
بسم الله الرحمن الرحيم

دانشکده برق و رباتیک

گروه الکترونیک

پایان نامه کارشناسی ارشد

تشخیص گوینده در محیط شامل چند گوینده با استفاده از ماشین بردار پشتیبان

دانشجو: مرضیه لشکربلوکی

استاد راهنما:

دکتر حسین مروی

استاد مشاور:

دکتر حسین صامتی

تیر ماه 1390

تقدیم به:

خانواده عزیزم که همواره نقش اساسی در موفقیت هایم داشته اند.

دوستان عزیزی که همواره همراه و مشوقم بوده اند.

تقدیر و تشکر

سپاس و تشکر از زحمات و راهنمایی های علمی استاد گرانقدر جناب آقای دکتر حسین مروی (استاد راهنمای اینجانب) و جناب آقای دکتر حسین صامتی (استاد مشاور اینجانب). برای ایشان همواره آرزوی سلامتی و موفقیت روز افزون دارم.

چکیده:

شناسایی گوینده یکی از مباحث مطرح در بحث پردازش گفتار می باشد. شناسایی گوینده عبارت است از فرآیندی که طی آن با استفاده از سیگنال صحبت تشخیص دهیم چه کسی چه موقع واقعا صحبت می کند. هدف طراحی سیستمی است که بتواند تغییر در گوینده را مشخص نماید و گفتار هرگوینده را برای سیستم برچسب گذاری نماید. یعنی مشخص نماید که کدام گوینده، در چه بازه هایی صحبت کرده است. امروزه این عمل با یک عنوان جدید که هر دو فرآیند جداسازی و برچسب گذاری را در بر می گیرد بنام Speaker Diarization مشهور گشته است. هدف از بخش بندی تقسیم سیگنال گفتاری به بخش هایی است که تنها شامل گفتار یک گوینده هستند و هدف از خوشه بندی نیز شناسایی بخش های گفتاری مربوط به یک گوینده و اختصاص یک برچسب واحد به آنهاست.

هدف از انجام این پایان نامه طراحی و پیاده سازی یک سیستم بخش بندی و خوشه بندی گوینده با استفاده از الگوریتم های جدید و همچنین بهبود نتایج این الگوریتم ها برای این موضوع می باشد. این سیستم باید بطور صحیح نقاط تغییر گوینده را بدون دانستن اطلاعات قبلی از گوینده تشخیص داده و در نهایت تمام قسمت های صوتی مربوط به یک گوینده را در یک خوشه قرار می دهد.

در این پایان نامه، سیستم تشخیص گوینده، از سه مرحله اصلی تشکیل شده است. درمرحله اول قسمت-های غیر گفتاری، از بخش های گفتاری فایل صوتی حذف می شوند، تا دقت و سرعت عملیات سیستم در مراحل بعدی افزایش پیدا کند. سپس فایل گفتاری به بخش هایی همگن که در آن فقط گفتار یک گوینده وجود دارد، تقسیم می شود. در مرحله سوم با استفاده از خوشه بندی مناسب، بخش های گفتاری مرحله قبل، که متعلق به یک گوینده هستند، در یک خوشه جای می گیرند. جهت پیاده سازی سیستم از چهار نوع بردار ویژگی TDC, root-MFCC, MFCC و root-TDC و سه نوع پایگاه داده استفاده شده است و دقت مرحله بخش بندی 80٪ بوده است و دقت مرحله خوشه بندی نیز 59٪ با استفاده از ماشین بردار پشتیبان بدست آمده است.

