تخمین سریع ضرایب پیچش در هنجارسازی طول مجرای صوتی با استفاده از امتیاز به‌دست‌آمده از مدل‌سازی جنسیت

یاسر شکفته[۱]، محمدمحسن گودرزی[۲]، حسن قلی‌پور[۳]، سیدجهانشاه کبودیان[۴]، فرشاد الماس‌گنج[۵]، شقایق رضا[۶] و ایمان صراف رضایی[۷]

[۱و۲و۳و۶و۷] گروه پردازش صوت و زبان طبیعی، پژوهشکده پردازش داده، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی، تهران، ایران

[۴] گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکدهٔ فنی و مهندسی، دانشگاه رازی، کرمانشاه، ایران

[۵] دانشکدهٔ مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

یکی از مشکلات عمدهٔ سامانه‌های خودکار بازشناسی گفتار (ASR)، تنوعات موجود در بین گوینده‌ها، کانال انتقال داده و محیط است که به‌علت وجود این تنوعات، کارایی این سامانه‌ها در شرایط کاربردی مختلف، به‌شدت تغییر می‌کند. مقاوم سازی سامانه‌های بازشناسی جهت مقابله با این تغییرات از جمله مسائل حال حاضر در حوزهٔ بازشناسی گفتار است. از جمله عواملی که باعث کاهش کارایی سامانه‌ها می‌شود، تمایز مشخصات صوتی آواهای یکسانِ تولیدشده از گوینده‌های مختلف است. یکی از عوامل اصلی این مشکل ناشی از تفاوت موجود در طول مجرای صوتی (VTL) بین گوینده‌های مختلف است. روش هنجارسازی طول مجرای صوتی (VTLN) از روش‌های رایج برای رفع این مشکل است که در آن برای هر گوینده یک ضریب پیچش فرکانسی تعیین می‌شود. در این مقاله روش متداول تعیین ضریب پیچش با رویکرد مبتنی بر جستجو در یک سامانهٔ بازشناسی گفتار پیوسته فارسی مبتنی بر مدل مخفی مارکوف معرفی و مشکلات محاسباتی استفاده از این روش شرح داده شده است. درنهایت روشی مبتنی بر رگرسیون خطی از روی امتیازِ محاسبه‌شده از مدل‌سازی جنسیت جهت تخمین ضرایب پیچش پیشنهاد شده است که منجر به کاهش قابل ملاحظهٔ هزینهٔ محاسباتیِ روش مبتنی بر جستجو می‌شود. علاوه‌براین، نتایج آزمایش‌ها بر روی دادگان آزمون گفتار تلفنی محاوره‌ای، بیان‌گر بهبود ۰/۵۴ درصدی دقت تشخیص کلمه روش پیشنهادی نسبت به روش متداول مبتنی بر جستجو است.

واژگان کلیدی: بازشناسی گفتار، هنجارسازی طول مجرای صوتی، تشخیص جنسیت، رگرسیون خطی، ضریب پیچش فرکانسی.

۱- مقدّمه

گفتار، روشی طبیعی برای ارتباط میان انسان‌ها است و می‌تواند نقش یک واسط را بین کاربر انسانی و ماشین ایفا کند. سامانه‌های خودکار بازشناسی گفتار (ASR) با هدف تبدیل گفتار به متن به وجود آمده‌اند. ورودی این سامانه‌ها، سیگنال گفتار و خروجی آن‌ها متن متناظر با گفتار ورودی است. این سامانه‌ها کاربردهای گوناگونی از قبیل سامانه‌های خودکار تلفنی، سامانه‌های خودکار دیکته‌نویسی و غیره دارند. کارایی این سامانه‌ها به میزان قابل توجهی وابسته به هماهنگی میان دادگان تعلیم و آزمایش آنها است؛ لذا از مشکلات عمده‌ی موجود در این سامانه‌ها می‌توان به تنوعات موجود در بین گوینده‌ها، کانال انتقال داده و محیط اشاره کرد که به‌علّت وجود این تغییرات، کارایی این سامانه‌ها در شرایط کاربردی به‌شدت تغییر می‌کند. بنابراین مقاوم‌سازی سامانه‌های بازشناسی جهت مقابله با این تغییرات از جمله مسائل حال حاضر حوزهٔ بازشناسی گفتار است. از جمله منابعی که باعث کاهش کارایی سیستم‌ها می‌شود، تغییرات بین آواهای تولیدشده از گوینده‌های مختلف است. این مسئله که تأثیر به‌سزایی در به‌دست‌آمدن ویژگی‌های متمایز از سیگنال صوتی آواهای یکسان دارد، نه‌تنها ناشی از تغییرات موجود در فیزیک دستگاه‌های تولید صوت و آناتومی متفاوت آنهاست، بلکه ناشی از عوامل دیگری همچون تغییرات زبانی بین گوینده‌های مختلف مانند نحوهٔ تلفظ، لهجه، استرس و ... است. یکی از عوامل اصلی تغییرات بین گوینده‌های مختلف، ناشی از این است که گویندگان

مختلف دارای طول مجرای صوتی[1] (VTL) متفاوتی هستند. این تفاوت باعث می‌شود صدای افراد مختلف از لحاظ شنیداری با هم متفاوت باشد. یک روش متداول در کاهش تنوعات گوینده، بهبود تحلیل بانک فیلترها است. در این روش که هنجارسازی طول مجرای صوتی[2] (VTLN) نام دارد، هدف جبران‌سازی اختلاف طول مسیر صوتی بین گویندگان مختلف است. حذف تغییرات ناشی از طول لولهٔ صوتی با اعمال پیچش فرکانسی روی طیف سیگنال گفتار گوینده انجام‌پذیر است. مقالات متعددی (لی، ۱۹۹۸؛ وگمان، ۱۹۹۶؛ پیتز، ۲۰۰۵؛ ایدی، ۱۹۹۶) به پیاده‌سازی و بررسی عملکرد روش‌های مختلف پیچش فرکانسی پرداخته‌اند. اگرچه تطبیق به گوینده VTLN موضوعی با قدمت حدود دو دهه در حوزهٔ پردازش و بازشناسی گفتار است؛ اما به‌علت کارایی اثبات‌شدهٔ آن، همچنان به‌عنوان یکی از فرآیندهای مؤثر در سامانه‌های مطرح و به‌روز بازشناسی گفتار (سائون، ۲۰۱۵؛ میترا، ۲۰۱۴؛ ساینا‌س، ۲۰۱۴) به کار می‌رود.

روش VTLN به‌طورمعمول با مقیاس‌دهی[3] محور فرکانس (وانگ، ۲۰۰۷؛ کویی، ۲۰۰۶؛ آسرو، ۱۹۹۱) و یا با شیفت‌دادن فرکانس‌های مرکزی فیلترهای میان‌گذر فیلتربانک (لی، ۱۹۹۸) پیاده‌سازی می‌شود. هر دو روش را می‌توان با استفاده از ضریب پیچش[4] انجام داد که این ضریب برای هر گوینده باید جداگانه محاسبه شود.

در مرجع (پیتز، ۲۰۰۵) نشان داده شده است که پیچش فرکانسی در حوزه فرکانس پیوسته معادل با یک انتقال خطی[5] (LT) در حوزهٔ کپسترال پیوسته است و از این رو می‌توان به جای مقیاس‌دهی محور فرکانس، تبدیل خطی متناظر را بر روی تبدیل کپسترال اعمال کرد. البته لازمهٔ این کار این است که یک رابطه تحلیلی برای انتقال از حوزهٔ فرکانس به حوزهٔ کپسترال به‌دست آورد و این رابطه به‌ناچار در حوزهٔ فرکانس پیوسته تعریف می‌شود. البته در مرجع (اومش، ۲۰۰۵) نشان داده شد که با فرض محدودکردن بازه فرکانسی مورد نظر به یک بازهٔ کوچک، این تبدیل خطی در حوزه فرکانس گسسته نیز به‌صورت تقریبی برقرار خواهد بود. به‌طور مشابه، در (کویی، ۲۰۰۶) نشان داده شده که انتقال فیلترهای مثلثی فیلتربانک مل در ضرایب MFCC با تقریب خوبی معادل با اعمال یک تبدیل خطی بر روی ضرایب کپسترال نهایی است و از آن برای هم‌ترازکردن فورمنت‌های[6] گفتار کودکان با بزرگسالان استفاده شده است. همچنین، در (مکدونوهت، ۲۰۰۴) این تبدیل خطی به‌صورت یک فیلتر تمام‌گذر[7] پیاده‌سازی شده است.

مزیت اصلی استفاده از تبدیل خطی در این است که فرآیند استخراج ویژگی تغییر نخواهد کرد و کافی است تا تبدیل مذکور را بر روی ویژگی‌های موجود اعمال کرد. این مزیت به‌خصوص در سامانه‌هایی که عملیات استخراج ویژگی در سامانه کاربر انجام شده اما عملیات بازشناسی در یک سامانهٔ دیگر (به‌عنوان‌مثال سرور مرکزی) انجام می‌شود بروز پیدا می‌کند. البته دقت این شیوه اندکی پایین‌تر از روش VTLN عادی است که در آن پیچش فرکانس به‌طورمستقیم در حوزهٔ فرکانس و یا بر روی فیلتربانک اعمال می‌شود (پانچاپاگسان، ۲۰۰۹). همچنین کارایی این شیوه در شرایط وجود نوفه کاهش پیدا می‌کند و لذا کوشش‌هایی برای بهبود آن صورت گرفته است (سنند، ۲۰۱۰).

