دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد
گرایش هوش مصنوعی

بررسی روش تشخیص صحبت در انسان و شبیه‌سازی آن

نگارنده:

سید کمال الدین غیاثی شیرازی

استاد راهنما:

دکتر سعید باقری شورکی

زمستان ۱۳۸۳

به نام خدا

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

رساله کارشناسی ارشد

عنوان: بررسی روش تشخیص صحبت در انسان و شبیه‌سازی آن

نگارش: سید کمال‌الدین غیاثی شیرازی

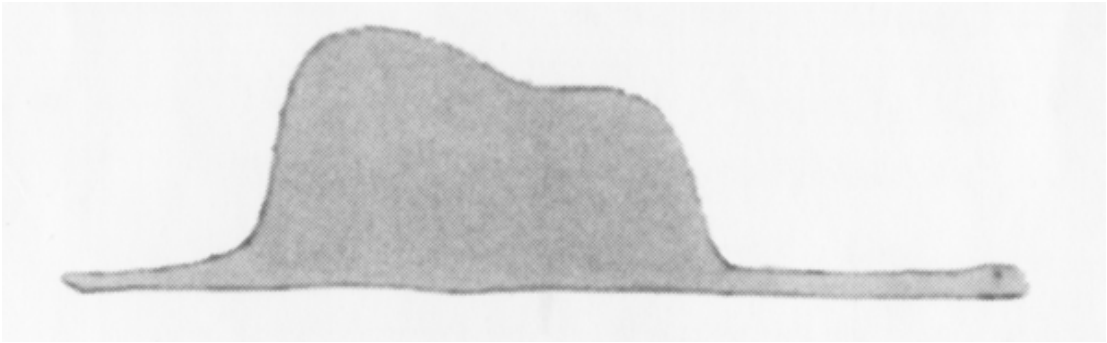
کمیته ممتحنین:

استاد راهنما: دکتر سعید باقری شورکی **امضاء**.....

استاد مشاور: دکتر حسین ثامتی **امضاء**.....

استاد مدعو: دکتر فرشاد الماس‌گنج **امضاء**.....

تاریخ:.....



تقدیم به مادر و پدر عزیزم و همه اساتیدم

تقدیم به دکتر باقری که انجام این پروژه بدون دلگرمی ایشان میسر نبود

تقدیم به دکتر باقری که امیدوارم فازی را از ایشان یاد گرفته باشم

تقدیم به دکتر رضوی‌زاده که سیگنال و سیستم را از ایشان آموختم

تقدیم به دکتر رضوی‌زاده استاد راهنمای پروژه دوره کارشناسی‌ام

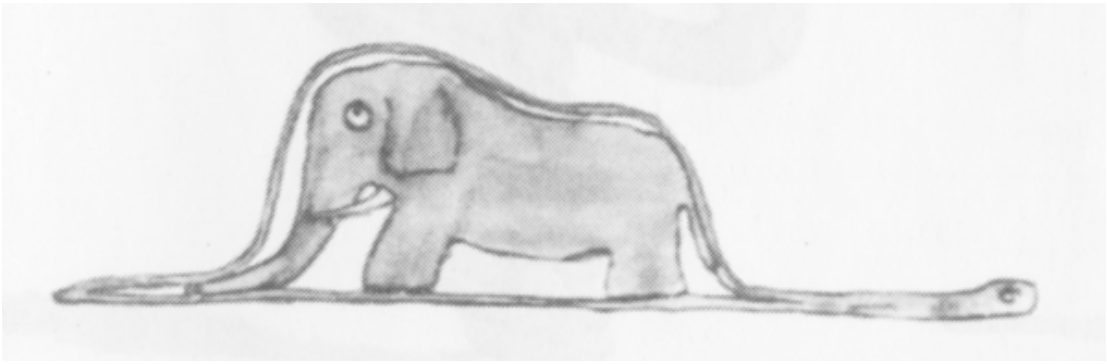
تقدیم به دکتر ثامتی که تجربیات بسیار با ارزشی را به من منتقل کردند

تقدیم به استاد حمید خورسند بخاطر نگاه زیبایشان به علم

تقدیم به دکتر قریشی که آمار و احتمال را از ایشان آموختم

تقدیم به تمام معلمینی که هرگز فراموششان نکرده‌ام

و تقدیم به تمام آدم‌کوچولوهای دنیا که به مار بوایی که فیل را خورده‌است کلاه نمی‌گویند



تشکر و سپاس گذاری

در اینجا لازم است از افراد زیادی تشکر کنم. اول از همه از دکتر باقری و دکتر رضوی‌زاده که اجازه تعریف پایان نامه هایی را به من دادند که در آن من توانستم به عنوان یک محقق بر روی یک مطلب کار کنم. جمله‌ای را که دکتر باقری در هنگام تعریف پایان‌نامه‌ام به من گفت فراموش نمی‌کنم. «می‌خواهی چه چیزی را به دنیای علم اضافه کنی؟». سپس باید از دکتر ثامتی که تمام امکاناتشان را در اختیار من قرار دادند و به من اهمیت مقایسه بین سیستم‌ها را یاد دادند تشکر کنم. رسیدن به نتایجی که در این پایان نامه به آن رسیدیم بدون این کمک میسر نبود. همچنین لازم است از آقای فاضل و آقای باباعلی و برادرم محمد امین و دیگر دوستانم در شرکت عصر گویش بخاطر گوش دادن به ایده‌هایم و بحث در مورد آنها تشکر کنم. همچنین از زحمات دکتر قدسی در آموختن روش ارائه مطلب و زحمات دکتر ثانی در آموختن روش نوشتن مقاله قدردانی می کنم (هر چند فکر می کنم هنوز نتوانسته‌ام ارائه کردن را یاد بگیرم).

همچنین بر خود لازم می‌دانم که از حمایت مالی مرکز تحقیقات مخابرات ایران تشکر کنم.

چکیده

در این پایان‌نامه سعی شد روش انسان در تشخیص صحبت مورد بررسی قرار گیرد. برای بررسی روش تشخیص صحبت در انسان ابزاری ساخته شد که اجازه اعمال انواع تغییرات در سیگنال صحبت به نحوی که طبیعی بودن صدا حفظ شود را می‌دهد. بدین ترتیب ما توانستیم ببینیم که انسان از چه ویژگی‌هایی برای تشخیص صحبت استفاده می‌کند. ما به این نتیجه رسیدیم که ویژگی‌هایی که انسان بر اساس آنها صحبت را تشخیص می‌دهد دارای انرژی بالایی هستند. بر اساس این نتیجه یک سیستم بخش‌بندی و استخراج ویژگی به نام OBSFE ساخته شد که به تغییرات جزئی در انرژی ویژگی‌ها حساس نیست. این روش استخراج ویژگی چند مزیت عمده دارد: اول اینکه مدل‌گرا نیست. یعنی از اطلاعات زبانی برای بخش‌بندی استفاده نمی‌شود. این مطلب برای سیستمی که می‌خواهد مشابه انسان زبان را یاد بگیرد بسیار حیاتی است. دومین ویژگی این روش این است که ویژگی‌ها را در زمان–فرکانس و نه فقط در فرکانس استخراج می‌کند. کارهای دیگر محققین نشان داده است که اولا این روش به روش انسان در استخراج ویژگی شبیه‌تر است و ثانیا نسبت به نویز مقاوم‌تر است. آزمایش‌های ما نیز مقاوم بودن این روش نسبت به نویز را تایید می‌کنند. روش OBSFE در نویزهای همهمه، اتومبیل، رستوران، فرودگاه و ایستگاه قطار و در نویزهای ۵ ، ۱۵ و ۲۰ دسیبل نرخ بازشناسی را ۲۱.۴۴٪ افزایش می‌دهد. ویژگی سوم روش OBSFE این است که بخش‌ها دارای همپوشانی هستند. تا آنجا که ما می‌دانیم این اولین روشی است که سیگنال صحبت را به بخش‌های دارای همپوشانی بخش‌بندی می‌کند. آزمایش نشان داد که این روش استخراج ویژگی در تشخیص کلمات مناسب است ولی برای تشخیص واج خوب عمل نمی‌کند. همچنین نظریه امکان به عنوان جایگزینی برای نظریه احتمال برای سیستم بازشناسی مورد بررسی قرار گرفت و نشان داده شد که از دیدگاه نظری این نظریه قابلیت یادگیری اشیاء جدید را دارد. همچنین با یک پیاده‌سازی ساده از یک سیستم بازشناسی مبتنی بر نظریه امکان نشان داده شد که نظریه امکان در شرایط نویزی بسیار خوب عمل می‌کند. از طرف دیگر یک اندازه‌گیری جدید امکانی که مبتنی بر کوانته شدن مقادیر دامنه به ۱۰۰ مقدار بر حسب صدک‌ها است پیاده‌سازی شد و نشان داده شد که این روش اندازه‌گیری دقت سیستم‌های تشخیص صحبت را در تشخیص کلمه و واج پایین نمی‌آورد. در نهایت ما توانستیم با استفاده از این سیستم بازشناسی مبتنی بر نظریه امکان به دقت ۴۹.۳۹٪ و درستی ۶۴.۵٪ در تشخیص واج در مجموعه تهرانی دادگان فارس‌دات دست یابیم.

کلمات کلیدی:

۱- تشخیص صحبت – Speech recognition

۲- منطق فازی – Fuzzy logic

۳- نظریه امکان – Possibility theory

۴- انسان – Human

۵- استخراج ویژگی - Feature extraction

فهرست مطالب

فهرست جدول ها

فهرست شکل ها

مقدمه

«یک فراری خیلی سریع‌تر از یک جیپ جنگی است، ولی جیپ جاهایی می‌تواند برود که فراری حتی فکرش را هم نمی‌تواند بکند» لطفی‌زاده

افرادی که در رشته‌های علمی کار می‌کنند را می‌توان به دو دسته تقسیم کرد. کسانی که هدفشان از علم کشف جهان است و کسانی که علم را می‌آموزند تا آن را در جهت حل مشکلات عملی اجتماع به کار گیرند. به عنوان نمونه‌ای از گروه اول از انشتین یاد می‌کنم که معتقد بود که هرچند نظریه کوانتوم در عمل مشکلات ما را حل کرده است اما چون جهان را خوب تفسیر نمی‌کند قابل پذیرش نیست. در مقابل او طرفداران نظریه کوانتوم می‌گویند که ما از علم پیش‌بینی را می‌خواهیم و نه بیشتر. از نظر این افراد هر مدلی که بتواند در عمل فرآیندهای فیزیکی را پیش‌بینی کند قابل پذیرش است. مشابه همین تقسیم‌بندی در هوش مصنوعی وجود دارد. قاطبه افرادی که در زمینه‌های تکنولوژیک هوش مصنوعی کار می‌کنند (کنترل، روباتیک، پردازش تصویر، پردازش صوت و ...)، وجهه همت خود را دستیابی به سیستم‌های کارا قرار داده‌اند. برای این افراد لغت هوش معنایی ندارد. این درحالی است که مبدعین مدل محاسباتی کامپیوتر و دانشمندان علوم مرتبط، آلن تورینگ، کرت گودل و ...، اهمیت زیادی به مقایسه بین انسان و ماشین می‌داده‌اند. تورینگ تست تورینگ را مطرح کرد و گودل با بحثی فلسفی ادعا کرد که قابلیت ماشین ذاتا از قابلیت انسان کمتر است.

این تحقیق متعلق به گروه اول است. هدف ما کشف روش انسان در برخورد با مسائل است. در حقیقت مسائل زیادی هستند که ماشین[1] بهتر از انسان با آنها برخورد می‌کند. هواپیما، خودرو، چاپگر، ماشین‌حساب و ... نمونه‌هایی از سیستم‌هایی ماشینی هستند که بهتر از انسان کار می‌کنند. اما واقعیت این است که انسان هواپیما را ساخته است و هواپیما نمی‌تواند انسان بسازد. با این جهت‌گیری دیگر هدف محاسبه حاصل‌ضرب دو عدد نیست که بگوییم ماشین بهتر انجام می‌دهد. هدف کشف روش انسان در حل این مساله است. البته اگر این مسیر به درستی طی شود، قطعا سیستم مبتنی بر روش انسان از بسیاری جهات بر سیستم‌های ماشینی برتری خواهد داشت.

در علم تشخیص صحبت نیز حیطه‌هایی وجود دارند که ماشین بهتر از انسان کار می‌کند:

۱- تشخیص واج‌های جدا شده از صحبت پیوسته

۲- تشخیص تعداد محدودی کلمه در یک محیط کاملا مشخص

۳- تشخیص صحبت در شرایط بسیار نویزی (0dB و -5dB)

اما در مجموع قابلیت هیچ سیستم تشخیص صحبت ماشینی را نمی‌توان با قابلیت انسان مقایسه کرد.

---

[1] ما نیز نهایتا سیستم خود را بر روی همین ماشین محاسباتی، رایانه، اجرا خواهیم کرد. اما با مشیی مشابه انسان و نه برگرفته از ریاضیات کلاسیک.

انسان صحبت را بدون توجه به گوینده، نوع صحبت کردن (صحبت معمولی، آواز و ...) و در شرایطی که انواع نویز بر روی صدا قرار گرفته است تشخیص می‌دهد. بعلاوه انسان از تشخیص خود مطمئن است. به عبارت دیگر انسان فقط بین چند گروه مشخص تفاوت قائل نمی‌شود بلکه آنها را می‌شناسد.

از دیدی دیگر عمر روش‌های متداول بسیار طولانی شده است و با وجود هزینه زیادی که شده است جهان هنوز نتوانسته است یک سیستم تشخیص صحبت قابل توجه ارائه دهد. از اوایل دهه ۱۹۷۰ که مدل HMM ارائه شد در نظر بگیریم و یا از دهه ۱۹۷۰ که HMM پیاده‌سازی شد و یا حتی از سال ۱۹۸۹ که رابینر مقاله مفصل خود راجع به HMM را نوشت[43]، باید بپذیریم که از عمر HMM دیرزمانی می‌گذرد. سیستم‌های تشخیص صحبت در برخی حیطه‌ها به کار گرفته شده‌اند ولی هنوز به‌جایی نرسیده‌اند که واقعا درآمدزا باشند. شاید یکی از دلایلی که سیستم‌های HTK و SPHINX اکنون در اختیار عموم قرار گرفته‌اند همین قطع امید تولیدکنندگان آنها از درآمدزا شدن آنها باشد. از دیدی دیگر حتی اگر این سیستم‌ها روزی بتوانند صحبت را مانند انسان تشخیص دهند، مساله ما یعنی شناخت روش انسان در تشخیص صحبت هنوز به قوت خود باقی است.

از طرف دیگر ما دلایل زیادی برای اصلاح جهت‌گیری خود برای رسیدن به سیستم‌های بهینه (که معمولا ریاضیات به دنبال آن است) داریم. نظریه زبان‌ها و ماشین‌ها، منطق و نظریه پیچیدگی به ما می‌گویند که روش‌های ریاضی کارآمدی انسان را ندارند:

۴- **قضیه church**: نوشتن برنامه‌ای که بتواند درستی قضایای ریاضی را بررسی کند غیر ممکن است[15].

۵- **مسائل NP–سخت**: بسیاری از مسائل (از جمله مساله ارضاپذیری فرمول و هامیلتونی بودن گراف) NP-Complete هستند. این بدین معنی است که اگر P≠NP آنگاه افزایش سرعت سخت‌افزار هیچ تاثیر قابل توجهی در زمان حل این مسائل بوجود نخواهد آورد [12].

٦- **قضیه گودل**: گودل در قضیه ناتمامیت خود نشان داد که در ریاضیات نمی‌توان هر جمله درستی را اثبات کرد. به عبارت دیگر جملات درستی هستند که غیر قابل اثبات هستند. اما اثبات گودل ساختنی بود. او جمله‌ای را بیان کرد و نشان داد که این جمله درست است و در ضمن نشان داد که ریاضیات نمی‌تواند درستی این جمله را ثابت کند. بدین ترتیب او ادعا کرد که انسان از ماشین بالاتر است زیرا او توانسته بود درستی جمله را نشان دهد در حالی که ریاضیات نمی‌تواند.[45]

۷- **قضایای تصمیم‌ناپذیری**: مسائلی هستند که ثابت شده است نمی‌توان برنامه‌ای برای حل آنها نوشت. جالب این است که این مسائل معمولا دارای کاربردهای زیادی

هستند. برای مثال برنامه‌ای وجود ندارد که پایان‌پذیر بودن یک برنامه را بر روی هر ورودی چک کند. همچنین برنامه‌ای وجود ندارد که مبهم بودن گرامرها را تشخیص دهد[30].

کار نگارنده در این پایان‌نامه مجازاً به دو بخش تقسیم می‌شود:

۱- ترتیب دادن آزمایش‌هایی برای بررسی روش انسان در تشخیص صحبت

۲- تست ایده‌های به دست آمده از این آزمایش‌ها با نوشتن یک سیستم بازشناسی گفتار

اگر ما علم منطق فازی را با هدف ساختن انسان دنبال کنیم می‌توانیم بگوییم که بخش دوم کار نگارنده می‌تواند منجر به تولید یا درک ایده‌هایی در منطق فازی نیز بشود. در حقیقت بخش مهمی از کار نگارنده بررسی منطقی (و تا حدودی تجربی) دلایلی بوده است که باعث می‌شود یک سیستم فازی مبتنی بر قانون بهتر از یک سیستم احتمالی کار کند. حیطه دیگری که بررسی شده است، میزان اهمیت دقت در نتایجی است که نگارنده خود به دست آورده است و نیز نتایجی که با روش‌های متداول به دست آمده است.

نکته دیگری که متاسفانه دامنگیر بسیاری از علوم شده است این است که مسیر آموزش از مسیر تحقیق و کشف جدا شده است و در دانشگاه‌ها و کتب عکس روش تحقیق را آموزش می‌دهند.

۱- انسان قضیه‌ای را به روشی انسانی (که معمولا همراه با دست و پا زدن و با زحمت زیاد است) برای اولین بار اثبات می‌کند. افراد روی این روش کار می‌کنند و اثبات‌هایی ارائه می‌دهند که به نظر زیباتر هستند ولی یک محقق برای بار اول اثبات نمی‌تواند چنین اثباتی را بیاورد. اساتید نیز این نوع اثبات‌ها را درس می‌دهند و کتاب‌ها نیز هر چند وقت یک بار اثبات‌های قدیمی را با این نوع اثبات‌ها عوض می‌کنند.

۲- انسان در طول زندگی زبان‌های مختلف و نویزها و ... را یاد می‌گیرد. در هنگام ساختن سیستم پردازش صوت واج‌ها و انواع نویزها و ... را به سیستم می‌دهیم. به عبارت دیگر انسان مبتنی بر داده کار می‌کند و ما سیستم را مبتنی بر مدل می‌سازیم.

۳- پس از اینکه پژوهشگران علم را کشف می‌کنند، نویسندگان کتب درسی آنها را به صورت کتاب در می‌آورند و به دانشجویان یاد می‌دهند. دانشجو باید مسیر برعکس پژوهشگر را بپیماید و هیچ تلاشی برای آموزش روش کشف این علوم صورت نمی‌گیرد.

در این پایان‌نامه سعی می‌شود که گامی در جهت تولید سیستم تشخیص صحبت به روشی مشابه انسان برداشته شود. خواننده با بحث‌های بالا باید متوجه شده باشد که چنین سیستمی بهینه نیست.

اهداف ما در این پایان‌نامه را می‌توان چنین برشمرد:

الف- ترتیب دادن آزمایش‌هایی برای بررسی روش انسان در تشخیص صحبت

قبل از ارائه یک روش برای بازشناسی گفتار لازم است ما با ویژگی‌های سیگنال صحبت و همچنین روش انسان در تشخیص صحبت آشنا شویم.

ب- تولید سیستمی که قابل تفسیر باشد

روش‌های متداول بازشناسی گفتار مانند مدل مخفی مارکوف و شبکه‌های عصبی اجازه درک مناسبی از نحوه عملکرد سیستم را نمی‌دهند. تجربه همکاری نگارنده با شرکت عصر گویش نیز نشان می‌داد که تنها راه بررسی درستی یک ایده در چنین سیستم‌هایی افزایش نتیجه است. در این تحقیق سعی می‌شود سیستمی تولید شود که قابل تفسیر باشد و بتوان از تجربیات انسان در کار با طیف‌نگار در بررسی صحت عملگرد سیستم بهره برد. ما برای رسیدن به سیستمی که قابل تفسیر باشد ویژگی‌ها را در بخش‌ها استخراج می‌کنیم.

ج- مشابهت با انسان

انسان دارای قابلیت‌هایی در تشخیص صحبت است که هیچ یک از سیستم‌های متداول این قابلیت‌ها را ندارند. مثال‌هایی از این قابلیت‌ها عبارتند از:

۱- یادگیری صحبت و زبان پس از به دنیا آمدن و بدون هیچ اطلاع قبلی

۲- تشخیص صحبت در محیط‌های نویزی.

۳- تشخیص صحبت با قطعیت. در صورتی که گوینده خوب صحبت کند ما می توانیم با قطعیت اعلام کنیم که آنچه فهمیده‌ایم همان است که گوینده گفته است.

ما در این پایان‌نامه سعی می‌کنیم گامی در جهت نزدیک شدن به این قابلیت‌ها برداریم. برای مثال برای آنکه یادگیری اولیه کودک ممکن باشد از بخش‌بندی خودکار استفاده می‌کنیم و برای رسیدن به قطعیت از منطق فازی بهره می‌گیریم. همچنین از تجربه محققین علم تشخیص صحبت در مورد روش انسان در بازشناسی گفتار نیز بهره می‌گیریم.

د- تولید سیستمی مبتنی بر نظریه امکان

به نظر ما منطق امکان به طرز تفکر انسان نزدیک‌تر است. اثبات‌هایی مبنی بر بهینه بودن سیستم بازشناسی احتمالی وجود دارد. اما همانطور که در مقدمه گفته شد، سیستم‌های بهینه نمی‌توانند به قابلیت انسان دست یابند. همچنین نظریه امکان قابلیت مدل‌سازی جهل را دارد که برای رسیدن به قابلیت انسان ضروری است.

سید کمال الدین غیاثی شیرازی ۱۳۸۳/۹/۲۱

فصل ۱
مروری بر کارهای انجام شده با هدف شباهت به انسان

یافتن فضای ویژگی مناسب با بررسی میزان فهمیدنی بودن صحبت

میزان فهمیدنی بودن صحبت با اطلاعات بسیار کم فرکانسی

میزان فهمیدنی بودن صحبت با وجود عدم همزمانی طیفی

میزان فهمیدنی بودن صحبت پس از فیلتر شدن خط سیر ویژگیها

ویژگیها و نمایش دانش قابل تفسیر و مبتنی بر دانش طیف نگار

شباهت به انسان

بررسی غیر دقیق روش انسان در تشخیص صحبت

استخراج ویژگی در بازههای حدود 200ms

طیف مدولاسیون

سیستمهای بازشناسی با هدف شباهت به انسان

منطق فازی

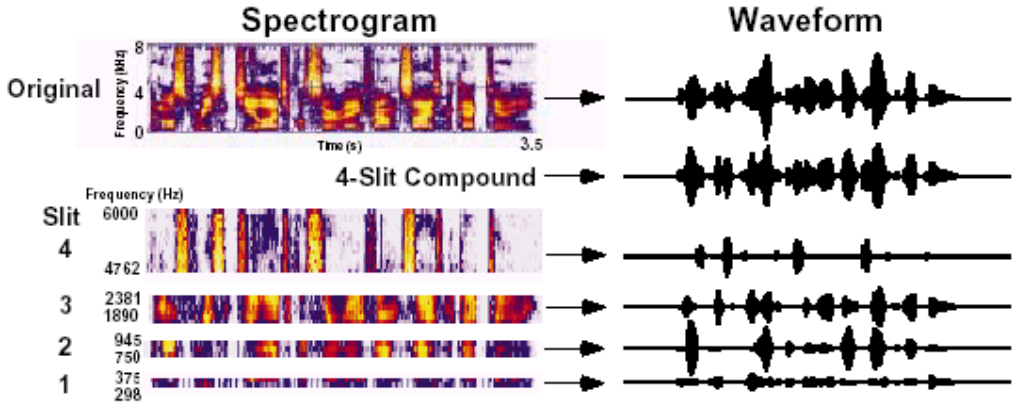
۱-۱) یافتن فضای ویژگی مناسب با بررسی میزان فهمیدنی[1] بودن صحبت

یکی از فازهای این پروژه شناخت روش انسان در تشخیص صحبت است. یک روش مناسب برای رسیدن به این شناخت، ایجاد تغییر در سیگنال صحبت و بررسی تاثیر آن بر میزان فهمیدنی بودن صدا است. در این بخش برخی کارهایی را که در این زمینه انجام شده است بررسی می‌کنیم.

۱-۱-۱) میزان فهمیدنی بودن صحبت با اطلاعات بسیار کم فرکانسی

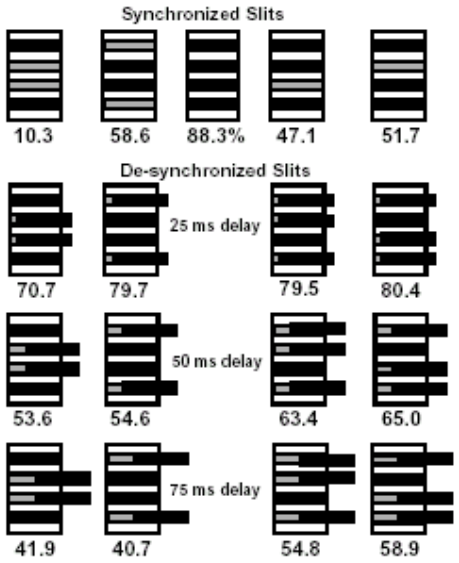
آقای گرینبرگ یکی از کسانی است که آزمایش‌های زیادی را بر روی میزان فهمیدنی بودن صحبت در شرایط متفاوت انجام داده است. یکی از این آزمایش‌ها [20] عبور دادن سیگنال صحبت از ۴ فیلتر مطابق شکل ۱ است. همانطور که دیده می‌شود خیلی از فرکانس‌ها در هیچ فیلتری حضور ندارند.

قابلیت تشخیص انسان در مورد سیگنال حاصل از مجموع این ۴ فیلتر برابر ۹۰٪ است. این درحالی است که از ۸۰۰۰ هرتز اطلاعات فرکانسی تنها از ۲۰۰۰ هرتز آن استفاده شده است. در صورتی که تنها اطلاعات دو یا سه فیلتر ترکیب شود، نرخ تشخیص بین ۶۰ تا ۸۳ درصد است و فهمیدنی‌بودن در هر فیلتر حداکثر ۹٪ است. این آزمایش‌ها نشان می‌دهند که قابلیت تشخیص صحبت در انسان مرهون ترکیب اطلاعاتی است که از فرکانس‌های مختلف به دست می‌آید.



شکل ۱: نمایش‌های طیف‌نگار و زمانی یک جمله و تجزیه آن به slit ها.

[1] Intelligibility

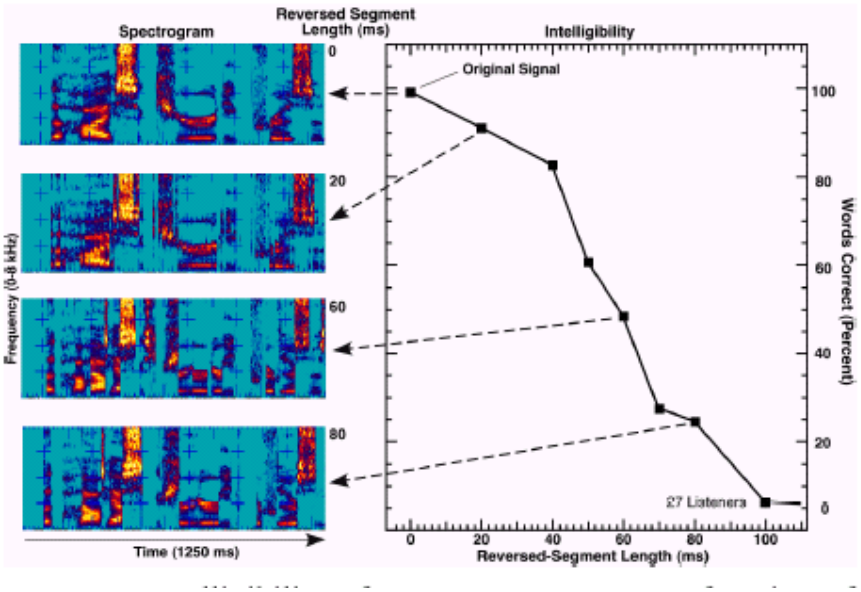


شکل ۲: تاثیر عدم هم‌زمانی در فهمیدنی بودن جملات تشکیل شده از ۴ slit.

۱-۱-۲) میزان فهمیدنی بودن صحبت با وجود عدم همزمانی طیفی

در بخش قبل دیدیم که می‌توان با استفاده از ۴ فیلتر که تنها ۲۵٪ فرکانس‌ها را می‌پوشانند به نرخ ۹۰٪ در تشخیص توسط انسان دست یافت[20]. افزونگی زیادی در منابعی که انسان‌ها از آن برای تشخیص صحبت استفاده می‌کنند وجود دارد. این نمایش ساده به ما اجازه می‌دهد که افزونگی اطلاعات را از بین ببریم و بتوانیم تاثیر ویژگی‌های مختلف را در بازشناسی انسان بررسی کنیم.

یکی از آزمایش‌های دیگر بر روی روش انسان در بازشناسی گفتار متوجه اهمیت همزمانی بین اطلاعات در فرکانس‌های مختلف است. همانطور که در شکل ۱ دیده می‌شود، پس از عبور سیگنال از ۴ فیلتر به ۴ سیگنال جدید می‌رسیم. در مقالات [50][4] [20] میزان اهمیت هم‌زمانی بین اطلاعات در این ۴ سیگنال بررسی شده است. شکل ۲ عدم همزمانی‌های مختلف و میزان فهمیدنی بودن صحبت در آنها را نشان می‌دهد. همچنین در [21] با معکوس کردن سیگنال صحبت در بازه‌های 0 تا 180ms نشان داده شده است که اطلاعات فاز طیف مدولاسیون نیز در تشخیص صحبت مؤثر است. شکل ۳ فهمیدنی بودن سیگنال صحبت را در طول‌های مختلف معکوس کردن سیگنال نشان می‌دهد.



شکل ۳: فهمیدنی بودن جملات **TIMIT** بر حسب طول بازه معکوس شده.

۱-۱-۳) میزان فهمیدنی بودن صحبت پس از فیلتر شدن خط سیر ویژگی‌ها [5]

کارهای زیادی نشان داده‌اند که استفاده از اطلاعات زمانی نتایج سیستم بازشناسی را بالا می‌برد. مثال ساده استفاده از اطلاعات زمانی همان محاسبه مشتق اول و دوم در ضرایب MFCC است. روش‌های پیچیده‌تر محاسبه اطلاعات زمانی شامل CMS[1]، ورودی‌های چندبرداره، RASTA و طیف مدولاسیون است. یکی از سوالاتی که مطرح می‌شود بررسی تاثیر فیلترهای زمانی بر روی قابلیت تشخیص سیگنال صحبت است. در مقاله [5] این آزمایش با استفاده از مدل RELP[2] [17]برای بازسازی سیگنال صحبت انجام شده است.

نتایج این آزمایش‌ها نشان می‌دهند که فیلتر کردن ضرایب کپسترال LPC قابلیت تشخیص صحبت در انسان را در موارد زیر خیلی پایین نمی‌آورد:

۱- تغییرات کندتر از ۱ هرتز فیلتر شوند.

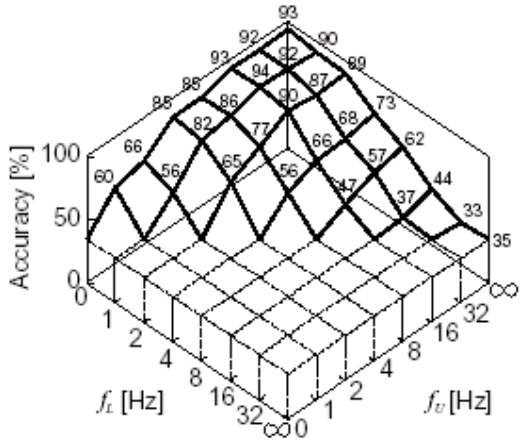
۲- تغییرات سریع‌تر از ۲۴ هرتز فیلتر شوند.

۳- تغییرات کندتر از ۱ هرتز و سریع‌تر از ۱۶ هرتز فیلتر شوند.

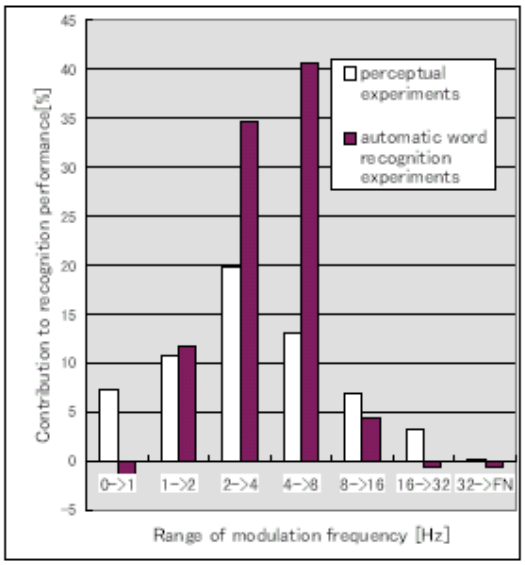
شکل ۴ میزان قابل فهمیدنی بودن سیگنال صحبت را پس از اعمال فیلترهای میان‌گذر مختلف نشان می‌دهد. جالب است که اعمال این فیلترهای زمانی تاثیر متفاوتی بر روی تشخیص انسان و ماشین دارد. شکل ۵ اهمیت فرکانس‌های مدولاسیون را در بازشناسی انسان و ماشین مقایسه می‌کند[35].

---

[1] Cepstral Mean Subtraction
[2] Residual-Excited Linear Prediction



شکل ۴: میزان قابل تشخیص بودن سیگنال صحبت پس از اعمال فیلتر میان‌گذر $[f_L,f_V]$.



شکل ۵: مقایسه اهمیت فرکانس‌های مدولاسیون برای سیستم‌های بازشناسی انسان و ماشین

۱-۲) ویژگی‌ها و نمایش دانش قابل تفسیر و مبتنی بر دانش طیف‌نگار

یکی از اهداف این پروژه تولید سیستمی قابل تفسیر است. دلیل ما برای تولید سیستمی قابل تفسیر، داشتن دیدی مناسب از سیستم برای مقایسه آن با روش انسان در تشخیص صحبت است. سیستم‌های قابل تفسیر و با نمایش دانش صریح معمولا مبتنی بر دانش خواندن طیف‌نگار و بخش‌بندی سیگنال صحبت هستند. امروزه کارهای زیادی در زمینه تولید سیستم‌های مبتنی بر دانش خواندن طیف-نگار انجام نمی‌شود. بیشترین مقالات مرتبط با این مطلب مربوط به ردیابی فرمنت‌ها[1] است که این نیز معمولا

[1] Formant Tracking

برای تولید سیگنال صحبت و نه بازشناسی آن استفاده می‌شود.

