
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

گرایش هوش مصنوعی

عنوان

شناسایی مستقل از متن گوینده از بین گویندگان متعدد

نگارش

حسین زینلی

استاد راهنما

دکتر حسین صامتی

شهریور ۱۳۹۱

به نام خدا

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

رساله کارشناسی ارشد

عنوان: شناسایی مستقل از متن گوینده از بین گویندگان متعدد

نگارش: حسین زینلی

کمیته ممتحنین:

استاد راهنما: دکتر حسین صامتی **امضاء: .....**

استاد مشاور: دکتر علی محمد افشین همت‌یار **امضاء: .....**

استاد مدعو: دکتر محمد مهدی همایون‌پور **امضاء: .....**

تاریخ: شهریور ۱۳۹۱

تقدیم به پدر زحمت کش، مادر فداکار و همسر مهربانم

سپاس گزاری

گرچه کلام در سپاس از آنان که مشوق و راهنمایم بوده اند ناتوان است، در اینجا فرصت را غنیمت می شمارم تا از تمامی آنان بخصوص آقای خرّم به خاطر گفتگوهای متعدد و راهنمایی های ارزشمندشان، آقای دکتر باباعلی و دکتر ویسی به خاطر مشورت-های سودمند، شرکت عصر گویش پرداز به خاطر در اختیار قرار دادن امکانات، مهم تر از همه استاد گرامی آقای دکتر صامتی به خاطر راهنمایی ها و کمک های بی دریغش و سایر دوستانی که بنده را در انجام این پایان نامه کمک کرده اند کمال تشکر را بنمایم. علاوه بر اینها، به خاطر کمک های فراوان همسر مهربانم و همچنین تحمل سختی های این مدت از ایشان نیز کمال سپاس گزاری را دارم. در پایان از اساتید محترمی که قبول زحمت کردند و در جلسه ی دفاع اینجانب شرکت کردند نیز سپاس گزارم.

چکیده

گفتار انسان حاوی اطلاعات زیادی از قبیل محتوای معنایی، احساسات و حتی هویت گوینده می‌باشد. هدف ما در این پایان‌نامه شناسایی گوینده مستقل از متن از بین گویندگان متعدد است. در سال‌های اخیر، زمان شناسایی (آزمایش) به یکی از مهمترین چالش‌های سامانه‌های بلادرنگ شناسایی گوینده تبدیل شده است. زمان شناسایی به روش محاسبه‌ی درستنمایی گفتار آزمایش در مدل گویندگان ثبت-نامی وابسته است. برای کاربردهای بلادرنگ، این سامانه‌ها باید خیلی سریع گوینده‌ی هدف را مشخص کنند و به همین دلیل نمی‌توان از روش‌های مرسوم شناسایی گوینده استفاده کرد. هدف اصلی این پایان-نامه ارائه روش‌هایی است که بتواند زمان شناسایی را کاهش دهد به طوری که روی دقت شناسایی تأثیر منفی نگذارد.

ابتدا، ساختار کلی سامانه‌های شناسایی گوینده و مراحل مختلف آن را معرفی کرده و روش مدل مخلوط گاوسی را به عنوان روش پایه برای این کار معرفی می‌کنیم. سپس روش‌های متداول مورد استفاده برای کاهش زمان شناسایی را مورد بررسی قرار می‌دهیم. در ادامه روشی دومرحله‌ای ارائه می‌کنیم که در مرحله اول از روش نمایش تنک برای شناسایی گوینده استفاده می‌کند. آزمایش‌های ما بر روی دادگان تیمیت نشان می‌دهند که این روش بدون کاهش دقت، سرعت شناسایی را تا ۱۸ برابر افزایش می‌دهد. علاوه بر این روش، روش دومرحله‌ای دیگری پیشنهاد شده است که در مرحله اول از فاصله بین مدل‌های مختلف برای شناسایی استفاده می‌کند. نتایج بر روی همان دادگان نشان می‌دهد که این روش نیز بدون کاهش دقت، سرعت شناسایی را ۲۸ برابر افزایش می‌دهد. در نهایت روشی دو مرحله‌ای بر مبنای فاصله نزدیکترین همسایه ارائه شده که قادر است سرعت را تا ۳/۴ برابر افزایش دهد.

واژه‌های کلیدی:

شناسایی گوینده، مستقل از متن، گویندگان متعدد، سامانه‌های بلادرنگ، زمان شناسایی، مدل مخلوط گاوسی، روش دومرحله‌ای، نمایش تنک، نزدیکترین همسایه.

فهرست مطالب

عنوان **صفحه**

فهرست اشکال

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فصل ۱

مقدمه

۱-۱ انگیزه

در سال‌های اخیر با پیشرفت علم تغییرات چشم‌گیری در تکنولوژی‌های مختلف بوجود آمده است. یکی از این تکنولوژی‌ها شناسایی افراد با استفاده از زیست‌سنجی[۱] است. زیست‌سنجی، علم شناختن افراد با استفاده از خصوصیات فیزیکی و ویژگی‌های رفتاری آنها است. نیاز به امنیت باعث شده که زمینه‌ی زیست‌سنجی در سال‌های اخیر پیشرفت سریعی داشته باشد. از روش‌های متداول زیست‌سنجی می‌توان به شناسایی با استفاده از اثر انگشت[۲]، صورت[۳]، عنبیه[۴]، شبکیه[۵] و گفتار[۶] اشاره کرد. در جدول ۱-۱ مقایسه‌ی این روش‌ها نشان داده شده است [de Luis-Garcia_03]. این مقایسه بر مبنای ویژگی‌های دقت یا قابلیت اطمینان، سادگی استفاده، سادگی پیاده‌سازی، کاربر پسندی و هزینه روش، انجام شده است.