از طرف دیگر، می‌توان از این تبدیل خطی، به جای اعمال بر روی ویژگی‌ها برای اصلاح مشخصه‌های مدل سامانه بازشناسی استفاده کرد. در این شیوه مشابه روش تطبیق به گوینده «حداکثر درست‌نمایی رگرسیون خطی[8] (MLLR)» عمل می‌شود و تبدیل خطی مذکور بر روی بردارهای میانگین مدل‌های HMM به کار رفته در سامانه بازشناسی اعمال می‌شود (کلیز، ۱۹۹۸؛ دینگ، ۲۰۰۲؛ ایموری، ۲۰۰۱). بدین ترتیب، این روش‌ها را می‌توان به‌عنوان یک مورد خاص از MLLR تفسیر کرد. البته می‌توان روش VTLN عادی را در کنار روش تطبیق به گوینده MLLR به کار برد و نشان داده شده است که این دو روش مکمل یکدیگر نیز هستند (گیولیانی، ۲۰۰۶).

در مرجع (باباعلی، ۱۳۸۶) روش‌های متداول استخراج و اعمال ضریب پیچش بر روی طیف گفتار جهت هنجارسازی اثر طول مسیر صوتی مورد بررسی و مقایسه قرار گرفته‌اند. در مرجع (عزیزی، ۱۳۹۱) یک سامانه بازشناسی کلمات مجزا بررسی شده است که هدف آن افزایش کارایی سامانه بازشناسی گفتار کودکان با استفاده از روش هنجارسازی طول مسیر صوتی است. این سامانه بازشناسی، برای استفاده در نرم‌افزار گفتاردرمانی ایجاد شده بود، به‌طوری که نرم‌افزار نهایی با استفاده از سامانه

[1] Vocal tract length
[2] Vocal tract length normalization
[3] Scaling
[4] Warping factor
[5] Linear Transform
[6] Formants
[7] All-pass
[8] MLLR

بازشناسی، درست یا نادرست‌بودن تلفظ کودک را تشخیص دهد و تلاش می کند تا با استفاده از بازخوردهای مناسب، گفتار کودک بهبود یابد. در مرجع (تبریزی، ۱۳۸۸) نیز روش VTLN برای بهبود کارایی بازشناسی گفتار کودکان فارسی زبان پیاده‌سازی شده است. همچنین یوما (یوما، ۲۰۱۳) یک روش VTLN فضای ویژگی ارائه می‌کند که پیچش فرکانسی را به‌صورت درون‌یابی خطی انرژی‌های فیلتر بانک‌های به‌هم‌پیوسته‌ی مِل (Mel) مدل می‌کند. هدف این روش کاهش اعوجاجات در تخمین انرژی فیلتر بانک مِل است که به‌علت ترکیب هارمونیک فواصل زمانی گفتار واکدار[1] و نمونه‌برداری DFT (تبدیل فوریه گسسته)، هنگامی که فرکانس مرکزی فیلتر باندگذر شیفت پیدا کرده، ایجاد شده است.

در تمامی روش‌های بالا لازم است تا ضریب پیچش فرکانسی برای تمامی گویندگان تخمین زده شود. برای این‌کار رویکردهای مختلفی پیشنهاد شده است. به‌عنوان مثال در (ایدی، ۱۹۹۶) پیشنهاد شده است که با استفاده از فورمنت‌های موجود در سیگنال گفتار، تخمینی از طول مجرای صوتی به‌دست آید و با استفاده از این تخمین، ضریب پیچش فرکانسی محاسبه شود. مشکل اصلی این دسته از روش‌ها که مبتنی بر تخمین طول مجرای صوتی هستند، خطای ذاتی موجود در همین تخمین است. چون این تخمین خود وابسته به آشکارسازی فورمنت‌ها بوده و درنتیجه نیازمند مشخص‌کردن نواحی صدادار گفتار است (وانگ، ۲۰۰۷).

رویکرد رایج دیگر، جستجوی شبکه‌ای[2] (لی، ۱۹۹۸؛ وگمان، ۱۹۹۶) است. در این روش، از میان مجموعه‌ای از ضرایب پیچش که به‌طورمعمول با فواصل یکسان در یک بازه مشخص می‌شوند، ضریبی که بیشترین دقت بازشناسی را به‌دست می‌دهد انتخاب می‌شود. این کار هزینهٔ محاسباتی بسیار بالایی دارد؛ چون برای هر ضریب باید یک‌بار عملیات بازشناسی انجام شود و همچنین محدودیت پیاده‌سازی از لحاظ نیاز به برچسب دقیق فایل صوتی مورد پردازش دارد. از این رو به‌طورمعمول به جای بازشناسی، از جستجوی شبکه‌ای مبتنی بر بیشینه‌سازی درست‌نمایی استفاده می‌شود که در این روش به جای بیشینه‌سازی دقت بازشناسی، درست‌نمایی بردارهای ویژگی که پیچش بر روی آنها اعمال شده به مدل بازشناسی، بیشینه می‌شود (لی، ۱۹۹۸). این کار باعث کاهش قابل توجه زمان محاسبات می‌شود اما همچنان این روش، از آنجا که مبتنی بر جستجو می‌باشد، کند است. البته درصورتی‌که از روش‌های مبتنی بر تبدیل خطی استفاده شود، می‌توان رابطه درست‌نمایی را تا حد ممکن ساده کرد و سپس از روش امید ریاضی-بیشینه‌سازی[3] (EM) و یا حتی روش‌های مبتنی بر کاهش گرادیان، عملیات تخمین ضریب پیچش را انجام داد (وانگ، ۲۰۰۷). اما به‌دلیل پایین‌بودن کارایی روش‌های مبتنی بر تبدیل خطی که در بالا ذکر شد، این روش‌ها مقبولیت کمتری دارند. در مرجع (ویلینگ، ۱۹۹۹) برای کاهش هزینهٔ محاسبات روشی مبتنی بر GMM[4] برای تخمین سریع VTLN ارائه شده است. در این روش، تمامی مقادیر مورد نظر ضریب پیچش بر روی داده‌های آموزش اعمال می‌شود و برای هر مقدار یک مدل GMM آموزش داده می‌شود. سپس در مرحلهٔ آزمون، ضریب پیچش متناسب با آن مدل GMM که بیشترین درست‌نمایی را دارد، انتخاب می‌شود. مطابق نتایج گزارش‌شده، اگرچه این روش هزینه محاسباتی را به میزان چشم‌گیری کاهش می‌دهد؛ اما همچنان روشی مبتنی بر جستجو است و در هر جستجو نیاز به محاسبهٔ درست‌نمایی یک مدل GMM دارد. همچنین این نتایج نشان می‌دهند که این روش منجر به کاهش دقت بازشناسی نسبت به روش (لی، ۱۹۹۸) می‌شود.

در مرجع (میمورا، ۲۰۱۱) روشی مبتنی بر جدول گویندگان برای تشخیص ضریب پیچش ارائه شده است. در این روش برای تعدادی گوینده مرجع ضریب پیچش محاسبه می‌شود؛ سپس برای داده‌های آزمایش با استفاده از روش ΔBIC نزدیک‌ترین گوینده از مجموعه فوق مشخص شده و از ضریب پیچش مربوط به آن گوینده برای داده آزمایش استفاده می‌شود. ادعا شده که این روش منجر به کاهش قابل توجه هزینه محاسباتی می‌شود؛ در حالی که میزان کاهش دقت بازشناسی محسوس نخواهد بود.

در مقالهٔ حاضر، هدف ما نیز کاهش قابل ملاحظهٔ هزینهٔ محاسباتی تخمین ضریب پیچش فرکانسی، بدون کاهش دقت بازشناسی، در روش‌های مبتنی بر مقیاس‌دهی محور فرکانس است. می‌توان نشان داد که ضرایب پیچش ارتباط مستقیمی با جنسیت گوینده دارند. از آنجا که تشخیص جنسیت گوینده می‌تواند با هزینهٔ محاسباتی اندکی انجام شود، در این مقاله، یک روش جدید برای تخمین

[1] Voiced Speech
[2] grid search
[3] Expectation-Maximization
[4] Gaussian Mixture Model

سریع ضرایب پیچش گوینده پیشنهاد شده است که مبتنی بر امتیازهای صوتی به‌دست‌آمده از سامانهٔ تشخیص جنسیت گوینده است و نه‌تنها دقت بازشناسی را نسبت به روش جستجوی شبکه‌ای کاهش نمی‌دهد، بلکه منجر به بهبود آن نیز می‌شود.

ساختار مقاله به شرح زیر است: در بخش ۲ روش هنجارسازی طول مجرای صوتی معرفی می‌شود. در بخش ۳ نحوه یافتن ضرایب پیچش و آموزش مدل هنجارسازی‌شده شرح داده می‌شود. در بخش ۴ سامانه تشخیص جنسیت گویندهٔ مورد استفاده بیان خواهد شد. در بخش ۵ روش پیشنهادی و نحوهٔ به‌دست‌آوردن ضریب پیچش با استفاده از سامانه تشخیص جنسیت توضیح داده می‌شود. دادگان مورد استفاده در بخش ۶ معرفی شده و در بخش ۷ سامانه پایه، آزمایش‌های پیاده‌سازی شده و بحث و بررسی نتایج به‌دست‌آمده گنجانده می‌شود و درنهایت در بخش ۸ به نتیجه‌گیری مقاله خواهیم پرداخت.

