Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال چهارم، شماره اول، بهار و تابستان ۱۴۰۵، صفحات ۱۸۱ – ۱۶۷

بهبود گشتار صدای گفتار احساسی با استفاده از یک تابع انرژی کارآمد

مجید نیک‌زر[۱]، حسن ختن‌لو[۲*]، میرحسین دزفولیان[۳]

[۱] دانشجوی دکتری، گروه مهندسی کامپیوتر، دانشگاه بوعلی‌سینا، همدان، ایران

[۲] استاد، گروه مهندسی کامپیوتر، دانشگاه بوعلی‌سینا، همدان، ایران

[۳] استادیار، گروه مهندسی کامپیوتر، دانشگاه بوعلی‌سینا، همدان، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۴/۰۹/۱۵

تاریخ پذیرش:
۱۴۰۴/۱۰/۳۰

کلیدواژه‌ها:
تبدیل گفتار احساسی، بازسازی گفتار، مدل‌سازی نوای گفتار، شدت صوت، طیف انرژی سیگنال، نقاط اکسترمم

نویسنده‌مسئول:
khotanlou@basu.ac.ir

چکیده

یکی از موضوعات مهم در حوزه پردازش صوت و گفتار، تغییر احساس در گفتار است. ازجمله چالش‌های موردتوجه، محاسبه مقدار دقیق ویژگی‌های اصلی شامل انرژی، گام و دیرش است. هرچند قبلاً روش‌های موثری ارائه‌شده‌اند، لیکن روش‌های موجود برای محاسبه انرژی به جهت اینکه فقط پارامتر دامنه را لحاظ می‌کنند، به‌تنهایی برای مدل‌سازی نوای گفتار کارایی مطلوبی ندارند. این پژوهش نشان می‌دهد که چگونه می‌توان از تابع جدیدی برای محاسبه انرژی جهت مدل‌سازی گشتار گفتار احساسی بهره برد. روش پیشنهادی برای محاسبه انرژی، بر اساس حساسیت سیستم شنوایی انسان در فرکانس‌های مختلف عمل می‌کند. در این روش انرژی بر اساس فاصله نقاط اکسترمم که مرتبط با دامنه و فرکانس است محاسبه می‌شود. جهت ارزیابی کارایی، یک سیستم گشتار گفتار احساسی با استفاده از تابع انرژی پیشنهادی بر روی پایگاه‌داده گفتار احساسی **Persian ESD** پیاده‌سازی و با روش‌های معمول محاسبه انرژی مقایسه شد. نتایج آزمایشی طبق نظرسنجی **CMOS** نشان می‌دهد که تابع پیشنهادی، کیفیت گفتار احساسی تولیدشده را در حد مطلوبی افزایش داده است.

: 10.22034/ABMIR.2026.24063.1196

E-ISSN: 2821-2037
/© 2026. Published by Yazd University This is an open access article under the CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).

۱- مقدمه

پردازش گفتار احساسی از اهمیت ویژه‌ای برخوردار است و در حال حاضر پژوهش‌های ارزشمندی در این زمینه در حال انجام است [۱]. برای اینکه این نوع ارتباط عاری از ابهام باشد لازم است گوینده واژه‌ها را دقیق و شمرده و مطابق با قواعد تلفظ کند. لیکن انسان‌ها اطلاعات فرازبانی مانند احساسات نیز به گفتار اضافه می‌کنند که باعث تغییرات اساسی در ویژگی‌های مهمی مانند انرژی، گام و دیرش می‌شود. گوینده با ایجاد تغییرات در لحن گفتار که به آن نوای گفتار گفته می‌شود، منظور و احساس خود را بیان می‌کند. تغییر اطلاعات غیر زبان‌شناختی گفتار از یک حالت احساسی به حالت دیگر همراه با حفظ هویت گوینده، گشتار صدای گفتار احساسی[1] نامیده می‌شود [۲]. در این فرآیند پارامترهای صدا تغییریافته و گفتار باید همچنان طبیعی و قابل‌فهم باقی بماند.

اولین پژوهش‌ها در زمینه اضافه کردن احساس به گفتار از اوایل دهه ۹۰ میلادی شروع شد [۳]. سیستم‌های پردازش گفتار موجود، قادر به پردازش گفتار احساسی به‌طور مؤثر و مطلوب نمی‌باشند. جهت طبیعی نمودن و انتقال اطلاعات فرازبانی به گفتار که عمدتاً مربوط به شش احساس پایه خشم، شادی، غم، ترس، نفرت و تعجب است، معمولاً از دو رویکرد استفاده می‌شود. در رویکرد اول با استفاده از روش‌های پیوندی[2] و یک پایگاه‌داده گفتار احساسی، پاره‌های صوتی احساسی یا دوآوایی‌های از قبل ضبط‌شده را به هم چسبانده و گفتار احساسی تولید می‌شود [٤]. این روش نتایج رضایت‌بخشی را حاصل می‌کند اما اگر کسی بخواهد گفتار احساسی با صدای خودش تولید کند یا احساسات جدیدی را اضافه کند یا حتی میزان احساسات را کنترل کند عملی نیست. در رویکرد دوم (گشتار صدای احساسی) با استفاده از پایگاه‌داده‌هایی از گفتار احساسی تولیدشده انسانی، مشخصه‌هایی چون منحنی گام، انرژی و دیرش مربوط به احساس مورد نظر استخراج‌شده و مدل گفتار احساسی تولید می‌شود. سپس مدل استخراج‌شده روی گفتار خنثی اعمال می‌شود. نتایج این روش چندان رضایت‌بخش نیست. گفتار نسبتاً غیرطبیعی بوده و احساسات، خیلی خوب توسط شنوندگان درک نمی‌شود [۵]. مزایای این رویکرد این است که میزان احساسات قابل‌کنترل بوده، وابستگی آن به زبان کم بوده، می‌توان احساسات جدید به سیستم اضافه کرد و گفتار احساسی با صدای افراد مختلف تولید کرد.

روش‌های پیوندی عمدتاً برای تولید گفتار خنثی از روی متن است. این روش‌ها هرچند که گفتار واضحی را تولید می‌کنند ولی گفتار تولیدشده طبیعی نیست. برای طبیعی بودن گفتار لازم است احساسات به گفتار افزوده شود. با توجه به اینکه در یک متن معمولی احساسات نگارش نمی‌شود، جهت تبدیل یک متن به گفتار احساسی، لازم است ابتدا با استفاده از اطلاعات زبان‌شناختی و تکنیک‌هایی مانند یادگیری عمیق، متن ورودی پردازش و احساسات استخراج شود و بر مبنای آن متن مورد نظر بر اساس نوع احساسات نشان‌گذاری شود [٦]؛ سپس با استفاده از روش پیوندی، گفتار احساسی را تولید کرد. راه حل دیگر این است که ابتدا متن بدون نشان را مستقیماً با روش پیوندی و پایگاه‌داده‌ای از پاره‌های گفتار خنثی به گفتار خنثی تبدیل کرد، سپس با استفاده از روش‌های گشتار صدای احساسی آن را به گفتار احساسی تبدیل نمود. جهت گشتار صدا نیاز به اطلاعات نوایی[3] گفتار احساسی است که در این راستا ابتدا لازم است اطلاعات نوایی مربوط به گفتار احساسی از یک پایگاه‌داده گفتار احساسی استخراج شود سپس با مدل‌سازی نوا و اعمال آن روی گفتار موجود، گفتار احساسی تولید نمود.

مهم‌ترین کاربردهای گشتار گفتار احساسی عبارت‌اند از: بهبود سیستم‌های تولید گفتار از حالت مصنوعی به حالت طبیعی، اضافه کردن اطلاعات فرازبانی به گفتار، صداگذاری روی فیلم و کارتون، روبات‌های سخنگو، استفاده در ترجمه ماشینیِ گفتار به گفتار، گفتاردرمانی و کمک به افراد معلول. رفع چالش یادشده و بهبود کارایی فناوری ایجاد احساس در گفتار ازنظر علمی و تجاری نیز اهمیت زیادی دارد [۵].

