دانشکده : برق و رباتیک

گروه : الکترونیک

پایان نامه دوره ی کارشناسی ارشد مهندسی برق – الکترونیک

کاربرد روش استخراج ویژگی RootMel جهت تخمین سن افراد با استفاده از سیگنال گفتار

عاطفه دهقانیان

استاد راهنما :

دکتر حسین مروی

استاد مشاور :

دکتر علی سلیمانی

شهریور ۱۳۹۱


بسم الله الرحمن الرحيم

تقدیم به پدر ومادرم

که هستی و همه ی وجودم از آن هاست

و همسرم

به پاس قدرانی و سپاسگذاری از زحمات و

دلسوزی هایشان

تشکر و قدردانی

در ابتدا از استاد راهنمای خود، جناب آقای دکتر حسین مروی به خاطر راهنمایی و حمایتی که از من در طی انجام این تحقیق کرده اند، نهایت تشکر را می نمایم.

و نیز از تمامی دوستانم در دانشگاه صنعتی شاهرود، که با حمایت و کمک بی دریغ خود در انجام این تحقیق، مرا همراهی کردند؛ تشکر می نمایم.

از پدر و مادر و همسرم به خاطر کمک ها و دلسوزی هایشان و به خاطر حمایت های معنوی که در طی انجام این تحقیق از من داشته اند، تشکر ویژه می نمایم.

چکیده :

تخمین سن بر اساس ویژگی های گفتار انسان، یک موضوع قابل توجه در سیستم های شناسایی گفتار اتوماتیک می باشد. مطالعاتی در زمینه ی تخمین سن گوینده صورت گرفته است ولی نیاز به کار های نوین بیشتری، خصوصا برای گوینده های فارسی زبان، می باشد. در تخمین سن، مانند سایر سیستم های پردازش گفتار، با دو چالش مهم مواجه هستیم: یافتن یک روش مناسب برای استخراج ویژگی و انتخاب یک روش قابل اطمینان برای کلاسه بندی.

هدف اصلی از این تحقیق استفاده از ویژگی Root Mel Ferequency Cepstral Coefficients جهت بخش استخراج ویژگی در یک سیستم تخمین سن و یافتن بهترین مقدار برای داشتن درصد خطای کمتر می باشد؛ و همچنین مقایسه ی عملکرد این ویژگی با سایر ویژگی های متداول نظیر MFCC معمول، PLP و LPC نیز مورد بررسی قرار گرفته است .

برای استخراج ویژگی، کل سیگنال صوت را به کوچک ترین جزء آن، یعنی واج ها، تفکیک می کنیم و عملیات استخراج ویژگی و رده بندی را بر سیگنال مربوط به واج ها انجام می دهیم.

از کلاسه بند به روش های تفکیک پذیری خطی و فواصل Mahalanobis استفاده شده است.

نتیجه آزمایشات انجام شده بر پایگاه داده ی FARSDAT به خطای کمترین ۲۸.۶۹ ٪ بازای ریشه ی ۰.۰۰۶ در استفاده از روش RootMFCC و تاثیر مثبت تفکیک سیگنال صوتی به واج های تشکیل دهنده ی آن، در کاهش خطا می باشد.

کلمات کلیدی : تخمین سن، گفتار، واج، استخراج ویژگی RootMel، FARSDAT

فهرست مطالب :

فهرست جدول ها:

فهرست شکل ها:

فهرست علائم و اختصارات:

Pitch : فرکانس صدا

Octave : فاصله ی زمانی(مدت) بین یک pitch و pitch دیگری با فرکانس دو برابر و یا نصف

Jitter : حداکثر مقدار انحراف از فرکانس پایه (F0)

Shimmer : تغییرات محلی توان

Formant :قله های موجود در طیف صوت (تجمع انرژی صوت در یک فرکانس خاص از شکل موج صوت)

linguistic information :

اطلاعات نمادین که با مجموعه ای از نمادهای مجزا و قواعدی برای ترکیب آن ها، تعریف می شوند.

paralinguistic information :

اطلاعاتی که از نوشته ها قابل استنتایج نمی باشند ولی توسط گوینده برای اصلاح و یا تکمیل اطلاعات نمادین، اضافه می شوند.

non-linguistic information :

در رابطه با سن، جنس، ویژگی شخصی و حالات فیزیکی و احساسی گوینده و... می باشد. اگرچه گوینده برای بیان احساسی خاص ، روش صحبت کردن خود را کنترل می کند، این ویژگی ها قابل کنترل نیستند.

