ارائه روشی مبتنی بر نرمال‌سازی اکوستیکی و خوشه‌بندی برای بهبود بازشناسی گفتار کودکان فارسی زبان

قمرناز تدین تبریزی*(۱) **سعید ستایشی(۲)**

(۱) دانشجوی دکتری مهندسی کامپیوتر- نرم افزار، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران ، گروه مهندسی کامپیوتر، تهران، ایران
(۲) دانشیار، دانشگاه صنعتی امیرکبیر، گروه مهندسی هسته ای، تهران، ایران

چکیده بررسی کاربردهای بازشناسی گفتار نشان دهندهٔ تفاوت‌های طیفی در سیگنال‌های گفتار کودکان می‌باشد. این تنوع، باعث ایجاد مشکلاتی در بازشناسی خودکار گفتار کودکان می‌شود. تجربه نشان داده در صورتی که از دادهٔ گفتار کودکان به عنوان ورودی در مدل‌های اکوستیکی استفاده شود که با گفتار بزرگسالان آموزش یافته‌اند، کارایی به اندازهٔ قابل توجهی کاهش می‌یابد. به طور میانگین نرخ خطای کلمه برای بازشناسی گفتار کودکان دو تا چهار بار بیشتر از بزرگسالان است. میزان درستی بازشناسی گفتار در کودکان به عواملی مثل سن، جنسیت، فرکانس مبنایی و قد بستگی دارد. در این مقاله برخی از روش‌های افزایش کارایی بازشناسی گفتار کودکان شامل هنجارسازی طول محدودهٔ صوتی (VTLN)، آموزش تطبیقی گوینده (SAT) و هنجارسازی گوینده بر اساس رگرسیون خطی با بیشترین درست نمایی محدود شده (CMLSN) مطرح و روش VTLN برای بهبود کارایی بازشناسی گفتار کودکان فارسی زبان پیاده سازی شده است. نهایتا روشی برمبنای ترکیب روش‌های هنجارسازی و خوشه بندی برای بازشناسی گفتار کودکان پیشنهاد شده است. با استفاده از خوشه بندی گفتار ورودی و تخصیص آن به مدل مناسب، درستی بازشناسی به طور متوسط ۵۰٪ افزایش می‌یابد.

واژه‌های کلیدی بازشناسی گفتار کودکان، تبدیل صوت، مدل سازی تطبیقی، نرمال سازی گوینده، خوشه‌بندی گفتار.

*عهده دار مکاتبات
نشانی: مشهد، قاسم آباد، دانشکدهٔ فنی و مهندسی دانشگاه آزاد اسلامی واحد مشهد، گروه فناوری اطلاعات
تلفن: ۰۵۱۱-۶۶۲۲۷۸۹ **پست الکترونیکی:** tadayon@mshdiau.ac.ir

۱- مقدمه

مطالعات انجام شده در زمینه بازشناسی گفتار نشان میدهد که محققین این حوزه اغلب با گفتار بزرگسالان سروکار داشته و تحقیقات کمی در زمینه بازشناسی گفتار کودکان انجام داده‌اند. در این زمینه همواره بررسی کاربردهای محاوره ای برای ارتباط با کودکان مورد توجه بوده است [۱]. کودکان با کامپیوتر به گونه ای متفاوت از بزرگسالان ارتباط برقرار و صحبت می‌کنند و قطعا سیستم‌های بازشناسی گفتاری که برای بزرگسالان طراحی شده‌اند نمی‌توانند در این رابطه به خوبی عمل کند. دلایل چندی وجود دارند که باعث کاهش کارایی سیستم‌های بازشناسی گفتار موجود برای کودکان شده است، از جمله تفاوت‌های فیزیکی، نحوهٔ تلفظ و متفاوت بودن فرهنگ لغت کودکان. بنابراین زمانی که از مدل‌های گفتار بزرگسالان برای بازشناسی گفتار کودکان استفاده می‌شود، سطح کارایی به اندازه‌ای پایین است که باعث می‌شود این گونه سیستم‌ها غیرقابل استفاده شوند. اولین تحقیقات انجام شده در این زمینه، بیشتر متمرکز بر تطبیق فناوری بازشناسی گفتار بزرگسالان برای به کارگیری آن در بازشناسی گفتار کودکان بود. مسألهٔ مهمی که در این جا مورد توجه قرار گرفت، تفاوت خصوصیات آکوستیکی بین دو گروه بود. Das و همکارانش از روش‌های هنجارسازی طول محدوده صوتی(VTLN) (Vocal tract length normalization) در یک شناسنده گفتار بزرگسالان استفاده کردند تا سیستم را برای کاربری آن در محدودهٔ کودکان آماده نمایند [۲]. Gustafson و همکارش نیز از شناسندهٔ گفتاری استفاده کردند که در آن به لایهٔ استخراج ویژگی‌ها دسترسی نبود و بنابراین نمی‌توانستند از VTLN استفاده کنند. در عوض، از روش‌های تبدیل صوت برای هنجارسازی سیگنال استفاده کردند [۳]. هردو روش باعث افزایش کارایی در بازشناسی گفتار شدند. Narayanan و همکارش از ترکیب روش‌های VTLN و تبدیل خطی پارامترها برای تطبیق دادن یک شناسنده استفاده کردند. در این تجربه بدون تغییر مدل، از روش‌های انتقال بردار ویژگی استفاده شد [٤].

یکی از روش‌های افزایش کارایی این گونه سیستم‌ها استفاده از داده‌های گفتار کودکان برای آموزش و تطبیق با مدل‌های صوتی موجود می‌باشد. این در حالی است که جمع آوری گفتار کودکان نیز خود با مشکلات زیادی همراه است که از آن جمله می‌توان به مواردی مثل عدم همکاری و نداشتن توانایی کافی برای خواندن متن آماده شده اشاره کرد [۵]. از این رو تحقیقات، بیشتر به سمت تطبیق فن آوری بازشناسی گفتار بزرگسالان برای کاربرد در گفتار کودکان متمایل شده اند.

در این مقاله ابتدا ویژگی‌های اکوستیکی گفتار کودکان بررسی خواهند شد و سپس روش‌های پیشنهادی برای تطبیق شناسنده‌های گفتار برای استفادهٔ کودکان مطرح و ارزیابی می‌شوند. یکی از مسائلی که به ویژه در بازشناسی گفتار کودکان اهمیت دارد، تنوع بسیار زیاد ویژگی‌های اکوستیکی و زبانی در سنین پایین است. به این جهت یک سیستم بازشناسی پیشنهاد شده که در آن دو مدل بر مبنای خوشه بندی واژگان آموزش می‌بینند و در مرحلهٔ باز شناسی نیز، کلمه ورودی بر مبنای HMMهای خوشه ای که در آن قرار می‌گیرد بازشناسی می‌شود.