[1] Emotional Voice Conversion/Transformation (EVC)
[2] Concatenation
[3] Prosodic

ویژگی انرژی نقش مهمی در تمام زمینه‌های پردازش گفتار ازجمله مدل‌سازی نوای گفتار احساسی دارد. در [۷] از انرژی برای خوشه‌بندی واج‌ها استفاده‌شده است. ویژگی انرژی در سایر حوزه‌ها مانند تعیین سطح نویز در آلودگی صوتی [۸]، جداسازی گفتار از نویز [۹] و پردازش تصویر [۱۰] نیز کاربرد دارد. در [۱۱] از دو الگوریتم بر مبنای انرژی برای آشکارسازی صدا (VAD) استفاده‌شده است. در این پژوهش یک تابع انرژی جدید برای استخراج و تولید نوای گفتار احساسی ارائه‌شده است. تابع انرژی پیشنهادی بر اساس حساسیت سیستم شنوایی انسان طراحی‌شده لذا علاوه بر مدل‌سازی نوای گفتار در سایر حوزه‌های پردازش گفتار نیز کاربرد دارد. جهت ارزیابی تابع انرژی از رویکرد دوم (گشتار صدای احساسی) برای تولید گفتار احساسی استفاده‌شده است، زیرا از یک طرف با چالش بیشتری مواجه است و از طرف دیگر نقش تابع انرژی در آن آشکارتر است. درواقع تمایز این روش با روش‌های موجود این است که در حوزه زمان و با پیچیدگی محاسباتی کم، ویژگی خاصی استخراج‌شده است که از یک طرف منعکس‌کننده فرکانس و از سوی دیگر منعکس‌کننده انرژی است.

در ادامه در بخش ۲، ضمن مرور ادبیات موضوع به تعریف اصطلاحات و مفاهیم موردنیاز پرداخته خواهد شد. در بخش ۳، تابع انرژی پیشنهادی همراه با شرح سایر ویژگی‌ها جهت بهبود گشتار گفتار احساسی ارائه و تشریح خواهد شد. در بخش ٤، نحوه پیاده‌سازی یک سیستم گشتار گفتار احساسی همراه با تابع انرژی پیشنهادی موردبررسی قرار خواهد گرفت. در بخش ۵، تبدیل گفتار احساسی با استفاده از تابع انرژی پیشنهادی و تابع انرژی مرسوم مقایسه و ارزیابی خواهد شد. در پایان در بخش ٦ نتیجه‌گیری و مسیر پژوهش‌های بعدی بیان خواهد شد.

۲- مرور ادبیات تحقیق

احساسات نقش مهمی در تعامل انسان و کامپیوتر مبتنی بر گفتار دارند. بنابراین، در سال‌های اخیر، پردازش گفتار توجه فزاینده‌ای را به خود جلب کرده است. بااین‌حال، علی‌رغم پیشرفت‌ها در این زمینه، هنوز یک کار چالش‌برانگیز است زیرا احساسات می‌توانند توسط افراد به روش‌های مختلف بیان شوند. علاوه بر این، عوامل بی‌ربط با احساسات، مانند جنسیت و سن، می‌توانند سیگنال گفتار را به روش‌های مختلفی تحت تأثیر قرار دهند [۱۲]. در این بخش ابتدا مدل‌سازی نوای گفتار احساسی را شرح داده سپس نحوه استخراج و اعمال ویژگی‌های مربوط به گفتار احساسی بررسی خواهد شد.

۲-۱ نوای گفتار و مدل‌سازی احساسات

منظور از احساس، تغییر در لحن گفتار با هدفِ رساندنِ منظور خاصی است. لذا احساسات شامل اطلاعات فرازبانی است که وابسته به زبان نیست. مشخصه‌های گفتار، هنگام تولید گفتار احساسی با زمان تغییر می‌کند. نوای گفتار بازتاب‌کننده مشخصه‌های گوینده، لهجه، احساسات و مشخصه‌های آوایی[1] است. برای تولید احساسات ابتدا باید نوای گفتار مدل‌سازی، سپس پردازش و تغییر داده شود. هنگام مدل‌سازی احساسات، عموماً دو مسئله پژوهشی وجود دارد: (۱) چگونگی توصیف و بازنمایی احساسات؛ (۲) چگونگی اثر دادن مدل احساسی روی گفتار بر اساس فرآیند بیان و درک احساسات توسط انسان [۲].

سیستم تولید گفتار توسط اندام گفتار در انسان را به دو بخش می‌توان تقسیم کرد. بخش اول شامل سخت‌افزار تولید گفتار است که شامل حنجره، زبان، لب‌ها، دندان‌ها و سایر اندام است که با یک سری فیلتر می‌توان آن را شبیه‌سازی کرد. بخش دوم شامل نرم‌افزار تولید گفتار است که شامل کنترل حرکاتی است که گوینده روی اندام گفتار خود انجام می‌دهد تا نوای گفتار را ایجاد کند. با فیلتر می‌توان اندازه طیف پاسخ فرکانسی سیستم اندام گفتار که حامل اطلاعاتی راجع به هویت گوینده است را تغییر داد. لذا با تغییر اندازه طیف اندام گفتار می‌توان هویت گوینده و همچنین احساسات را کنترل نمود [۱۳].

برای اعمال تغییرات روی فیلتر، رویکردهای قطعی گسسته و رویکردهای احتمالاتی را می‌توان نام برد. نگاشت چندی‌سازی برداری، رویکرد درون‌یابی گوینده و رویکرد استفاده از فیلترهای تصحیح [١٤] نمونه‌هایی از رویکردهای قطعی گسسته است. از رویکردهای احتمالاتی، هم در مبحث اعمال تغییرات روی فیلتر، هم در مبحث دسته‌بندی و هم در بحث مدل‌سازی نوای گفتار

[1] Phonetic

استفاده می‌شود. اغلب روش‌های مدل‌سازی بر مبنای فرضیات آماری استوار می‌باشند [۱۵]. به‌طورمعمول از تکنیک‌های مدل‌سازی احتمالاتی مانند مدل ترکیبی گوسین [۱۶] و مدل مخفی مارکوف [۱۷] برای بازشناسی گوینده، زبان، احساس و گفتار استفاده می‌شود. ردی و رائو نوای گفتار احساسی را با استفاده از تغییر پارامترهای فیلترینگ معکوس که توسط مدل مخفی مارکوف مدل شده است به دست آورده‌اند [۱۸]. در مدل‌های آماری هر احساس به‌عنوان یک منبع احتمالاتی همراه با یک تابع چگالی احتمال ناشناخته معین مدل می‌شود. در فاز آموزش پارامترهای تابع چگالی احتمال توسط تعداد کافی از نمونه‌های آموزشی تخمین زده می‌شود. مدل‌های ترکیبی گوسی، ترکیب خطی از توزیع‌های گوسی چند متغیره می‌باشند که به‌صورت P(X/C) مدل می‌شوند. مدل‌های ترکیبی گوسی توسط قانون بیز می‌توانند به دسته‌بندهای پسین تبدیل شوند [۱۵]. این روش‌ها دارای مزایایی چون قابلیت آموزش مدل روی یک پایگاه‌داده بزرگ و تطبیق‌پذیری روی داده‌های جدید می‌باشند.

۲-۲ استخراج ویژگی‌های مربوط به نوای گفتار

مهم‌ترین ویژگی‌ها جهت دسته‌بندی احساسات بر مبنای اطلاعات آماری عبارت‌اند از فرکانس گام، منحنی انرژی، دیرش، نرخ عبور از صفر، زمان سکوت، سازه‌ها[1]، انرژی‌های Mel-band، ضرایب پیش‌بینی خطی، ضرایب کپسترال، ضرایب MFCC و کیفیت صدا است [۱۹]، [۲۰]. نوع ویژگی‌ها و نحوه استخراج آن‌ها از دیگر چالش‌های گشتار صدای احساسی است که در این پژوهش به آن پرداخته‌شده است. در این پژوهش از سه ویژگی اصلی «انرژی، دیرش و گام»[2] در حوزه زمان جهت تبدیل گفتار خنثی[3] به گفتار احساسی مورد نظر استفاده‌شده است. در ادامه نحوه معمول استخراج این ویژگی‌ها را بررسی می‌کنیم.

۲-۲-۱ پنجره‌گذاری

ازآنجا که سیگنال‌های گفتار ناایستان[4] می‌باشند جهت استخراج ویژگی‌ها لازم است گفتار به فریم‌های زمانی کوتاه‌مدت تقسیم شود که به این عمل پنجره‌گذاری گفته می‌شود. سیگنال ناایستان به سیگنالی گفته می‌شود که فرکانس و دامنه آن در طول زمان متغیر است. این پنجره‌ها به‌صورت متوالی و با اندازه ثابت یا متغیر بر روی سیگنال گفتاری اعمال می‌شوند. طول یک پنجره را معمولاً بین ۱۰ تا ۳۰ میلی‌ثانیه اختیار می‌کنند. برای کارایی بیشتر معمولاً پنجره‌ها بیش از ۵۰٪ روی هم افتادگی[5] دارند.