[1] Biometric
[2] Finger-Print
[3] Face
[4] Iris
[5] Retina
[6] Voice

جدول ۱-۱ مقایسه روش‌های مختلف زیست‌سنجی

ویژگی / روش	دقت	سادگی استفاده	کاربر پسند بودن	سادگی پیاده‌سازی	هزینه
اثر انگشت	زیاد	متوسط	کم	زیاد	متوسط
صورت	کم	زیاد	زیاد	متوسط	کم
عنبیه	متوسط	متوسط	متوسط	متوسط	بالا
شبکیه	زیاد	کم	کم	کم	متوسط
گفتار	متوسط	زیاد	زیاد	زیاد	کم

همان طور که از جدول بالا مشخص است استفاده از گفتار یکی از روش‌های مناسب زیست‌سنجی است که علاوه بر کاربرپسند بودن و سادگی پیاده‌سازی بالا، هزینه پایینی دارد. استفاده از گفتار برای شناسایی افراد در مقایسه با روش‌های دیگر مزایایی دارد که از آنها می‌توان به موارد زیر اشاره کرد:

- گفتار همیشه با انسان است و در زمان نیاز می‌توان از آن استفاده کرد.
- گفتار قابل سرقت نیست و به همین دلیل گزینه مناسبی می‌باشد.
- گفتار نسبت به اثر انگشت کمتر در سوانح آسیب می‌بیند.
- شناسایی با استفاده از گفتار نیاز به تجهیزات و سنسورهای پیچیده ندارد.
- به راحتی در کاربردهای راه دور مثل پشت تلفن قابل استفاده می‌باشد.

با توجه به این ویژگی‌ها، در سال‌های اخیر توجه زیادی به این روش شده است. در کاربردهای تلفنی که قابلیت استفاده از روش‌های دیگر بویژه اثر انگشت نیست، این روش بیشتر مورد توجه قرار گرفته است. با وجود مزایای زیادی که استفاده از گفتار دارد، این روش دارای مشکلات و محدودیت‌هایی است. از این محدودیت‌ها می‌توان به کاهش دقت این روش در وجود نویزهای مختلف به خصوص نویز کانال اشاره کرد. علاوه بر این مشکل می‌توان به زمان‌بر بودن فرایند شناسایی وقتی که تعداد گویندگان موجود در سامانه زیاد باشند و در نتیجه عدم استفاده از این روش در سامانه‌های بلادرنگ اشاره کرد. مشکل دیگری که تمام سامانه‌های شناسایی با آن روبه‌رو هستند مشکل جعل و سوء استفاده است. در

سامانه‌های بازشناسی گوینده نیز این مشکل وجود دارد. امکان جعل صدا توسط انسان و ماشین امنیت سامانه‌های بازشناسی گفتار را تحت تأثیر قرار داده است.

۱-۲ تعریف مسئله

شناسایی گوینده از بین گویندگان متعدد را از دو منظر می‌توان مورد بررسی قرار داد. اول اینکه در سامانه‌های شناسایی گوینده با افزایش تعداد گویندگان ثبت‌نامی، احتمال اشتباه شناسایی کردن یک گوینده بیشتر شده و در نتیجه دقت شناسایی کاهش پیدا می‌کند. یکی از دلایل این امر این است که با افزایش تعداد گویندگان، احتمال وجود گویندگانی که ویژگی‌های صوتی شبیه هم داشته باشند بیشتر می‌شود. این شباهت، احتمال اشتباه شناسایی کردن سامانه را بیشتر می‌کند.

دومین منظر این است که در این سامانه‌ها با افزایش تعداد گویندگان ثبت‌نامی، تعداد درست-نمایی‌هایی که در مرحله‌ی شناسایی باید محاسبه شوند بیشتر شده و در نتیجه زمان پاسخ‌گویی افزایش می‌یابد. در روش‌های مرسوم شناسایی گوینده نیاز است که یک جستجوی خطی روی تمام گویندگان ثبت‌نامی انجام شود، که با افزایش تعداد گویندگان ثبت‌نامی این جستجو طولانی‌تر می‌شود. با وجود این مشکل به نظر می‌رسد که در کاربردهای بلادرنگ نمی‌توان از این سامانه‌ها استفاده کرد. در این پایان‌نامه تمرکز ما روی کاهش زمان شناسایی است و همچنین بدلیل اینکه دقت شناسایی مهمترین فاکتور سامانه‌های شناسایی گوینده است، در روش‌های ارائه شده سعی بر این است که افزایش سرعت باعث کاهش دقت نشود.

۱-۳ رئوس مطالب

در فصل ۲، یک بررسی اجمالی بر سامانه‌ی شناسایی گوینده انجام می‌دهیم و دسته‌بندی‌های مختلف شناسایی گوینده را معرفی می‌کنیم. سپس اجزاء و مراحل مختلف سامانه‌های شناسایی گوینده را مورد بررسی قرار می‌دهیم. در ادامه‌ی این فصل، ویژگی‌های مختلف را معرفی کرده و مراحل استخراج ضرایب

کپسترال مقیاس فرکانس مل را به تفصیل شرح می‌دهیم. سپس مدل‌های مختلف برای مدل‌سازی فضای ویژگی را معرفی می‌کنیم. مدل مخلوط گاوسی را به عنوان روش اصلی مدل‌سازی، شرح داده و روش تخمین پارامترهای آن را نیز توضیح می‌دهیم. علاوه بر این روش، دو روش تخمین بیشینه‌ی پسین و درون‌یابی خطی بیشینه‌ی درستنمایی را به عنوان روش‌های موفق برای تطبیق پارامترهای مدل مخلوط گاوسی شرح می‌دهیم. سپس در فصل ۳ روش‌های متداول برای کاهش زمان مرحله‌ی شناسایی را معرفی کرده و تعدادی از آنها را شرح می‌دهیم. در فصل ۴ روش‌های پیشنهادی برای افزایش سرعت را توضیح داده و سپس نتایج این روش‌ها را در فصل ۵ می‌آوریم. در پایان در فصل ۶ نتایج این پایان‌نامه را به صورت خلاصه آورده و زمینه‌هایی برای کار و پژوهش بیشتر در این موضوع را بیان می‌کنیم.

فصل ۲

مروری بر روش‌های شناسایی گوینده

۲-۱ مقدمه

همان‌طور که در فصل قبل شرح داده شد، بازشناسی گوینده یکی از روش‌های زیست‌سنجی است که به دلیل مزایای متعددی که دارد در سال‌های اخیر مورد توجه قرار گرفته است. در این فصل ابتدا دسته-بندی‌های مختلف شناسایی گوینده را معرفی کرده و اجزاء و مراحل سامانه‌های شناسایی گوینده را شرح می‌دهیم. ویژگی‌های مختلف مورد استفاده برای شناسایی گوینده را بیان کرده و ضرایب کپسترال مقیاس فرکانس مل را به تفصیل شرح می‌دهیم. در ادامه روش‌های مختلف مدل‌سازی گویندگان را معرفی و مدل مخلوط گاوسی را به عنوان مدل اصلی استفاده شده، توضیح می‌دهیم.