یکی از کارهای مهم در زمینه تولید سیستمی با نمایش دانش صریح توسط آقای Philip Schmid انجام شده است. ایشان در تز دکترای خود [46]کار نسبتا جامعی را در زمینه تولید یک سیستم بازشناسی قابل تفسیر مبتنی بر دانش خواندن طیف‌نگار انجام داده است. می‌دانیم که دانش ما در مورد صحبت معمولا به شکل رخدادهایی شنیداری در مرز بین واج‌ها و محل، شکل و خط سیر فرمنت‌ها بیان می‌شود. ایشان با الهام از این مطلب سیستمی مبتنی بر بخش‌بندی ارائه کرده است و در تشخیص واکه‌ها از اطلاعات فرمنت‌ها به‌تنهایی و در کنار ویژگی‌های MFCC استفاده کرده است. ایشان در بخش اول کار خود یک سیستم بخش‌بندی تولید کرده است که حاصل آن شبکه‌ای از فرضیه‌های بخش‌بندی است. ایشان توانسته‌اند با دقت بالایی گروه کلی هر بخش را نیز تعیین کند. بدین ترتیب از دو سیستم بازشناسی مجزا برای تشخیص حروف صدادار و بی‌صدا استفاده شده است. برای تشخیص حروف بی‌صدا از همان ویژگی‌های MFCC استفاده شده است. اما برای تشخیص حروف صدادار از اطلاعات فرمنت‌ها استفاده شده است. برای کاهش خطای الگوریتم ردیابی فرمنت، یک الگوریتم مستحکم مبتنی بر تشخیص N-بهترین تفسیر سازگار از فرمنت‌ها ارائه شده است. بدین ترتیب ایشان توانسته است تشخیص تفسیر صحیح از فرمنت‌ها را تا مراحل تشخیص واج و کلمه به تعویق بیاندازد. برای استخراج ویژگی در واکه‌ها از اطلاعات دامنه فرمنت، پهنای فرمنت، pitch و طول واکه استفاده شده است. نتایج بازشناسی مبتنی بر این روش با نتایج بازشناسی مبتنی بر ویژگی‌های MFCC قابل مقایسه است. همچنین ترکیب این ویژگی‌ها با ویژگی‌های MFCC نتایج بازشناسی را بالا برده است.

کار مهم دیگری که در جهت تولید یک سیستم قابل تفسیر انجام شده است، سیستم SPREXII است که اکنون به صورت یک محصول توسط شرکتی به همین نام عرضه می‌شود. SPREXII یک سیستم خبره برای تشخیص گفتار پیوسته است. در این سیستم نیز سعی شده است دانش خواندن طیف‌نگار به صورت قوانینی فازی در سیستم خبره گنجانده شود. هدف از استفاده از مفاهیم فازی، کاهش خطایی که به علت اعمال مقادیر آستانه‌ای ایجاد می‌شود و نیز ارتباط راحت‌تر شخص خبره با سیستم است. در این سیستم از روش قاب[45]، که یک روش نمایش دانش شیءگرا است، استفاده شده است. بخش‌بندی در این سیستم بر مبنای واحدهای شنیداری است و رابطه‌ای با واحدهای آوایی مانند واج ندارد. هر واحد شنوایی یک تغییر در سیگنال صحبت را نشان می‌دهد. این نوع بخش‌بندی با بخش‌بندی متداول که به دنبال یافتن حالات ایستای[1] سیگنال صحبت است متفاوت است.

نگارنده کار دیگری را که به این صراحت دانش را مدل کرده باشد نمی‌شناسد. اما در بسیاری از کارها تلاش شده است که از واحدهای با مفهوم‌تر استفاده شود. یکی از این کارها استفاده از واحدهای آوایی

[1] Stationary

شبه سیلاب در تشخیص کلمه است[32]. در این کار به‌جای واحد آوایی واج از واحد سیلاب برای بازشناسی ماه‌های سال استفاده شده است و نشان داده شده است که نتایج هر دو روش یکسان است. همچنین نشان داده شده است که بخش‌بندی خودکار برای تشخیص واحدهای شبه سیلاب قابل انجام است ولی تولید سیستمی برای بخش‌بندی در حد واج میسر نیست.

۱-۳) شباهت به انسان

همانطور که از نام این پایان‌نامه برمی‌آید، هدف اصلی ما در این پروژه شناخت بیشتر انسان و تولید سیستمی با شباهت بیشتر به انسان است. آقای هرمانسکی در [27] در مورد اینکه آیا یک سیستم بازشناسی باید از انسان الهام بگیرد یا خیر بحث می‌کند. یک نظر این است که هواپیما بال نمی‌زند ولی پرواز می‌کند، پس ما نیز نباید برای تولید سیستم بازشناسی گفتار الزاما از انسان الهام بگیریم. ایشان سه نفر را با هم مقایسه می‌کنند:

۱- شخصی که به تقلید از انسان به دست و پای خود بالی آویخته تا بتواند با بال زدن پرواز کند.

۲- شخصی که ماشینی سنگین ساخته است و با علاقه مشغول کار بر روی آن است.

۳- شخصی که نیروی برنولی را کشف کرده است و می‌داند که این قانون است که پرنده‌ها را در هوا نگه می‌دارد. او اکنون می‌تواند کایت را بسازد که بال هم نمی‌زند.

ما نیز با ایشان هم‌عقیده هستیم و بیشتر به دنبال آزمایش‌هایی بر روی انسان هستیم که بتوان از آنها پندهای مناسبی گرفت. به نظر ما هم علم تشخیص صحبت نیازمند کشف روش انسان در بازشناسی گفتار است و هم تقلید کورکورانه از مسیر انسانی درست نیست. در ادامه برخی نظرات در مورد روش تشخیص صحبت در انسان را بررسی می‌کنیم. لازم به ذکر است که روش‌های متداول استخراج ویژگی مانند MFCC نیز تاحدود زیادی از انسان الهام گرفته‌اند.

۱-۳-۱) بررسی غیر دقیق روش انسان در تشخیص صحبت

اطلاعات دقیقی که ما در مورد انسان داریم بسیار کم است. همچنین معمولا آزمایش‌های شنوایی که انجام می‌شوند برای نتایجی که گرفته می‌شوند کافی نیستند. در این بخش مقالاتی را بررسی می‌کنیم که با مطالعه آزمایش‌های انجام شده و برخی فرض‌ها مدل‌هایی از روش تشخیص صحبت در انسان ارائه کرده‌اند.

آقای آلن در مقاله خود [3] به توضیح کار آقای فلچر[1] در بررسی تشخیص صحبت در انسان می‌پردازد. او نتیجه می‌گیرد که سیستم‌های بازشناسی گفتار باید در جهات زیر اصلاح شوند:

---

[1] Fletcher

۱- پردازش‌های زمانی جایگزین پردازش‌های فرکانسی شوند.
۲- اطلاعات هر باند فرکانسی مستقل از دیگر باندها در بازشناسی استفاده شود و تنها ارتباط بین فرکانس‌های مجاور در نظر گرفته شود.
۳- از دقت موجود در همزمانی بین ویژگی‌ها در فرکانس‌های مختلف کاسته شود.

آقای گرینبرگ نیز در [18] دلایلی در دفاع از ادعاهای زیر می‌آورد:

۱- فرآیند تشخیص صحبت یک فرآیند پایین به بالا نیست. منظور از فرآیند پایین به بالا لایه‌های محاسبه طیف، تشخیص واج، تشخیص سیلاب، تشخیص کلمه و جمله است.
۲- فرض فلچر در مورد استقلال اطلاعات باندهای فیلتر صحیح نیست. دلیل ایشان آزمایش‌هایی است که در مورد اثر ترکیب اطلاعات باندهای مختلف در فهمیدنی بودن سیگنال صحبت انجام شده است[1].
۳- زمان اهمیت زیادی دارد و باید اطلاعات طیفی را در فواصل بزرگتری (مثلا 300ms) بررسی کرد. این زمان متناسب با طول سیلاب است.
۴- اطلاعات بخش‌بندی اهمیت زیادی در تشخیص صحبت انسان دارد. دلیل این مدعا این است که افرادی که دارای مشکل شنوایی هستند از دیدن طیف سیگنال صحبت کمک زیادی می‌گیرند.
۵- سیلاب در تشخیص صحبت نقش اساسی دارد. دلایل زیادی وجود دارد که واحد اصلی در درک صحبت سیلاب است، نه واج.
۶- طیف مدولاسیون نمایش مقاوم‌تری برای سیگنال صحبت است (نسبت به طیف‌نگار).

آقای هرمانسکی نیز در [27]به دلایل عدم موفقیت مدل‌های شنوایی[2] ارائه شده می‌پردازد و از ایده‌های زیر دفاع می‌کند.

۱- تجدید نظر در مورد تحلیل سیگنال در زمان‌های کوتاه برای استخراج ویژگی و روی آوردن به بازه‌های زمانی بزرگ‌تر در حد 200ms و توجه به ساختار صحبت در زمان.
۲- توجه به آنچه انسان‌ها نمی‌شنوند و ارائه مدل‌هایی که توانایی نشنیدن را داشته باشند.
   a. کاهش دقت در فرکانس‌های بالاتر.
   b. پوشش زمانی[3].

---

[1] نگارنده معتقد است که این دلیل برای مدعا درست نیست. زیرا کار آقای فلچر بر روی articulation بوده است در حالی که آزمایش‌های آقای گرینبرگ بر روی intelligibility انجام شده است.

[2] Auditory model
[3] Temporal masking

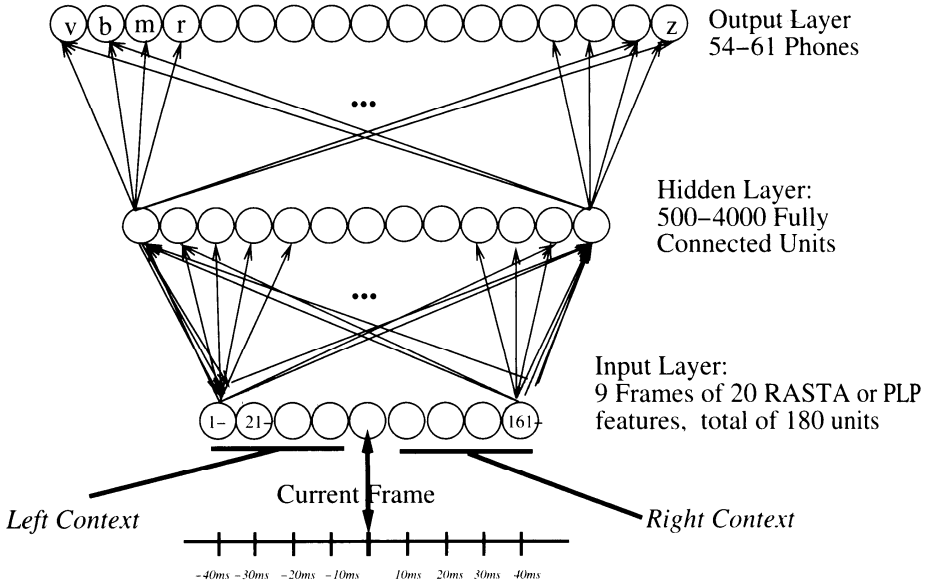
۳- بازشناسی مبتنی بر اطلاعات جزئی و نادیده گرفتن اطلاعات خراب.

۴- کاهش دقت در طیف فرکانسی

در [62] نیز نمایشی برای صوت با ایده نرمال‌سازی طیف ارائه شده و استحکام آن نسبت به نویز نشان داده شده است.

۱-۳-۲) استخراج ویژگی در بازه‌های حدود 200ms

کارهای زیادی بر روی استخراج ویژگی در بازه‌های زمانی حدود 200ms انجام شده است و نشان داده شده است که این کار استحکام سیستم بازشناسی را بالا می‌برد. روش‌های مبتنی بر بخش‌بندی که قبلا مرور شد نمونه‌هایی از روش‌های استخراج ویژگی در زمان و فرکانس هستند. علاوه بر این، سیستم‌های ترکیبی شبکه‌های عصبی و مدل مخفی مارکوف نیز از اطلاعات زمانی-فرکانسی بهره می‌گیرند. شکل ۶ نمونه‌ای از یک شبکه عصبی که ویژگی‌ها را در بازه‌های 90ms استخراج می‌کند نشان می‌دهد. در بخش بعد طیف مدولاسیون را معرفی می‌کنیم و خواهیم دید که در این روش نیز ویژگی‌ها در زمان و فرکانس استخراج می‌شوند. همچنین می‌توان نشان داد [25] که روش RASTA نیز به نوعی از اطلاعات زمانی بهره می‌گیرد.

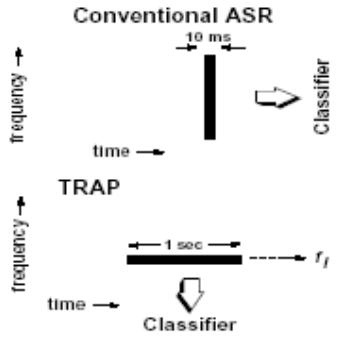


شکل ۶: یک شبکه عصبی نمونه برای تخمین زدن احتمال هر واج بر اساس ویژگی‌هایی که در 90ms استخراج می‌شوند.

آقای هرمانسکی در [28] و [29] روش TRAPS را در استخراج ویژگی ارائه می‌دهد. در این روش به‌جای ترکیب ویژگی‌ها در فرکانس، ابتدا ویژگی‌های هر باند فرکانسی در زمان ترکیب می‌شوند. در آزمایش انجام شده ۲۹ واج وجود داشته است. برای هر باند فرکانسی ابتدا یک شبکه عصبی MLP احتمال تعلق بردار ورودی (که شامل یک ثانیه اطلاعات از آن باند است) را به هر یک از ۲۹ واج تخمین

می‌زند. سپس نتایج این ۱۵ باند به عنوان ورودی به شبکه عصبی ترکیب کننده نتایج داده می‌شود. بدین ترتیب تعداد ورودی‌های شبکه عصبی دوم برابر ۴۳۵=۲۹×۱۵ است. بدین ترتیب نشان داده شده است که در برخی محیط‌های نویزی و در تشخیص واج نتایج بازشناسی بالا می‌رود. جالب است که در این روش نتایج بازشناسی کلمه بالا نرفته است.

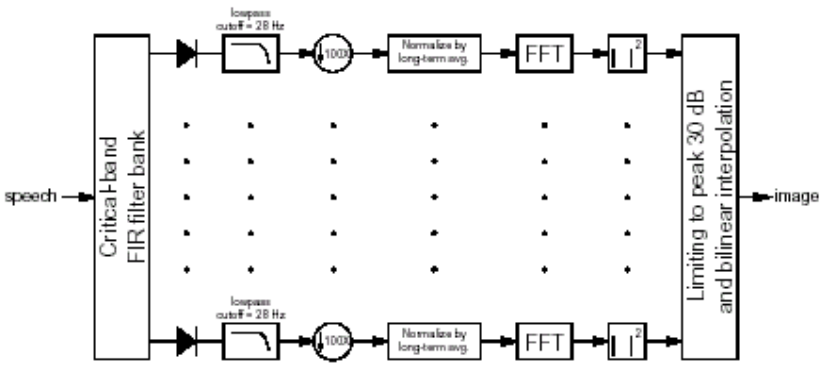
همچنین مقالات زیادی سعی کرده‌اند در ماتریسی زمانی–فرکانسی به استخراج ویژگی بپردازند [47] [56] [57]. نتایج کار این مقالات نیز نشان می‌دهد که نرخ بازشناسی در محیط‌های نویزی به شدت بالا رفته است.



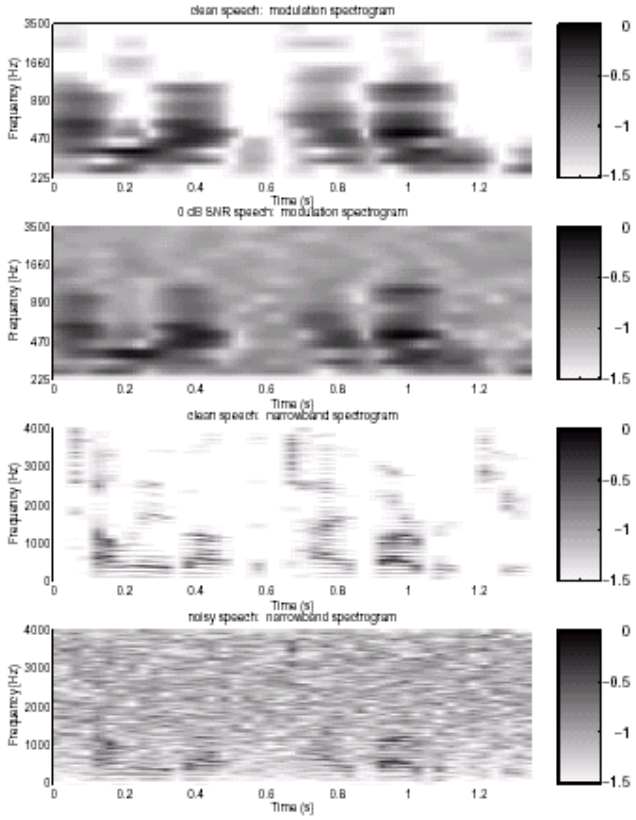
شکل ۷: نحوه استخراج ویژگی در بازشناسی گفتار متداول و TRAP.

۱-۳-۳) طیف مدولاسیون

طیف–نگار نمایشی از سیگنال صحبت است که هیچ بهره‌ای از خواص سیگنال صحبت نبرده است. به عبارت دیگر طیف–نگار صدا را نشان می‌دهد، نه صحبت را. یکی از روش‌های افزایش استحکام سیستم تشخیص صحبت حذف اطلاعات اضافی غیر مربوط به سیگنال صحبت است. آقای گرینبرگ با توجه به آزمایش‌هایی که بر روی انسان انجام داد به این نتیجه رسید که نمایش طیف–نگار دارای دقت زیادی است و باید نمایشی با دقت کمتر را به جای آن به‌کار برد. همچنین آزمایش‌های مختلف نشان داده‌اند که انسان تنها به فرکانس‌های مدولاسیون بین ۱ تا ۱۶ هرتز توجه دارد. بدین ترتیب ایشان نمایش طیف مدولاسیون را به عنوان جانشینی برای نمایش طیف‌نگار پیشنهاد دادند [19] و نشان دادند که این نمایش دارای استحکام بسیار بیشتری است. نحوه محاسبه طیف مدولاسیون در شکل ۸ نشان داده شده است.



شکل ۸: نمایش مراحل محاسبه طیف مدولاسیون



شکل ۹: مقایسه‌ای بین نمایش طیف مدولاسیون و نمایش طیف‌نگار باند باریک[1] در سیگنال نویزی و تمیز.

به نظر می‌رسد که استفاده از یکسو کننده تزئینی است زیرا خروجی بانک فیلتر انرژی است و مثبت است. پس از محاسبه ضرایب بانک فیلتر، خط سیر هر ویژگی در زمان از یک فیلتر پایین‌گذر با فرکانس قطع ۲۸ هرتز عبور می‌کند. تبدیل فوریه انتهایی در یک پنجره 250ms از نوع hamming محاسبه می‌شود. سپس انرژی موجود در فرکانس مدولاسیون ۴ هرتز برای فرستاده شدن به تصویر آماده می‌شود.

[1] Narrow-band spectrogram

در نهایت از دو مقدار آستانه‌ای استفاده می‌شود و انرژی به رنگ نگاشته می‌شود. ما تنها انرژی را در باندهای فیلتر داریم و باید راهی برای تولید شکلی مشابه طیف‌نگار پیدا کنیم. برای محاسبه انرژی در هر فرکانس و زمان از میان‌یابی دوخطی[1] [33] استفاده شده است. در این روش فرض می‌شود که مقدار تابع در ۴ نقطه مشخص است و مقدار تابع در یک نقطه بین آنها مد نظر است. تابع $f(x,y) = a_1 + a_2x + a_3y + a_4xy$ را دوخطی گویند زیرا با فرض ثابت بودن هر یک از متغیرها یک تابع خطی می‌شود. شکل ۹ نشان می‌دهد که نمایش طیف مدولاسیون نسبت به نمایش طیف‌نگار مقاوم‌تر است.

۱-۳-٤) سیستم‌های بازشناسی با هدف شباهت به انسان

سیستم‌های بازشناسی عموما مبتنی بر مدل مخفی مارکوف و شبکه عصبی هستند که شباهتی به روش بازشناسی انسان ندارند. همانطور که گفتیم آقای آلن با توجه به آزمایش‌های آقای فلچر مدعی شده بود که خطاهایی که در باندهای مختلف رخ می‌دهند از یکدیگر مستقل هستند [3]. به عبارت دیگر اگر سیگنال صحبت را از دو فیلتر مکمل عبور دهیم، خطای فهمیدن[2] صحبت در سه سیگنال از رابطه زیر تبعیت می‌کند:

$$e = e_1.e_2$$

که در آن e خطای فهمیدن سیگنال اولیه و e_1 و e_2 خطای فهمیدن سیگنال‌های فیلتر شده را نشان می‌دهد. این بدین معنی است که خطاهای باندهای مختلف مستقل هستند. به بیانی دیگر، ما زمانی یک واج را غلط می‌شنویم که قادر به تشخیص صحبت در هیچ یک از دو سیگنال فیلتر شده نباشیم.

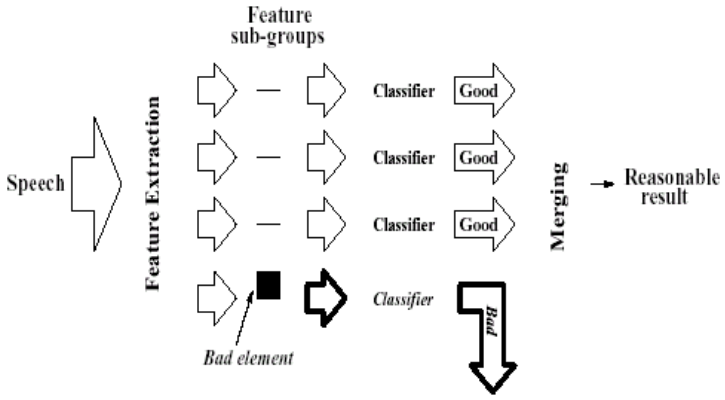
آقای هرمانسکی با الهام از کار آقای فلچر بازشناسی مبتنی بر زیرباندها[3] را مطرح کرد[54]. در این روش سعی می‌شود در شرایطی که برخی از باندهای فرکانسی خراب شده‌اند، از اطلاعات بقیه باندها برای تشخیص صحبت استفاده شود. طرح کلی سیستم بازشناسی مبتنی بر زیرباند در شکل ۱۰ نشان داده شده است. در این روش احتمال مشاهده به شرط مدل در تمام باندها و برای تمام واج‌ها جداگانه محاسبه می‌شود. سپس این احتمال‌ها با هم ترکیب می‌شوند تا نتیجه نهایی به دست آید. نتایج به دست آمده نشان می‌دهند که در محیط تمیز نتایج این روش بازشناسی حداقل به اندازه روش‌های متداول است. همچنین نشان داده شده است که این روش در برخی محیط‌های نویزی بهتر از روش‌های متداول عمل می‌کند.

---

[1] Bilinear Smoothing

[2] لغت intelligibility برای درصد تشخیص کلمات و لغت articulation برای درصد تشخیص سیلاب‌های بی‌معنی استفاده می‌شود. در اینجا منظور articulation است.

[3] Sub-Band Based Recogniztion



شکل ۱۰: مدل بازشناسی مبتنی بر زیرباند

۱-٤) منطق فازی

طرفداران منطق فازی (از جمله نگارنده) معتقدند که منطق فازی به منطق انسان‌ها نزدیک‌تر است. به همین دلیل ما در این پروژه تصمیم گرفتیم که یک سیستم فازی تشخیص صحبت تولید کنیم. در این بخش کارهای مشابهی را که در زمینه تشخیص صحبت فازی انجام شده است بررسی می‌کنیم. البته ما معتقدیم که کارهای دیگران عموما محدود به جایگزین کردن معادل‌های فازی در الگوریتم‌های شناخته شده بوده است. برای مثال به‌جای k-means از c-means و به‌جای مخلوطی از توزیع‌های احتمالی از انتگرال فازی استفاده شده است و اساس تفکر که مبتنی بر نظریه احتمال است تغییر نکرده است. در حقیقت ما معتقدیم بسیاری از کارهایی که به نام فازی انجام نشده‌اند، مانند کارهای آقای هرمانسکی و گرینبرگ، فازی و بسیاری از کارهایی که به اسم فازی انجام می‌شود از ماهیت فازی تهی هستند. برای مثال در کارهای آقای گرینبرگ تلاش در جهت کاهش دقت به وضوح به چشم می‌خورد که با تفکر فازی سازگار است. حال برخی مقالات مرتبط با تشخیص صحبت فازی را مرور می‌کنیم.

در بخش‌های قبل با سیستم خبره فازی SPREXII آشنا شدیم. در حقیقت این سیستم نسخه فازی سیستم SPREX است و نرخ بازشناسی در آن ۱۰٪ بیش از سیستم خبره غیر فازی است. علت این امر را می‌توان کاهش خطای ناشی از اعمال آستانه دانست.

بسیاری از کارهای فازی در تشخیص صحبت به صورت ترکیب با HMM هستند. می‌دانیم که الگوریتم باوم–ولش را می‌توان یک الگوریتم بیشترین شباهت و یا تعمیمی از الگوریتم k-means دانست[31]. بدین ترتیب برخی از محقیقین الگوریتم fuzzy c-means را که یک الگوریتم فازی است به‌جای k-means استفاده کرده‌اند. در [55] روش جدیدی به نام SMM[1] ارائه شده است و نشان داده شده است که علاوه بر الگوریتم بیشترین شباهت برای یادگیری مدل می‌توان از الگوریتم fuzzy c-means نیز

[1] State Mixture Modeling

استفاده کرد. آزمایش‌های ارائه شده در این مقاله از افزایش توام سرعت و نرخ بازشناسی حکایت دارد. مقاله [41] نیز الگوریتم‌های k-means و c-means را مقایسه می‌کند. در این مقاله نتایج الگوریتم k-means بالاتر است، اما نشان داده شده است که نتایج بازشناسی در سگمنت‌بندی مناسب در c-means بیشتر است.

نوع دیگر ترکیب فازی با HMM در تخمین تابع شباهت است. این کار مشابه ترکیب سیستم‌های شبکه عصبی با HMM است. در [9] از انتگرال فازی به‌جای شبکه عصبی برای تخمین احتمال استفاده شده است و نتایجی مشابه شبکه عصبی گرفته شده است. حسن استفاده از منطق فازی به‌جای شبکه عصبی قابل تفسیر بودن آن است.

یک نمونه از کارهای انجام شده در زمینه تشخیص صحبت که به روش ما شباهت دارد در [52] آمده است. این سیستم از شبکه عصبی HRCNN استفاده می‌کند. این شبکه عصبی فضای بردارهای ویژگی را به مکعب‌هایی (به ابعاد بردار ویژگی) تقسیم می‌کند. همچنین این شبکه عصبی بر مبنای عملگرهای Max و Min کار می‌کند. هر مکعب نماینده چندین داده آموزشی است که در یک گروه قرار دارند. در هنگام تست، فضای مربوط به هر مکعب به شکل فازی پخش می‌شود و نقاطی که داخل مکعب نیستند نیز با یک درجه تعلق داخل مکعب فرض می‌شوند. این سیستم برای تشخیص کلمات مجزا استفاده شده است و یکی از حرف‌های اصلی آن حذف تطابق زمانی موجود در HMM و DTW است. در حقیقت این روش اجازه می‌دهد که هر قاب کلمه تست با تمام قاب‌های آموزشی مقایسه شود و هیچ ترتیب زمانی بین بردارهای ویژگی قائل نمی‌شود.

در [11] نسخه‌ای فازی برای HMM به نام GFHMM[1] ارائه شده است. هدف از این مقاله جایگزینی منطق فازی به‌جای نظریه احتمال است. بدین ترتیب مدل مخفی مارکوف حالت خاص GFHMM در صورت استفاده از عملگرهای احتمالی است. نتایج بازشناسی در هر دو روش یکسان است اما محاسبات روش فازی کمتر است. در پیاده‌سازی GFHMM از نظریه امکان که دارای عملگرهای Max و Min است استفاده شده است.

[1] Generalized Fuzzy Hidden Markov Model

فصل ۲
مروری بر روش‌های متداول استخراج ویژگی

پیش پردازش‌های قبل از استخراج ویژگی

حذف مقدار ثابت DC

حذف پیش‌تاکید

پنجره‌بندی

ویژگی‌های MFCC **(**mel cepstrum**)**

ویژگی‌های PLP

ویژگی‌های RASTA

۲-۱) پیش پردازش‌های قبل از استخراج ویژگی

۲-۱-۱) حذف مقدار ثابت DC

در اولین مرحله ابتدا مقدار ثابت DC از سیگنال صحبت حذف می‌شود. بدین منظور از فرمول زیر استفاده می‌شود:

$$y[n] = x[n] - x[n-1] + 0.999 * y[n-1]$$

که در آن y[n] مقدار جدید سیگنال و x[n] سیگنالی است که می‌خواهیم مقدار DC آن حذف شود. ایده این است که مقدار سیگنال را همواره با ضرب کردن در عددی کوچک‌تر از ۱ به سمت صفر میل دهیم و همچنین تغییرات بین هر دو نمونه از سیگنال را نیز تقریبا حفظ کنیم.

۲-۱-۲) حذف پیش‌تاکید

انسان نسبت به فرکانس‌های مختلف به یک اندازه حساس نیست. بدین منظور سیگنال توسط معادله تفاضلی زیر فیلتر می‌شود ($0 \le k < 1$).

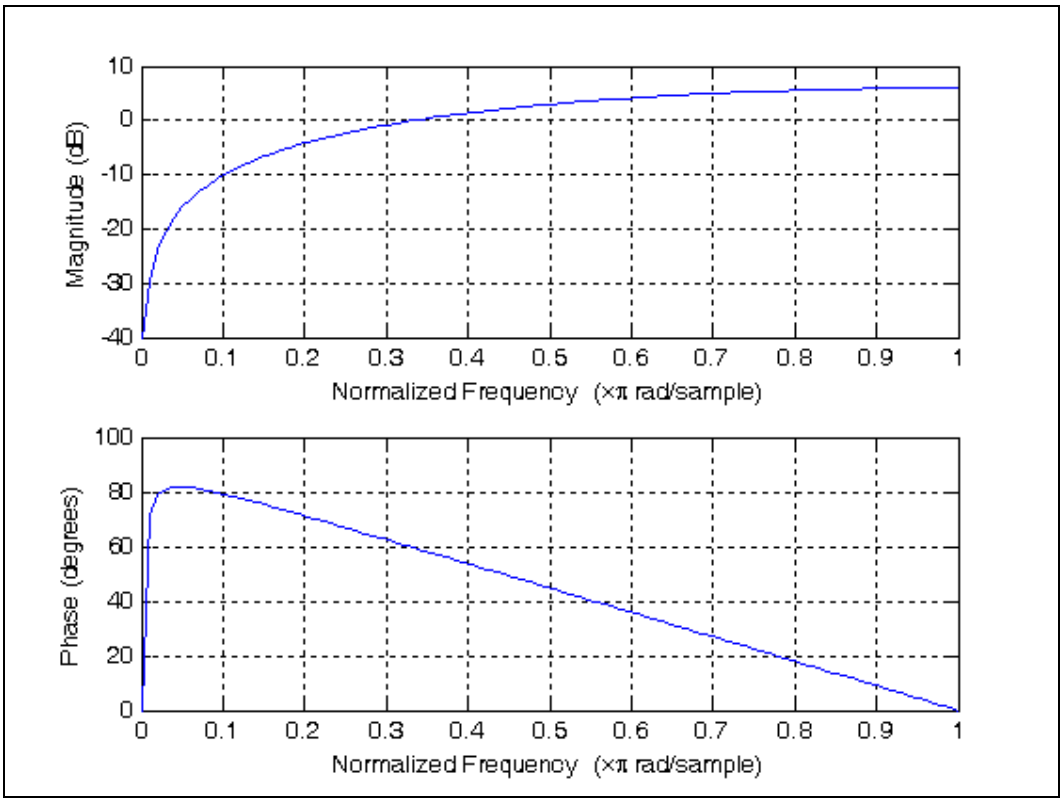
$$y(n) = x(n) - x(n-1)$$

معمولا k عددی بین ۰.۹۷ تا ۰.۹۹ است.

برای بررسی خواص فرکانسی این فیلتر از تبدیل Z استفاده می‌کنیم.

$$y(n) = x(n) - kx(n-1)$$
$$Y(z) = X(z) - kz^{-1}X(z)$$
$$Y(z) = (1 - kz^{-1})X(z)$$
$$H(z) = 1 - kz^{-1}$$

می‌دانیم که اگر متغیر z بر روی دایره واحد حرکت کند تبدیل z با تبدیل فوریه برابر می‌شود و بنابراین مفهوم فرکانس می‌دهد. با محاسبه دامنه و فاز H(z) بر روی نقاط دایره واحد می‌توان پاسخ فرکانسی این فیلتر را دید. شکل ۱۱ دامنه و فاز این فیلتر را برای k=0.99 نشان می‌دهد. همانطور که دیده می‌شود، این فیلتر فرکانس‌های پایین را تضعیف می‌کند و انرژی فرکانس‌های بالا را افزایش می‌دهد.



شکل ۱۱: دامنه و فاز فیلتر پیش‌تاکید با k=0.99

۲-۱-۳) پنجره‌بندی

می‌دانیم که هنگامی که تبدیل فوریه N نمونه را حساب می‌کنیم، در حقیقت تبدیل فوریه سیگنالی که از تکرار این N نمونه به دست می‌آید را حساب کرده‌ایم. از طرف دیگر می‌دانیم که تفاوت شدید بین مقدار دو نمونه مجاور منجر به ظهور فرکانس‌های بالا در تبدیل فوریه سیگنال می‌شود. فرض کنیم ما N نمونه از سیگنال صحبت را جدا کرده‌ایم و آن را x(n) نامیده‌ایم. اگر x(N-1)-x(0) زیاد باشد فرکانس‌های بالا در تبدیل فوریه سیگنال ظاهر خواهد شد. علت این است که همانطور که ذکر شد ما تبدیل فوریه را بر روی یک سیگنال متناوب محاسبه می‌کنیم. برای حل این مشکل هر قاب را در پنجره‌ای ضرب می‌کنند. شکل کلی این پنجره‌ها چنین است که مقادیر کناری سیگنال را به صفر نزدیک می‌کند. جدول ۱ پنجره‌های گوناگونی را که پیشنهاد داده شده‌اند نشان می‌دهد. در پردازش صحبت معمولا از پنجره‌های Hamming و Hanning استفاده می‌شود.

جدول ۱: پنجره‌های گوناگون

نام پنجره	فرمول ریاضی
Rectangular	$w(n) = 1, n = 0, \ldots, N-1$
Triangular	$w(n) = 2(n+0.5)/N$ $w(N-1-n) = w(n)$ $n = 0,1,\ldots, N/2$
Hanning	$w(n) = \frac{1}{2}\left[1 - \cos\frac{2(n+0.5)\pi}{N}\right],$ $n = 0,1,2,\ldots, N-1$
Hamming	$w(n) = 0.54 - 0.46\cos\frac{2(n+0.5)\pi}{N},$ $n = 0,1,2,\ldots, N-1$
Kaiser-Bessel	$w(n) = I_0(\pi\alpha\beta)/I_0(\pi\alpha)$ *where* $\beta = \sqrt{1-\left(\frac{2n+1}{N}-1\right)^2}$ $n = 0,1,2,\ldots, N-1$ $I_0(x) = J_0(jx) = \sum_{k=1}^{\infty}\left\{\frac{x^k}{k!2^k}\right\}^2$
Gausian	$w(n) = e^{0.5\left(\alpha\frac{k-\frac{N}{2}}{\frac{N}{2}}\right)^2},$ $\alpha \geq 2$ $n = 0,1,2,\ldots, N-1$
Blackman-Harris	$w(n) = 0.42 - 0.5\cos(\frac{2\pi n}{N-1}) + 0.08\cos(\frac{4\pi n}{N-1})$
Bartlett	$w(n) = \begin{cases} \frac{2n}{N-1}, & n \in \left[0, \frac{N-1}{2}\right] \\ 2-2\frac{n}{N-1}, & n \in \left[\frac{N-1}{2}+1, N-1\right] \end{cases}$

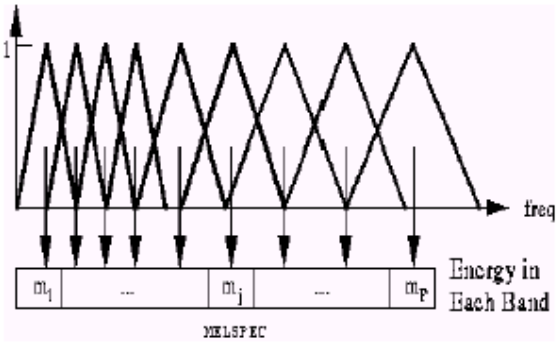
۲-۲) ویژگی‌های MFCC (*mel cepstrum*) *[17]*

این ویژگی‌ها یکی از متداول‌ترین ویژگی‌هایی است که در تشخیص صحبت استفاده می‌شود. معمولا ابتدا ۱۲ یا ۱۳ ویژگی استخراج می‌شود و سپس مشتق اول و دوم آنها هم به ویژگی‌ها اضافه می‌شود.