یک پنجره دارای تابع مشخصه است که در تمام نمونه‌های سیگنال ضرب می‌شود لذا با ضرب یک تابع به شکل w[n] با طول زمانی محدود در سیگنال آن را به بازه‌های زمانی کوتاه‌تری تبدیل می‌کند. برخی از انواع پنجره‌گذاری معمول در پردازش سیگنال گفتار عبارت‌اند از: پنجره مستطیلی، پنجره مثلثی، پنجره همینگ و پنجره هنینگ. معمولاً از پنجره مستطیلی به جهت اینکه تاثیری روی نمونه‌های درون پنجره نمی‌گذارد برای استخراج گام و از پنجره هنینگ برای استخراج منحنی انرژی و تغییر گام استفاده می‌شود [۲۱]. تابع مشخصه پنجره هنینگ مطابق رابطه (۱) است [۲۲].

$$w[n] = \begin{cases} 0.5(1 - cos(\frac{2\pi n}{N})), 0 \le n \le N \\ 0 \; otherwise \end{cases} \quad (۱)$$

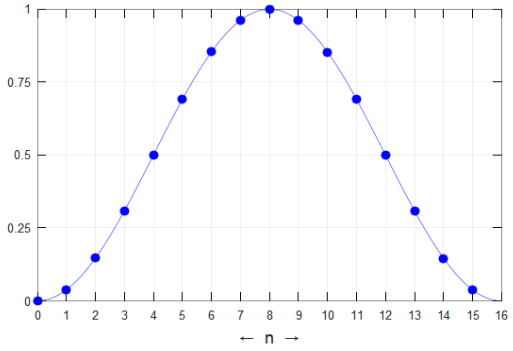
در رابطه (۱)، پارامتر N طول پنجره برحسب تعداد نمونه‌ها است. نمودار پنجره هنینگ در حوزه زمان در شکل (۱) دیده می‌شود.

۲-۲-۲ استخراج منحنی انرژی

انرژی صوت همان بلندی صدا است که به‌طورمعمول از روی نمونه‌های صوتی محاسبه می‌شود. انرژی یک ویژگی مهم در پردازش گفتار احساسی است که لازم است به‌طور دقیق استخراج و اثر داده شود. از تغییرات انرژی سیگنال به‌عنوان یک مشخصه مهم جهت شناسایی محل هسته هجا یا واکه‌ها و همچنین جهت تشخیص سکوت استفاده می‌شود [۲۳]. برای محاسبه انرژی یک سیگنال گسسته، روش‌های مختلفی در حوزه زمان و فرکانس ارائه‌شده است [۲۴]. برای محاسبه انرژی در حوزه زمان معمولاً از روش مجموع مربعات نمونه‌ها یا مجموع قدرمطلق نمونه‌ها و برای محاسبه انرژی در حوزه فرکانس، از روش جمع مقدار حقیقی

[1] Formants
[2] Intensity / Duration / Pitch
[3] Neutral
[4] Nonstationary
[5] Overlap

تبدیل فوریه گسسته نمونه‌ها استفاده می‌شود. برای محاسبه انرژی در فرکانس‌های مختلف از انرژی موجک نسبی[1] استفاده می‌شود [۲۵]. تمام این روش‌ها فقط پارامتر دامنه سیگنال را لحاظ می‌کنند.



شکل (۱): پنجره هنینگ در حوزه زمان

منظور از توان، انرژی لحظه‌ای صوت است. به‌عبارت‌دیگر اگر انرژی محاسبه‌شده در یک پنجره را بر تعداد نمونه‌ها در آن تقسیم کنیم توان به دست می‌آید. با توجه به اینکه طول زمانی پنجره‌ها معمولاً ثابت در نظر گرفته می‌شود، توان و انرژی به جای یکدیگر می‌توان استفاده کرد. اگر سیگنال نمونه‌برداری شده را با s[.] و پنجره را با تابع مشخصه آن یعنی w[.] نمایش دهیم، با توجه به تقارن تابع مشخصه پنجره، انرژی زمان-کوتاه سیگنال در لحظه یا نمونه n طبق رابطه (۲) محاسبه می‌شود [۲۶].

$$E_n = \sum_{m=-\infty}^{+\infty} (s[m].w[n-m])^2 \qquad (۲)$$

درواقع $w^2[.]$ یک فیلتری است که سیگنال $s^2[.]$ از آن عبور می‌کند (رابطه (۳)).

$$E_n = \sum_{m=-\infty}^{+\infty} s^2[m].h[n-m] \ , \quad h[n] = w^2[n] \qquad (۳)$$

برای محاسبه انرژی چنانچه از پنجره مستطیلی استفاده شود، با حذف نقاط ابتدایی و انتهایی اضافی رابطه (۳) به رابطه (٤) تبدیل خواهد شد.

$$E_n = \sum_{m=n-N/2}^{n+N/2} s^2[m] \qquad (٤)$$

در رابطه فوق پارامتر N طول پنجره و n شماره نمونه وسط پنجره است.

[1] Relative wavelet energy

ازآنجایی که در پردازش گفتار، مقدار نسبی انرژی اهمیت دارد نه مقدار واقعی آن لذا یک روش ساده برای محاسبه انرژی متوسط در حوزه زمان به‌صورت رابطه (۵) است.

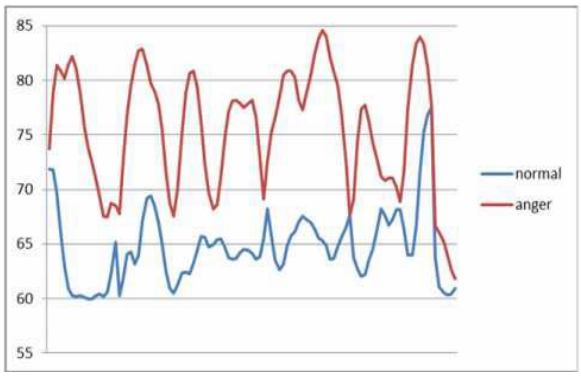
$$\bar{E} = \frac{1}{N} \sum_{m=n1}^{n2} |s[m]| \qquad (۵)$$

در رابطه (۵)، n1 شماره اولین نمونه و n2 شماره آخرین نمونه از یک پنجره مستطیلی با طول N است. برای محاسبه انرژی لحظه‌ای، بهتر است به جای پنجره مستطیلی از پنجره هنینگ در رابطه (۲) یا (۳) استفاده شود. روش دیگر برای محاسبه انرژی، این است که ابتدا از پنجره مورد نظر تبدیل فوریه زمان کوتاه[2] گرفته سپس مجموع مقادیر فرکانس را محاسبه نمود.

مفهوم شدت و انرژی صوت خیلی به هم نزدیک است لذا می‌توان به جای یکدیگر استفاده کرد. واحد شدت صوت دسیبل (db) بوده و روی سیگنال صوتی $x(t)$ به‌صورت رابطه (٦) تعریف می‌شود [۲۷].

$$I(t) = 10 \log x(t)^2 \qquad (٦)$$

در شکل (۲) تأثیر احساس خشم روی منحنی انرژی سیگنال گفتار مطابق با رابطه (٦) دیده می‌شود.



شکل (۲): نمودار شدت صوت روی گفتار خنثی و گفتار متناظر با حس خشم

۲-۲-۳ استخراج ویژگی گام

فرکانس اصلی گفتار واک‌دار مربوط به صوتی است که توسط تارآواها در حنجره تولید می‌شود که یک سیگنال صوتی نیمه سینوسی است. به فرکانس این صوت، گام یا فرکانس پایه گفته

[2] Short Time Fourier Transform-STFT

می‌شود که با F_0 نمایش داده می‌شود. بر اثر عبور این صوت از مجرای گفتار شامل گلو، چاکنای، زبان، دندان‌ها، لب‌ها و دماغ یک سری نویز با فرکانس و انرژی‌های مختلف افزوده می‌شود که گفتار را به وجود می‌آورد. لذا گفتار از مجموع یک سیگنال نیمه سینوسی و یک منبع نویز تشکیل‌شده است. فرکانس پایه یا گام نقش مهمی در احساسات دارد. همچنین گام در لحن گفتار بسیار مؤثر است. عمده اطلاعات نظیر لهجه، احساس و آهنگ در گام نهفته است. مطالعات نشان می‌دهد با تغییر نرخ گفتار، فرکانس گام روی واکه‌ها و همخوان‌ها به اشکال مختلف تغییر می‌کند [۲۸].