۲-۲ دسته‌بندی شناسایی گوینده

شناسایی گوینده را می‌توان از جهت‌های مختلفی دسته‌بندی کرد که یکی از این جهت‌ها اهمیت متنی است که گوینده بیان می‌کند. از این جهت شناسایی گوینده به دو دسته‌ی وابسته به متن[۱] و مستقل از متن[۲] تقسیم می‌شود [Beigi_11].

[1] Text-Dependent
[2] Text-Independent

۲-۲-۱ شناسایی گوینده‌ی وابسته به متن

در شناسایی گوینده‌ی وابسته به متن چیزی که گوینده بیان کرده مهم است. در این دسته در مرحله‌ی شناسایی، گوینده باید همان متنی که در مرحله‌ی ثبت‌نام گفته را تکرار کرده باشد. این روش بیشتر برای تصدیق هویت گوینده کاربرد دارد و در عمل برای شناسایی گوینده قابل استفاده نیست. مزیت اصلی سامانه‌هایی که از این روش استفاده می‌کنند، گفتار کم برای آموزش و آزمایش و همچنین دقت بالاتر آنها است.

۲-۲-۲ شناسایی گوینده‌ی مستقل از متن

در این دسته اینکه گوینده چه بیان کرده است و یا اینکه در مرحله‌ی شناسایی متنی متفاوت از متن مرحله‌ی ثبت‌نام بیان کرده باشد، مهم نیست. مستقل از متن بودن چندین سطح دارد. بیشترین استقلال حالتی است که سامانه هم مستقل از متن بیان شده و هم مستقل از زبان[1] بیان متن باشد. حالت دیگر این است که سامانه، مستقل از متن اما وابسته به زبان باشد. در حالت مستقل از متن و زبان، سامانه باید کار شناسایی را فقط با استفاده از خصوصیات مسیر صوتی گوینده انجام دهد و نباید هیچ فرضی در رابطه با محتوای بیان شده بکند [Beigi_11]. سامانه‌های مستقل از متن نسبت به سامانه‌های وابسته به متن گفتار بیشتری هم برای آموزش و هم برای شناسایی نیاز دارند. به طور کلی این محدودیت باعث شده است که سامانه‌های مستقل از متن نسبت به وابسته به متن دقت کمتری داشته باشند.

علاوه بر این دسته‌بندی، شناسایی گوینده با توجه به نوع مجموعه‌ای که در آن عمل شناسایی انجام می‌شود، به دو دسته‌ی مجموعه‌ی بسته[2] و مجموعه‌ی باز[3] تقسیم می‌شود [Beigi_11]:

[1] Language-Independent
[2] Close-Set
[3] Open-Set

۲-۲-۳ شناسایی گوینده‌ی مجموعه‌ی بسته

در شناسایی گوینده‌ی مجموعه‌ی بسته فرض می‌کنیم که گفتار آزمایش حتماً متعلق به یکی از گویندگانی است که در سامانه ثبت شده‌اند. در این حالت گفتار آزمایش با مدل تمام گویندگان ثبت‌نامی مقایسه می‌شود و گوینده‌ای که مدل آن بیشترین شباهت را با گفتار آزمایش داشته باشد به عنوان گوینده‌ی هدف انتخاب می‌شود. توجه داشته باشید که در این حالت همیشه یک گوینده به عنوان گوینده‌ی هدف انتخاب می‌شود.

شناسایی گوینده مجموعه‌ی بسته بجز کاربردهای خاص، در عمل کمتر مورد استفاده قرار می-گیرد. به عنوان مثالی از کاربرد این دسته، محیطی را در نظر بگیرید که برای کاربران مختلف شخصی سازی شده است. در این محیط، سامانه‌ای به صورت خودکار با استفاده از صدای فرد تنظیمات شخصی آن را بکار می‌برد. در این مثال اشتباه تشخیص دادن گوینده جریمه‌ی زیادی ندارد. در حقیقت اگر گوینده در مجموعه‌ی گویندگان ثبت‌نام شده وجود نداشته باشد، فرقی نمی‌کند که تنظیمات چه گوینده‌ای انتخاب شده است.

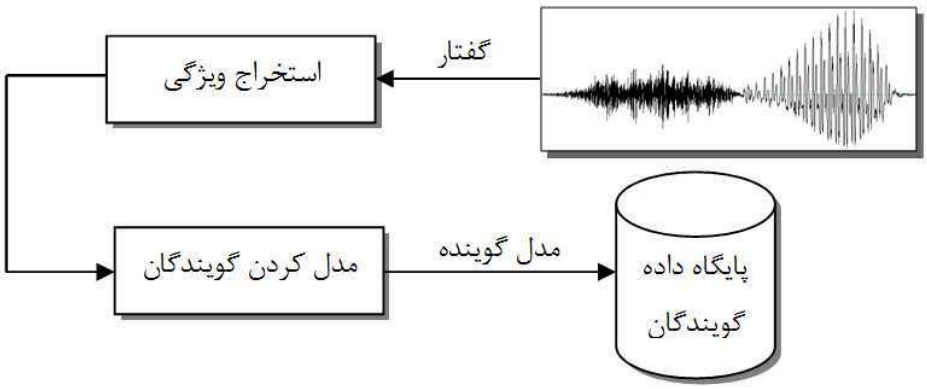
۲-۲-۴ شناسایی گوینده‌ی مجموعه‌ی باز

در این حالت بر خلاف حالت قبلی ممکن است که گفتار آزمایش مربوط به هیچ یک از گویندگان ثبت-نامی در سامانه نباشد و باید به این سوال که آیا گفتار آزمایش توسط یکی از گویندگان ثبت‌نامی تولید شده است یا نه؟ پاسخ بدهیم. به این دلیل کار شناسایی گوینده در این حالت سخت‌تر از حالت قبل است. شناسایی گوینده‌ی مجموعه‌ی باز ترکیبی از شناسایی گوینده‌ی مجموعه‌ی بسته و تصدیق هویت گوینده است و در حقیقت تصدیق هویت گوینده حالت خاصی از شناسایی گوینده‌ی مجموعه‌ی باز می-باشد که در هر لحظه فقط یک گوینده در سامانه وجود دارد. به همین دلیل شناسایی گوینده‌ی مجموعه-ی باز از تصدیق هویت گوینده پیچیده‌تر است.

