نگاهی بر تشخیص احساس در گفتار مبتنی بر استنتاج فازی

محمدمهدی مختاری
دانشگاه جامع امام حسین (ع)
محمدعلی جوادزاده
دانشگاه جامع امام حسین (ع)
میثم میرزایی
دانشگاه جامع امام حسین (ع)

چکیده

انسان‌ها در طول زندگی خود، بر اساس دلیل و احساس تصمیم‌گیری می‌کنند. احساسات بخش مهمی از انسان بودن است. این احساسات به‌وسیله رفتار و گفتار ادا می‌شوند. در دهه‌های اخیر احساسات توسط کامپیوترها با استفاده از تشخیص گفتار و پردازش سیگنال پیش‌بینی و تشخیص داده شده‌اند. سیستم‌های مؤثر در زمینه تشخیص احساسات به طور فزاینده‌ای موردتوجه محققان و شرکت‌های دارای فناوری پیشرفته قرار می‌گیرند تا بتوانند با خلق‌وخوی کاربر تصدیق یا سازگاری داشته باشند. مسئله مهم در این میان انتخاب روش‌های متناسب با مسائل تحت بررسی، با درنظرداشتن مزایا و معایب آن‌ها است. تشخیص احساسات از سیگنال گفتار عمدتاً با استفاده از دسته‌بندهای نظارت‌شده به‌دست‌آمده است. بااین‌حال، به نظر می‌رسد تکنیک‌های خوشه‌بندی برای حل چنین مشکلی مناسب هستند، به‌ویژه در پایگاه‌های داده بزرگ، جایی که برچسب‌گذاری گفتار ممکن است یک کار سخت و خسته‌کننده باشد. پس از تصمیم‌گیری دررابطه‌با نوع تکنیک مورداستفاده و پس از انجام آزمایش‌ها و نتایج آن‌ها، محققان دریافته‌اند که عمده تکنیک‌ها نیازمند تشخیص بهتر و دقیق هستند. سیستم‌های فازی یکی از مفیدترین روش‌های بهینه‌سازی در این زمینه به‌حساب می‌آیند. سیستم‌های فازی نیز در هر دو حالت دسته‌بندی و خوشه‌بندی قابل استفاده‌اند. ما در این مقاله به بیان چند کاربرد از سیستم‌های فازی در تشخیص احساسات گفتار پرداخته‌ایم.

واژگان کلیدی: تشخیص احساسات، پردازش سیگنال، سیستم‌های فازی، بهینه‌سازی

مقدمه

احساسات یک جنبه تا حد زیادی تأثیرگذار از زندگی انسان است که بر بسیاری از تعاملات و فرایندهای تصمیم‌گیری تأثیر می‌گذارد. این ممکن است به دلیل مسیر تکاملی ما باشد که منجر به رشد احساسات به‌عنوان ابزاری برای سازگاری رفتاری با سناریوهای متمایز شده است. تحقیقات در مورد تعامل انسان و کامپیوتر[1] برای تشخیص احساسات به طور فزاینده‌ای تبدیل به یک موضوع داغ می‌شود. در زمینه تشخیص احساسات، نیاز به توسعه ماشین‌هایی است که بتوانند احساسات انسان را بهتر درک کنند. در زمینه تعامل انسان و کامپیوتر، کامپیوتر می‌تواند احساسات را از طریق ژست، سیگنال‌های صوتی، حالت بدن، حالت چهره، سیگنال‌های فیزیولوژیکی و روش‌های تصویربرداری عصبی و غیره تشخیص دهد. ازاین‌رو، تحقیق در مورد رشد عاطفی و شبیه‌سازی بالقوه آن در ماشین‌ها ممکن است تلاشی ارزشمند به‌منظور افزایش پویایی پاسخ و سازگاری آن‌ها باشد. زبان بخش مهمی از تمدن است و گفتار راحت‌ترین راه برای برقراری ارتباط است. سیگنال گفتار اطلاعات مفید زیادی را منتقل می‌کند. مطالعه زبان‌شناختی مدرن بر اطلاعات زبان تمرکز می‌کند و به نتایج بسیاری دست می‌یابد. بااین‌حال، بسیاری از اطلاعات زبانی موازی وجود دارد که در تشخیص سنتی گفتار، مانند وضعیت عاطفی، جنسیت و سن گوینده، بیش از حد مورد بررسی قرار می‌گیرند. این نوع اطلاعات گفتار برای تعامل انسان و رایانه، تشخیص بیماری‌های روان‌پزشکی، تنظیم رفتار رانندگی از حالت رانندگان، بازی‌های اینترنتی، ارتباط با مراکز تماس اورژانس و واقعیت مجازی بسیار مهم است. بسیاری از سیستم‌هایی مانند دستیار خانگی Google Home که بخشی از استفاده روزمره هستند، می‌توانند از این مزیت بهره ببرند.

استخراج و انتخاب ویژگی

استخراج ویژگی، یکی از جنبه‌های مهم در یک سیستم تشخیص احساسات گفتار است و ویژگی‌های استخراج شده باید به طور دقیق اطلاعات احساسی گفتار را منعکس کند. در واقع انتخاب ویژگی به مشکل بستگی دارد. به‌طورکلی در تشخیص گفتار، ویژگی‌ها را می‌توان به عروضی در مقابل آکوستیک یا طیفی تقسیم کرد. ویژگی‌های عروضی شامل f0، شدت و مدت‌زمان یک صوت هستند، درحالی‌که ویژگی‌های صوتی از طیفی مانند ضرایب فرکانس مغزی مل[2]، پارامترهای طیفی خطی[3]، و مشتقات زمانی اول و دوم آن‌ها استخراج می‌شوند. در برخی موارد نیز مشاهده شده است که از دو ویژگی لرزش، میزان تغییر در گفتار f0 از یک چرخه به چرخه دیگر و تاب، وسعت تنوع در دامنه گفتار از یک چرخه به چرخه دیگر، در تحقیقات مورداستفاده قرار گرفته‌اند.

توضیحات بیان‌شده در زمانی رخ می‌دهند که به‌صورت کلاسیک ویژگی‌ها مورداستفاده واقع شوند. این نکته دررابطه‌با تحقیقات فازی متفاوت است پس از استخراج ویژگی از تکنیک‌هایی استفاده نموده‌اند که متناسب با مسئله و نوع مدل فازی و دسته‌بندی، ویژگی-های مناسب را انتخاب نمایند که موفق نیز بوده است. البته در برخی موارد نیز این تکنیک‌ها در دقت سامانه تأثیر چندان مثبتی نداشته‌اند. به طور مثال در تحقیقاتی مانند پژوهش رووتا و همکارانش (Rovetta et el, 2019) از تکنیک‌های تحلیل واریانس[4] و اطلاعات متقابل[5] استفاده شده است. اما بااین‌حال تأثیر مثبت آن‌ها به میزانی نبوده است که بتوان گفت استفاده از آن‌ها مؤثر بوده است.