پس از حذف DC و انجام پیش‌تاکید تعدادی قاب به دست می‌آید که هر کدام شامل N نمونه است. سیگنال قاب i ام را با $x_i(n)$ نشان می‌دهیم. در ادامه بحث اندیس قاب را حذف می‌کنیم. سپس با استفاده از الگوریتم FFT تبدیل فوریه گسسته سیگنال x(n) که آن را با X(n) نشان می‌دهیم به دست می‌آید. چون مقادیر X(i) و X(N-i) مزدوج هستند عملا تنها مقادیر ۰ تا N/2 از سیگنال دارای اطلاعات هستند. در حقیقت X(0) مقدار ثابت DC را نشان می‌دهد و X(N/2) مؤلفه فرکانسی متناظر با نصف فرکانس نمونه‌برداری است. بدین ترتیب می‌توان X(0) تا X(N/2) را متناظر با فرکانس‌های ۰ تا نصف فرکانس نمونه‌برداری قرار داد. سپس توان دوم انرژی در هر مؤلفه فرکانسی محاسبه می‌شود. تا این مرحله انرژی در هر مؤلفه فرکانسی مشخص شده است.

جدول ۲: رابطه مل و بارک و فرکانس. فرکانس زاویه‌ای از ۰ تا π تغییر می‌کند و متناظر با 0Hz تا نصف فرکانس نمونه‌برداری است

رابطه فرکانس (هرتز) و مل	$F_{mel} = \frac{1000}{\log 2}\left[1 + \frac{F_{Hz}}{1000}\right]$
رابطه فرکانس (رادیان بر ثانیه) و بارک	$\Omega(\omega) = 6\ln\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right\}$



شکل ۱۲: نمایی از فیلترهای Mel و نحوه محاسبه آنها

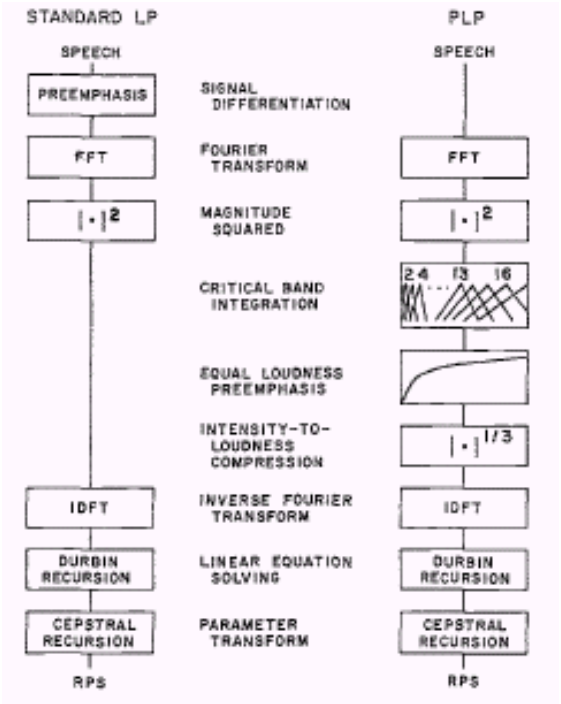
سپس انرژی در فیلترهای مل یا بارک محاسبه می‌شود و بدین ترتیب ضرایب بانک فیلتر به دست می‌آید.

جدول ۲ نحوه محاسبه فرکانس‌های مل و بارک را نشان می‌دهد. در هر دو حالت هدف شبیه‌سازی حساسیت انسان به فرکانس‌های مختلف است. دقت انسان در فرکانس‌های بالا کمتر از فرکانس‌های پایین است.

تا این مرحله تعدادی ضریب (مثلا ۲۵ تا) که انرژی را در فیلترهای مل یا بارک نشان می‌دهند به دست آمده است. توجه داریم که نوعی نمونه‌برداری مجدد[1] (در حیطه فرکانس) نیز انجام گرفته است و تعداد نمونه‌ها به ۲۵ عدد کاهش یافته است. سپس تبدیل کسینوسی گسسته[2] بر روی این ضرایب اعمال می‌شود و تعداد آنها را به ۱۲ یا ۱۳ عدد کاهش می‌دهد. حسن دیگر DCT در تولید ویژگی‌هایی است که از نظر احتمالی نسبتا مستقل هستند. با محاسبه مشتق اول و دوم این ویژگی‌ها که با توجه به اطلاعات قاب‌های مجاور تعیین می‌شود به ۳۶ یا ۳۹ ویژگی می‌رسیم.

۲-۳) ویژگی‌های *PLP [17] [24]*

شکل ۱۳ مراحل الگوریتم PLP را در مقایسه با LPC نشان می‌دهد. ما در اینجا راجع به الگوریتم LPC بحث نمی‌کنیم و در عوض به مقایسه بین PLP و MFCC می‌پردازیم. همچنین شکل ۱۴ تاثیر هر مرحله را نشان می‌دهد.



شکل ۱۳: مقایسه LPC با PLP

[1] Re-sampling
[2] DCT

مراحل الگوریتم PLP عبارتند از:

۱- **محاسبه طیف سیگنال**. در هر دو روش PLP و MFCC این کار با قاب‌بندی سیگنال صحبت و ضرب هر قاب در یک پنجره (مثلا Hamming) و سپس محاسبه FFT و در نهایت توان دوم مقدار هر مؤلفه انجام می‌شود.

۲- **محاسبه انرژی عبوری از بانک‌های فیلتر**. این فیلترها شکل‌های متفاوتی دارند، ولی همگی مبتنی بر نوعی مقیاس فرکانسی هستند که در فرکانس‌های زیر ۱۰۰۰ هرتز خطی و در فرکانس‌های بالاتر از آن تقریبا لگاریتمی عمل می‌کند. دو مقیاس متداول برای این منظور مل و بارک هستند که در بخش قبل معرفی شدند.

۳- **پیش تاکید طیف برای تقریب زدن حساسیت غیر یکنواخت شنوایی انسان در فرکانس‌های مختلف**. پیش تاکید در روش MFCC پیش از محاسبه طیف با مشتق‌گیری از سیگنال صحبت انجام می‌شود. در تحلیل PLP این مرحله با یک وزن‌دهی صریح به مؤلفه‌های طیف بانک فیلتر انجام می‌شود.

۴- **فشرده‌سازی دامنه طیف طبق قانون توان شنیدن**[1]. در MFCC این مرحله با محاسبه لگاریتم انرژی صورت می‌گیرد ولی در PLP از قانون ریشه سوم استفاده می‌شود.

۵- **محاسبه تبدیل فوریه معکوس**. این مرحله هم در MFCC و هم در PLP عملا با DCT انجام می‌شود. در کتب معمولا ادعا می‌شود که چون مقادیر طیف توان حقیقی و زوج هستند، می‌توان به‌جای IDFT از DCT استفاده کرد. به نظر نگارنده این استدلال صحیح نیست چون ما می‌توانستیم از فازی که از تبدیل فوریه به دست آمده است استفاده کنیم. به نظر می‌رسد دلیل اصلی استفاده از DCT رسیدن به مؤلفه‌های منستقل آماری باشد.

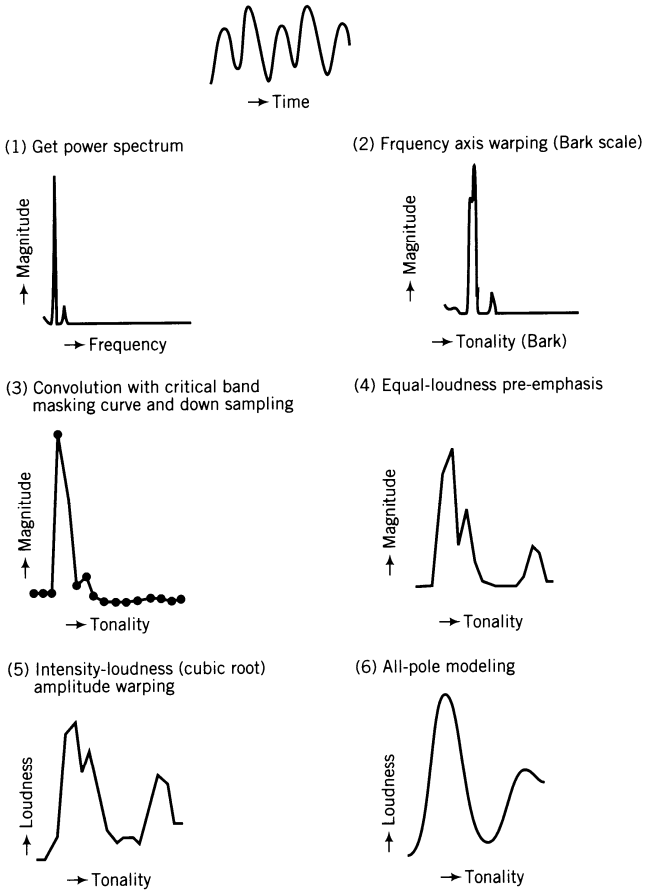
۶- **نرم کردن طیف**[2]. هرچند استفاده از بانک فیلتر تا حدی جزئیات فرکانسی را کاهش می‌دهد اما در عمل از یک مرحله دیگر از یکسان‌سازی[3] برای کم کردن تاثیر صداهایی که از منابعی غیر از صحبت تولید شده‌اند، استفاده می‌شود. در MFCC این کار با حذف مؤلفه‌های پایین تبدیل کسینوسی گسسته انجام می‌شود. از ۲۵ مؤلفه بانک فیلتر ۲۵ مؤلفه تبدیل کسینوسی گسسته به دست می‌آید که نصف آنها دور

---

[1] Power law of hearing
[2] Spectral smoothing
[3] Integration

ریخته می‌شوند. بدین ترتیب تاثیر تغییرات شدید در طیف سیگنال نادیده گرفته می‌شود. در PLP طیف با یک مدل تمام‌قطب تقریب زده می‌شود. از آنجا که هر قطب معادل یک قله در طیف است، این مدل برای تقریب زدن قله‌ها مناسب‌تر است تا دره‌ها. تجربه نشان داده است که روش PLP از نظر استحکام در مقابل نویز و همچنین استقلال از گوینده بهتر از حذف مؤلفه‌های انتهایی در MFCC عمل می‌کند.

۷- **استفاده از نمایش متعامد.** در مورد MFCC نیازی به انجام کار اضافی نیست زیرا خروجی تبدیل کسینوسی گسسته این خاصیت را دارد. در PLP ضرایب خودپسرو[1] به ضرایب کپسترال تبدیل می‌شوند.



شکل ۱۴: تاثیر گام‌های PLP بر روی طیف

[1] Autoregressive

۲-٤) ویژگی‌های RASTA *[17] [23]*

همانطور که دیدیم روش‌های PLP، MFCC و PCA با نرم کردن شکل طیف تاثیرات نامطلوب نویز بر روی شکل طیف در فرکانس را حذف می‌کردند. RASTA تاثیرات نویز را در شکل طیف در زمان حذف می‌کند.

فرض کنیم که سیگنال صحبت با طیف کوتاه-زمان $S(\omega,t)$ از فیلتر خطی مستقل از زمان $H(\omega,t)$ عبور کرده است. بنابراین اگر $X(\omega,t)$ سیگنال مشاهده شده باشد، داریم:

$$X(\omega,t) = S(\omega,t)H(\omega,t)$$

بنابراین لگاریتم طیف توان چنین است:

$$\ln\left|X(\omega,t)\right|^2 = \ln\left|S(\omega,t)\right|^2 + \ln\left|H(\omega,t)\right|^2$$

بنابراین نویز پیچشی در فضای زمان به صورت نویز ضربی در فضای فرکانس و نویز جمعی در فضای لگاریتم فرکانس ظهور می‌کند. اگر ویژگی‌های نویز، $H(\omega,t)$، در زمان با سیگنال صحبت متفاوت باشد می‌توان آن را به سادگی جدا کرد. برای مثال اگر H مقدار ثابتی در زمان باشد و مؤلفه‌های ثابت S مهم نباشند، می‌توان تخمین این مؤلفه ثابت را با میانگین‌گیری بر روی طیف لگاریتم توان به دست آورد. همچنین می‌توان نویز را با توجه به اطلاعات زمانی ویژگی‌هایی مانند MFCC و یا PLP به دست آورد. بدین ترتیب RASTA-PLP به دست آمد. در حقیقت روش CMS[1] نوع خاصی از تحلیل فوق است که در آن مؤلفه ثابت ویژگی‌های MFCC حذف می‌شود.

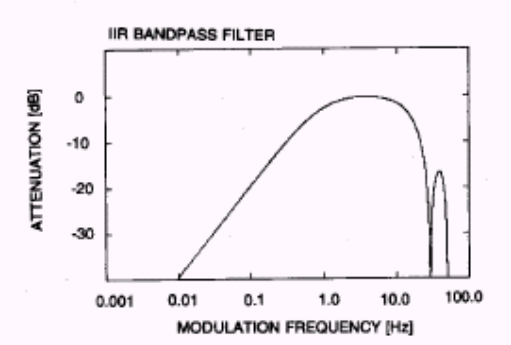
تحلیل فرکانسی
محاسبه بانک فیلتر
عبور دادن خط سیر لگاریتم انرژی در ویژگی‌های بانک فیلتر از فیلتر میان‌گذر
برگرداندن لگاریتم انرژی در بانک فیلتر به انرژی در بانک فیلتر
پردازش‌های اختیاری

شکل ۱۵: پردازش RASTA

[1] Cepstral Mean Subtraction

در روش RASTA هر مؤلفه بانک فیلتر از یک فیلتر میان‌گذر عبور می‌کند. در حقیقت فرکانس‌های پایین و بالا که در شنیدنی بودن[1] صحبت اهمیت کمی دارند حذف می‌شوند. سپس خط سیر فیلتر شده با تابع نمایی از لگاریتم انرژی به انرژی تبدیل می‌شود. بدین ترتیب دیده می‌شود که روش RASTA را می‌توان به اول هر الگوریتمی که در الگوریتم خود بانک فیلتر را محاسبه می‌کند اضافه کرد.

شکل ۱۵ مراحل اصلی پردازش RASTA را نشان می‌دهد. این پردازش را می‌توان به اول PLP یا MFCC اضافه کرد. شکل ۱۶ فیلتر میان‌گذری را که بر اساس آزمایش به دست آمده است نشان می‌دهد. در طراحی این فیلتر سعی شده است که قطب‌های فیلتر در مکان‌هایی قرار داده شوند که نرخ بازشناسی در یک سیستم تشخیص کلمه مجزا[2] بیشینه شود. با کمال تعجب این مشخصه شباهت زیادی به یک اندازه‌گیری مستقل در مورد حساسیت انسان به فرکانس مدولاسیون که در شکل ۱۷ نشان داده شده است، دارد.

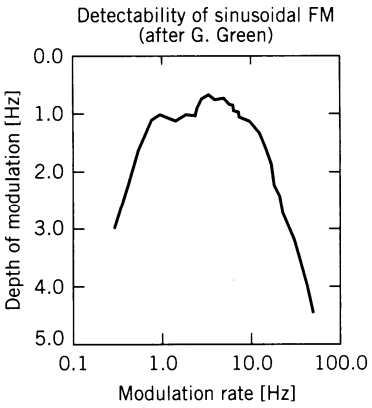


شکل ۱۶: مشخصه فیلتر میان‌گذر RASTA.

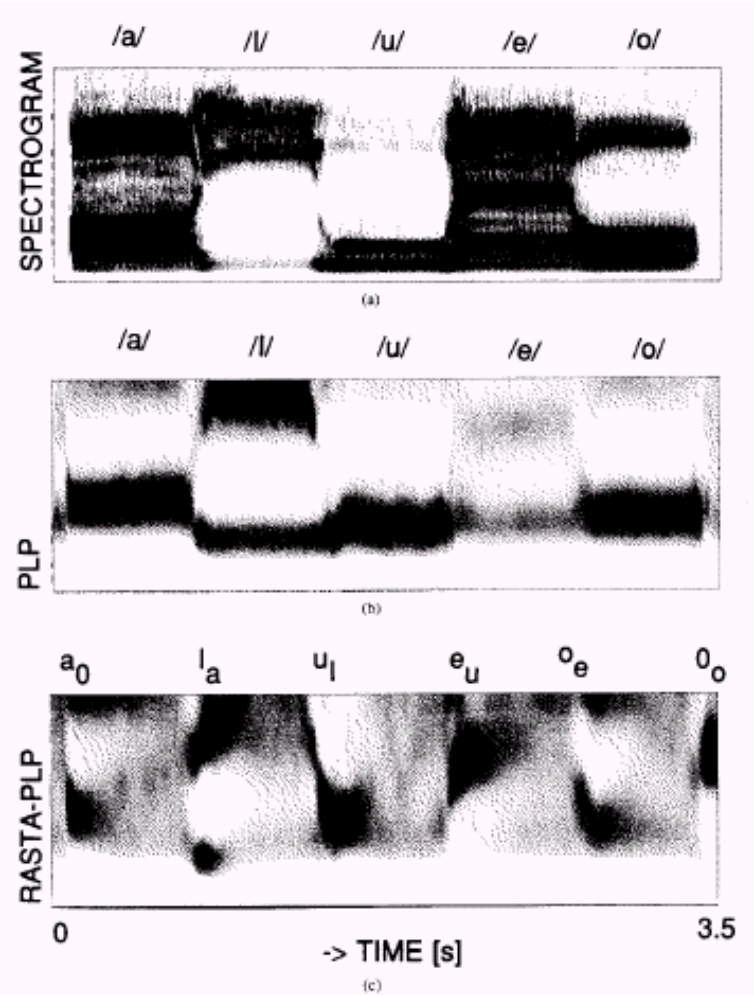
نکته دیگر در مورد RASTA این است که RASTA حالات پایدار سیگنال صحبت را از بین می‌برد و تغییرات را برجسته می‌کند. شکل ۱۸ این اثر را نشان می‌دهد. این امر باعث می‌شود که RASTA برای تشخیص واج مناسب نباشد زیرا در پردازش RASTA اطلاعات واج تقریبا پاک شده‌اند. RASTA را باید در مدل‌های دوواجی یا سه‌واجی استفاده کرد.

ضمنا اگر تشابه شکل ۱۶ و شکل ۱۷ را نشانه‌ای بر تشابه RASTA با روش انسان در تشخیص صحبت بدانیم، باید بپذیریم که انسان نیز نسبت به تغییرات حساس است و به لحظات پایدار سیگنال صحبت توجه ندارد.

[1] Intelligibility
[2] Isolated Word Recognition



شکل ۱۷: حساسیت انسان به فرکانس مدولاسیون



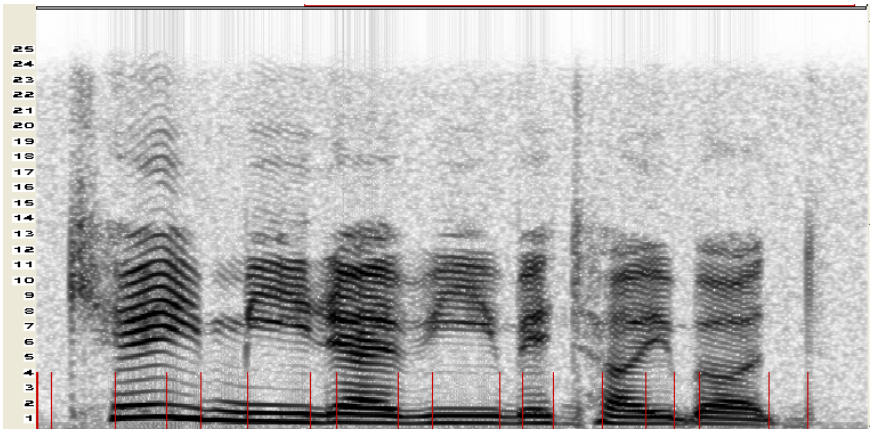
شکل ۱۸: طیف پنج آوای زبان چک که نگه داشته شده‌اند.(a) طیف سیگنال، (b) PLP و (c) RASTA–PLP

فصل ۳
شناخت روش انسان در تشخیص صحبت

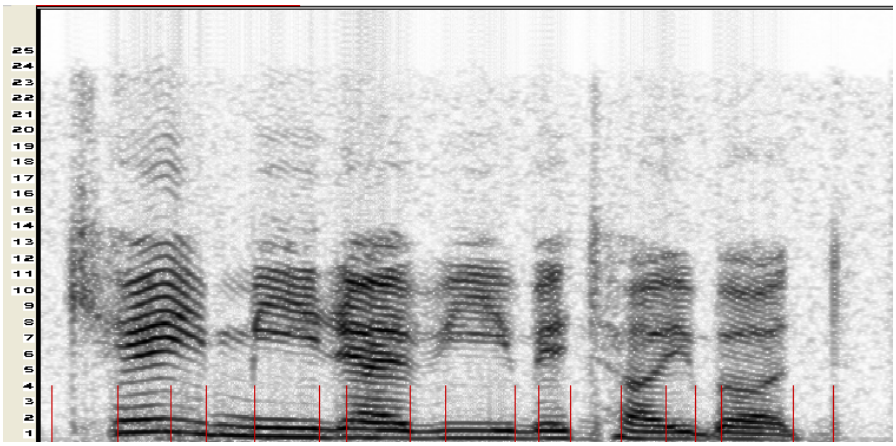
اثر کوانته کردن مقدار لگاریتم انرژی در هر فرکانس بر روی صدا
اثر کوانته کردن مقدار انرژی در هر باند فیلتر بر روی صدای شنیده شده
بررسی روش انسان در تشخیص تفاوت بین «ما و نا» و «با و دا» و ...
بررسی مفهوم امکان در تشخیص صحبت
حذف نویز – پذیرفتن یک حالت ممکن
دقیق– غیر دقیق
حساس به مقدار – حساس به تغییرات
ویژگیهای شنوایی – ویژگیهای گویایی
مبتنی بر یک مدل پیچیده یا چند مدل ساده
مبتنی بر یادگیری بامعلم– بدون معلم
مبتنی بر اطلاعات سطوح گرامر و کلمه یا مبتنی بر اطلاعات سطح سیگنال

۳-۱) اثر کوانته کردن[۱] مقدار لگاریتم[۲] انرژی در هر فرکانس بر روی صدا

ابتدا لگاریتم انرژی در هر فرکانس را در سطح قاب محاسبه می‌کنیم. برای ترکیب قاب‌ها از روش Add-Overlap در فصل ۵ استفاده می‌نماییم. در بحث ما $\hat{s}_m(n)$ سیگنال قاب m ام است که لگاریتم انرژی آن کوانته شده است. سوال این است که اگر لگاریتم انرژی در $s_m(n)$ را به L سطح کوانته کنیم، چقدر در صدای شنیده شده تغییر ایجاد می‌شود. آزمایش‌های ما نشان می‌دهد که برای L=5,10,20 این گسسته کردن باعث می‌شود که انسان احساس کند که صدا کمی ماشینی شده است ولی صدا هنوز کاملا قابل تشخیص[۳] است. اما برای L=100 می‌توان گفت که انسان چیزی را احساس نمی‌کند. شکل ۱۹ طیف صدای اولیه و شکل ۲۰ طیف همان صدا را پس از گسسته کردن به ۵ سطح نشان می‌دهد.



شکل ۱۹: طیف فایل s11881.wav از دادگان فارس‌دات.



[۱] البته اگر کوانته کردن برحسب صدک‌ها باشد نتایج باید خیلی بهتر باشد. البته اکنون نیز عمل کوانته کردن پس از پیدا کردن ضریب مناسب برای طیف-نگار انجام می‌شود.

[۲] ما برای منفی نشدن لگاریتم از فرمول کلی $\log(1+Jx)$ به‌جای $\log(x)$ استفاده می‌کنیم.

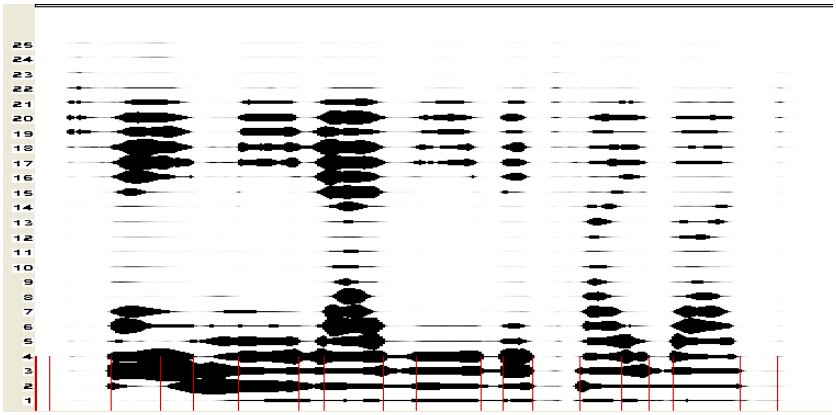
[3] Intelligible

شکل ۲۰: طیف فایل s11881.wav از دادگان فارس‌دات پس از کوانته شدن به ۵ سطح.

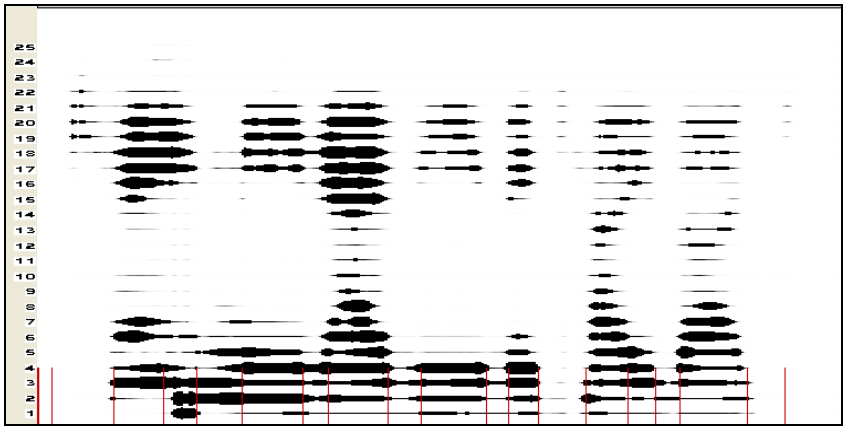
۳-۲) اثر کوانته کردن مقدار انرژی در هر باند فیلتر بر روی صدای شنیده شده

روش ما برای کوانته کردن مقدار لگاریتم انرژی در هر باند فیلتر در فصل ۵ توضیح داده شده است. متاسفانه نگارنده خود از نتایج این الگوریتم راضی نیست و معتقد است که اولا خروجی الگوریتم دقیقا بر آنچه انتظار می‌رود منطبق نیست، ثانیا می‌توان مسائل بهتری را طرح کرد (مثلا به‌جای $\sum \log(x)$ یا $\log(\sum x)$ از $\log(\sum \log(x))$ استفاده کرد که نگارنده معتقد است شکل زیباتری دارد) و ثالثا می‌توان بهبودهای زیادی در الگوریتم داد تا آن را از جهات سرعت و دقت قابل استفاده‌تر کند. به هر حال در اینجا نمونه‌ای از کوانته کردن در فضای بانک فیلتر به حدود۵۰ سطح نمایش داده می‌شود. شکل ۲۱ نمایش بانک فیلتر فایل s11881.wav از دادگان فارس‌دات را نشان می‌دهد. شکل ۲۲ نمایش بانک فیلتر این فایل را و شکل ۲۳ طیف این فایل را پس از کوانته کردن به ۵۰ سطح نشان می‌دهند.

از آنجا که روش پردازش سیگنال برای ایجاد تغییر در فضای ویژگی‌های بانک فیلتر کار بیشتری را می‌طلبد، نمی‌توان ادعا کرد که نتایجی بهتر از این قابل حصول نیست. به هرحال به نظر می‌رسد که قابل فهم بودن صدا با کوانته کردن به ۵۰ سطح از بین نمی‌رود.

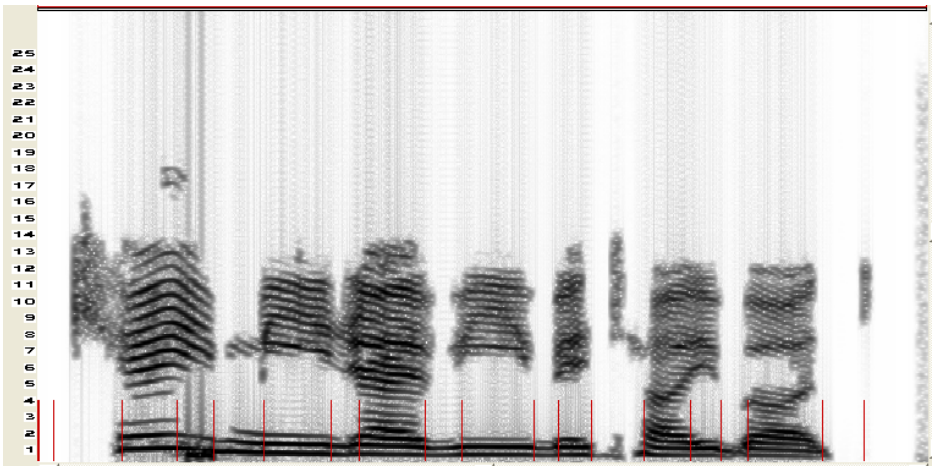


شکل ۲۱: نمایش بانک فیلتر فایل s11881.wav از دادگان فارس‌دات.



شکل ۲۲: نمایش بانک فیلتر فایل s11881.wav از دادگان فارس‌دات پس از کوانته شدن لگاریتم انرژی در بانک فیلتر به ۵۰ سطح. فکر

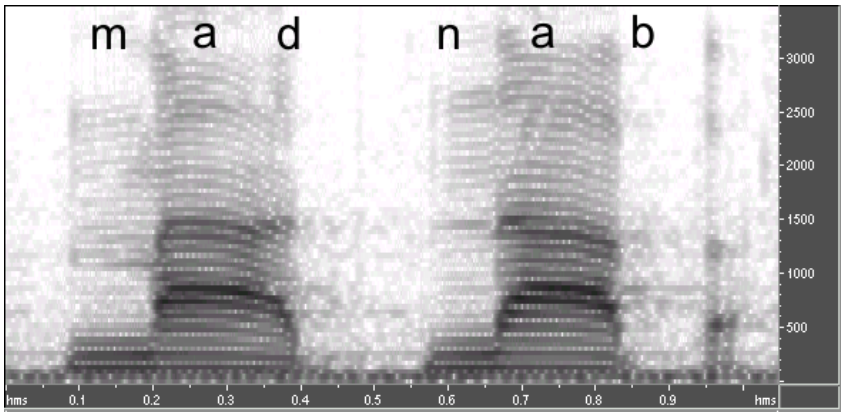
می‌کنم خواننده نیز با من هم‌عقیده باشد که تغییری که رخ داده است بیش از کوانته شدن به ۵۰ سطح است.



شکل ۲۳: نمایش طیف فایل s11881.wav **از دادگان فارس‌دات پس از کوانته شدن لگاریتم انرژی در بانک فیلتر به ۵۰ سطح.**

۳-۳) بررسی روش انسان در تشخیص تفاوت بین «ما و نا» و «با و دا» و ...

با استفاده از ابزار HearView[1] می‌توانیم اثر ایجاد تغییر در ویژگی‌های بانک فیلتر را بر روی صدای شنیده شده بررسی کنیم. در اینجا سعی می‌کنیم تفاوت بین واج‌های m و n را در کلمات "mad" و "nab" بررسی کنیم. طیف این کلمات در شکل ۲۴ نمایش داده شده است. همانطور که دیده می‌شود، تفاوت‌های بسیار کمی بین m و n در طیف-نگار است. جالب‌تر اینکه حتی اگر با یک نویز سفید، تفاوت‌های دیده شده بین n و m را از بین ببریم، باز هم می‌توانیم تفاوت این دو واج را بشنویم.

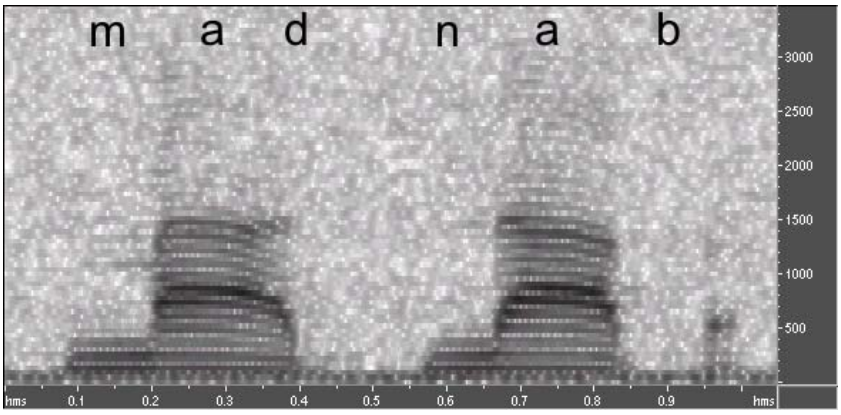


شکل ۲۴: طیف کلمات "mad" و "nab"

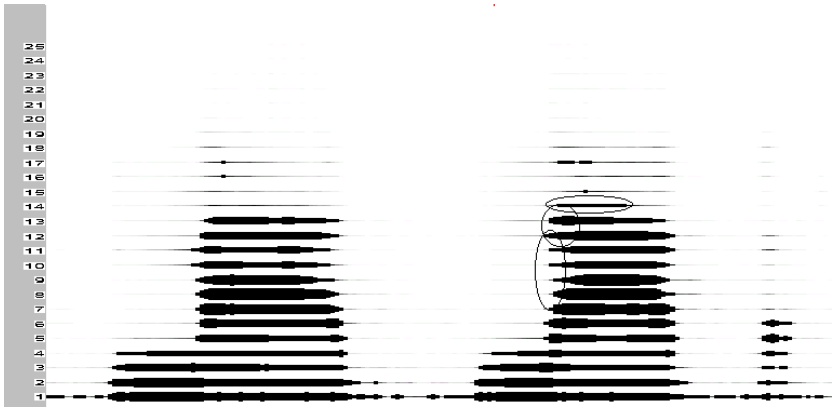
طیف این دو کلمه پس از اضافه شدن نویز سفید به آن در شکل ۲۵ نشان داده شده است. حدس می‌زنیم که تفاوت بین m و n در گذر از این واج‌ها به واج /æ/ باشد. برای تست این حدس، اصلا جای m و n

[1] این ابزار توسط نگارنده برای بررسی روش انسان در تشخیص صحبت ساخته شده است.