جهت استخراج گام روش‌های متعددی مانند روش آشکارساز پردازش موازی گام در دامنه زمان (PPROC[1])، روش کدگذاری پیش‌بینی خطی (LPC[2])، فیلترینگ معکوس ساده‌شده (SIFT[3])، روش همبستگی خودکار (AUTOC[4])، روش کپسترال (CEP[5])، تابع متوسط اختلاف دامنه (AMDF[6]) و غیره به کار می‌رود. تابع متوسط اختلاف دامنه دارای دقت بسیار خوبی است که در این پژوهش استفاده‌شده است. تابع متوسط اختلاف دامنه از همبستگی نمونه‌های سیگنال مربوط به واکه‌ها استفاده کرده و دوره تناوب اصلی صدا را محاسبه می‌کند. این تابع به‌صورت رابطه (۷) تعریف می‌شود.

$$\Delta M_s(\eta, m) = \frac{1}{N}\sum_{n=m-N+1}^{m} |s[n] - s[n-\eta]|w[m-n] \quad (۷)$$

در این رابطه $w[.]$ تابع مشخصه پنجره، N طول آن و m محل نقطه پایان پنجره است. $s[.]$ سیگنال گفتار درحوزه زمان و η متغیری است که نمایانگر اختلاف فاصله زمانی میان دو نمونه است. نقطه‌ای که در آن مقدار تابع AMDF به ازای $\eta \neq 0$ به صفر نزدیک‌تر است (مینیمم مطلق)، نمایانگر دوره تناوب گام است. اگر این نقطه را با η_1 نمایش دهیم و فرکانس نمونه‌برداری F_s باشد. دوره تناوب گام برابر با η_1/F_s و فرکانس گام برابر با F_s/η_1 خواهد بود.

۲-۲-۴ ویژگی دیرش

منظور از دیرش، طول بازه زمانی تلفظ یک واج است و محاسبه آن ساده است. البته مرز واج‌ها در یک واژه به‌طور دقیق مشخص نیست ولی با بررسی تغییرات انرژی و نرخ عبور از صفر می‌توان مرز واج‌ها را به‌طور تقریبی تخمین زد. دیرش با نرخ گفتار رابطه عکس دارد. برخی احساسات مانند خشم تأثیر زیادی روی دیرش برخی واج‌ها دارند. مثلاً وقتی شخصی عصبانی می‌شود تندتر صحبت می‌کند و نرخ گفتار (تعداد واج‌های تلفظ شده در واحد زمان) افزایش می‌یابد و به‌تبع آن دیرش کاهش می‌یابد.

۲-۲-۵ نرخ عبور از صفر

تابع انرژی به‌تنهایی کارایی لازم جهت تشخیص سکوت را ندارد. همچنین انرژی بعضی از همخوان‌ها آن‌قدر کم است که به‌صورت سکوت ظاهر می‌شوند. برای حل این مشکل از تابع نرخ عبور از صفر زمان-کوتاه متوسط[7] استفاده می‌شود. نرخ عبور از صفر به‌صورت تعداد تغییر علامت‌های سیگنال نرمال شده نسبت به صفر در واحد زمان تعریف می‌شود. محاسبه نرخ عبور از صفر سیگنال $s[.]$ در بازه زمانی که با پنجره مستطیلی L نقطه‌ای جدا می‌شود و انتهای این پنجره در نقطه m قرار دارد به‌صورت رابطه (۸) است [۲۹].

$$ZCR(x) = \frac{1}{2(N-1)}\sum_{i=1}^{N-1}|sign(x[i]) - sign(x[i-1])| \quad (۸)$$

و تابع علامت $sign$ به‌صورت رابطه (۹) می‌باشد.

$$sign(x[i]) = \begin{cases} 1, if\ x[i] > 0 \\ 0, \quad if\ x[i] > 0 \\ -1, \ if\ x[i] > 0 \end{cases} \quad (۹)$$

نرخ عبور از صفر در مورد تعداد زیادی از همخوان‌ها ازجمله اصوات سایشی دارای مقدار زیادی نسبت به واکه‌ها است. همچنین مقدار آن در سکوت به حداقل می‌رسد. بنابراین نرخ عبور از صفر معیاری برای جداسازی واکه‌ها از همخوان‌ها بوده و می‌تواند سکوت را از همخوان‌ها بخصوص همخوان‌های سایشی با انرژی

[1] Parallel PROCessing time-domain pitch detector
[2] Linear Predictive Coding
[3] Simplified Inverse FilTering
[4] AUTO Correlation method
[5] CEPstrum method
[6] Average Magnitude Difference Function
[7] Short- term average zero crossing

کم تمیز دهد. ترکیب دو تابع نرخ عبور از صفر و تابع انرژی می‌تواند با دقت بالایی کلمات و هجاها را از یکدیگر و حتی بعضی از واکه‌ها را از همخوان‌ها جدا کند. از این ویژگی معمولاً برای تغییر نوای گفتار استفاده نمی‌شود.

۲-۳ نگاشت نوای گفتار

جهت تغییر نوای گفتار باید سه منحنی انرژی، گام و دیرش مطابق با احساس مورد نظر با استفاده از تابع نگاشت تغییر یابد. برای این کار نیاز به پایگاه‌داده گفتار احساسی است تا از روی آن مدل تبدیل و تابع نگاشت برای هر ویژگی به دست آید. تاکنون در این زمینه پایگاه‌داده‌های گفتار احساسی مختلفی در زبان‌های مختلف مانند Persian ESD به زبان فارسی [۳۰] و پایگاه‌داده برلین به زبان آلمانی تهیه‌شده است. یک پایگاه‌داده گفتار احساسی شامل تلفظ چندین جمله به‌صورت خنثی و احساس‌های مختلف است.

در فاز اول که فاز یادگیری نام دارد سیگنال گفتار به‌صورتی نمایش داده می‌شود که ویژگی‌های آن قابل‌ویرایش باشد. گفتار عادی و گفتار احساسی متناظر، متناسب با زمان هم‌تراز شده تا بخش‌هایی با محتوای آوایی مشابه با یکدیگر مرتبط می‌شوند. تابع نگاشت بر روی ویژگی‌های هم‌تراز شده آموزش داده می‌شود. در فاز دوم که فاز تبدیل نامیده می‌شود، پس از استخراج ویژگی‌های گفتار خنثی، مدل نوای گفتار احساسی توسط تابع نگاشت حاصل می‌شود. برای به دست آوردن یا آموزش تابع نگاشت روش‌های مختلفی مانند نگاشت کتاب رمز[1]، نگاشت مخلوطی از خطی‌ها[2]، نگاشت شبکه عصبی، نگاشت فرهنگ لغت[3]، نگاشت پیچش فرکانسی[4]، تکنیک‌های تطبیق[5] و سایر نگاشت‌ها وجود دارد [۳۱].

۲-۴ اثر دادن مدل نوای گفتار

برای ایجاد گفتار احساسی باید مدل نوای گفتار احساسی را روی گفتار خنثی اعمال نمود. برای این کار پس از استخراج مدل نوای گفتار خنثی، لازم است آن را با مدل نوای گفتار احساسی نسبت به زمان هم‌تراز نمود. در مرحله بعد منحنی‌های سه ویژگی انرژی، گام و دیرش باید مطابق با مدل نوای احساسی تغییر یابد. برای تغییر هر ویژگی روش‌های مختلفی وجود دارد. برای تغییر منحنی انرژی جمله خنثی، انرژی فقط روی واکه‌ها و همخوان‌های واک‌دار به‌طور غیریکنواخت بر اساس موقعیت آن‌ها باید تغییر داد و انرژی مناطق بی‌واک نباید تغییر داد. جهت تغییر انرژی روی گفتار خنثی از مقیاس بندی استفاده می‌شود [۲۸].

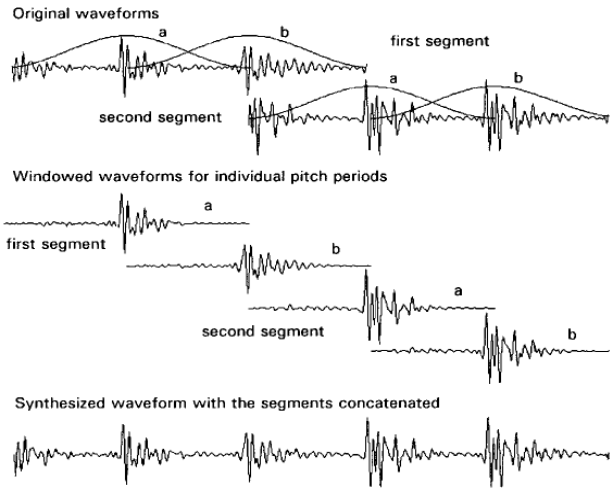
با تغییر نرخ نمونه‌برداری[6] می‌توان گام را تغییر داد ولی این روش نرخ گفتار را نیز تغییر می‌دهد. چندین روش برای تغییر مدل گام و دیرش از سوی پژوهشگران ارائه‌شده است. این روش‌ها نباید تغییر ناخواسته روی سایر ویژگی‌های گفتار مانند نرخ گفتار یا تغییر واکه‌ها ایجاد کنند. دو رویکرد غیرپارامتری و پارامتری وجود دارد. برای مثال دو نمونه از رویکرد غیرپارامتری عبارت‌اند از: روش «همپوشانی هم‌زمان گام و افزودن در دامنه زمان (TD-PSOLA)[7]» و روش «همپوشانی هم‌زمان گام و افزودن با استفاده از پیش‌گویی خطی (LP-PSOLA)[8]» [٤]. در رویکرد پارامتری، سیگنال گفتار به شکل پارامتریک نشان داده می‌شود. مدل «هارمونیک به‌اضافه نویز(HNM)[9]» [۱۹]، «گشتار گفتار و بازنمایی با استفاده از درون‌یابی تطبیقی طیف وزن‌دار[10] (STRAIGHT)» و «مدل‌سازی سینوسی» [۳۲] نمونه‌هایی از رویکرد پارامتری می‌باشند.