[1] Human – Computer Interaction
[2] Mel-Frequency Cepstral Coefficients or MFCC
[3] Line Spectrum Pairs or LSP
[4] Analysis of Variance or ANOVA
[5] Matual Information

جدول ۱.نمونه‌ای از انواع ویژگی‌های مورد استفاده در پژوهش‌های مورد مطالعه

Feature Type	Feature Number	Feature Name
Pitch	1	Median pitch
	2	Mean pitch
	3	Standard deviation
	4	Minimum pitch
	5	Maximum pitch
Pulses	6	Number of pulses
	7	Number of periods
	8	Mean period
	9	Standard deviation of period
Voice	10	Fraction of locally unvoiced frames
	11	Number of voice breaks
	12	Degree of voice breaks
Jitter	13	Jitter (local)
	14	Jitter (local, absolute)
	15	Jitter (rap)
	16	Jitter (ppq5)
	17	Jitter (ddp)
Shimmer	18	Shimmer (local)
	19	Shimmer (local, dB)
	20	Shimmer (apq3)
	21	Shimmer (apq5)
	22	Shimmer (apq11)
	23	Shimmer (dda)
Harmonics	24	Mean autocorrelation
	25	Mean noise-to-harmonics ratio
	26	Mean harmonics-to-noise ratio

باتوجه‌به دو حالت گفته شده در استخراج و انتخاب ویژگی، در اکثر پژوهش‌ها در زمینه تشخیص احساسات گفتار، از ویژگی‌های کلاسیک، مخصوصاً ویژگی‌های آکوستیک استفاده می‌گردد. در ادامه نیز به کاربرد سیستم فازی در هر کدام از پژوهش‌های موردمطالعه می‌پردازیم.

سیستم‌ها و مدل‌های فازی مورداستفاده

در این بخش به سراغ آن می‌رویم که چه مدل‌هایی که شامل منطق فازی هستند در تحلیل احساسات گفتار به کار گرفته شده‌اند و چه تأثیری در نتایج پژوهش‌ها داشته‌اند.

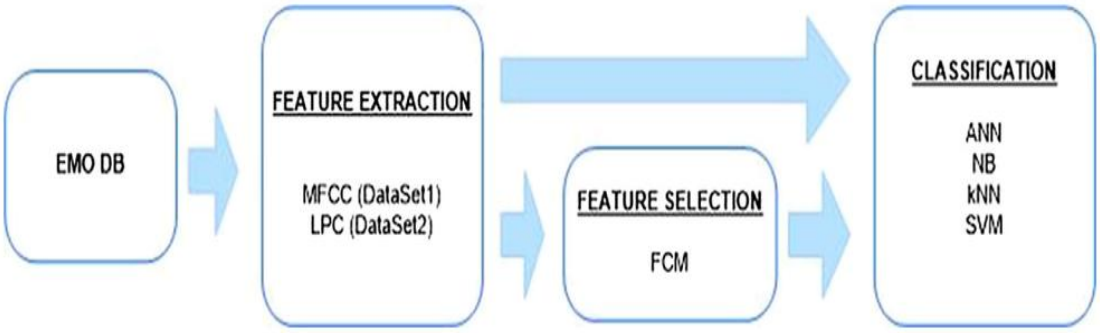
C-means فازی

این مدل در اصل یک روش جهت خوشه‌بندی داده‌ها است که کاربرد قابل‌توجهی در انواع پژوهش‌های داده‌کاوی همانند تشخیص احساسات و در پژوهش‌های موردمطالعه علاوه بر خوشه‌بندی، وظیفه انتخاب ویژگی مناسب، پس از استخراج ویژگی توسط پژوهشگران را برعهده داشته است. در پژوهش رووتا و پژوهش دمیرکان و همکارانشان (Demircan et el, 2018) از مجموعه‌داده یکسان EmoDB مورداستفاده قرار گرفته است. در پژوهش رووتا از C-means فازی به‌عنوان خوشه‌بند استفاده شده است. ازآنجایی‌که انتخاب ویژگی‌های مناسبی توسط تکنیک‌های تحلیل واریانس و اطلاعات متقابل صورت نگرفته است، باتوجه‌به اینکه مدل فازی تغییراتی در افزایش دقت داشته، دقت تشخیص بالایی نسبت به سایر پژوهش‌ها نداشته است. نتایج مقایسه‌ای مربوط به مدل در جدول ۲ قابل‌مشاهده است.

جدول ۲. دقت تشخیص با استفاده از C-means فازی

# of classes	# of clusters	Feature selection method	Proportion of selected features	t	α	kmeans rate (%)	FCM rate (%)
7	7	ANOVA-Group	50%	0.1	0.9	51.9	69.6
7	14	ANOVA	25%	0.1	0.9	56.6	55.9
7	21	ANOVA	25%	0.1	0.9	60.9	63.0
4	4	ANOVA	75%	0.1	0.9	62.4	61.3
4	8	ANOVA	50%	0.1	0.9	69.9	77.4
4	12	ANOVA	50%	0.1	0.9	73.9	75.1

در پژوهش دمیرکان این مدل به‌عنوان انتخاب‌گر ویژگی از میان ویژگی‌های استخراجی توسط ضرایب فرکانس مغزی مل و پارامترهای طیفی خطی بوده است. طبق شکل ۱ پس از استخراج ویژگی توسط هر کدام از تکنیک‌ها، به دو صورت ویژگی‌ها تحت بررسی قرار گرفته‌اند. اول به‌صورت مستقیم ویژگی‌ها به دسته‌بندهای شبکه عصبی[۶]، بیز ساده[۷]، k نزدیک‌ترین همسایه[۸] و ماشین بردار پشتیبان[۹] جهت تشخیص منتقل شده‌اند. دوم، این ویژگی‌ها به مدل C-means فازی به‌عنوان ورودی منتقل شده‌اند تا انتخاب ویژگی و کاهش ویژگی بر روی داده‌ها صورت پذیرد. پس از انتخاب و کاهش نیز خروجی مدل به دسته‌بندهای مورداستفاده داده شده که تشخیص را انجام دهند.



شکل ۱. معماری مربوط به روش پیشنهادی دمیرکان

پس اتمام آزمایش‌ها بسیار مبرهن است که این انتخاب و کاهش ویژگی در هر دو تکنیک ضرایب فرکانس مغزی مل و پارامترهای طیفی خطی تأثیر بسزایی داشته است. نتایج این پژوهش در دو جدول ۳ و ۴ قابل مشاهده است.