را در این دو کلمه عوض می‌کنیم. باز هم همان صدای اولیه شنیده می‌شود. حدس ما درست است. تفاوت اصلی بین m و n در خود آنها نیست بلکه در نحوه شروع واج بعدی است. شکل ۲۶ نمایش بانک فیلتر شکل ۲۴ است. ما سه حدس راجع به تفاوت بین m و n می‌زدیم که با بیضی‌هایی در شکل ۲۶ نشان داده شده است. توجه داریم که تفاوت‌های دیگری نیز می‌توانست پیشنهاد شود (برای مثال می‌توانستیم تفاوت‌هایی را که در باند ۱۷ ام[1] وجود دارد در نظر بگیریم).



شکل ۲۵: طیف سیگنال شکل ۲۴ پس از اضافه شدن نویز سفید.



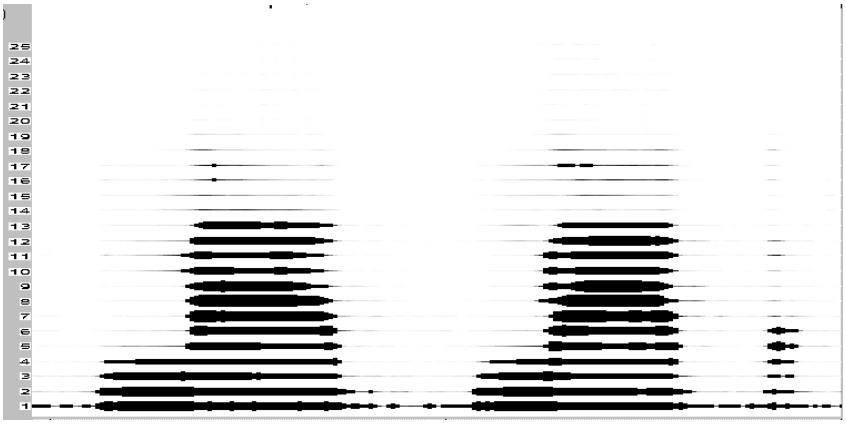
شکل ۲۶: ویژگی‌های بانک فیلتر سیگنال شکل ۲۴. کلفتی خط‌ها نشان‌دهنده میزان انرژی در هر باند است.

برای تست حدس‌هایمان، از نرم افزار HearView استفاده می‌کنیم و سعی می‌کنیم که گذر /næ/ در کلمه nab را شبیه گذر /mæ/ در کلمه mab کنیم. سیگنال تغییر یافته در شکل ۲۷ نشان داده شده است. اکنون سیگنال به صورت "mad-mab" شنیده می‌شود. همچنین بررسی ما نشان داد که اهمیت ناحیه دایره‌ای از ناحیه بیضوی عمودی بیشتر است و همین‌طور اهمیت ناحیه بیضوی عمودی از ناحیه بیضوی افقی بیشتر است. برای آزمودن استحکام ویژگی‌های خود، آنها را بر روی داده نویزی امتحان کردیم. همانطور که

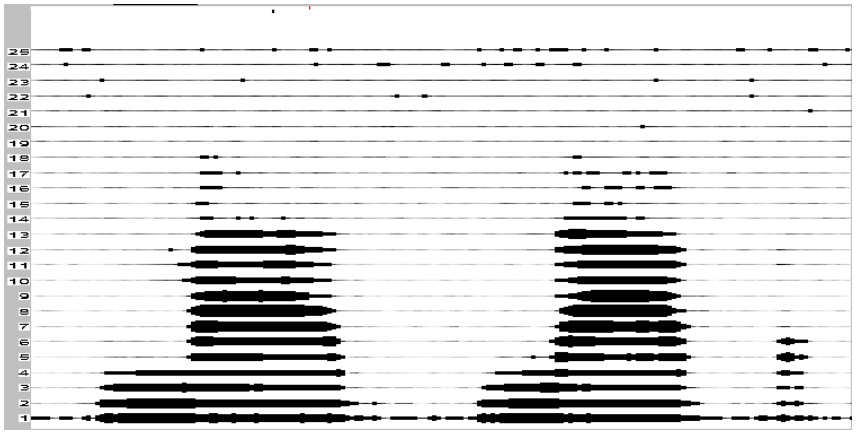
[1] که معادل فیلتری به مرکز 2KHz و پهنای باند 600Hz است.

شکل ۲۸ نشان می‌دهد، ویژگی‌های ما دارای استحکام خوبی هستند و نویز نتوانسته‌است آنها را از بین ببرد.

اگر مشابه این آزمایش‌ها را برای یافتن تفاوت بین واج‌های /b/ و /d/ و اصولا واج‌های کوتاه (واج‌های صدادار و انفجاری) انجام دهیم می‌بینیم که انسان در بسیاری از موارد از روی نحوه شروع واج بعدی یا نحوه اتمام واج قبلی این واج‌ها را می‌شناسد. نکته دیگر این است که اگر واج /m/ از کلمه mad حذف شود و فقط اثر آن بر روی واج بعدی باقی بماند، صدای bad شنیده خواهد شد. به طور مشابه nab به dab تبدیل می‌شود. یکی از درس‌های این آزمایش این است که ویژگی مستحکم ویژگی‌ای است که دارای انرژی زیادی است (حتی اگر در واج بعدی باشد).



شکل ۲۷: سیگنال تغییر یافته. اکنون کلمه دوم mab شنیده می‌شود.



شکل ۲۸ : نمایش بانک فیلتر سیگنال شکل .۲۵

۳-٤) بررسی مفهوم امکان در تشخیص صحبت

یک سیگنال داده شده، مثلا s، را در نظر بگیرید. توجه کنید که ما نمی‌دانیم که این سیگنال از کجا آمده است. آیا این سیگنال یک سیگنال طبیعی است و یا اینکه یک سیگنال مصنوعی است و یا حتی صدای

بوق یک ماشین است. می‌خواهیم راجع به انتساب این سیگنال به یکی از واج‌های شناخته شده و یا رد کردن[1] آن تصمیم‌گیری کنیم. فرض کنید که دو گروه A و B را می‌شناسیم و X به معنای رد کردن است. روش‌های متداول سعی می‌کنند این واج را به گروهی نسبت دهند که احتمال اینکه چنین سیگنالی توسط آن گروه تولید شده باشد بیشتر است[2]. از همین‌جا معلوم است که نمی‌توان معیاری برای رد کردن داشت، زیرا ما از احتمال اینکه چیزی که نشناسیم این سیگنال را تولید کرده باشد اطلاع نداریم. تئوری امکان به شکل دیگری فکر می‌کند. ما می‌خواهیم میزان شدنی بودن این فرض را که سیگنال داده شده را بتوان به‌جای نمونه‌ای از گروه A یا B استفاده کرد، بررسی کنیم. دیگر برای ما اهمیتی ندارد که در اکثر موارد، این سیگنال توسط چه کسی تولید شده است. از دید عینی، این سیگنال می‌تواند به عنوان نمونه‌ای از گروه A به کار رود و شنونده هم ایراد نخواهد گرفت.

در این بخش می‌خواهیم ببینیم که امکان اینکه واج نوعی X و یا حتی یک نویز را به جای واج نوعی Y به‌کار بریم و شنونده هنوز هم همان صحبت قبلی را بشنود چقدر است. از دیدی دیگر می‌توان بررسی کرد که اصلا در یک سیگنال طبیعی، چقدر واج‌های ادا شده به تنهایی به مفهوم آن واج شبیه هستند.

در اینجا لیستی کوچک از موارد امکان را ذکر می‌کنیم که خواننده می‌تواند خودش درستی آنها را بیازماید.

۱- می‌توان[3] به جای بخش انفجاری حرف t یک نویز سفید گذاشت.

۲- می‌توان صدای «دا» را در صداهای «نا» و «سا» پیدا کرد.

۳- می‌توان صدای «با» را در صدای «ما» پیدا کرد.

۴- می‌توان نمونه‌هایی از صدای /æ/ را پیدا کرد که اگر به تنهایی شنیده شوند صدای /a:/ بدهند.

۵- می‌توان قبول کرد که کلمه «فرید» که بر روی آن نویز سفید مناسب قرار گرفته است، کلمه «سعید» بوده است.

۶- اگر کسی کلمه‌ای را بشنود و تشخیص دهد، نمی‌توان با تغییرات کم در سیگنال او را متقاعد کرد که کلمه دیگری ادا شده است.

۷- می‌توان بسیاری از واج‌ها را به‌جای یکدیگر به‌کار برد، به شرط آنکه اثراتی که این واج‌ها بر روی واج‌های کناری می‌گذارند حفظ شود (مثلا م و ن. ولی لیست بلندتر از این است).

---

[1] Rejection

[2] ممکن است کسی در پیاده‌سازی خود امکان رد کردن را نیز اضافه کند. در این صورت دیگر به نظریه احتمال پایبند نبوده است و همچنین دیگر اثبات‌های نظریه احتمال در مورد بهینه بودن سیستم صدق نمی‌کنند.

[3] می‌توان یعنی امکان دارد.

۳-۵) حذف نویز – پذیرفتن یک حالت ممکن

یکی از روش‌های متداول در تشخیص صحبت تشکیل مدلی برای نویز و یادگیری آن و سپس حذف نویز است. دلایل زیر نشان می‌دهد که این روش با روش انسان متفاوت است:

۱- در این روش نویز باید از قبل شناخته شده باشد. سیستم نمی‌تواند در محیطی که نویزی ناشناخته وجود دارد کار کند.

۲- علاوه بر اینکه نویز باید مشخص باشد، نوع آن نیز باید مشخص باشد. سیستم نمی‌تواند بطور خودکار نویز را طبقه‌بندی کند.

۳- سیستم نمی‌تواند نویزهای جدید را یاد بگیرد.

۴- روش‌های حذف نویز کاملا بر اساس روش‌های ریاضی و بدون توجه به شباهتشان به انسان ارائه می‌شوند. البته ممکن است در آینده نشان داده شود که مغز انسان قابلیت برخی از این پردازش‌ها را داشته است (برای مثال تصور الگوریتمی طبیعی برای spectral mean subtraction دور از ذهن نیست)، ولی در برخی موارد روش‌هایی که ارائه شده اند بیش از حد مصنوعی هستند (مانند روش [6] برای حذف پژواک صدا).

اصول روشی که ما برای شناسایی و حذف نویز ارائه می‌دهیم (و ادعا می‌کنیم که به روش انسان شبیه‌تر است) چنین است: «انسان در ابتدا که به دنیا می‌آید هیچ صدایی را نمی‌شناسد. اما انسان می‌تواند سیگنال صوتی را به بخش‌هایی تقسیم کند که ما آنها را شیئ می‌نامیم[1]. به عبارت دیگر انسان یک بخش‌بندی شنیداری انجام می‌دهد. سپس بر اساس شباهت بین اشیاء (که یک اندازه‌گیری فازی[2] است) یک طبقه‌بندی از نمونه‌ها انجام می‌دهد. برای مثال ممکن است در طبقه‌بندی اولیه واج‌های «ب، د، ق، ء و گ» در یک گروه قرار گیرند. به مرور با توجه به آموزش‌های بامعلم، انسان یاد می‌گیرد که این طبقه بندی اولیه را اصلاح کند. بدین منظور برخی از اشیاء نام مشترک می‌گیرند و برخی از اشیاء به چند نوع جدید تجزیه می‌شوند. ادامه فرآیند یادگیری بر اساس یادگیری اشیاء بزرگ‌تری مانند واج و کلمه است. نکته مهم این است که همه اینها شیئ هستند (البته مسلم است که شیء پایه نیستند). خواننده علاقمند می‌تواند مبانی روانشناسی و فرمول‌بندی دقیق‌تر این روش را در کتاب [51] جستجو کند. حال روش حذف نویز را شرح می‌دهیم.»

نکته مهم این است که انسان اشیائی را که دیده است خیلی خوب می‌شناسد و از جهات مختلف می‌تواند تشخیص دهد که سیگنالی که می‌شنود با نمونه آموزشی تفاوت دارد. انسان به‌قدری این اشیاء را خوب

---

[1] به نظر می‌رسد که بخشی از قابلیت بخش‌بندی ذاتی و بخشی دیگر نتیجه تمرین و یادگیری است.

[2] Fuzzy measure

می‌شناسد که می‌تواند علاوه بر صحبت صداهای دیگری مانند صدای سوت، صدای پرندگان، صدای ماشین و ... را نیز تشخیص دهد. بدین ترتیب انسان می‌تواند تشخیص دهد که برخی از اشیاء دیده شده شبیه هیچ شیئی نیستند. در این مرحله انسان سعی می‌کند طور دیگری به صدا گوش دهد. برای مثال ممکن است با دور و نزدیک شدن به منبع صوتی، شرایط بهتری را بوجود آورد (خواننده اکنون اهمیت شناخت قوی از اشیاء را درک می‌کند. اگر انسان نمی‌توانست میزان مناسب بودن تشخیص خود را بسنجد، نمی‌توانست بفهمد که تغییراتی که بوجود می‌آورد بهتر بوده‌اند یا بدتر). ما معتقدیم که امکانات دیگری نیز در اختیار انسان وجود دارد که به کمک آنها می‌تواند ویژگی‌هایی که به مغز می‌رسند را تغییر دهد. بدین ترتیب انسان ابتدا تلاش می‌کند که پارامترهای گوش دادن را تنظیم کند. اگر این مرحله با موفقیت انجام شود، انسان اطمینان پیدا می‌کند که با امکان بسیار بالا صدا را درست می‌شنود. اگر هنوز هم اشیائی باشند که از نظر انسان زیادی هستند، انسان آنها را به عنوان نویز می‌شناسد و به مرور آنها را نیز یاد می‌گیرد.»

پس به طور خلاصه:

۱- در این روش سیستم خودش نویز را کشف می‌کند.

۲- در این روش نویز حذف نمی‌شود؛ بلکه ابتدا شناسایی می‌شود و سپس با شناخت کامل دور انداخته نمی‌شود. همچنین علم به اینکه نویز در لحظه‌ای حضور داشته است امکان وجود اشیائی را که زیر نویز له شده‌اند را بالا می‌برد. از آنجا که انسان بر اساس تئوری امکان کار می‌کند، مشکلی در تشخیص صحبت برای او پیش نمی‌آید. برای مثال اگر یک نویز سفید بسیار شدید داشته باشیم، «ممکن است کسی در همان لحظه به آرامی گفته باشد: من هستم».

۳- در مرحله بعد، پارامترهایی که برای آن محیط کشف شده‌اند بخاطر سپرده می‌شوند. یعنی انسان می‌داند که هر صحبتی را چگونه باید گوش دهد. به همین دلیل اگر منتظر صحبتی به زبان فارسی باشید و کسی انگلیسی صحبت کند، برای چند لحظه غافلگیر می‌شوید. همچنین وقتی ما به محیط جدیدی می‌رویم، پارامترهای شنیدن خود را بهنگام می‌کنیم.

۳-۶) دقیق- غیر دقیق

سوال اول این است که آیا اصولا اطلاعات دقیقی وجود دارد که انسان از آنها استفاده نکند یا خیر؟ به عبارت دیگر آیا انسان از همه ویژگی‌های اصواتی که توسط انسان تولید می‌شود استفاده می‌کند یا اینکه ویژگی‌های شنیداری بخشی از ویژگی‌هایی هستند که وجود دارند؟ پاسخ این سوال مثبت است. توانایی ماشین در تشخیص واج بیش از انسان است اما انسان می‌تواند صحبت را در شرایط نویزی و در حالات

مختلف (مثلا با آواز) هم تشخیص دهد. این نشان می‌دهد که اطلاعاتی بیش از آنچه انسان می‌داند در سیگنال صحبت وجود دارد.

آزمایش‌های مختلف نشان می‌دهند که انسان بیشتر از سازگاری امکان‌های مختلف برای تشخیص صحبت استفاده می‌کند تا اینکه به مقدار دقیق ویژگی‌های زمانی- فرکانسی حساس شود:

۱- اگر دور نوار را تند کنیم باز هم انسان صحبت را تشخیص می‌دهد.

۲- همین آزمایش نشان می‌دهد که انسان به ۲ برابر شدن فرکانس‌ها نیز حساس نیست.

۳- مشخص است که انرژی نیز تاثیر کمی در تشخیص صحبت دارد.

البته با توجه به بحثی که در مورد روشی وفق یافتن در انسان انجام شد، می‌توان این توانایی‌ها را به وفق یافتن در انسان نسبت داد (و گفت یک پارامتر بیرونی برای کش دادن و یا جمع کردن محورهای زمان، فرکانس و انرژی وجود دارد). اما آنچه قطعی است این است که انسان به بیش از ۱۰۰ سطح[1] حساس نیست و با دیدی واقع‌بینانه‌تر باید گفت که انسان در بسیاری از موارد تنها به حدود ۵ سطح حساس است (این سطوح همان توابع عضویت[2] هستند). خواننده از لغت توابع عضویت باید متوجه شده باشد که این سطوح معمولا دارای همپوشانی هستند و دارای فواصل یکسان نیز نیستند.

۳-۷) حساس به مقدار - حساس به تغییرات

نکته مهم دیگر این است که اصولا انسان به تغییرات در ویژگی‌ها بیش از مقدار آنها حساس است. برخی دلایلی که بر این مدعا وجود دارد عبارتند از:

۱- در برنامه HearView امکان این وجود دارد که یک تناوب از یک صدای تناوبی تکرار شود و صدای آن پخش شود. آزمایش نشان می‌دهد که در این حالت تشخیص واج ادا شده بسیار دشوار است.

۲- اگر در آزمایش قبل شخص نتواند لحظه تغییر از سکوت به واج مورد نظر را بشنود، تشخیص واج ادا شده مشکل‌تر می‌شود.

۳- تجربیات دیگر محققین نیز مؤید اهمیت تغییرات در تشخیص صحبت است[61]

۴- البته این بدین معنا نیست که انسان اصلا به مقدار ویژگی توجه ندارد، بلکه برای تاکید بر اهمیت تغییرات است.

---

[1] Level
[2] Membership functions

۳-۸) ویژگی‌های شنوایی - ویژگی‌های گویایی

آیا ویژگی‌هایی که انسان بر اساس آنها صحبت را تشخیص می‌دهد ویژگی‌های شنیداری هستند و یا ویژگی‌های گویایی. از دید نحوه تولید صدا می‌توان برای هر واج ویژگی‌های گویایی و شنیداری قائل شد. چنین ویژگی‌هایی در علم تجوید قرآن[۱] و علم فونتیک ذکر شده‌اند. به عنوان مثالی از ویژگی‌های گویایی می‌توان به وضعیت دهان، وضعیت لب‌ها و دندان‌ها، وضعیت زبان و ... اشاره کرد. به عنوان ویژگی‌های شنیداری می‌توان به ویژگی تکریر[۱] در حرف «راء»، صدای صفیر در حروف «س و ز» و همچنین صفات اصلی حروف اشاره کرد. از آنجا که صفات شنیداری برای اطمینان از صحت ادا شدن واج استفاده می‌شوند، در برخی از موارد می‌توان ویژگی‌هایی گویایی نیز برای این صفات پیدا کرد. برای مثال می‌توان صفت تکریر را به گونه خاصی از اتصال زبان به سقف دهان نسبت داد.

با وجود آنکه نمی‌توان اهمیت صحبت کردن را در یادگیری بهتر زبان انکار کرد، اما روشن است که صحبت کردن لازمه یادگیری زبان نیست. برخی معتقدند که انسان هنگامی که به صحبت شخص دیگری گوش می‌دهد، در حال شبیه‌سازی صحبت او است[18][37] . اما اگر چنین باشد، شخصی که گوش سالمی دارد ولی نمی‌تواند صحبت کند[۲]، باید متوجه حرف دیگران نشود.

به نظر می‌رسد که صحبت کردن نقش مهمی در یادگیری با معلم دارد زیرا افراد اجتماع می‌توانند اشتباهات گوینده را به او تذکر دهند. از طرفی لازمه این امر این است که انسان بتواند تفاوت بین آنچه باید گفته شود و آنچه گفته می‌شود را درک کند. بدین ترتیب، می‌توان گفت که قبل از آنکه ویژگی‌های گویایی درک شوند، انسان باید تفاوت را در صدای شنیده شده درک کند تا پس از آن بتواند صدای صحیح را تولید نماید. دلیل دیگری که بر این مدعا وجود دارد این است که انسان تفاوت بسیاری از صداهای مصنوعی که نمی‌تواند آنها را تولید نماید را نیز به خوبی درک می‌کند (صدای شیر آب، انواع موسیقی و ...). در نهایت به نظر می‌رسد (با وجود اهمیت صحبت کردن در یادگیری بامعلم) که ویژگی‌هایی که انسان برای تشخیص صحبت استخراج می‌کند ویژگی‌های شنیداری هستند و نه گویایی.

۳-۹) مبتنی بر یک مدل پیچیده یا چند مدل ساده

یکی از دلایلی که مخالفان روش‌های فازی در تشخیص صحبت برای رد روش‌های ساده فازی می‌آورند این حقیقت است که مدلی که بتواند گوناگونی صحبت را بپوشاند باید مدلی پیچیده باشد. انسان می‌تواند صحبت را در نویزهای مختلف، با دامنه‌های مختلف، با گوینده‌های مختلف، با ریتم‌های مختلف (آواز، دکلمه، صحبت تند و ...) و ... بشنود. وقتی به ویژگی‌های شنیداری صحبت دقت می‌شود، به نظر

---

[۱] این ویژگی را ذکر می‌کنند که بگویند حرف «را» اگر صحیح ادا شود این ویژگی را ندارد.

[۲] نگارنده از وجود چنین شخصی اطمینان ندارد. زیرا افراد لال، معمولا بدین علت نمی‌توانند صحبت کنند که ناشنوا نیز هستند.

می‌رسد که این گوناگونی باید حتما توسط مدلی پیچیده پوشش داده شود. اما ما معتقدیم که انسان برای هر گونه از صحبت مدلی دارد که پیچیده نیست. دلیل ما نیز این است که انسان علاوه بر اینکه تشخیص می‌دهد که چه چیزی گفته شده است، جنسیت گوینده، سن او، حالت او و خیلی ویژگی‌های دیگر را نیز درک می‌کند. به نظر می‌رسد اگر صحبت توسط یک مدل پیچیده درک می‌شد، ما نباید متوجه این تفاوت‌ها می‌شدیم. به بیان دیگر ما می‌فهمیم که:

۱- چگونه به صحبت گوش دهیم. همانطور که ذکر شد انسان پارامترهایی دارد که نحوه گوش کردن را کنترل می‌کند.

۲- بر اساس کدام مدل ساده گوش دهیم. کسانی که چند زبان بلدند ابتدا باید تشخیص دهند که گوینده به چه زبانی صحبت می‌کند. آنها ادعا می‌کنند که این دو زبان را در دو محل متفاوت از مغزشان ذخیره کرده‌اند.

۳- چگونه پارامترهای یک صحبت ناآشنا را یاد بگیریم. برای مثال نگارنده یک‌بار با شخصی آشنا شد که لهجه شهرستانی خاصی داشت. حدود ۳ دقیقه طول کشید تا من یاد بگیرم که چگونه باید به صحبت او گوش دهم و در ادامه صحبت می‌توانستم به‌راحتی صحبت او را بفهمم.

۳-۱۰) مبتنی بر یادگیری بامعلم- بدون معلم

مسلم است که کسی برای یادگیری صحبت به مدرسه نمی‌رود. بنابر این فرض یادگیری با معلم در مورد صحبت درست نیست. از طرف دیگر می‌دانیم که تقسیم‌بندی واج‌ها در زبان های مختلف یکسان نیست. برای مثال در زبان فارسی دری[1] واج‌های «ق و غ» ، «س و ث»، «ز و ذ» یکی گرفته می‌شوند، در حالی که عرب‌زبان و انگلیسی‌زبان بین برخی از آنها تفاوت قائل می‌شود. همچنین در زبان ژاپنی تفاوت بین «ر» و «ل» درک نمی‌شود. به این ترتیب به نظر می‌رسد که یادگیری صحبت در انسان هم شامل بخش‌هایی است که یادگیری با معلم صورت می‌گیرد و هم شامل بخش‌هایی است که یادگیری بدون معلم صورت می‌گیرد.

۱- استخراج ویژگی‌ها و بخش‌بندی صحبت بدون معلم انجام می‌شود.

۲- نام دهی به گروه‌های یادگرفته شده و ترکیب و تجزیه این گروه‌ها با معلم صورت می‌گیرد.

۳- پس از یادگیری اولیه، انسان می‌تواند از خودش به عنوان معلم استفاده کند و صحبت را در محیط‌های پیچیده‌تر نیز یاد بگیرد.

---

[1] در زبان فارسی پهلوی واج‌های دیگری نیز وجود داشته است.

۳-۱۱) مبتنی بر اطلاعات سطوح گرامر و کلمه یا مبتنی بر اطلاعات سطح سیگنال

برخی ادعا می‌کنند که ناتوانی سیستم‌های متداول در تشخیص صحبت در سطح واج به علت عدم وجود اطلاعات کافی در این سطح است و باید از اطلاعات سطوح کلمه و جمله نیز استفاده شود. در اینکه اطلاعات سطوح کلمه و جمله کمک زیادی به تشخیص صحبت می‌کنند، شکی نیست. اما آزمایشی ساده نشان می‌دهد که انسان می‌تواند صحبت را بدون استفاده از اطلاعات زبانی نیز تشخیص دهد. کافی است دنباله‌ای بی معنی از واج‌را برای شخص دیگری بگویید. خواهید دید که او به‌خوبی قادر است که صحبت شما را تشخیص دهد. بدین ترتیب به نظر نگارنده اطاعات زبانی برای تشخیص صحبت ضروری نیستند.

فصل ۴
نظریه‌های موجود برای برخورد با عدم قطعیت

نظریه احتمال

اپراتورهای TNorm **و** SNorm

نظریه مدرک شافِر دمپستر

اپراتورهای TNorm **و** SNorm

نظریه مجموعه‌های فازی

اپراتورهای TNorm **و** SNorm

نظریه امکان

توزیع امکان

امکان به معنای شدنی بودن

امکان به عنوان شروع یک منطق جدید برای اثبات ریاضی

عملگرهای غیر قابل جبران min **و** max

مساله بازشناسی گفتار: امکان یا احتمال

نظریه امکان یک نظریه دوبانده

اندازه‌گیری امکانی و عملگرهای TNorm **و** SNorm

اندازه‌گیری امکانی پیشنهادی نگارنده

آمار چیست؟

٤-١) نظریه احتمال

این نظریه برای اولین بار در قرن ۱۶ میلادی توسط گرومارو کاردانو[1] ریاضیدان ایتالیایی ارائه شد[36]. او بیان کرد که اگر پدیده‌ای دارای n برآیند ممکن باشد و یک رخداد متناظر با m نمونه از این برآمدها باشد، آنگاه احتمال این رخداد $\frac{m}{n}$ است. او کتابش را در سال ۱۵۲۵ نوشت، ولی انتشار آن تا سال ۱۶۶۳ طول کشید. اما اکثر تاریخ‌دان‌ها سال ۱۶۵۴ را سال پیدایش نظریه احتمال می‌دانند. در این سال، در پاریس یک قمارباز ثروتمند به نام چِوالیر دِمِیر[2] از ریاضیدانان مطرح آن دوره و از جمله بلیز پاسکال[3] یک سری سوال پرسید که معروف‌ترین آنها مساله نقاط است:

> A و B قرار می‌گذارند که یک سری بازی جوانمردانه انجام دهند تا اینکه یکی از آنها ۶ دست برده باشد. قرار است برنده تمام پول را بردارد. به هر دلیل بازی در حالی که A، ۵ دست و B، ۳ دست برده است، متوقف می‌شود. پول چگونه باید تقسیم شود؟

به این ترتیب دیده می‌شود که نظریه احتمال برای یافتن مدلی ریاضی برای شانس بنا شد. به این ترتیب تعریف کلاسیک احتمال پدید آمد:

تعریف کلاسیک احتمال[4]: آزمایشی را در نظر بگیرید که دارای n برآمد ممکن است که همگی به یک اندازه شانس دارند. اگر رخداد A متناظر با m مورد از آن n برآمد باشد، داریم:

$$\text{Probability of A} = P(A) = \frac{m}{n}$$

این تعریف محدودیت‌هایی دارد. اگر برآمدها دارای شانس یکسان نباشند چه؟ اگر تعداد برآمدها نامتناهی باشد چه؟ این سوال‌ها منجر به تعریف احتمال تجربی در قرن بیستم شد. این تعریف را اغلب به ریاضیدان آلمانی قرن بیستم، ریچارد فُن میسز[5] نسبت می‌دهند.

تعریف احتمال تجربی[6]:

فضای نمونه‌ای S و رخداد A را در نظر بگیرید. در هر بار آزمایش یا A و یا A^c ظاهر می‌شود. اگر این آزمایش را n بار انجام دهیم و n به اندازه کافی بزرگ شود، توقع داریم که نسبت تعداد دفعاتی که

---

[1] Geromaro Cardano
[2] Chevalier Demere
[3] Blaise Pascal
[4] Classical, or *a priori*, definition of probability
[5] Richard von Misses
[6] Empirical probability

رخداد A ظاهر شده است، m، به n به سمت احتمال A میل کند.

$$\text{Probability of A} = \text{P(A)} = \lim_{n\to\infty} \frac{m}{n}$$

تلاش بعدی برای تعریف احتمال، محصول طبیعی قرن بیستم بود که در آن ریاضیدانان سعی می‌کردند که تمام ریاضیات را براساس نظریه مجموعه‌ها بنا کنند. کار عمده را در این زمینه، آندره کولموگوروف[1] با چاپ کتاب «اصول نظریه احتمال» در سال ۱۹۳۳ انجام داد. به این ترتیب نظریه احتمال بر اساس ۴ اصل بنا شد.

اصل ۱: فرض کنیم A رخدادی باشد که بر روی S تعریف شده است. آنگاه $P(A)\geq 0$

اصل ۲: $P(S) = 1$

اصل ۳: فرض کنیم A و B دو رخداد ناسازگار بر روی S باشند (یعنی $A\cap B=\emptyset$). آنگاه:

$$P(A \cup B) = P(A) + P(B)$$

و اگرS نامتناهی باشد، اصل زیر نیز اضافه می‌شود:

اصل ۴: فرض کنیم ... $A_1, A_2, A_3,$ رخدادهایی تعریف شده بر روی S باشند. اگر برای هر i و j، $A_i \cap A_j = \phi$ باشد، آنگاه:

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

تا اینجا احتمال را به صورت شانس (که یک مفهوم عینی است) تعریف کردیم. اما در واقعیت ما معمولا از احتمال برای اندازه‌گیری اعتقاد خودمان به یک رخداد استفاده می‌کنیم. برای مثال پرسیده می‌شود که «احتمال اینکه تیم فوتبال ایران از ژاپن ببرد چقدر است؟». در یک سیستم تشخیص صحبت نیز هدف تولید سیستمی است که اعتقادش به اینکه سیگنال داده شده، فلان واج باشد درست باشد. ما در اینجا می‌خواهیم فرض کنیم که شانس یک مفهوم عینی و اعتقاد یک مفهوم ذهنی است. اما از نظر فلسفی این سوال مطرح است که آیا شانس واقعا وجود دارد (همانطور که طرفداران مکانیک کوانتوم می‌گویند) و یا اینکه مفهوم شانس از جهل ما نشات می‌گیرد (همانطور که انشتین و شرودینگر می‌گفتند). همچنین مشخص است که در بسیاری از مسائلی که ما با آنها سروکار داریم، عدم قطعیت از ناآگاهی ما نشات می‌گیرد و نه عدم قطعیت در طبیعت. برای مثال تاسی انداخته شده است و می‌دانیم که زوج آمده است. عدم قطعیت راجع به اینکه تاس دقیقا چه آمده است، از جهل ما نشات می‌گیرد نه از ناقص انداخته شدن

[1] Andrei Kolmogorov

تاس. به این ترتیب دیده می‌شود که هرچند ما در اینجا فرض می‌کنیم که شانس مفهومی عینی است و به جهان بیرون ربط دارد، اما از دیدی دیگر می‌توان گفت که هیچ مفهومی کاملا عینی نیست و تحت قضاوت بیننده و تفسیر او قرار گرفته است. در این پایان‌نامه با توجه به اینکه هدف ما بررسی روش تشخیص صحبت در انسان است، ما به تابع احتمال به عنوان تابع اعتقاد نگاه می‌کنیم و بررسی می‌کنیم که اعتقادی که تئوری احتمال به یک پدیده دارد تا چه حد به اعتقاد انسان شبیه است.

٤-١-١) اپراتورهای TNorm و SNorm [1]

در آینده نشان خواهیم داد که یک اندازه‌گیری فازی[2] تعمیمی از اندازه‌گیری احتمالی[3] (و حتی اندازه‌گیری نظریه شافر-دمپستر) است[68]. به این ترتیب می‌توان به نظریه احتمال به عنوان حالت خاص نظریه فازی نگریست. از این منظر، یکی از سوالاتی که مطرح می‌شود نوع اپراتورهای برهم‌نهی و فصل مشترک‌گیری این نظریه است. فرض کنیم A و B دو پیش‌آمد مستقل باشند. داریم:

$$P(A \cup B) = P(A) + P(B) - P(A).P(B)$$
$$P(A \cap B) = P(A).P(B)$$

بدین ترتیب دیده می‌شود که عملگر TNorm در نظریه احتمال ضرب جبری[4] و عملگر SNorm جمع جبری[5] است.

٤-٢) نظریه مدرک شافِر دمپستر[6]

خصوصیت اصلی این نظریه این است که اجازه نمایش جهل را می‌دهد. تئوری احتمال، اعتقاد را تنها به برآمدها می‌دهد ولی تئوری مدرک اجازه می‌دهد که مقداری از اعتقاد در سطح رخدادها بماند و به سطح برآمدها کشیده نشود.

مثال ۱: آیا در مدارهای نزدیک ستاره شعرای یمانی حیات وجود دارد یا خیر؟ ما راجع به این مساله چیزی نمی‌دانیم. می‌توانیم فضای تشخیص زیر را بسازیم:

$$\Theta = \{\theta_1, \theta_2\}$$

---

[1] البته به نظر نگارنده نمی‌توان این اپراتورها را بدون توجه به مفهومی که دارند استفاده کرد. برای مثال اپراتور ضرب جبری در نظریه احتمال تنها می‌تواند برای ترکیب دو پیش‌آمد مستقل به کار رود و مفهوم ترکیب چند اعتقاد با مفهوم اعتقادات اولیه متفاوت است. اما چون ما معمولا تعدادی نمونه آموزشی داریم که می‌خواهیم بر اساس آنها به یک اعتقاد کلی برسیم، فرض استقلال نمونه‌ها درست است و هدف از ترکیب اعتقادات پیدا کردن اعتقادی است که داده‌های آموزشی را بهتر خلاصه کند.

[2] Fuzzy measure
[3] Probability measure
[4] Algebraic product
[5] Algebraic sum
[6] Shafer-Dempster's theory of evidence

که در آن θ1 متناظر با امکان وجود حیات و θ2 متناظر با امکان عدم وجود حیات است.
ما همچنین می‌توانستیم فضای تشخیص زیر را بسازیم:

$$Z = \{\zeta_1, \zeta_2, \zeta_3\}$$

که در آن ζ_1 متناظر با امکان عدم وجود سیاره‌ای در نزدیکی ستاره شعرای یمانی، ζ_2 متناظر با امکان وجود سیاره‌ای در نزدیکی ستاره شعرای یمانی و عدم وجود حیات در این سیاره، و ζ_3 متناظر با امکان وجود حیات در نزدیکی ستاره شعرای یمانی است. در این تئوری داریم (Bel مخفف Belief است):

$$Bel(Z = \{\zeta_1, \zeta_2, \zeta_3\}) = 1$$
$$Bel(\{\zeta_1, \zeta_2\}) = 0$$
$$Bel(\{\zeta_1, \zeta_3\}) = 0$$
$$Bel(\{\zeta_2, \zeta_3\}) = 0$$
$$Bel(\{\zeta_1\}) = 0$$
$$Bel(\{\zeta_2\}) = 0$$
$$Bel(\{\zeta_3\}) = 0$$
$$Bel(\{\}) = 0$$

$$Bel(\Theta = \{\theta_1, \theta_2\}) = 1$$
$$Bel(\{\theta_1\}) = 0$$
$$Bel(\{\theta_2\}) = 0$$
$$Bel(\{\}) = 0$$

همانطور که دیده می‌شود، در اینجا تنها چیزی که می‌دانیم این است که یکی از برآمدها درست است.

