آشکارسازی وجود استرس در گفتار با استفاده از ویژگی‌های مبتنی بر مدل غیرخطی تولید گفتار

شهلا ترابی[۱،۲]، فرشاد الماس گنج[۱] ، امین محمدیان[۲]
۱- دانشگاه صنعتی امیرکبیر، دانشکده‌ی مهندسی پزشکی
۲ – پژوهشکده‌ی پردازش هوشمند علائم

نویسنده‌ی عهده‌دار مکاتبات: شهلا ترابی

چکیده

مطالعات نشان می‌دهند که استرس روانی گوینده در نحوه‌ی تولید گفتار او اثر می‌گذارد. آشکارسازی وجود استرس در گفتار دارای کاربردهای متعدد می‌باشد. در کارهای اخیر، ویژگی‌های صوتی مختلف، به‌صورت جداگانه، توسط طبقه‌بندی‌کننده‌ی HMM مورد ارزیابی قرار گرفته‌اند و از میان آن‌ها، ویژگی غیرخطی TEO-CB-Auto-Env کارآمدترین پارامتر بوده است. در این مقاله، یک ویژگی جدید که آن را (TEO-Pch-LFPC) می‌نامیم، پیشنهاد شده است. دادگان گفتار استرسی (حالات خنثی، عصبانی، بلند و لمبارد) از پایگاه داده‌ی SUSAS برداشته شده و نقطه‌ی قوّت کار حاضر این است که در آن، از طبقه‌بندهای ساده‌تری نسبت به HMM استفاده شده است، یعنی طبقه‌بندهای استاتیک (KNN، LDA و SVM)، و روش ارزیابی نیز RRM می‌باشد. با استفاده از ویژگی TEO-Pch-LFPC و طبقه‌بند SVM، در تفکیک دو حالته، درصد صحت ۹۳.۷۸٪ و در طبقه‌بندی چندحالته ۷۰.۲۲٪ می باشد.

کلید واژه: آشکارسازی استرس، آشکارسازی احساس، پردازش گفتار

۱- مقدمه

یکی از کارکردهای گفتار، انتقال مفاهیم غیرزبانی و حالت روحی شخص است؛ طوری که گاهی نقش احساسات در برقراری ارتباط، فراتر از اطلاعات منطقی منتقل می‌شود [۱]. هنگامی که یک لغت توسط یک فرد تلفّظ می‌شود، بسته به این‌که چگونه ادا شود معانی مختلفی خواهد داشت؛ به‌طور مثال، کلمه‌ی «really» در زبان انگلیسی، می‌تواند معانی مختلفی داشته باشد (سؤالی باشد، مفهوم «تحسین‌کردن» را داشته باشد و یا با حالت تعجّب بیان گردد). بنابراین فهمیدن مفهوم لغوی یک کلمه به تنهایی، برای تعبیر معنای آن کافی نیست. برقراری ارتباط را از طریق ارسال و دریافت پیام‌های غیرزبانی، یا به عبارتی «پیام‌های غیروابسته به کلمات»، ارتباط غیرکلامی[۱] (NVC) می‌گویند [۲]. در این زمینه، یکی از مسائلی که در سال‌های اخیر به‌طور تخصّصی مورد توجّه قرار گرفته، بازشناسی استرس از گفتار شخص است.

در تعاملات انسان و رایانه طبیعی‌تر است سعی کنیم با رایانه‌ها همان‌گونه که با انسان‌ها رابطه برقرار می‌کنیم، ارتباط برقرار کنیم. برای این منظور رایانه‌ها باید به گونه‌های احساس و استرس انسانی حساسیت داشته باشند. مطالعاتی که در مورد تعاملات انسان و رایانه انجام شده است، نشان می‌دهد که آشکارسازی استرس گوینده، حاوی اطّلاعات مهمّی می‌باشد [۳]. اگر بتوان به طریقی وجود استرس و انواع آن را تعیین کرد، آن‌گاه این اطّلاعات می‌توانند با سیستم‌های بازشناسی گفتار تلفیق شوند و عملکرد این سیستم‌ها را بهبود بخشند. همچنین، آشکارسازی استرس می‌تواند کاربرد پزشکی، نظامی و یا قضایی (دروغ سنجی) داشته باشد [۱].

به‌طور کلی، استرس به یکی از حالت‌های زیر یا ترکیبی از آن‌ها اطلاق می‌شود [۴] :"احساسات منفی (عصبانیت، ترس، تاسف)، اجبار برای انجام یک وظیفه در یک زمان محدود و شرایط محیطی نامطلوب (مقدار زیاد نویز پس‌زمینه)". در این میان حالات عصبانی، بلند و اثر لمبارد

[۱] Non-Verbal Communication

پرکاربردترین دادگان گفتار استرسی مورد استفاده در مقالات هستند [۱، ۳، ۴، ۵]. اثر لمبارد به شرایطی گفته می‌شود که شخص در یک محیط نویزی قرار دارد و سعی می‌کند حالت تولید گفتارش را به‌گونه‌ای تغییر دهد که صدایش بهتر به گوش شنونده برسد [۵].

محققان سعی کرده‌اند با بررسی متغیرهای صوتی از قبیل فرکانس پایه، دامنه، شدّت، تراکم انرژی طیفی، مکان سازه‌ها[1] [۳، ۴، ۵] و موارد دیگر، شاخص‌های قابل اعتمادی را برای استرس تعیین نمایند (اکثر پارامترهای نامبرده، مبتنی بر تئوری خطی تولید گفتار هستند). در تئوری خطی فرض می‌شود که جریان هوا از داخل تارهای صوتی به‌صورت یک موج خطّی منتشر می‌شود و انقباض یا حرکت تارهای صوتی، به‌عنوان منبع تولید گفتار شناخته می‌شوند.

ارزیابی‌ها نشان می‌دهند که ویژگی‌های مبتنی بر مدل خطی تولید گفتار، همواره برای آشکارسازی استرس مناسب نیستند و نشان داده شده است که از بین این ویژگی‌های خطّی، ویژگی گام گفتار بهترین عملکرد را دارد. از طرف دیگر، مطالعات *تیگر* [۱] نشان می‌دهند که در داخل ناحیه‌ی صوتی، جریان هوا به‌طور یکنواخت در تارهای صوتی منتشر نمی‌شود، بلکه به‌صورت تفکیک شده درمی‌آید و در نزدیکی دیواره‌ها متمرکز می‌گردد؛ به این ترتیب گرداب‌هایی در سرتاسر ناحیه‌ی صوتی پدیدار می‌شوند. منبع تولید صوت، تعاملات بین جریان این گرداب‌هاست و ماهیت غیرخطّی دارد. از آنجایی‌که تغییرات فیزیولوژیک ناشی از استرس، روی الگوی تعاملات بین جریان گرداب‌ها تأثیر می‌گذارند، برای تفکیک گفتار خنثی و استرسی، نیاز به ویژگی‌های غیرخطّی داریم [۳].

