیت	تنفر	ترس	خوشحالی	ناراحتی	تعجب
عصبانیت	**۴۷**	۲	۱۵	۳	۱۰	۵
تنفر	۲	**۸۶**	۵	۰	۵	۲
ترس	۷	۷	**۵۵**	۵	۱۷	۱۰
خوشحالی	۲	۱۲	۵	**۶۷**	۵	۱۰
ناراحتی	۵	۲	۱۴	۲	**۶۷**	۱۰
تعجب	۱۰	۵	۱۷	۱۰	۱۵	۴۴

کرده‌اند مورد مقایسه قرار گرفت. در جدول ۲ نتیجه این مقایسه نشان داده شده است.

همان‌طور که در جدول ۲ مشاهده می‌شود با استفاده از 3D-CNN دقت بازشناسی حالات چهره بسیار بهتر از سایر روش‌های استخراج ویژگی دستی مانند نقاط حالات چهره، ماتریس تصویر کوانتیزه شده (QIM)، LPQ و Gabor Wavelet در (۱۳، ۳۹، ۴۰) بود. علاوه بر این، مدل یادگیری عمیق دیگری برای بازشناسی حالات چهره به نام CNN-RNN ارائه شد و بر روی پایگاه داده موجود مورد بررسی و ارزیابی قرار گرفت. با وجود اینکه در مقایسه با مدل 3D-CNN پارامترهای کمتری داشت و حافظه کمتری مصرف می‌کرد اما دقت بازشناسی آن کمتر بود. در روش CNN-RNN، ابتدا تصاویر حالات چهره ایستا از پایگاده داده fer2013 به شبکه Vgg16 برای پیش آموزش داده شده و وزن‌ها و پارامترهای شبکه ذخیره می‌شدند، سپس فریم‌های تصاویر حالات چهره پایگاه داده eNterface به شبکه عصبی Vgg16 از قبل آموزش یافته داده می‌شدند تا ویژگی‌های هر فریم بدست آید. در انتها ویژگی‌های آخرین لایه تماماً اتصال بدست آمده از هر کدام از فریم‌های یک نمونه ویدئو در ماتریس دو بعدی ذخیره می‌شدند که این ماتریس در مرحله بعد به شبکه LSTM با ۴۰۹۶ نورون لایه ورودی داده می‌شد.

نتایج بازشناسی هیجان گفتار

ابتدا سیگنال گفتار تبدیل به تصویر مل_اسپکتروگرام به اندازه ۹۶×۲۰۰ می‌شود. سپس این تصویر به شبکه CNN برای یادگیری توالی مکان_زمان سیگنال گفتار داده می‌شد. در جدول ۲، ماتریس درهم‌ریختگی مدل ارائه شده برای بازشناسی هیجان گفتار برای ۶ کلاس هیجان پایه بر روی پایگاه داده eNterface نشان داده شده است. به طور متوسط پس از ۵ بار اجرا، میانگین دقت بازشناسی هیجان گفتار با استفاده از مدل CNN، ۶۷/۷ درصد بدست آمد.

مطابق جدول ۳، حالت هیجان «ناراحتی» و «عصبانیت» بالاترین و «تنفر» کمترین دقت بازشناسی هیجان گفتار بر روی پایگاه داده eNterface را داشتند. برای نمایش مزیت مدل پیشنهادی بازشناسی هیجان گفتار، کارایی آن با سایر روش‌ها مورد مقایسه قرار گرفت. نتایج این مقایسه در جدول ۴ آورده شده است. همچنین در این جدول، روش استفاده شده استخراج ویژگی و نوع طبقه‌بندی به کار رفته در مدل ارائه شده در مقالات مختلف به همراه دقت بازشناسی کل نشان داده شده است. علاوه بر این، جدول ۴ به این نکته دلالت می‌کند که عملکرد روش CNN به عنوان استخراج ویژگی بهتر از عملکرد سایر روش‌های استخراج ویژگی دستی مانند استخراج ویژگی‌های عروضی و طیفی گفتار به صورت دستی می‌باشد.