در این پایان‌نامه به علت اینکه روش وابسته به متن برای شناسایی گوینده کاربرد عملی خاصی ندارد ما روش مستقل از متن را انتخاب کرده‌ایم. علاوه بر این، چون هدف ما در این پایان‌نامه سرعت بخشیدن به مرحله‌ی شناسایی است و از آنجایی که فرآیند زمان‌بر در این مرحله، جستجوی خطی روی تمام گویندگان سامانه است، حالت مجموعه بسته انتخاب شده است. این فرآیند زمان‌بر در هر دو کاربرد مجموعه‌ی باز و مجموعه‌ی بسته مشترک است.

۲-۳ اجزاء و مراحل سامانه‌های شناسایی گوینده

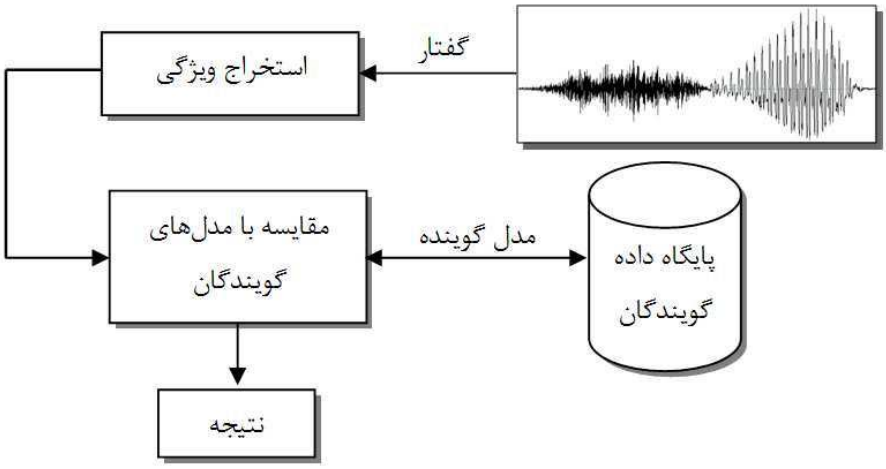
به طور کلی هر سامانه‌ی شناسایی گوینده شامل دو مرحله‌ی مجزا است که مرحله‌ی اول مرحله‌ی ثبت-نام (آموزش) و مرحله‌ی دوم مرحله‌ی شناسایی (آزمایش) است. در مرحله‌ی ثبت‌نام، ابتدا از گفتار آموزشی هر گوینده ویژگی‌هایی استخراج می‌شود. سپس این ویژگی‌ها برای مدل کردن خصوصیات صوتی گوینده استفاده می‌شود. در مرحله‌ی مدل‌سازی، برای هر گوینده مدل واحدی ساخته می‌شود و در نهایت این مدل‌ها در پایگاه‌داده‌ای از گویندگان ذخیره می‌شوند. شکل ۲-۱ نمودار جعبه‌ای مرحله‌ی ثبت‌نام را نشان می‌دهد.



شکل ۲-۱ نمودار جعبه‌ای مرحله‌ی ثبت‌نام

بعد از اینکه مدل گویندگان ساخته شد، سامانه آماده‌ی انجام عمل شناسایی است. در مرحله‌ی شناسایی ابتدا از گفتار ورودی ویژگی‌هایی استخراج می‌کنیم. استخراج ویژگی در این مرحله نیز شبیه مرحله‌ی ثبت‌نام است. بعد از استخراج ویژگی نوبت به انجام عمل شناسایی می‌رسد. در این مرحله،

ویژگی‌های استخراج شده با مدل تمام گویندگان مقایسه می‌شود و گوینده‌ای که مدل آن بیشترین شباهت را با این ویژگی‌ها داشته باشد به عنوان گوینده‌ی هدف انتخاب می‌شود. عمل مقایسه بسته به نوع مدلی که برای گویندگان انتخاب می‌شود متفاوت است. شکل ۲-۲ نمودار جعبه‌ای مرحله‌ی شناسایی را نشان می‌دهد.



شکل ۲-۲ نمودار جعبه‌ای مرحله‌ی شناسایی

بیشترین محاسبات در مرحله‌ی شناسایی را عمل مقایسه به خود اختصاص می‌دهد. برای افزایش سرعت سامانه باید تا می‌توان محاسبات این قسمت را کاهش داد.

۲-۴ پردازش اولیه و استخراج ویژگی

۲-۴-۱ مقدمه

هدف از بلوک پردازش اولیه و استخراج ویژگی، این است که یک سری ویژگی‌هایی که توانایی نشان دادن اطلاعات و ویژگی‌های فردی هر گوینده را دارد از سیگنال گفتار استخراج کند. در این بلوک بسته به کاری که قرار است انجام شود ممکن است عملیات دیگری مثل حذف نویز و سایر پیش‌پردازش‌ها[1] انجام گیرد.

[1] Pre-Processing

ویژگی‌هایی که در پردازش گفتار و بویژه شناسایی گوینده استفاده می‌شوند به دو دسته‌ی کوتاه مدت[1] و بلند مدت[2] تقسیم می‌شوند. ویژگی‌های کوتاه مدت از قاب‌های کوتاهی از گفتار استخراج می‌شوند و نشان‌دهنده‌ی ساختار فیزیکی مسیر صوتی گوینده می‌باشند. از طرف دیگر، ویژگی‌های بلند مدت از محدوده‌ی طولانی‌تری از گفتار استخراج می‌شوند و نشان دهنده‌ی اطلاعات آوایی و لغوی گوینده می‌باشند. برای انتخاب و استخراج هرگونه ویژگی، باید نیازمندی‌هایی مورد توجه قرار گیرد. این نیازمندی‌ها ممکن است استفاده یا عدم استفاده از ویژگی‌های مختلف را تحت تأثیر قرار دهد. از این نیازمندی‌ها می‌توان به موارد زیر اشاره کرد [Naik_90]، [Atal_76]:

- قابلیت جداسازی گویندگان را داشته باشد در حالی که نسبت به تغییرات ویژگی‌های صوتی یک گوینده حساس نباشد.
- به راحتی و در زمانی معقول قابل استخراج باشد.
- نسبت به گذشت زمان و تغییرات سنی و همچنین حالات گوینده (مثل سرماخوردگی) مقاوم باشد.
- به صورت طبیعی و متداوم در گفتار تکرار شود.
- تغییرات کمی در محیط‌های مختلف و نویزی داشته باشد.
- قابل تقلید توسط افراد دیگر نباشد.