تعریف: فرض کنید Θ یک مجموعه متناهی باشد و 2^Θ مجموعه توانی Θ را نشان دهد. آنگاه هر تابع اعتقاد بر روی Θ، $Bel : 2^\Theta \to [0,1]$، در شرایط زیر صدق می‌کند.

اصل ۱: $Bel(\emptyset) = 0$

اصل ۲: $Bel(\Theta) = 1$

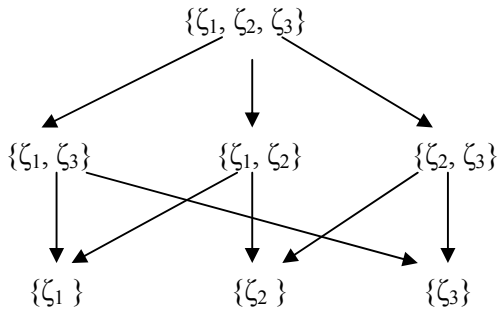
اصل ۳: برای هر عدد مثبت n و هر مجموعه $A_1,...A_n$ از زیرمجموعه‌های S داریم:

$$Bel(A_1 \cup ... \cup A_n) \geq \sum_i Bel(A_i) - \sum_{i>j} Bel(A_i \cap A_j) ... + (-1)^{n+1} Bel(A_1 \cap ... \cap A_n)$$

مجموعه Θ، مجموعه تمام حالات ممکن است که تنها یکی از آنها درست است و Bel(A) بیانگر میزان اعتقاد ما به این است که برآمد درست در A باشد (عضو مجموعه A باشد).

با استفاده از رابطه زیرمجموعه بودن، بین رخدادها یک ترتیب جزئی به دست می‌آید. به این ترتیب

شکلی به دست می‌آید که مجموعه تمام حالات ممکن در بالاترین نقطه آن و برآمدها در پایین‌ترین نقطه آن قرار دارند. شکل ۲۹ رابطه ترتیب جزئی را برای مجموعه Z نشان می‌دهد.



شکل ۲۹: رابطه ترتیب جزئی بین رخدادهای مختلف

تعریف: فرض کنید Θ یک مجموعه متناهی باشد. آنگاه هر تابع احتمال/اعتقاد اولیه بر روی Θ، $m: 2^{\Theta} \to [0,1]$، در شرایط زیر صدق می‌کند.

۱: $m(\emptyset) = 0$

۲: $\sum_{A \subseteq \Theta} m(A) = 1$

m(A) میزان اعتقاد دقیقا به خودِ A را نشان می‌دهد.

مشخص است که:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

به این ترتیب می‌توان دید که در این تئوری احتمال/اعتقاد به‌جای اینکه بین برآمدها توزیع شود، بین رخدادها توزیع شده است.

مثال ۲: مجموعه $\Theta = \{\gamma_1, \gamma_2, \gamma_3\}$ را در نظر بگیرید. یک نمونه از مقادیر تابع احتمال اولیه و تابع اعتقاد چنین است:

$m(\phi) = 0$	$Bel(\phi) = 0$
$m(\{\lambda_1\}) = 0.1$	$Bel(\{\lambda_1\}) = 0.1$
$m(\{\lambda_2\}) = 0.2$	$Bel(\{\lambda_2\}) = 0.2$
$m(\{\lambda_3\}) = 0.3$	$Bel(\{\lambda_3\}) = 0.3$
$m(\{\lambda_1, \lambda_2\}) = 0.05$	$Bel(\{\lambda_1, \lambda_2\}) = 0.1 + 0.2 + 0.05 = 0.35$
$m(\{\lambda_1, \lambda_3\}) = 0.15$	$Bel(\{\lambda_1, \lambda_3\}) = 0.1 + 0.3 + 0.15 = 0.55$
$m(\{\lambda_2, \lambda_3\}) = 0.2$	$Bel(\{\lambda_2, \lambda_3\}) = 0.2 + 0.3 + 0.2 = 0.7$
$m(\{\lambda_1, \lambda_2, \lambda_3\}) = 0$	$Bel(\{\lambda_1, \lambda_2, \lambda_3\}) = 1$

مدل‌سازی جهل باعث می‌شود که Bel(A) به تنهایی برای نمایش میزان اعتقاد ما به یک گزاره کافی نباشد. به این ترتیب برای توصیف بهتر میزان اعتقاد به یک گزاره، A، باید اعتقاد به نقیض آن گزاره، A^c، نیز مشخص باشد. توجه داریم که هر گزاره در حقیقت یک رخداد را مشخص می‌کند. اعتقاد به A^c میزان شک ما را در مورد A نشان می‌دهد.

$$Dou(A) = Bel(A^c)$$

به‌جای Dou(A) معمولا از کمیت $P^*(A)=1-Bel(A^c)$ استفاده می‌شود که بیانگر میزان سختی شک کردن در A است. از طرفی می‌توان نشان داد که $P^*(A) \geq Bel(A)$ و به همین دلیل به آن احتمال بالای A نیز گفته می‌شود. می‌توان نشان داد که:

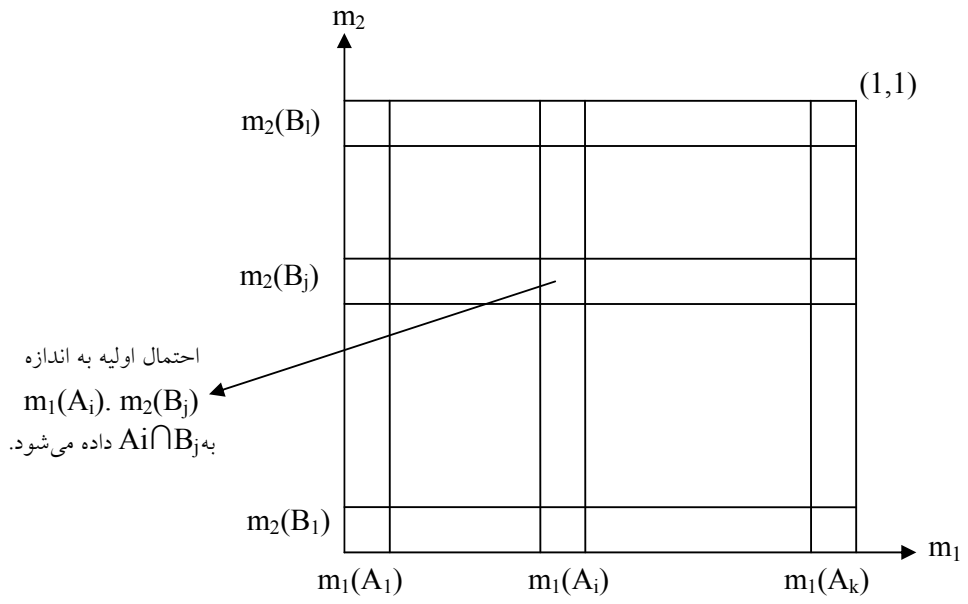
$$P^*(A) = \sum_{B \cap A \neq \phi} m(B)$$

یعنی احتمال بالای A، حداکثر مقدار اعتقادی که ممکن است به A پیدا شود را نشان می‌دهد. زیرا اعتقاد به یک مجموعه یعنی اعتقاد به اینکه گزینه صحیح عضو آن مجموعه باشد. پس حداکثر اعتقادی که می‌توان به مجموعه A داشت برابر مجموع اعتقادات تمام مجموعه‌هایی است که با مجموعه A حداقل یک عضو مشترک دارند. به این ترتیب زوج $<Bel(A), P^*(A)>$ مشخص کننده یک بازه هستند که در آن، Bel(A) اعتقاد کنونی به A را نشان می‌دهد و $P^*(A)$ حداکثر اعتقادی که ممکن است در آینده به A پیدا شود را نشان می‌دهد.

٤-٢-١) اپراتورهای TNorm و SNorm

ابتدا مساله ترکیب اعتقادات را در این نظریه بررسی می‌کنیم. فرض کنید A و B دو تابع اعتقاد باشند و تابع اعتقاد اولیه آنها m_A و m_B باشد. همچنین فرض کنید که $A_1 \ldots A_k$ و $B_1 \ldots B_l$ رخدادهای با اعتقاد اولیه غیر صفر از فضای قابل تشخیص[1] Θ باشند. در این صورت می‌توان شکلی مانند شکل ۳۰ کشید که در آن یک مربع به مساحت واحد به l.k خانه تجزیه شده است. حاصل‌ضرب دو اعتقاد عمودی و افقی به اشتراک مجموعه‌های آنها نسبت داده می‌شود. بدین ترتیب دیده می‌شود که در این نظریه نیز TNorm ضرب جبری است.

[1] Frame of discernment



شکل ۳۰: ترکیب اعتقادات در تئوری مدرک

اگر X زیر مجموعه Θ باشد، اعتقاد جدید به مجموعه X برابر مجموع اعتقادات تمام خانه‌هایی است که اعتقاد آنها به مجموعه X نسبت داده شده است (یعنی اشتراک مجموعه محور افقی با مجموعه محور عمودی X شده است). به این ترتیب دیده می‌شود که عملگر SNorm همان جمع معمولی (و نه جمع جبری) است. نکته دیگر اینکه ممکن است بخشی از اعتقاد به مجموعه تهی نسبت داده شود. در این صورت مجموع اعتقادات اولیه در تابع اعتقاد جدید برابر یک نمی‌شود. برای رفع این مشکل ضریبی در اعتقادهای به دست آمده به روش فوق ضرب می‌شود تا مجموع اعتقادات یک شود. در نهایت اعتقاد به مجموعه X چنین به دست می‌آید:

$$m(A) = \frac{\sum_{\substack{i,j \\ A_i \cap B_j = X}} m_1(A_i) m_2(B_j)}{1 - \sum_{\substack{i,j \\ A_i \cap B_j = \phi}} m_1(A_i) m_2(B_j)}$$

البته نگارنده پیشنهاد می‌دهد که از میزان اعتقاد به مجموعه تهی برای اعتقاد به غلط بودن فرض اولیه خود درباره وجود تمام اعتقادات در مجموعه Θ استفاده کنیم. این شبیه زمانی است که ما به اطلاعات اولیه خودمان شک می‌کنیم. می‌توان از این روش برای کشف اشیاء جدید استفاده کرد.

٤-۳) نظریه مجموعه‌های فازی

این نظریه در سال ۱۹۶۵ توسط آقای لطفی زاده ارائه شد. ریاضیات کلاسیک مبتنی بر نظریه

مجموعه‌های کلاسیک و منطق کلاسیک است که مبتنی بر دو مفهوم تعلق و عدم تعلق به مجموعه و درستی یا غلطی یک گزاره است[1]. این محدودیت‌ها در موارد زیر ما را از نظر مدل‌سازی دچار مشکل می‌کند.

۱- مفهومی واقعا مبهم است و تعریف روشنی ندارد. برای مثال مفهوم قد بلند را نمی‌توان با مجموعه‌های کلاسیک به درستی مدل کرد.

۲- اصلی مانند اصل عدم قطعیت هایزنبرگ بین *میزان درستی گفته*[2] و دقت[3] وجود دارد. وقتی ما مسافتی را بر حسب کیلومتر بیان می‌کنیم جمله ما معنی‌دارتر از زمانی است که آن را بر حسب میلی‌متر بیان می‌کنیم. زمانی که سیستم‌ها پیچیده می‌شوند، دیگر نمی‌توان دقت و درستی گفته را همزمان داشت. آقای زاده پیشنهاد می‌دهد که ما این رابطه را در وضعیتی متعادل‌تر قرار دهیم و جملات درست‌تری که خیلی‌هم دقیق نیستند بیان کنیم.

تعریف مجموعه فازی: اگر X مجموعه‌ای از اشیاء مانند x باشد، آنگاه یک مجموعه فازی مانند Φ در X به صورت مجموعه‌ای از زوج‌های مرتب چنین تعریف می‌شود:

$$\Phi = \left\{ \left(x, \mu_{\Phi}(x) \right) \middle| x \in X \right\}$$

که به $\mu_{\Phi}(x)$ تابع تعلق یا درجه تعلق x در Φ گویند. تابع تعلق مجموعه X را به فضای تعلق M (که معمولا [0,1] است) می‌نگارد.

یکی از مفاهیمی که در این پایان‌نامه دارای اهمیت زیادی است مفهوم اندازه‌گیری فازی است. می‌خواهیم ببینیم که تحت چه شرایطی درجه تعلق‌هایی که ما به زیرمجموعه‌های یک مجموعه مانند X نسبت می‌دهیم سازگارند (یعنی تناقض ندارند). در نظریه احتمال برای هر دو مجموعه ناسازگار[4] مانند A و B داشتیم:

$$P(A \cup B) = P(A) + P(B)$$

به این خاصیت، خاصیت جمع‌پذیری[5] گویند [22]. در حقیقت تفاوت اندازه‌گیری فازی با اندازه‌گیری کلاسیک در همین است که اندازه‌گیری فازی خاصیت جمع‌پذیری را ندارد.

---

[1] به هرحال از نظر نگارنده اینکه انسان بر اساس چنین منطقی صحبت می‌کند بسیار حائز اهمیت است.

[2] significance

[3] Precision

[4] دو مجموعه ناسازگارند اگر اشتراک آنها تهی باشد.

[5] Additivity

تعریف اندازه‌گیری فازی (از Sugeno):

تابع g را بر روی میدان بورل $\mathcal{B}$ که بر روی مجموعه X تعریف شده است یک اندازه‌گیری فازی گوییم هرگاه دارای خواص زیر باشد:

1. $g(0) = 0, g(X) = 1$
2. if $A,B \in \mathcal{B}$ and $A \subseteq B$, then $g(A) \leq g(B)$.
3. if $A_n \in \mathcal{B}, A_1 \in \mathcal{B}, \ldots,$ then $\lim_{n\to\infty} g(A_n) = g(\lim_{n\to\infty} A_n)$

همانطور که دیده می‌شود، بر خلاف نظریه احتمال دیگر نمی‌توان بر اساس خواص اندازه‌گیری $g(A \cup B)$ را بر اساس $g(A)$ و $g(B)$ به دست آورد. به همین دلیل مفهوم عملگرهای فازی بوجود می‌آید. ضمنا این مطلب نشان می‌دهد که اندازه‌گیری احتمالی حالت خاصی از اندازه‌گیری فازی است.
نکته مهم دیگر این است که این تعریف از اندازه‌گیری فازی منجر به بی‌معنا شدن مفهوم مقداری عدد می‌شود. در حقیقت یک عدد تنها مفهومی که دارد رابطه ترتیبی است که با بقیه اعداد دارد. به عنوان مثال درجه تعلق 10^{-6} بیانگر درجه تعلق کم نیست زیرا تنها چیزی که از تعریف اندازه‌گیری فازی نتیجه می‌شود این است که درجه تعلق 10^{-6} از درجه تعلق 1 کمتر است. در بخش ۴-۴-۲ ما روش اندازه‌گیری خود را ارائه می‌کنیم که در آن مقدار اعداد نیز دارای مفهوم است.

٤-٣-١) اپراتورهای TNorm و SNorm

همانطور که دیدیم از مفهوم اندازه‌گیری فازی نمی‌توان استفاده کرد و درجه تعلق زیرمجموعه‌های مجموعه جهانی X را به دست آورد. TNorm ها برای تعیین درجه تعلق $A \cap B$ و SNorm ها برای تعیین درجه تعلق $A \cup B$ استفاده می‌شوند.

تعریف TNorm: تابع t از $[0,1]\times[0,1]$ به $[0,1]$ که دارای خواص زیر است یک TNorm است:

1. $t(0,0) = 0; \qquad t(\mu_{\tilde{A}}(x),1) = \mu_{\tilde{A}}(x)$
2. $if\ \ \mu_{\tilde{A}}(x) \leq \mu_{\tilde{C}}(x)\ \ and\ \ \mu_{\tilde{B}}(x) \leq \mu_{\tilde{D}}(x)\ then$
   $t(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \leq t(\mu_{\tilde{C}}(x), \mu_{\tilde{D}}(x)) \quad (monotonicity)$
3. $t(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) = t(\mu_{\tilde{B}}(x), \mu_{\tilde{A}}(x)) \qquad (commutativity)$
4. $t(\mu_{\tilde{A}}(x), t(\mu_{\tilde{B}}(x), \mu_{\tilde{C}}(x))) = t((\mu_{\tilde{A}}(x), t(\mu_{\tilde{B}}(x)), \mu_{\tilde{C}}(x))$
   $(associativity)$

به طریق مشابه می‌توان SNorm ها را تعریف کرد. توجه داریم که $\max(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x))$ کمترین مقداری است که SNorm(A,B) می‌تواند به خود بگیرد و $\min(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x))$ بیشترین مقداری است

که TNorm(A,B) می‌تواند داشته باشد.

٤-٤) نظریه امکان

این نظریه برای اولین بار توسط آقای لطفی زاده در سال ۱۹۷۷ بر مبنای نظریه فازی ارائه شد[64]. برای یک بحث جامع راجع به نظریه امکان می‌توانید به [14] مراجعه نمایید. هدف این نظریه بررسی مفهوم امکان که یک مفهوم منطقی است به‌جای مفهوم احتمال است.

٤-٤-١) توزیع امکان

در این نظریه آقای زاده توزیع امکان را برای یک متغیر X که مقدارش را از مجموعه U می‌گیرد مشابه توزیع احتمال برای یک متغیر تصادفی تعریف می‌کند. فرض کنیم F یک زیر مجموعه فازی (مثلا بلند) از جهان مورد بحث[1] U={u} (مثلا قدهای بین ۰ تا ۳ متر) باشد که دارای تابع عضویت μ_F است. همچنین فرض کنید که X (مثلا قد علی) یک متغیر فازی است که مقادیرش را از مجموعه U می‌گیرد. در این صورت گزاره‌ای به شکل "X is F" (مثلا «قد علی» «بلند» است) دارای توزیع امکان Π_X است که می‌گوید امکان اینکه X مقدار u را بگیرد برابر $\mu_F(u)$ است. توجه نمایید که در عین حال متغیر X می‌تواند دارای یک توزیع احتمال نیز باشد که بسامد اینکه متغیر X مقدار u را بگیرد نشان می‌دهد.

٤-٤-١) امکان به معنای شدنی بودن

منظور ما از امکان یک پدیده میزان شدنی بودن آن است. بدین ترتیب مشخص است که مفهوم امکان با عملگر max پیوند خورده است (برای آنکه چیزی ممکن باشد، تنها یک راه کافی است؛ همان بهترین راه). آنچه آقای زاده بر آن تاکید دارد این است که عدم قطعیتی که در زبان طبیعی از آن صحبت می‌شود بیشتر جنبه امکانی دارد تا احتمالی. مثال زیر را در نظر بگیرید:

علی احتمالا[2] امروز به مدرسه نمی‌آید.

۱- **احتمال:** ۱۰۰۰ بار مادر علی مریض شده است و از این ۱۰۰۰ بار علی ۶۳۰ بار به مدرسه نیامد پس به احتمال ۰.۶۳ علی به مدرسه نمی‌آید.

۲- **امکان:** با توجه به مریضی مادر علی، برای او مشکل است که امروز به مدرسه بیاید.

در حقیقت در بسیاری موارد منظور ما از احتمال، امکان است. مثال‌های زیر رابطه بین احتمال و امکان را بیشتر روشن می‌کند:

۱- احتمال اینکه یک نقطه بر کنج یک مربع بیافتد صفر است ولی امکان دارد.

---

[1] Universe of discourse

[2] در زبان معمولا برای بیان عدم قطعیت از لغت احتمالا استفاده می‌شود ولی این لغت هیچ ارتباطی به نظریه احتمال ندارد.

۲- اگر چیزی غیر ممکن باشد، احتمال آن نیز صفر است. به این رابطه، ارتباط ضعیف بین امکان و احتمال می‌گویند.

٤-٤-٢) امکان به عنوان شروع یک منطق جدید برای اثبات ریاضی[1]

نکته دیگر در رابطه با نظریه امکان، رابطه آن با منطق است. وقتی انسان‌ها قضیه‌ای ریاضی را اثبات می‌کنند، به دنبال این هستند که نشان دهند برخی از حالاتی که در ذهن آنها وجود دارد ممکن و برخی غیر ممکن هستند. به همین دلیل:

۱- انسان‌ها باید متوجه شوند که کدام قسمت اثبات سخت است. اگر تمام قسمت‌های اثبات بدیهی باشند، ما می‌پرسیم که چه چیزی اثبات شد. در حالی که از نظر ریاضی نیازی به دانستن این حالات نیست.

۲- اگر انسان متوجه حالت خاصی نشود، اثبات غلط را نیز می‌پذیرد. به همین دلیل قضایای غلطی بوده‌اند که برای سال‌ها کسی متوجه غلط بودن آنها نشده بود.

٤-٤-٣) عملگرهای غیر قابل جبران min و max

یکی از ویژگی‌های مهم نظریه امکان عملگرهای این نظریه است که خاصیت غیرقابل جبرانی[2] دارند. می نیمم ۱۰۰۰ عدد برابر کوچک‌ترین آنها است. ما به اینکه ۹۹۹ نفر می‌گویند ۱ کاری نداریم و اعتقادمان را بخاطر یک نمونه صفر می‌کنیم. از نظر منطقی وقتی یک‌جای کار خراب است، کل کار را می‌تواند خراب کند. همچنین اگر یک راه برای انجام کاری باشد، آن کار شدنی است (امکانش ۱ است). به همین دلیل ما از این عملگرها برای کار با مفهوم امکان استفاده می‌کنیم. آقای لطفی‌زاده در [64] بر روی این خاصیت اپراتورهای تئوری امکان تاکید زیادی می‌کند.

٤-٤-٤) مساله بازشناسی گفتار: امکان یا احتمال

نکته دیگر رابطه تئوری امکان با تشخیص صحبت است. مساله ما در حقیقت پیدا کردن پاسخ این سوال است که آیا سیگنال داده شده می‌تواند فلان جمله باشد یا خیر؟ تلاش نگارنده این است که سیستمی بسازد که تنها موارد ممکن را تشخیص دهد.

در حقیقت مساله ما یافتن محتمل‌ترین واجی که ممکن است سیگنال داده شده را تولید کرده باشد نیست. در حقیقت این مساله بسیار سخت است و نیازمند مقایسه سیگنال داده شده با تمام گروه‌های شناخته شده است. مساله ما ساده‌تر از این است. ما به دنبال تمام برداشت‌های ممکن از سیگنال داده شده هستیم. در عمل چون تعداد کمی برداشت ممکن (معمولا یکی) بیشتر وجود ندارد، این مساله ساده‌تر

---

[1] به نظر می‌رسد که این برداشت از نظریه امکان کار نگارنده است.

[2] Noncompensability

است.

٤-٤-٥) نظریه امکان یک نظریه دوبانده مناسب برای مدل‌سازی جهل

یکی از ویژگی‌های مهم نظریه امکان توانایی این نظریه در مدل‌سازی جهل است. بگذارید برخورد نظریه امکان را با چیزی که راجع به آن اطلاعی ندارد بررسی کنیم:

سوال: آیا حیات در ستاره شعرای یمانی وجود دارد؟

پاسخ: ممکن است داشته باشد و ممکن است نداشته باشد.

در نظریه امکان می‌توان به یک گزاره و نقیض آن همزمان اعتقاد داشت (اعتقادی از نوع ممکن است). به عبارت دیگر مجموع Bel(A) و Bel(¬A) یک نیست. در نظریه امکان مفهومی به نام الزام[1] تعریف می‌شود که به نظر می‌رسد تنها در حالتی که اعتقاد محدود به ۰ و ۱ باشد درست کار می‌کند. در حقیقت ما ترجیح می‌دهیم که به نظریه امکان اجازه کار با هر مقدار Bel(A) و Bel(¬A) را بدهیم.

٤-٤-٦) اندازه‌گیری امکانی[2] و عملگرهای TNorm و SNorm

اندازه‌گیری امکانی بر روی مجموعه X به عنوان نوع خاصی[3] از اندازه‌گیری فازی به صورت تابع $\Pi : 2^X \to [0,1]$ چنین تعریف می‌شود:

1. □(0) = 0, □ (X) = 1
2. if A⊆B then □ (A) ≤ □ (B).
3. $\Pi\left(\bigcup_{i\in I} A_i\right) = \sup \Pi(A_i)$ *with index set I*

در حقیقت در این نظریه از SNorm ماکزیمم‌گیری استفاده شده است. جالب است که اندازه‌گیری امکانی محدودیتی بر روی روش محاسبه اشتراک دو مجموعه معرفی نمی‌کند. در حقیقت اگر ما اصراری نداشته باشیم که قانون دمورگان درست باشد، می‌توانیم از TNorm دیگری غیر از min استفاده کنیم. این نکته بسیار مهمی است زیرا بسیاری از مسائل بهینه‌سازی به شکل Max-Product (یعنی ماکزیمم‌گیری بر روی ضرب چند عدد) ظاهر می‌شوند و نه Max-Min (در حقیقت مساله بازشناسی گفتار [36][13] نمونه‌ای از همین مساله است). یکی از معایب نظریه امکان این است که مقدار اعداد در این نظریه نیز بی‌معنی است.

٤-٤-٧) اندازه‌گیری امکانی پیشنهادی نگارنده[4]

ما معتقدیم که یک اندازه‌گیری خوب باید دارای ویژگی‌های زیر باشد:

---

[1] Necessity

[2] Possibility measure

[3] [68] مطالب متناقضی را در مورد اینکه آیا یک اندازه‌گیری امکانی نوع خاصی از توزیع اندازه‌گیری فازی است یا خیر بیان می‌کند.

[4] Possibility measure

۱- یک اندازه‌گیری فازی باشد.
۲- مقادیر اندازه‌گیری شده باید دارای یک مفهوم جهانی باشند. مقدار 0.1 باید نمایانگر یک درجه تعلق پایین و مقدار 0.9 باید نشان‌گر درجه تعلق بالا باشد.
۳- هم در جهان مورد اندازه‌گیری و هم در مقدار درجه تعلق نباید نیاز به دقت‌های بالا باشد.

در حقیقت توابع عضویت فازی این نقش را در اندازه‌گیری فازی معمولی ایفا می‌کنند. از آنجا که این توابع عضویت معمولا در رابطه با متغیرهای «خیلی کم»، «کم»، «متوسط»، «زیاد» و «خیلی زیاد» هستند، می‌توان گفت که آنها دارای مفهومی جهانی هستند. ولی یافتن توابع عضویت در جهانی مانند X با مشکلات زیر همراه است:

۱- نیازمند جستجوی تابع عضویت مناسب در اعداد حقیقی است.
۲- نیازمند تعریف دقیق تابع عضویت است.
۳- این روش ما را مجبور به استفاده از روش‌های مبتنی بر مدل می‌کند.

مجموعه جهانی X را در نظر بگیرید و فرض کنید که بر روی آن اندازه‌گیری امکانی □ انجام شده است. بدین ترتیب می‌دانیم که x های شدنی‌تر دارای امکان بیشتری هستند. حال فرض کنید که P_x یک اندازه‌گیری احتمالی بر روی جهان X است و P_y یک اندازه‌گیری احتمالی بر روی مقادیری است که □ به نقاط مجموعه X نسبت می‌دهد. ما ابتدا مقادیر مجموعه X را به اعداد حقیقی، ℝ، می‌نگاریم. در عمل معمولا X همان ℝ است و نیازی به این مرحله نیست. در ادامه فرض می‌کنیم که X=ℝ.

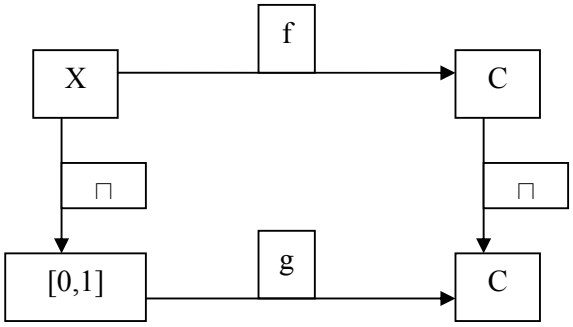
فرض کنیم که C مجموعه اعداد صحیح بین ۰ تا ۹۹ باشد[۱]. فرض کنید f تابعی باشد که مقادیر ℝ را به اندیس نزدیک‌ترین صدک p_x می‌نگارد. یادآوری می‌کنم که صدک i نقطه‌ای مانند t است که $P_x(X \le t) = \frac{i}{100}$. بطور مشابه g را تابعی تعریف کنید که مقادیر [0,1] را به نزدیک‌ترین صدک در p_y می‌نگارد.

ما یک اندازه‌گیری امکانی اصلاح‌شده □ را بر روی مجموعه C (که شامل اعداد صحیح بین ۰ تا ۹۹ است) و با مقادیر امکان C چنین تعریف می‌کنیم:

$$\Psi(i) = g(\sup_{i \in C}\{\Pi(x) \mid f(x) = i\})$$

این بدان معنا است که ما به تک‌نمونه‌هایی که احتمالی کمتر از 1% دارند علاقمند نیستیم. شکل ۳۱ نحوه محاسبه □ را نشان می‌دهد.

[۱] در صورت نیاز می‌توان دقت را افزایش داد. ولی ما فکر می‌کنیم که داده‌ای که احتماتل آن کمتر از ۱٪ است قابل صرف نظر است.



شکل ۳۱: رابطه بین Ψ (اندازه‌گیری اصلاح‌شده) و Π (اندازه‌گیری امکانی).

٤-٥) آمار چیست؟

به علت سیطره نظریه احتمال بر دنیای عدم قطعیت، در ذهن بسیاری از افراد مفهوم آمار[1] با مفهوم احتمال پیوند خورده است. واقعیت این است که آمار، اطلاعاتی است که به نوعی خلاصه‌ای از داده‌های زیادی را نگه می‌دارد. اینکه ما چگونه بر اساس آمار عقایدمان را تنظیم می‌کنیم هیچ ارتباطی به آمار ندارد. در حقیقت نظریه احتمال، یک روش برای بهنگام‌سازی عقاید ما بر اساس آمار ارائه می‌دهد. نظریه مدرک، نظریه فازی و نظریه امکان نیز هر یک روش‌های آماری خود را دارند. همچنین بسیاری از روش‌های مورد علاقه نگارنده (مانند صدک‌ها) در حقیقت روش‌های آماری هستند و نه احتمالی.

ذکر این نکته ضروری است که از نظر نگارنده میانگین و واریانس روش‌های مناسبی برای نگهداری خلاصه‌ای از اطلاعات نیستند. در حقیقت پیشنهاد نگارنده[2] این است که به‌جای میانگین و واریانس، خط سیرها ذخیره شوند. ذخیره خط سیرها دارای این برتری است که رابطه بین ویژگی‌های مختلف را نیز حفظ می‌کند. به عنوان مثال برای ذخیره خلاصه‌ای از انواع موجودات، بهتر است نمونه‌هایی از آنها را نگه داریم تا اینکه از آنها میانگین بگیریم!.

[1] Statistics

[2] این پیشنهاد قبلا توسط استاد راهنمای پروژه داده شده است و نگارنده نیز مدعی است که توانسته است اهمیت این پیشنهاد را درک کند.

فصل ۵
پردازش سیگنال

روش Add-Overlap **برای ترکیب تغییرات اعمال شده در قاب‌ها**

تغییر سیگنال صحبت برای رسیدن به شکل مشخصی در فضای بانک فیلتر

روش اول: پخش انرژی بانکهای فیلتر

روش دوم: ضرب طیف در ۲۵ فیلتر

۵-۱) روش *Add-Overlap* برای ترکیب تغییرات اعمال شده در قاب‌ها*[10]*

فرض کنید سیگنال s(n) به یک دنباله از سیگنال‌های کوچک دارای همپوشانی $s_m(n)$ تجزیه شده است و داریم:

$$s_m(n) = h(n)s(n+mK)$$

که در آن h(n) یک پنجره تحلیل (مثلا hanning)[۲ [8] [67]] به طول N نمونه و متقارن حول n=0 است و K پرش بین قاب‌ها است.

پس از اعمال تغییر (مثلا کوانته کردن انرژی در هر فرکانس) در هر قاب، سیگنال‌های $\hat{s}_m(n)$ به دست می‌آیند که باید از روی آنها سیگنال $\hat{s}(n)$ به دست آید. سیگنال $\hat{s}(n)$ از روی سیگنال‌های $\hat{s}_m(n)$ چنین محاسبه می‌شود (یادآوری: x(n-t) برابر شیفت x(n) به اندازه t واحد به سمت راست است):

$$\hat{s}(n) = \frac{\sum_m \hat{s}_m(n-mK)h(n-mK)}{\sum_m h^2(n-mK)}$$

در حالت حدی اگر $\hat{s}_m(n) = s_m(n)$ (یعنی تغییری در سیگنال ندهیم) آنگاه داریم:

$$\hat{s}(n) = \frac{\sum_m \hat{s}_m(n-mK)h(n-mK)}{\sum_m h^2(n-mK)}$$

$$= \frac{\sum_m s_m(n-mK)h(n-mK)}{\sum_m h^2(n-mK)}$$

$$= \frac{\sum_m \left[h(n-mK)s(n)\right]h(n-mK)}{\sum_m h^2(n-mK)}$$

$$= \frac{\sum_m s(n)\,h^2(n-mK)}{\sum_m h^2(n-mK)}$$

$$= s(n)\frac{\sum_m h^2(n-mK)}{\sum_m h^2(n-mK)}$$

پس اگر h(n) دارای نقاط صفر با تناوب K نباشد داریم: $\hat{s}(n) = s(n)$.

۵-۲) تغییر سیگنال صحبت برای رسیدن به شکل مشخصی در فضای بانک فیلتر[1]

در این بخش الگوریتمی ارائه می‌شود که به ما اجازه دستکاری سیگنال صحبت را در فضای بانک فیلتر می‌دهد. بدین ترتیب ما می‌توانیم تاثیر ویژگی‌های مختلف را در صدایی که شنیده می‌شود بررسی کنیم. ابتدا روش محاسبه ضرایب بانک فیلتر را توضیح می‌دهیم. فرض کنیم سیگنال صحبت s(n) داده شده است. برای به دست آوردن ضرایب بانک فیلتر ابتدا سیگنال به یک دنباله از سیگنال‌های کوچک دارای همپوشانی $s_m(n)$ تجزیه می‌شود:

$$s_m(n) = h(n)s(n + mK)$$

که در آن h(n) یک پنجره تحلیل (مثلا hanning) به طول N نمونه و متقارن حول n=0 است و K پرش بین قاب‌ها است. ادامه بحث برای هر یک از سیگنال‌های کوچک $s_m(n)$ است.

برای هر سیگنال کوچک $s_m(n)$ کارهای زیر انجام می‌شود:

۱- محاسبه تبدیل فوریه.

۲- به دست آوردن انرژی در هر فرکانس.

۳- فیلتر کردن انرژی با ۲۵ فیلتر مختلف و محاسبه انرژی عبوری از هر فیلتر. این فیلترهای در دو مقیاس Mel و Bark هستند.

۴- محاسبه لگاریتم انرژی در هر فیلتر. به اعداد حاصل، ضرایب بانک فیلتر می‌گویند.

تا اینجا ضرایب بانک فیلتر محاسبه شد. حال می‌خواهیم سیگنال صحبت را در فضای این ضرایب تغییر دهیم. مساله این است که چه تغییری در سیگنال $s_m(n)$ بوجود آوریم تا ضرایب بانک فیلتر مطلوب به دست آید. همچنین می‌خواهیم سیگنال صحبت را طوری تغییر دهیم که حتی‌الامکان صدا طبیعی بماند. مشخص است که برای طبیعی ماندن سیگنال صحبت باید فاز آن حفظ شود.