جدول ۲. مقایسه دقت بازشناسی حالات چهره بین روش‌های مختلف و مدل پیشنهادی به ازای ۶ کلاس در پایگاه داده eNterface'05

دقت بازشناسی	بازنمایی ویژگی حالات چهره	مراجع
۳۷ درصد	نقاط برجسته حالات چهره	Mansoorizadeh و همکاران (۱۳)
۳۹/۲۷ درصد	QIM	Bejani و همکاران (۴۰)
۴۲/۱۶ درصد	LPQ	Zhalephpour و همکاران (۳۹)
۵۷/۸ درصد	CNN-RNN	مدل پیشنهادی اول
۶۲ درصد	3D-CNN	مدل پیشنهادی ما

جدول ۳. ماتریس درهم‌ریختگی بازشناسی هیجان گفتار با استفاده از مدل CNN بر روی پایگاه داده eNterface-05 (درصد)

	عصبانیت	تنفر	ترس	خوشحالی	ناراحتی	تعجب
عصبانیت	**۸۱**	۷	۱	۳	۳	۵
تنفر	۱۲	**۵۵**	۹	۷	۷	۱۰
ترس	۲	۶	**۵۹**	۱۴	۱۴	۵
خوشحالی	۱۲	۲	۵	**۶۹**	۲	۱۰
ناراحتی	۵	۵	۱۰	۱	**۷۰**	۱۰
تعجب	۵	۱۰	۲	۱۵	۵	**۶۳**

جدول ۴. مقایسه دقت بازشناسی هیجان گفتار مدل پیشنهادی با سایر روش‌های قبلی بر روی پایگاه داده eNterface’05 و Berlin

پایگاه داده	مراجع	ویژگی‌های صوتی	دقت (درصد)
eNterface’05	Mansoorizadeh و همکاران (۱۳)	عروضی، LDA	۴۳
	Sahoo و همکاران (۴۱)	MFCC	۵۷
	Zhang و همکاران (۴۲)	عروضی+ طیفی	۶۲/۷
	Bejani و همکاران (۴۰)	عروضی + MFCC	۵۴/۹
	Zhalehpour و همکاران (۳۹)	MFCC-RASTA-PLP	۷۲/۹
	مدل پیشنهادی	مل اسپکتروگرام+ CNN	۶۷/۷
Berlin	Badshah و همکاران (۴۳)	اسپکتروگرام + CNN	۶۵/۵
	Mansoorizadeh و همکاران (۴۴)	عروضی	۷۱
	Farhoudi و همکاران (۳۶)	عروضی + MFCC	۶۶
	مدل پیشنهادی	مل اسپکتروگرام+ CNN	۷۴/۵

علاوه بر این، ما مدل CNN را بر روی پایگاه داده Berlin نیز اعمال کرده‌ایم و نتیجه عملکرد این مدل را با نتایج کارهای قبلی مانند استفاده از تبدیل اسپکتروگرام و استخراج ویژگی CNN (۴۳)، استخراج ویژگی‌های عروضی (۴۴)، ترکیب استخراج ویژگی‌های عروضی و ضرایب MFCC و استفاده از مدل BEL برای بازشناسی هیجان گفتار در کار قبلی (۳۶) مورد مقایسه قرار دادیم. مطابق جدول ۴، ثابت شده است که هم کارایی لگاریتم مل_اسپکتروگرام بهتر از اسپکتروگرام است و هم مدل CNN که یک مدل شناخته شده در دسته‌بندی تصاویر است قادر به یادگیری همبستگی زمان_مکان سیگنال گفتار بود و کارایی آن بهتر از سایر روش‌های قبلی است.

نتایج همجوشی اطلاعات

در این پژوهش، عملکرد مدل پیشنهادی همجوشی شبکه عصبی MoBEL با دو روش همجوشی، یعنی همجوشی سطح ویژگی‌ها (اولیه) و همجوشی سطح تصمیم‌گیری (آخر) مورد مقایسه و ارزیابی قرار گرفت.

همجوشی سطح تصمیم‌گیری

در این آزمایش، خروجی نهایی به دست آمده از دو فرآیند شنیداری_دیداری با یکدیگر ترکیب می‌شوند. برای همجوشی در سطح تصمیم‌گیری، انواع روش‌های قواعد ترکیب مورد آزمایش قرار گرفته است که عبارتند از: کمترین، بیشترین، میانگین و حاصل ضرب. نتایج این آزمایش در جدول ۵ نشان داده شده است. مطابق جدول ۵، قانون ترکیب «حاصل ضرب» بالاترین دقت بازشناسی را دارد. زیرا که حاصل ضرب با ضرب امتیاز خروجی وجه‌های مختلف، بیشینه امتیاز کلاس را به دست می‌آورد.