۲- ویژگی‌های فیزیکی و آوایی گفتار در کودکان

تأثیر سن گوینده بر درستی شناسنده‌های گفتار ابتدا توسط Wilpon و Jacobson مورد توجه قرار گرفت [٦]. بهترین کارایی در بازشناسی گفتار زمانی به دست می‌آید که برای هر گروه سنی مقدار دادهٔ کافی در مجموعهٔ آموزش وجود داشته باشد، هرچند که این مقدار برای کودکان و افراد سالخورده بسیار پایین است. در جدول (۱) ضریب همبستگی بین برخی از ویژگی‌های کودکان و نرخ بازشناسی نشان داده شده

است [۷].

جدول ۱ همبستگی بین برخی ویژگی‌های گوینده و نرخ خطای کلمه [۷]

ویژگی	همبستگی
سن	-۰/۵۱
قد	-۰/٤۹
میانگین F2 (Hz)	۰/٤۰
میانگین F0 (Hz)	۰/۳٦
جنسیت	-۰/۰۵
میانگین F1 (Hz)	۰/۰۵

با بررسی جدول مشاهده می‌شود که عامل جنسیت کودک، ضریب همبستگی بسیار پایینی با دقت بازشناسی گفتار دارد بنابراین نیازی نیست که تفاوتی بین گفتار دختر و پسر درنظر گرفته شود. دلیل این مسأله، عدم تفاوت بین ویژگی‌های گفتار دختر و پسر تا حدود سن ۱۰ سالگی می‌باشد. بالا بودن همبستگی بین قد کودک و نرخ خطا نشان می‌دهد که استفاده از روش‌هایی مثل VTLN در بازشناسی گفتار کودکان می‌تواند مفید باشد. سن، با بسیاری از ویژگی‌های کودک مثل قد و زمن یادگیری صحبت رابطه مستقیم دارد و همچنین همبستگی بالایی با نرخ خطای کلمه دارد، لذا توجه به سن کودکان یک عامل کلیدی در هنجار سازی گوینده و بازشناسی گفتار می‌باشد. این عامل، دلیل اصلی به کارگیری روش خوشه بندی در بازشناسی گفتار است که در این مقاله پیشنهاد شده است.

مطالعات در زمینهٔ گفتار کودکان تفاوت‌هایی را در خصوصیات اکوستیکی گفتار مثل سازه‌ها (Formant) ، طنین (Pitch) و تداوم صوت (Duration) نشان می‌دهد که این تفاوت‌ها وابسته به سن می‌باشند [٦۸]. دامنهٔ وسیع تر پارامترهای اکوستیکی و افزایش تنوع اکوستیکی، به طور چشمگیری کارایی بازشناسی را کاهش می‌دهند. این تفاوت‌ها را می‌توان به شکل زیر خلاصه کرد:

۱- با توجه به این که کودکان از محدوده صوتی (Vocal tract) کوتاهتر (فاصله بین تارهای صوتی تا لب‌ها) و دریچهٔ صوتی (Vocal fold) کوچکتری برخوردارند، فرکانس مبنایی (Fundamental frequency) (F_0) و پهنای باند سازه‌های آنها نسبت به بزرگسالان بالاتر است. این مسأله باعث می‌شود که فاصله زیادی بین همنواها (Harmonics) باشد و در نتیجه استخراج ویژگی‌های طیفی (Spectral) که یک مرحلهٔ اصلی در پردازش سیگنال است برای گفتار کودکان بسیار مشکل‌تر می‌باشد. به علاوه، در یک پهنای باند سیگنال مشخص مثل باند ٤ kHz تلفنی، سازه‌های کمتری در طیف گفتار کودکان نسبت به بزرگسالان وجود دارد. بنابراین، نمونه برداری پراکنده از طیف و سازه‌های نسبتاً کمتر در یک پهنای باند معین باعث ایجاد محدودیت‌های عمده در مقدار اطلاعات به دست آمده از صداها می‌شود.

۲- متغیر بودن مقادیر سازه باعث می‌شود محدودهٔ گروه‌های آوایی در کودکان فصل مشترک بزرگتری داشته باشند و بنابراین مسألهٔ دسته بندی الگو مشکل تر می‌شود.

۳- کودکان نسبت به بزرگسالان طنین صدای بالاتری دارند، طول بخش‌های آوایی آنها بیشتر است و گوناگونی طیف بیشتری دارند.

با توجه به موارد فوق، قطعا سیستم‌های شناسایی گفتاری که برای بزرگسالان طراحی شده اند نمی‌توانند در بازشناسی گفتار کودکان به خوبی عمل کند. بسیاری از سیستم‌های بازشناسی گفتار وقتی بر روی گفتار کودکان آزمایش می‌شوند موفق نبوده و مشاهده می‌شود که نرخ خطا نسبت به بزرگسالان حتی تا ۲۰۰٪ افزایش می‌یابد. یکی از روش‌های افزایش کارایی این گونه

سیستم‌ها استفاده از داده‌های گفتار کودکان برای آموزش و تطبیق با مدل‌های صوتی موجود می‌باشد که با توجه به مواردی که در بخش قبل ذکر شد با مشکلات زیادی همراه است. از این رو تحقیقات، بیشتر به سمت تطبیق فن آوری بازشناسی گفتار بزرگسالان و یا استفاده از دادهٔ نسبتا محدود برای کاربرد در گفتار کودکان متمایل شده اند.

۳- روش‌های افزایش کارایی بازشناسی گفتار در کودکان

استفاده از دادهٔ آموزش بیشتر در سیستم‌های بازشناسی گفتار متداول که هزینه و زمان زیادی صرف می‌کند، تنها راه حل برای افزایش کارایی در سیستم‌های بازشناسی گفتار کودکان نمی‌باشد، بلکه به دلیل وجود تفاوت‌های عمده موجود در گفتار کودکان، استفاده از روش‌های تبدیل صوت و تطبیق شناسنده برای کاربرد در گفتار کودکان پیشنهاد می‌شوند که در ادامه مورد بررسی قرار خواهند گرفت.