۵-۲-۱) روش اول: پخش کردن انرژی بانک‌های فیلتر

چون فاز سیگنال نباید تغییر کند ما تنها مجاز به تغییر در دامنه مؤلفه‌های فرکانسی هستیم. یکی از روش‌های پیدا کردن دامنه مؤلفه‌های فرکانسی، پخش کردن انرژی موجود در هر فیلتر بین تمام فرکانس‌های شرکت‌کننده در آن فیلتر و سپس محاسبه تبدیل فوریه معکوس است. مشکل این روش این است که یک pitch مصنوعی تولید می‌کند. علت این است که ما از یک پنجره به طول ثابت (مثلا 20ms) برای محاسبه طیف سیگنال استفاده می‌کنیم. قضیه زیر نشان می‌دهد که اگر اندازه پنجره دوبرابر

---

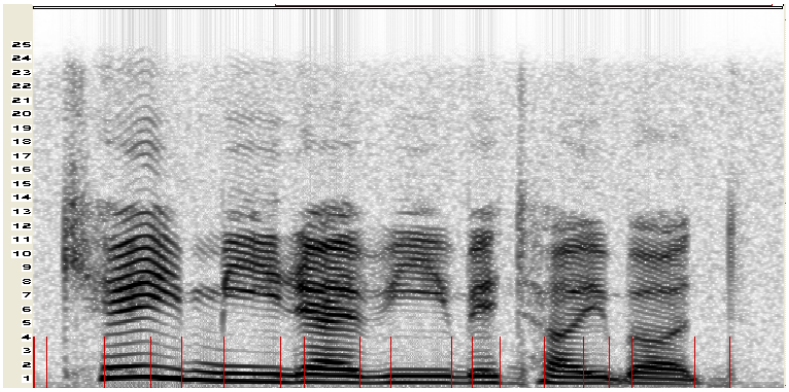
[1] راه حلی که در اینجا ارائه شده است توسط نگارنده به دست آمده است. متاسفانه به علت عدم احساس نیاز دانشمندان علم پردازش صوت به شناخت روش انسان، این مقاله پذیرفته نشده است.

فرکانس طبیعی باشد، انرژی در فرکانس‌های فرد صفر است.

قضیه: فرض کنیم که سیگنال x(n) تناوبی و با تناوب N نمونه است. اگر $X_N(n)$ تبدیل فوریه x(n) به اندازه N باشد، آنگاه:

$$X_{2N}(n) = \begin{cases} X_N(n/2) & if\ \ n\ \ is\ even \\ 0 & if\ \ n\ \ is\ odd \end{cases}$$

شکل ۳۲ نیز رگه‌هایی که به همین دلیل در طیف سیگنال بوجود آمده‌اند را نشان می‌دهد. ولی روش ذکر شده باعث می‌شود که انرژی به داخل فرکانس‌های فرد[۱] نیز رسوخ کند. در نتیجه وقتی به زمان برگردیم مشاهده می‌کنیم که pitch صدا نصف شده است. یک راه حل برای رفع این مشکل این است که DFT را در اندازه متناسب با pitch محاسبه کنیم. روش دیگر این است که نسبت انرژی در مؤلفه‌های فرکانسی مجاور را نیز همانند فاز حفظ کنیم.



شکل ۳۲: طیف فایل s11881.wav از دادگان فارس‌دات. رگه‌های انرژی در این شکل دیده می‌شود.

۵-۲-۲) روش دوم: ضرب طیف در ۲۵ فیلتر

تاکنون نشان داده‌ایم که فاز و نسبت مؤلفه‌های فرکانسی مجاور باید حفظ شوند. بنابراین فرض می‌کنیم که هر مؤلفه فرکانسی در ضریبی ضرب می‌شود. اگر فرض کنیم که فرکانس مرکزی فیلترها تنها در همان فیلتر حضور دارند، می‌توان با تغییر دامنه فرکانس مرکزی هر فیلتر و حذف تمام فرکانس‌های دیگر به سیگنال مطلوب رسید. اما این روش صدایی غیر طبیعی تولید می‌کند.

ما برای حل این مساله به هر یک از ۲۵ فیلتر یک ضریب نسبت دادیم. ضریب هر فرکانس برابر مجموع وزنی ضرایبی است که از فیلترهای مختلف به دست می‌آورد. بنابراین مساله به یافتن ضرایب مناسب برای فیلترها ساده می‌شود. برای رسیدن به سیگنال مطلوب، سیگنال‌ها را در قاب‌ها تغییر می‌دهیم و از

[۱] توجه نمایید که مشابه این قضیه برای $X_{3N}(n)$ و ... نیز وجود دارد. منظور ما از فرکانس‌های فرد، فرکانس‌هایی است که طبق یکی از این قضایا انرژی در آنها صفر یا تقریبا صفر است.

روش Add-Overlap برای ترکیب قاب‌ها استفاده می‌کنیم. به همین دلیل در ادامه بحث اندیس زمان را حذف کرده‌ایم. ابتدا علامت‌گذاری را تعریف می‌کنیم:

x(n): سیگنال

N: اندازه قاب

X[n]: انرژی در فرکانس n ام.

fb[i]: مقدار بانک فیلتر i ام که برابر لگاریتم انرژی عبوری از فیلتر iام است.

dfb[i]: مقدار مطلوب فیلتر iام.

c[i]: dfb[i]/ fb[i]

w[i,x]: مقدار فیلتر i ام در مؤلفه فرکانسی xام

α[i] : ضریب فیلتر iام که همواره بزرگ‌تر از صفر است.

ما فرض می‌کنیم که فیلترها مثلثی هستند و مرکز هر فیلتر پایان فیلتر قبلی و شروع فیلتر بعدی است. بعلاوه ما به دلایلی که در [23] ذکر شده است، به‌جای محاسبه log(x) از log(1+Jx) استفاده می‌کنیم. پس از یافتن ضرایب α[i]، مقدار جدید انرژی در هر فرکانس از فرمول زیر محاسبه می‌شود:

$$X[i]_{new} = X[i]_{old}.coef[i],$$

که coef[i] از فرمول زیر به‌دست می‌آید.

$$coef[x] = \sum_{i=1}^{\max\_band} w[i,x].\alpha[i]$$

مساله یافتن α[i]ها به نحوی است که برای هر فیلتر i داشته باشیم:

$$c[i]\,fb[i]_{old} = fb[i]_{new}$$

$$c[i]\log\left(1+\sum_{n=1}^{N} X[n]_{old}\,w[i,n]\right)$$

$$= \log\left(1+\sum_{n=1}^{N} X[n]_{new}\,w[i,n]\right)$$

همانطور که قبلا ذکر شد، ما بجای log(x) از log(1+Jx) استفاده می‌کنیم.

$$1+J\sum_{n=1}^{N} X[n]_{new}\,w[i,n] = \exp(c[i].\,fb[i]_{old})$$

$$\sum_{n=1}^{N} X[n]_{new}\,w[i,n] = \frac{\exp(c[i].\,fb[i]_{old})-1}{J}$$

$$\sum_{n=1}^{N} X[n]_{new}\,w[i,n] = \beta[i]\sum_{n=1}^{N} X[n]_{old}\,w[i,n]$$

که β[i] چنین تعریف شده است:

$$\beta[i] = \frac{\exp(c[i].fb[i]_{old}) - 1}{J\sum_{n=1}^{N} X[n]_{old}\, w[i,n]}$$
$$= \frac{\exp(c[i].fb[i]_{old}) - 1}{\exp(fb[i]_{old}) - 1}$$

حال ما باید ضرایب α[i] را به‌دست آوریم:

$$\beta[i]\sum_{n=1}^{N} X[n]_{old}\, w[i,n] = \sum_{n=1}^{N} X[n]_{new}\, w[i,n]$$
$$= \sum_{n=1}^{N} coef[i].X[n]_{old}\, w[i,n]$$

برای محاسبه $X[i]_{new}$، باید توجه داشت که هر فرکانسی دقیقا از دو فیلتر عبور می‌کند. فرض کنیم CF[i] فرکانس مرکزی فیلتر iام باشد. بنابراین:

$$\beta[i]\sum_{n=1}^{N} X[n]_{old}\, w[i,n] = \sum_{n=1}^{N} coef[i].X[n]_{old}\, w[i,n]$$
$$= \sum_{n=1}^{CF[i]} (\alpha[i]w[i,n] + \alpha[i-1]w[i-1,n]).X[n]_{old}\, w[i,n]$$
$$+ \sum_{n=CF[i]}^{N} (\alpha[i]w[i,n] + \alpha[i+1]w[i+1,n]).X[n]_{old}\, w[i,n]$$
$$= \sum_{n=1}^{CF[i]} (\alpha[i]w[i,n] + \alpha[i-1](1 - w[i,n])).X[n]_{old}\, w[i,n]$$
$$+ \sum_{n=CF[i]}^{N} (\alpha[i]w[i,n] + \alpha[i+1](1 - w[i,n])).X[n]_{old}\, w[i,n]$$

حال چند متغیر جدید تعریف می‌کنیم:

$$A^1[i] = \sum_{n=1}^{CF} X[n]_{old}\, w[i,n]$$
$$A^2[i] = \sum_{n=1}^{CF} X[n]_{old}\, w[i,n]^2$$
$$B^1[i] = \sum_{n=CF}^{N} X[n]_{old}\, w[i,n]$$
$$B^2[i] = \sum_{n=CF}^{N} X[n]_{old}\, w[i,n]^2$$

و در نهایت به رابطه زیر می‌رسیم:

$$\beta[i]\left(A^1[i] + B^1[i]\right) = \alpha[i]\left(A^2[i] + B^2[i]\right)$$
$$+ \alpha[i-1]\left(A^1[i] - A^2[i]\right) + \alpha[i+1]\left(B^1[i] - B^2[i]\right)$$

توجه داریم که α[0] و α[last_band+1] طبق تعریف صفر هستند. چون α[i] نمی‌تواند منفی باشد، یافتن فرمولی که مقادیر دقیق α[i] را بدهد کار دشواری است. در حقیقت اگر ما به دنبال تغییرات شدیدی در ضرایب بانک فیلتر باشیم، ممکن است اصلا چنین α[i] ای وجود نداشته باشد. بنابراین ما

مقدار اولیه $\alpha[i]$ها را صفر می‌کنیم و از فرمول فوق برای بهتر کردن تقریب خود استفاده می‌کنیم.

فصل ۶
بخش‌بندی سیگنال صحبت

۶-۱) مروری بر روش‌های بخش‌بندی سیگنال صحبت

بخش‌بندی سیگنال صحبت به عنوان واحدی که استخراج ویژگی بر اساس آن انجام می‌شود، مورد نیاز تمام سیستم‌های بازشناسی گفتار می‌باشد. در اینجا بخش را بازه‌ای از سیگنال صحبت تعریف می‌کنیم که ویژگی‌ها در آن بازه استخراج می‌شوند. به این ترتیب بخش با واحد آموزشی متفاوت است. سیستم‌های بازشناسی گفتار، آموزش را بر اساس یکی از واحدهای زیر انجام می‌دهند (واحدهای آموزشی همواره واحدهای آوایی هستند- حتی اگر بدون ناظر یادگرفته شوند):

۱- واج

۲- دو-واجی[1]: در این حالت از وسط یک واج تا وسط واج بعدی مدل می‌شود.

۳- سه-واجی[2]: در این روش، از نیمه واج اول تا نیمه واج سوم مدل می‌شود.

۴- سیلاب[49] [9]

۵- واحدهای آوایی قابل شنیدن و با بسامد حضور مناسب: در این پروژه، واحدهای آوایی از ابتدا مشخص نیستند و واحدهای آموزشی در فرآیند یادگیری تعیین می‌شوند. یک واحد آموزشی ممکن است یک تک‌واج و یا یک دوواجی باشد.

اما با توجه به تعریف ما از بخش، مشخص است که در سیستم‌های متداول، بخش عبارت است از چندین قاب مجاور که ویژگی‌های انرژی، مشتق اول و مشتق دوم از آنها قابل استخراج هستند. البته عمده اطلاعات از فریم مرکزی استخراج می‌شود و به همین دلیل گفته می‌شود که در سیستم‌های متداول از اطلاعات زمانی به خوبی استفاده نشده است[26][18][27].

در اینجا ما به دنبال استخراج ویژگی‌ها از بخش‌های بامعناتر و همچنین دارای دقت مناسب در زمان و فرکانس هستیم و به همین دلیل بخش‌بندی برای ما اهمیت زیادی دارد. روش‌های بخش‌بندی عبارتند از:

۱- **بخش‌بندی برای یافتن واحدهای آوایی (واج-دو-واجی-سه-واجی-سیلاب):** نگاه به شکل ۳۳ نشان می‌دهد که رابطه شدیدی بین گذر از یک واحد آوایی به واحد آوایی دیگر و تغییرات سیگنال صحبت وجود دارد. برخی از روش‌های بخش‌بندی به دنبال یافتن واحدهای آوایی بدون توجه به اطلاعات زبانی هستند. یکی از کاربردهای این روش، تشخیص صحبت در محیط‌های چندزبانی است که استفاده از اطلاعات زبان برای بخش‌بندی مشکل است. آزمایش‌های انجام شده بر روی افرادی که زبان خاصی را بلد نیستند نشان می‌دهد که انسان‌ها می‌توانند یک زبان خارجی را در حد سیلاب تقطیع کنند[59]. حسن دیگر این نوع بخش‌بندی، مقاومت آن نسبت به نویزهایی است

---

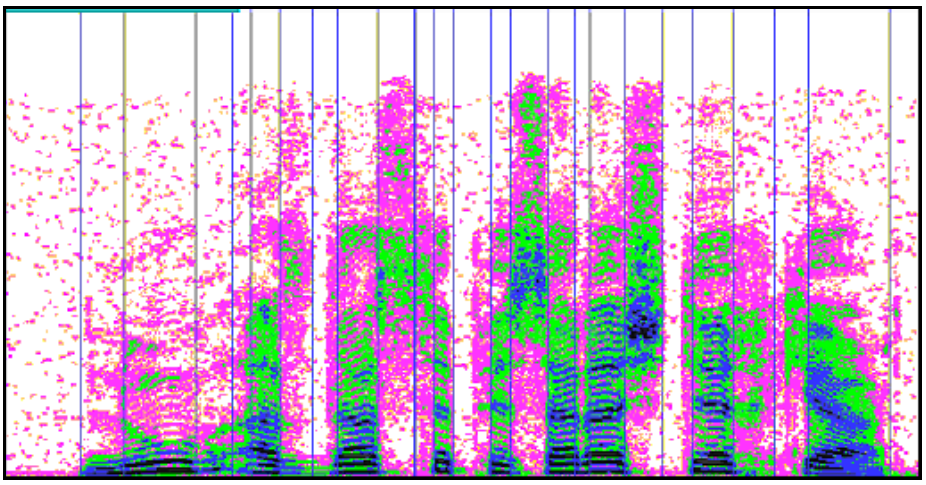
[1] Diphone
[2] Triphone

که بر روی تغییرات طیف تاثیر زیادی ندارند (مانند نویز میکروفن). از آنجا که خطاهای بخش‌بندی بویژه خطای حذف تقریبا دیگر قابل جبران نیستند[46] [61]، سیستم‌های بخش‌بندی نمی‌توانند به عنوان بخش اولیه یک سیستم بازشناسی گفتار استفاده شوند. به همین دلیل این سیستم‌ها چندین برابر (معمولا ۱ تا ۷ برابر) مرزهای واقعی، مرز بین واجی پیدا می‌کنند و تعیین دقیق بخش‌بندی را به مراحل بعد واگذار می‌کنند.

۲- بخش‌بندی بر اساس واحد های شنوایی[1]

- **یک بخش پایدار از سیگنال:** در این حالت هر بخش متناظر با یک حالت پایدار در سیگنال صحبت است و مرز بین دو بخش بیانگر یک تغییر است. مشکل این روش این است که بسیاری از واج‌ها (مانند «د» و «ب») تنها یک تغییر هستند و این مدل‌سازی برای آنها مناسب نیست.
- **تغییرات بین دو بازه پایدار سیگنال:** در این حالت هر بخش متناظر با یک تغییر در سیگنال صحبت است و مرز بین دو بخش بیانگر یک حالت پایدار است[61].
- **شیئ:** در این پایان‌نامه سعی‌شده است که بخش‌بندی به روش انسان در تشخیص صحبت نزدیک باشد. شیئ کوچک‌ترین بخش از سیگنال صحبت است که انسان می‌تواند راجع به آن صحبت کند. برای مثال، یک شیء می‌تواند صدای جهش در حرف «د» و یا تحریرها در هنگام آواز خواندن باشد.

در ادامه سه روش بخش‌بندی صحبت که از بین منابع دیده شده انتخاب شده‌اند، شرح داده می‌شوند.



شکل ۳۳: در این شکل رابطه بین تغییرات در طیف و گذر بین واج‌ها به خوبی دیده می‌شود.[2]

[1] Auditory units

[2] جمله S110.wav از دادگان فارس‌دات: «نوح از دست پسرش دق کرد»

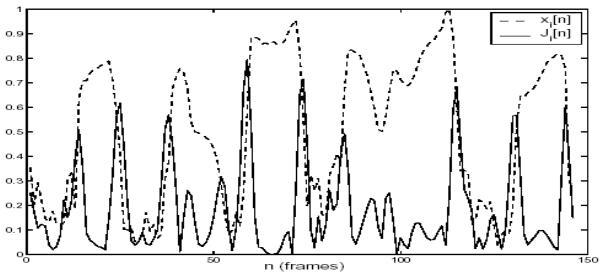
۶-۱-۱) بخش‌بندی پایگاه داده TIMIT با دقت ۷۴ درصد[7]

این مقاله روشی برای بخش‌بندی سیگنال صحبت بر اساس واحد آوایی واج ارائه می‌کند. این روش چون اطلاعاتی از واج‌های زبان ندارد (مستقل از زبان است)، در سدد یافتن مرز بین واج‌ها بر اساس یافتن قاب‌هایی است که در آنها ویژگی‌های مرحله پیش‌پردازش به سرعت و به مقدار زیادی تغییر می‌کنند. به همین دلیل برای هر ویژگی مانند $x_i[n]$، تابع پرش چنین تعریف می‌شود:

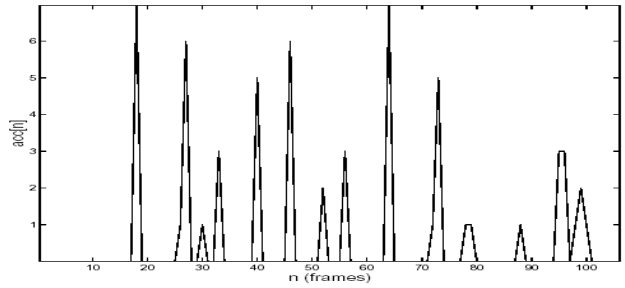
$$J_i^a[n] = \left| \sum_{m=n-a}^{n-1} \frac{x_i[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x_i[m]}{a} \right|$$

که پارامتر a پارامتری است که باید تنظیم شود. فرمول فوق فاصله بین مقدار متوسط ویژگی x_i را در a فریم قبل با a فریم بعد نشان می‌دهد (یکی از بهبودهایی که ما به این الگوریتم دادیم، این بود که اهمیت نقاط نزدیک به فریم فعلی را از نقاط دور بیشتر کردیم). شکل ۳۴ مقدار یک ویژگی و تابع پرش آن را نشان می‌دهد. ارتفاع هر قله در تابع $J_i^a[n]$ برابر اختلاف ارتفاعش با نزدیک‌ترین دره تعریف می‌شود. یک قله معتبر است اگر و فقط اگر ارتفاع آن از یک مقدار آستانه b بیشتر باشد. البته می‌توان قبل از اعمال آستانه، مقدار تابع $J_i^a[n]$ را بین [0,1] نرمال کرد. در این مقاله از ویژگی‌های PLP [24] استفاده شده است.

مرحله بعدی الگوریتم، ترکیب قله‌های یافت شده در خط سیر ویژگی‌های مختلف به منظور تعیین مرزهای واج‌ها است. بدین منظور، از بین هر c قاب مجاور، یک قاب به عنوان محتمل‌ترین نقطه به عنوان مرز انتخاب می‌شود. برای این منظور فاصله این نقطه تا قله‌های ویژگی‌های مختلف در این c قاب باید کمینه شود. سپس تابع acc[n] که برابر تعداد دفعات برنده شدن قاب n است به دست می‌آید. قله‌های تابع acc[n] که دارای ارتفاع مناسبی هستند، به عنوان مرزهای واج‌ها تشخیص داده می‌شوند. شکل ۳۵ یک مثال از تابع acc[n] را نشان می‌دهد.



شکل ۳۴: مثالی از تابع $x_i[n]$ و تابع $J_i^5[n]$ مرتبط با آن.



شکل ۳۵: یک تابع acc[n] نوعی. هر قله با یک مرز در بخش‌بندی معادل است.

خلاصه الگوریتم چنین است:

۱- محاسبه تغییرات هر ویژگی در هر فریم. برای این منظور متوسط مقدار ویژگی در a فریم قبل و پس از فریم مورد نظر محاسبه می‌شود. این مقدار را J می‌نامیم.

۲- به دست آوردن قله‌های تابع J.

۳- ارتفاع هر قله برابر کم‌ترین فاصله از ارتفاع دو دره مجاور است.

۴- حذف تمام قله‌های با ارتفاع کمتر از b

۵- برای هر پنجره به طول c قاب برنده قابی است که کم‌ترین فاصله را از قله‌های موجود در این پنجره داشته باشد.

۶- تعداد برنده شدن هر قاب را ثبت کن.

۷- به سادگی می‌توان آستانه‌ای پیدا کرد که قاب‌هایی که زیاد برنده شده‌اند را به عنوان مرز معرفی کند.

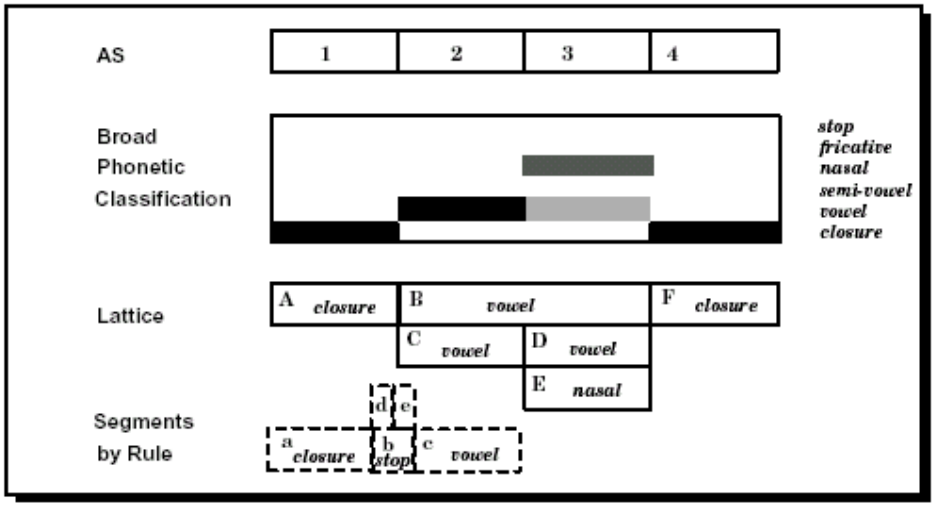
برای ارزیابی این سیستم از پایگاه داده TIMIT استفاده شده است. یک مرز زمانی درست است که حداکثر 20ms از یک مرز صحیح فاصله داشته باشد. بر اساس پارامترهای a، b و c می‌توان بخش‌بندی‌های متفاوتی داشت. اگر بخواهیم تمام مرزهای به دست آمده صحیح باشند می‌توان تا ۷۴٪ مرزها را پیدا کرد. با پیدا کردن ۶۳٪ مرز اضافی، می‌توان ۹۰٪ مرزها را پیدا کرد.

۶-۱-۲) استفاده از بخش‌بندی افزونه برای پیشنهاد به سیستم بازشناسی گفتار[46]

در پایان‌نامه دکترای آقای philipp schmid، نیز بخش‌بندی به عنوان یک مرحله اولیه وجود دارد. در این تز، ایشان سعی می‌کنند که از دانش خواندن طیف-نگار در تشخیص صحبت استفاده کنند. به عنوان بخش اولیه سیستم، ایشان شبکه‌ای[1] تولید می‌کنند که در آن تمام بخش‌بندی‌های ممکن مشخص شده است. خروجی بخش‌بندی، بخش‌های ممکن و نوع کلی هر بخش (بست، صدادار، شبه صدادار، خیشومی، سایشی و انفجاری) است. شکل ۳۶ نمونه‌ای از یک شبکه بخش‌بندی را نشان می‌دهد.

[1] lattice

همچنین برای جبران خطای حذف، از یک‌سری قانون برای درج واج‌های انفجاری b و d در بین یک بست و یک صدادار استفاده می‌شود.



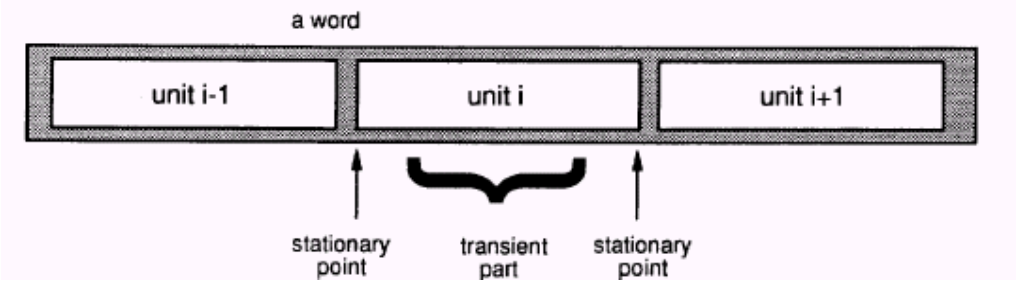
شکل ۳۶: نمونه‌ای از فرآیند بوجود آمدن شبکه بخش‌ها. در بالا بخش‌بندی بر اساس ویژگی‌های شنیداری انجام می‌شود. سپس شباهت هر بخش با گروه‌های آوایی تعیین می‌شود. این شباهت در شکل با تیرگیِ بیشتر مشخص شده است. سپس شبکه بخش‌های ممکن به دست می‌آید. در نهایت بر اساس یک سری قانون، در این شبکه تغییراتی بوجود می‌آید.

ما در اینجا از ذکر جزئیات تعیین بخش‌های ممکن صرف نظر می‌کنیم. آنچه در مورد این تز مهم است این است که هم آموزش و هم تست بر اساس بخش‌بندی خودکار توسط ماشین انجام می‌شود. علت امر این است که آزمایش نشان داد که آموزش بر اساس بخش‌بندی دستی و تست بر اساس بخش‌بندی ماشینی منجر به اختلاف بین شرایط آموزش و تست و در نتیجه کاهش نتیجه می‌شود.

۶-۱-۳) سیستم خبره SPREX II [61]

SPREX یک شرکت کره‌ای است که در زمینه پردازش صوت فعالیت دارد. یکی از محصولات این شرکت، سیستم تشخیص صحبت SPREX[1] است که یک سیستم خبره است. جزئیات روش بخش‌بندی صحبت در این سیستم توضیح داده نشده است و اهمیت چندانی نیز ندارد. هدف سیستم بخش‌بندی به دست آوردن واحدهایی است که دارای تغییرات شدید هستند. در حقیقت همانطور که شکل ۳۷ نشان می‌دهد، یک بخش عبارت است از گذر بین دو حالت ایستا. این ابتکار بسیار مهم است و توانایی این سیستم را در استخراج ویژگی‌های با اهمیت نشان می‌دهد. این ابتکار خود را در جای دیگری نیز نشان می‌دهد. این سیستم تلاشی در جهت یافتن واج‌هایی که سیستم توانایی استخراج آنها را ندارد نمی‌کند و به جای تشخیص واج‌ها، گروه‌های واجی را تشخیص می‌دهد. جزئیات بیشتر سیستم SPREX خارج از بحث بخش‌بندی است.

[1] جالب این است که این شرکت زبان فارسی را به عنوان اولین زبان غیر کره‌ای برای محصول بازشناسی گفتار انتخاب کرده است.



شکل ۳۷: تعریف یک بخش در سیستم SPREX

۶-۲) روش پیشنهادی برای یافتن بخش‌های در حد واج

در ابتدای پروژه (پیش از تصمیم‌گیری قطعی در مورد واحد استخراج ویژگی) تلاش‌هایی برای به دست آوردن مرز بین واج‌ها انجام شد. تصویر طیف-نگار و رابطه آن با تقطیع واجی (شکل ۴۷) انسان را تشویق می‌کند که مرز واج‌ها را از روی طیف-نگار تشخیص دهد.

در این روش نیز واحد بخش‌بندی واج است و تغییرات در ویژگی‌های بانک فیلتر را به عنوان دلیلی برای امکان گذر بین واجی در نظر می‌گیریم. در اینجا نیز هدف یافتن تمام مرزهای ممکن به همراه امکان متناظر آنها است. امکان یک مرز، بر اساس شدت تغییرات در این مرز مشخص می‌شود. ممکن است به علت فراوانی بیشتر برخی از واج‌ها مرزهای ضعیف‌تر دارای احتمال وقوع بیشتری باشند، ولی ما تنها امکان مرز بودن را در نظر می‌گیریم و نه احتمال آن را. اگر مرز را لحظه گذر از مقداری به مقداری دیگری در نظر بگیریم، آنگاه امکان مرز بودن یک قاب با داشتن B بردار ویژگی برابر است با:

{ امکان مرز بودن یکی از ویژگی‌ها}Snorm = امکان مرز بودن

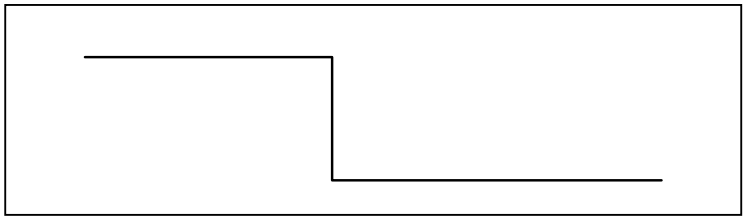
همچنین امکان مرز بودن هر ویژگی چنین حساب می‌شود(شکل ۳۸ را ببینید):

Tnorm{A, B, C}= امکان مرز بودن

A=امکان ثابت بودن مقدار قبلی

B=امکان شدید بودن تغییرات

C= امکان ثابت بودن مقدار جدید



شکل ۳۸: مفهوم مرز عبارت است از: مدتی بدون تغییر، یک تغییر شدید و دوباره مدتی بدون تغییر

حال مساله به به‌دست آوردن امکان A، B و C ساده می‌شود. امکان A بر اساس واریانس محاسبه می‌شود. اگر واریانس مقادیر قبلی صفر باشد، امکان A یک است و اگر واریانس مقادیر قبلی ∞ باشد، آنگاه امکان A صفر است. از آنجا که اعدادی که به عنوان امکان بیان می‌شوند تنها دارای ترتیب هستند و مقدارشان اهمیتی ندارد، می‌توان از هر شکل محدبی (مثلا یک تابع گوسی) برای مقدار دهی به امکان A بر اساس واریانس مقادیر قبلی استفاده کرد. اما نکته مهم این است که انتخاب مقادیر امکان‌ها باید به گونه‌ای باشد که امکان‌هایی که با عملیات Tnorm و Snorm و بر اساس امکان‌های سطوح پایین به دست می‌آیند نیز دارای ترتیب صحیح باشند. این به معنی این است که مثلا اگر می‌خواهیم امکان پدیده D را بر اساس امکان پدیده‌های A و B به دست آوریم، امکان D بودن در زمانی که A دارای امکان ۰.۳ است (بدون توجه به امکان B) با امکان D بودن در زمانی که B دارای امکان ۰.۳ است (بدون توجه به امکان A) برابر است. در عمل ما تنها با شکل تابع گوسی بازی کردیم تا نتایج قابل قبولی بگیریم. امکان B بر حسب نسبت تغییر دیده شده به شدیدترین تغییر ممکن محاسبه می‌شود. محاسبه امکان C نیز مانند A است. در ادامه پروژه ما به یک روش اندازه‌گیری جدید دست یافتیم که مشکل ترکیب امکان‌ها را حل می‌کند.

۶-۲-۱) نتایج

جدول زیر نتایج بخش بندی صحبت را بر روی دایرکتوری‌های FCJF0، FECD0، FMEM0، FTBR0، FVMH0، MEDR0، MPGH0، MRWS0 و MWAR0 از دایرکتوری DR1 از دادگان TIMIT نشان می‌دهد. لازم به ذکر است که دادگان TIMIT با نرخ نمونه‌برداری 16kHz ضبط شده است.

جدول ۳: نتایج بخش‌بندی صحبت توسط نرم‌افزار نوشته شده اول.

تعداد لبه‌هاي مشخص شده به کل لبه‌ها	درصد تشخیص با تعریف فاصله بیش از 30ms به عنوان خط	درصد تشخیص با تعریف فاصله بیش از 25ms به عنوان خط	درصد تشخیص با تعریف فاصله بیش از 20ms به عنوان خط	حداقل طول يك واج	حداقل فاصله دو لبه	حداقل درجه تعلق الزام‌آور[1]	بیشترین resolution	شماره آزمایش
0.15	0.99	0.98	0.84	10	256	3	3	1
0.036	1	1	0.9	10	256	6	3	2
0.14	0.99	0.98	0.84	10	768	3	3	3
0.3	0.97	0.92	0.73	10	256	3	5	4
0.14	0.99	0.98	0.85	10	256	4.5	5	5
0.11	0.995	0.99	0.88	10	256	6	5	6

[1] همانطور که گفته شد، یکی از راه‌های افزایش الزام یک پدیده، تعداد دلایلی است که در تایید آن می‌آید. به همین دلیل هرچند در تئوری امکان اپراتور صحیح اپراتور max است، ولی در اینجا از جمع معمولی (که تضمینی به خارج نشدن عقیده از حد ۱ نمی‌دهد) استفاده شده است.

8	7	3	768	10	0.72	0.92	0.97	0.3
9	7	4.5	768	10	0.79	0.94	0.98	0.23
10	7	6	768	10	0.85	0.98	0.99	0.14
11	7	6	256	10	0.85	0.98	0.99	0.14
12	3	3	768	20	0.84	0.93	0.96	0.02
13	5	3	768	20	0.8	0.96	0.98	0.05
14	7	4.5	768	20	0.8	0.96	0.98	0.03
15	3	3	768	8	0.83	0.98	0.995	0.19
16	5	3	768	8	0.74	0.93	0.97	0.36
17	7	4.5	768	8	0.78	0.94	0.97	0.3
18	7	4.5	256	8	0.76	0.93	0.97	0.32
19	5	3	768	5	0.64	0.88	0.95	0.43
20	5	5	768	5	0.79	0.95	0.98	0.26

۶-۲-۲) روش بهبوددادہ شده

بررسی نشان داد که اگر تنها تغییرات در هر باند فیلتر را در نظر بگیریم، نتایج بهتری به دست می‌آید. در این روش امکان مرز بین دو واج بودن برابر بیشترین شدت تغییراتی است که در باندهای فیلتر دیده شده است. این نتایج بر روی بخش تست لهجه تهرانی دادگان فارس‌دات که شامل ۱۰۰ فایل و ۳۴۵۸ واج است به دست آمده است.

جدول ۴: نتایج بخش‌بندی صحبت توسط نرم‌افزار نوشته شده دوم.