همجوشی سطح بالای ویژگی

در این آزمایش که مدل پیشنهادی و محور اصلی و نوآوری پژوهش بود از مدل بهبود یافته و ترکیبی BEL به نام MoBEL (مطابق شکل ۱) برای همجوشی ویژگی‌های بدست آمده از حالات چهره و هیجان گفتار استفاده شد. به این صورت که در مسیر بینایی، دنباله‌ای از تصاویر

جدول ۵. مقایسه دقت بازشناسی هیجان مبتنی بر همجوشی سطح تصمیم‌گیری بین روش‌های مختلف ترکیب قواعد (درصد)

انواع روش‌های قواعد ترکیب	بیشترین	کمترین	جمع	میانگین	حاصل ضرب
دقت بازشناسی هیجان	۶۸/۶	۷۱	۷۱/۴	۷۱/۴	۷۴/۳

چهره به شبکه 3D-CNN داده می‌شدند و پس از اعمال لایه‌های مختلف، در لایه تماماً اتصال آخر یک بردار ویژگی ۲۵۶ تایی بدست می‌آمد. در مسیر شنوایی نیز تصاویر مل_اسپکتروگرام گفتار با اندازه ۱×۲۰۰×۹۶ به شبکه عصبی CNN داده شده و پس از اعمال لایه‌های مختلف یک بردار ویژگی ۲۵۶ تایی از گفتار بدست می‌آمد. سپس این ویژگی‌ها الحاق شده و یک بردار ۵۱۲ تایی به شبکه عصبی MoBEL داده شد. این شبکه پس از همجوشی ویژگی‌ها، خروجی کلاس نهایی هیجان را می‌دهد.

از آنجا که همجوشی ویژگی‌های بدست آمده از دو جریان شنیداری_دیداری یک مسأله با پیچیدگی بالا بود و ارتباط وجه‌های مختلف به صورت غیرخطی بود باید از یک مدل ترکیبی استفاده شد. به همین دلیل، در این پژوهش از ترکیب شبکه‌های عصبی ترکیبی مبتنی بر اختلاط خبره‌ها که در هر خبره یک مدل BEL بود؛ برای افزایش کارایی سیستم استفاده شد. برای نشان دادن مزیت استفاده از مدل اختلاط خبره‌ها مبتنی بر مدل BEL، این مدل با طبقه‌بندهای دیگر مورد مقایسه قرار گرفت. در جدول ۶ این مقایسه نشان داده شده است. مطابق این جدول، دقت بازشناسی مدل پیشنهادی MoBEL از همه بهتر بود.

از دیگر مزایای مدل MoBEL این است که از نظر حافظه مصرفی، تعداد پارامترهای آموزش و سرعت پردازش بسیار کارامدتر از سایر طبقه‌بندها از جمله مدل اختلاط خبره‌های مبتنی بر شبکه‌های عصبی است. پیچیدگی محاسباتی مدل BEL، $O(n)$ می‌باشد اما پیچیدگی محاسباتی در مدل MoBEL، $O(3n+3)$ است. مدل پیشنهادی MoBEL از دو مدل BEL

جدول ۶. مقایسه دقت بازشناسی هیجان مبتنی بر همجوشی سطح ویژگی بین طبقه‌بندهای مختلف

انواع طبقه‌بندها	دقت بازشناسی (درصد)
MLP	۷۸
BEL	۷۸/۷
SVM	۷۷/۹
Weighted KNN	۷۸
RBF	۷۱
Mixture of NN (MoE)	۸۰
MoBEL	۸۰/۷

به عنوان خبره تشکیل شده بود که این خبره‌ها می‌توانند به صورت موازی اجرا شوند. بنابراین در حالت کلی تعداد اتصالات برابر است با $n+1$ وزن برای آمیگدال و $n+1$ وزن برای OFC و n وزن برای خروجی شبکه میانجی می‌باشد. همچنین، شبکه عصبی MoBEL می‌تواند در کنار مدل‌های یادگیری عمیق استخراج ویژگی‌های هیجان به صورت end-to-end و بدون دخالت کاربر آموزش ببیند.