کلمات کلیدی:

بخش بندی آماری گوینده

بخش بندی گویندگان

تشخیص بخش های صوتی

خوشه بندی گویندگان

فهرست مطالب

فصل اول: معرفی سیستم های تشخیص گوینده

فصل دوم: تشخیص گفتار از نواحی غیرگفتاری

فصل سوم: آشکارسازی تغییر گوینده

فصل چهارم: روش‌های دسته‌بندی

فصل پنجم: پیاده سازی و مشاهدات سیستم ترکیبی پیشنهادی

فهرست شکل ها

فهرست جداول

فصل اول :

معرفی سیستم های

تشخیص گوینده

1-1-مقدمه

امروزه داده های چند رسانه ای بخش قابل توجهی از دانش انسان را در بر می گیرند. حجم پرونده های چند رسانه ای آرشیو شده در موسسه های مختلف در سال های اخیر افزایش چشمگیری داشته است. دسترسی و وضوح بالای این پرونده ها می تواند کمک شایانی به افرادی کند که در جستجوی اطلاعات باشند. بنابراین عملیات جستجو و بازیابی اطلاعات در این حجم بالا کاری است که خود احتیاج به سیستم کامپیوتری دارد. و درنتیجه یکی از حوزه های تحقیقاتی که به تازگی مورد توجه قرار گرفته است، مربوط به ساختاربندی پرونده- های چند رسانه ای است. در میان این داده ها، اطلاعات صوتی اهمیت بالاتری دارد. زیرا بخش اعظم آرشیوها حاوی داده های صوتی از گزارش های تلویزیونی، رادیویی و همچنین مکالمات تلفنی می باشد. در سالهای اخیر تحقیقات وسیعی در این حوزه آغاز شده و نتایج قابل قبولی نیز حاصل شده است. از دیگر کاربردهای این حوزه در تشخیص مجرم، جدا کردن صحبت های مهم یک شاهد یا متهم در دادگاه و ... میتوان اشاره نمود.

در کاربرد صوتی، عمده اطلاعات موجود در پرونده ها، صحبت های تعدادی گوینده است و هدف از سیستم نهایی، پاسخ به این سوال است که چه کسی در چه زمانهایی صحبت کرده است؟ بخش های مختلف این حوزه تحقیقاتی به نامهای مختلفی مانند: قطعه بند گوینده ای[1]، تشخیص گوینده[2] ،رونویسی قوی[3]، و اندیس گذاری گوینده ای[4] نامیده شده اند. از چنین سیستم هایی برای جابجایی راحت در داده های صوتی، در فایل های صوتی طولانی (مانند: اخبار و ملاقات ها و جلسات یک شرکت و ...) که متعلق به چند گوینده باشند بهره- برداری می شود. مکالمات و محاسبات رادیویی طولانی از محیط هایی هستند که در آنها چند گوینده حضور داشته و با هم صحبت می کنند. هدف نهایی چنین سیستم هایی، پیاده سازی روش هایی مناسب برای افراز پرونده صوتی به نواحی است که در آنها گوینده ای خاص صحبت کرده باشد. دسترسی راحت به

1.Speaker Segmentation
2.Speaker Diarization
3.Rich Transcription
4.Speaker Indexing

بخش هایی از صحبت یک گوینده توسط این سیستم فراهم می گردد. با داشتن حجم بالایی از داده های صوتی اهمیت این سیستم ها بیشتر می گردد.

با افزایش تعداد مدارک متنی موجود در اینترنت، نیاز به تکنیک هایی نظیر فهرست نگاری متن به منظور تسهیل دسترسی و جستجو در این مدارک افزایش پیدا کرد. نظیر همین نیاز نیز با افزایش تعداد مدارک صوتی نظیر سخنرانی ها، مصاحبه ها و گردهمایی ها و ... ایجاد شد. بطور مشخص دسترسی به مدارک صوتی بسیار سخت تر از دسترسی به متن است و گوش دادن به یک فایل صوتی ضبط شده بیشتر از خواندن متن زمان بر است و فهرست نگاری دستی مدارک صوتی در مقایسه با فهرست نگاری متن، مشکل است. راه حل پیشنهادی جهت رفع این مشکل، فهرست نگاری خودکار مدارک صوتی[5] است.