۲- معرفی روش هنجارسازی طول مجرای صوتی (VTLN)

روش VTLN می‌تواند با پیچش[1] محور فرکانسی (FW) در بخش تحلیل فیلتر بانک در دو روش متداول استخراج ویژگی گفتاری MFCC یا PLPC پیاده‌سازی شود. در شکل (۱) روندنمای روش VTLN در روش استخراج ویژگی PLPC نشان داده شده است.

پیچش محور فرکانسی می‌تواند با ضرایب مختلفی انجام شود که این ضرایب بیان‌گر تغییرات موجود در طول لولهٔ صوتی گوینده‌ها می‌باشند. درواقع هر گوینده دارای ضریب پیچش خاص خود می‌باشد که مقدار آن متناسب با عکس طول لولهٔ صوتی آن گوینده است. درنتیجه برای پیاده‌سازی روش VTLN باید طیف گفتار هر گوینده با ضریب پیچش خاص آن گوینده مقیاس‌دهی شود. برای پیاده‌سازی این روش، به‌طور متداول سه مرحله متوالی زیر انجام می‌گیرد:

۱- پیداکردن ضریب پیچش بهینه برای سیگنال گفتار هر گوینده.

۲- پیچش فرکانسی طیف گفتار و هنجارسازی آن با ضریب پیچش بهینه.

۳- استفاده از طیف هنجارسازی‌شده در بخش استخراج ویژگی.

[1] Warping

(شکل-۱): به‌کارگیری روش VTLN در روش استخراج ویژگی PLPC

هر سامانهٔ بازشناسی دارای دو مرحله آموزش و آزمون (رمزگشایی) است. در مرحلهٔ آموزش مدل صوتی، ابتدا بایستی ضرایب پیچش تمام گوینده‌های آموزشی را به‌دست آورد؛ سپس در مرحلهٔ استخراج ویژگی، بعد از محاسبه طیف، این ضرایب را جهت پیچش طیف اعمال کرد تا بردار ویژگی‌های جدید محاسبه شوند و سپس با استفاده از این بردارهای ویژگی، مدل صوتی نهایی تعلیم یابد. در مرحلهٔ به‌کارگیری سامانهٔ (آزمون) هم ابتدا باید برای سیگنال ورودی ضریب پیچش مناسب را پیدا کرد و بعد از اعمال آن، بایستی با بهره گیری از مدل هنجار شده صوتی، فرایند بازشناسی انجام گیرد.

۳- تعیین ضریب پیچش

همان‌طور که در بخش ۲ ذکر شد، برای پیاده‌سازی روش VTLN ابتدا باید ضریب پیچش را برای هر گوینده از دادگان آموزش محاسبه کرد و سپس ویژگی هنجار شده به طول مسیر صوتی استخراج شود.

تعیین مقدار مناسب ضریب پیچش با یک الگوریتم جستجو انجام می‌شود. می‌دانیم که مقدار ضریب پیچش در یک محدوده کوچک اطراف مقدار یک قرار دارد (ویلینگ، ۲۰۰۲) ؛ لذا با تغییر مقدار ضریب پیچش در این محدوده و با گام تغییرات به اندازهٔ کافی کوچک (به‌طورمعمول برابر با

۰/۰۱) در مرحله استخراج ویژگی و سپس محاسبهٔ امتیاز لگاریتم درست‌نمایی[1] (LL) در مرحلهٔ بازشناسی، ضریب پیچش بهینه تعیین می‌شود. کوهن در (کوهن، ۱۹۹۴) نشان داد که ضریب پیچش مناسب منجر به بیشینه‌کردن امتیاز LL می‌شود.

نحوه اعمال ضریب پیچش در روش استخراج ویژگی PLPC بدین شرح است: فرض کنیم $X(\omega)$ طیف نخستین به‌دست‌آمده از یک قاب گفتاری باشد. از آنجا که در روش استخراج ویژگی PLPC از فیلترهای ذوزنقه ای بارک $T(\omega)$ در مرحلهٔ فیلتر بانک استفاده می‌شود، خروجی فیلتر nام بدون پیچش به‌صورت رابطه زیر درمی‌آید:

(۱)

$$O(n)=\sum_{\omega=l_n}^{\omega=h_n} T_n(\omega).X(\omega) \quad 0\le n\le N-1$$

که در آن N تعداد فیلترها، l_n و h_n به‌ترتیب فرکانس قطع پایین و فرکانس قطع بالای فیلتر n اُم است.

اگر تابع f وظیفه پیچش محور فرکانسی را داشته باشد، در این صورت برای طیف انتقال یافته داریم:

(۲)

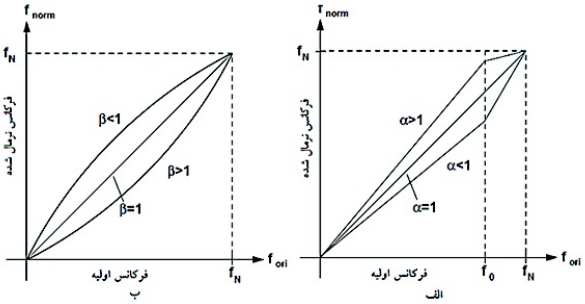
$$Y(\omega)=X(f(\omega))$$

این مسئله (پیچش محور فرکانسی) در شکل ۲ نشان داده شده است که در آن فرکانس نخستین f_{ori} بعد از اعمال پیچش به فرکانس f_{norm} تبدیل شده است.

درنهایت خروجی فیلتر بانک پیچش‌خورده به‌صورت زیر محاسبه می‌شود:

$$O_f(n)=\sum_{\omega=l_n}^{\omega=h_n} T_n(\omega).X(f(\omega)) \quad 0\le n\le N-1 \qquad (۳)$$

به‌طورمعمول برای تابع f دو نوع رابطهٔ متداول استفاده می‌شود (مولائو، ۲۰۰۰):



(شکل- ۲): هنجارسازی طول مسیر صوتی و پیچش طیف سیگنال گفتاری. الف)تابع تکه‌ای- خطی. ب) تابع دوسویه (مولائو، ۲۰۰۰).

[1] Log-Likelihood

تابع تکه‌ای- خطی: این رابطه خطی بوده و به‌صورت زیر مشخص می‌شود:

$$f_\alpha(w)=\begin{cases}\alpha\omega & \omega\le\omega_0\\ \alpha\omega_0+\dfrac{\pi-\alpha\omega_0}{\pi-\omega_0}(\omega-\omega_0) & \omega\ge\omega_0\end{cases} \qquad (۴)$$

این تابع دو پارامتر α و ω_0 دارد که ω_0 نقطه شکستگی در شکل (۲.الف) می‌باشد و به‌طورمعمول توسط رابطهٔ زیر تعیین می‌شود:

$$\omega_0=\begin{cases}\dfrac{7}{8}\pi & \alpha\le 1\\ \dfrac{7}{8\alpha}\pi & \alpha\ge 1\end{cases} \qquad (۵)$$

تابع دوسویگی: این رابطه غیر خطی بوده و فقط شامل یک پارامتر β می‌باشد و در شکل (۲.ب) نشان داده شده است.

$$f_\beta(w)=\omega+2.\tan^{-1}\left\{\frac{(1-\beta)\sin\omega}{1-(1-\beta)\cos\omega}\right\} \qquad (۶)$$

در قسمت‌هایی از منحنی که مشتق نخست تابع بالا دارای مقدار عددی کمتر از یک باشد، منجر به فشرده[2]شدن محور فرکانسی (بیش‌تر در صداهای زیر مانند گفتار زنان) می‌شود و برای مقادیر بزرگتر از یک، معادل با بسط‌دادن[3] (باز کردن) محور فرکانسی (بیش‌تر در صداهای بم مانند گفتار مردان) خواهد بود.

درصورت استفاده از توابع f_α یا f_β برای پیچش فرکانسی، منظور از تخمین ضریب پیچش، تخمین پارامترهای این توابع (α و β) خواهد بود. با توجه به اینکه استفاده از تابع تکه‌ای-خطی (f_α) رایج‌تر است، در این مقاله نیز از این تابع استفاده می‌شود و در ادامه مقاله منظور از ضریب پیچش، پارامتر α است.

بعد از به‌دست آوردن ضریب پیچش برای هر گوینده، این ضریب به طیف گفتار جهت پیچش اعمال شده و ویژگی هنجارشدهٔ گفتار استخراج می‌شود؛ سپس با استفاده از این ویژگی‌های هنجارشده، مدل صوتی هنجارشده ساخته می‌شود.

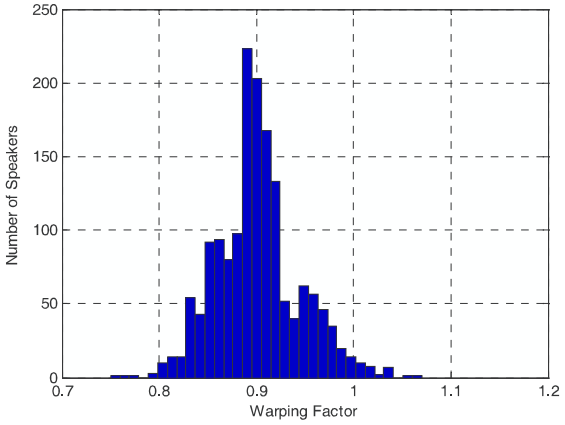
ضریب پیچش گفتار در مرحله آزمون نیز همانند مراحل آموزش محاسبه می‌شود؛ سپس با استفاده از این ضرایب، ویژگی‌های هنجارشده به‌دست می‌آیند و درنهایت بازشناسی گفتار بر روی ویژگی‌های هنجارشده و با استفاده

[2] Compressed spectrum
[3] Expanded spectrum

از مدل صوتی آموزش‌یافته بـر روی دادگـان گفتـاری تعلـیم هنجارشده، انجام می‌گیرد.