ANN : artificial NEURAL NETWORK

HMM : hidden markove modle

GMM : Gussian mixture modle

KNN:K- nearest neighber

SVM: support vector machine

SVR:support vector regression

PNN : PROBABILISTIC NEURAL NETWORK

DFT : discrite fuorier transform

KLT : Karhunen-Loeve transform

ICA : INDEPENDENT COMPONENT ANALYSIS

EM : EXPECTATION MAXIMIZATION

MAP : MAXIMUM A POSTERIORI TRAINING ALGORITHM

RBF(Radial basic function)

RASTA: Relative Spectral

MAXR : MAX-Relevance (feature selection method)

فصل اول

مقدمه

در تولید گفتار، انواع متعددی از اطلاعات نیز به طور موازی تولید می شوند، هم اطلاعات زبان شناسی و هم تعداد زیادی از صفات شخصی گوینده. که اطلاعات اخیر شامل نشانه هایی است که به ویژگی های فیزیکی دستگاه صوتی گوینده مربوط می شود. بعلاوه تعدادی از ویژگی های غیر فیزیولوژیکی گوینده نیز تاثیر خود را بر سیگنال صوتی می گذارند. مانند اطلاعاتی درباره ی حالت احساسی گوینده، ویژگی های مذهبی و اجتماعی.

تعیین کیفی دقیق این اطلاعات کار دشواری است زیرا اکثر این اطلاعات با مجموعه نشانه های یکسانی علامت گذاری می شوند. مثلا فرکانس اصلی هم به رابطه ی بین محتوی علمی _معنایی عبارت و آهنگ تلفظ بستگی دارد و هم به بسیاری فاکتورهای دیگر مانند جنس و گروه سنی و حالت احساسی گوینده .

امروزه شاهد گسترش استفاده ی افراد مختلف جامعه از کودک و خردسال تا میان سال و کهنسال و همچنین قشرهای گوناگون جامعه کودکان و دانش آموزان، دانشجویان و محققان دانشگاهی، صنعتی و کارکنان ادارات شرکتی، استفاده های خانگی تا مدیریت های شرکتی و کارخانجات بزرگ صنعتی و کشاورزی از کامپیوتر هستیم.

این واقعیات به این معنی است که برای ارتباط سریع تر و آسان تر افراد از کامپیوتر، سیستم های مکالمه ای- گفتاری باید گسترش یابند تا بتوانند با کاربر ارتباط برقرار کرده و راهبرد های مکالمه ی صمیمانه با همه ی نسل ها را بهبود بخشد. اگرچه این کار با به کارگیری یک شیوه ی واحد و عمومی برای همه ی نسل ها امکان پذیر است، ولی کنترل دینامیک، انعطاف پذیر و دقیق بر ارتباط با کاربر و طرح های مکالمه ای، را می توان با تخمین اتوماتیک سن افراد و گوینده ها، محقق ساخت.

بیشتر سیستم های گفتاری – مکالمه ای رایج تنها دنباله ای از کلمات را از صدای گوینده استخراج می کنند. این کار به طور وسیعی از اطلاعات مفید دیگری که می توان از گفتار بدست آورد، چشم پوشی می کند. انسان می تواند یک سری صفات مهم را درباره ی گوینده، مانند جنسیت، سن، لهجه، نژاد، احساسات، سطح تحصیلات و حتی قد و وزن را استنباط کند. این نوع ویژگی های صوت می تواند برای طبیعی ساختن رابط های گفتاری انسان - ماشین، به طور بهینه ای مورد استفاده قرار گیرند. مثلا یک سیستم مشتری - محور با دانستن اینکه یک شخص مسن از قسمت جنوبی کشور که ناراحت و افسرده است، می تواند با ساختن یک دنباله ی مکالمه ای مناسب به لهجه ی جنوبی، مایه ی آرامش کاربر شود.

در ارتباطات گفتاری انسان - با - انسان، مخصوصا در ارتباطات شخص – با – شخص، انتظار می رود که گوینده، شیوه ی سخن گفتن خود را با توجه به خصوصیات و پاسخ های شنونده، تغییر دهد. در ارتباطات بر پایه ی گفتار، به ویژه در اولین ملاقات برای تنظیم رفتار مناسب، یک فاکتور مهم برای هر کسی سن می باشد. گفتار مردان و زنان نه تنها حامل بار معنایی عبارات گفته شده است بلکه شامل ویژگی هایی است که اطلاعاتی غیر زبانی، وابسته به گوینده را تامین می کنند مانند جنس، سن گوینده، حالت احساسی گوینده و ...با استخراج این ویژگی ها برای هر فرد، می توانیم نحوه ی صحبت کردن خود را با طرف مقابل تنظیم کنیم.