اولین بار سیستم هایی تشخیص گوینده توسط کمپانیNIST در سال 1999 ارایه شد. در سال2001، پلکان و سیدهارون به همراه گروهشان با استفاده از کم کردن اثر نویز بر روی سیگنال بهبودهایی در نتایج سیستم دادند و جداسازی بهتر گویندگان را باعث شدند. در سال 2005، بولیان و کنی با بکارگیری بردارهای ویژگی دیگر (یا ادغام روش های قبلی) و استفاده از مدل های گوسی در سیستم نتایج متفاوتی بدست آوردند. در سال 2005 توسط یاماشیتا و ماتسوناگا با استفاده از ویژگی های سیگنال صوتی مانند فرکانس پیچ سیگنال، انرژی، فرکانس های ماکزیمم سیگنال، و سه ویژگی دیگر نتایج در قسمت بخش بندی گوینده این سیستم بهبود داده شد.[1] و در سال های بعدی با انجام روش های مختلف برروی قسمت های متفاوت آن تا به امروز این سیستم ها در حال تکمیل شدن و بهتر شدن نتایج بوده اند.

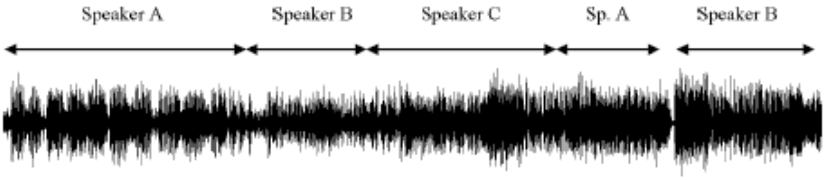
هدف از این پایان نامه، طراحی و پیاده سازی سیستمی است که بتواند در یک فایل صوتی که شامل گفتار چندین گوینده می باشد، تغییر در گوینده را مشخص نماید و تا حد امکان، گفتار هر گوینده را بدون دانستن اطلاعات قبلی از وی، دسته بندی نماید. این سیستم می تواند شامل دو بخش اساسی باشد که عبارتند از:

-بخش بندی گوینده

[1] .Automatic Audio Indexing

-خوشه بندی گوینده

کار قسمت بخش بندی[6]، تقسیم سیگنال گفتاری به سگمنت هایی است که تنها شامل گفتار یک گوینده هستند. در مرحله خوشه بندی[7]، شناسایی و دسته بندی بخش های گفتاری مربوط به یک گوینده و اختصاص یک برچسب واحد به آن انجام می شود. این مطلب در بسیاری از کاربردهای گفتاری که مربوط به بازشناسی یا فهرست نگاری[8] گفتار در محیطی که چندین گوینده ممکن است در آن اقدام به سخن گفتن بنمایند، مانند یک جلسه، کنفرانس، اخبار و نظایر آن کاربرد دارد. این کار نه تنها می تواند به سیستم های بازشناسی گفتار پیشرفته جهت بهبود نتایج بازشناسی گروهی کمک نماید بلکه در شناسایی و متن نگاری مکالمه ها نیز به آنها کمک می نماید. همانطور که قبلا نیز ذکر شد، امکان استفاده از آن در فهرست نگاری صوتی که امکان جستجو در فایل های صوتی را فراهم می نماید نیز ممکن است. شکل (1-1) نحوه کار این سیستم را بخوبی نشان می-دهد.



شکل (1-1): نمایش بخش بندی گویندگان روی گفتار ورودی

فایل صوتی مورد بررسی یک صوت ضبط شده تک کاناله است که شامل چندین منبع صوتی است. این منابع صوتی متفاوتند و می توانند شامل چند گوینده، موسیقی، انواع نویز و ... باشند. نوع و جزییات منابع صوتی موجود در فایل به ویژگی کاربردی آن فایل بستگی دارد.