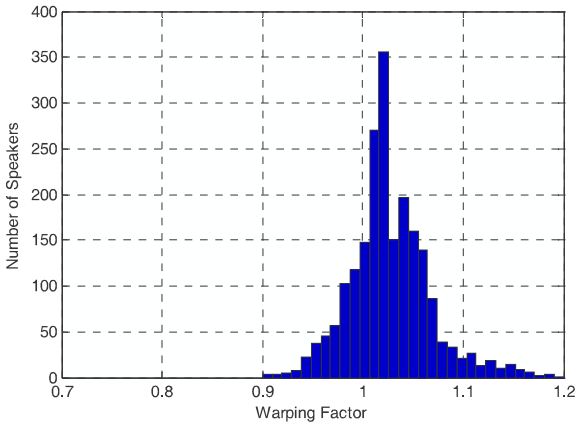
از دید عملیاتی، مهم‌ترین و پرهزینه‌ترین بخش بـرای پیاده‌سـازی VTLN، مرحلـهٔ نخسـت، یعنـی یـافتن ضـریب پیچش بهینه بـرای هـر گوینـده اسـت (جسـتجوی بهتـرین ضریب از میان ۵۱ ضریب که به معنی ۵۱ بار محاسبه امتیاز درست‌نمایی است)؛ لذا روش VTLN مبتنی بر جسـتجو بـه روشی با هزینه محاسباتی بالا بدل خواهـد شـد. بنـابراین در این مقاله روشی پیشنهاد می‌شـود کـه هزینـهٔ محاسـباتی و زمان لازم برای تعیین ضـریب پـیچش مناسـب را بـه‌شـدت کاهش دهد.

تجربه نشان داده است کـه ضـرایب پـیچش، ارتبـاط مستقیمی با جنسیت گوینده دارند و ضـرایب پـیچش بـرای گویندگان زن به‌طورمعمول کمتر از ۱ و برای گویندگان مرد به‌طورعمومی بیشتر از یک اسـت. در شـکل‌هـای ۳ و ۴ ایـن ضرایب برای گویندگان زن و مرد مربوط به دادگان تعلیم بـا استفاده از روش جستجو کـه در ایـن مقالـه بـه‌عنـوان روش بهینه در نظر گرفته شده است، به‌دست آمـده و توزیـع آنهـا رسـم شـده اسـت. محـور عمـودی ایـن توزیـع‌هـا، تعـداد گوینده‌هایی را نشان می‌دهد که ضریب پیچش آنها برابر عدد مشخص‌شده در محور افقی است. (مشخصات دادگان به کـار رفته و شرایط آزمایش‌ها در بخش ۶ آمده است).



(شکل- ۳): نمودار هیستوگرام توزیع ضرایب پیچش بهینه برای گویندگان زن

در این مقاله با استفاده از این واقعیت که نحوه توزیـع ضرایب پیچش ارتباط مستقیمی با جنسـیت گوینـده دارد و همچنین کم‌هزینه‌بودن تعیـین خودکـار جنسـیت گوینـده، برای کاهش هزینه محاسباتی تعیین ضـرایب پـیچش، روش جدیدی پیشنهاد شده است که با استفاده از امتیاز به دسـت آمده از سامانه تشخیص جنسیت گوینده، ضریب پیچش هـر گوینده تخمین زده شود. در ادامهٔ مقاله سامانه مورد استفاده برای تشخیص جنسیت گوینده معرفی می‌شود.



(شکل- ۴): نمودار هیستوگرام توزیع ضرایب پیچش بهینه برای گویندگان مرد

۴- سامانه تشخیص جنسیت گوینده

در ایـن بخـش روش و اجـزای سـامانه طراحـی‌شـده بـرای تشخیص جنسیت گوینده معرفی می‌شود. سـامانه تشـخیص جنسیت گوینده مبتنی بـر روش بیـز[1] از سـه بخـش اصـلی استخراج ویژگی، محاسـبه امتیـاز و تصـمیم‌گیـری تشـکیل می‌شود. روش کـار بـدین صـورت اسـت کـه بـا اسـتفاده از بردارهای ویژگی به‌دست‌آمده از بخش‌های گفتاریِ نمونه‌های متنوعی از گویندگان زن و مرد، برای هـر دو جـنس بـه‌طـور مجزا مدل مخلوط گوسـی (GMM) تعلـیم داده مـی‌شـود و سپس با محاسبه احتمال تعلق داده، آزمون به دو مدل، مدل برنده تعیین و جنسیت گوینده مشخص مـی‌شـود. در شـکل (۵) شماتیک کلی این سامانه نمایش داده شده است.

(شکل- ۵): بلوک دیاگرام سیستم تشخیص جنسیت گوینده

در زیربخش‌های بعدی ابتدا نحوهٔ اسـتخراج ویژگـی و سـپس روش تعلـیم تمـایزی GMM-MMI بـرای آمـوزش مدل‌های جنسیتی توضیح داده؛ سپس دادگان مورد استفاده

[1] Bayes

برای تعلیم و آزمون و دقت سیستم تشخیص جنسیت گوینده بیان می‌شود.

۴-۱- استخراج ویژگی

استخراج ویژگی: برای سامانه تشخیص جنسیت گوینده از ویژگی‌های کپستروم متداول در بازشناسی گفتار استفاده شده است. برای بهبود کیفیت ویژگی‌های استخراج‌شده، فیلتر RASTA به بردارهای ویژگی استخراج‌شده اعمال شده است. این فیلتر با حذف فرکانس‌های نزدیک به صفر (در حوزه فرکانس مدولاسیون)، موجب کاهش برخی فرکانس‌های کوچک مخرب (نوفه‌های ایستان و اعوجاجات ثابت کانال) از سیگنال می‌شود. از طرف دیگر این فیلتر فرکانس‌های بالا و غیر مفید در حوزه فرکانس مدولاسیون را نیز حذف می‌کند.

۴-۲- روش مدل‌سازی جنسیتی مبتنی بر GMM-MMI

تشخیص جنسیت گوینده یک مسئله طبقه‌بندی الگوی دو طبقه (دو جنس زن و مرد) است که به‌طورمعمول با روش بیز پیاده‌سازی می‌شود. یکی از روش‌های معمول برای مدل‌سازی توزیع دادگان هر طبقه، استفاده از مدل مخلوط گوسی تعلیم‌یافته با روش ML است. در روش تعلیم استاندارد ML، مجموع لگاریتم احتمال کل دادگان تعلیم، به‌عنوان تابع هدف در نظر گرفته می‌شود و هدف بیشینه‌کردن تابع ذیل است (رینالد، ۲۰۰۰):

$$F_{ML}(\lambda) = \sum_{r=1}^{R} \log\, p(x_r \mid G_r, \lambda) \qquad (۷)$$

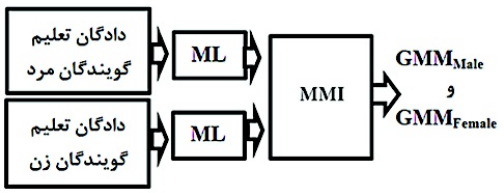
در این رابطه λ پارامترهای مدل، x_r نشان‌گر r امین قاب گفتار تعلیم، R تعداد قاب‌های گفتار تعلیم و G_r برچسب طبقهٔ داده تعلیم r ام (جنسیت گوینده) است. برای بیشینه‌کردن تابع هدف، پارامترهای GMM به‌صورت تکراری و با استفاده از روابط تخمین الگوریتم EM محاسبه می‌شود.

پس از آموزش مدل با روش ML، با استفاده از تعلیم تمایزی، دقت بازشناسی بهبود داده می‌شود. یکی از الگوریتم‌های تعلیم تمایزی مشهور، الگوریتم تعلیم MMI[1] است که در آن تابع هدف زیر بیشینه می‌شود (بورگت، ۲۰۰۶):

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(x_r \mid G_r)p(G_r)}{\sum_{\forall G} p_\lambda(x_r \mid G)p(G)} \qquad (۸)$$

در مخرج رابطه فوق، عبارت ($\sum_{\forall G} p(x_r \mid G)$)، احتمال گفتار x_r با درنظرگرفتن مدل رقیب (جنس مخالف) است. بیش‌تر احتمال پیشین کلیهٔ طبقه‌ها یکسان در نظر گرفته می‌شود. بنابراین جمله‌های $p(G_r)$ و $p(G)$ از رابطهٔ بالا حذف می‌شوند.

در این مقاله ابتدا مدل مربوط به دو جنس با استفاده از الگوریتم ML تعلیم یافته و سپس با استفاده از روش MMI پارامترهای مدل‌ها مجدداً تعلیم داده می‌شود تا تمایز مدل دو جنس افزایش یابد (شکل ۶).



(شکل- ۶): نحوه تعلیم تمایزی مدل وابسته به دو جنس مرد (GMM_{Male}) و زن (GMM_{Female}).