۳-۱- مدل سازی اکوستیکی قابل تطبیق با گوینده

یکی از روش‌های موفق در افزایش نرخ بازشناسی گفتار برای بزرگسالان و کودکان، مدل سازی قابل تطبیق با گوینده می‌باشد که هدف از آن کاهش تفاوت‌های اکوستیکی است که در اثر ویژگی‌های متفاوت گویندگان در مراحل آموزش و آزمایش ایجاد می‌شود. از جمله روش‌هایی که بر این اساس عمل می‌کنند می‌توان به VTLN، آموزش قابل تطبیق با گوینده (Speaker Adaptive Training) (SAT) و هنجارسازی گوینده بر اساس رگرسیون خطی با بیشترین درست نمایی (Maximum Likelihood Linear Regression) (MLLR) محدود شده (Constrained MLLR based Speaker Normalization) (CMLSN) اشاره کرد:

الف) SAT: هدف از روش آموزش قابل تطبیق با گوینده، تنظیم و کاهش تنوع آکوستیکی بین گویندگان مختلف با استفاده از تبدیل‌های خاص گوینده است که با استفاده از MLLR در توزیع‌های خروجی HMMها تخمین زده می‌شود [۹]. در این روش به جای تغییر در پارامترهای مدل، مشاهدات اکوستیکی هر گوینده برای آموزش و آزمایش تبدیل می‌شوند. یک مجموعه از مدل‌ها که کاملاً با دادهٔ غیر نرمال آموزش دیده اند به عنوان مدل هسته درنظر گرفته می‌شوند و پارامترهای گوسین این مدل‌ها به طور تکرار شونده با بکارگیری تبدیل‌های تخمین زده شده با استفاده از دادهٔ آموزش مجدداً تخمین زده می‌شوند. در مرحلهٔ بازشناسی، پارامترهای انتقال مجددا با توجه به مدل‌های استفاده شده در کدگشایی تخمین زده می‌شوند.

ب) CMLSN: یک روش هنجارسازی گوینده است که از انتقال بردارهای مشاهدات اکوستیکی با به کار گیری تبدیل MLLR خاص گوینده استفاده می‌کند. این انتقال با هدف کاهش عدم تطابق دادهٔ گویندگان با توجه به مجموعه ای از HMM های هدف انجام می‌شود که با مجموعهٔ HMM هایی که برای بازشناسی استفاده می‌شوند متفاوتند. این فرایند در هر دو مرحلهٔ آموزش و رمزگشایی انجام می‌شود [۱۰].

ج) VTLN: بدیهی است که طول محدودهٔ صوتی در گویندگان متفاوت است. افرادی که طول محدودهٔ صوتی آنها کوتاهتر است صدای زیرتری دارند و برعکس. این تفاوت، باعث عدم تطابق بین آموزش و آزمایش شناسنده می‌شود و بنابراین کارایی بازشناسی کاهش می‌یابد. در طول مرحلهٔ استخراج ویژگی می‌توان این تفاوت‌ها را کاهش داد. به این فرایند، کاهش محدوده صوتی گفته می‌شود. اساس این روش گسترش یک لوله (tube) به اندازه فاکتور α است فرکانس طیف را تغییر می‌دهد. مثلا با دو برابر کردن طول لوله، تشدید (resonance) صدا از ۱۰۰Hz به ۲۰۰ Hz تغییر می‌کند. همچنین برای این منظور می‌توان مقیاس بردار فرکانس را تغییر داد. VTLN تغییر مقیاس بردار فرکانس به صورت خطی، دوخطی یا غیرخطی می‌باشد. در این

مقالـه از یـک روش تغییـر مقیـاس اسـتفاده شـده تـا فیلتربانـک بهینـه بـرای اسـتخراج خطـی ویژگـی‌هـای اکوستیکی از گفتار کودکان تعیین شود. در مرحلهٔ بعد با اسـتفاده از هنجارسـازی نـرخ گفتـار کودکـان، کـارایی افزایش می‌یابد. هنجارسازی را مـی‌تـوان بـا اسـتفاده از تطبیق طیف در مرحله استخراج ویژگی‌ها با تغییر فـضا و پهنای فیلترهای مقیاس مل انجام داد. VTLN هـم در طی مرحلهٔ آموزش و هم در طی فرایند بازشناسی قابـل اجرا است. در مرحلهٔ آمـوزش، توابـع تطبیـق فرکـانس وابسته به گوینده برای ایجاد مدل‌های اکوستیکی نرمـال شده اجرا می‌شوند. در مرحلـهٔ بازشناسـی از یـک تـابع تطبیق وابسته به گفتـار اسـتفاده مـی‌شـود کـه در زمـان محاسبهٔ بردارهای ویژگـی بـه کـار مـی‌رود. بـرای ایـن منظور نیاز به کنترل کامل بر روال‌هـای آمـوزش و رمـز گشایی می‌باشد.

VTLN بر اساس این فرضیه عمـل مـی‌کنـد کـه بخش قابل توجهی از تفاوت‌های ویژگی‌های گویندگان را می‌تـوان بـا اسـتفاده از تغییـر خطـی مقیـاس محـور فرکانس کاهش داد. در بسیاری از کاربردهـا لازم اسـت تطبیق بـه گوینـدهٔ جـاری بـسیار سـریع انجـام شـود و بنابراین عملیات محاسباتی و جستجو برای یافتن تغییـر مقیاس فرکانس خطی هزینه زیادی خواهـد داشـت. در این مقاله، تغییر مقیاس یک بار برای کل گروه گویندگان تعیین می‌شود و بیشتر تاکید بر تفاوت بین صدای اطفال و بزرگسالان است و نه گویندگان خاص. بنابراین تعیین مقیاس فرکاس به صورت بـرون خطـی و نـه بلادرنـگ انجام می‌شود. در ادامـه الگـوریتم مـورد اسـتفاده ارائـه می‌شود [۱۱]. تغییر مقیاس محور فرکانس با بهینه سازی منحنی مل (Mel-curve) انجام می‌شود که فیلتربانک را برای محاسبه ضرایب MFCC تعیین می‌کند.

منحنی مل با رابطهٔ زیر داده می‌شود:

$$\text{mel}(f) = 1125\ln(1+\frac{f}{700}) \qquad (۱)$$

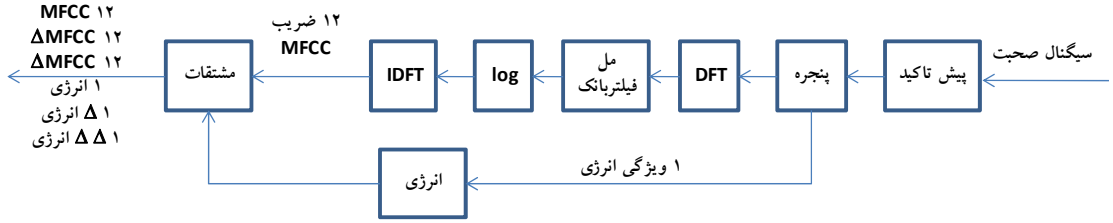
به منظور تطبیق منحنـی بـا ویژگـی‌هـای داده، n نقطـهٔ نمونـه (fi,p(fi)) تعیـین مـی‌شـوند کـه بـه طـور یکنواخت در مقیاس مل قرار گرفته اند:

$$f_i = \text{mel}^{-1}\left((i-1).\frac{f_{max}}{n-1}\right), i \in 1, \ldots, n \qquad (۲)$$

F_{max} وابسته به فرکانس نمونـه بـرداری اسـت و ۸۰۰۰Hz درنظر گرفته شده است. مقداردهی اولیهٔ نقاط نمونه با منحنی مل داده شده است:

$$p(f_i) = \text{mel}(f_i),\ i \in 1, \ldots, n \qquad (۳)$$

منحنـی بهینـه شـده opt(f) شـامل نقـاط نمونـه می‌باشد. الگوریتم بهینـه سـازی بـه طـور تکرارشـونده مقادیر $p(f_i)$ را تغییر می‌دهد.