جدول ۳. جدول دقت تشخیص با استفاده از ویژگی‌های MFCC

	Classification	Accuracy (%)	Average precision	Average recall	*F*-score	Area under the ROC
DataSet1	ANN	72.11	0.72	0.72	0.72	0.94
	NB	61.86	0.61	0.62	0.61	0.90
	kNN	69.64	0.70	0.70	0.69	0.93
	SVM	71.72	0.72	0.70	0.71	0.93
DataSet1 with FCM	ANN	90.00	0.91	0.90	0.90	0.95
	NB	91.43	0.94	0.91	0.92	0.95
	kNN	**92.86**	0.95	0.93	0.93	0.96
	SVM	**92.86**	0.95	0.93	0.93	0.94

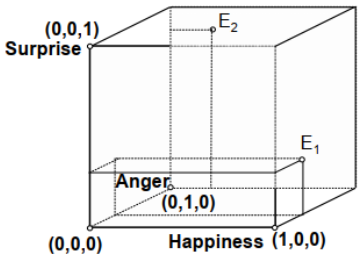
[6] Neural Network
[7] Naïve Bayes
[8] k-Nearest Neighbor or k-nn
[9] Support Vector Machine or SVM

جدول ۴ .جدول دقت تشخیص با استفاده از ویژگی‌های LSP

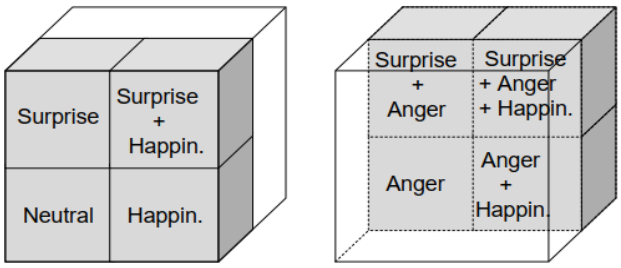
	Classification	Accuracy (%)	Average precision	Average recall	*F*-score	Area under the ROC
DataSet2	ANN	46.73	0.45	0.467	0.46	0.81
	NB	35.0	0.37	0.35	0.33	0.73
	kNN	42.31	0.44	0.42	0.40	0.76
	SVM	42.5	0.38	0.43	0.38	0.76
DataSet2 with FCM	ANN	44.23	0.42	0.44	0.42	0.79
	NB	27.63	0.20	0.28	0.22	0.66
	kNN	32.12	0.30	0.32	0.29	0.70
	SVM	33.46	0.22	0.34	0.25	0.66

PROSBER

PROSBER یک مدل تشخیص احساسات مبتنی بر منطق فازی است که احساس را از ویژگی‌های عروضی گفتار تشخیص می‌دهد. جملات منفرد را به‌عنوان ورودی می‌گیرد و آن‌ها را در دسته‌های احساسات شادی، غم، خشم و ترس دسته‌بندی می‌کند. علاوه بر این یک حالت عاطفی خنثی متمایز می‌شود. PROSBER به طور خودکار مدل‌های فازی را برای تشخیص احساسات تولید می‌کند. براین‌اساس دو حالت آموزش و شناخت متمایز می‌شود. در پژوهش انجام‌شده توسط اسائو و همکاران (Esau et el, 2005) از این مدل استفاده شده است. ابتدا از مفهوم یک ابرمکعب فازی n بعدی برای نشان‌دادن حالات عاطفی متشکل از n احساس اساسی استفاده کرده‌اند. از منظر پژوهشگران برخلاف سایر رویکردها، این روش نه‌تنها امکان بازنمایی و شناسایی مجموعه ثابتی از احساسات اساسی را فراهم می‌کند، بلکه از مدیریت احساسات مشتق‌شده نیز پشتیبانی می‌کند. همچنین پس از آن کاربرد این مدل را با استفاده از سیستم تشخیص احساسات فازی PROSBER ارائه نموده‌اند. به‌عنوان اولین گام، تقسیمی از ابرمکعب واحد را در زیرمکعب‌هایی با اندازه مساوی برای تشخیص احساسات اساسی و ترکیب آن‌ها، طبق شکل ۲ و ۳، پیشنهاد شده است. نکته جالب برای بررسی بیشتر این است که آیا تقسیم‌بندی صورت‌گرفته با شناخت انسان مطابقت دارد یا خیر. برای مثال، تقسیم‌بندی می‌تواند با استفاده از یک رویکرد یادگیری انجام شود که به طور خودکار چنین زیربخش‌هایی را پیدا کرده و آن را با تفسیرهای انسانی از حالات احساسی مربوطه مقایسه نماید. اما بااین‌وجود تنها مشکل این پژوهش را می‌توان عدم ارائه میزان کمی دقت این روش بیان نمود.



شکل ۲. نمونه تقسیم‌بندی ابرمکعب حالت‌های احساسی

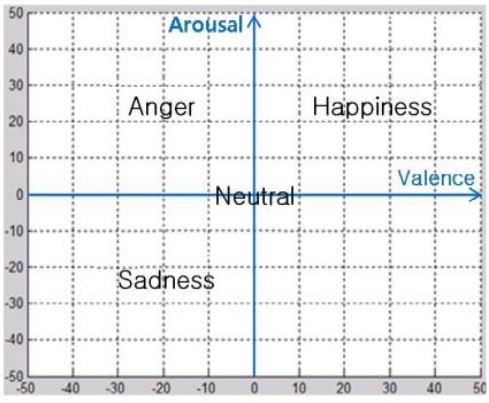


شکل ۳. نمونه تقسیم‌بندی نهایی و مساوی ابرمکعب حالت‌های احساسی

شبکه عصبی فازی با استفاده از تابع NEWFM

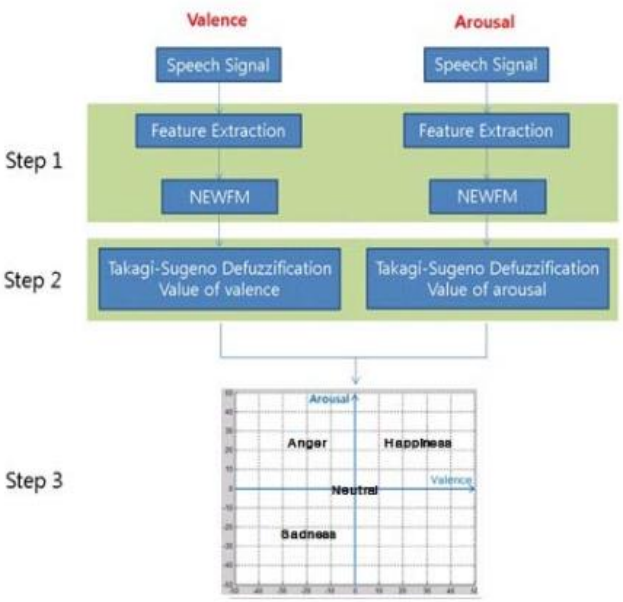
ژانگ و همکاران (Zang et el, 2015) با به‌کارگیری یک شبکه عصبی فازی بر اساس تابعی با عضویت فازی وزنی[10]، یک تشخیص احساسات دوبعدی جدید از گفتار ارائه شده است. در این مورد نیز همانند C-Means فازی، از مجموعه‌داده EmoDB استفاده شده است. NEWFM یک سیستم دسته‌بندی مبتنی بر شبکه عصبی فازی است که از طریق آن می‌توان توابع عضویت فازی را از ویژگی‌های ورودی با پردازش آموزش نتیجه گرفت. تعداد زیادی از توابع عضویت فازی وزنی کران‌دار[11] در این پژوهش استفاده شده است. در طول آموزش شبکه عصبی فازی، مقدار غیرفازی‌سازی تاکاگی-سوگنو برای مدل بصری دوبعدی توسط BSWFM محاسبه شده است.