ماتریس درهم‌ریختگی بازشناسی هیجان چندوجهی در مدل پیشنهادی همجوشی ویژگی‌های فرآیند شنیداری_دیداری با استفاده از مدل MoBEL در جدول ۷ نشان داده شده است. جالب است که در پایگاه داده eNterface مطابق شکل ۴، هیجان‌های «تنفر» و «عصبانیت» بالاترین دقت بازشناسی و هیجان‌های تعجب و ترس کمترین دقت بازشناسی با کمتر از ۸۰ درصد را دارند در صورتی که سایر هیجان‌ها دقت بالای ۸۰ درصد را داشتند. همچنین هیجان ترس و ناراحتی، همپوشانی بیشتری با هم دارند. به طوری که ۱۴ درصد از نمونه‌های هیجان ترس در ناراحتی گذاشته شده است. علاوه بر این از این جدول استنباط می‌شود که بازشناسی هیجان «تنفر» با دقت تقریباً ۹۰ درصد یعنی راحت‌تر از سایر هیجان‌ها است.

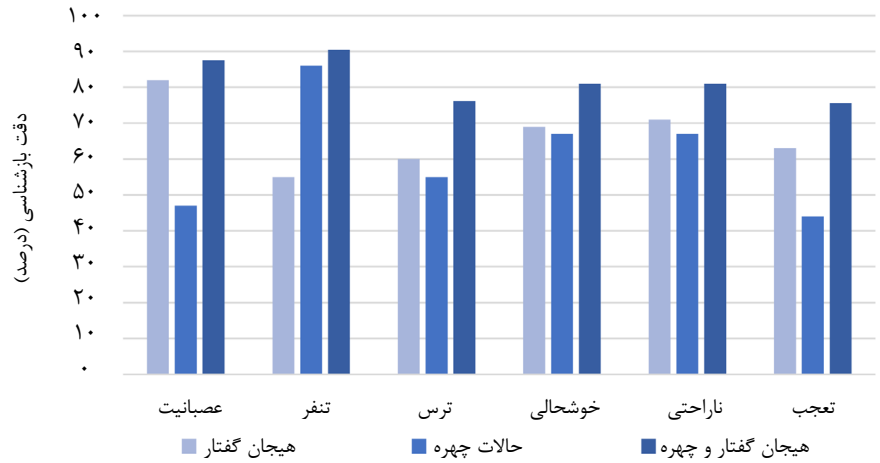
در شکل ۳ دقت بازشناسی هیجان از روی حالات چهره، هیجان گفتار و همجوشی ویژگی‌ها نشان داده شده است. همان طور که مشاهده می‌شود دقت بازشناسی هیجان چندوجهی بیشتر از هر کدام از وجه‌ها بود. همچنین نشان داد که می‌تواند با دقت بازشناسی کم در هر یک از وجه‌های مختلف مقابله کند. به عنوان مثال در بازشناسی حالات چهره، هیجان «تعجب» دارای کمترین دقت است در حالی که در بازشناسی هیجان گفتار، هیجان «تنفر» دارای کمترین دقت بازشناسی بود و پس از همجوشی دو فرآیند شنیداری_دیداری با استفاده از مدل پیشنهادی از سایر تک وجهی‌ها در هر کدام از کلاس‌های هیجان بهتر بود.

علاوه بر این، به منظور نمایش سرعت همگرایی مدل پیشنهادی MoBEL، در شکل ۴ میانگین مربعات خطای آموزش، تست و اعتبارسنجی مدل پیشنهادی به ازای تعداد تکرار مشخص نشان داده

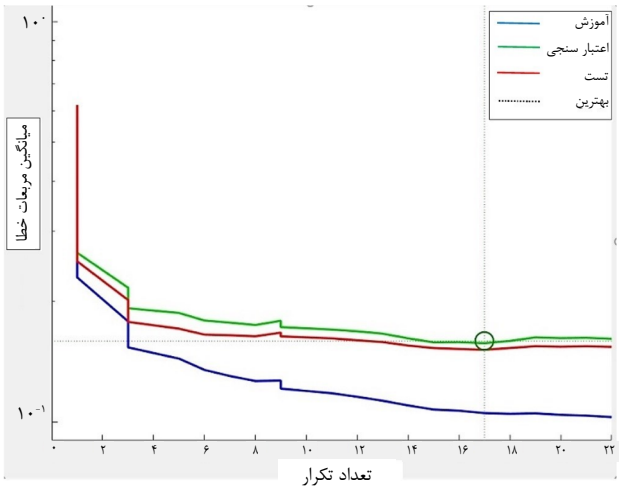
شده است. مطابق شکل، مدل MoBEL بعد از ۲۲ مرحله تکرار به سرعت یاد می‌گیرد و وزن‌های مدل همگرا می‌شوند. همه بهتر بود. علاوه بر این، به منظور نمایش سرعت همگرایی مدل پیشنهادی MOBEL، در شکل ۴ میانگین مربعات خطای آموزش، تست و اعتبارسنجی مدل پیشنهادی به ازای تعداد تکرار مشخص نشان داده شده است. مطابق شکل، مدل MOBEL بعد از ۲۲ مرحله تکرار به سرعت یاد می‌گیرد و وزن‌های مدل همگرا می‌شوند.