تیگر یک اپراتور انرژی به‌نام TEO[2] را به‌صورت زیر ارائه می‌کند که در آن $x(n)$ سیگنال گفتار نمونه‌برداری شده است:

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (۱)$$

در مرجع [۱] سه ویژگی غیرخطّی که از اپراتور TEO مشتق شده‌اند برای طبقه‌بندی استرس، پیشنهاد شده‌است که از بین آن‌ها ویژگی TEO-CB-Auto-Env[3] به‌عنوان کارآمدترین پارامتر شناخته می‌شود. اگرچه این ویژگی مبتنی بر TEO می‌تواند در تفکیک دو حالته (گفتار خنثی و یکی از حالات استرسی) عملکرد مناسبی داشته باشد، امّا مشاهده می‌شود که در طبقه‌بندی چندحالته‌ی استرس، درصد صحت طبقه‌بندی، به‌شدت افت می‌کند.

هنگامی که یک کلمه در حالات مختلف استرسی بیان می‌شود سیگنال، هم در حوزه‌ی زمان و هم در حوزه‌ی فرکانس دستخوش تغییر می‌گردد. یک معیار ممکن برای اندازه‌گیری استرس موجود در گفتار، بررسی توزیع انرژی طیفی در باندهای فرکانسی مختلف است. فرض می‌شود سیستم شنوایی انسان یک فرآیند فیلترینگ است که در آن کل محدوده‌ی فرکانسی قابل شنیدن به تعداد زیادی باندهای بحرانی قسمت‌بندی می‌شود [۱]. بنابراین توزیع انرژی در لگاریتم باندهای فرکانسی مختلف می‌تواند یک ویژگی مناسب برای طبقه‌بندی استرس باشد. به‌همین دلیل ویژگی‌های زیرباندی مبتنی بر FFT می‌توانند برای طبقه‌بندی استرس مناسب باشند؛ زیرا تبدیل فوریه (بدون این‌که به زمان وابستگی داشته باشد) خطی بودن در رزولوشن فرکانسی را حفظ می‌کند. ضرایب LFPC[4] برای این منظور مناسب‌اند [۶،۷].

در مراجع مختلف [۳، ۸]، ویژگی‌های طیفی حاصل از تجزیه‌ی مبتنی بر ویولت، به‌عنوان نشان‌گرهایی از استرس معرفی شده‌اند. امّا، این ویژگی‌ها وابسته به زمان هستند و بیشتر برای کاربرد بازشناسی گفتار مناسب‌اند تا طبقه‌بندی استرس. چراکه در بازشناسی لغات، توالی زمانی آواها در کلمه مهم است امّا در یک حالت استرسی (مثلا عصبانیت)، توالی زمانی خاصّی در سیگنال مد نظر نیست. به‌طور مثال، اگر بلندی صدای مرتبط با عصبانیت را در نظر بگیریم هیچ زمان مشخّص و ثابتی در کلمه موجود نیست که این بلندی در آن زمان خاص باشد. این رویداد می‌تواند در ابتدا، وسط یا انتهای کلمه رخ دهد. تا زمانی‌که بلندی صدا هست، حالت عصبانیت توصیف می‌شود [۳].

در مقالات سه روش برای طبقه‌بندی استرس موجود می‌باشد [۲، ۹]:

۱- روش مبتنی بر شبکه‌های عصبی مصنوعی (ANNs)
۲- مدل مخفی مارکوف چند کاناله
۳- ترکیب مدل‌های مخفی مارکوف

واماک و هانسن[5] [۹] برای طبقه‌بندی استرس از روش شبکه‌ی عصبی استفاده کرده‌اند و این‌گونه گزارش کرده‌اند

[1] Formant
[2] Teager Energy Operator
[3] Critical band based TEO autocorrelation envelope area
[4] Log-Frequency Power Coefficients
[5] Womack , Hansen

که، عملکرد طبقه‌بندی استرس توسط شبکه‌ی عصبی، از طریق اندازه‌ی فاصله‌ی تفکیک‌پذیری ویژگی‌هاست و با تغییر اندازه‌ی لغت و نیز تغییر شخص گوینده، نتیجه‌ی این عملکرد، به‌طور جدّی تغییر می‌کند؛ و با توجّه به عملکرد ضعیف این طبقه‌بند در آشکارسازی استرس، لازم است تست‌ها به‌صورت وابسته به متن و وابسته به گوینده انجام گردند. درنتیجه، روش شبکه‌ی عصبی در این زمینه کارآیی چندانی ندارد؛ و دو روش طبقه‌بندی دیگر هم دینامیک هستند [۳].

در کار حاضر، سعی داریم از طبقه‌بندهای دیگری که تاکنون در این زمینه مورد استفاده قرار نگرفته‌اند (یعنی طبقه‌بندهای استاتیک (KNN، LDA و SVM)) استفاده کنیم. لازم به ذکر است که پیاده‌سازی و کاربرد طبقه‌بندهای استاتیک، ساده‌تر و سریع‌تر است.

در کار جاری، در درجه‌ی اول، تلاش ما در جهت تعیین یک مجموعه ویژگی مفید و کاربردی است که قادر باشد «وجود استرس در گفتار» را تشخیص دهد و درضمن، ایده‌آل خواهد بود اگر بتوانیم حالات مختلف استرسی را در سیگنال‌های گفتاری از هم تفکیک نماییم. ویژگی پیشنهاد شده، TEO-Pch-LFPC نام دارد و توسط آن عملکرد توأم پارامترهای (۱)گام گفتار، (۲) TEO-CB-Auto-Env و (۳) ضرایب LFPC را در آشکارسازی استرس روانی گوینده بررسی می‌کنیم. اگرچه اکثر مطالعات اخیر از طبقه‌بند مبتنی بر HMM که نسبتاً پیچیده است استفاده کرده‌اند [۱، ۳، ۵، ۱۰]، امّا از آن‌جایی‌که تمرکز ما روی انتخاب بهترین مجموعه ویژگی کاربردی است، ابتدا از دو طبقه‌بند معمول LDA و KNN که خوش‌رفتار هستند و پیاده‌سازی آن‌ها ساده است، استفاده می‌کنیم [۱۱]؛ سپس، نتایج طبقه‌بندی را، به‌ازای طبقه‌بند SVM که قوی‌تر است، بیان خواهیم کرد.

کار حاضر به‌صورت زیر سازماندهی شده است:

در بخش ۲ مفهوم استرس از دیدگاه پردازش گفتار استرسی تعریف می‌شود. بخش ۳ پایگاه داده مورد استفاده را توضیح می‌دهد. بخش ۴ ویژگی‌های صوتی منفرد را برای طبقه‌بندی استرس توصیف می‌کند. بخش ۵ این ویژگی‌ها را ترکیب می‌کند تا به یک بردار ویژگی کارآمدتر دست یابد. بخش ۶ روی نتایج به‌دست آمده از بخش ۵ بحث می‌کند و آن‌ها را با بهترین نتایج کارهای قبلی مقایسه می‌کند. در نهایت در بخش ۷، پیشنهادهایی برای کارهای آینده ارائه می‌گردد.

۲- تعریف استرس

از دیدگاه پردازش گفتار استرسی، استرس به حالتی گفته می‌شود که در آن تولید گفتار یک گوینده از حالت طبیعی خارج شود [۱۳،۱۲،۳]. در مقابل، اگر گوینده در یک محیط آرام و بدون هیجان باشد، فشار کاری نداشته باشد و دچار احساسات و یا اضطراب نشده باشد، آنگاه گفتار تولید شده حالت طبیعی خواهد داشت. با این تعریف، دو حوزه‌ی استرسی پدیدار می‌شوند [۱۳]: "ادراکی" و "فیزیولوژیکی".