روش TD-PSOLA یک روش مرسوم در حوزه زمان برای تغییر هم‌زمان گام و دیرش است. در این روش ابتدا طول دوره تناوب و مقدار بیشینه در تمام بخش‌های متناوب از گفتار مشخص می‌شود که به آن نشانگرهای گام گفته می‌شود. از این نشانگرهای گام می‌توان برای تولید یک بخش پنجره شده از شکل موج برای هر دوره گام استفاده کرد. برای هر دوره تناوب، پنجره باید بر روی ناحیه بیشینه دامنه قرار داده شود. طول پنجره بیشتر از طول یک دوره تناوب تنظیم می‌شود تا مقداری همپوشانی میان پنجره‌های

[1] Codebook
[2] Mixture of linear
[3] Dictionary
[4] Frequency warping
[5] Adaptation techniques
[6] Resampling
[7] Time Domain-Pitch Synchronous Overlap and Add (TD-PSOLA)
[8] Linear Prediction Pitch Synchronous Overlap and Add (LP-PSOLA)
[9] Harmonic plus Noise Model (HNM)
[10] Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT)

مجاور ایجاد شود. با اضافه کردن بخش‌های پنجره شده با هم در نقاط علامت‌گذاری شده مربوط به بیشینه گام می‌توان سیگنال را بازسازی نمود. در شکل (۳) مراحل انجام این کار نمایش داده‌شده است [۲۱]؛ برای دو قطعه صدادار، نشانگرهای گام و محل قرارگیری پنجره در نمودار بالایی و خروجی‌های پنجره‌های اثر داده‌شده در نمودارهای میانی نشان داده‌شده‌اند. نمودار پایین شکل موجی را نشان می‌دهد که با استفاده از روش PSOLA با پیوستن آخرین پنجره پردازش‌شده قطعه اول به اولین پنجره پردازش‌شده قطعه دوم به‌دست‌آمده است.

با کاهش یا افزایش فضای میان نشانگرهای گام، مقدار گام را می‌توان به ترتیب افزایش یا کاهش داد. برای افزایش دیرش می‌توان پنجره‌هایی از گفتار با مرکزیت نشانگرهای گام و طول مناسب تکرار نمود یا با حذف پنجره‌ها یک‌درمیان یا چند در میان، دیرش کاهش می‌یابد.



شکل (۳): تجزیه سیگنال گفتار به دنباله‌ای از پنجره‌های هم‌زمان هم‌گام و افزودن آن‌ها به هم

۳- روش پیشنهادی

هدف از این پژوهش، بهبود گشتار صدا از حالت عادی به احساسی از طریق بازنگری در ویژگی‌ها و نحوه استخراج آن از سیگنال گفتار و کم نمودن خطای مدل نوای گفتار بوده است. یک ویژگی مهم در حوزه پردازش گفتار ویژگی انرژی است. روش‌هایی که برای محاسبه انرژی تاکنون ارائه‌شده‌اند عمدتاً در کاربردهایی مانند پردازش تصویر برای پیدا کردن مرز یک ناحیه در تصویر [۱۰] یا پوش سیگنال‌های رادیویی استفاده می‌شوند. این دسته از سیگنال‌ها برای انتقال نیاز به محیط مادی مانند هوا ندارند و در خلاء منتشر می‌شوند ولی سیگنال‌های صوتی از جنس مکانیکی بوده و در یک محیط فیزیکی مانند هوا منتشر می‌شوند. از سوی دیگر، سیگنال‌های صوتی توسط سیستم شنوایی انسان شامل گوش و اعصاب دریافت و پردازش می‌شوند. لذا روش‌هایی که برای محاسبه انرژی سیگنال‌های غیرصوتی استفاده می‌شوند ممکن است مناسب برای محاسبه انرژی سیگنال‌های صوتی نباشند.

در این راستا با توجه به مدل سیستم شنوایی انسان، تابع جدیدی برای محاسبه انرژی سیگنال‌های گفتار ارائه‌شده است. تابع انرژی پیشنهادی می‌تواند در سایر حوزه‌های پردازش گفتار مانند بازشناسی گفتار، تقلید صدا، سیستم‌های تبدیل متن به گفتار و غیره استفاده شود. تابع انرژی پیشنهادی بر روی گشتار گفتار احساسی پیاده‌سازی شد. نتایج آزمایشی بهبود کیفیت گفتار احساسی تولیدشده را نشان می‌دهد. کلیه الگوریتم‌ها و رویه‌های موردنیاز با استفاده از زبان متلب پیاده‌سازی و کارایی روش پیشنهادی مورد ارزیابی قرار گرفت. جهت ارزیابی، یک بار محاسبه انرژی با توابع معمول و یک بار با تابع انرژی پیشنهادی انجام و پیاده‌سازی شد.

۳-۱ محاسبه انرژی به روش پیشنهادی

در حوزه گشتار گفتار احساسی عموماً از سه ویژگی انرژی، گام و دیرش استفاده می‌شود. در بخش ۲-۲-۲ روش‌های معمول استخراج ویژگی انرژی شرح داده شد. در روش‌های مذکور برای محاسبه انرژی سیگنال، فقط پارامتر دامنه لحاظ شده است ولی سیستم شنوایی انسان علاوه بر دامنه به فرکانس نیز حساس است. به‌عنوان مثال مطابق با روش‌های مرسوم، انرژی یک سیگنال صوتی با فرکانس ۳۰ هرتز و دامنه A تقریباً برابر با انرژی یک سیگنال صوتی با فرکانس ۳۰۰ هرتز و دامنه A است، ولی صوت ۳۰ هرتزی شنیده نمی‌شود درحالی‌که صوت ۳۰۰ هرتزی به‌خوبی شنیده می‌شود. با بررسی‌هایی که انجام شد به نظر می‌آید سیستم

شنوایی انسان نسبت به نقاط اکسترمم[1] سیگنال‌های صوتی حساس است. طبق قضیه ارائه‌شده در [۳۳] ثابت‌شده است که تعداد نقاط اکسترمم در یک سیگنال سینوسی با فرکانس آن در رابطه است. مطابق با نتایج این مقاله، دستگاه شنوایی انسان به دو پارامتر دامنه و فرکانس صوت حساس است به‌نحوی‌که هر چه دامنه یا انرژی یک صدا بیشتر باشد بهتر شنیده می‌شود. به همین ترتیب، هر چه فرکانس سیگنال در یک بازه معین بیشتر باشد، بهتر شنیده می‌شود. بنابراین، اجزای یک سیگنال گفتاری که حاصل‌ضرب دامنه و فرکانس آن‌ها بیشتر از سایرین است، بهتر شنیده می‌شوند. همچنین، طبق قضیه مذکور، تعداد اکسترمم‌ها مربوط به مولفه‌ای است که حاصل‌ضرب دامنه در فرکانس آن بیشتر از بقیه باشد. بنابراین با استفاده از نقاط اکسترمم می‌توان جزئی از سیگنال گفتار را که سیستم شنوایی به آن حساس‌تر است را پیدا کرد. به‌عبارت‌دیگر، سیستم شنوایی به نرخ اکسترمم‌های سیگنال گفتار حساس است.