در روش بیان‌شده، سیگنال‌های فیزیولوژیکی را می‌توان جمع‌آوری نمود و به‌وسیله این روش، احساسات به‌صورت هدفمند می‌توانند در چهار قسمت مدل احساسات دوبعدی منعکس شوند. عواطف دارای سه ویژگی ظرفیت، برانگیختگی و کنترل هستند. ژانگ تنها شامل ظرفیت و برانگیختگی است. تحریک و غم به ظرفیت کم نسبت داده می‌شود، درحالی‌که شادی به ظرفیت بالا تعلق دارد. هر دو تحریک و شادی به‌عنوان برانگیختگی زیاد در نظر گرفته می‌شوند درحالی‌که غم به برانگیختگی کم تعلق دارد. NEWFM برای تشخیص ظرفیت کم یا بالا و همچنین برای میزان برانگیختگی به ترتیب به‌کارگرفته‌شده است تا با استفاده از مدل بصری دوبعدی، مقدار غیرفازی‌سازی تاکاگی-سوگنو را به دست آورد. شکل ۴ نمودار دوبعدی بر اساس ظرفیت و برانگیختگی را نمایش می‌دهد. همچنین در شکل ۵ ساختار مدل فازی استفاده‌شده قابل‌مشاهده است.



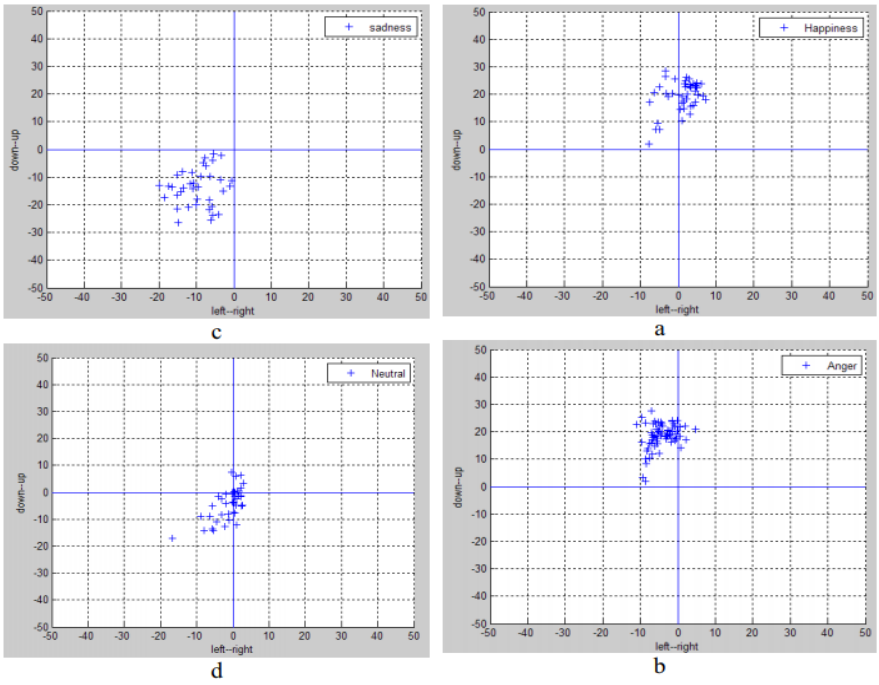
شکل ۴. مدل بصری دوبعدی ظرفیت و برانگیختگی NEWFM

[10] Neural Network with Weighted Fuzzy Membership Functions or NEWFM
[11] Bounded Sum of Weighted Fuzzy Membership Functions or BSWFM



شکل ۵. ساختار مدل فازی

با مقادیر غیرفازی‌سازی تاکاگی-سوگنو، احساسات در صفحه دوبعدی توصیف شدند. سیستم در این تحقیق به دقت بالای ۸۵ درصد در دسته‌بندی با استفاده از شناسایی احساسات دوبعدی مبتنی بر NEWFM دست یافت. دقت دسته‌بندی شادی، خشم، غم و خنثی به ترتیب ۷۰.۵، ۹۱، ۱۰۰ و ۸۱ درصد بود. نتایج نشان‌دهنده این است که روش مورداستفاده، در مواردی که محققان تنها نیازمند تشخیص غم هستند، می‌تواند بدون خطا عمل نماید. شکل ۶ مربوط به نتایج بصری پژوهش است.



شکل ۶. نتایج بصری تشخیص احساسات در پژوهش ژانگ و همکاران

k نزدیک‌ترین همسایه فازی

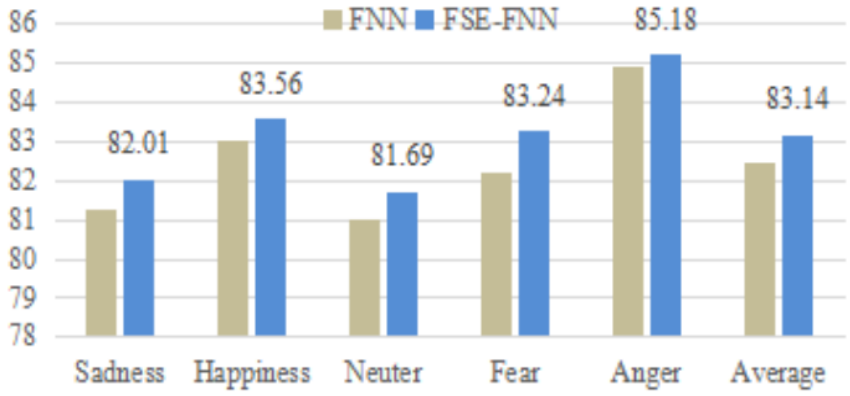
در پژوهش‌های زبانسیاک (Zbancioc et el, 2012) و یوآن (Yuan et el, 2016) از الگوریتم k نزدیک‌ترین همسایه استفاده شده است. پژوهش زبانسیاک و همکاران بر جنبه یادگیری نظارت‌شده k نزدیک‌ترین همسایه به‌منظور دسته‌بندی احساسات گفتار تمرکز دارد. الگوریتم دسته‌بند فازی k نزدیک‌ترین همسایه در مقایسه باحالت کلاسیکش این مزیت را دارد که «قدرت» عضویت در یک کلاس را تعیین کند. در الگوریتم کلاسیک، تصمیم در مورد تخصیص یک نمونه به یک کلاس تنها بر اساس اکثریت تعداد همسایگان در یک کلاس خاص گرفته می‌شود. هر همسایه اهمیت یکسانی در فرایند دسته‌بندی دارد؛ بنابراین نتایج به‌دست‌آمده با الگوریتم فازی در مقایسه با نتایج به‌دست‌آمده در مطالعات قبلی بهبود یافته است. هدف این پژوهش تجزیه‌وتحلیل درصدهای دسته‌بندی احساسات با استفاده از پارامترهای آماری استخراج‌شده از پایگاه‌داده عاطفی SROL است. بردارهای ویژگی شامل ۱۷ پارامتر هستند.