شماره آزمایش	تابع تعیین لبه‌ها	حداقل درجه تعلق الزام‌آور	حداقل فاصله دو لبه	حداقل طول یك واج	درصد تشخیص با تعریف فاصله بیش از 20ms به عنوان خط	درصد تشخیص با تعریف فاصله بیش از 25ms به عنوان خط	درصد تشخیص با تعریف فاصله بیش از 30ms به عنوان خط	تعداد لبه‌هاي مشخص شده به کل لبه‌ها
1	Necessary	2	512	5	95%	97%	98%	58%
2	Possible	0.5	512	5	80%	86%	89%	79%
3	Necessary	1	512	5	90%	93%	94%	73%
4	Possible	0.3	512	5	62%	68%	74%	84%
5	Necessary	4	512	5	98%	99%	99.5%	33%
6	Possible	0.7	512	5	92%	95%	97%	62%

۶-۳) روش پیشنهادی اول برای یافتن اشیاء (*OBSFE*[1])

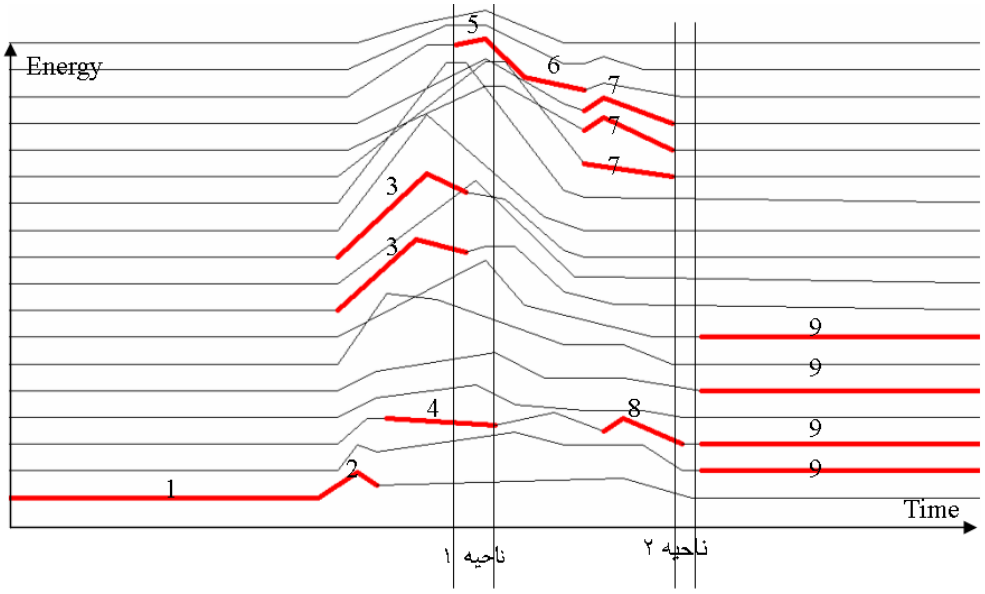
اگر به روش تشخیص صحبت در انسان توجه کنیم متوجه می‌شویم که واحد تشخیص در انسان

[1] Object Based Segmentation and Feature Extraction

کوچک‌تر از واج است. برای مثال افرادی که در زمینه موسیقی و آواز کار می‌کنند می‌توانند تعداد تحریرها را در یک آواز بشمرند. همچنین انسان می‌تواند از وقوع یک پدیده در صدا صحبت کند. برای مثال حرف «د» شامل یک بخش انفجاری است. این بسیار جالب است که با وجود اینکه تشخیص واج‌های کوتاهی مانند «د» و «ب» برای روش‌های متداول تشخیص صحبت دشوار است، اما انسان‌ها از کودکی قادرند این واج‌ها را تشخیص داده و ادا کنند. جالب است که در کلمات مادر و پدر در زبان‌های مختلف، چنین واج‌هایی به چشم می‌خورند. همچنین آزمایش نشان می‌دهد که اطلاعاتی که انسان از گذر بین واج‌ها و تغییرات در صحبت به دست می‌آورد بسیار مهم‌تر از اطلاعاتی است که از شنیدن یک صدای ممتد به دست می‌آید. برای مثال اگر واج «اُ» را به‌صورت ممتد پخش کنیم و کسی لحظه گذر به واج «اُ» را نشنود، تشخیص واجی که در حال ادا شدن است برای او بسیار مشکل می‌شود.

وقتی به سیگنال صحبت در فضای زمان نگاه می‌کنیم، از شدت تغییرات آن متعجب می‌شویم. برعکس اگر به طیف‌نگار نگاه کنید، متوجه می‌شوید که تغییرات انرژی در هر فرکانس نسبتا کند است. این مطلب نوید می‌دهد که ما توانسته‌ایم نمایشی از سیگنال صحبت را بیابیم که در آن قابلیت پیش‌بینی ما بسیار بهتر است. آزمایش‌های فصل ۳ نشان دادند که تغییرات جزئی در مقدار انرژی در باندهای فیلتر تاثیر بسیار کمی در صدای شنیده شده دارد. به این ترتیب ما در اولین مرحله، خط سیر انرژی در باندهای مختلف را با خط تقریب می‌زنیم.

در این روش هدف یافتن اشیاء است. یک شیئ کوچک‌ترین بخش از سیگنال صوتی است که انسان می‌تواند راجع به آن صحبت کند. در یک تعریف عینی‌تر، یک شیئ بخشی از سیگنال است که در آن انرژی در حداقل یکی از باندهای فیلتر، از یک مقدار می‌نیمم به مقدار می‌نیمم بعدی می‌رسد. شکل ۳۹ این مفهوم را نشان می‌دهد.



شکل ۳۹: تقریب خطی، برخی اشیاء و بخش‌بندی کلمه 'five' با ۱۰۴ قاب. خطوط کلفت اشیائی را نشان می‌دهند که متناظر با یک بخش‌هستند. ناحیه ۱ چند قاب را نشان می‌دهد که در بین چند بخش (بخش‌های ۴ و ۵) مشترک هستند. ناحیه ۲ چندین قاب را نشان می‌دهد که توسط هیچ بخشی پوشانده نشده‌اند.

به نظر می‌رسد که این روش تشخیص اشیاء برای واج‌های «ر» و «و» مناسب نیست. زیرا همانطور که در شکل ۴۳ دیده می‌شود، این واج‌ها مترادف با یک افت ناگهانی در سیگنال انرژی هستند. همین مطلب نشان می‌دهد که بررسی مفهوم شیء کار بیشتری می‌طلبد.

مراحل سیستم بخش‌بندی صحبت عبارتند از:

۱- محاسبه انرژی باندهای فیلتر در قاب‌ها.

۲- تقریب زدن خط سیر انرژی در هر باند فیلتر با خط.

۳- به دست آوردن اشیاء. با استفاده از تقریب خطی سیگنال خط سیر.

۴- بخش‌بندی سیگنال صحبت.

۵- استخراج ویژگی در هر بخش.

۶- [در مرحله آموزش] به دست آوردن صدک‌ها برای هر ویژگی.

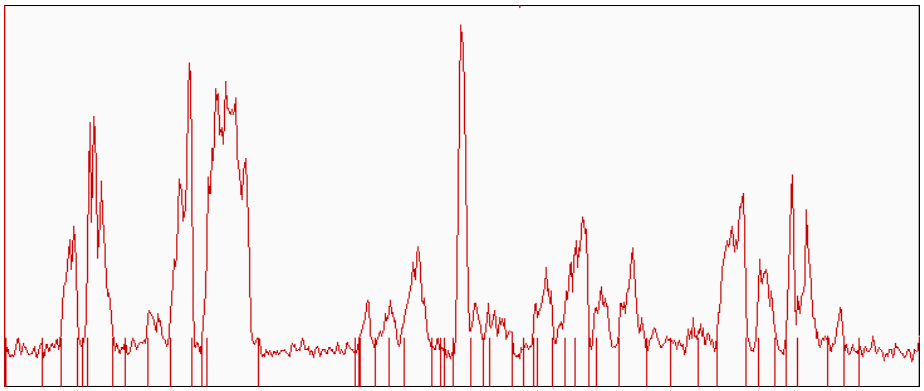
۷- بیان مقدار هر ویژگی با عددی صحیح بین ۰ تا ۱۰۰.

۶-۳-۱) محاسبه انرژی باندهای فیلتر در قاب‌ها

برای محاسبه انرژی، ابتدا سیگنال به قاب‌هایی با اندازه FRAMESIZE[۱] و فواصل SBF تقسیم می‌شود. سپس هر قاب از یک فیلتر Hanning عبور می‌کند و تبدیل فوریه آن محاسبه می‌شود. در مرحله بعد قدر

[۱] این پارامترها می‌توانند در فایل Definitions.ini تعریف شوند.

مطلق انرژی محاسبه می‌شود. پس از محاسبه انرژی در هر طیف، مقدار انرژی در ۲۴ باند فیلتر محاسبه می‌شود. تعریف این باندها توسط پارامتر BAND مشخص می‌شود. سپس اگر کاربر پارامتر FEATURETYPE را برابر CEPSTRUM قرار داده‌باشد، ویژگی‌های فیط محاسبه می‌شوند. در غیر این صورت ویژگی‌ها به همان شکل انرژی در باندهای فیلتر باقی می‌مانند. پارامتر NFRONTENDFEATURES تعیین می‌کند که بالاخره چند ویژگی استخراج شده است.



شکل ۴۰: انرژی در باند فیلتر [6800Hz-7770Hz-8860Hz] از فایل s21849.wav در دادگان فارس‌دات. همانطور که دیده می‌شود، خط سیر انرژی در فرکانس‌های بالا معمولا دارای مقدار ثابت DC است.

شکل ۴۰ خط سیر انرژی در فیلتر [6800Hz-7770Hz-8860Hz] را نشان می‌دهد. همانطور که دیده می‌شود، خط سیر انرژی در فرکانس‌های بالا دارای مقدار ثابت DC است که باعث بی‌معنی شدن مفهوم انرژی می‌شود. به همین دلیل در انتهای الگوریتم محاسبه انرژی، این مقدار DC حذف می‌شود. برای حذف مقدار DC، خط صفر جایی تعریف می‌شود که ۲٪ مقادیر انرژی زیر آن باشند.

۶-۳-۲) تقریب زدن خط سیر انرژی در هر باند فیلتر با خط[۱]

ابتدا مقدار انرژی با ضرب در عدد مناسب که توسط YCOEFFOR22050 مشخص می‌شود با مقدار زمان قابل مقایسه می‌شود. هر تقریب خطی شامل تعدادی نقطه در قاب‌های مختلف است که به ترتیب با خط به یکدیگر وصل می‌شوند. این نقاط را اتصال می‌نامیم. الگوریتم تقریب خطی به دنبال تقریبی می‌گردد که دارای شرایط زیر باشد:

۱- تعداد اتصال‌ها از MAXPOINTPERSECOND نقطه در ثانیه بیشتر نشود.

۲- اتصال‌ها باید در نقاطی باشد که خط سیر زاویه می‌سازد. در هر اتصال، زاویه‌ای که خط سیر انرژی بین سه نقطه در DX قاب قبل، قاب فعلی و DX قاب بعد ایجاد می‌کند باید بیش از MINBENDDEGREE باشد.

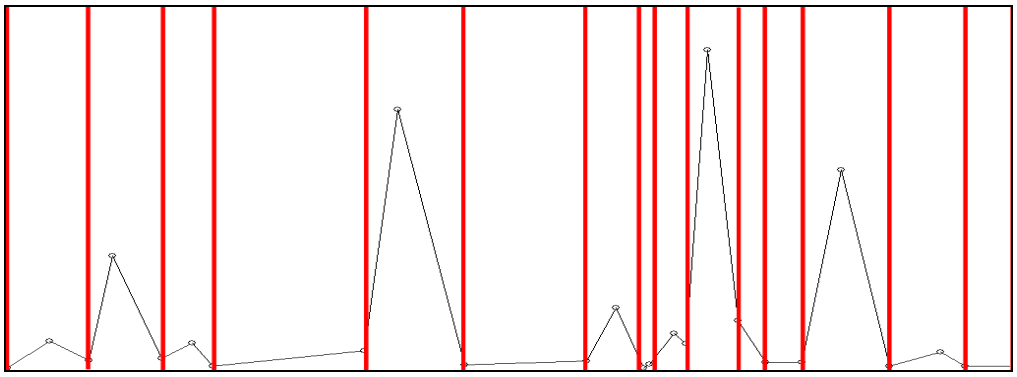
۳- فاصله نقاط اولیه از خط سیر تقریب زده شده باید کم‌تر از

[۱] یکی از زمینه‌هایی که باید بیشتر بررسی شود، تاثیر مدل‌های پیچیده‌تر تقریب زدن مانند مدل درجه ۲ و یا استفاده از SPline ها است.

ACCEPTABLEERROR باشد. در صورتی که نتوان خط سیر را طوری تقریب زد که متوسط تعداد اتصال‌ها در ثانیه از MAXPOINTPERSECOND اتصال کمتر باشد و در عین حال فاصله نقاط اولیه از خط سیر تقریب زده شده کم‌تر از ACCEPTABLEERROR باشد، برنامه یک واحد به مقدار ACCEPTABLEERROR اضافه می‌کند و الگوریتم را دوباره اجرا می‌کند. این کار در حذف نویز در محیط‌هایی که نویز اشیاء اضافی تولید می‌کند بسیار مفید است.

۶-۳-۳) به دست آوردن اشیاء

با استفاده از تقریب خطی سیگنال خط سیر، می‌توان بزرگ‌ترین بخش‌های محدب را پیدا کرد. ایده اصلی این است که بیان یک مطلب جدید حتما باید با پیش‌بینی ما از خط سیر متفاوت باشد. تغییر تحدب نیز با یک زاویه بیش از ۱۸۰ درجه در یک اتصال مشخص می‌شود. شکل ۴۱ اشیائی که در یکی از باندهای فیلتر تشخیص داده شده‌اند را نشان می‌دهد. هر شیئ می‌تواند از ۲، ۳ یا ۴ نقطه اتصال تشکیل شده باشد.

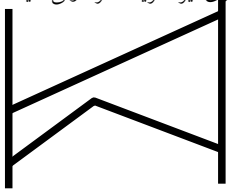


شکل ۴۱: اشیاء پیدا شده در باند فیلتر [1000-900-800] در فایل s11881.wav در دادگان فارس‌دات.

۶-۳-٤) بخش‌بندی سیگنال صحبت

اشیائی که در باندهای مختلف فیلتر به دست آمده‌اند کاندیدهای ایجاد بخش هستند. اینکه ما چگونه کاندیدهای مناسب را انتخاب کنیم کار بیشتری می‌طلبد و به توانایی‌های سیستم تشخیص‌دهنده نیز وابسته است. می‌دانیم که در پیاده‌سازی باید بین اشیاء یک ترتیب کلی وجود داشته باشد[1]. هر بخش کاندید بازه‌ای از قاب‌ها را مشخص می‌کند. بازه‌ای با شروع قاب a و پایان قاب b را با [a,b] نشان می‌دهیم. می‌گوییم که [a,b]≤[c,d] اگر و فقط اگر a≤c و b≤d. به این ترتیب مشخص است که برخی از بازه‌ها قابل مقایسه نیستند. در این صورت بین هر دوبازه‌ای که قابل مقایسه نیستند، بازه کوچکتر (که

[1] البته می‌توان رابطه ترتیب جزئی و حتی ترتیب جزئی فازی نیز تعریف کرد ولی سیستم بازشناسی گفتار دارای پیچیدگی‌هایی می‌شود.



حتما درون بازه بزگ‌تر است) را انتخاب می‌کنیم و بازه بزرگ‌تر را از لیست بازه‌های کاندید حذف می‌کنیم. در پیاده‌سازی جزئیات بیشتری وجود دارد تا بخش‌بندی بهبود یابد. خط سیر انرژی در هر باند فیلتر در هر بخش می‌تواند از ۲ یا سه نقطه اتصال تشکیل شده باشد. به بیان دیگر، خط سیر انرژی در یک بخش یا یک خط است یا یک مثلث (شکل ۴۲).

شکل ۴۲: خط سیر انرژی در هر بخش از یک یا دو پاره‌خط تشکیل می‌شود. خط سیر دوپاره‌خطی شبیه مثلث است.

۶-۳-۵) استخراج ویژگی در هر بخش.

برای استخراج ویژگی در هر بخش، خط سیر سیگنال در باندهای فیلتر در آن بخش را در نظر می‌گیریم. خط مبنا را خطی که اولین نقطه و آخرین نقطه خط سیر را به هم وصل می‌کند، تعریف می‌کنیم. ویژگی‌ها عبارتند از[۱]:

۱- طول بخش (بر حسب تعداد قاب[۲]).

۲- بیشینه انرژی در هر باند فیلتر.

۳- میزان تقعر/ تحدب خط سیر انرژی در هر باند که برابر با فاصله علامت‌دار دورترین نقطه از خط مبنا تعریف می‌شود.

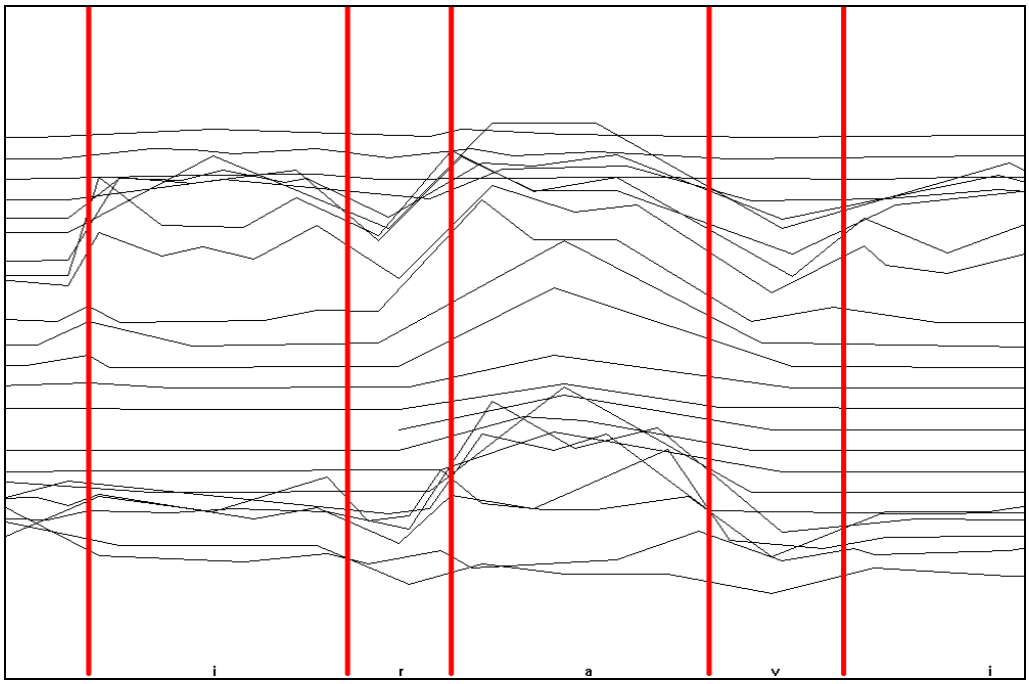
۴- زاویه خط مبنا با محور زمان.

۵- مرکز ثقل انرژی در هر باند فیلتر. علاوه بر دو نقطه ابتدایی و انتهایی، خط سیر در هر بخش حداکثر دارای یک نقطه در وسط است.

---

[۱] ما برای سرعت بیشتر و زمان کم ویژگی‌ها را از روی تقریب خطی سیگنال استخراج می‌کنیم. می‌توان برای دستیابی به جزئیات بیشتر، ویژگی‌ها را مستقیما از دنباله قاب‌ها استخراج کرد.

[۲] یکی از کارهای تکمیلی این پروژه استخراج ویژگی‌ها در واحد SI است تا از جزئیات پیاده‌سازی و نرخ نمونه‌برداری مستقل شود. البته سعی شده است که تا حدودی این هدف تحقق یابد.



شکل ۴۳: خط سیر انرژی در باندهای مختلف فیلتر در کلمه «می‌روی». همانطور که دیده می‌شود، واج‌های «ر» و «و» به صورت دو دره در خطی سیر انرژی ظاهر می‌شوند.

۶-۳-۶) *[در مرحله آموزش]* به دست آوردن صدک‌ها برای هر ویژگی.

در این مرحله با آمارگیری بر روی مقادیری که هر ویژگی می‌تواند به خود بگیرد، صدک‌ها را برای هر ویژگی به دست می‌آوریم. یادآوری می‌کنم که صدک i ام برای یک تابع توزیع از متغیر X برابر برابر x ای است که i% مساحت تابع توزیع در بازه $[-\infty,x]$ است. برای این منظور ابتدا مقدار هر ویژگی به عددی صحیح بین ۰ تا ۱۰۰۰۰ نرمال می‌شود و سپس صدک‌ها در این فضای گسسته به دست می‌آیند.

۶-۳-۷) بیان مقدار هر ویژگی با عددی صحیح بین ۰ تا ۱۰۰.

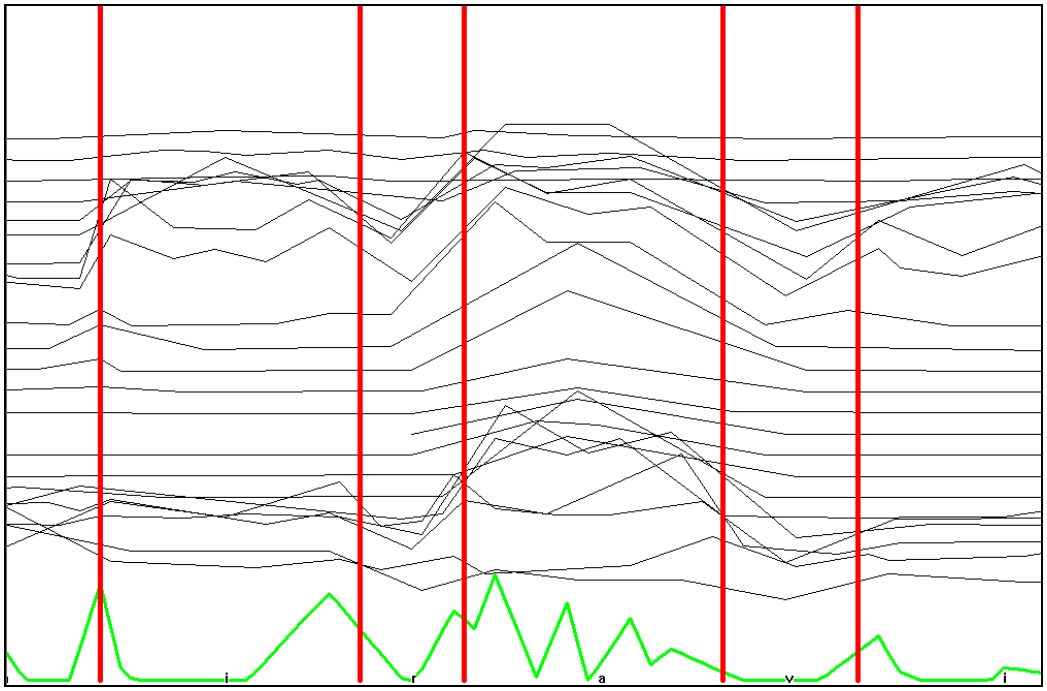
در این مرحله هر ویژگی را به اندیس نزدیک‌ترین صدک آن ویژگی منتسب می‌کنیم.

۶-٤) روش پیشنهادی دوم برای یافتن اشیاء (*OBSFE2*)

بخش بندی روش قبل دارای این خاصیت است که بر اساس رأی‌گیری نیست. روش دوم مبتنی بر ترکیب نظر باندهای مختلف برای تشخیص بخش‌ها است. در این روش نیازی به تشخیص اشیاء نیست و به‌جای آن از روی تقریب خطی خط سیر باندهای مختلف، پارامتری به نام اهمیت برای هر قاب حساب می‌شود[۱]. پارامتر IMPORTANCEDISTRIBUTION تعیین می‌کند که ناحیه‌ای که اهمیت قاب را تعیین می‌کند چند میلی‌ثانیه است.

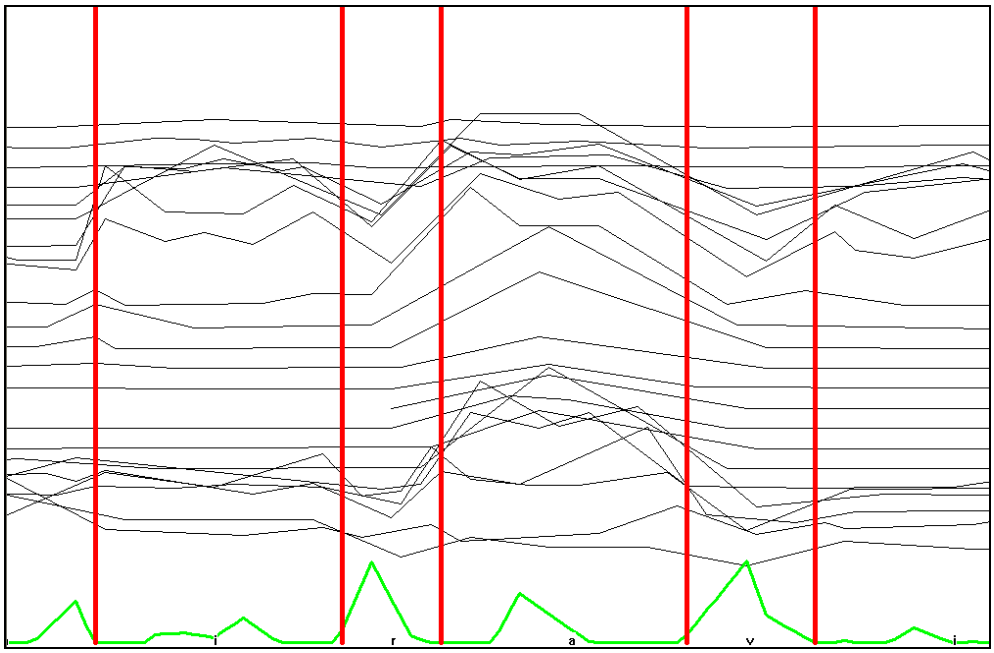
[۱] البته در پیاده‌سازی از اشیائی که طبق روش اول به دست آمده‌اند استفاده می‌شود ولی این استفاده صرفا بخاطر تغییرات کمتر در کد بوده‌است.

اهمیت هر نقطه اتصال بر حسب انرژی آن نقطه و همچنین جهت اتصال تعیین می‌شود. اگر جهت نقطه اتصال رو به بالا باشد اهمیت مثبت و در غیر این صورت اهمیت منفی است. از روی اهمیت نقاط اتصال مجاور یک قاب می‌توان اهمیت آن قاب را به دست آورد. از آنجا که ترکیب کردن اتصال‌های با اهمیت مثبت و منفی پاسخ خوبی نداد، برای هر قاب دو مقدار اهمیت مثبت و منفی حساب می‌شود. اهمیت مثبت هر قاب برابر با مجموع وزن‌دار اهمیت نقاط اتصال مثبت مجاور آن است. وزن هر نقطه اتصال بر حسب فاصله آن از قاب تعیین می‌شود. اهمیت منفی هر قاب نیز برابر مجموع وزن‌دار اهمیت نقاط اتصال منفی مجاور آن است. شکل ۴۴ اهمیت مثبت و شکل ۴۵ اهمیت منفی را برای فایل s11181.wav از پایگاه داده فارس‌دات نشان می‌دهد. اهمیت منفی در تشخیص واج‌های «ر» و «و» دارای اهمیت زیادی است. چون سیستم بازشناسی ما هنوز مدلی زمانی از واج‌ها ندارد[1]، بخشی که به عنوان نماینده یک واج انتخاب می‌شود باید کاملا به آن واج شبیه باشد. انتساب بخش‌های با اهمیت منفی به یک واج با مفهومی که مد نظر نگارنده است سازگار نیست (هرچند نتایج آن حدود ۲٪ بیشتر است). اما زمانی که بتوانیم روشی برای مدل‌سازی دنباله واجی ارائه دهیم، می‌توانیم از این بخش‌ها نیز استفاده کنیم.



شکل ۴۴: بخشی از فایل s11881.wav از دادگان فارس‌دات و اهمیت مثبت هر قاب که در پایین شکل نشان داده شده است.

[1] منظور این است که حالات واج‌ها در مدلی چپ به راست که ترتیب زمانی را نشان می‌دهد ذخیره نشده‌اند.



شکل ۴۵: بخشی از فایل s11881.wav از دادگان فارس‌دات و اهمیت منفی هر قاب که در پایین شکل نشان داده شده است.

فصل ۷
سیستم تشخیص صحبت پیاده‌سازی شده

سیستم پیاده‌سازی شده از دیدگاه نظریه امکان
سیستم پیاده‌سازی شده از دیدگاه شباهت با انسان
بخش‌بندی و استخراج ویژگی
آموزش سیستم
- **نام‌دهی ۴ نامی به اشیاء**
- **نام‌دهی تک‌نامی به اشیاء**
- **محاسبه توزیع امکان یک گروه** ARU
- **محاسبه شباهت یک شیء به یک توزیع امکان**

استفاده از توزیع امکان منفی برای اطمینان از تصمیم‌گیری اولیه
تشخیص نویز
- **مخلوط**
- **استفاده از مدل اولیه بر اساس** VQ

بازشناسی
بخشهای پیاده‌سازی نشده
- **روش حذف نویز پیچشی**
- **روش حذف اثر دامنه سیگنال**

امتیاز دهی

۷-۱) سیستم پیاده‌سازی شده از دیدگاه نظریه امکان

اگر به مسائل پیرامون خود نگاه کنیم، می‌بینیم که بسیاری از آنها مسائلی هستند که به صورت حاصل اعمال یک SNorm بر روی یک TNorm بیان می‌شوند. یکی از این مسائل مساله بازشناسی در HMM است. فرض کنیم $M_1, M_2,\ldots, M_n$ تعدادی HMM هستند. همچنین فرض کنید که دنباله $O = o_1, o_2, ...o_T$ توسط یکی از HMM ها تولید شده است. می‌دانیم که مساله بازشناسی در HMM یافتن HMM ای است که احتمال اینکه دنباله O توسط آن تولید شده باشد بیش‌تر است[60]:

$$P(O \mid M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

در فرمول فوق X دنباله‌ای از حالات HMM را مشخص می‌کند. $a_{x(t)x(t+1)}$ احتمال گذر از حالت x(t) به حالت x(t+1) را نشان می‌دهد و $b_{x(t)}$ احتمال تولید شدن رخداد o_t توسط x(t) را نشان می‌دهد (که به نوعی می‌توان آن را شباهت o_t به x(t) در نظر گرفت). توجه نمایید که فرمول فوق به شکل SNorm-TNorm است که SNorm آن جمع و TNorm آن ضرب می‌باشد. در عمل برای سادگی تعمیم یافتن به بازشناسی گفتار پیوسته، به‌جای عملگر جمع از عملگر max استفاده می‌شود:

$$\hat{P}(O \mid M) = \max_{X} \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right\}$$

اگر به تعریف اندازه‌گیری امکانی مراجعه کنید، مشاهده می‌کنید که max-product در تعریف اندازه‌گیری امکانی صدق می‌کند. حال برای تشخیص گفتار پیوسته کافی است دنباله‌ای از حالات HMMهای مختلف مانند X را طوری در تناظر با دنباله مشاهدات O قرار دهیم که اولا گذر از هر حالت به حالت بعدی مجاز باشد و ثانیا حاصل‌ضرب $a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(o_t) a_{x_t x_{t+1}}$ بیشینه شود. در عمل معمولا احتمال گذر بین حالات را صفر و یک فرض می‌کنند و مساله به یافتن دنباله‌ای مجاز از حالات با بیشترین مقدار $\prod_{t=1}^{T} b_{x_t}(o_t)$ ساده می‌شود.

در این پایان‌نامه ما تلاش کردیم که یک قدم به نظریه امکان نزدیک‌تر شویم و به‌جای max-product از max-min استفاده کنیم. بنابراین مساله ما چنین تعریف می‌شود:

$$\Pi(O \mid M) = \max_{X}\left\{\min_{t=1}^{T} b_{x(t)}(o_t)\right\}$$

اما ما به دنبال واج (کلمه‌ای) هستیم که امکان اینکه مشاهده ما متعلق به آن باشد، بیشینه باشد. داریم:

$$\arg\max_{M}\{\Pi(M \mid O)\} = \arg\max_{M}\{\min\{\Pi(O \mid M), \Pi(M)\}\}$$
$$\forall M \in \textit{phonemes}: \ \Pi(M) = 1$$
$$\arg\max_{M}\{\Pi(M \mid O)\} = \arg\max_{M}\{\Pi(O \mid M)\}$$

دیده می‌شود که در مساله بازشناسی گفتار، بیشینه شدن $\Pi(M \mid O)$ معادل بیشینه شدن $\Pi(O \mid M)$ است. حال ببینیم مفهوم استفاده از عملگر min چیست؟ یکی از مهم‌ترین خواص عملگر min خاصیت غیرقابل جبران بودن آن است. از همین ابتدا معلوم است که عملگر min ما را دچار مشکل می‌کند. اعتقاد ما به یک دنباله از حالات برابر با کمترین مقدار $b_{x_t}(o_t)$ است. بدین ترتیب ما به دنبال دنباله‌ای از حالات[1] می‌گردیم که کمترین مقدار $b_{x_t}(o_t)$ در آن از بقیه دنباله‌های حالت بیشتر باشد. عیب این روش این است که اگر در مشاهدات ما نویز وجود داشته باشد مساله به یافتن نزدیک‌ترین حالت به مشاهده نویزی تبدیل می‌شود. بدین ترتیب دیده می‌شود که جایگزینی عملگر ضرب با عملگر می‌نیمم‌گیری نشدنی است. به همین دلیل ما فرض می‌کنیم که مساله ما یافتن یک دنباله شدنی از حالات با امکان مشخص □ (مثلا ۰.۷) است. یعنی می‌خواهیم دنباله حالات X را طوری بیابیم که شباهت هر مشاهده o_t به حالت $x(t)$ از □ کم‌تر نباشد. مشخص است که مساله تنها یک پاسخ ندارد. برای مثال ممکن است تمام حالات به o_t شبیه باشند. این با برداشت ما از نظریه امکان سازگار است. در حقیقت صحبت روزمره افراد را نمی‌توان بدون دانستن اطلاعات زبانی درک کرد. اگر مشاهده o_t به هیچ حالتی شبیه نبود آن را نویز فرض می‌کنیم. این شاهکار نظریه امکان است. این نظریه می‌تواند متوجه اشیاء ناشناخته شود! همین علم به جهل خود به این نظریه امکان می‌دهد که دامنه اشیائی را که می‌شناسد گسترش دهد.

۷-۲) سیستم پیاده‌سازی شده از دیدگاه شباهت با انسان

یکی از مهم‌ترین مباحثی که در رابطه با تشخیص صحبت در انسان وجود دارد، قطعیت تشخیص آدمی است. سیستم‌های تشخیص صحبت معمولا قادرند با دقت‌های بالا به تفکیک تعدادی واج یا کلمه بپردازند. اما این سیستم‌ها در مقابل ورودی‌های بی‌ربط (مثلا صدای شیر آب) بسیار شکننده هستند.

[1] در عمل سیستم ما بر مبنای HMM پیاده‌سازی نشده است و حالتی هم وجود ندارد. ولی برای سازگار بودن بحث، از همان مفاهیم HMM استفاده می‌کنیم.

اساس سیستم ما مبتنی بر تشخیص صحبت قطعی است[1]. از طرف دیگر تشخیص موارد ناشناخته به سیستم اجازه می‌دهد که در یک محیط پویا به یادگیری اشیاء جدید بپردازد. همچنین چون خروجی سیستم شامل تمام دنباله‌های واجی ممکن است، این سیستم نیاز کمی به تعامل با سیستم تشخیص کلمه دارد. از دیدی دیگر، سیستم بخش‌بندی ویژگی‌ها را در یک پنجره زمانی-فرکانسی (و نه صرفا فرکانسی) استخراج می‌کند که با روش استخراج ویژگی در انسان مشابهت دارد[3]. همچنین این خاصیت سیستم ما که اشیاء را خودش پیدا می‌کند برای تولید سیستمی که صحبت را از کودکی یاد بگیرد مناسب‌تر است.