با توجه به مطالب فوق، روش پیشنهادی برای محاسبه انرژی سیگنال‌های گفتار، بر اساس مجموع مربعات اختلاف دامنه نقاط اکسترمم متوالی در بازه مورد نظر تعریف می‌شود؛ به این صورت که ابتدا تمام نقاط اکسترمم سیگنال گفتار $s[.]$ بر اساس تغییر علامت مشتق پیدا و اندیس آن‌ها را در بردار $e[.]$ ذخیره نموده، سپس مطابق رابطه (۱۰) انرژی سیگنال روی نمونه n محاسبه می‌شود.

$$E_n = \sum_{m=k1}^{k2-1} ((s[e[m+1]] - s[e[m]]).w[e[m] - n1])^2 \quad (۱۰)$$

در رابطه (۱۰)، $w[.]$ تابع مشخصه پنجره، n نقطه وسط پنجره، n_1 نقطه شروع پنجره، k_1 اندیس اولین و k_2 اندیس آخرین عنصر بردار $e[.]$ داخل پنجره قرارگرفته روی سیگنال است. اگر از پنجره مستطیلی استفاده شود، با حذف نقاط ابتدایی و انتهایی اضافی رابطه (۱۰) به رابطه (۱۱) تبدیل می‌شود.

$$E_n = \sum_{m=k1}^{k2-1} (s[e[m+1]] - s[e[m]])^2 \quad (۱۱)$$

در رابطه فوق پارامتر N طول پنجره و n شماره نمونه وسط پنجره است. با توجه به رابطه میان نقاط اکسترمم با فرکانس، محاسبه انرژی به روش پیشنهادی، هر دو ویژگی دامنه و فرکانس را در برمی‌گیرد. در روش پیشنهادی، مشهود است که مقدار انرژی محاسبه‌شده برای سیگنال‌های با فرکانس پایین که در محدوده شنوایی نیست. بسیار کم است. البته این نکته را هم باید در نظر داشت که مقدار انرژی برای اجزای فرکانسی بالای ۱۰ کیلوهرتز که به‌سختی شنیده می‌شوند، زیاد است. برای رفع این مشکل و همچنین رفع نویز قبل از محاسبه انرژی ابتدا باید توسط یک فیلتر پایین گذر، اجزای فرکانسی بالای ۱۰ کیلوهرتز از سیگنال گفتار حذف گردند.

ازآنجا که ویژگی گام در رابطه با فرکانس است، محاسبه انرژی به روش پیشنهادی تا حدودی این ویژگی را پوشش می‌دهد با این تفاوت که ویژگی گام فقط در مورد واکه‌ها و آواهای واک دار مطرح است ولی ویژگی انرژی پیشنهادی تحت تأثیر فرکانس همه اجزای گفتار قرار می‌گیرد.

۳-۲ استخراج مدل گفتار

در این پژوهش برای مدل‌سازی نوای گفتار از سه ویژگی انرژی، گام و دیرش استفاده‌شده است. منظور از استخراج مدل گفتار، به دست آوردن و ذخیره نمودار تغییرات[2] ویژگی‌های مورد نظر در طول زمان ادای یک جمله است. برای استخراج مدل گام از روش اختلاف متوسط دامنه (AMDF) شرح داده‌شده در بخش ۲ استفاده شد. استخراج منحنی انرژی یک بار با استفاده از رابطه (۳) و یک بار با استفاده از روش پیشنهادی مطابق با رابطه (۱۰) جهت مقایسه انجام شد. از پنجره‌گذاری هنینگ مطابق با رابطه (۱) برای بخش‌بندی گفتار و استخراج ویژگی‌ها استفاده شد.

۳-۳ تابع نگاشت نوای گفتار احساسی

پس از اینکه مدل نوای گفتار خنثی استخراج شد، باید تغییراتی در مدل داده شود تا مدل نوای احساسی حاصل شود. برای آموزش تابع نگاشت برای هر ویژگی، نیاز به پایگاه‌داده احساسی است. هرچند که موضوع این پژوهش محدود به گشتار گفتار احساسی

[1] Extrema Points

[2] Contour

به زبان خاصی نیست ولی از گفتار زبان فارسی به دلیل در دسترس بودن شنوندگان فارسی‌زبان جهت ارزیابی روش پیشنهادی استفاده‌شده است. در اینجا از پایگاه‌داده گفتار احساسی Persian ESD استفاده‌شده است [۳۰] که یک پایگاه‌داده نسبتاً کوچک بوده و در کشور آلمان تهیه‌شده است. پایگاه‌داده Persian ESD شامل یک مجموعه از ۹۰ جمله فارسی است که توسط دو گوینده یک مرد و یک زن فارسی‌زبان و هر جمله به‌صورت خنثی و با پنج احساس خشم، شادی، غم، ترس و نفرت تلفظ شده است. این جملات توسط ۱۱۲۶ شنونده فارسی‌زبان تائید شده‌اند.

برای ارزیابی دقیق‌تر روش پیشنهادی برای محاسبه انرژی، مسئله را یک مقدار محدود کردیم. در اینجا روی جملاتی کار کردیم که در پایگاه‌داده معادل آن وجود داشت. ۵ زوج جمله خنثی و احساسی با دو جنسیت مرد و زن برای ۵ احساس مختلف از پایگاه‌داده Persian ESD انتخاب شد (جمعاً ۱۰۰ جمله). ابتدا کلیه جملات منتخب بر اساس ویژگی‌های نقاط اکسترمم، انرژی و نرخ عبور از صفر به‌صورت خودکار در سطح واج و هجا تقطیع و سپس برای دقت بیشتر به‌صورت دستی بررسی و تصحیح شد. برای به دست آوردن تابع نگاشت دیرش، مدت هر واج یا سکوت در جمله خنثی و احساسی متناظر محاسبه سپس نسبت آن‌ها در بردار مربوط به آن جمله ذخیره شد. تابع نگاشت دو ویژگی انرژی و گام، برای هر دو هجای متناظر در گفتار خنثی و احساسی محاسبه شد. برای هر هجا نیاز به یک تابع نگاشت منحنی جداگانه برای هر کدام از این دو ویژگی است. در اینجا به جهت کوچک بودن پایگاه‌داده از نگاشت مخلوطی از خطی‌ها استفاده شد و نتیجه آن به‌صورت برداری از بردارها برای هر جمله و هر ویژگی ذخیره شد. پس از به دست آوردن تابع نگاشت برای هرکدام از ویژگی‌ها، برای نگاشت نوای جمله خنثی به احساسی از تبدیل خطی استفاده گردید.

۳-٤ اثر دادن مدل نوای گفتار

پس از استخراج مدل نوای گفتار خنثی، محدوده واج‌ها و هجاها در جمله خنثی با استفاده از ویژگی‌های نقاط اکسترمم، انرژی و نرخ عبور از صفر تعیین گردید. در مرحله بعد مدل نوای گفتار خنثی در سطح واج و هجا با مدل نوای احساسی نسبت به زمان هم‌تراز گردید. تابع نگاشت حاصل از مرحله آموزش، با استفاده از روش TD-PSOLA روی ویژگی گام و دیرش هم‌زمان و توام اثر داده شد. برای تغییر انرژی جمله خنثی، انرژی روی سیگنال گفتار بر اساس تابع نگاشت از طریق مقیاس بندی تغییر داده شد.

٤- پیاده‌سازی روش پیشنهادی

با توجه به اینکه فرکانس نمونه‌برداری ٤٤۱۰۰ و با عمق ۱٦ بیتی مقادیر استاندارد، با کیفیت خوب و رایج برای ضبط گفتار انسان است در اینجا از این مقادیر برای پیاده‌سازی استفاده‌شده است. قبل از هرگونه پردازش لازم است سیگنال گفتار پنجره‌گذاری شود. طول یک پنجره را معمولاً بین ۱۰ تا ۳۰ میلی‌ثانیه اختیار می‌کنند. در اینجا بر اساس سعی و خطا پنجره هنینگ به طول ۲۰ میلی‌ثانیه مناسب دیده شد و روی سیگنال گفتار استفاده شد. روش پیشنهادی به کمک توابع کتابخانه‌ای نرم‌افزار Praat و همچنین جعبه‌ابزارها و توابع MATLAB روی پایگاه‌داده گفتار احساسی زبان فارسی Persian ESD[1] پیاده‌سازی و بررسی شد. گوینده مورد نظر یکی از جملات منتخب را به‌صورت عادی تلفظ می‌کند. گشتار گفتار احساسی شامل فاز «آموزش یا به دست آوردن تابع نگاشت» و فاز «اجرا» است. در فاز آموزش، تابع نگاشت مطابق بخش قبل از روی ویژگی‌های انرژی، گام و دیرش ازجمله خنثی و جمله احساسی متناظر (برای ۵۰ زوج جمله) استخراج و در بردارهای مربوطه ذخیره شد.

فاز اجرای الگوریتم پیشنهادی گشتار گفتار احساسی برای بررسی کارایی تابع انرژی پیشنهادی:

۱) اعمال پنجره هنینگ به طول ۲۰ میلی‌ثانیه روی سیگنال گفتار خنثی ورودی.
۲) استخراج ویژگی‌های گام، انرژی، تعداد نقاط اکسترمم و نرخ عبور از صفر از هر پنجره.
۳) تقطیع خودکار جمله ورودی با استفاده از ویژگی‌های حاصل از قدم قبل، در سطح واج و هجا.
٤) استخراج منحنی انرژی، گام و دیرش از گفتار تقطیع شده

[1] Persian Emotional Speech Database

۵) هم‌تراز کردن جمله خنثی تقطیع شده با جمله احساسی مربوطه در پایگاه‌داده نسبت به زمان.