در کاربردهای عملی امکان اینکه یک ویژگی تمام نیازهای بالا را برآورده کند، وجود ندارد و همیشه یک رابطه‌ی سبک و سنگینیی بین آنها وجود دارد و بسته به کاربردهای مختلف اهمیت این نیازها تغییر می‌کند.

[1] Short-term
[2] Long-term

۲-۴-۲ ویژگی‌های کوتاه مدت

اکثر سامانه‌های بازشناسایی گوینده از ویژگی‌های کوتاه مدت که از بخش کوتاهی از گفتار استخراج می-شوند، استفاده می‌کنند. از ویژگی‌های کوتاه مدت که بیشتر مورد استفاده قرار می‌گیرند می‌توان به ضرایب کپسترال مقیاس فرکانس مل (MFCCs[1]) [Davis_80]، ضرایب کپسترال پیشگویی خطی (LPCCs[2]) [Makhoul_75] و ضرایب پیشگویانه خطی ادراکی (PLP[3]) [Hermansky_90] اشاره کرد. از آنجایی که ضرایب کپسترال مقیاس فرکانس مل در بازشناسایی گوینده زیاد استفاده شده است و دقت بهتری نسبت به سایر ویژگی‌های کوتاه مدت دارد ما نیز در این پایان‌نامه از آن استفاده کرده‌ایم. در ادامه این ویژگی به تفصیل شرح داده می‌شود.

۲-۴-۳ ویژگی‌های بلند مدت

در سال‌های اخیر تلاش‌های زیادی برای استفاده از ویژگی‌های بلند مدت به منظور بالا بردن دقت و همچنین مقاوم‌سازی روش‌های بازشناسی گوینده در سامانه‌های مستقل از متن شده است. طرفداران این ویژگی‌ها معتقد هستند که با استفاده از ویژگی‌های کوتاه مدت به تنهایی نمی‌توان دقت سامانه‌های بازشناسی را بالاتر برد [Campbell_03] و به همین دلیل استفاده از ویژگی‌های بلند مدت را امری ضروری می‌دانند.

ویژگی‌های بلند مدت به سه طبقه‌ی آوایی[4]، لغوی[5] و نوایی[6] تقسیم می‌شوند [Buyuk_11]. دنباله‌ی زمانی آواها و تلفظ‌های منحصر به فرد یک گوینده در دسته آوایی قرار می‌گیرند. این ویژگی‌ها اغلب با استفاده از خروجی یک سامانه‌ی بازشناسی گفتار استخراج می‌شوند. ویژگی‌های لغوی شامل تکرار کلمات خاص یا تکه کلام‌های یک گوینده می‌باشد که اینها هم از خروجی سامانه‌ی بازشناسی گفتار

[1] Mel-frequency cepstral coefficients
[2] Linear prediction cepstral coefficients
[3] Perceptual linear predictive
[4] Phonetic
[5] Lexical
[6] Prosodic

استخراج می‌شوند. ویژگی‌های نوایی شامل تغییرات آهنگ گفتار، بلندی و طول زمانی قسمت‌های کلمه و حتی جمله می‌باشد. در این دسته فرکانس گام[1]، دیرش[2] و انرژی از مهمترین ویژگی‌ها می‌باشند.

به طور معمول استخراج ویژگی‌های بلند مدت نیاز به داده‌های آموزشی بیشتری دارند. علاوه بر این، برای استخراج بیشتر این ویژگی‌ها نیاز به یک سامانه‌ی بازشناسی گفتار با دقت بالا است. این محدودیت‌ها باعث شده که استفاده از این ویژگی‌ها رونق چندانی پیدا نکند. با توجه به این محدودیت‌ها ما در این پایان‌نامه فقط از ویژگی‌های کوتاه مدت استفاده کرده‌ایم.

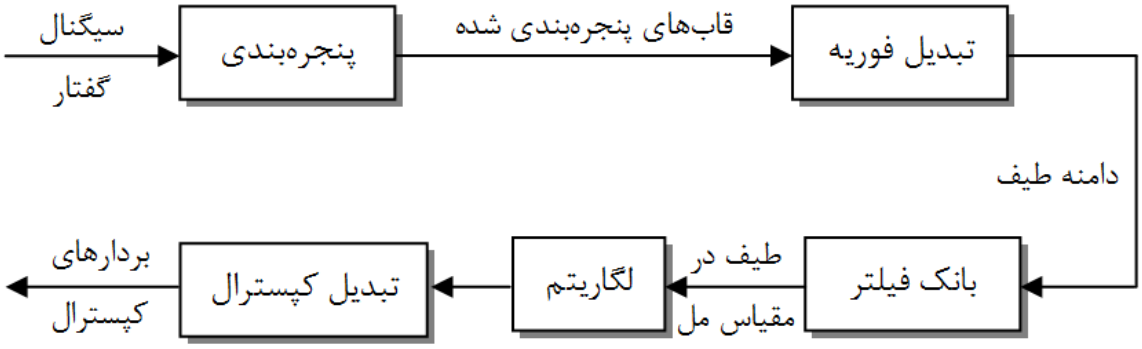
۲-۴-۴ ضرایب کپسترال مقیاس فرکانس مل

شکل ۲-۳ نمودار جعبه‌ای مراحل استخراج ضرایب MFCC را نشان می‌دهد. برای استخراج این ضرایب، ابتدا گفتار را به قاب‌های هم‌پوشان تقسیم کرده و سپس از هر قاب یک بردار ضرایب استخراج می‌کنند. طول قاب‌ها در حدود ۲۰ تا ۳۰ میلی‌ثانیه در نظر گرفته می‌شود. این طول را طوری در نظر می‌گیرند که ویژگی ایستا بودن سیگنال گفتار در هر قاب حفظ شود. جابه‌جایی قاب‌ها برای حفظ هم‌پوشانی قاب‌های مجاور، در حدود ۱۰ الی ۱۵ میلی‌ثانیه است. بعد از اینکه سیگنال گفتار را به قاب‌های هم‌پوشان تقسیم کردند، برای اینکه تأثیر ناپیوستگی در لبه قاب‌ها را کم کنند هر قاب را در پنجره‌ای ضرب می‌کنند. معمولاً از پنجره‌های همینگ[3] یا هنینگ[4] برای این کار استفاده می‌شود. ما در اینجا از پنجره‌ی همینگ استفاده کرده‌ایم. فرمول (۲-۱) معادله‌ی حوزه‌ی زمان این پنجره را نشان می‌دهد:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \qquad (۲-۱)$$

در معادله‌ی بالا N نشان دهنده‌ی طول پنجره و n نشان دهنده‌ی عنصر n-ام از پنجره‌ی همینگ است.