جدول ۷. ماتریس درهم‌ـریختگی بازشناسی هیجان چندوجهی با استفاده از مدل MoBEL بر روی پایگاه داده eNterface'05 (درصد)

	عصبانیت	تنفر	ترس	خوشحالی	ناراحتی	تعجب
عصبانیت	**۸۵**	۷/۱	۲/۳	۲/۳	۰	۲/۳
تنفر	۲/۳	**۸۸/۱**	۲/۳	۲/۳	۴/۷	۰
ترس	۲/۳	۰	**۷۳/۸**	۷/۱	۱۴	۲/۳
خوشحالی	۷/۱	۲/۳	۴/۷	**۷۸/۶**	۲/۳	۴/۷
ناراحتی	۲/۳	۰	۹/۵	۰	**۸۱**	۹/۵
تعجب	۲/۳	۰	۴/۷	۷/۱	۷/۱	**۷۸**



شکل ۳. نمودار مقایسه دقت بازشناسی هیجان از روی حالات چهره، هیجان گفتار و همجوشی فرآیند شنیداری‌دیداری به ازای هیجان‌های مختلف



شکل ۴. میانگین مربعات خطا (MSE) در طول ۲۲ مرحله تکرار از آموزش، تست و اعتبارسنجی مدل MOBEL

بحث

در این پژوهش، ما روش جدیدی برای بازشناسی هیجان چندوجهی مبتنی بر همجوشی فرآیند شنیداری_دیداری با استفاده از شبکه عصبی MoBEL الهام گرفته شده از سیستم لیمبیک مغز ارائه کردیم. مدل پیشنهادی ویژگی‌های یاد گرفته شده از جریان‌های بینایی و شنوایی را با استفاده از الگوریتم‌های یادگیری عمیق، به ترتیب 3D-CNN و CNN با هم ادغام می‌کند و در شبکه عصبی جدید MoBEL ویژگی‌های ادغام شده آموزش می‌بینند. شبکه MoBEL هم همبستگی موجود بین ویژگی‌های شنوایی و بینایی را یاد می‌گیرد و هم به عنوان یک طبقه‌بند، خروجی کلاس هیجان ویدیو را می‌دهد. این شبکه از ترکیبی از چند BEL تشکیل شده است که الهام گرفته شده از سیستم لیمبیک مغز و نیز قشر انجمنی مغز (Associative cortex) می‌باشد که ادغام اطلاعات از منابع حسی مختلف در آنجا صورت می‌گیرد (۴۵). مدل MoBEL از چند شبکه خبره و یک شبکه میانجی مبتنی بر مدل BEL تشکیل شده است. همچنین یکی دیگر از چالش‌های بازشناسی هیجان در ویدیو، تحلیل سری زمانی بروز هیجان در چهره و گفتار بود. به همین منظور ما از روش یادگیری عمیق 3D-CNN برای بازنمایی اطلاعات مکان_زمان هیجان چهره و CNN برای بازنمایی اطلاعات مکان_زمان هیجان گفتار استفاده کرده‌ایم که قدرت بالایی در تفکیک و دسته‌بندی کلاس هیجان نسبت به روش‌های استاتیک و دستی دارند. آزمایشات انجام شده بر روی پایگاه داده صوتی_تصویری eNterface'05 نشان می‌دهد که مدل پیشنهادی بسیار بهتر از سایر مدل‌های قبلی کار شده در زمینه استخراج ویژگی‌های دستی هیجان گفتار و حالات چهره و همچنین همجوشی اطلاعات به منظور بازشناسی هیجان در ویدیو می‌باشد.