یک کاربرد رایج تخمین سن، هر روزه در مکالمات تلفنی صورت می گیرد. هم چنین بعضی شرکت ها نیاز به سیستم تخمین سن اتوماتیک دارند تا مثلا بتوانند برای گروه های سنی مختلف از مشتریان خود، موسیقی های مناسب پخش کنند.

علاوه بر سیستم های مکالمه ای، سن به عنوان یک پارامتر کلاسه بندی مهم در بسیاری دیگر از کاربردها، در نظر گرفته می شود. مثلا اجازه دادن به گروه سنی خاصی برای دسترسی به حقوقی ویژه و یا لحاظ کردن قیمت های متفاوتی در خرید کالاها برای گروه های سنی مختلف.

نماینده های اجرای قوانین برای تشخیص هویت یک فرد، تکنیک های بیومتریک مختلفی را مد نظر قرار می دهند. مشخصه های مختلف بیومتریک، می توانند برای هویت شناسی قانونی مورد استفاده قرار گیرند. مانند الگوهای اثر انگشت، مشخصه های چهره، ترکیب هندسی دست، تغییرات امضا ء و الگوهای صدا . انتخاب یک روش مناسب به اطلاعات موجود و نیز به درجه اطمینان آن روش در یک کاربرد مشخص، بستگی دارد. در بعضی از جُرم ها، مدرک در دسترس ممکن است به شکل مکالمات ضبط شده باشد. الگوهای گفتاری می توانند دارای اطلاعات مهمی برای ماموران قانون باشد. برای مثال، نمونه ی گفتاری یک شخص می تواند اطلاعاتی درباره ی سن، جنس، لهجه، حالات فیزیولوژیکی باشد ویا حتی اینکه آن شخص عضو یک گروه اجتماعی یا مذهبی خاصی باشد. در نتیجه، از گفتار می توان برای شناسایی گوینده که در موارد زیادی مانند آدم دزدی، تماس های تهدید آمیز و ... مورد نیاز است، استفاده کرد.

۱-۱: خلاصه ای از چگونگی تغییر صدای انسان با افزایش سن

هر موجود زنده ای گذر عمر و افزایش سن را تجربه می کند. این یک مکانیزم پیچیده است که از جنبه های مختلفی بر یک شخص تاثیر می گذارد. در نتیجه مورد بررسی قرار دادن مفهوم تغییر سن در اکثر علوم طبیعی و قوانین بشری، گریز ناپذیر است .

بالارفتن سن در شیوه ی سخن گفتن ما تغییراتی ایجاد می کند. صدای ما و الگوهای گفتاری ما از کودکی تا پیری تغییر می کند. اگرچه اکثر تغییرات در کودکی و بلوغ اتفاق می افتد، تغییرات وابسته به سن را می توان از دوران بزرگسالی تا پیری نیز مشاهده کرد. درنتیجه، سن ما در گفتار ما منعکس می شود. بدین ترتیب سن گوینده با به کارگیری روش های متعددی، عمدتا تحلیل صوتی و آزمایشات ادراکی، مورد مطالعه قرار می گیرد.

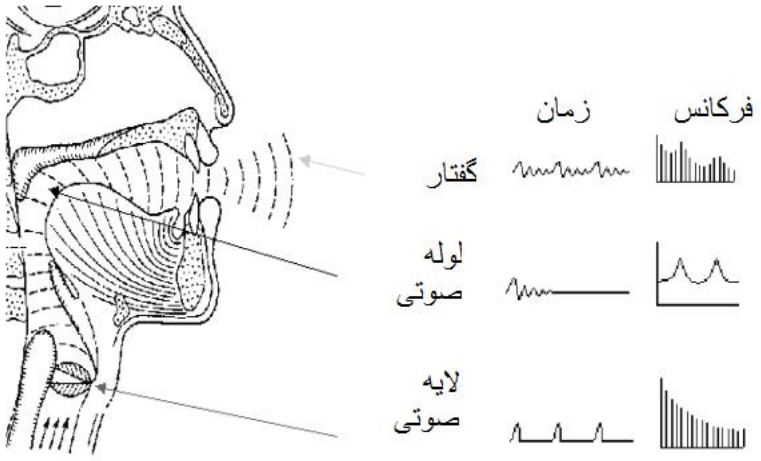
تغییرات وابسته به سن در گفتار بزرگسالان به طور گسترده از ۱۹۶۰ مورد مطالعه قرار گرفته است. اکثر مطالعات جنبه های صوتی و ادراکی را در نظر گرفته اند، اگرچه بعضی روش های تکنولوژی گفتار نیز دنبال شده است. بدلیل پیچیدگی فرایند افزایش سن، در بیشتر تحقیقات نیاز است تا تغییرات وابسته به سن در گفتار به خوبی درک شود. [۱]