[1] Pitch
[2] Duration
[3] Hamming
[4] Hanning



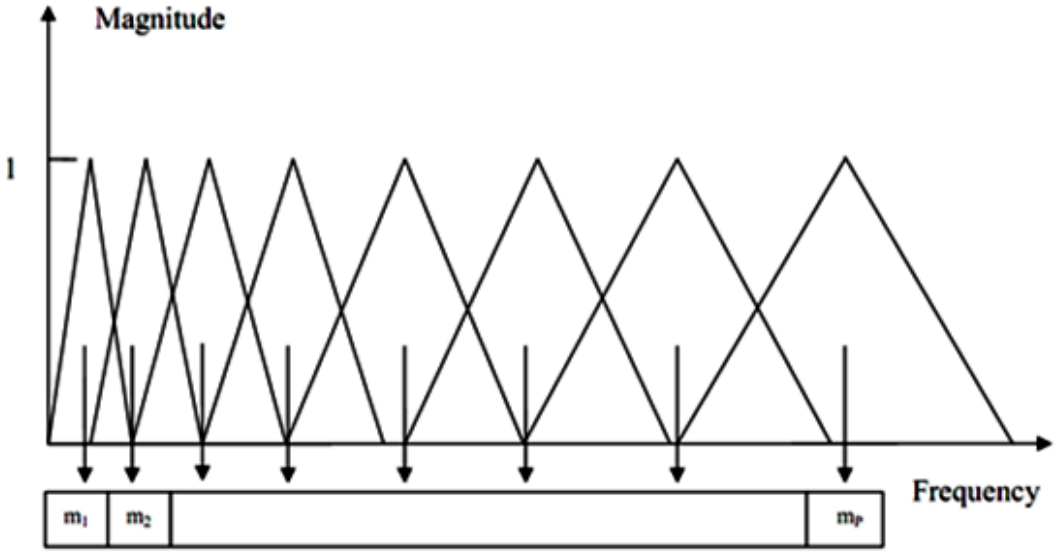
شکل ۲-۳ نمودار جعبه‌ای مراحل استخراج ضرائب کپسترال

بعد از اینکه قاب را در پنجره ضرب کردند از آن تبدیل فوریه گسسته (DFT[1]) گرفته و دامنه‌ی تبدیل را در حوزه‌ی فرکانس حساب می‌کنند. تا اینجا دامنه‌ی DFT قاب گفتار بدست آمده است که آن را با X_k نشان می‌دهند، (k اندیس DFT است). حال X_k را در بانک فیلتر مل متناظر ضرب کرده و نتایج را با هم جمع می‌کنند. نتیجه اینکه هر ضریب بانک فیلتر (m_i) یک جمع وزن‌دار از دامنه‌ی طیف در کانال i–ام از بانک فیلتر می‌باشد که مطابق فرمول زیر محاسبه می‌شود:

$$m_i = \sum_k w_{k,i} X_k \qquad (۲-۲)$$

جایی که $w_{k,i}$ وزن فیلتر در کانال i–ام را نشان می‌دهد. این وزن‌ها در خارج از باندفرکانسی فیلتر صفر در نظر گرفته می‌شوند. بانک فیلتر مل استفاده شده در شکل ۲-۴ نشان داده شده است:

[1] Discrete Fourier transform



شکل ۲-۴ بانک فیلتر مل

در محاسبات ضرایب MFCC، بانک فیلترها دارای توزیع یکسانی در حوزه‌ی فرکانس نیستند ولی آنها به صورت متساوی الفاصله در طول مقیاس مل که توسط فرمول (۲-۳) تعریف می‌شود پخش شده‌اند.

$$Mel(f) = 2595 * \log(1 + \frac{f}{700}) \qquad (۲-۳)$$

علت عدم توزیع یکنواخت این است که پاسخ فرکانسی گوش انسان به صورت غیریکنواخت در حوزه‌ی فرکانس توزیع شده است. همچنین، علت استفاده از توزیع غیرخطی این است که اطلاعاتی که توسط مؤلفه‌های فرکانس پایین منتقل می‌شوند برای انسان‌ها اهمیت بیشتری نسبت به اطلاعات مؤلفه-های فرکانس بالا دارند. توزیع غیرخطی مل نسبت به سایر توزیع‌های خطی در حوزه‌ی فرکانس باعث افزایش دقت سامانه‌های بازشناسی می‌شود. در پایان با استفاده از فرمول (۲-۴) از لگاریتم ضرایب بانک فیلتر، تبدیل کسینوسی گسسته (DCT[1]) گرفته می‌شود. با این کار ضرایب نهایی بدست می‌آیند.

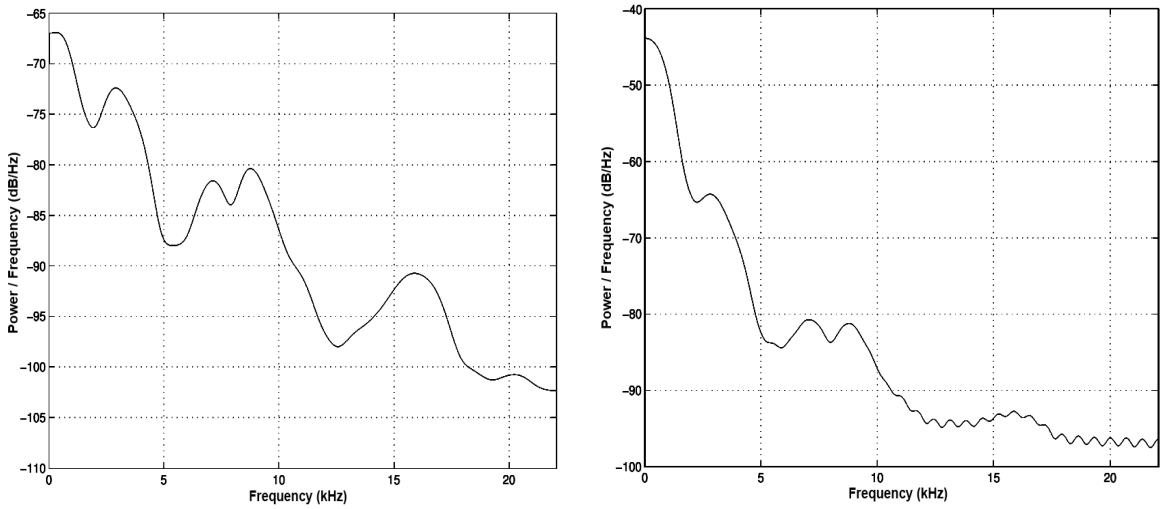
$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log\left(m_j\right) \cos\left(\frac{\pi i}{N}\left(j - 0.5\right)\right) \qquad (۲-۴)$$