انگیزه استفاده از الگوریتم فازی این است که همه بردارهای الگوریتم کلاسیک اهمیت یکسانی در فرایند خوشه‌بندی دارند. هیچ اطلاعاتی برای تعیین کمیت "قدرت" عضویت در یک کلاس وجود ندارد. اطلاعات فازی، اهمیت هر بردار را در تصمیم‌گیری نهایی توصیف می‌کند. بردارهایی با مقدار بالای عضویت درجه فازی، وزن بیشتری در ایجاد کلاس نسبت به بردارهایی با درجه عضویت پایین‌تر خواهند داشت. مجموعه‌داده طراحی‌شده، با عنوان DB100 نام‌گذاری شده است. DB100 شامل ۱۴۵ فایل انتخاب‌شده از مجموعه اعتبارسنجی احساسی SROL است. از هر فایل صوتی که حاوی ۴ یا ۵ تلفظ با حالات احساسی مختلف است، فقط مصوت‌هایی استخراج شده‌اند که شامل "a"، "e"، "i"، "o"، "u"، "ă" با "a+" و "î" با "a" است. پایگاه عاطفی شامل ۱۴ سخنران مرد و ۱۱ سخنران زن است.

درصد تشخیص احساسات به‌دست‌آمده با الگوریتم فازی، بین ۶۱.۲۱ درصد برای حرف مصوت «a» و ۷۵.۸۱ درصد برای حرف مصوت «o» است. همچنین میانگین دقت تشخیص به‌دست‌آمده در این روش ۷۰ درصد است، که می‌توان گفت به دلیل انتخاب نامناسب ویژگی با آنکه دقت تشخیص افزایش داشته است اما درصدی کم نسبت به سایر پژوهش‌ها است.

در پژوهش یوآن و همکاران به دلیل نویز، محیط آکوستیک اتاق چندکاناله و چالش‌برانگیز بودن تشخیص عملی احساسات گفتار، ابتدا با استفاده از همسان‌ساز فاصله‌های کسری کور[۱۲]، پیش‌پردازش گفتار را عملی ساخته‌اند. با استفاده از این روش، تداخل نویز به طور مؤثر حذف شده و ویژگی‌های احساسی دقیق‌تری رزرو شده‌اند. استحکام نویز بهتری را می‌توان برای سیگنال گفتار پردازش‌شده توسط همسان‌ساز به دست آورد و برای تشخیص عملی احساسات گفتار مناسب‌تر است. تشخیص سنتی احساسات گفتار را می‌توان بهبود بخشید و همچنین سیگنال تقویت‌شده برای الگوریتم فازی مناسب است. نتایج تجربی نشان می‌دهد که همسان‌سازی فاصله کسری برای تشخیص عملی احساسات گفتار مؤثر بوده است. محققان این پژوهش به‌منظور تأیید اثربخشی روش پیشنهادی همسان‌ساز به همراه شبکه عصبی فازی نتایج تشخیص را با شبکه عصبی فازی منفرد مقایسه کرده است. شکل ۷ نمودار عملکرد تشخیص احساسات مربوط به مجموعه‌داده مورداستفاده بر اساس هر دو روش است.

[12] Fractional Spaced Blind Equalizer



شکل ۷. نمودار عملکرد تشخیص احساسات با/بدون همسان ساز

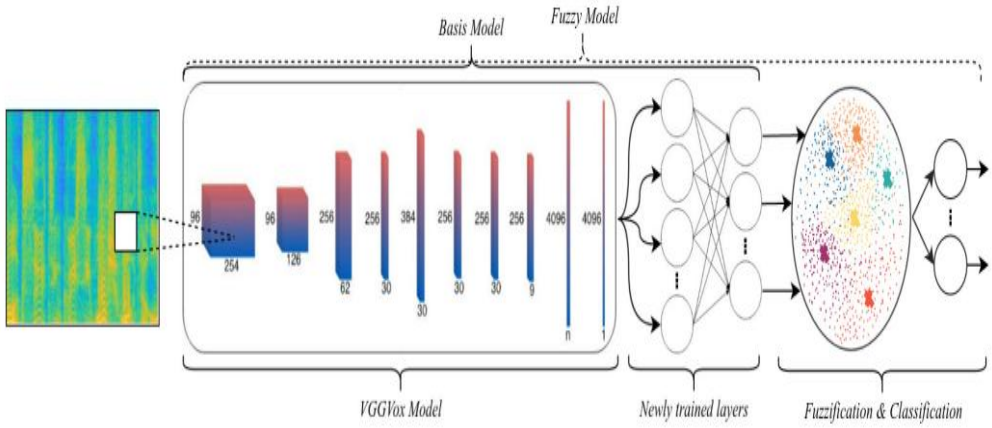
شبکه عصبی فازی مبتنی بر VGGVox

در پژوهش آسانسائو و همکاران (Assunção et el, 2020) گنجاندن یک لایه فازی واسطه در یک شبکه عصبی مبتنی بر VGGVox پیشنهاد شده است که هدف آن انتقال مه‌آلود بین حالت‌های احساسی است. این مدل عصبی-فازی در بر اساس چهار پایگاه‌داده گفتار احساسی آموزش و ارزیابی شده و بهبودهایی را در عملکرد دسته‌بندی نسبت به همتای غیرفازی خود نشان داده است. مدل VGGVox یک معماری VGG-M بر اساس شبکه عصبی پیچشی[۱۳] است که به طور خاص برای تجزیه‌وتحلیل صدا در شکل طیفی آن، با هدف نهایی شناسایی احساسات گویندگان، طراحی شده است. معماری آن به طور گسترده با بیش از ۲۰۰۰ ساعت صدا آموزش دیده و باتوجه‌به عملکرد بالای آن کاملاً قادر به استخراج ویژگی‌های صوتی بسیار قوی از طیف‌نگارهای گفتاری است. برای این منظور، فقط لایه‌های دسته‌بندی نهایی آن باید برای تشخیص احساسات جایگزین شوند، درحالی‌که بقیه لایه‌های پیچشی و تلفیقی که پیش‌تر آموزش دیده‌اند ممکن است حفظ شوند.

لایه فازی استفاده شده در این پژوهش نیز همان C-means فازی است. این لایه فازی از تعبیه داده‌های آموزشی دریافتی از لایه قبلی برای انجام خوشه‌بندی به‌گونه‌ای استفاده می‌کند که هر نمونه داده ممکن است به بیش از یک خوشه تعلق داشته باشد. علاوه بر این، باتوجه‌به اینکه چگونه تعداد خوشه‌ها ممکن است بیشتر از کلاس‌ها باشد، این فرایند خوشه‌بندی با دسته‌بندی نهایی برابری نمی‌کند. بیشترین دقتی که این لایه در خوشه‌بندی خود داشته، در خوشه‌بندی ۱۵۰ خوشه‌ای است که میزان آن ۶۸.۹۳ درصد است.