در استرس ادراکی، شخص از طریق مشاهده درمی‌یابد که محیط پیرامونش از حالت طبیعی خارج شده و به‌همین دلیل، "آهنگ" گفتارش نسبت به حالت خنثی تغییر می‌کند. دلایل استرس ادراکی عبارتند از احساسات، نویز محیطی (اثر لمبارد) و استرس کاری.

استرس فیزیولوژیکی نتیجه‌ی تأثیر یک عامل فیزیکی روی بدن شخص است که موجب می‌شود آهنگ تولید گفتار او از حالت خنثی خارج شود. عوامل ایجاد کننده‌ی این نوع استرس عبارتند از: لرزش بدن، تعاملات شیمیایی داروها، بیماری، اثر G-Force (تأثیر جاذبه‌ی زمین روی بدن شخص در فرود یا صعود ناگهانی) و یا سنگینی هوا.

همان‌طور که می‌دانیم در علم گفتار، به تأکیدی هم که روی سیلاب‌های یک کلمه می‌شود، استرس گفته می‌شود. برای این‌که این دو نوع استرس، با هم اشتباه نشوند در تحقیقات از عبارت «گفتار تحت استرس[1]» استفاده می‌شود. مفهوم «تحت استرس» این است که فشاری روی گوینده وجود دارد و این فشار روی فرآیند تولید گفتارش اثر می‌گذارد.

۳- پایگاه داده

در این مطالعه ارزیابی‌های صورت گرفته به‌منظور طبقه‌بندی استرس، روی پایگاه داده SUSAS[2] [۱۲] انجام شده‌اند. این پایگاه داده قبلاً توسط *هانسن* جمع‌آوری شده است. تمام کلمات با یک A/D 16 بیتی با نرخ نمونه‌برداری ۸کیلو هرتز، نمونه‌برداری شده‌اند. ما ارزیابی‌های خود را روی بخش شبیه‌سازی شده SUSAS انجام می‌دهیم. علّت این امر این است که استرس ظاهر شده در شرایط واقعی شدیدتر و آسان‌تر قابل تشخیص می‌باشد [۱]؛ یعنی با همین روند آشکارسازی، در شرایط واقعی به درصد صحت بالاتری دست خواهیم یافت، جزئیات بیشتر در [۱] آمده است.

[1] Speech under stress
[2] Speech Under Simulated and Actual Stress

دلیل استفاده از پایگاه داده‌ی SUSAS این است که در حال حاضر این مجموعه، به‌عنوان استانداردترین و پرکاربردترین داده‌ی گفتاری استرس‌دار مورد استفاده قرار می‌گیرد. مزیتی که این پایگاه داده دارد این است که اکثر مقالات مرتبط با موضوع «پردازش گفتار استرس‌دار»، تحقیقات خود را روی این دادگان انجام داده‌اند و این امر ما را قادر می‌سازد که بتوانیم نتایج خود را با کارهای قبلی مقایسه کنیم.

در حوزه‌ی شبیه‌سازی شده هر کلمه توسط هر گوینده دو بار گفته شده است. از آن‌جایی که عملکرد اپراتور TEO برای واکدارها بهتر از بی‌واک‌ها است در هر مورد فقط، از بخش واکدار هر کلمه استفاده می‌شود [۵،۱۲].

۴- استخراج ویژگی

۴-۱- TEO-CB-Auto-Env:

همان‌طور که اشاره شد، به‌طور تجربی فرض می‌شود که سیستم شنوایی انسان یک فرآیند فیلترینگ است و کل محدوده‌ی فرکانسی قابل شنیدن را به تعداد زیادی باندهای بحرانی قسمت‌بندی می‌کند [۱]. بر پایه‌ی این فرضیه، ویژگی TEO-CB-Auto-Env از یک بانک فیلتر با ۱۶ باند بحرانی، برای سیگنال گفتار استفاده می‌کند و سپس پردازش TEO انجام می‌شود [۱]. هر فیلتر موجود در بانک فیلتر، یک فیلتر گابور میان‌گذر است که پهنای باند RMS مؤثر آن به اندازه‌ی باند بحرانی مربوطه است (شکل ۱). دلیل انتخاب فیلتر گابور این است که این فیلتر هم در حوزه‌ی زمان و هم در حوزه‌ی فرکانس، به‌صورت بهینه فشرده شده است. شکل گوسین $H(f)$ مانع از تولید لوب‌های کناری (بزرگ) می‌شود [۱۴]. پاسخ ضربه و پاسخ فرکانسی این فیلتر در حالت پیوسته به‌صورت زیر تعریف می‌شوند [۱۵]:

$$g(t) = \exp(-\alpha^2 t^2)\cos(2\pi f_c t) \qquad (۲)$$

$$G(f) = \frac{\sqrt{\pi}}{2\alpha}[\exp(-\frac{\pi^2 (f-f_c)^2}{\alpha^2}) + \exp(-\frac{\pi^2 (f+f_c)^2}{\alpha^2}] \qquad (۳)$$

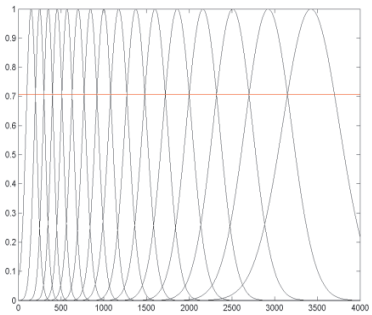
f_c فرکانس مرکزی فیلتر و α پارامتری است که پهنای باند را تنظیم می‌کند (۱۶ مقدار دارد):

$$\alpha_1 = 266, \alpha_2 = 266, \alpha_3 = 266, \alpha_4 = 294, \alpha_5 = 321,$$
$$\alpha_6 = 374, \alpha_7 = 401, \alpha_8 = 427, \alpha_9 = 507, \alpha_{10} = 561,$$
$$\alpha_{11} = 641, \alpha_{12} = 747, \alpha_{13} = 854, \alpha_{14} = 1014,$$
$$\alpha_{15} = 1201, \alpha_{16} = 1467,$$

با جایگزینی t توسط nT، معادله‌ی $g(t)$ گسسته‌سازی می‌شود (T پریود نمونه‌برداری است). $g(n)$ یک فیلتر FIR متقارن در فاصله‌ی $-N \le n \le N$ است که با سیگنال گفتار، کانوالو می‌شود [۱۴]:

$$g(n) = \begin{cases} \exp(-b^2 n^2)\cos(\Omega_c n), & |n| \le N \\ 0 & |n| > N \end{cases} \qquad (۴)$$

در این رابطه، $b = \alpha t$ و $\Omega_c = 2\pi f_c$ است. N طوری انتخاب می‌شود که پوش گوسین $g(n)$ در $n = N$، لزوماً به صفر برسد. معادله‌ی $\exp(-b^2 N^2) \approx 10^{-5}$ یک انتخاب مناسب برای N به‌دست می‌دهد.