شکل ۱ استخراج دنباله بردار ویژگی MFCC، ۳۹ بعدی با استفاده از شکل موج دیجیتالی شده کوانتیزه

هدف بهینه سازی، به حداکثر رساندن نرخ بازشناسی می‌باشد. خلاصهٔ این روال در الگوریتم (۱) ذکر شده است.

الگوریتم (۱): تطبیق پارامترها

f_i و $p(f_i)$ با توجه به معادلات (۱) و (۲) مقداردهی می‌شوند.

با توجه به هدف بهینه سازی زیر، مقدار p(fi) به طور تکرارشونده تطبیق می‌یابد.

الف) منحنی opt(f) برای نقاط (fi,p(fi)) محاسبه می‌شوند

ب) فیلتربانک با توجه به opt(f) تعیین می‌شود.

ج) ویژگی‌های اکوستیکی با استفاده از فیلتربانک جدید استخراج می‌شوند.

د) نرخ بازشناسی محاسبه می‌شود.

ه) روال تکرار می‌شود.

۳- مقدار opt(f) برگردانده می‌شود.

مزیت الگوریتم این است که تعداد پارامترهایی که باید بهینه شوند را می‌توان مستقل از تعداد فیلترهایی که در استخراج ویژگی به کار می‌روند انتخاب کرد.

۲-۳- روش‌های تبدیل صوت

در سیستم‌های بازشناسی گفتار در صورتی که دسترسی به لایهٔ ویژگی‌ها امکان پذیر نباشد امکان استفاده از روش VTLN وجود ندارد. در این صورت می‌توان روش‌های تبدیل صوت را برای هنجارسازی سیگنال به کار برد [۳]. در این روش، تبدیل‌ها بر روی اصواتی که در ۱۶ kHz نمونه برداری شده انجام می‌گیرد و سپس سیگنال پیش از ورود به شناسنده که از مدل‌های اکوستیکی با پهنای باند تلفنی استفاده می‌کند در ۸ kHz نمونه برداری می‌شود. با توجه به این فرکانس صدای کودکان بالاتر از بزرگسالان است، امکان به کارگیری برخی از اطلاعات طیفی ایجاد می‌شود که در صورت استفاده از روش ضبط در پهنای باریک از بین می‌رفتند. یکی از روش‌های متداول در ای زمینه، الگوریتم TD-PSOLA (Pitch-sinchronous overlap-add) می‌باشد. این الگوریتم که در ترکیب گفتار برای ایجاد تغییرات در طنین و تداوم صوت سیگنال گفتار به کار می‌رود، با یک کاهش ۲۳ درصدی در طنین و یک فشرده سازی طیفی نتایج قابل قبولی را تولید می‌کند[۱۲]. همچنین کاهش چشمگیری در نرخ خطای بازشناسی گفتار کودکان به خصوص زیر ۱۰ سال نتیجه می‌دهد، هر چند که هنوز هم برای کاربرد بازشناسی گفتار کودکان در سنین پایین تر مناسب نیست.

٤- پیاده سازی و نتایج

درحال حاضر، متداول ترین روش طراحی سیستم‌های بازشناسی گفتار، برمبنای مدل مخفی مارکوف (HMM) می‌باشد [۱۳]. در این مقاله از ابزار HTK برای پیاده سازی این سیستم استفاده شده است [۱٤].

اولین مرحله در طراحی یک سیستم بازشناسی گفتار، فاز استخراج ویژگی است که در این مقاله از ضرایب MFCC (Mel Frequency Cepstral Coefficients) استفاده شده است. مراحل ایجاد مجموعه ویژگی‌ها در شکل (۱) نشان داده شده است که نهایتاً شامل ۱۲ ضریب کپسترال و یک انرژی به همراه مشتق‌های اول و دوم آنها می‌باشد.

برای بررسی نتایج حاصله از طراحی شناسنده، سیستم‌های زیر پیاده سازی شدند [۱۵،۱٦]:

۱- مدل آموزش یافته با استفاده از گفتار بزرگسالان.

۲- مدل آموزش یافته با استفاده از گفتار کودکان .

۳- سیستمی که با استفاده از گفتار کودکان آموزش یافته است و در آن از روش‌های تطبیق استفاده شده است.

٤- سیستم آموزش یافته با گفتار بزرگسالان که در آن از روش‌های تطبیق استفاده شده است. در این بخش دو زیر سیستم نیز طراحی شد که یکی بر اساس گفتار بزرگسالان زن و دیگری بزرگسالان مرد می‌باشد.

٤-١- دادگان مورد استفاده

استفاده از واژگان مناسب در نتایج سیستم‌های بازشناسی گفتار تأثیر زیادی دارد. پایگاه داده مورد استفاده در این پایان نامه از دو بخش تشکیل شده است، پایگاه داده گفتار بزرگسالان و کودکان. داده بزرگسالان برای آموزش شناسنده بزرگسالان و همچنین در ارزیابی نتایج به عنوان مرجع به کار رفته است. داده کودکان در آموزش، تطبیق و ارزیابی شناسنده به کار رفته است.

برای پایگاه داده بزرگسالان از داده گفتار گسسته موجود در پایگاه داده فارس دات استفاده شده است که در اتاق اکوستیک آزمایشگاه زبان شناسی دانشگاه تهران تهیه شده است.