بطور کلی سیستم های بخش بندی و خوشه بندی گوینده در سه حوزه زیر دارای کاربرد می باشند:

- دادگان اخباری
- جلسات ضبط شده
- مکالمات تلفنی

---

[6] .Segmentation
[7] .Clustering
[8] .Indexing

همانطور که قبلا نیز اشاره شد این سه حوزه تفاوت هایی مانند کیفیت ضبط صوت (پهنای باند، میکروفون ها و نویز) و میزان و نوع منابع غیرگفتاری، تعداد گویندگان، سبک و ساختار گفتار (طول مدت گفتار، ترتیب گویندگان) دارند و هر حوزه جهت کار بخش بندی و خوشه بندی گوینده، مسائل و مشکلات خاص خود را دارد. البته در سیستم های تشخیص گوینده سعی بر آن است تا برای هر سه حوزه کاری، نتایج قابل قبول و مناسبی حاصل شود.[1]

در سطح پایین تر کار چنین سیستمی دسته بندی داده های صوتی در خوشه هایی است که هر یک متعلق به یک گوینده باشد. در همین جا به راحتی میتوان دید که دو دیدگاه ناظرانه[9] (با سرپرست) و غیر ناظرانه[10] (بدون سرپرست) در این بخش مشاهده می شود. در دیدگاه اول از پیش اطلاعاتی از اینکه چه کسانی در فایل صوتی صحبت می کنند، وجود دارد. ولی در دیدگاه دوم کار سیستم دسته بندی فایل به بازه های زمانی است که در آنها تنها یک گوینده که هویت آن بر ما پوشیده است، صحبت می کند. توجه شود که میتوان از خروجی یک دسته بند غیرناظرانه به عنوان ورودی سیستم های شناسایی[11]، استفاده کرد و به این ترتیب یک سیستم دسته بندی ناظرانه خواهیم داشت. بنابراین کارایی و همچنین زمان اجرای سیستم ناظرانه بدست آمده بهتر است. از سوی دیگر، عملکرد این سیستم ها، به میزان اطلاعات قبلی مجاز نیز بستگی دارد. این اطلاعات قبلی می تواند نمونه گفتار از گویندگان، تعداد گویندگان موجود در فایل صوتی، یا اطلاعاتی از ساختار فایل ضبط شده باشد. ولی در اکثر سیستم های بخش بندی و خوشه بندی گوینده فرض بر نبود هیچگونه اطلاعات قبلی راجع به گویندگان و تعداد آنهاست. در این پروژه نیز با روش های بکار گرفته شده، فرض بر اینست که هیچگونه اطلاعات قبلی از گویندگان، مانند تعداد آنها، هویت آنها و داده آموزشی موجود نمی باشد و بنابراین مدل های گویندگان را نمیتوان از قبل آماده کرد. شکل (1-2) ساختار کلی سیستم های بخش بندی و خوشه بندی گوینده را نشان می دهد.

1.Supervised
2.Unsupervised
3.Identification

شکل (1-2): ساختار کلی سیستم های بخش بندی و خوشه بندی گوینده

چنین سیستمی شامل مراحل کاری مختلفی است و میتوان بخش های ذکر شده در قسمت های بعدی را برای آنها در نظر گرفت.[6-5]

1-2-مراحل مختلف کاری سیستم های بازشناسی گوینده

بطور کلی مراحل مختلف یک سیستم بازشناسی گوینده، بصورت زیر خلاصه می گردد:

1-قطعه بندی آکوستیکی[12]

2-تشخیص گفتار از غیر گفتار[13]

3-تشخیص جنسیت گوینده

4-تشخیص تغییر گوینده

5-جمع زدن گوینده های مشابه

1. Acoustic Segmentation Module
2.Speech Detection

این سیستم دارای بلوک های کاری مستقل از هم می باشد که هر بلوک ورودی خود را از خروجی بلوک قبلی دریافت می کند و ورودی لازم برای بلوک کاری پس از خود را تهیه می کند. در برخی سیستم ها، از بلوک سوم کاری صرف نظر می شود. در ادامه شرح مختصری از بخش های مختلف داده شده است.[4-2]