شکل۱: دامنه‌ی نرمالیزه شده‌ی پاسخ فرکانسی فیلترهای گابور [۱]

برای استخراج بردار ویژگی TEO-CB-Auto-Env، هر پروفایل TEO به فریم‌هایی به طول ۲۵ میلی‌ثانیه، با هم‌پوشانی ۱۲.۵ میلی‌ثانیه بین فریم‌های مجاور، بخش‌بندی می‌شود. برای هر فریم زمانی، M تا مساحت نرمالیزه شده زیر پوش بالایی خودهمبستگی TEO، استخراج می‌شود (یعنی برای هر باند بحرانی یکی) و M تعداد کل باندهای بحرانی است (M=16). در محاسبه‌ی این ویژگی نیازی به تخمین F_0 نمی‌باشد. روابط ریاضی مربوط به فرآیند فوق به صورت زیر خلاصه می‌شوند:

$$u_j(n) = s(n) * g_j(n), \quad j = 1,2,3,...,16 \qquad (۵)$$

$s(n)$ سیگنال گفتار، $u_j(n)$ خروجی هر فیلتر گابور میانگذر و "*" اپراتور کانولشن می‌باشد. $\Psi_j(n)$، یعنی

پروفایل TEO مربوط به باند بحرانی jام مطابق تعریف، به‌صورت زیر است:

$$\Psi_j(n) = \Psi[u_j(n)] = u_j^2(n) - u_j(n-1)u_j(n+1) \quad (۶)$$

و

$$R_{\Psi_j^{(i)}(n)}(k) = \sum_{n=1}^{N-k} \Psi_j^{(i)}(n)\,\Psi_j^{(i)}(n+k) \quad (۷)$$

تابع خودهمبستگی فریم iام از $\Psi_j^{(i)}(n)$ می‌باشد. به این ترتیب بردار ویژگی TEO-CB Auto-Env به‌ازای هر فریم، ۱۶ پارامتر دارد.

۴-۲- مقدار متوسط گام گفتار:

نشان داده شده است [۱] که مقدار متوسط گام گفتار در شرایط غیرخنثی به‌طور محسوسی تغییر می‌کند. ما از مقدار متوسط گام گفتار به‌عنوان یک ویژگی در آشکارسازی استرس گوینده، استفاده می‌کنیم. به‌ازای هر واکه، مقدار متوسط گام گفتار به‌روش Cross-Correlation محاسبه می‌شود.

۴-۳- LFPC:

ویژگی‌های LFPC مبتنی برFFT [۶،۷،۱۶]، برای شبیه‌سازی مشخّصه‌های فیلترینگ لگاریتمی سیستم شنوایی انسان، با استفاده از اندازه‌گیری باندهای طیفی، طراحی شده‌اند. ابتدا سیگنال گفتار به پنجره‌های زمانی زمان کوتاه ۲۵ میلی ثانیه، بخش‌بندی می‌شود. سپس محتوای فرکانسی هر فریم با استفاده از روش FFT محاسبه می‌گردد. این طیف توان به یک بانک فیلتر با فرکانس لگاریتمی داده می‌شود. در [۱۷]، ایده‌ی طراحی یک بانک فیلتر ۱۲ بانده ارائه می‌شود که برای شکستن سیگنال گفتار به باندهای شنوایی، با گوش انسان مطابقت دارد. فرکانس مرکزی f_i و پهنای باند b_i به‌صورت زیر تعریف می‌گردند:

$$\begin{aligned} &b_1 = C \\ &b_i = \alpha b_{i-1},\ 2 \le i \le 12 \\ &f_i = f_i + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_j)}{2} \end{aligned} \quad (۸)$$

به ازای $F_s = 8000$ (فرکانس نمونه‌برداری پایگاه داده مورد استفاده) مقادیر $f_1 = 100, C = 24, \alpha = 1.3$ هستند. این بانک فیلتر از طریق ۱۲ فیلتر میان‌گذر با محدوده‌ی فرکانسی ۱۰۰ هرتز تا فرکانس نایکوئیست (نصف فرکانس نمونه‌برداری)، سیگنال ورودی را به تعداد زیادی خروجی می‌شکنند. انرژی در بانک فیلتر mام از طریق معادله‌ی زیر محاسبه می‌گردد:

$$S_t(m) = \sum_{k=f_m-b_m/2}^{k=f_m+b_m/2} (X_t(k))^2 \ ,\ m = 1,2,...,12. \quad (۹)$$

که در آن $X_t(k)$، kامین مؤلّفه‌ی طیفی سیگنال قطعه‌بندی شده است. t تعداد فریم‌ها و $S_t(m)$ خروجی بانک فیلتر mام است و f_m و b_m به‌ترتیب فرکانس مرکزی و پهنای باند زیرباند mام می‌باشند. انرژی در خروجی بانک فیلتر mام به‌صورت زیر محاسبه می‌شود:

$$LFPC_t(m) = 10\frac{\log_{10}(S_t(m))}{N_m} \quad (۱۰)$$

که در آن N_m تعداد مؤلّفه‌های طیفی در زیرباند mام است. برای هر فریم ۱۲ پارامتر LFPC بدست می‌آید.

۵- ارزیابی

ارزیابی‌ها در دو مرحله انجام می‌شوند [۱۸]: در مرحله‌ی اوّل سیگنال گفتار ورودی به‌عنوان گفتار خنثی و یا استرس‌دار طبقه‌بندی می‌گردد. بعد از این‌که حالت غیرخنثی آشکارسازی شد، می‌توان روی آن پردازش‌هایی به‌منظور تفکیک حالات استرسی مختلف انجام داد.

ما در آزمایش‌های خود از بخش واکدار مجموعه لغات زیر که از پایگاه داده‌ی SUSAS برداشته شده‌اند، استفاده می‌کنیم: "east"، "eight"، "eighty"، "fix"، "enter"، "help"، "mark"، "nav"، "no" و "oh". برای هر حالت استرسی (خنثی، عصبانی، بلند و لمبارد) ۲۰ کلمه (دو ضرب در ۱۰) از ۹ گوینده داریم؛ در مجموع ۷۲۰ کلمه.

بعد از استخراج ویژگی‌های ذکر شده در بالا از بخش واکدار هر کلمه، ابتدا پارامترهای آماری ویژگی TEO-CB-Auto-Env (یعنی میانگین و انحراف معیار آن) در آن واکه تخمین زده می‌شوند[۱] و سپس مقدار متوسط گام گفتار به این پارامترهای آماری اضافه می‌گردد (تا اینجا بُعد بردار ویژگی ما ۳ است). عملکرد این بردار ویژگی که آن را TEO-Env-Pch می‌نامیم برای طبقه‌بندی دوحالته و چندحالته استرس

[۱] توجه شود که در کارهای قبلی که از ویژگی TEO-CB Auto-Env به‌صورت منفرد استفاده می‌شد، بُعد این ویژگی به تنهایی ۱۶ بود. در ضمن عملکرد پارامترهای آماری این ویژگی بهتر است.

بررسی می‌شود. سپس بردار ویژگی پیشنهاد شده با اضافه-کردن LFPCها (۱۲ پارامتر) تکمیل می‌گردد (بردار ویژگی نهایی TEO-Pch-LFPC، نامیده می‌شود و بُعد آن ۱۵ است). برای تعیین طبقه‌ی هر نمونه، ابتدا از طبقه‌بندهای ساده‌ی LDA و KNN استفاده می‌کنیم. به‌ازای هر طبقه‌بند، نتایج ویژگی LFPC هم گزارش شده است. روش ارزیابی RRM[1] می‌باشد [۵]. به این معنی که، هر بار یک کلمه (از دو تلفظ مربوط به یک سوژه) به‌عنوان داده‌ی تست کنار گذاشته می‌شود و از بقیه‌ی کلمات برای تعلیم استفاده می‌گردد. این فرآیند به‌ازای تمام داده‌ها تکرار می‌شود و سپس روی نتایج، میانگین گرفته می‌شود.