در فرمول بالا N تعداد کانال‌های بانک فیلتر می‌باشد. تنها تعداد کمی از ضرایب DCT نگه داشته می-شوند تا یک بردار ویژگی را نشان دهند. این ضرایب را ضرایب ایستا می‌گویند.

[1] Discrete cosine transform

۲-۴-۵ فیلتر پیش-تاکید[1]

خواص سیستم صوتی انسان، جریان هوای حنجره و تشعشع[2] لب باعث تعدیل رفتن مؤلفه‌های فرکانس بالای گفتارهای صدادار می‌شود [Harrington_99]. برای گفتارهای صدادار، طیف سیگنال خروجی از دهانه‌ی حنجره تقریباً دارای شیب 12db/octave- است. زمانی که این سیگنال‌ها از لب‌ها انتشار پیدا می‌کنند، طیف آنها در فرکانس‌های بالا تقویت شده و تقریباً 6db/octave+ از شیب قبلی جبران می‌شود و در نتیجه طیف سیگنال خروجی در مجموع با شیب 6db/octave- از فرکانس‌های پایین به فرکانس-های بالا کاهش می‌یابد. در شکل ۲-۵ (الف) کاهش شدید توان سیگنال در فرکانس‌های بالا قابل مشاهده است.



(ب) (الف)

شکل ۲-۵ نمودار چگالی طیف توان، (الف) سیگنال اصلی با فرکانس نمونه‌برداری ۴۴۱۰۰، (ب) پیش-تاکید شده همان سیگنال [Beigi_11]

در سیستم شنوایی انسان، حلوزن گوش[3] بر اساس بازخورد مغز پردازش‌هایی روی گفتار انجام می-دهد و بعضی از فرکانس‌ها را تقویت می‌کند. این کار باعث می‌شود که انسان‌ها به راحتی قسمت کم انرژی سیگنال در فرکانس‌های بالا را دریافت کنند. برای شبیه‌سازی سیستم شنوایی انسان و جبران این اثر و جلوگیری از افزایش اثر مؤلفه‌های فرکانس پایین گفتار، قبل از استخراج ویژگی باید فیلتر پیش-تاکید اعمال شود. در شکل ۲-۵ (ب) همان سیگنال در حوزه‌ی فرکانس نشان داده شده است با این

[1] Preemphasis
[2] Radiations
[3] Cochlea

تفاوت که سیگنال از فیلتر پیش-تاکید عبور کرده است. توجه کنید که مقدار توان برای تمام فرکانس‌ها کاهش پیدا کرده است، اما توان نسبی در فرکانس‌های مختلف دارای توزیع بهتری است.

معمولاً فیلتر پیش-تاکید بوسیله‌ی فیلتر کردن سیگنال گفتار توسط یک فیلتر با پاسخ ضربه‌ی محدود (FIR) مرتبه‌ی اول اعمال می‌شود. معادله‌ی این فیلتر به فرم زیر است:

$$F(z) = 1 - kz^{-1} \qquad (0 < k < 1) \qquad (۵-۲)$$

در فرمول بالا، k فاکتور پیش‌تاکید است که مقدار متداول برای آن ۰.۹۷ می‌باشد. همان طور که مشخص است سیگنال خروجی در حوزه‌ی زمان به فرم زیر بدست می‌آید:

$$y(n) = s(n) - k * s(n-1) \qquad (۶-۲)$$

جایی که s(n) سیگنال ورودی و y(n) سیگنال خروجی فیلتر می‌باشد.

۲-۴-۶ ویژگی‌های پویا

معمولاً دقت سامانه‌های بازشناسی با اضافه کردن اطلاعات پویا (مشتق زمانی) به پارامترهای ایستا بطور چشم‌گیری افزایش می‌یابد. بعد از اینکه پیش‌پردازش‌های اولیه روی ویژگی‌های ایستا اعمال شد، اطلاعات پویا که نمایانگر تغییرات زمانی بردارهای کپسترال هستند به ویژگی‌ها اضافه می‌شوند. بطور معمول از مشتق اول و دوم استفاده می‌شود و مشتقات بالاتر به دلیل اینکه تأثیر اندکی بر دقت دارند مورد استفاده قرار نمی‌گیرند. برای بدست آوردن ضرایب پویای مرتبه‌ی اول (مشتق اول) از فرمول درون-یابی زیر استفاده می‌شود:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta \left(c_{t+\theta} - c_{t-\theta} \right)}{2 \sum_{\theta=1}^{\Theta} \theta^2} \qquad (۷-۲)$$

جایی که dt ضریب دلتا در زمان t می‌باشد که با استفاده از ضرایب ایستای قبلی و بعدی محاسبه می-شود. پارامتر Θ اندازه‌ی پنجره برای بدست آوردن ضرایب دلتا را مشخص می‌کند. برای بدست آوردن

ضرایب مراتب بالاتر از فرمول مشابه‌ای بر روی ضرائب پویای مرتبه‌ی پایین‌تر استفاده می‌شود [Young_06].