برای ادامه خط اصلی این پژوهش، در زمینه بازشناسی هیجان چندمدالیتی، ایده‌های مختلفی در زمینه‌های نظری می‌توان ارائه داد. ایده اول برای بازشناسی هیجان در ویدیوهای طولانی این است که ابتدا کل فیلم به بخش‌هایی حدوداً ۴ ثانیه تقسیم شده و سپس مدل پیشنهادی در هر بخشی اعمال شود و خروجی هیجان بدست آید. ایده دوم این است که در مدل پیشنهادی MoBEL از مدل BEL با ناظر استفاده شد در حالی که مدل اصلی BEL مبتنی بر پاداش/جریمه است. بنابراین به عنوان کار آینده برای بهبود مدل BEL می‌توان از الگوریتم‌های یادگیری عمیق در سیگنال پاداش/جریمه مدل BEL استفاده کرد. همچنین با انجام یکسری روش‌های پیش‌پردازش اولیه و حذف نویز می‌توان در آینده، مدل پیشنهادی را بر روی مجموعه پایگاه داده‌های طبیعی مانند AFEW (۴۶) و BAUM-1s (۴۷) نیز آزمایش کرد.

نتیجه‌گیری

نوآوری این مقاله در ارائه مدل MoBEL برای بازشناسی هیجان در ویدیو بود. مدل MoBEL از چند شبکه خبره و یک شبکه میانجی مبتنی بر مدل BEL تشکیل شده است. هدف از این مدل، همجوشی ویژگی‌های هیجان گفتار و حالات چهره در سطوح بالاتر است به طوری که شبکه یاد می‌گیرد به کدام وجه (گفتار یا چهره) وزن بیشتری اختصاص دهد. نتایج این آزمایش‌ها نشان داد که دقت بازشناسی هیجان مدل پیشنهادی برای پایگاه داده eNterface نسبت به سایر روش‌های همجوشی بهبود یافته است.

تشکر و قدردانی

این مقاله حاصل از پایان‌نامه در مقطع دکتری هوش مصنوعی در دانشگاه علوم و تحقیقات آزاد اسلامی واحد تهران است.

References

1. Peter C, Beale R. The role of affect and emotion in HCI. In Peter C, Beale R, editors. Affect and emotion in HCI. Lecture notes in computer science. Vol 4868. Berlin:Springer;2008. pp. 23–34.

2. Szwoch M, Szwoch W. Emotion recognition for affect aware video games. In: Choraś R, editor. Image processing & communications challenges. Advances in Intelligent Systems and Computing. Vol 313. New York:Springer;2015. pp. 227–236.

3. Shen L, Wang M, Shen R. Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society*. 2009;12(2):176-189.

4. Torous J, Friedman R, Keshavan M. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth and uHealth*. 2014;2(1):e2.

5. Lucey S, Ashraf AB, Cohn JF. Investigating spontaneous

facial action recognition through AAM representations of the face. In Delac K, Grgic M, editors. Face Recognition. I-Tech Education and Publishing: Rijeka, Croatia;2007. pp. 275-286.

6. Guo G, Dyer CR. Learning from examples in the small sample case: Face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. 2005;35(3):477-488.

7. Chang Y, Hu C, Feris R, Turk M. Manifold based analysis of facial expression. *Image and Vision Computing*. 2006;24(6):605-614.

8. Anderson K, McOwan PW. A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics*). 2006;36(1):96-105.

9. Pantic M, Patras I. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. 2006;36(2):433-349.

10. Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis And Machine Intelligence*. 2007;29(6):915-928.

11. Dhall A, Asthana A, Goecke R, Gedeon T. Emotion recognition using PHOG and LPQ features. Face and Gesture. Santa Barbara, CA:IEEE;2011. pp. 878-883.

12. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, et al. Adieu features end-to-end speech emotion recognition using a deep convolutional recurrent network. 2016 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). Shanghai;2016. pp. 5200-5204.

13. Mansoorizadeh M, Charkari NM. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*. 2010;49(2):277-297.

14. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (ICCV). Santiago, Chile;2015. pp. 4489-4497.

15. Fan Y, Lu X, Li D, Liu Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction;2016. pp. 445-450.

16. Rong J, Li G, Chen YP. Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management*. 2009;45(3):315-328.

17. Wu S, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*. 2011;53(5):768-785.

18. Dave N. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology*. 2013;1(6):1-4.

19. El Ayadi MM, Kamel MS, Karray F. Speech emotion recognition using Gaussian mixture vector autoregressive models. International Conference on Acoustics, Speech, and Signal Processing. 16-20 April 2007; Honolulu, Hawaii. Vol 4. pp. IV-957. IEEE.