در ادبیات گفتار استرسی، آشکارسازی استرس را «تفکیک دو حالته» و تفکیک حالات مختلف استرسی از هم را «طبقه‌بندی چندحالته» می‌گویند. درضمن، منظور از تشخیص استرس در این تحقیق، آشکارسازی خودکار وجود استرس ادراکی در یک کلمه است.

در بخش ۵-۱ فقط تفکیک دو حالته‌ی (خنثی/ استرسی) را توصیف می‌کنیم. در حالی‌که در بخش ۵-۲ تلاش می‌کنیم حالات مختلف استرسی را از هم جدا کنیم.

۵-۱- طبقه بندی دو حالته

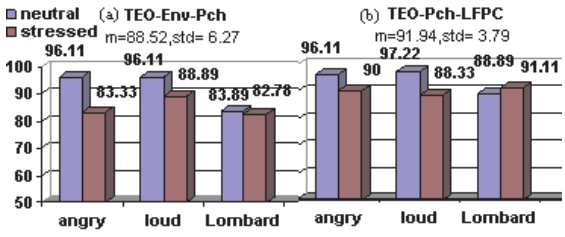
در این رویکرد برای هر طبقه‌بند سه بار عمل تفکیک را انجام می‌دهیم و هر بار روی دو حالت کار می‌کنیم (حالت خنثی و یکی از حالات استرسی). به این ترتیب، در هر طبقه‌بندی دوحالته، RRM به ۳۶۰ کلمه اعمال می‌شود که نصف آن‌ها دادگان گفتاری خنثی و نصف دیگر مربوط به دادگان گفتاری آن حالت استرسی خاص می‌باشند. جدول (۱) درصد صحت متوسط‌گیری شده‌ی کل را، به‌ازای هر ویژگی، گزارش می‌کند.

جدول ۱: طبقه بندی دو حالته‌ی استرس

Accuracy		LFPC	TEO-Env-Pch	TEO-Pch-LFPC
LDA	Mean(٪)	۸۶.۵۰	۸۸.۵۲	۹۱.۹۴
	std	۳.۸۹	۶.۲۷	۳.۷۹
KNN	Mean(٪)	۸۵.۲۷	۸۸.۱۴	۹۱.۷۶
	std	۵.۱۷	۶.۳۱	۴.۳۲

به‌منظور مقایسه با کارهای قبلی شکل (۲) بهترین نتایج این جدول (یعنی نتایج طبقه‌بند LDA) را با جزییات بیشتر نشان می‌دهد.

[1] Round-Robin Method



شکل ۱: طبقه بندی دوحالته با استفاده از LDA، در هر مورد std به‌معنای انحراف معیار و m به‌معنای مقدار میانگین است.

۵-۲- طبقه‌بندی چندحالته (چهار طبقه)

بعد از آشکارسازی حالات استرسی، هر حالت می‌تواند به‌صورت جداگانه تفکیک شود [۱۸]. در جدول (۲) درصد صحت‌های طبقه‌بندی چندحالته‌ی استرس، برای ویژگی‌های پیشنهاد شده، بیان می‌شوند و در جدول (۳) ، بهترین نتایج جدول (۲) (یعنی نتایج مربوط به ویژگی TEO-Pch-LFPC توسط طبقه‌بند KNN) با جزییات بیشتر آورده شده است. در اینجا RRM به ۷۲۰ کلمه اعمال می‌شود.

جدول ۲: طبقه‌بندی چند حالته‌ی استرس (۴طبقه)

	LFPC	TEO-Env-Pch	TEO-Pch-LFPC
LDA	۵۵.۱۴	۵۸.۴۷	۶۱.۲۵
KNN	۵۵.۵۶	۵۵.۱۴	۶۵.۸۵

جدول ۳: جزییات نتایج طبقه‌بندی۴حالته‌ی استرس توسط ویژگی TEO-Pch-LFPC

KNN	a) Correct Detection Rate (٪)		b) Distribution of STRESS Detection Rate (٪) across 3 Styles		
	Neutral	Stress	Angry	Loud	Lombard
Neutral	۸۶.۱۱		۰	۴	۹۶
Angry		۹۵	۵۵.۵۶	۲۵.۱۵	۱۹.۳
Loud		۹۳.۸۹	۲۰.۱۲	۵۲.۶۶	۲۷.۲۲
Lombard		۸۷.۷۸	۵.۰۳	۹.۴۳	۸۵.۵۳
Total Rate	Test		Train		
	۶۵.۸۵		۸۶.۱۰		

هدف این ارزیابی این است که ابتدا دریابیم ویژگی پیشنهاد شده تا چه حد می‌تواند گفتار خنثی و استرس‌دار را از هم تفکیک کند و درضمن تا چه حد قادر است حالات مختلف استرسی را از هم جدا نماید. در جدول(۳)ابتدا نتایج آشکارسازی صحیح خنثی و استرسی گزارش می‌شوند [بخش (a)]. به‌منظور پیاده‌سازی این فرآیند، کلّ دادگان استرسی به‌عنوان "گفتار استرسی" در یک گروه قرار می‌گیرد. اگر یک کلمه‌ی خنثی به‌عنوان داده‌ی تست به طبقه‌بند داده شود، آشکارسازی صحیح زمانی اتفاق می‌افتد

که طبقه‌بند برچسب خنثی را آشکار کند (۸۶/۱۱٪). در حالی‌که برای یک کلمه‌ی استرسی اگر برچسب آشکار شده متعلق به هر یک از حالات استرسی باشد، آن‌گاه فرض بر این است که حالت استرسی درست تشخیص داده شده است. (به‌عنوان مثال ۹۵٪ داده‌های عصبانی به‌عنوان گفتار استرسی تشخیص داده می‌شوند، نه لزوماً به‌عنوان داده‌ی عصبانی).

در بخش (b) نتایج تفکیک چندحالته گزارش می‌شود؛ یعنی با این فرض که کلمه به‌عنوان استرسی تشخیص داده شده باشد، می‌خواهیم ببینیم که چند درصد این حالت استرسی، درست همان حالتی که بوده است، تشخیص داده شده است.

برای تست خنثی (بخش (b)) نرخ خطا را می‌دهد. بخش عمده‌ای از کلمات خنثی که به‌طور اشتباه استرس‌دار تشخیص داده شده‌اند، متعلق به حالت لمبارد هستند (۹۶٪؛ جدول۳).

جدول ۴: جزییات نتایج طبقه‌بندی۴حالته استرس توسط ویژگی TEO-Pch-LFPC و طبقه‌بند SVM

SVM classifier	a) Correct Detection Rate (٪)		b) Distribution of STRESS Detection Rate (٪) across 3 Styles		
	Neutral	Stress	Angry	Loud	Lombard
Neutral	۹۱.۰۵				
Angry		۹۷.۲۲	۵۳.۳۳		
Loud		۹۷.۵۳		۵۶.۰۱	
Lombard		۹۴.۷۵			۸۷.۹۵
Total Rate	Test		Train		
	۷۰.۲۲		۷۷.۰۱		

توضیحات مربوط به جدول (۴) در بخش بعدی آورده شده است.