ادغام لایه فازی در معماری شبکه عصبی پیچشی به کار گرفته شده کاملاً ساده است. برای استفاده کامل از وزن‌های از پیش آموزش‌دیده VGGVox، لایه‌های میانی آن نگهداری شده و دو لایه کاملاً متصل برای کاهش ابعاد ویژگی اضافه شده است. متعاقباً، اندازه بردارهای کاهش‌یافته این لایه فازی را دریافت می‌کند و می‌تواند فازی‌سازی آن‌ها را با موفقیت انجام دهد. سپس بردارهای درجه عضویت حاصل از طریق یک لایه کاملاً متصل دیگر برای دسته‌بندی نهایی پیش می‌روند. شکل ۸ معماری کلی این سیستم را نشان می‌دهد.

[13] Nonvolutional Neural Network or CNN



شکل ۸. معماری مدل شبکه عصبی فازی مبتنی بر VGGVox

این پژوهش شامل مجموعه‌داده‌های مهمی مانند EmoDB در آلمانی، EMOVO در ایتالیایی، SAVEE در انگلیسی بریتانیایی و ELRA-S0329 در اسپانیایی است. نتایج این پژوهش در سه حالت دوره یادگیری به میزان‌های ۳۰، ۵۰ و ۱۰۰ دوره اندازه‌گیری شده است. بیشترین دقت تشخیص مربوط به حالت ۱۰۰ دوره‌ای برروی مجموعه S0329 است. جدول ۵ تمامی نتایج را در حالت‌های فازی و غیرفازی نمایش می‌دهد.

جدول ۵. جدول عملکرد تشخیص مدل VGGVox با/بدون لایه فازی

Epochs	Non-Fuzzy				Fuzzy			
	EMODB	SAVEE	EMOVO	S0329	EMODB	SAVEE	EMOVO	S0329
30	67.05% (5.61%)	58.54% (8.08%)	51.02% (8.02%)	88.38% (3.47%)	68.93% (5.35%)	61.25% (5.88%)	50.85% (7.19%)	90.61% (3.03%)
50	71.94% (5.30%)	63.54% (4.61%)	55.42% (4.55%)	88.97% (2.97%)	74.17% (5.71%)	66.04% (6.27%)	55.93% (3.26%)	90.31% (2.11%)
100	76.62% (8.41%)	68.54% (2.98%)	62.20% (7.42%)	90.91% (3.19%)	78.48% (7.86%)	71.04% (4.76%)	64.07% (7.72%)	91.51% (2.68%)

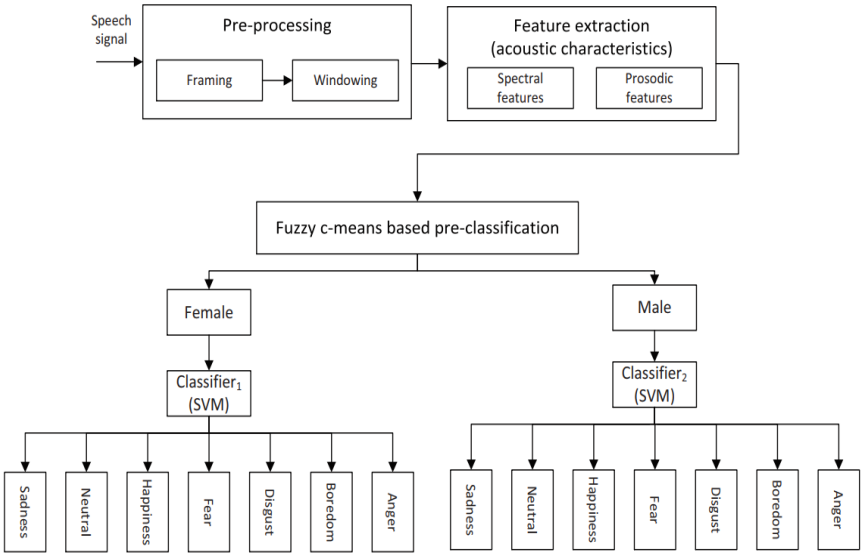
ماشین بردار پشتیبان فازی ترکیبی چندلایه[14]

هوآنگ و همکاران (Huang et el, 2021) در پژوهش خود یک مدل ماشین بردار پشتیبان فازی ترکیبی چندلایه را پیشنهاد می‌کنند که شامل سه لایه استخراج ویژگی، پیش دسته‌بندی و دسته‌بندی است. مدل MLHF-SVM چالش‌های پژوهش‌های قبل را با استفاده از C-means فازی بر اساس اطلاعات تشخیص دسته‌بندهای انسانی و چندلایه ماشین بردار پشتیبان حل می‌کند. علاوه بر این، برای غلبه بر ضعف سقوط به حداقل‌های محلی که C-means فازی دارد، یک الگوریتم بهینه‌سازی ازدحام ذرات وزن با اینرسی نمایی طبیعی پیشنهاد شده و با C-means فازی برای بهینه‌سازی ادغام می‌شود.

مجموعه‌داده‌های مورداستفاده در ارزیابی این مدل شامل سه مجموعه‌داده EmoDB، SAVEE و eNTERFACE'05 است. همان‌طور که در تمامی پژوهش‌ها قابل‌مشاهده است، به‌طورکلی، تشخیص احساسات گفتار از سه مرحله پیش‌پردازش، استخراج ویژگی و

[14] Multi-Layer Hybrid Fuzzy Support Vector Machine or MLHF-SVM

دسته‌بندی تشکیل شده است. با درنظرگرفتن اطلاعات شناسایی، این سیستم قبل از مرحله دسته‌بندی یک قسمت پیش دسته‌بندی را اضافه می‌کند. این فرایند از C-means فازی برای خوشه‌بندی پایگاه‌داده در کلاس‌های متعدد استفاده می‌کند و سپس هر یک از آن‌ها را به ترتیب در ماژول دسته‌بندی ویژگی‌ها به‌کارگیری می‌کند. در شکل ۹، فرض بر این است که C-means فازی سیگنال‌های گفتار را بر اساس جنسیت به دو دسته مرد و زن تقسیم می‌کند و دسته‌بند، احساسات را بر اساس EmoDB به هفت دسته تقسیم می‌کند.



شکل ۹. معماری مدل ماشین بردار پشتیبان فازی

نتایج به‌دست‌آمده از استفاده این مدل برای مجموعه‌داده‌های EmoDB، SAVEE و eNTERFACE’05 به ترتیب ۹۰، ۸۰.۶۶ و ۷۲.۴۷ درصد می‌باشد. در صورتی که دقت تشخیص روش بدون وجود لایه فازی و بهبوددهنده به همان ترتیب به‌صورت ۷۷.۶، ۶۵.۴ و ۵۸.۶۵ درصد است که این آمار نشان دهنده موفقیت این مدل است.