۲-۵ مدل کردن گویندگان

۲-۵-۱ مقدمه

در بازشناسی گوینده، روش‌های مختلفی برای مدل کردن گویندگان وجود دارد. این روش‌ها به دو دسته-ی کلی مولد[1] و تمایزی[2] تقسیم می‌شوند. در روش مدل‌سازی مولد، مدل گویندگان به سادگی توسط تابع چگالی احتمال نشان داده می‌شود. از روش‌های مختلف این دسته می‌توان به چندی‌ساز برداری (VQ[3]) [Soong_85]، مدل مخلوط گاوسی، مدل مخفی مارکوف [Deshpande_08] و پیچش زمانی پویا (DTW[4]) [Wutiwiwatchai_99] اشاره کرد. در طرف دیگر، در روش‌های مختلف دسته‌ی تمایزی از جمله ماشین بردار پشتیبان (SVMs[5]) [Xing_12]، [Ruiling_11] و شبکه‌های عصبی مصنوعی (ANNs[6]) [Hossain_07]، [Rao_10] یک مرز بین گویندگان مختلف تخمین زده می‌شود.

اگرچه در سال‌های اخیر روش‌های متفاوتی که از مدل‌سازی تمایزی استفاده کرده و در حالت مستقل از متن کار بازشناسی گوینده را انجام می‌دهند موفقیت‌های قابل قبولی داشته‌اند، ما در این پایان‌نامه از روش‌های مولد استفاده کرده‌ایم. علت این انتخاب این است که در روش‌های مولد از جمله مدل مخلوط گاوسی حجم محاسبات در مرحله‌ی شناسایی بسیار بالا است و این در حالی است که محاسبات مورد نیاز برای روش‌های تمایزی کمتر است. از آنجا که هدف اصلی ما در این پایان‌نامه کم کردن محاسبات مرحله‌ی شناسایی و در نتیجه افزایش سرعت این مرحله است، از مدل‌های مولد استفاده شده است.

[1] Generative
[2] Discriminative
[3] Vector quantization
[4] Dynamic time warping
[5] Support vector machine
[6] Artificial neural networks

۲-۵-۲ مدل مخلوط گاوسی

مدل مخلوط گاوسی پرکاربردترین روش مدل‌سازی بویژه در کاربردهای مستقل از متن است. این مدل به ویژگی‌های زمانی و دنباله‌ای گفتار حساس نیست و به همین خاطر در کاربردهای وابسته به متن کمتر استفاده می‌شود. خصوصیات زمانی و چگونگی تغییرات گفتار در طول زمان حاوی اطلاعاتی برای شناسایی گوینده است که در کاربردهای وابسته به متن مورد توجه قرار می‌گیرد. تابع چگالی احتمال در مدل مخلوط گاوسی، مجموع وزن‌داری از M تابع گاوسی (مخلوط) به صورت زیر است:

$$\Pr(\mathbf{x}\mid\lambda)=\sum_{i=1}^{M} w_i g\left(\mathbf{x}\mid\mu_i,\Sigma_i\right) \qquad (۸-۲)$$

جایی که $\mathbf{x}$ یک بردار D بعدی با مقدار پیوسته و w_i وزن مخلوط i-ام است. وزن مخلوط‌ها باید به گونه-ای بدست بیایند که شرط زیر را ارضاء کنند:

$$\sum_{i=1}^{M} w_i = 1 \qquad (۹-۲)$$

در رابطه (۲-۸)، هر مخلوط یک توزیع گاوسی می‌باشد، $g(x|\mu_i,\Sigma_i)$ ، که با بردار میانگین μ_i و ماتریس کواریانس[1] Σ_i مشخص می‌شود. فرمول توزیع گاوسی در رابطه‌ی زیر نشان داده شده است:

$$g(\mathbf{x}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} * \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)'\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right] \qquad (۱۰-۲)$$

مجموعه‌ی تمام پارامترهای مدل مخلوط گاوسی توسط سه‌تایی زیر نشان داده می‌شود:

$$\lambda=\left\{w_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right\} \quad i=1,\ldots,M$$

[1] Covariance matrice

فرم کلی مدل‌سازی در این روش از ماتریس کواریانس کامل استفاده می‌کند، اما در عمل از ماتریس کواریانس قطری استفاده می‌شود. دلیل استفاده از ماتریس قطری این است که ماتریس کواریانس قطری از نظر محاسباتی نسبت به ماتریس کواریانس کامل بهینه‌تر است. علاوه بر این مزیت، دقت شناسایی در این حالت نسبت به ماتریس کامل بیشتر است [Reynolds_00]. زمانی که از ماتریس کواریانس کامل استفاده شود، تعداد پارامترهایی که باید تخمین زده شوند بیشتر می‌شود و این افزایش تعداد پارامترها نیاز به داده‌های آموزشی بیشتری نسبت به حالت قطری دارد. در کاربردهای عملی معمولاً میزان داده‌های آموزشی کم می‌باشد که در این حالات بهتر است از ماتریس کواریانس قطری استفاده کنیم. علاوه بر این، از آنجایی که مخلوط‌های گاوسی با یکدیگر تمام فضای ویژگی‌ها را مدل می‌کنند، استفاده از ماتریس کواریانس کامل حتی اگر ویژگی‌هایی که استفاده می‌شوند به صورت آماری مستقل نباشند ضرروری نیست. ترکیب خطی توابع گاوسی با کواریانس قطری امکان مدل کردن رابطه بین ابعاد بردارهای ویژگی را دارد [Reynolds_95a]. تاثیر استفاده از مجموعه‌ای از توابع گاوسی با ماتریس کواریانس کامل را می‌توان با مجموعه‌ای بزرگتر از توابع گاوسی با ماتریس کواریانس قطری ایجاد کرد.

همان طور که مشاهده نمودید هر مدل مخلوط گاوسی دارای پارامترهای نامعلوم زیادی است. این پارامترها باید برای هر گوینده به روشی تخمین زده شوند. برای تخمین پارامترهای مدل، دو رویکرد کلی وجود دارد. در رویکرد اول پارامترهای مدل هر گوینده به صورت مجزا تخمین زده می‌شوند. رویکرد دوم به این صورت می‌باشد که ابتدا یک مدل جهانی از داده‌های آموزشی تمام گویندگان می‌سازیم. سپس با استفاده از پارامترهای این مدل برای هر گوینده، یک مدل منحصر به فرد بدست می‌آوریم. برای هر کدام از این دو رویکرد روش‌های مختلفی وجود دارد. موفقترین روش برای بدست آوردن مستقیم پارامترهای مدل هر گوینده الگوریتم بیشینه‌سازی امید ریاضی (EM[1]) است. از روش‌های رویکرد دوم می‌توان به

[1] Expectation-Maximization (EM) algorithm