۶- بحث

مطالعات نشان می‌دهند [۱۸] که طبقه‌بندی چند حالته‌ی استرس فقط در حالت وابسته به گوینده عملکرد مناسبی دارد و این عملکرد در کاربردهای مستقل از گوینده افت می‌کند. در مرجع [۱] هر یک از ویژگی‌های TEO-CB-Auto-Env وگام گفتار به‌طور جداگانه توسط طبقه‌بندهای مبتنی بر HMM پنج‌حالته (۴ مدل HMM) با ۴۳۲ کلمه از بخش شبیه‌سازی SUSAS در مرحله‌ی تعلیم بررسی شده بودند. برای هر حالت استرسی ۲۷۰ کلمه (به‌جز آن‌هایی که در مرحله‌ی تعلیم بودند) به‌صورت خودکار برای تست استفاده شده‌اند. بدین ترتیب مجموعه دادگان تست و تعلیم در ارزیابی‌های آن‌ها ثابت بوده است. نتایج طبقه‌بندی دو حالته‌ی آن‌ها در جدول(۵) گزارش شده است:

جدول۵: نتایج تفکیک دو حالته با طبقه‌بند HMM [۱]

	m	std
گام گفتار	۸۱.۸۳	۱۷.۴۶
TEO-CB-Auto-Env	۸۷.۲۳	۸.۷۶

همچنین، برای طبقه‌بندی چندحالته درصد صحت کل و نرخ آشکارسازی صحیح حالت خنثی به ترتیب بصورت زیر هستند [۱]:

جدول۶: نتایج طبقه‌بندی چهارحالته با طبقه‌بند HMM [۱]

	درصد صحت کل	نرخ آشکارسازی صحیح حالت خنثی
گام گفتار	٪۵۸.۵	٪۵۲.۲
TEO-CB-Auto-Env	٪۵۶.۱۶	٪۷۰.۶

به‌نظر می‌رسد این ویژگی‌ها قادر نیستند، حالات لمبارد و خنثی را به‌طور مناسبی از هم تفکیک کنند [۱]. دلیل این مسأله این‌گونه توجیه شده است که با بررسی نرخ توزیع آشکارسازی استرس، متوجّه می‌شویم حالات مختلف استرسی تا حدودی با هم، همپوشانی دارند. علّت این امر این است که شرایط استرسی، بسیار پراکنده هستند و نمی‌توان آن‌ها را در تعداد محدودی داده‌ی تعلیم گنجاند. معمولاً جفت‌های (عصبانی و بلند) با هم و جفت‌های (خنثی و لمبارد) با هم اشتباه گرفته می‌شوند؛ چرا که به‌طور مثال، اشخاص اغلب سعی می‌کنند عصبانیت خود را با بلند کردن صدایشان نشان دهند. حال می‌خواهیم با ثابت نگه‌داشتن ویژگی، طبقه‌بند را از نوع دینامیک به نوع استاتیک تغییر دهیم؛ برای این منظور، ۱۶ پارامتر ویژگی TEO-CB-Auto-Env را به عنوان ورودی به طبقه‌بند KNN می‌دهیم، نتایج زیر بدست می‌آید (طبقه‌بند LDA در این زمینه بسیار ضعیف عمل می‌کند و نتایج حتی از شانس هم کمتر است):

جدول ۷: جزئیات نتایج طبقه بندی۴حالته استرس توسط ویژگی TEO-CB-Auto-Env

KNN	a) Correct Detection Rate (٪)		b) Distribution of STRESS Detection Rate (٪) across 3 Styles		
	Neutral	Stress	Angry	Loud	Lombard
Neutral	۶۸.۳۳				
Angry		۸۷.۲۲	۵۷.۳۲	۳۱.۲۱	۱۱.۴۶
Loud		۷۹.۴۴	۲۱.۸۸	۴۶.۸۸	۳۱.۲۵
Lombard		۸۸.۸۹	۹.۷۹	۳۳.۵۷	۵۶.۶۴
Total Rate	Test		Train		
	۵۱.۲۵		۷۵.۹۷		

همان‌طور که مشاهده می‌شود نتایج این ویژگی با طبقه‌بند استاتیک هم، برای تفکیک چند حالته‌ی استرس مناسب نمی‌باشد. مرجع [۱] بحث خود را این‌گونه خاتمه می‌دهد: "کیفیت عملکرد TEO-CB-Auto-Env و گام گفتار با هم متفاوت است، به این ترتیب که TEO-CB-Auto-Env در جداکردن گفتار خنثی از استرسی، عملکرد بهتری دارد؛ درحالی‌که گام گفتار در آشکارسازی حالات استرسی بهتر عمل می‌کند. به این ترتیب پیشنهاد می‌شود که ترکیب این دو ویژگی، عملکرد طبقه‌بند را بهبود می‌دهد." بر اساس این پیشنهاد، در کار حاضر این دو ویژگی با هم تلفیق شدند. یادآوری این نکته ضروری است که اگرچه ترکیب گام گفتار و TEO-CB-Auto-Env تا حدودی مؤثر بوده است (ستون یکی مانده به آخر جدول‌های ۱و۲) اما دلیل اصلی بهبود نتایج، دخیل کردن ضرایب LFPC می‌باشد (ستون آخر جدول‌های۱ و۲). توجّه شود که این بهبود، به‌ازای ساده‌ترین طبقه‌بند است و هنگامی که از طبقه‌بندهای استاتیک دیگر چون SVM استفاده می‌کنیم، نتایج به‌دست آمده، بسیار بهتر می‌شوند. در کار حاضر از SVM با هسته‌ی RBF استفاده شده است. نتایج مربوط به این طبقه‌بند در جدول۴، آمده است. در اینجا هم روش ارزیابی، RRM است و افزایش چشمگیر نرخ آشکارسازی حالت خنثی (۹۱.۰۵) و درصد صحت کلّ(۷۰.۲۲) مشهود می‌باشد. با توجّه به این‌که در هر دو نوع طبقه‌بندی (دو حالته و چند حالته)، عملکرد طبقه‌بند SVM از دو طبقه‌بند استاتیک دیگر بهتر است، برای مقایسه‌ی نهایی نتایج این کار با ویژگی قبلی (TEO-CB-Auto-Env)، از مجموعه (ویژگی TEO-Pch-LFPC و طبقه‌بند SVM) استفاده می‌کنیم (درطبقه‌بندی چهارکلاسه با استفاده از طبقه‌بند SVM، ۱۳.۵۸٪+ بهبود مشاهده می‌شود، جدول۹).