٦) ایجاد تناظر یک‌به‌یک بین جمله خنثی و مدل نوای گفتار احساسی.

۷) نگاشت مدل استخراج‌شده به مدل احساسی با استفاده از تابع نگاشت برای هرکدام از ویژگی‌ها

۸) اثر دادن مدل قدم قبل روی واج‌ها و هجاهای متناظر در جمله خنثی تقطیع شده و تولید گفتار احساسی مورد نظر.

فاصله سکوت میان هجاها نیز به‌عنوان یک ویژگی در نظر گرفته شد زیرا در احساس نقش دارند. البته این ویژگی در سطح پنجره یا هجا نیست و در سطح جمله است و به‌صورت جداگانه پردازش شد.

۵- ارزیابی روش پیشنهادی

روش‌های مختلفی برای ارزیابی یک سیستم تولید یا گشتار گفتار وجود دارد. از یک دیدگاه این روش‌ها را می‌توان به دو دسته عینی[1] و ذهنی[2] تقسیم کرد. در ارزیابی عینی با استفاده از معیارها و روش‌های علمی بدون دخالت نظر شخصی ارزیابی انجام می‌شود. سیگنال گفتار را می‌توان به‌صورت یک سری زمانی فرض کرد و بر اساس یکی از روش‌های اندازه‌گیری شباهت، میزان شباهت گفتار تولیدشده را با گفتار واقعی سنجید [۳۳]. در ارزیابی ذهنی با شنیدن صدای خروجی سیستم توسط افراد مختلف، عملکرد آن مورد ارزیابی قرار می‌گیرد. آزمون‌های ارزیابی کلی کیفیت مانند آزمون میانگین امتیازات نظردهی MOS[3]، و آزمون CMOS[4] رایج‌ترین می‌باشند[۳٤].

در روش MOS، شنوندگان با توجه به کیفیت گفتار، یک نمره بین ۱ تا ۵ می‌دهند. هم کیفیت صدا و هم شباهت به گفتار احساسی موردنظر را می‌توان ارزیابی کرد. نمرات به این صورت است:

۱ = بد، ۲ = ضعیف، ۳ = متوسط، ٤ = خوب، ۵ = عالی.

روش CMOS می‌تواند به‌طور مستقیم کیفیت گفتار دو سیستم گشتار گفتار را مقایسه کند. در این روش از شنونده خواسته می‌شود گفتار بهتر را انتخاب کند. اندازه‌گیری به‌صورت درصد روی نظرات کل شنوندگان است. امتیازدهی در مقیاس ۵ به این صورت است: ۵ = قطعاً بهتر، ٤ = بهتر، ۳ = شبیه هم، ۲ = بدتر، ۱ = قطعاً بدتر [۳۱].

روش‌های مذکور، کیفیت و میزان قابل‌فهم بودن گفتار را با توجه به نظر تعدادی از شنوندگان مورد ارزیابی قرار می‌دهند و نیاز به مجموعه داده خاصی جهت ارزیابی ندارند لذا از روش‌های متداول ارزیابی تولید گفتار در تمام زبان‌ها هستند. با توجه به اینکه در این جا می‌خواهیم تابع انرژی پیشنهادی را با روش مرسوم مقایسه کنیم از آزمون CMOS استفاده‌شده است. ویژگی انرژی برای هر احساس یک بار با استفاده از رابطه (۳) و بار دیگر با استفاده از رابطه پیشنهادی (۱۰) محاسبه شد، سپس الگوریتم ارائه‌شده در بخش قبل برای تولید گفتار احساسی به کار گرفته شد.

۵-۱ تحلیل نتایج

جهت ارزیابی روش پیشنهادی، ۵ شنونده زن و ۵ شنونده مرد شرکت کردند. این شنوندگان به‌صورت تصادفی و با سنین مختلف انتخاب شدند و فقط یک توضیحات کلی به آن‌ها داده شد، لیکن در این مرحله در رابطه با تشخیص احساسات هیچ آموزش قبلی به آن‌ها داده نشد. شنوندگان، عملکرد سیستم تبدیل گفتار احساسی با تابع انرژی مرسوم و تابع انرژی پیشنهادی را مطابق با آزمون CMOS ارزیابی کردند. نظرات روی گفتار تولیدشده توسط هر دو روش جمع‌آوری گشت. جدول (۱) خلاصه نتایج به‌دست‌آمده را نشان می‌دهد.

در جدول (۱) ستون‌ها به ترتیب در رابطه با جملات مرتبط با احساس‌های خشم، شادی، غم، ترس و نفرت می‌باشند. سطرهای ۱ تا ۵ مربوط به ارزیابی شنوندگان مرد و سطرهای ٦ تا ۱۰ مربوط به ارزیابی شنوندگان زن است. به هر شنونده برای هر احساس ۵ جمله آزمایشی داده‌شده و نظرات ثبت‌شده است. درنهایت نیز میانگین هر ستون که متناظر با یک احساس است در سطر آخر درج‌شده‌اند. مشاهده می‌شود که میانگین نظرات برای تمام احساسات بالای ۳ است، و این به معنای بهتر بودن روش پیشنهادی برای محاسبه انرژی روی تبدیل گفتار احساسی است. قبلاً ضمن

[1] Objective
[2] Subjective
[3] Mean Opinion Score
[4] Comparative MOS

انتشار نتایج اولیه پژوهش مربوط به این تحقیق یک اثبات ریاضی ارائه‌شده است [۳۵].

جدول (۱): امتیازات ۱۰ شنونده برای پنج نوع جمله احساسی

	خشم	شادی	غم	ترس	نفرت
۱	۳/۱	۳/۳	۳	۳/۲	۳/۵
۲	۳/۲	۳/۳	۳/۲	۳/۲	۳/۳
۳	۳/۳	۳/۲	۳/۲	۳/۳	۳/۵
٤	۳	۳/۱	۲/۹	۳/۱	۳/۲
٥	۲/۸	۳	۳	۳/۱	۳/۲
٦	۲/۹	۲/۹	۲/۸	۳	۳
۷	۳/۲	۳/۳	۳	۳/۱	۳/۳
۸	۳/۳	۳/۲	۳/٤	۳/٥	۳/٤
۹	۳/٤	۳/۸	۳/۲	۳/۷	۳/۱
۱۰	۳/۱	۳	۲/۹	۳/۲	۳
میانگین	۳/۱۳	۳/۲۱	۳/۰٦	۳/۲٤	۳/۲٥

٦- نتیجه‌گیری و پیشنهادها

پژوهش‌هایی که اخیراً در حوزه پردازش سیگنال‌های گفتار انجام می‌شود در زمینه مدل‌سازی یا استخراج خودکار ویژگی‌ها است و کمتر به نحوه استخراج ویژگی‌های بنیادین مانند انرژی پرداخته‌شده است. این پژوهش نشان داد که در هر حوزه برای استخراج ویژگی‌ها لازم است یک بازنگری انجام شود؛ زیرا روش استخراج یا تعریف یک ویژگی برای یک کاربرد ممکن است برای کاربردهای دیگر مؤثر نباشد. این پژوهش نگاه جدیدی در حوزه پردازش گفتار در رابطه با بازنگری ویژگی‌ها در اختیار می‌گذارد. فرمول پیشنهادی برای استخراج ویژگی انرژی، کارایی خود را در تبدیل گفتار احساسی به‌خوبی نشان داد. پیشنهاد می‌شود کارایی این روش در سایر حوزه‌های پردازش صوت و گفتار مانند بازشناسی گفتار بررسی شود. بازه فرکانس صوت قابل‌شنیدن برای انسان بین ۵۰ الی ۲۰۰۰۰ هرتز است ولی سیستم شنوایی انسان روی فرکانس در بعضی بازه‌ها حساس‌تر است. برای کارایی بیشتر بهتر است بازه مذکور به چند زیربازه تجزیه‌شده و در هر زیربازه با توجه به میزان حساسیت شنوایی انسان، یک ضریب مناسب برای فرمول پیشنهادی در نظر گرفته شود. به‌عنوان ادامه پژوهش می‌توان آزمایش‌ها با تعداد بیشتری از شنوندگان انجام شود و نیز مقایسه می‌تواند با روش‌هایی مانند Mel-band energy [۱]، Spectral centroid / roll-off، یا Wavelet energy صورت بگیرد.