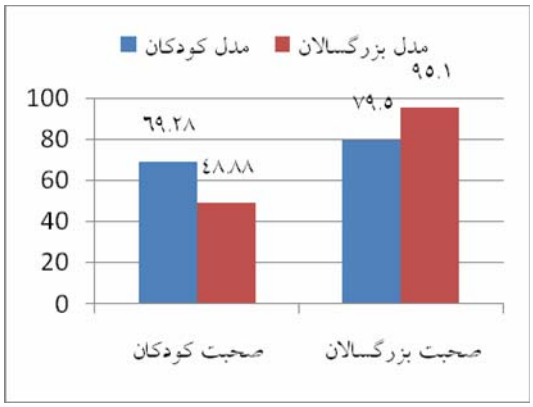
متأسفانه پایگاه داده استانداردی برای گفتار کودکان فارسی زبان وجود ندارد. واژگان به کاررفته در این پایان نامه از داده گفتار ۵۰ کودک سنین ۲ تا ۱۰ سال در محیطی با نویز کم جمع آوری شده است. به دلیل مشکلات موجود در انتقال کودکان به محل ضبط صدا، برای افزایش داده گفتار، هر کودک دنباله ای از ارقام را ۲۰ مرتبه (بدون در نظر گرفتن حالت‌های غیر قابل قبول) تکرار کرده است.

برای بررسی نقش هر یک از عوامل، آزمایش‌های متفاوتی اجرا و نتایج آنها به شرح زیر مورد بررسی قرار گرفت.

٤-٢- نقش دادهٔ آموزش

در ابتدا، شناسنده با استفاده از داده بزرگسالان آموزش یافته و با یک بار با استفاده از داده بزرگسالان و یک بار با استفاده از داده کودکان نیز آزمایش شد. در شکل (۲) نتایج بازشناسی گفتار برای دو مدل آکوستیکی نشان داده شده است. همانگونه که انتظار می‌رود، کارایی برای شرایطی که داده‌های آموزش و آزمایش با هم منطبق نیستند پایین تر است.

نکته دیگری که مورد بررسی قرار گرفته است جنسیت دادهٔ آموزش می‌باشد. درستی شناسنده بر مبنای سن با چهار مدل در جدول (۲) نشان داده شده است.



شکل ۲ کارایی شناسندهٔ کلمه با دو مدل آکوستیکی

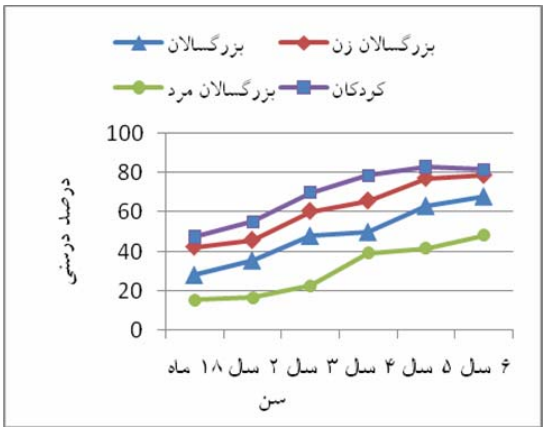
با بررسی میانگین درستی شناسنده مشاهده می‌شود که به طور کلی مدل کودکان بالاترین درستی را دارد، در حالیکه درستی مدل بزرگسالان و بزرگسالان زن حدود ۲۹٪ و ۱۱٪ کمتر از مدل مبنا (کودکان) می‌باشد و مدل بزرگسالان مرد کمترین کارایی (۵۶٪ کمتر از مدل مبنا) را دارد. این نتایج نشان می‌دهد که ویژگی‌هایی اکوستیکی گفتار کودکان به گفتار بزرگسالان زن شباهت بیشتری دارد.

جدول ۲ تغییرات درستی شناسنده آموزش داده شده با گروه‌های سنی مختلف

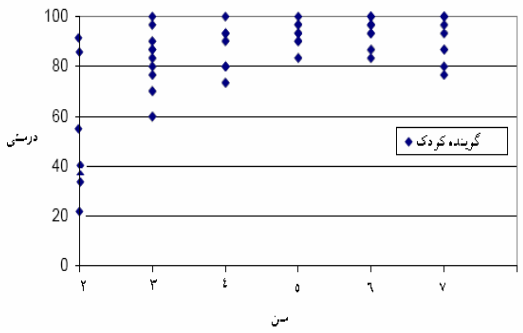
سن	بزرگسالان	بزرگسالان زن	بزرگسالان مرد	کودکان
میانگین	٤٨,٨٨	٦١,٧٣	٣٠,٥٧	٦٩,٢٨
تفاوت دقت با مدل مبنا (کودکان)	٪٢٩	٪١١	٪٥٦	٪٠

٤-٣- نقش سن

درستی شناسنده با چهار مدل بدون استفاده از هنجارسازی در شکل (۳) مشاهده می‌شود. نکته ای که در این نتایج قابل توجه است، تغییرات درستی شناسنده کودکان با سن می‌باشد. دقت بازشناسی در میانهٔ این بازه (۳ تا ۵ سال) بالاتر و در کرانه‌ها (زیر ۲ سال و حدود ٦ سال) پایین تر می‌باشد. یک دلیل عمده کاهش کارایی در حدود سن ٦ سال، از دست دادن دندان‌های شیری و تلفظ متفاوت کلمات می‌باشد.



شکل ۳ دقت بازشناسی گفتار کودکان به صورت تابعی از سن با چهار مدل مختلف

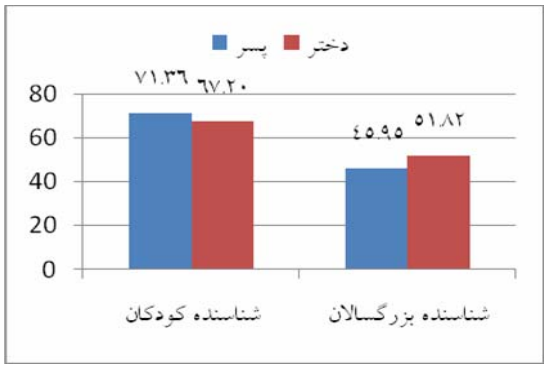


شکل ٤ درصد درستی بازشناسی برای هر گوینده

در شکل (٤) نمودار دقت بازشناسی برای هر کودک نشان داده شده که تفاوت قابل توجهی را در میزان بازشناسی برای هر کودک در یک گروه سنی خاص نشان می‌دهد که به دلیل قابلیت‌های فردی کودکان مختلف در بیان کلمات می‌باشد. درصورتی که ضریب همبستگی بین سن و دقت شناسنده محاسبه شود، عدد ۰/۸٦ به دست می‌آید که نشان دهندهٔ تأثیر بسیار قابل توجه سن در دقت بازشناسی گفتار است.