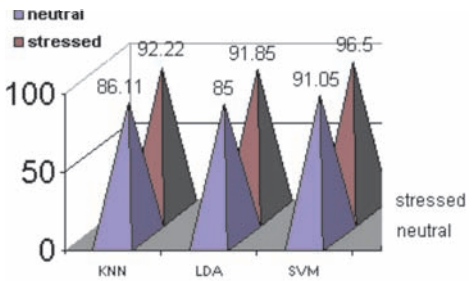
نکته‌ی قابل تأمّل این است که در کارهای اخیر، در رویکرد تفکیک دو حالته، هر بار روی دو حالت کار شده است (حالت خنثی و تنها یکی از حالات استرسی). اگرچه، در این تحقیق، نتایج با این شیوه هم گزارش شده‌اند، امّا این مسأله قدری اشکال دارد! چرا که هیچگاه، در کاربرد عملی، نمی‌توان یک حالت استرسی خاص داشت و فقط روی آن حالت تمرکز نمود. معقول‌تر این است که به‌منظور پیاده‌سازی این فرآیند، کلّ دادگان استرسی به‌عنوان «گفتار استرس‌دار» در یک گروه قرار گیرند. بنابراین، گزارش نهایی نتایج تفکیک دوحالته را بر اساس این معیار ارائه می‌دهیم (جدول ۸)؛ جدول ۹ نتایج مربوط به طبقه‌بندی چهارحالته را خلاصه می‌کند. درضمن، برای مقایسه، نتایج ویژگی TEO-CB-Auto-Env هم آمده است. دقّت شود که در مجموعه‌ی حاضر، در مقایسه با کارهای اخیر، هم ماهیت ویژگی‌ها و هم ماهیت طبقه‌بند (طبقه‌بند HMM از نوع دینامیک است) تغییر کرده‌اند. در شکل(۲)، خلاصه‌ی نتایج آورده شده است.

جدول۸: خلاصه‌ی نتایج تفکیک دو حالته با ویژگی TEO-Pch-LFPC (N مخفف Neutral و S مخفف Stressed می‌باشد).

		Neutral	Angry	Loud	Lombard	نرخ آشکارسازی خنثی / استرس	درصد صحت کل
KNN	N	۸۶.۱۱				۸۶.۱۱	۸۹.۱۷
	S		۹۵	۹۳.۸۹	۸۷.۷۸	۹۲.۲۲	
LDA	N	۸۵				۸۵	۸۸.۴۲۵
	S		۹۴.۴۴	۹۲.۷۸	۸۸.۳۳	۹۱.۸۵	
SVM	N	۹۱.۰۵				۹۱.۰۵	۹۳.۷۸
	S		۹۷.۲۲	۹۷.۵۳	۹۴.۷۵	۹۶.۵	

جدول۹: خلاصه‌ی نتایج طبقه‌بندی چهار حالته

	KNN	LDA	**SVM**
TEO-CB-Auto-Env	۵۱.۲۵	-	**۵۶.۶۴**
TEO-Pch-LFPC	۶۵.۸۵	۶۱.۲۵	**۷۰.۲۲**



شکل ۲: خلاصه‌ی نتایج در تفکیک دوحالته (ویژگی TEO-Pch-LFPC)

با وجود این‌که شرایط با کارهای پیشین یکسان نیست، یعنی تعداد دادگان در مرحله‌ی تعلیم افزایش یافته است، امّا به‌نظر می رسد رویکرد پیشنهاد شده در این مقاله در مقایسه با تحقیقات اخیر عملکرد بهتر و ساده‌تری دارد.

۷- نتیجه‌گیری و پیشنهادها

یک راه برای ارتقای نرخ آشکارسازی استرس گوینده، تلفیق شیوه‌های قدیمی با روش‌های موجود است. با استفاده از نتایج به‌دست آمده از طبقه‌بندی ویژگی‌های منفرد، در کار حاضر چند ویژگی خطّی و غیرخطّی را (که در مطالعات اخیر بهترین عملکرد را داشته‌اند)، با هم ترکیب کرده‌ایم و از طبقه‌بندهای استاتیک LDA، KNN و SVM استفاده نموده‌ایم. به‌عنوان یک پیشنهاد برای کارهای آینده، می‌توان طبقه‌بند مبتنی بر HMM را با این طبقه‌بندهای استاتیک ترکیب کرد. همچنین می‌توان قبل از ترکیب ویژگی‌ها، به هر کدام یک وزن مناسب اختصاص داد و یا یک تابع مناسب اعمال کرد تا نرخ طبقه‌بندی در عملکرد توأم آن‌ها، بهبود یابد (این کار باید بر اساس نظر فیزیولوژیست‌ها و اشخاص کارشناس در این زمینه انجام گردد). اعمال طبقه‌بندهای HMM یا DBN به ویژگی پیشنهاد شده، می‌تواند یک شانس جدید برای مدل‌سازی دینامیک گفتار استرسی را فراهم نماید.

یکی از مشکلات عمده‌ای که در پردازش گفتار استرسی (به‌خصوص هنگام کاربرد عملی) با آن مواجه می‌شویم مشکل کمبود داده‌ی برچسب‌خورده[1] است. روش‌های طبقه‌بندی سنّتی نمی‌توانند از دادگان بدون برچسب استفاده نمایند. از طرف دیگر، برچسب‌زدن داده‌های این حوزه، هزینه‌بر و زمان‌بر است و به‌نظر افراد متخصّص نیاز دارد. روش‌های یادگیری نیمه‌سرپرستی [۲۰،۱۹] می‌توانند برای حلّ این مشکل مفید باشد. الگوریتم طبقه‌بندی نیمه‌سرپرستی با استفاده از دادگان بدون‌برچسب، فرضیات به‌دست آمده توسط دادگان برچسب‌دار را اصلاح می‌کنند و حتّی گاهی فرضیات جدید و قوی‌تری می‌سازند.

درضمن، از آن‌جایی‌که اکثر راه‌کارهای بازشناسی و طبقه‌بندی استرس گفتاری، وابسته به گوینده هستند [۳]، باید سعی شود این وابستگی تا حد امکان کاهش یابد. به نظر می‌رسد، ساده‌ترین راه برای این منظور، تلفیق روش‌های طبقه‌بندی استرس با الگوریتم‌های بازشناسی گوینده باشد [۲۱]. به‌طور کلی، اگر قرار باشد مقوله‌ی پردازش گفتار استرسی، جنبه‌ی کاربردی پیدا کند، باید مستقل از گوینده و حتّی مستقل از زبان باشد. در مجموعه‌ی حاضر، ما روی انتخاب «مناسب‌ترین مجموعه‌ی ویژگی-طبقه‌بند» متمرکز شدیم و روی بحث استقلال از گوینده تمرکز نکردیم؛ چراکه هدف ما نخست، بررسی قابلیت آشکارسازی استرس گوینده و سپس، بهبود نتایج در مقایسه با کارهای قبل بود. از آن‌جایی‌که تاکنون، در مبحث تشخیص استرس گفتاری به‌صورت مستقل از گوینده، کاری انجام نشده است؛ پیشنهاد می‌شود در کارهای آینده این امر هم مورد توجه قرار گیرد. با به‌کارگیری رویکردهای مختلف و مقایسه‌ی نتایج آن‌ها با هم، این مهم دست یافتنی به‌نظر می‌رسد.

برای داشتن یک سیستم بازشناسی احساس «مستقل از زبان» باید مجموعه‌ای از گفتارهای احساسی به زبان‌های مختلف گردآوری شوند. امّا ابتدا باید به این سؤال پاسخ داد که «آیا می‌توان مستقل از زبان کار کرد یا نه؟!». برای پاسخ به این سوال، اوّلین مرحله این است که یک واکه‌ی ثابت را، که توسط گویندگان بومی زبان‌های مختلف گفته شده است، مورد پردازش قرار دهیم.