۴-۳- مرحله ارزیابی در تشخیص جنسیت

در روش GMM-MMI، برای تصمیم‌گیری در مورد داده آزمون، امتیاز تشخیص جنسیت زیر محاسبه می‌شود:

(۹)

$$\mathrm{GD}_{test} = \log p(x \mid GMM_{Male}) - \log p(x \mid GMM_{Female})$$
$$\text{if } \mathrm{GD}_{test} > \theta \Rightarrow \text{Male}$$
$$\text{if } \mathrm{GD}_{test} < \theta \Rightarrow \text{Female}$$

در این رابطه x ویژگی‌های گفتاری استخراج‌شده از دادهٔ آزمون است که شباهت آن به مدل گویندگان مرد (GMM_{Male}) و مدل گویندگان زن (GMM_{Female}) محاسبه می‌شود. امتیاز نهایی (GD_{test}) با یک مقدار آستانه از پیش تعیین شده (θ) مقایسه و سپس در مورد جنسیت دادهٔ آزمون تصمیم‌گیری می‌شود.

۵- معرفی روش پیشنهادی: بررسی ارتباط ضرایب پیچش بهینه با امتیاز مدلسازی جنسیت

همان طور که درقبل اشاره شد، در روش‌های متداول پیاده‌سازی الگوریتم VTLN تعیین مقدار بهینه ضریب

[1] Maximum Mutual Information

پیچش با یک الگوریتم جستجو انجام می‌شود که از لحاظ محاسباتی کاری هزینه‌بر است. روش پیشنهادی این مقاله برای کاهش محاسبات بدین صورت است که با استفاده از امتیاز به‌دست‌آمده از سامانه تشخیص جنسیت گوینده، پارامتر بهینهٔ پیچش تخمین زده شود. برای این کار به دنبال پیداکردن تابع مناسبی هستیم که با استفاده از امتیاز تشخیص جنسیت به‌دست‌آمده از رابطهٔ (۹)، تخمین مناسبی از مقدار ضریب پیچش ارائه کند. برای این منظور از معیار کمینه مربعات خطا[1] برای تخمین تابع بر روی تمامی دادگان آموزشی استفاده می‌کنیم. فرض کنیم تابع انتخابی به‌صورت چندجمله‌ای درجه یک، به‌صورت زیر در نظر گرفته شود:

$$WF(GD) = a_1 * GD + a_0 \qquad (۱۰)$$

که در آن WF ضریب پیچش بهینه، GD امتیاز تشخیص جنسیت، a_0 و a_1 ضرایب رگرسیون خطی می‌باشند.

با فرض اینکه R نقطه به‌صورت (gd_r, wf_r) داشته باشیم، تقریب حداقل مربعات خطا به‌صورت زیر محاسبه می‌شود:

$$S = \sum_{r=1}^{R}(wf_r - WF(gd_r))^2 = \sum_{r=1}^{R}(wf_r - (a_1 * gd_r + a_0))^2 \qquad (۱۱)$$

در این رابطه R تعداد دادگان آموزشی، wf_r ضریب پیچش بهینه برای داده آموزشی شماره r، gd_r امتیاز تشخیص جنسیت برای همان داده آموزشی و S میانگین مربعات خطا است. دراین روابط متغیرهای مستقل با حروف کوچک و متغیرهای وابسته با حروف بزرگ مشخص شده‌اند. مقدار ضرایب a_0 و a_1 باید طوری محاسبه شود که S کمینه شود. در نتیجه ضرایب از حل معادله زیر به‌دست می‌آید:

$$\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0 \qquad (۱۲)$$

درنتیجه بعد از ساده‌سازی، ضرایب به‌صورت زیر به‌دست می‌آیند:

$$a_0 = \frac{\sum_{r=1}^{R} wf_r \sum_{r=1}^{R} gd_r^{\,2} - \sum_{r=1}^{R} gd_r \sum_{r=1}^{R} gd_r wf_r}{R\sum_{r=1}^{R} gd_r^{\,2} - (\sum_{r=1}^{R} gd_r)^2} \qquad (۱۳)$$

$$a_1 = \frac{R\sum_{r=1}^{R} gd_r wf_r - \sum_{r=1}^{R} gd_r \sum_{r=1}^{R} wf_r}{R\sum_{r=1}^{R} gd_r^{\,2} - (\sum_{r=1}^{R} gd_r)^2} \qquad (۱۴)$$

در ادامه با استفاده از تابع به‌دست‌آمده، مقدار ضریب پیچش برای هر گوینده، از روی امتیاز تشخیص جنسیت تخصیص‌یافته به آن گوینده محاسبه شده و این ضریب به طیف گفتار جهت پیچش اعمال شده و ویژگی هنجارشدهٔ گفتار استخراج می‌شود.

۶- معرفی دادگان

دادگان گفتاری مورد استفاده برای تعلیم مدل صوتی، شامل سه مجموعه از گفتار فارسی است که برای کانال تلفنی آماده شده است. مجموعهٔ دادگان فارس‌دات تلفنی کوچک (بی‌جن‌خان، ۲۰۰۳) که شامل جملات بیان‌شده از ۶۴ نفر گویندهٔ زن و مرد است و درمجموع حدود ۸ ساعت گفتار را در بر می‌گیرد. دادگان فارس دات بزرگ (۷۴ ساعت) (شیخ‌زادگان، ۲۰۰۶) که حاوی یکصد گویشور است و به‌طور متوسط هر گویشور ۲۰ تا ۲۵ فایل ۱/۵ تا ۲ دقیقه‌ای از متن‌های رسمی که بیش‌تر روزنامه‌ای هستند، خوانده است. همچنین از یک مجموعهٔ ۱۰۲ ساعتهٔ دادگان فارسی حاوی دویست گوینده استفاده شده است که در آن هر گوینده شش فایل درحدود پنج دقیقه‌ای را از روی متون رسمی و نیمه‌محاوره‌ای خوانده است. این دادگان به‌منظور گسترش و تکمیل دادگان فارس‌دات بزرگ و با همان شرایط تهیه شده است. با توجه به اینکه آزمایش‌های این مقاله در شرایط تلفنی گزارش شده‌اند دو مجموعه داده آخر از خطوط تلفن عبور داده شده‌اند تا شرایط کانال آن را داشته باشند. دادگان گفتاری مورد استفاده برای آزمون نیز شامل ۹۴ دقیقه از گفتار نیمه‌محاوره‌ای تلفنی فارسی است که توسط ۱۰ گوینده زن و ۱۰ گوینده مرد (متفاوت از گویندگان آموزشی) از روی متون نیمه‌محاوره‌ای خوانده شده است.

مدل زبانی به‌کاررفته برای بازشناسی دادگان آزمون نیمه‌محاوره‌ای، ۳-گرامی[2] است که حاوی حدود ۱۱/۵ میلیون ۳-گرام است که از متون رسمی- محاوره‌ای با حجم درحدود یکصد میلیون کلمه‌ای (برگرفته از پیکرهٔ متنی فارسی) و چندین متن محاوره‌ای درحدود ۱۴ میلیون کلمه-ای ساخته شده است. مجموعه واژگان مورد استفاده نیز شامل حدود ۵۳ هزار کلمه با میانگین تنوع تلفظی ۲/۴ تلفظ برای هر کلمه است. تنوعات تلفظی هر کلمه در مجموعه واژگان براساس ۲۹ واج متداول فارسی تهیه شده است (شکفته، ۲۰۱۳).

[1] Least Square Error

[2] Trigram

برای تعلیم سامانه تعیین جنسیت گوینده از دادگان فارس‌دات بزرگ (تلفنی‌شده) استفاده شده است. دادگان مورد استفاده برای آزمون سامانه نیز دادگان فارس‌دات کوچک تلفنی بوده است. دقت سامانه تشخیص جنسیت گوینده بر روی این دادگان ۹۵/۹۴ درصد است که از تقسیم تعداد فایل‌های درست به کل فایل‌ها به‌دست آمده است.

۷- معرفی سامانه پایه، آزمایش‌های انجام‌شده و بحث و بررسی نتایج

برای پیاده‌سازی روش VTLN و تعلیم مدل صوتی و بازشناسی دادگان آزمون در این مقاله از سامانه بازشناسی گفتار پیوسته با مدل‌های صوتی مبتنی بر مدل مخفی مارکوف (HMM) به‌کمک جعبه ابزار HTK استفاده شده است (http://htk.eng.cam.ac.uk). این سامانه، شامل مجموعه‌ای از مدل‌های HMM سه‌آوایی[1] چپ به راست برای هر یک از ۲۹ واج فارسی است، که به‌طور متوسط برای هر واج تعداد چهار حالت (State) در نظر گرفته می‌شود. توزیع احتمالاتی ویژگی‌ها در هر حالت نیز با مدل مخلوط گوسی (GMM) با تعداد ۳۲ گوسی در نظر گرفته شده است. ویژگی‌های استخراج شده برای هر قاب گفتار تلفنی، شامل ۱۲ ضریب PLPC به همراه ضریب کپسترال c_0 است و از مشتقات مرتبه نخست تا سوم ضرایب کپستروم به دست آمده برای مدل‌سازی دینامیک ویژگی‌های گفتاری استفاده شده است. همچنین در جهت مقاوم سازی بردار ویژگی به دست آمده نسبت به نوفه‌های جمع‌شونده (مانند صدای محیط) و اثر انتقال کانال تلفنی از روش پس پردازش MVA با درجه فیلتر یک استفاده شده است (شکفته و همکاران، ۱۳۸۹). نتایج بازشناسی بر حسب درصد دقت (Accuracy) بازشناسی کلمه گزارش شده است.