1-2-1-قطعه بند آکوستیکی

در اولین مرحله، باید جریان داده های صوتی به قطعات همگن آکوستیکی تقسیم شود. برای این امر باید نقاطی که تغییر در خواص آکوستیکی داده های صوتی روی میدهد را، بدست آورد. در واقع این نقاط شکست[14] بعنوان ورودی به بلوک کاری بعدی داده می شود. در بسیاری از کاربردهای چند رسانه ای که داده ها علاوه بر صدا دارای تصویر نیز می باشند، عمل تشخیص نقاط تغییر، هم از روی صدا و هم از روی تصویر امکان پذیراست.[2] بنابراین کارایی چنین سیستم هایی نسبت به داده هایی که تنها شامل صوت یا تصویر هستند، بالاتر خواهد بود.

امروزه روش های کاربردی تعیین نقاط تغییر آکوستیکی، همگی بر پایه ی محاسبه فاصله آماری بین دو قطعه مجاور استوار هستند. تفاوت عمده ی میان آنها معیار فاصله ای است که در آنها بکار می رود. از روش های غیر آماری مورد استفاده میتوان به شبکه عصبی[15] و ماشین بردار پشتیبان[16] اشاره نمود، که در بخش های بعدی توضیح داده خواهند شد.

از دیدگاهی قطعه بندی، یک مساله بهینه سازی[17] است. زیرا هدف نهایی یافتن نقاطی است که در آنها معیار فاصله به ماکزیمم محلی[18] برسد. یکی از پرکاربردترین معیارهایی که امروزه برای تعیین نقاط شکست آکوستیکی بکار می رود، معیار بیزین[19] است. پیش از این، روش های آماری دیگری از سال 1997 ابداع شده بود، که همگی آنها در مقایسه با معیار بیز جواب مناسبی نمی داده اند.[1] ارایه این روش اعتبار روش های دیگر را تا حدودی کمتر نمود.

1.Break Point
2.Artificial Neural Network
3.Support Vector Machine
4.Optimization
5.Local Maximum
1.Bayesian Information Criterion

1-2-2-تشخیص گفتار از غیر گفتار(دسته بندی[20] صوتی)

برای پیاده سازی این سیستم ها، قبل از هر کار دیگری بخش های گفتاری صوت ضبط شده را از بخش های غیر گفتاری آن مانند (سکوت، موسیقی، نویز خیابان، صدای سرفه ، صدای ورق زدن و ...) جدا می نمایند. با حذف بخش های غیرگفتاری میزان بار محاسباتی سیستم کاهش پیدا می کند و سرعت سیستم بیشتر می شود و سپس مراحل بخش بندی و خوشه بندی اجرا می شود. بعد از یافتن نقاط تغییر آکوستیکی، میتوان جریان داده های صوتی را مانند قطعات همگن در نظر گرفت. به عبارت دیگر یک قطعه نباید هم شامل گفتار، هم موسیقی و سکوت با هم باشد. اگر یک قطعه شامل گفتار دو گوینده باشد، باز هم همگن نخواهد بود. بنابراین این بلوک کاری خروجی قطعه بند صوتی را دریافت کرده و از آن قطعاتی را که حاوی داده های صوتی غیرگفتاری اند را حذف می کند. در یک سیستم تشخیص گفتار، معمولا داده های صوتی به 5 کلاس[2] زیر تقسیم می شوند:

1-موسیقی خالص

2-گفتار خالص

3-گفتار همراه با نویز

4-سکوت

5-سکوت همراه با نویز

البته در یک سیستم تشخیص گوینده، تنها احتیاج به تشخیص موارد 2و3 وجود دارد. زیرا هدف سیستم کار با گفتار بوده و هر چیزی غیر از گفتار از جریان داده ی صوتی حذف می شود تا بلوک های کاری پس از این بلوک با تمرکز برروی گفتار عمل نمایند. روشی که برای رسیدن به هدف این سیستم وجود دارد، بیشترین میزان شباهت(ML)[21] مبتنی بر مدل مخلوط گوسی(GMM)[22] می باشد.

2.Classification

[1].Maximum Likelihood.

[2]. Gaussian Mixture Model