درنهایت، تعیین میزان و سطح استرس گوینده (کم، متوسط و یا زیاد) می‌تواند یک موضوع قابل بررسی برای تحقیقات آینده باشد. روش‌های موجود برای آشکارسازی استرس، فقط از یک رویکرد تصمیم‌گیری دوحالته (وجود یا عدم وجود استرس) بهره می‌گیرند. چون میزان استرس متغیر است، این‌گونه به‌نظر می‌رسد که برای تشخیص صحیح حالت استرسی گوینده، نیاز است که سطح استرس فرد نیز به‌درستی تشخیص داده شود.

[1] منظور از این برچسب، حالات مختلف استرسی (عصبانی، بلند، خنثی و اثر لمبارد) است و با برچسب مربوط به جداسازی واج‌ها تفاوت دارد.

۸ ـ مراجع

[1]G. Zhou, J. H. L. Hansen and J. F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress," IEEE Trans. Speech and Audio Processing, vol. 9, No. 3, March 2001.

[2] T. Polzin, "Detecting verbal and non-verbal cues in the communication of emotions". Doctoral Dissertation, School of Computer Science, Carnegie Mellon University,2000.

[3] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features and Methods," Artificial Intelligence and Information Analysis Laboratory, Aristotle University of Thessaloniki, April 2006.

[4] D. A. Cairns and J. H. L. Hansen, "Nonlinear Analysis and Classification of Speech Under Stressed Conditions," Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, July 1994.

[5] G. Zhou, J. H. L. Hansen and J. F. Kaiser, "A New Nonlinear Feature for Stress Classification," Robust Speech Processing Laboratory, Duke University, 1998.

[6] T. L. Nwe, Y. Wang, "Automatic Detection of Vocal Segments in Popular Songs," School of Computing National University of Singapore, 2004.

[7] K. H. Hyun, E. H. Kim and Y. K. Kwak, “Robust Speech Emotion Recognition Using Log Frequency Power Ratio,” SICE-ICASE International Joint Conference, 2006.

[8] R. Sarikaya, N. Gowdy, “Wavelet Based Analysis of Speech under Stress,” Digital Speech and Audio Processing Laboratory of Clemson University, 1998.

[9] B.D. Womack and J.H.L Hansen. N-channel hidden markov models for combined stressed speech. IEEE Trans Speech Audio Proc, 7(6):668–677, 1999.

[10] G. Zhou, J. H. L. Hansen and J. F. Kaiser, “Classification of Speech Under Stress Based on Features Derived from the Nonlinear Teager Energy Operator,” IEEE ICASSP-98, 1998.

[11] A. Mohammadian and V. Aboutalebi, "Single Trial Classification of Event-Related Potentials for Detection of Target stimulus" Journal of Biannual Letter of Research Center of Intelligent Signal Processing, No. 1, serial.5, pp.3-12, 2006.

[12] J. H. L. Hansen and S. Bou-Ghazale, “Getting Started with SUSAS: A Speech under Simulated and Actual stress,” in proc. EUROSPEECH ’97, vol. 4, pp. 1743-1746, 1997.

[13] B. D. Womak and J. H. L. Hansen, “N-Channel Hidden Markov Model for Combined Stressed Speech Classification and Recognition,” IEEE Transactions on Speech and Audio Processing, vol. 7, NO. 6, November 1999.

[14] P. Maragos, J. F. Kaiser and T. F. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis,” IEEE Trans. Signal Proc., vol. 41, NO. 10, pp. 3025-3051, October 1993.

[15] A. Potamianos and P. Maragos, “Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation,” School of Electrical and Computer Engineering, Georgia of Technology, June 1994.

[16] T.L. Nwe, S.W. Foo, and L.C. De Silva. “Detection of stress and emotion in speech using traditional and FFT based log energy features,” In Fourth Pacific Rim Conference on Multimedia, Information, Communications and Signal Processing, volume 3, pages 1619–1623, December15-18 2003.

[17] K. Hyun, E. Kim and Y. Keun Kwak, “robust speech recognition using log frequency power ratio”, SICE-ICACE International Joint Conference, 2006.

[18] G. Zhou, J. H. L. Hansen and J. F. Kaiser, “Linear and Nonlinear Speech Feature Analysis for Stress Classification,” Robust Speech Processing Laboratory, Duke University, 1998.

[19] W. Du, K. Inoue, and K. Urahama, “Unsupervised and semi-supervised extraction of fuzzy clusters in similarity data,” IEICE Technical Report, PRMU 2005-177, Jan. 2006.

[20] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, andM. Figueiredo, “On semi-supervised classification,” Proc. NIPS, pp.721–728, 2005.

[21] P. Thevenaz, and H. Hugli, “Usefulness of the LPC Residue in Text-independent Speaker Verification,” Speech Communication, Vol.17, pp. 145-157,1995 .

شهلا ترابی (متولّد۱۲ تیر ۱۳۶۳) در سال ۱۳۸۵، در رشته‌ی مهندسی پزشکی، گرایش بالینی(دوره‌ی کارشناسی)، از دانشگاه امیرکبیر فارغ التحصیل شده است. سپس دوره‌ی کارشناسی ارشد در همین رشته (گرایش بیوالکتریک)، را تا سال ۱۳۸۸، ادامه داده است. زمینه‌ی تحقیقاتی مورد علاقه‌ی وی، پردازش گفتار و نیز پردازش علائم حیاتی می‌باشد.

نشانی پست الکترونیکی ایشان عبارت است از:

Shahla_t5024@yahoo.com

فرشاد الماس گنج در سال ۱۳۶۳ در رشتهٔ برق، گرایش الکترونیک، از دانشگاه امیرکبیر فارغ التحصیل گردیده است. سپس دورهٔ کارشناسی ارشد در همین رشته را تا سال ۱۳۶۷ ادامه داده است. وی با یک فاصلهٔ ۴ ساله، دورهٔ دکتری برق (گرایش مهندسی پزشکی) را در دانشگاه تربیت مدرس آغاز نموده و از سال۱۳۷۷ با سمت استادیار در دانشکدهٔ مهندسی پزشکی از دانشگاه امیرکبیر مشغول به کار است. زمینهٔ تحقیقاتی اصلی او پیرامون پردازش سیگنال و عمدتاً در زمینهٔ بازشناسی گفتار فارسی و بازشناسی خصوصیات پروزودیک گفتار می باشد.

نشانی پست الکترونیکی ایشان عبارت است از:

almas@aut.ac.ir

امین محمدیان مدرک کارشناسی خود را در رشته‌ی مهندسی پزشکی در سال ۱۳۸۱ از دانشگاه صنعتی امیرکبیر و مدرک کارشناسی ارشد در رشته‌ی مهندسی برق- بیوالکتریک را در سال ۱۳۸۴ از دانشگاه تربیت مدرس اخذ نمود. وی هم اکنون دانشجوی دکتری مهندسی پزشکی- بیوالکتریک در دانشگاه صنعتی امیرکبیر و و مدیر گروه پردازش علائم حیاتی پژوهشکده‌ی پردازش هوشمند علائم می‌باشد. زمینه‌های تحقیقاتی مورد علاقه‌ی وی پردازش سیگنال‌های EEG، ERP در حالات ذهنی، پردازش تصاویر حراراتی، پردازش اطلاعات فرا گفتار، شناسایی الگو و محاسبات نرم می‌باشد.

نشانی پست الکترونیکی ایشان عبارت است از:

Mohammadian@Rcisp.ac.ir