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

21. Zeng Z, Tu J, Pianfetti BM, Huang TS. Audio–visual affective expression recognition through multistream fused HMM. *IEEE Transactions on Multimedia*. 2008;10(4):570-577.

22. Ntalampiras S, Fakotakis N. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing*. 2011;3(1):116-125.

23. Zhang S, Zhang S, Huang T, Gao W, Tian Q. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. 2017;28(10):3030-3043.

24. Potamianos G, Neti C, Gravier G, Garg A, Senior AW. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*. 2003;91(9):1306-1326.

25. Kahou SE, Bouthillier X, Lamblin P, Gulcehre C, Michalski V, Konda K, Jean S, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on*

Multimodal User Interfaces. 2016;10(2):99-111.

27. Lan ZZ, Bao L, Yu SI, Liu W, Hauptmann AG. Multimedia classification and event detection using double fusion. *Multimedia Tools and Applications.* 2014;71(1):333-347.

28. Schuller B, Müeller R, Höernler B, Höethker A, Konosu H, Rigoll G. Audiovisual recognition of spontaneous interest within conversations. In Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI). Nov 12 2007; Aichi, Japan. pp. 30–37.

29. Wang Y, Guan L, Venetsanopoulos AN. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*. 2012;14(3):597-607.

30. Gurban M, Thiran JP, Drugman T, Dutoit T. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In Proceedings of the 10th international conference on Multimodal Interfaces. Oct 20; 2008. pp. 237-240.

31. Chen S, Jin Q. Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Oct 26 2015. pp. 49-56.

32. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning. 2011 June 28-July 2; Bellevue, WA, USA;2011.

33. Balkenius C, Morén J. A computational model of emotional conditioning in the brain. In Proceedings of Workshop on Grounding Emotions in Adaptive Systems. Zurich;1998.

34. Lucas C. BELBIC and its industrial applications: towards embedded neuroemotional control codesign. In Integrated systems, design and technology 2010-2011. Berlin, Heidelberg:Springer. pp. 203-214.

35. Lotfi E, Akbarzadeh-T MR. Brain emotional learning-based pattern recognizer. *Cybernetics and Systems*. 2013;44(5):402-421.

36. Farhoudi Z, Setayeshi S, Rabiee A. Using learning automata in brain emotional learning for speech emotion recognition. *International Journal of Speech Technology*. 2017;20(3):553-562.

37. Lotfi E, Khazaei O, Khazaei F. Competitive brain emotional learning. *Neural Processing Letters*. 2018;47(2):745-764.

38. Fino E, Yuste R. Dense inhibitory connectivity in neocortex. *Neuron*. 2011;69(6):1188-1203.

39. Zhalehpour S, Onder O, Akhtar Z, Erdem CE. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*. 2016;8(3):300-313.

40. Bejani M, Gharavian D, Charkari NM. Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Computing and Applications*. 2014;24(2):399-412.

41. Sahoo S, Routray A. Emotion recognition from audio-visual data using rule based decision level fusio. 2016 IEEE Students' Technology Symposium (TechSym). Kharagpur, India;2016. pp. 7-12.

42. Zhang SH, Wang XI, Zhang GA, Zhao XI. Multimodal emotion recognition integrating affective speech with facial expression. *WSEAS Transactions on Signal Processing*. 2014;10(2014):526-537.

43. Badshah AM, Ahmad J, Rahim N, Baik SW. Speech emotion recognition from spectrograms with deep convolutional neural network. In 2017 International Conference On Platform Technology And Service (PlatCon). Feb 13 2017;Busan. pp. 1-5. IEEE.

44. Mansoorizadeh M, Charkari NM. Speech emotion recognition: Comparison of speech segmentation approaches. In Proceedings of IKT. Mashad, Iran;2007.

45. Stein BE, Stanford TR, Rowland BA. The neural basis of multisensory integration in the midbrain: Its organization and maturation. *Hearing Research.* 2009;258(1-2):4-15.

46. Dhall A, Goecke R, Lucey S, Gedeon T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*. 2012;19(3):34-41.

47. Zhalehpour S, Onder O, Akhtar Z, Erdem CE. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*. 2016;8(3):300-313.