٤-٤- نقش جنسیت

برای بررسی این که آیا جنسیت کودکان نقش قابل توجهی در نتایج بازشناسی دارد یا خیر، دقت برای هر جنس به طور جداگانه به دست آمده که نتایج آن در شکل (۵) دیده می‌شود. این که دقت بازشناسی گفتار کودکان پسر در شناسنده‌های بزرگسال بالاتر است به این دلیل می‌باشد که تفاوت‌های آکوستیکی گفتار کودکان دختر نسبت به پسر بیشتر می‌باشد. از طرفی معمولاً کودکان دختر واضح تر صحبت می‌کنند، بنابراین دقت بازشناسی شناسندهٔ کودکان بر گفتار کودکان دختر بالاتر است. البته همانطور که مشاهده می‌شود این تفاوت‌ها ناچیز می‌باشد.



شکل ۵ درستی شناسنده برای کودکان دختر و پسر

٤-٥- کاربرد روش‌های تبدیل صوت در بازشناسی گفتار

در این پایان نامه از اجرای PSOLA در نرم افزار PRAAT برای تغییر در طول گفتار کودکان استفاده شده است. جدول (۳) نتایج بازشناسی گفتار را در یک آزمایش نمونه که برای بازشناسی گفتار کودکان بین ۲ تا ۱۰ سال انجام گرفته نشان می‌دهد نرخ خطای پرس و

جو حدود ٤٥٪ کاهش یافته است. برای مشخص شدن این مسأله که آیا روش برای کودکان کم سن ترنتایج متفاوتی دارد یا خیر، داده‌ها به دو گروه سنی ٦-٢ سال و ١٠- ٦ سال تقسیم شده است.

جدول ٣ نرخ خطای کلمه به درصد برای دو گروه سنی

روش تبدیل	٦-٢سال	١٠-٦ سال
-	٥٣/٢	١٤/٧
PSOLA	٣٩/٩	٩/٣

٤-٦- نتایج تجربی حاصل از اجرای VTLN

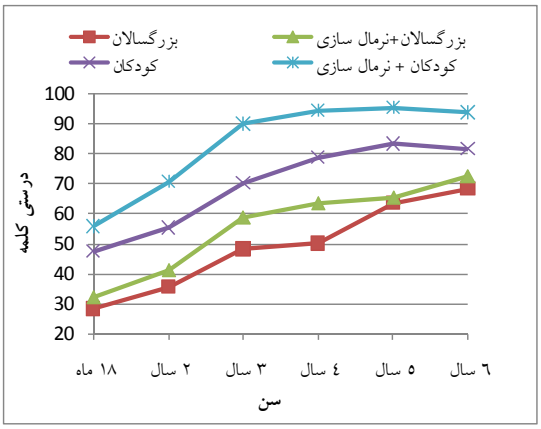
در جدول (٤) نتایج به دست آمده از بازشناسی گفتار کودکان با استفاده از دادهٔ آموزش کودکان، بزرگسالان و ترکیبی از هردو برای مدل مخفی مارکوف (HMM) قبل و بعد از هنجارسازی نشان داده شده است.

جدول ٤ نرخ خطای بازشناسی ارقام برای گویندگان کودک قبل و بعد از هنجارسازی

	قبل از هنجارسازی	بعد از هنجارسازی	تغییر کارایی
HMM بزرگسالان	٤٨,٨٨	٥٥,٥٨	٪١٤
HMM کودکان	٦٩,٢٨	٨٣,١٣	٪٢٠

در این مدل، ترکیبی از روش‌های VTLN و تبدیل خطی پارامترها برای تطبیق دادن شناسندهٔ ارقام به کار رفته است. بعد از اجرای فرایند هنجارسازی گوینده، نرخ درستی بازشناسی برای کودکان در سنین بالاتر (حدود ٩ سال) قابل مقایسه با بزرگسالان است، هرچند که برای سنین پایین تر هنوز هم کارایی مطلوب به دست نمی‌آید. استفاده از روش‌های تنظیم فرکانس برای هنجارسازی گفتار کودکان قبل از آموزش مدل‌های اکوستیکی نرخ خطای بازشناسی را تا ٥٥٪ کاهش می‌دهد.

با استفاده از روش‌های تطبیق مدل و VTLN می‌توان کارایی مدل‌های بزرگسال را برای استفادهٔ گفتار کودکان افزایش داد، هرچند که استفاده از دادهٔ آموزش کودکان نتیجهٔ بهتری دارد.



شکل ٦ درستی کلمه با توجه با افزایش سن قبل و بعد از هنجارسازی

در شکل (٦) دقت بازشناسی با استفاده از دو مدل کودکان و بزرگسالان قبل و بعد از هنجارسازی که روش‌های آن در بخش‌های قبل بررسی شدند، مشاهده می‌شود. در اینجا فقط از روش VTLN استفاده شده است.

٤-٧- کاربرد خوشه بندی برای افزایش کارایی

یکی از نکاتی که در نحوه صحبت کودکان به چشم می‌خورد تنوع در نحوه بیان صداها و کلمات توسط آنها می‌باشد. برخلاف بزرگسالان، تفاوت بین گفتار بین کودکان مختلف حتی در بازه‌های سنی محدود بسیار گسترده است. نتایج تجربی نشان می‌دهند که تنوع در ویژگی‌های آکوستیکی گفتار کودکان را نمی‌توان به سادگی با مدل‌های آکوستیکی مستقل از سن نشان داد. از این جهت، در این بخش طراحی و پیاده سازی سیستمی که قادر باشد شناسنده مناسب را وابسته به صدای گوینده انتخاب کند پیشنهاد می‌شود. برای رسیدن به این هدف، از خوشه بندی گفتار کودکان برای نمایش دقیق تر تنوع در ویژگی‌های اکوستیکی کودکان،

و همچنین افزایش کارایی به خصوص در سنین پایین تر استفاده شده است. تعداد خوشه‌ها ثابت و برابر ۲ درنظر گرفته شده است ({۲ تا ٤ سال}، {٥ تا ٦ سال}). بنابراین به جای استفاده از یک شناسنده، از دو شناسنده استفاده می‌شود که براساس نتیجه حاصل از خوشه بندی می‌توان برای بازشناسی هر وردوی، مناسب ترین مدل را انتخاب نمود. این روش قابل تعمیم به k مدل نیز می‌باشد.

در مرحلهٔ خوشه بندی که در شکل (۷) نشان داده شده است، به صورت زیر عمل شده است:

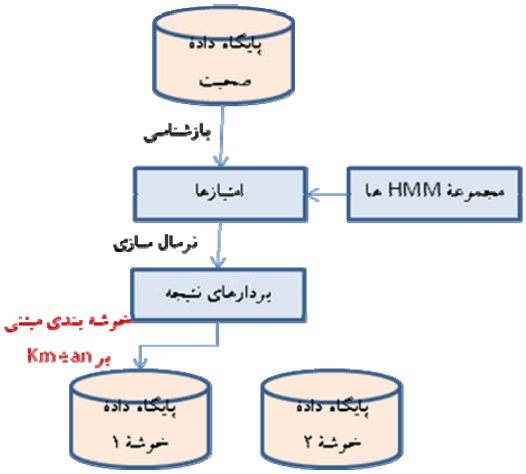
الف) HMM ها برای هر سن با استفاده از دادهٔ همان بازهٔ سنی آموزش دیده شده اند.