برای پیاده سازی روش VTLN، برای هر گفتار گوینده، ضریب پیچش بهینه با استفاده از روش مبتنی بر جستجو، روش مبتنی بر جدول گویندگان و روش پیشنهادشده، به شرح زیر محاسبه شد:

۷-۱- روش مبتنی بر جستجو ((باباعلی، ۱۳۸۶)):

ابتدا با استفاده از ویژگی PLPC با بُعد ۵۲ و با ضریب پیچش یک (معادل با عدم اعمال ضریب پیچش)، یک مجموعه مدل صوتی تک‌آوایی[2] بر مبنای مدل مخفی مارکوف تعلیم داده می‌شود؛ سپس در مرحله استخراج ویژگی، با تغییر مقدار ضریب پیچش در محدوده (۱/۲۰-۰/۷۰) و با گام تغییرات ۰/۰۱، ۵۱ سری ویژگیِ مختلفِ به‌دست می‌آید و برای هر کدام از این سری ویژگی‌ها، با استفاده از مدل صوتی تک‌آوایی، بازشناسی در سطح واج انجام می‌گیرد و امتیاز لگاریتم درست‌نمایی محاسبه می‌شود. درنهایت ضریب پیچشی که امتیاز لگاریتم درست‌نمایی را بیشینه کند، به‌عنوان ضریب پیچش بهینه انتخاب می‌شود. این عملیات بر روی کل مجموعه آموزش انجام می‌شود.

در ادامه با استفاده از ضریب پیچش بهینه، ویژگی‌های هنجارشده کل گفتار هر گوینده استخراج و با استفاده از ویژگی‌های هنجارشدهٔ دادگان آموزشی، مدل صوتی نهایی تعلیم داده می‌شود. درنهایت نیز ویژگی‌های هنجارشده دادگان آزمون با استفاده از مدل صوتی ساخته‌شده در سطح کلمه بازشناسی می‌شود.

۷-۲- روش مبتنی بر جدول گویندگان (میمورا، ۲۰۱۱)

در این روش برای تعدادی گویندهٔ مرجع، ضریب پیچش با استفاده از روش مبتنی بر جستجو محاسبه می‌شود؛ سپس برای داده‌های آزمایش با استفاده از روش ΔBIC نزدیک‌ترین گوینده از مجموعه بالا مشخص شده و از ضریب پیچش مربوط به آن گوینده برای دادهٔ آزمایش استفاده می‌شود. برای پیاده‌سازی این روش از تمامی گویندگان موجود در دادگان آموزش (۳۶۴ گوینده) به‌عنوان گویندگان مرجع استفاده کرده‌ایم؛ سپس به هر یک از ۲۰ گوینده دادگان آزمون، ضریب پیچش نزدیک‌ترین گوینده از مجموعه مرجع اختصاص داده شده است.

۷-۳- روش پیشنهادی-مبتنی بر امتیاز جنسیت

با استفاده از رابطه ۱۰ در بخش ۵، تخمین ضرایب پیچش جدید از روی امتیاز تشخیص جنسیت تعیین می‌شود. ویژگی‌های هنجارشده داده‌های آموزش با استفاده از این ضرایب پیچش جدید استخراج می‌شوند و مدل صوتی با استفاده از آنها ساخته می‌شود. در نهایت ویژگی‌های هنجارشدهٔ دادگان آزمون با استفاده از این مدل، بازشناسی می‌شوند. این روش در دو حالت پیاده‌سازی شده است: در حالت نخست برای کل گویندگان از یک مدل رگرسیون (ضرایب a_0 و a_1 در رابطهٔ ۱۰)

[1] Triphone

[2] monophone

و در حالت دوم از دو مدل رگرسیون جداگانه برای گویندگان مرد و زن استفاده شده است.

۷-۴- پیاده‌سازی آزمایش‌ها

برای پیاده‌سازی آزمایش‌ها چهار نوع ضریب پیچش متفاوت محاسبه خواهد شد که در جدول زیر آورده شده است:

(جدول- ۱): معرفی انواع ضریب پیچش پیاده سازی شده در آزمایش‌ها

حالت پایه	بدون ضریب پیچش (WF=1)
نوع ۱	**محاسبه ضریب پیچش با استفاده از امتیاز لگاریتم درست نمایی از روی کل واج‌های بازشناسی‌شده (روش پایه VTLN مطابق با مرجع** (باباعلی و همکاران ۱۳۸۶))
نوع ۲	**محاسبه ضریب پیچش با روش مبتنی بر جدول گویندگان** (میمورا، ۲۰۱۱)
نوع ۳	**محاسبه ضریب پیچش با روش تخمین رابطه ۱۰ (روش پیشنهادی مقاله)**
نوع ۴	**محاسبه ضریب پیچش با روش تخمین رابطه ۱۰ به‌صورت جداگانه برای گویندگان مرد و زن (روش پیشنهادی مقاله)**

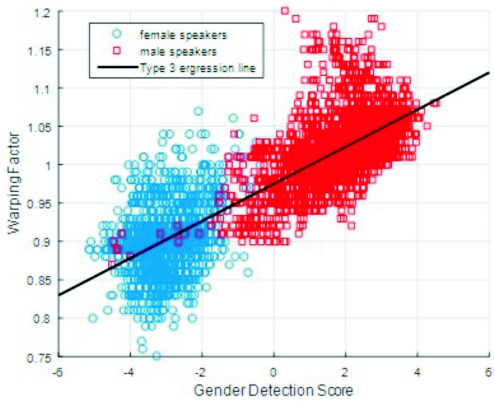
جدول (۲)، زمان پردازش یک دقیقه گفتار را برای هر یک از روش‌های جدول (۱) و به تفکیک بخش‌های مختلف آنها بر حسب ثانیه نشان می‌دهد. همچنین در این جدول سهم هر بخش از پردازش کل در مرحله آزمون نیز به صورت درصد بیان شده است.

همان طور که درقبل نیز اشاره شد، به‌دست‌آوردن ضریب پیچش بهینه با استفاده از روش مبتنی بر الگوریتم جستجو (نوع ۱) زمان‌بر است؛ به‌طوری که برای محاسبهٔ ضریب پیچش بهینه، برای یک فایل گفتار یک دقیقه‌ای زمانی در حدود ۱۱۵ ثانیه لازم است که معادل ۴۸ درصد از زمان کل پردازش است؛ در حالی که برای به‌دست‌آوردن ضریب پیچش هر دقیقه گفتار با استفاده از روش پیشنهادی (نوع ۳ و ۴)، زمانی در حدود ۰/۶۸ ثانیه مورد نیاز خواهد بود (کمتر از ۱ درصد از پردازش کل). روش مبتنی بر جدول گویندگان (نوع ۲) نیز از نظر هزینهٔ محاسبات، عملکرد مناسبی داشته و برای هر دقیقه گفتار حدود ۰/۸۹ ثانیه زمان صرف کرده است. برنامه‌های مورد نیاز برای به‌دست‌آوردن ضریب پیچش در محیط MATLAB پیاده‌سازی شدند و آزمایش‌ها بر روی یک ماشین مجازی با ۱۲ هسته ۲/۵ گیگاهرتزی و ۱۲ گیگابایت حافظه RAM انجام شده‌اند. این ماشین مجازی بر روی یک رایانهٔ سرور با پردازنده Xenon ساخته شده است. آموزش مدل‌های صوتی بر روی این سامانه با استفاده از قابلیت پردازش موازی جعبه ابزار HTK، در حدود پانزده ساعت زمان صرف کرده است.

(جدول- ۲): زمان صرف شده برای پردازش یک دقیقه از گفتار بر حسب ثانیه به همراه درصد سهم هر بخش.

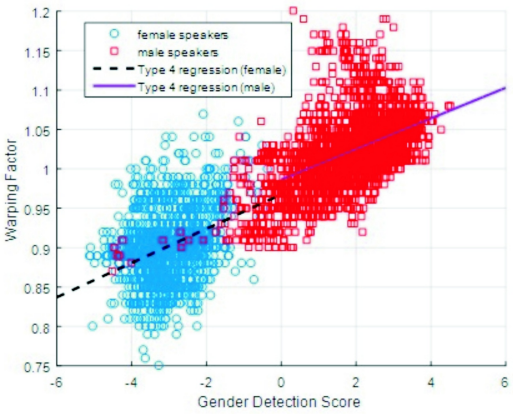
روش	زمان استخراج ویژگی		زمان تخمین ضریب پیچش		زمان بازشناسی	
	زمان	درصد	زمان	درصد	زمان	درصد
نوع ۱	۰/۱۷	٪۰/۰۷	۱۱۵/۰۶	٪۴۸/۳۹	۱۲۲/۵۵	٪۵۱/۵۴
نوع ۲	۰/۱۷	٪۰/۱۴	۰/۸۹	٪۰/۷۵	۱۱۸/۰۹	٪۹۹/۱۱
نوع ۳ و ۴	۰/۱۷	٪۰/۱۵	۰/۶۸	٪۰/۵۹	۱۱۴/۸۹	٪۹۹/۲۶

در شکل (۷) توزیع ضریب پیچش گفتارهای آموزشی برحسب امتیاز تشخیص جنسیت آنها به همراه مدل خطی به‌دست‌آمده (نوع ۳)، نشان داده شده است. همان‌طور که مشاهده می‌شود برای گویندگان زن، ضریب پیچش آنها مقداری کمتر از یک و امتیاز جنسیت آنها منفی است. همچنین برای اکثریت گویندگان مرد، ضریب پیچش آنها مقداری بیشتر از یک و امتیاز جنسیت آنها مثبت است.