با توجه به اینکه در این پژوهش تنها دسترسی به شنوندگان فارسی‌زبان داشتیم، ارزیابی نتایج روی زبان فارسی انجام شد. این روش روی پایگاه‌داده برلین به زبان آلمانی نیز پیاده‌سازی شد ولی با توجه به در دسترس نبودن شنوندگان آلمانی‌زبان و اینکه احساسات تا حدی وابسته به زبان می‌باشند ارزیابی روی زبان آلمانی انجام نشد؛ ولی با توجه به شباهت‌های زیاد احساسات در زبان‌های مختلف، انتظار می‌رود این روش روی زبان‌های دیگر نتایج قابل قبولی داشته باشد.

References

[1] T. Qi, S. Wang, C. Lu and T. Song, "PromptEVC: Controllable Emotional Voice Conversion with Natural Language Prompts," in Interspeech, Rotterdam, The Netherlands, 2025.

[2] K. Zhou, B. Sisman, R. Liu and H. Li, "Emotional voice conversion: Theory, databases and ESD," Speech Communication, vol. 137, pp. 1-18, 2022.

[3] L. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," The Journal of the Acoustical Society of America, 1993.

[4] K. Waghmare, S. Kayte and B. Gawali, "Analysis of Pitch and Duration in Speech Synthesis using PSOLA," Communications on Applied Electronics (CAE), vol. 4, no. 4, pp. 10-18, 2016.

[5] P. Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," International Journal of Human-Computer Studies, vol. 59, pp. 157-183, 2003.

[6] S. S. Sadeghi, H. Khotanlou and M. R. Mahand, "Automatic Persian Text Emotion Detection using Cognitive Linguistic and Deep Learning," Journal of Artificial Intelligence and Data Mining (JAIDM), vol. 9, no. 2, pp. 169-179, 2021.

[7] N. Esfandian, "Phoneme Classification using Temporal Tracking of Speech Clusters in spectro- temporal domain," International Journal of Engineering (IJE), IJE Transactions A: Basics, vol. 33, no. 1, pp. 105-111, 2020.

[8] M. Aliabadi, R. Golmohammadi, M. Mansoorizadeh, H. Khotanlou and A. O. Hamadani, "An empirical technique for predicting noise exposure level in the typical embroidery workrooms using artificial neural networks," Applied Acoustics, vol. 74, p. 364–374, 2013.

[9] M. Karami Mollaei and M. Eshaghi, "A NEW ALGORITHM FOR VOICE ACTIVITY DETECTION BASED ON WAVELET PACKETS," International Journal of Engineering(IJE), IJE Transactions A: Basics, vol. 22, no. 3, pp. 225-232, 2009.

[10] A. A. Kiaei and H. Khotanlou, "Segmentation of Medical Images using Mean Value Guided Contour," Medical Image Analysis, 2017.

[11] S. Özaydın, "Examination of Energy Based Voice Activity Detection Algorithms for Noisy Speech Signals," European Journal of Science and Technology, pp. 157-163, 2019.

[12] K. Aghajani and I. Esmaili Paeen Afrakoti, "Speech Emotion Recognition Using Scalogram Based Deep Structure," International Journal of Engineering (IJE), IJE TRANSACTIONS B: Applications, vol. 33, no. 2, pp. 285-292, 2020.

[13] Y. Stylianou, "VOICE TRANSFORMATION: A SURVEY," in IEEE International Conference on Acoustics Speech and Signal Processing, 2009.

[14] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," Computer Speech and Language, vol. 20, p. 441–467, 2006.

[15] L. Mary, Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition, 2 ed., SpringerBriefs in Speech Technology, 2019.

[16] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models," in 2005 IEEE International Symposium on Circuits and Systems, Kobe, Japan, 2005.

[17] D. Verma, S. K. Barnwal, A. Barve, M. K. J. Kannan, R. Gupta and R. Swaminathan, "Multimodal Sentiment Sensing and Emotion Recognition Based on Cognitive Computing Using Hidden Markov Model with Extreme Learning Machine," International Journal of Communication Networks and Information Security, vol. 14, no. 2, pp. 155-167, 2022.

[18] M. K. Reddy and K. S. Rao, "Excitation Modeling Method Based on Inverse Filtering for HMM-Based Speech Synthesis," Machine Intelligence and Signal Analysis, vol. 748, pp. 85-91, 2019.

[19] J. B. Singh and P. K. Lehana, "Emotional speech analysis using harmonic plus noise model and Gaussian mixture model," International Journal of Speech Technology, vol. 22, p. 483–496, 2019.

[20] S. Karimi and M. H. Sedaaghi, "How to categorize emotional speech signals with respect to the speaker's degree of emotional intensity," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 24, p. 1306–1324, 2016.

[21] J. Holmes and W. Holmes, Speech Synthesis and Recognition, Second ed., London: Taylor & Francis, 2001.

[22] L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing, Boston: Now the essence of knowledge, 2007.

[23] M. Mansoorizadeh and N. Moghaddam Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," Multimedia Tools and Applications, vol. 49, p. 277–297, 2010.

[24] A. V. Oppenheim, A. S. Willsky and H. Nawab, Signals and systems, New Jersey: Prentice-Hall, 1996.

[25] S. Hadiyoso, I. D. Irawati and A. Rizal, "Epileptic Electroencephalogram Classification using Relative Wavelet Sub-band Energy and Wavelet Entropy," International Journal of Engineering(IJE), Transactions A: Basics, vol. 34, no. 1, pp. 75-81, 2021.

[26] M. Jalil, A. Butt and A. Malik, "Short-Time Energy, Magnitude, Zero Crossing Rate and Autocorrelation Measurement for Discriminating Voiced and Unvoiced segments

of Speech Signals," in International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), Konya, Turkey, 2013.

[27] M. Hamidi and M. Mansoorizade, "EMOTION RECOGNITION FROM PERSIAN SPEECH WITH NEURAL NETWORK," International Journal of Artificial Intelligence & Applications (IJAIA), vol. 3, no. 5, pp. 107-112, 2012.

[28] H.K. Vydana, S.R. Kadiri and A.K. Vuppala, "Vowel-Based Non-uniform Prosody Modification for Emotion Conversion," Circuits, Systems and Signal Processing, vol. 35, p. 1643–1663, 2016.

[29] P. R. Hill, Audio and Speech Processing with MATLAB, New York: CRC Press, 2019.

[30] N. Keshtiari, M. Kuhlmann, M. Eslami and G. Klann-Delius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," Behavior Research Methods, vol. 47, p. 275–294, 2015.

[31] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," Speech Communication, vol. 88, p. 65–82, 2017.

[32] K. S. Rao, Predicting Prosody from Text for Text-to-Speech Synthesis, New York: Springer Briefs in Electrical and Computer Engineering, 2012.

[33] M. Nikzar, H. Khotanlou and M. Dezfoulian, "THE RELATIONSHIP BETWEEN THE NUMBER OF EXTREMA OF COMPOUND SINUSOIDAL SIGNALS AND ITS HIGH-FREQUENCY COMPONENT," Journal of Mahani Mathematical Research (JMMR), vol. 13, no. 1, pp. 181-195, 2023.

[34] A. Salarpour and H. Khotanlou, "An Empirical Comparison of Distance Measures for Multivariate Time Series Clustering," International Journal of Engineering (IJE), IJE TRANSACTIONS B: Applications, vol. 31, no. 2, pp. 250-262, 2018.

[35] R. C. Streijl, S. Winkler and D. S. Hands, "Mean Opinion Score (MOS) revisited:Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, p. 213–227, 2016.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 4, No. 1, Spring & Summer 2026, pp. 167 - 181

Improving the quality of emotional speech conversion using an efficient energy function

Majid Nikzar [1], Hassan Khotanlou [2]*, Mirhossein Dezfoulian[3]

[1] PhD Student, Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran

[2] Professor, Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran

[3] Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran

Article Information

Original Research Paper

Received:
2025 December 06

Accepted:
2026 January 20

Keywords:
Emotional Speech Transformation, Speech Synthesis, Speech Prosody Modeling, Sound Intensity, Energy Spectrum

Corresponding Author*:
khotanlou@basu.ac.ir

Abstract

An important issue in speech processing is how to beneficially add or change the emotion of a speech. Among the important challenges is calculating the exact value of the main features including energy, pitch, and duration. This research shows that energy feature extraction can be promoted using a proper energy function. The proposed method for energy calculation is based on the sensitivity of the human hearing system at different frequencies. In this method, the energy is calculated based on the distance between the extrema points in the speech signal, which is related to the amplitude and frequency. To evaluate the efficiency, a simple emotional speech conversion system was implemented using the proposed energy function on the Persian ESD emotional speech dataset and the results are compared with the conventional energy function. The experimental results based on the CMOS assessment show that using the proposed method produced better results in compare with the state of the art methods.

: 10.22034/ABMIR.2026.24063.1196

E-ISSN: 2821-2037 /The Author 2026. Published by Yazd University This is an open access article under the CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).