با گذشت سن مقداری تغییرات فیزیکی و هورمونی در مرد ها و زن ها رخ می دهد که می تواند بر صدای آنها تاثیر گذارد. برای مثال یک موجی ازهورمون های فعال سازی در زمان بلوغ بوجود می آید که بر انتقال صدا از کودکی به یک فرد بالغ تاثیر می گذارد .

در زمان بلوغ، بر اثرافزایش هورمون تستوسترون، صدای یک مرد با تقریب ۱ octave در pitch ، کاهش می یابد. در خانم ها نیز در زمان بلوغ بر اثر هورمون های استروژن و پروژسترون، صدای آنها کاهش می یابد ولی با درجه ی کمتر تقریبا یک چهارم octave در pitch .

با بالا رفتن سن در مردها، مقدار هورمون تستوسترون کاهش می یابد و میانگین حجم صدا و تغییرپذیری در حجم صدای آنها افزایش می یابد. با بالا رفتن سن در خانم ها، کاهش فعالیت هورمون های جنسی، مخصوصا در دوران یائسگی، بر لایه های صوتی خانمها و عملکرد حنجره تاثیر می گذارد. صدای خانم ها عمیق تر می شود، فرکانس آوائیِ حداکثر کاهش می یابد و بازه ی صوتی گسترده می شود که باعث ادا شدن pitch های پایین تر می شود. تغییرپذیری pitch و jitter بر اثر مرور سن افزایش می یابد. [۲]

۱-۱-۱ مکانیزم تولید گفتار با گذشت سن :



شکل ۱-۱ : تولید گفتار صدا . فشار هوای شش ها تولید ارتعاش لایه های صوتی می کند ، که یک سیگنال منبع صدا ی ضربه ای شبه – متناوب را نتیجه می دهد . سیگنال منبع صدا باعث تحریک لوله ی صوتی ، که مانند یک بدنه ی رزونانسی عمل می کند که فرکانس های معینی را تقویت و یا تضعیف می کند و باعث تولید گفتار می شود .[۳]

مکانیزم تولید صدا در انسان را می توان به ۳ بخش تقسیم کرد: شش ها، لایه های صوتی و لوله ی صوتی. در شکل ۱-۱ مکانیزم صوتی انسان نشان داده شده است. فشار هوای شش ها باعث جریان یافتن هوا از بین حنجره، که فضای بین لایه های صوتی است، می شود . لایه های صوتی دو توده از گوشت، رباط و ماهیچه هستند که بین جلو و عقب حنجره کشیده شده و بسته به رانش و ربایش لایه های صوتی، در حالت های مختلف ارتعاشی (صوت های صدا دار) قرار می گیرند و یا اینکه اصلا ارتعاشی صورت نمی گیرد (صوت های بی صدا).

برای صوت های صدادار، لایه های صوتی به صورت شبه تناوبی باز و بسته می شوند و در نتیجه جریان هوای حنجره ای را به جریان ضربه هایی تبدیل می کند که سیگنال منبع صدا نامیده می شود. سپس سیگنال منبع صدا از لوله ی صوتی، که از حنجره آغاز و به لب ها ختم می شود، عبور می کند. لوله ی صوتی مانند یک بدنه با رزونانس ها (فرکانس های فرمنت) و ضد رزونانس ها (صفر ها) عمل می کند. لوله ی صوتی به عنوان فیلتر صوتی که طیف صوت را شکل می دهد، انجام وظیفه می کند. صداهای متنوع صوت با تنظیم شکل لوله ی صوتی و هم چنین سیگنال منبع صدا، تولید می شوند.

گفتار براثر ارتعاش تارهای صوتی هنگامی که نفس از شش ها بیرون داده می شوند، تولید می شود. تنوع صداهای گفتار عمدتا به دلیل تنوع در جرم و طول تارهای صوتی و بندهای موجود در لوله ی صوتی می باشد.