(شکل- ۷): توزیع ضریب پیچش بهینه مجموعه متنوعی از دادگان آموزشی برحسب امتیاز تشخیص جنسیت آن‌ها. مربع‌های قرمز و دایره‌های آبی رنگ به ترتیب نشان‌دهندهٔ گویندگان مرد و زن هستند و خط سیاه نشان دهنده منحنی به‌دست‌آمده از رگرسیون در حالت نوع سه است.

همان‌طور که در شکل (۷) مشخص است، پراکندگی غالب گویندگان مرد و زن در یک راستا نیست. این موضوع نشان می‌دهد که درصورت استفاده از دو مدل خطی جداگانه برای گویندگان مرد و زن، تخمین بهتری از ضریب پیچش به‌دست خواهد آمد. این موضوع در جدول (۱) با نام نوع چهار بررسی شده است. برای این منظور، برای گویندگانی که امتیاز جنسیت بزرگ‌تر از صفر دارند، یک خط رگرسیون و برای گویندگان با ضریب جنسیت کمتر از صفر، خطی دیگر تخمین زده شده است. شکل (۸) خطوط رگرسیون به‌دست‌آمده را نشان می‌دهد.



(شکل- ۸): مدل‌سازی جداگانه توزیع ضریب پیچش بهینه برای گویندگان مرد و زن (نوع ۴). مربع‌های قرمز و دایره‌های آبی‌رنگ به‌ترتیب نشان دهنده گویندگان مرد و زن هستند و خط و خط چین سیاه به‌ترتیب نشان‌دهنده منحنی به‌دست آمده از رگرسیون برای هر یک از آنها است.

با توجه به استخراج دو دسته ویژگی متناظر با دو روش تخمین ضریب پیچش، و همچنین حالت پایه (WF=1)، در کل چهار مدل صوتی می‌توان آموزش داد که نتیجه دقت بازشناسی آنها بر روی دادگان آزمون در جدول (۳) آورده شده است.

همان‌طور که مشاهده می‌شود با استفاده از روش پیشنهادی در حالت نوع سه در دادگان آزمون، دقت بازشناسی کلمه کل به حدود ۵۶/۱۴ درصد رسیده است که منجر به افزایش حدود ۱/۶ درصدی نسبت به روش پایه شده است. دقت روش پیشنهادی در حالت نوع چهار، به ۵۶/۴۶ درصد رسیده است که افزایشی در حدود ۱/۹۱ درصد را نسبت به روش پایه نشان می‌دهد. همچنین، این روش سرعت تخمین ضریب پیچش را به میزان قابل توجه ۱۶۹ برابر افزایش داده است (جدول ۲). دقت‌های بازشناسی در روش مبتنی بر جدول گویندگان (نوع ۲)، بسیار نزدیک به روش مبتنی بر جستجو (نوع ۱) بوده است. این موضوع بدین صورت قابل توجیه است که در این روش برای گوینده آزمون به‌طوردقیق از ضریب پیچش گوینده‌ای استفاده می‌شود که ضریب پیچش آن توسط روش نوع یک به‌دست آمده است. با این حال، سرعت این روش نیز بسیار بالا بوده و در حدود ۱۲۹ مرتبه سریع‌تر از روش مبتنی بر جستجو است.

(جدول- ۳): نتایج درصد دقت بازشناسی کلمه با انواع متفاوت تعیین ضریب پیچش در دادگان تعلیم و آزمون و به‌تفکیک جنسیت گویندگان

نوع ضریب پیچش دادگان آموزش و آزمون	درصد دقت بازشناسی کلمه		
	کل	مرد	زن
حالت پایه (WF=1)	۵۴/۵۵	۵۸/۶۷	۵۰/۳۲
نوع ۱	۵۵/۹۲	**۵۹/۴۰**	۵۲/۳۳
نوع ۲	۵۵/۹۴	۵۹/۳۸	۵۲/۴۰
نوع ۳	۵۶/۱۴	۵۸/۸۱	۵۳/۳۹
نوع ۴	**۵۶/۴۶**	۵۹/۲۵	**۵۳/۵۸**

اگرچه هدف روش پیشنهادی در این مقاله بهبود کارایی روش پایه VTLN از جهت زمان محاسبات بود، اما نتایج جدول ۳ نشان می‌دهد که علاوه‌بر این هدف، این روش درصد دقت بازشناسی کلمه را نیز حدود ۰/۵۴ درصد نسبت به سامانه نوع یک بهبود داده است. بررسی تفکیکی این نتایج بر حسب جنسیت گویندگان نشان می‌دهد باتوجه به اینکه روش پیشنهادی باعث بهبود دقت بازشناسی نسبت به سامانه نوع یک، بر روی کل داده‌ها شده است، اما برتری این روش از جهت گویندگان زن بوده است. حتی استفاده از مدل رگرسیون جداگانه برای گویندگان مرد نیز این موضوع را تغییر نداده و با وجود بهبود بازشناسی گویندگان مرد در حالت نوع چهار نسبت به نوع سه، همچنین این حالت ضعیف‌تر از نوع یک بوده است. این موضوع را می‌توان به نوع پراکندگی گویندگان مرد ارتباط داد؛ به‌نحوی که پخش این گویندگان به‌صورت خطی نبوده و ممکن است درصورت استفاده از مدل‌های با مرتبه بالاتر، نتایج بهتری حاصل شود.

۸- جمع‌بندی

در این مقاله به جبران‌سازی مشکلات مربوط به تفاوت طول مجرای صوتی گویندگان مختلف در حوزه بازشناسی گفتار پرداخته شد. هنجارسازی طول مجرای صوتی به‌عنوان راه حلی مؤثر برای این مشکل بررسی و نحوهٔ اعمال پیچش فرکانسی بر روی طیف گفتار تشریح شد. در ادامه روش متداول به‌دست آوردن ضریب پیچش با استفاده از الگوریتم مبتنی بر جستجو توضیح داده و مشاهده شد که عمده مشکلات این روش زمان‌بربودن و هزینهٔ محاسباتی بالای آن است. با بررسی توزیع گویندگان مرد و زن بر حسب ضریب پیچش نشان دادیم که ارتباط معنی‌داری بین جنسیت و ضریب پیچش وجود دارد و مبنای روش پیشنهادی را بر همین موضوع استوار کردیم. از این رو سامانه‌ای برای تشخیص جنسیت گوینده ارائه شد که خروجی این سامانه، امتیاز وابسته به جنسیت هر گوینده است؛ سپس با استفاده از امتیاز جنسیتی حاصل و ضرایب پیچش بهینهٔ به‌دست‌آمده از روش جستجو، روشی پیشنهاد شد که در آن ضرایب پیچش از روی امتیاز تشخیص جنسیت با روش کمینه‌کردن خطا، تخمین زده شود. نتایج پیاده سازی‌ها کارایی بهتر و سرعت بیشتر این روش را نسبت به روش متداول مبتنی بر الگوریتم جستجو نشان می‌دهد.

۹- مراجع

باقر باباعلی، حسین صامتی ، هادی ویسی، "بکارگیری نرمالسازی اثر طول مسیرصوتی گوینده ها (VTLN) در سیستم بازشناسی گفتار پیوسته فارسی مبتنی بر مدل مخفی مارکوف"، *سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران*، جزیره کیش، اسفند ۱۳۸۶.

شهلا عزیزی، فرزاد توحیدخواه، فرشاد الماس گنج، "بررسی اثر استفاده از روش تطبیق هنجارسازی طول مسیر صوتی به منظور تشخیص اختلالات گفتاری رایج و گفتاردرمانی کودکان فارسی زبان"، *فصلنامه مهندسی پزشکی زیستی*، سال ششم، شماره ۴، صص. ۲۵۷-۲۶۵، زمستان ۱۳۹۱.

قمرناز تدین تبریزی، سعید ستایشی، "ارائه روشی مبتنی بر نرمالسازی اکوستیکی و خوشه بندی برای بهبود بازشناسی گفتار کودکان فارسی زبان"، *مجله فنی مهندسی دانشگاه آزاد اسلامی مشهد*، صص. ۱۱۳-۱۲۵ ، زمستان ۱۳۸۸.

یاسر شکفته، جهانشاه کبودیان، محمدمحسن گودرزی و ایمان صراف رضائی، "بهبود کارایی سیستم کاوشگر کلمات تلفنی با استفاده از نرمالیزاسیون امتیاز اطمینان مبتنی بر روش برنامه‌ریزی خطی"، *دوفصلنامه پردازش علائم و داده‌ها*، سال چهارم، شماره پیاپی ۱۴، صص. ۳۷-۴۸، زمستان ۱۳۸۹.

Acero A. and Stern R. 1991, Robust speech recognition by normalization of the acoustic space, in Proc. ICASSP '91, vol. 2, pp. 893–896.

Bijankhan M., Sheykhzadegan J., Roohani M.R., Zarrintare R., Ghasemi S.Z., and Ghasedi M.E. 2003, TFarsDat – The Telephone Farsi Speech Database, In Proc. Eurospeech, Geneva, Switzerland, pp. 1525-1528.

Burget L., Matejka P., and Cernocky, J. 2006, Discriminative training techniques for acoustic language identification. In Proc. ICASSP, vol. 1, pp. I-I.

Claes T., Dologlou I., Bosch L., and Compernolle D. 1998, A novel feature transformation for vocal tract length normalization in automatic speech recognition, IEEE Trans. Speech Audio Process., vol. 11, no. 6, pp. 603–616.

Cohen, J., Kamm, T., and Andreou, A. 1994, An Experiment in systematic speaker variability, Speech Workshop on Robust Speech Recognition.