ب) یک HMM بر اساس دادهٔ بزرگسالان مستقل از جنسیت ایجاد شده است.

ج) روال خوشه بندی بر روی هر واژه در پایگاه داده اجرا شده است که نشان دهندهٔ گوناگونی در ویژگی‌های گفتاری بین گویندگان می‌باشد. به ویژه این که کودکان برخی کلمات را کاملاً واضح و برخی را گنگ بیان می‌کنند.

د) HMM ها براساس هر واژهٔ پایگاه داده آموزش، یک مقدار امتیاز تولید می‌کنند.

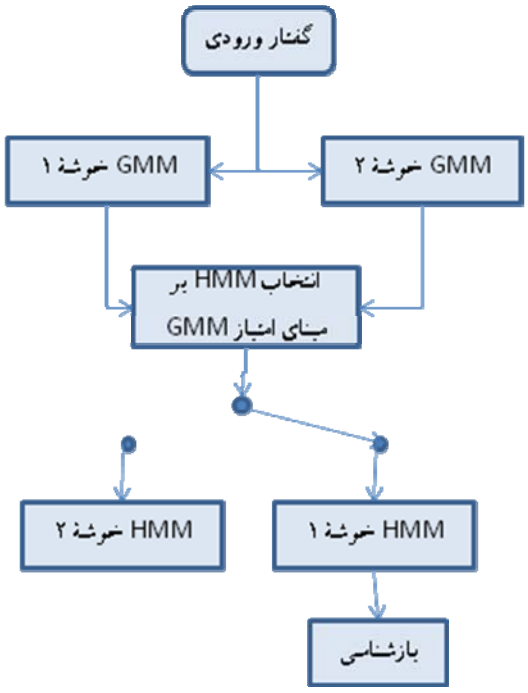
ه) خروجی امتیازها به صورت یک بردار می‌باشد که با توجه به مقدار امتیاز حداکثر، نرمال می‌شوند.



شکل ۷ اجزاء سیستم پیشنهادی

و) با تکرار روال فوق برای n دادهٔ آموزش، n بردار نرمال شده به دست می‌آید.

ز) از روال خوشه بندی براساس Kmeans برای ایجاد ۳ خوشهٔ مورد نظر استفاده می‌شود.

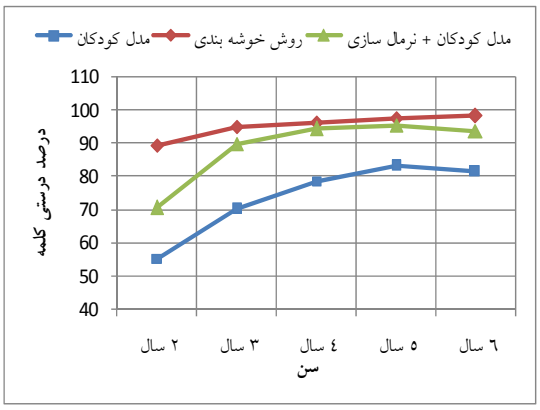


شکل ۸ بازشناسی مبتنی بر خوشه بندی

در این مرحله، دو مدل اکوستیکی برای هر خوشه آموزش داده شده اند. با توجه به محدود بودن دادهٔ آموزش برای هر خوشه، از مدل کودکان مستقل از سن استفاده شده است که با استفاده از روش‌های تطبیق که در بخش‌های قبل به آنها اشاره شد، برای هر خوشه تطبیق یافته اند. مدل ترکیبی گوسین (GMM) (Gaussian Mixture Models) با استفاده از تخمین بیشترین درست نمایی(MLE) (Maximum Likelihood Estimation) با استفاده از داده‌های گفتار هر خوشه آموزش یافته است. در مرحلهٔ بازشناسی، هر گفتار ورودی به خوشه ای که GMM آن بالاترین تطبیق را با گفتار ورودی دارد مرتبط می‌شود و با استفاده از

HMMهای همان خوشه بازشناسی می‌شود (شکل (۸)).

همانطور که مشاهده می‌شود، استفاده از روش خوشه بندی علیرغم افزایش پیچیدگی سیستم، کارایی شناسنده را افزایش می‌دهد. دلیل این امر وابسته بودن گفتار به محدودهٔ سنی گوینده به ویژه در سنین پایین می‌باشد. بنابراین با طراحی مدل‌های وابسته به سن یا ویژگی‌های آوایی میتوان بازشناسی هر کلمه را با استفاده از نزدیکترین مدل به آن که توسط خوشه بندی تعیین می‌شود انجام داد. با افزایش تعداد خوشه‌ها پیچیدگی سیستم و همچنین کارایی افزایش می‌یابد. در این مقاله براساس نتایج تجربی و آزمایش‌های انجام گرفته، دو خوشه پیش بینی شده است. نکته قابل توجه در این روش این است که در مرحله بازشناسی بدون توجه به سن گوینده، ورودی به نزدیکترین خوشه تخصیص می‌یابد. بنابراین کودکانی که سن بالاتری دارند ولی هنوز در بیان کلمات دچار مشکل هستند در خوشه اول قرار میگیرند و به این ترتیب، دقت بازشناسی افزایش می‌یابد.



شکل ۹ دقت بازشناسی مدل مبتنی بر خوشه بندی

نتایج بازشناسی در شکل (۹) نشان داده شده‌اند.در این محدودهٔ سنی، جنسیت گوینده تأثیری در نرخ بازشناسی ندارد، هرچند که با افزایش سن، ویژگی‌های اکوستیکی وابسته به جنسیت آشکارتر خواهند شد. در سنین بالاتر با توجه به افزایش قابلیت‌های گفتاری کودکان می‌توان از مدل‌های گفتار بزرگسالان استفاده نمود.