الگوریتم کرم شب‌تاب

الگوریتم کرم شب‌تاب[15] یک الگوریتم مبتنی بر جمعیت است. این الگوریتم به دنبال یافتن تابع هدف جهانی بهینه بر اساس رفتار اکتشافی کرم شب‌تاب است. در الگوریتم کرم شب‌تاب، عامل به طور تصادفی در فضای مسئله توزیع می‌شود. عوامل به نام کرم شب‌تاب شناخته می‌شوند و کیفیت نور را شدت نور می‌گویند. هر کرم شب‌تاب توسط همسایه‌های درخشان‌تر جذب می‌شود. جذابیت با افزایش فاصله بین آن‌ها کاهش می‌یابد. اگر هیچ کرم شب‌تاب روشن‌تر از بقیه نباشد، به طور تصادفی حرکت می‌کنند. در اعمال خوشه‌بندی، متغیر تصمیم‌گیری الگوریتم، خوشه‌ها هستند. هدف، مجموع فاصله اقلیدسی تمام نمونه‌های داده‌های آموزشی در فضای n بعدی است. عوامل بر اساس این تابع هدف، به‌صورت تصادفی توزیع شده و در ابتدا کمی‌سازی خواهند شد.

در پژوهش اقسانورد و رستمی (Aghsanavard and Rostami, 2016) سیگنال‌های ورودی به یک سیستم فازی داده شده و این سیگنال‌ها برای تشخیص بهتر سیگنال‌های گفتار احساسی، خوشه‌بندی می‌شوند. فرایند خوشه‌بندی بر اساس الگوریتم کرم شب‌تاب انجام می‌شود. سپس سیگنال خوشه‌بندی را با کلاس نویز تولیدشده، رفع نویز شده تا مدل هرکدام از سیگنال‌های نهایی احساسات

[15] Firefly Algorithm

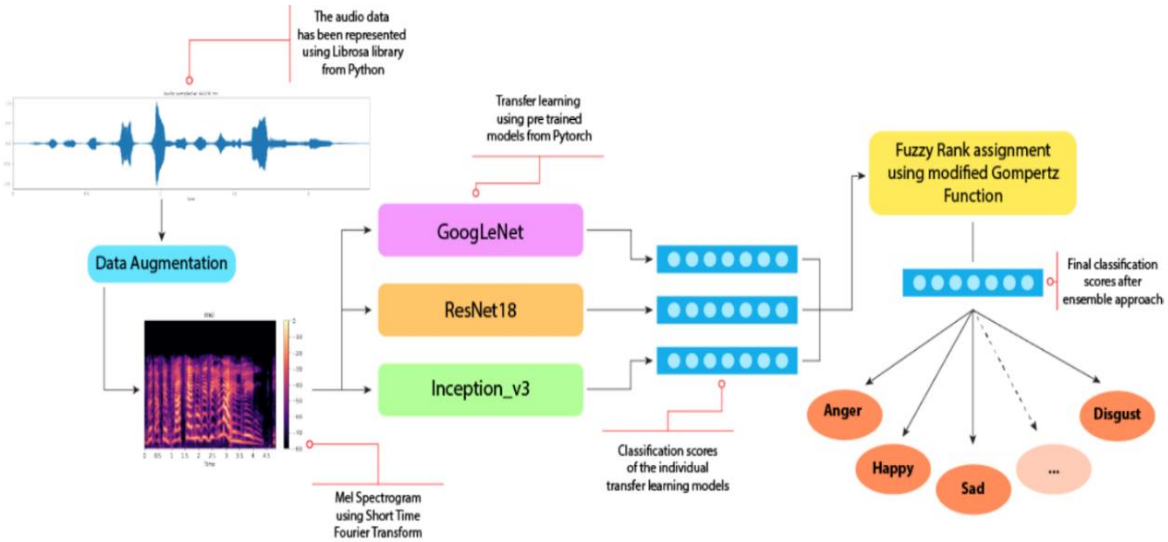
متعلق به گفتار را بدون توجه به نویز تشخیص دهد. این روش ترکیبی علاوه بر افزایش دقت، سرعت بسیار بالایی در تشخیص احساسات گفتار دارد.

اثر کلی مبتنی بر رتبه فازی

در پژوهش ساهو و همکاران (Sahoo et el, 2021) مدلی معرفی شده است که TLEFuzzyNet نام دارد. این مدل شامل سه خط برای تشخیص است که هرکدام به‌صورت یکسان ویژگی‌های طیفی-زمانی را استخراج می‌نمایند . ویژگی‌های استخراج شده به‌عنوان ورودی به‌صورت جداگانه به سه مدل از پیش آموزش‌دیده پیچشی انتقال می‌یابند. این سه مدل ResNet18، Inception_v3 و GoogleNet نام دارند. پس از تشخیص‌های انجام‌شده توسط این سه مدل، یک تابع فازی به میان می‌آید.

تابع گومپرتز[16]، یک تابع سیگموئید در حوزه زمان است که در ابتدا و انتهای یک دوره معین به کندترین حالت اشباع می‌شود. تابع در ابتدا برای مرگ‌ومیر انسان مدل‌سازی شد، زیرا مرگ‌ومیر به طور تصاعدی با افزایش سن، افزایش می‌یابد، پس از آن به طور جانبی اشباع می‌شود. به امتیازهای پیش‌بینی سه مدل در مجموعه آزمایشی با استفاده از تابع گومپرتز، رتبه‌های فازی داده شد که دقت بهتری نسبت به مدل‌های تشکیل‌دهنده ارائه می‌دهد. این بدین معناست که پس انجام تشخیص توسط سه مدل، نتایج توسط تابع فازی، فازی‌سازی می‌شود و بر اساس اشتراک میان نتایج تشخیص هر سه، تشخیص نهایی تعیین می‌گردد. شکل ۱۰ معماری کلی مدل معرفی شده در این پژوهش را بیان می‌نماید.

این مدل پیشنهادی، بر روی سه مجموعه‌داده SAVEE، RAVDESS و EmoDB آزمایش شده و نتایج بسیار شگفتی‌آوری به‌دست‌آمده است. دقت به‌دست‌آمده برای هر سه مجموعه‌داده به ترتیب ۹۸.۵۷، ۹۹.۶۶ و ۹۹.۳۸ است. البته می‌توان ذکر نمود که علاوه بر وجود یک تابع فازی، مدل‌های معرفی‌شده علاوه بر آموزش بر روی مجموعه‌داده‌ها، از پیش نیز آموزش‌دیده بودند که خود تأثیر بسزایی در دقت تشخیص دارد.



شکل ۱۰. معماری کلی مدل TLEFuzzyNet

[16] Gompertz

بحث و نتیجه‌گیری

استفاده از سیستم‌های فازی در تشخیص احساسات گفتار، به‌عنوان یک رویکرد قدرتمند و انعطاف‌پذیر، در سال‌های اخیر موردتوجه بسیاری از پژوهشگران قرار گرفته است. این سیستم‌ها با توانایی مدل‌سازی عدم قطعیت و ابهام در داده‌ها، به‌ویژه در حوزه‌هایی مانند تشخیص احساسات که ماهیتی ذهنی و پیچیده دارند، عملکرد قابل‌توجهی از خود نشان داده‌اند. در این پژوهش‌ها، روش‌های مختلفی از جمله خوشه‌بندی فازی، شبکه‌های عصبی فازی، و ترکیب آن‌ها با الگوریتم‌های بهینه‌سازی و یادگیری عمیق مورد بررسی قرار گرفته‌اند. هر یک از این روش‌ها مزایا و چالش‌های خاص خود را دارند که در ادامه به طور خلاصه بررسی می‌شوند.