Cui X.and Alwan A. 2006, Adaptation of children's speech with limited data based on formant-like peak alignment, Comput. Speech Lang., vol. 20, no. 4, pp. 400–419.

Ding G., Zhu Y., Li C., and Xu B. 2002, Implmenting vocal tract length normalization in the MLLR framework, in Proc. ICSLP '02, pp. 1389–1392.

Eide E. and Gish H. 1996, A parametric approach to vocal tract length normalization, in Proc. ICASSP '96, pp. 346–349.

Emori T. and Shinoda K. 2001, Rapid vocal tract length normalization using maximum likelihood estimation, in Proc. Eurospeech '01, Aalborg, Denmark.

Farsdet Speech Database, Research Center of Intelligent Signal Processing, http://www.rcisp.com

Giuliani D., Gerosa M., and Brugnara F. 2006, Improved automatic speech recognition through speaker normalization, Comput. Speech Lang., vol. 20, no. 1, pp. 107–123.

HTK, HMM ToolKit and HTK book, Available from: http://htk.eng.com.ac.uk.

Lee L. and Rose R. 1998, A frequency warping approach to speaker normalization, IEEE Trans. Speech Audio Process., vol. 6, no. 1, pp. 49–60.

Wang S., Cui X., and Alwan A. 2007, Speaker adaptation with limited data using regression-tree-based spectral peak alignment, IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2454–2464.

Wegmann S., McAllaster D., Orloff J., and Peskin B. 1996, Speaker normalization on conversational telephone speech, In Proc. ICASSP '96, pp. 339–341.

Welling L., Ney H., and Kanthak S. 2002, Speaker Adaptive Modeling by Vocal Tract Normalization, IEEE Trans. Speech and Audio Processing, vol. 10, no. 6, pp. 415-426.

Welling L., Kanthak S. and Ney H. 1999, Improved Methods for Vocal Tract Normalization, In proc. ICASSP, pp. 761-764.

Yoma, N. B., Garreton, C., Huenupan, F., Cetalan, I., and Wuth Sepulveda, J. 2013, On Reducing Harmonic and Sampling Distortion in Vocal Tract Length Normalization, IEEE Trans. Audio, Speech, Lang. Proc., vol. 21, no. 1, pp 110-121.

یاسر شکفته تحصیلات خود را در مقطع کارشناسی در دو رشته مهندسی پزشکی-بیوالکتریک و مهندسی برق-الکترونیک به‌ترتیب در سال‌های ۱۳۸۴ و ۱۳۸۵ در دانشگاه صنعتی امیرکبیر به پایان رساند. ایشان در سال‌های ۱۳۸۷ و ۱۳۹۲ مدارک کارشناسی ارشد و دکتری خود را در رشته مهندسی پزشکی (گرایش بیوالکتریک) از همان دانشگاه اخذ کرد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش‌های خطی و غیرخطی سیگنال، شناسایی الگو و پردازش صوت و زبان طبیعی است. نشانی رایانامهٔ ایشان عبارت است از:
y_shekofteh@{aut.ac.ir, yahoo.com}

محمد محسن گودرزی تحصیلات خود را در مقطع کارشناسی در دو رشتهٔ مهندسی پزشکی-بیوالکتریک و مهندسی برق-کنترل به‌ترتیب در سال‌های ۱۳۸۶ و ۱۳۸۸ در دانشگاه صنعتی امیرکبیر به پایان رساند. در سال ۱۳۸۹ در مقطع کارشناسی ارشد رشته مهندسی پزشکی-بیوالکتریک از همان دانشگاه فارغ التحصیل شد. وی هم اکنون در مقطع دکترای مهندسی پزشکی-بیوالکتریک در دانشگاه صنعتی امیرکبیر در حال تحصیل است. زمینه‌های پژوهشی مورد

Leggetter C. J. and Woodland P. C. 1995, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov modals, Comput. Speech Lang., vol. 9, pp. 171–185.

McDonough J., Schaaf T., and Waibal A. 2004, Speaker adaptation with all-pass transforms, Speech Commun., vol. 41, no. 1, pp. 75–91.

Mimura M. and Kawahara T. 2011, Fast Speaker Normalization and Adaptation based on BIC for Meeting Speech Recognition, IEICE Transaction on information and systems, vol J95-D.

Mitra V., Wang W., Franco H., Lei Y., Bartels C., and Graciarena M. 2014, Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions, In proc. Interspeech, pp. 895-899.

Molau S., Kanthak S., and Ney H. 2000. Efficient vocal tract normalization in automatic speech recognition. In Proc. of the ESSV'00.

Panchapagesan S. and Alwan A. 2009, Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC, Comput. Speech Lang., vol. 23, no. 1, pp. 42–46.

Pitz M. and Ney H. 2005, Vocal tract normalization equals linear transformation in cepstral space, IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 930–944.

Reynolds, D.A., Quatieri, T.F., and Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1), pp. 19-41.

Sainath T. N., Kingsbury B., Mohamed A., Saon G., and Ramabhadran B. 2014, Improvements to filterbank and delta learning within a deep neural network framework., In proc. ICASSP, pp. 6839-6843.

Sanand D.R., Schlüter R., and Ney H. 2010, Revisiting VTLN using linear transformation on conventional MFCC, In Proc. Interspeech '10, pp. 538–541.

Saon G., Kuo J., Rennie S. and Picheny M. 2015, The IBM 2015 English Conversational Telephone Speech Recognition System, arXiv:1505.05899v1.

Sheikhzadegan J., Bijankhan M. 2006, Persian speech databases. In 2nd Workshop on Persian Language and Computer, pp. 247-261.

Shekofteh Y., Almasganj F. 2013, Autoregressive modeling of speech trajectory transformed to the reconstructed phase space for ASR purposes
Digital Signal Processing, vol. 23, pp. 1923–1932

Umesh S., Zolnay A., and Ney H. 2005, Implementing frequency-warping and VTLN through linear transformation of conventional MFCC., In Proc. Interspeech '05, pp. 269–272.

علاقهٔ ایشان شناسایی الگو، پردازش سیگنال‌های تصادفی و پردازش و بازشناسی گفتار است.
نشانی رایانامهٔ ایشان عبارت است از:
mm.goodarzi@aut.ac.ir

حسن قلی‌پور مدارک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی پزشکی (گرایش بیوالکتریک) به‌ترتیب در سال‌های ۱۳۸۹ و ۱۳۹۱ از دانشگاه صنعتی امیرکبیر اخذ کرده است. وی در حال حاضر عضو گروه پردازش صوت و زبان طبیعی پژوهشکده پردازش داده پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی است. زمینه‌های تخصصی مورد علاقه ایشان پردازش و بازشناسی گفتار و پردازش زبان طبیعی است.
نشانی رایانامهٔ ایشان عبارت است از:
h.gholipour@aut.ac.ir

سیدجهانشاه کبودیان تحصیلات خود را در مقاطع کارشناسی، کارشناسی ارشد و دکترا در دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) گذراند و مدرک دکترای خود را در سال ۱۳۸۹ از دانشگاه مذکور دریافت کرد. وی هم‌اکنون استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه رازی کرمانشاه است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: پردازش سیگنال، پردازش گفتار و صوت، پردازش زبان طبیعی، شناسایی الگو، یادگیری ماشین و الگوریتم‌های فراابتکاری.
نشانی رایانامهٔ ایشان عبارت است از:
kabudian@{razi, aut}.ac.ir

فرشاد الماس‌گنج در سال ۱۳۶۳ در رشته برق، گرایش الکترونیک، از دانشگاه امیرکبیر فارغ التحصیل شد؛ سپس دوره کارشناسی ارشد خود را در همین رشته تا سال ۱۳۶۷ ادامه داد و با یک فاصله چهار ساله، دوره دکترای برق (گرایش مهندسی پزشکی) را آغاز کرد. وی در حال حاضر دانشیار دانشکدهٔ مهندسی پزشکی دانشگاه امیرکبیر است. زمینه پژوهشی اصلی ایشان پردازش سیگنال و بیش‌تر در زمینه بازشناسی گفتار فارسی و بازشناسی خصوصیات پروزودیک گفتار است.
نشانی رایانامهٔ ایشان عبارت است از:
almas@aut.ac.ir

شقایق رضا تحصیلات خود را در مقطع کارشناسی در رشته مهندسی پزشکی (بیوالکتریک) در دانشگاه صنعتی امیرکبیر (۱۳۸۵) و کارشناسی ارشد را در همان رشته در دانشگاه صنعتی امیرکبیر (۱۳۸۷) به پایان رساند. ایشان هم اکنون دانشجوی مقطع دکتری بیوالکتریک در دانشگاه صنعتی امیرکبیر هستند. از موضوعات مورد علاقهٔ ایشان می‌توان به پردازش گفتار، پردازش سیگنال و تصویر اشاره کرد.
نشانی رایانامهٔ ایشان عبارت است از:
shaghayegh.reza@gmail.com

ایمان صراف رضایی مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی پزشکی (گرایش بیوالکتریک) به‌ترتیب در سال‌های ۱۳۸۴ و ۱۳۸۷ از دانشگاه صنعتی امیرکبیر اخذ کرده است. وی در حال حاضر مدیر گروه پردازش صوت و زبان طبیعی پژوهشکده پردازش دادهٔ پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی است. مدل‌سازی و پردازش سیگنال گفتار، زمینهٔ عمومی پژوهش‌های ایشان است.
نشانی رایانامهٔ ایشان عبارت است از:
imansarraf@aut.ac.ir