۵- نتیجه‌گیری

گفتار کودکان از نظر خصوصیات اکوستیکی و زبانی با بزرگسالان متفاوت است. تغییرات در کودکان به خصوص رشد محدودهٔ صوتی باعث ایجاد تنوع در پارامترهای طیفی سیگنال گفتار کودکان می‌شود. این موارد باعث بروز مشکلاتی در بازشناسی خودکار گفتار و افزایش نرخ خطا می‌شود. در این مقاله، چندین مدل بر مبنای واژگان گروه‌های سنی مختلف طراحی و نتایج آنها مقایسه شد. استفاده از مدل‌های مبتنی بر گفتار بزرگسالان، با توجه به کارایی بسیار پایین، عملا غیر ممکن و ناکارامد است که با استفاده از روش‌های هنجارسازی و تطبیق گفتار می‌توان این کارایی را تا حدی افزایش داد. باید توجه داشت که خصوصیات گفتاری کودکان بسیار با بزرگسالان متفاوت و دامنه تغییرات آنها وسیع تر است، بنابراین لازم است روش‌هایی برای تطبیق و هنجارسازی گفتار کودکان به کار روند. با توجه به این که عمدهٔ تفاوت کودکان به لحاظ خصوصیات فیزیکی و درنتیجه اندازهٔ تارهای صوتی و محدودهٔ صوتی آنها می‌باشد، استفاده از روش هنجارسازی محدوده صوتی پیشنهاد گردید. همچنین از روش‌های تغییر در نرخ گفتار برای هنجارسازی گفتار ورودی استفاده شد. در نهایت با توجه به تنوع وسیع گفتار کودکان با یکدیگر و همچنین با بزرگسالان استفاده از یک مدل مبتنی بر خوشه بندی گفتار پیشنهاد شد. در این مدل، گفتار کودکان به دو خوشه مبتنی بر سن تقسیم شد و دو شناسنده برمبنای این دو خوشه گفتاری طراحی شد. با دریافت هر ورودی، خوشه مناسب انتخاب و بازشناسی براساس آن انجام می‌گیرد. با استفاده از این روش، دقت بازشناسی تا ۵۰٪ افزایش یافت.

ادامه این تحقیقات می‌تواند منشاء تحولی فراگیر در عرصه‌های روان شناسی، آموزشی و کامپیوتر قرار گیرد که برخی از آنها به شرح زیر می‌باشد:

الف) در حوزه آموزشی: با توجه به علاقه ای که کودکان به استفاده از سیستم‌های کامپیوتری نشان می‌دهند ایجاد رابطهای نرم افزاری که توجه کودکان را جلب کند به عنوان ابزار بسیار مناسبی برای ابزارهای کمک آموزشی مانند آموزش زبان‌های خارجی، به کار می‌رود. متأسفانه از رهاوردهای زندگی ماشینی امروزه، گرفتاری‌های والدین و نبودن وقت کافی برای ارتباط با کودکان می‌باشد. با استفاده از ابزارهای آموزشی میتوان تا حد کمی این خلا را جبران نمود هرچند که در بعد روان شناسی هنوز مشکلات بسیاری باقی خواهد ماند.

ب) در حوزهٔ پزشکی: ایجاد ابزارهای مکانیزهٔ تعیین نارسایی‌های گفتاری و شنوایی

ج) در حوزهٔ سرگرمی: ایجاد بازی‌هایی که همزمان با ایجاد سرگرمی، بعد آموزشی و فرهنگی در آنها در نظر گرفته شده باشد.

در انتها، مواردی که می‌توانند به عنوان زمینه ای برای ادامه پژوهش و تکمیل بحث انتخاب شوند به شرح زیر پیشنهاد می‌گردند:

الف) بسط محدودهٔ سنی و ایجاد سیستم‌های هوشمند که قادر به بازشناسی آواها و اصوات کودکان از زمان آغاز به تکلم باشد.

ب) ایجاد سیستم بازشناسی کلمات پیوسته.

ج) افزایش تحمل نویز: با توجه به این که جمع آوری داده گفتار کودکان و ضبط صدای کنترل شده در محیط بدون نویز بسیار مشکل می‌باشد، بهتر است جمع آوری داده در محیط‌های معمول مثل کودکستان و مدرسه انجام شود.

٦- مراجع

1. M. Gerosa, D. Giuliani and F. Brugnara, "Speaker adaptive acoustic modeling with mixture of adult and children's speech", In *p*roceedings of the european conference on speech communication and technology [INTERSPEECH2005], Lisbon, Portugal, pp. 2193-2196, 2005.
2. S. Das, D. Nix, M. Picheny, "Improvements in children's speech recognition performance". Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume:1 , 12-15. pp. 433 – 436, May 1998.
3. J .Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system", In the Proceedings of the International Conference on Spoken Language Processing, pp 297 – 300, 2002.
4. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children", IEEE Transactions on speech and audio processing, Vol. 10, No. 2, February 2002.
5. M. Blomberg and D. Elenius, "Collection and recognition of children's speech in the PF-Star project", Ume University, department of philosophy and linguistics, PHONUM 9, 81-84, 2003.

6. J.G. Wilpon, C.N. Jacobsen, "A study of speech recognition for children and the elderly", In proceedings: ICASSP-96. IEEE international conference on acoustics, speech, and signal processing, vol. 1, 3527-10 May 1996.
7. D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children" Interspeech, Portugal, September 2005.
8. S. Lee, A. Potamianos and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters", Journal of Acoust. Soc. Amer., vol. 105, no. 3, pp. 1455–1468, March 1999.
9. G. Stemmer, C. Hacker, S. Steidl and E. N¨oth, "Acoustic Normalization of Children's Speech", In EUROSPEECH, Geneva , Switzerland , pp. 1313-1316, 2003.
10. Giuliani, D., Gerosa, M. and Brugnara, F. "Speaker Normalization through Constrained MLLR Based Transforms" in Proc. of INTERSPEECH/ICSLP, Jeju Island, Korea, Oct. 2004, pp. 2893–2897.
11. G. Stemmer, C. Hacker, S. Steidl, and E. N¨oth, "Acoustic Normalization of Children's Speech", In EUROSPEECH, Geneva , Switzerland , pp. 1313-1316, 2003.
12. J .Cabral and L. Oliveira, "Pitch-Synchronous Time-Scaling for Prosodic and Voice Quality Transformations", Interspeech ,Lisbon, Portugal, September 2005.
13. A. Ogawa and S. Takahashi, "Children's Speech Recognition Based on Clustering Techniques." NTT Cyber Space Laboratories, 3 (12), 75-81, 2005.
14. S. Young, "The HTK Book", Revised for HTK Version 3.4, Cambridge University Engineering Department, Dec 2006.
15. G. Tadayon Tabrizi, S. Setayeshi and M. Molavi, "HMM-based recognition and adaptation of Persian children's speech". submitted to: ELEX Journal, 2009.
16 G. Tadayon Tabrizi, S. Setayeshi and M. Molavi, "Applying acoustic normalization to improve recognition of children's speech", Proc. 16th Iranian conference on electrical engineering, May 2008 (In Persian).