سیستم‌های فازی به دلیل توانایی در مدیریت داده‌های مبهم و نادقیق، به‌ویژه در تشخیص احساسات که ذاتاً پیچیده و چندبعدی هستند، بسیار مؤثر عمل می‌کنند. این ویژگی باعث می‌شود که سیستم‌های فازی بتوانند حالات احساسی را با دقت بیشتری تشخیص دهند. به‌علاوه روش‌های فازی قابلیت ادغام با سایر روش‌های یادگیری ماشین و یادگیری عمیق را دارند. برای مثال، در پژوهش‌هایی مانند پژوهش ساهو و همکاران، ترکیب شبکه‌های عصبی عمیق با توابع فازی منجر به دقت بسیار بالایی در تشخیص احساسات شده است. در بسیاری از پژوهش‌ها، استفاده از سیستم‌های فازی منجر به افزایش دقت تشخیص احساسات شده است. برای مثال، در پژوهش ژانگ و همکاران، دقت تشخیص احساسات به بیش از ۸۵ درصد رسید و در پژوهش ساهو و همکاران، دقت تشخیص در برخی مجموعه‌داده‌ها به بیش از ۹۹ درصد افزایش یافت. هنگامی که سیستم‌های فازی با الگوریتم‌های بهینه‌سازی مانند الگوریتم کرم شب‌تاب ترکیب می‌شوند، توانایی بالایی در کاهش تأثیر نویز و بهبود کیفیت تشخیص دارند. این ویژگی در پژوهش اقسانورد و رستمی به‌خوبی نشان داده شده است.

باید افزود که علاوه بر مزیت‌های این سیستم‌ها، محدویت‌های نیز وجود دارد. برخی از روش‌های فازی، به‌ویژه هنگامی که با شبکه‌های عصبی عمیق ترکیب می‌شوند، ممکن است از نظر محاسباتی پیچیده و زمان‌بر باشند. این موضوع می‌تواند استفاده از این روش‌ها را در سیستم‌های بلادرنگ با محدودیت مواجه کند. همچنین عملکرد سیستم‌های فازی تا حد زیادی به کیفیت و کمیت داده‌های آموزشی وابسته است .در برخی پژوهش‌ها، مانند پژوهش رووتا، انتخاب نامناسب ویژگی‌ها منجر به کاهش دقت تشخیص شده است. در برخی پژوهش‌ها، مانند پژوهش اسائو و همکاران، میزان دقت سیستم به‌صورت کمی گزارش نشده است که این موضوع ارزیابی و مقایسه روش‌ها را دشوار می‌کند.

جمع‌بندی

به‌طورکلی، سیستم‌های فازی به‌عنوان یک ابزار قدرتمند در تشخیص احساسات گفتار، توانایی بالایی در مدیریت عدم قطعیت و بهبود دقت تشخیص دارند. بااین‌حال، برای دستیابی به نتایج بهتر، نیاز به ترکیب این سیستم‌ها با روش‌های بهینه‌سازی و یادگیری عمیق وجود دارد. همچنین، توجه به چالش‌هایی مانند پیچیدگی محاسباتی و وابستگی به داده‌های آموزشی، می‌تواند زمینه‌ساز پژوهش‌های آینده در این حوزه باشد. در نهایت، می‌توان گفت که سیستم‌های فازی به‌عنوان یک رویکرد مکمل، نقش مهمی در پیشبرد فناوری‌های تشخیص احساسات ایفا می‌کنند.

منابع

Rovetta, Stefano, et al. "Emotion recognition from speech signal using fuzzy clustering." 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019). Atlantis Press, 2019.

Demircan, Semiye, and Humar Kahramanli. "Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech." Neural Computing and Applications 29.8 (2018): 59-66.

Esau, Natascha, Lisa Kleinjohann, and Bernd Kleinjohann. "An Adaptable Fuzzy Emotion Model for Emotion Recognition." EUSFLAT Conf. 2005.

Zhang, Zhenxing, and Joon S. Lim. "Emotion Recognition Algorithm Based on Neural Fuzzy Network and the Cloud Technology." 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA). IEEE, 2015.

Zbancioc, Marius, and Silvia Monica Feraru. "Emotion recognition of the SROL Romanian database using fuzzy KNN algorithm." 2012 10th International Symposium on Electronics and Telecommunications. IEEE, 2012.

Yuan, Tao, Chunhong Deng, and WangYang Shi. "Speech emotion recognition based on Fuzzy K-NN algorithm with fractionally spaced blind equalization." 2016 2nd Workshop on Advanced Research and Technology in Industry Applications (WARTIA-16). Atlantis Press, 2016.

Assunção, Gustavo, and Paulo Menezes. "Intermediary fuzzification in speech emotion recognition." 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2020.

Huang, S., Dang, H., Jiang, R., Hao, Y., Xue, C., & Gu, W. (2021). Multi-layer hybrid fuzzy classification based on SVM and improved PSO for speech emotion recognition. Electronics, 10(23), 2891.

Aghsanavard, Daniar, and Vahid Rostami. "Speech Emotion Recognition Using Fuzzy Logic Classifier." International Journal of Advanced Networking and Applications 7.4 (2016): 2817.

Sahoo, Karam Kumar, et al. "TLEFuzzyNet: fuzzy rank-based ensemble of transfer learning models for emotion recognition from human speeches." IEEE Access 9 (2021): 166518-166530.

A Review on Emotion Recognition in Speech Based on Fuzzy Inference

Mohammad Mahdi Mokhtari
Imam Hossein University
Mohammad Ali Javadzade
Imam Hossein University
Meysam Mirzai
Imam Hossein University

Abstract

Humans make decisions based on both reason and emotion throughout their lives. Emotions are a fundamental aspect of being human. These emotions are expressed through behavior and speech. In recent decades, emotions have been predicted and detected by computers using speech recognition and signal processing. Effective systems in the field of emotion detection are increasingly attracting the attention of researchers and advanced technology companies to align with or adapt to the user's mood. A key issue in this regard is selecting methods appropriate to the problems under investigation, considering their advantages and disadvantages. Emotion detection from speech signals has primarily been achieved using supervised classifiers. However, clustering techniques appear to be suitable for solving such problems, especially in large datasets where labeling speech can be a tedious and challenging task. After deciding on the type of technique to use and conducting experiments, researchers have found that most techniques require better and more accurate detection. Fuzzy systems are among the most useful optimization methods in this field. Fuzzy systems can also be applied in both classification and clustering scenarios. In this article, we discuss several applications of fuzzy systems in speech emotion detection.

Keywords: Emotion Recognition, Signal Processing, Fuzzy Systems, Optimization