۷-۳) بخش‌بندی و استخراج ویژگی

در این سیستم از روش OBSFE2 استفاده شده است. جزئیات این زیرسیستم در ۶-۴ بیان شده است. می‌توان روش بخش بندی و استخراج ویژگی را در مراحل زیر خلاصه کرد:

۱- استخراج ویژگی‌های بانک فیلتر
۲- تقریب زدن خط سیر ویژگی‌های بانک فیلتر با خط
۳- محاسبه تابعی به نام تابع اهمیت که برای بخش‌بندی صحبت به‌کار می‌رود
۴- بخش‌بندی صحبت بر اساس تابع اهمیت و به دست آوردن اشیاء
۵- استخراج ویژگی‌ها (۹۷ ویژگی)
۶- به دست آوردن صدک‌ها (این کار تنها در داده آموزشی انجام می‌شود و داده تست از همان صدک‌های داده آموزشی استفاده می‌کند)
۷- کوانته کردن مقدار ویژگی‌ها به عددی بین ۰ تا ۱۰۰ برحسب صدک‌ها.

۷-٤) آموزش سیستم

در سیستم آموزش موارد زیر مورد یادگیری واقع می‌شوند:

۱- واج‌ها یا دوواجی‌هایی که می‌خواهیم مدل کنیم که اصطلاحا به آنها ARU[2] می‌گوییم. البته همه واج‌ها مدل می‌شوند ولی تنها برخی از دوواجی‌ها مدل می‌شوند
۲- تعداد مدل‌هایی که می‌خواهیم برای هر ARU بسازیم که آن‌ها را مخلوط می‌نامیم (مفهومی مشابه مخلوط[3] در HMM ولی با تفاوت‌های اساسی که در بخش‌های بعدی ذکر می‌شود).
۳- نمونه‌های هر یک از مخلوط‌ها.

---

[1] البته ممکن است ویژگی‌هایی که امروزه برای تشخیص صحبت استخراج می‌شوند برای تشخیص صحبت قطعی مناسب نباشند.
[2] Articulatory and auditory unit
[3] Mixture

۴- برای هر مخلوط مانند A، توزیع امکان A و نیز توزیع امکان A~.

۵- توزیع امکان اصلاح شده که در آن مقدار عددی امکان نیز مهم است.

۶- اشیائی که نویز هستند.

فرآیند آموزش چنین است:

۱- شماره مخلوط را صفر قرار بده.

۲- ابتدا اشیاء پیدا شده در پایگاه داده آموزشی به همراه ۴ نام ممکن برای آنها در فایل examples.4 ذخیره می‌شوند. هر نام یک واج یا یک دوواجی است. هر نام به یک عدد صحیح نگاشته می‌شود که به آن نام عددی می‌گوییم. فاصله دو نام عددی همواره از ۹ بیشتر است. در ادامه از این ویژگی برای تولید مخلوط‌ها استفاده می‌کنیم. هر مدخل فایل شامل یک بردار ویژگی به همراه ۴ نام مختلف و درجه اعتقاد به هر نام است. درجه اعتقاد به نام‌ها نرمال شده است، بدین معنی که حتما یکی از نام‌ها دارای امکان یک[1] است.

۳- از روی فایل examples.4 فایل examples.1 ساخته می‌شود که هر مدخل آن یک بردار ویژگی به همراه نام آن است. نام هر شیء برابر نامی است که در فایل examples.4 دارای درجه تعلق یک بوده است.

۴- تمام واحدهای ARU که دارای حداقل مشخصی نمونه در فایل examples.1 هستند، برای مدل‌سازی انتخاب می‌شوند و مدل مثبت آنها (یعنی توزیع امکان هر واحد ARU) ساخته می‌شود. نام واحدهای ARU و توزیع امکان آنها در فایل poss.txt ذخیره می‌شود.

۵- با استفاده از مدل مثبت هر واحد ARU می‌توان نمونه‌هایی که ممکن است به یک واحد ARU، مثلا A، تعلق داشته باشند را پیدا کرد. برخی از این نمونه‌ها واقعا نمونه‌ای از A هستند و برخی به اشتباه A تشخیص داده شده‌اند. از روی نمونه‌هایی که به اشتباه A تشخیص داده شده‌اند مدل منفی A (A~) ساخته می‌شود. از روی توزیع شباهت می‌توان شباهت هر نمونه داده شده را به این توزیع حساب کرد. تابع امکان اولیه را Π می‌نامیم و تابع امکان اصلاح‌شده را با □ نشان می‌دهیم. برای سادگی فرض می‌کنیم که $\Psi(x) = 0.5 + \frac{\Pi(x) - \mu}{\sigma} 0.3$. μ و σ میانگین و واریانس Π هستند و می‌خواهیم به تابع امکانی برسیم که میانگین آن ۰.۵ و واریانس آن ۰.۰۹ است. در فایل recognition.info برای هر واحد ARU نام و توزیع امکان مثبت و

[1] در پیاده‌سازی همه‌چیز اعدادی بین ۰ تا ۱۰۰ هستند.

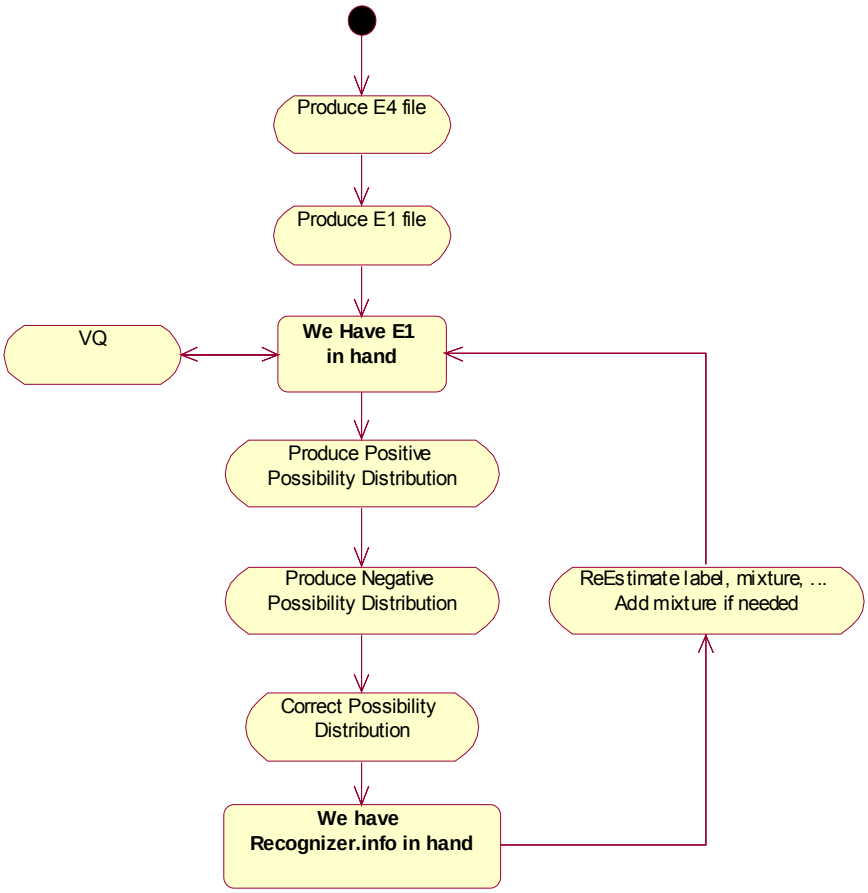
منفی آن به همراه پارامترهای μ و σ برای هر مدل ذخیره می‌شود.

۶- با استفاده از فایل recognition.info مرحله قبل و فایل examples.4 مجددا یک فایل examples.1 به دست می‌آید. برای این منظور از بین ۴ نام ممکن برای هر شیئ، نامی که سیستم مرحله قبل (recognition.info) بیشترین اعتقاد را به آن دارد انتخاب می‌شود. اگر شیء به هیچ یک از ۴ امکان شبیه نباشد، نام عددی شبیه‌ترین ARU بر اساس بخش‌بندی دستی بعلاوه شماره مخلوط بعنوان نام عددی این شیء انتخاب می‌شود. بدین ترتیب به مرور زمان برای نمونه‌هایی که توسط ARU های مدل شده تشخیص داده نمی‌شوند مخلوط‌های جدید اضافه می‌شود.

۷- اگر تعداد تکرارهای داخلی کافی نیست الگوریتم به مرحله ۳ می‌رود.

۸- شماره مخلوط را افزایش بده و اگر تعداد تکرارهای خارجی کافی نیست به مرحله ۳ برو.

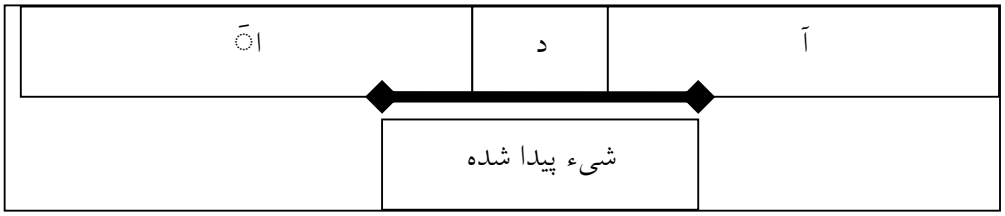
شکل ۴۶ فرآیند آموزش و رابطه بین اجزای آن را نشان می‌دهد.



شکل ۴۶: فرآیند آموزش

۷-٤-۱) نام‌دهی ٤ نامی به اشیاء

در روش‌های متداول چون شیء معادل با یک واحد آوایی مانند واج، دوواجی و یا سه‌واجی است نام آن مشخص است. اما در سیستم ما اشیاء بدون توجه به بخش‌بندی دستی پیدا می‌شوند. از طرف دیگر برای بررسی میزان تشخیص سیستم نیازمند نام‌دهی به این اشیاء هستیم. همچنین می‌دانیم که حتی اگر از بخش‌بندی دستی نیز استفاده کنیم سیستم خطای زیادی خواهد داشت که باید با تخمین دوباره[1] مرز بین واج‌ها خطا را کاهش داد. به همین دلیل برای هر شیء تمام نام های ممکن را می‌نویسیم و درجه اعتقادمان به آن را نیز مشخص می‌کنیم. شکل ۴۷ یک نمونه از رابطه بین شیء و واحد آوایی را نشان می‌دهد.



شکل ۴۷: نمونه‌ای از رابطه بین اشیاء پیدا شده و بخش‌بندی دستی.

برای هر دو بازه زمانی، میزان اعتقاد ما به یکی بودن آنها از فرمول زیر حساب می‌شود:

$$\text{امکان یکی بودن بازه A و B} = \frac{|A \cap B|}{\max(|A|,|B|)}$$

۷-٤-۲) نام‌دهی تک‌نامی به اشیاء

در این حالت بسته به عوامل زیر به هر یک از ۴ نام ممکن امکانی نسبت داده می‌شود. نام هر شیء نامی است که بیشترین امکان را داشته باشد. عواملی که بر اساس آنها امکان هر نام تعیین می‌شود عبارتند از:

۱- اعتقاد به نام در بخش‌بندی دستی
۲- اعتقاد به نام بر اساس سیستم بازشناسی مرحله قبل.
۳- آیا نام مخلوط هم باید درست تشخیص داده شود و یا تشخیص نام گروه ARU کافی است؟
۴- آیا دوواجی‌ها هم مدل می‌شوند؟
۵- آیا به واج‌های کوچک (مانند ب و د) بیشتر توجه شود؟

[1] Re-estimation

۷-٤-۳) محاسبه توزیع امکان یک گروه *ARU*

فرض کنیم n بردار ویژگی f بعدی برای گروه A داریم. می‌دانیم که مقادیری که این f بعد می‌توانند بگیرند بین ۰ تا ۱۰۰ هستند. توزیع امکان A را همانند توزیع احتمال با آمارگیری از نمونه‌ها به دست می‌آوریم. تنها تفاوت این توزیع با توزیع احتمال در این است که به‌جای اینکه مساحت زیر منحنی توزیع یک شود، توزیع نرمال می‌شود (امکان حداقل یکی از نقاط برابر یک می‌شود). با آمارگیری بر روی n نمونه می‌توان تعداد نمونه‌هایی که ویژگی i ام (i بین ۰ تا f) آنها مقدار j (j بین ۰ تا ۱۰۰) را دارد پیدا کرد. سپس برای هر ویژگی تابع امکانی به دست می‌آید که امکان اینکه این ویژگی مقدار خاصی را داشته باشد نشان می‌دهد. این تابع امکان نرمال است. همچنین همانطور که قبلا ذکر شد، این تابع امکان به یک تابع امکان اصلاح شده تبدیل می‌شود که در آن مقدار امکان نیز معنی‌دار است. برای توزیع امکان A، امکان اینکه ویژگی i ام مقدار j را بگرد با $A_{i,j}$ نشان می‌دهیم.

۷-٤-٤) محاسبه شباهت یک شیء به یک توزیع امکان

حال فرض کنیم توزیع امکان گروه‌های A و B داده شده است. می‌خواهیم شباهت نمونه x داده شده را به این گروه‌ها به دست آوریم. سپس با مقایسه میزان شباهت به هر گروه، می‌توان نمونه را به گروه شبیه‌تر منتسب کرد. در ادامه بحث فرض می‌کنیم که می‌خواهیم شباهت x به A را به دست آوریم. مقدار ویژگی i ام را در بردار ویژگی x با x_i نشان می‌دهیم.

ابتدا به هر ویژگی در هر گروه وزنی برابر معکوس واریانس توزیع امکان آن ویژگی نسبت می‌دهیم. سپس برای هر ویژگی i، $A_{i,x(i)}$ را به دست می‌آوریم و آن را با p_i نشان می‌دهیم. سپس لگاریتم p_i را محاسبه می‌کنیم. اعمال لگاریتم دو تعبیر می‌تواند داشته باشد:

۱- p_i مقدار اعتقاد را نشان می‌دهد و ما چون از عملگر ضرب برای ترکیب اعتقادها استفاده می‌کنیم، به‌جای ضرب اعتقادات لگاریتم آنها را جمع می‌کنیم. البته ما در عمل به‌جای تابع لگاریتم از تابع $\log(1+Jx)$ استفاده می‌کنیم.

۲- مقادیر p_i دارای یک توزیع احتمال نرمال هستند. بدین معنی که احتمال اینکه مقدار p_i کم باشد بیشتر است. از آنجا که مقدار امکان نیز تنها ۱۰۰ مقدار می‌تواند داشته باشد، اعمال تابع لگاریتم موجب بهتر نسبت داده شدن ۱۰۰ مقدار ممکن به مقادیر p_i می‌شود.

سپس مجموع $\log(p_i)$ ها محاسبه می‌شود. به عبارت دیگر در محاسبه امکان از عملگر Sum-Product استفاده کردیم. تا اینجا یک اندازه‌گیری امکان داریم که از روی آن با استفاده از مدل $\Psi(x) = 0.5 + \frac{\Pi(x) - \mu}{\sigma} 0.3$ تابع امکان اصلاح شده را به دست می‌آوریم.

۷-٤-۵) استفاده از توزیع امکان منفی برای اطمینان از تصمیم‌گیری اولیه

پس از ساخته شدن توزیع امکان مثبت A، برخی از نمونه‌ها به اشتباه از این فیلتر عبور می‌کنند و A تشخیص داده می‌شوند. به توزیع این نمونه‌ها مدل منفی A می‌گوییم. حالات زیر را در مورد $A_{i,j}$ و $A_{i,j}$~ در نظر بگیرید.

۱- امکان مدل مثبت برابر ۰ و امکان مدل منفی برابر ۰ است. در این صورت دیدن نمونه x با مقدار $x_i=j$ دلیلی بر هیچکدام از مدل‌های مثبت و منفی نیست. بعلاوه این نشان می‌دهد که ممکن است نمونه دیده شده نمونه‌ای باشد که قبلا در داده آموزشی دیده نشده است و به همین دلیل شبیه هیچ‌یک از مدل‌ها نیست.

۲- امکان مدل مثبت برابر ۰ و امکان مدل منفی برابر ۱ است. در این صورت دیدن نمونه x با مقدار $x_i=j$ دلیلی بر مدل منفی A است و یکی از وجوه تمایز A و A~ است.

۳- امکان مدل مثبت برابر ۱ و امکان مدل منفی برابر ۰ است. در این صورت دیدن نمونه x با مقدار $x_i=j$ دلیلی بر مدل مثبت A است و یکی از وجوه تمایز A و A~ است.

۴- امکان مدل مثبت برابر ۱ و امکان مدل منفی برابر ۱ است. در این صورت دیدن نمونه x با مقدار $x_i=j$ دلیلی بر هر دوی مدل‌های مثبت و منفی است. اما این ویژگی وجه تمایز A و A~ نیست و نمی‌تواند برای تفکیک نمونه‌های این دو گروه به‌کار رود. پس بهتر است اعتقادمان به A و A~ را خیلی زیاد نکنیم.

برای تصمیم‌گیری راجع به پذیرش نمونه x که از مدل مثبت عبور کرده است توزیع دیگری به نام امتیاز می‌سازیم و در آن امتیازی را که در هر گره ویژگی-مقدار قرار دارد می‌نویسیم. شکل ۴۸ برنامه محاسبه توزیع امتیاز مثبت و منفی را نشان می‌دهد. برای تعیین اینکه آیا x متعلق به A است یا خیر، مجموع امتیازات خانه‌هایی که زوج ویژگی-مقدار x را نشان می‌دهند را حساب می‌کنیم. این مجموع یک توزیع امکان است که دوباره با استفاده از مدل $\Psi(x)=0.5+\frac{\Pi(x)-\mu}{\sigma}0.3$ تابع امکان اصلاح شده را به دست می‌آوریم.

```
        poss0 = m_PositiveMF[i].GetLogPossibility (j);
        poss1 = m_NegativeMF[i].GetLogPossibility (j);
        if (poss0 < poss1)
                {
                        coef0 = 0;
                        coef1 = (poss1+T) / (poss0+T) - 1;
                }
        else
                {
                        coef0 = (poss0+T) / (poss1+T) - 1;
                        coef1 = 0;
                }
m_PositiveCoefs [i][j] = coef0;
m_NegativeCoefs [i][j] = coef1;
```

شکل ۴۸: الگوریتم محاسبه امتیازهای مثبت و منفی از روی توزیع‌های امکان مثبت و منفی

۷-٤-۶) تشخیص نویز

همانطور که در ۴-۴-۵ ذکر شد، یکی از مهم‌ترین مزایای نظریه امکان این است که به ما امکان مدل‌سازی جهل را می‌دهد. امکان A و A~ را در نظر بگیرید که امکان اینکه x متعلق به گروه X باشد/نباشد را نشان می‌دهند.

۱- $\Pi(A)=0$ و $\Pi(\sim A)=0$: در این صورت x نه به X شبیه است و نه به گروه‌های دیگری که می‌شناسیم[1]. بدین ترتیب بهتر است بگوییم که x یک داده جدید است که آن را نویز می‌نامیم. بعدا یاد می‌گیرم که این داده‌های جدید را نیز مدل کنیم. بدین ترتیب دیده می‌شود که نظریه امکان قابلیت کشف و یادگیری نمونه‌های جدید را نیز دارد که بسیار حائز اهمیت است.

۲- $\Pi(A)=0$ و $\Pi(\sim A)=1$: یعنی x یا یک نویز است و یا یکی از گروه‌ها بجز X.

۳- $\Pi(A)=1$ و $\Pi(\sim A)=0$: یعنی x یا یک نویز است و یا X.

۴- $\Pi(A)=1$ و $\Pi(\sim A)=1$: یعنی x یا یک نویز است و یا یکی از گروه‌هایی که می‌شناسیم (بیان ساده‌تر آن این است که بگوییم نمی‌دانم).

حالت اول به ما امکان می‌دهد که گونه‌های جدید را کشف و سپس مدل کنیم. زمانی که نمونه دیده شده شبیه هیچ کدام از گروه‌هایی که می‌شناسیم نیست، آن را نویز معرفی می‌کنیم. نکته مهم این است که بدون اندازه‌گیری امکانی اصلاح شده تعیین اینکه آیا یک نمونه به یک گروه شبیه است و یا خیر، شدنی نیست.

---

[1] توجه داریم که A~ بر اساس داده آموزشی به دست آمده است.

در عمل، ما از تمام واحدهای ARU می‌پرسیم که آیا نمونه x به آنها شبیه است یا خیر. اگر پاسخ تمام گروه‌ها منفی باشد ما آن نمونه را به عنوان نویز در نظر می‌گیریم و حذف می‌کنیم. توجه نمایید که نویز همواره به معنای دیده شدن یک نمونه جدید نیست. در بسیاری از موارد نویز بیانگر خطایی در بخش‌بندی است که باعث پرسیدن این سوال نامربوط شده است.

۷-٤-۷) مخلوط

در HMM فرض می‌شود داده هر واج دارای توزیع نرمال است. از آنجا که این فرض اصلا درست نیست، فرض می‌شود که داده مربوط به هر حالت در HMM دارای انواعی است که به آنها مخلوط می‌گویند و این انواع دارای توزیع نرمال هستند.

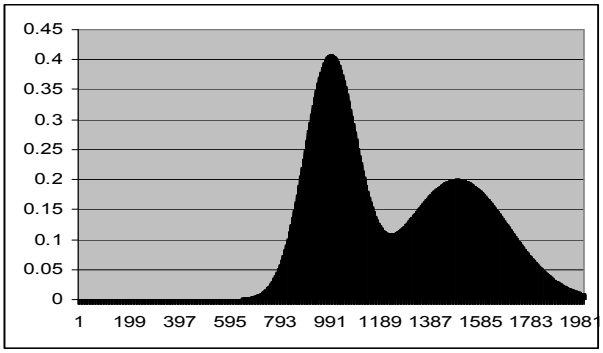
یکی از ویژگی‌های روش ما این است که مبتنی بر مدل (توزیع نرمال) نیست و در نتیجه نیازی به استفاده از مخلوط برای تقریب زدن بهتر توزیع ندارد. ما این قابلیت را به دست آورده‌ایم چون دامنه ویژگی را از $\mathbb{R}$ به $\{0,1,\ldots,99\}$ محدود کردیم.

اما ما نیز از مفهومی مشابه مخلوط استفاده می‌کنیم. هدف از استفاده از مخلوط، پوشش دادن گوناگونی موجود در انواع مختلف صحبت (صدای مرد، صدای زن و ...) است. ما برای هر گروه ARU چند مدل می‌سازیم تا ارتباطی که بین ویژگی‌های مختلف وجود دارد نیز حفظ شود. در عمل این کار باعث افزایش نتیجه شده است. برای مثال اگر مدل صدای مرد را A_m و مدل صدای زن را A_f بنامیم، داشتن چند مدل به ما اجازه نمی‌دهد که یک ویژگی از صدای مرد را با یک ویژگی از صدای زن مخلوط کنیم. در حالی که وقتی ما تنها یک مدل برای A می‌سازیم امکان چنین اشتباهی وجود دارد. در حقیقت این نوع مدل‌سازی معادل قوانین فازی است که در آن مجموعه‌ای از شروط نتیجه خاصی را تولید می‌کنند. به عبارت دیگر رمز برتری روش مبتنی بر قانون در این است که غیر مستقیم از مدل‌سازی درجه بالاتر استفاده می‌کند.

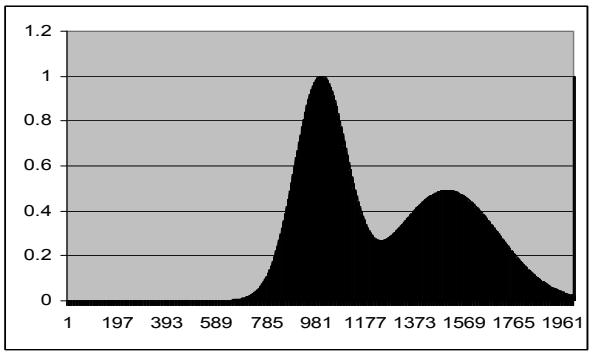
نکته مهم این است که تولید این مدل جدید از دیدگاه نظریه احتمال غلط است. علت این است که ما از توزیع امکان نرمال استفاده می‌کنیم که معادل حذف P(w) در قانون بیز است:

$$P(w \mid A) = \frac{p(A \mid w)P(w)}{P(A)}$$

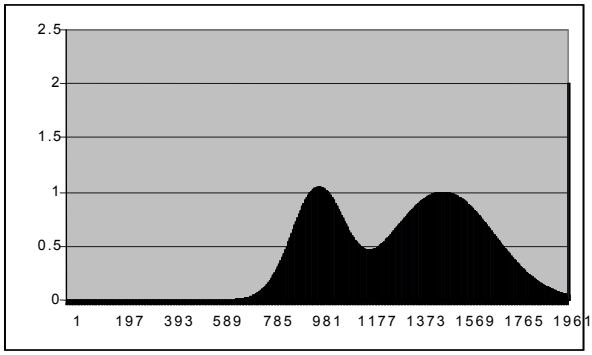
به عبارت دیگر ما به مدلی که با تعداد کمتری نمونه ساخته شده است به اندازه مدلی که با تعداد بیشتری نمونه ساخته شده است اهمیت می‌دهیم. شکل‌های زیر تفاوت نظریه احتمال را با نظریه امکان نشان می‌دهند.



شکل ۴۹: توزیع احتمال برای یکی از بردارهای ویژگی در گروه A



شکل ۵۰: توزیع امکان معادل شکل ۴۹



شکل ۵۱: هدف از تولید مخلوط در نظریه امکان رسیدن به این شکل است.

۷-٤-۸) استفاده از مدل اولیه بر اساس VQ

یکی از معایب روش فوق این است که نمونه‌های مخلوط‌ها به مرور کم می‌شوند و به مخلوط جدید داده می‌شوند. این کار گاهی آن‌قدر ادامه پیدا می‌کند که مخلوط‌های قبلی از بین می‌روند. در این بخش

یک روش مبتنی بر مدل را شرح می‌دهیم. فرض کنید که واحدهایARU همان واج‌ها هستند و هر یک ۸ مخلوط دارند. می‌خواهیم با استفاده از VQ[13]، توزیع امکان مخلوط‌ها را به دست آوریم.
مراحل کار عبارتند از:

۱- استخراج نمونه‌های هر واج
۲- به دست آوردن بردارهای حاصل از اعمال VQ بر نمونه‌های هر واج
۳- پخش کردن بردارهای به دست آمده برای رسیدن به توزیع امکان هر مخلوط
۴- استفاده از توزیع امکان هر مخلوط و یک فایل examples.1 برای به دست آوردن نمونه‌های هر مخلوط در پایگاه داده (examples.1). به عبارت دیگر، پایگاه داده‌ای که در آن مخلوط مشخص نشده است عوض می‌شود و پایگاه داده‌ای به دست می‌آید که در آن مخلوط مشخص شده است.
۵- استفاده از فایل examples.1 مشابه روش آموزش معمولی. تنها تفاوت این است که مخلوط جدیدی اضافه نمی‌شود.

برای به دست آوردن بردارهای VQ از ابزار HQuant در HTK استفاده شده است. خروجی HQuant تعدادی codebook است. هر ویژگی در هر گروه مانند A، دارای پراش مشخصی است که آن را با σ_A^2 نشان می‌دهیم. هر ویژگی را در هر codebook به اندازه ضریبی از واریانس به شکل یک تابع نرمال در فضای ویژگی–مقدار پخش می‌کنیم. بدین ترتیب از روی بردار codebook به توزیع امکان مربوط به آن می‌رسیم. در مرحله بعد هر نمونه از فایل examples.1 ورودی را که در آن فقط نام گروه مشخص شده است و مخلوط آن معلوم نیست را به مخلوطی که بیشترین شباهت را به آن دارد نسبت می‌دهیم و به فایل examples.1 خروجی می‌رسیم.

۷-۵) بازشناسی

ورودی سیستم بازشناسی دنباله‌ای از اشیاء است و هدف این است که از روی آن دنباله‌ای واجی تولید شود. ابتدا با استفاده از سیستم تشخیص دو مرحله‌ای برای هر شیء لیستی از واج‌های ممکن تولید می‌شود. در این پیاده‌سازی از میزان اعتقاد به این واج‌ها استفاده نشده است. اگر لیست واج‌های ممکن یکی از اشیاء تهی باشد، آن شیء را نویز در نظر می‌گیریم و مجازا از لیست اشیاء کنار می‌رود. اگر ARUها همگی واج باشند می‌توان دنباله واج‌های اول لیست اشیاء را به عنوان خروجی تولید کرد. اما آزمایش نشان می‌دهد که در این صورت خطای درج بسیار بالا خواهد بود. برای حل این مشکل از برنامه‌نویسی پویا برای یافتن کوتاه‌ترین دنباله واجی استفاده شد. یعنی ما ترجیح می‌دهیم که عنصر دوم

لیست را برداریم تا اینکه یک واج جدید درج کنیم. بدین ترتیب دقت ۱۰٪ افزایش یافت و دقت[1] و درستی[2] تقریبا مساوی شدند.

اگر برخی از ARUها دوواجی باشند، مجبوریم که ابتدا آنها را به دنباله واجی تبدیل کنیم. روش‌هایی که برای رسیدن به دنباله واجی آزمایش شدند عبارتند از:

۱- تبدیل هر دو واجی به دو واج تشکیل‌دهنده آن و سپس حذف واج‌های تکراری برای رسیدن به دنباله واجی.

۲- حذف واج دوم دوواجی‌ها. ایده اصلی این است که دوواجی‌های مهم به شکل «دا»، «نا» و ... هستند که واج دوم آنها از طرق دیگر نیز قابل تشخیص است و هدف از مدل‌سازی دوواجی تشخیص واج‌های کوچک (مانند «د») بوده است.

۳- حذف واکه‌های دوواجی‌ها و نگه‌داشتن واج‌های کوچک. در این حالت مهم نیست که واج‌های کوچک در کجا ظاهر می‌شوند.

اما مشکل اصلی این بود که افزایش تعداد گروه‌های آوایی احتمال اشتباه شدن بین دو گروه را افزایش می‌داد که این به نوبه خود منجر به پایین بودن نتایج شد[3]. به هر حال نتایج مدل‌سازی تک‌واجی از دوواجی بیشتر است.

۷-۶)بخش‌های پیاده‌سازی نشده

نمای کلی سیستم پیشنهادی ما در شکل ۵۲ نشان داده شده است. ما ادعا می‌کنیم که این سیستم نسبت به نویز پیچشی و تغییرات دامنه سیگنال مقاوم است.

۷-۶-۱) روش حذف نویز پیچشی

یکی از مشکلاتی که امروزه سیستم‌های تشخیص صحبت با آن مواجهند، طولانی بودن فرآیند وفق پیدا کردن با محیط استفاده است. روش‌های معمول در HMM برای حذف نویز MLLR و MAP هستند. روش MAP مبتنی برقانون بیز و فرض داشتن توزیعی از پارامترهای HMM است که پیدا کردن کوچک‌ترین ارتباطی بین آن و روش تشخیص صحبت در انسان بسیار مشکل است. اما اساس روش MLLR ایده بیشترین شباهت است که به نظر ما شباهت زیادی به روش انسان دارد. هدف تغییر دادن پارامترهای مدل اولیه به نحوی است که احتمال اینکه این مجموعه مشاهدات توسط این مدل تولید شده باشند بیشینه شود. مشکل این روش این است که باید دنباله مشاهدات صحیح را داشته باشیم که انسان‌ها برای وفق یافتن معمولا از این دنباله بی اطلاع هستند. همچنین این روش نسبتا کند است.

---

[1] Accuracy
[2] Correctness

[3] در حقیقت این سیستم برای استفاده از مزیت دوواجی‌ها ساخته شده بود اما در تک‌واجی بهتر پاسخ داد.

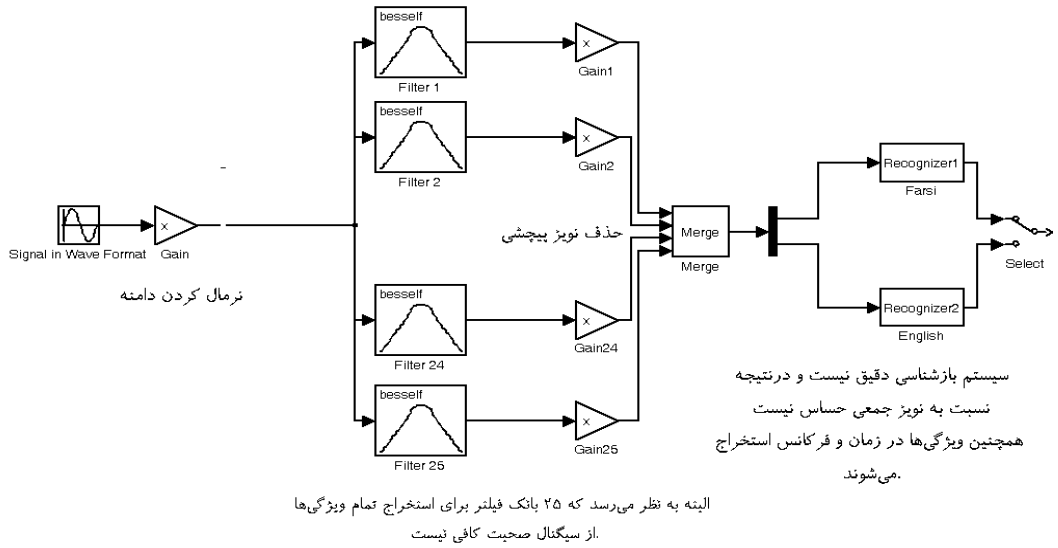
معمولا هدف از این تطبیق با محیط دو چیز است:

۱- حذف اثر نویز پیچشی

۲- یادگیری صدای گوینده

تجربه نشان می‌دهد که عامل اول از عامل دوم مؤثرتر است. یعنی پس از اینکه مدل HMM با محیط جدید وفق پیدا کرد و اثر نویز پیچشی حذف شد، صدای گوینده‌های دیگر نیز قابل بازشناسی است.

می‌دانیم که پیچش در زمان معادل ضرب در فرکانس است. به عبارت دیگر می‌توان فرض کرد که نویز پیچشی به صورت ضریبی در مقادیر بانک فیلتر (که ماهیت فرکانسی دارند) ظاهر شده است. حذف نویز نیز معادل ضرب کردن مقادیر بانک فیلتر در معکوس ضرایب نویز است. بدین ترتیب مساله وفق پیدا کردن با محیط را می‌توان به مساله یافتن این ضرایب ساده کرد.



شکل ۵۲: نمای کلی سیستم بازشناسی پیشنهادی. در این طرح نرمال‌سازی دامنه، حذف نویز پیچشی و پوشش گوناگونی صحبت دیده شده‌است. در حقیقت مساله تنظیم دامنه و حذف نویز پیچشی مسائلی کنترلی هستند.

به نظر ما انسان این ضرایب را با سعی و خطا و استفاده از قانون گرادیان پیدا می‌کند. تابع ارزیابی انسان برای محاسبه گرادیان چیست؟ همانطور که گفتیم، انسان صحبت را خوب می‌شناسد. پس اگر فرآیند حذف نویز موفقیت آمیز باشد، انسان توانسته است عبارت گفته شده را تشخیص دهد. در حقیقت انسان ضرایب حذف نویز را در جهتی تغییر می‌دهد که درجه شباهت جمله شنیده شده به سیگنال گفته شده بیشینه شود. هرگاه این تغییرات منجر به شنیدن یک عبارت ممکن شود نویز حذف شده است. شباهت

روش MLLR نیز در همین است. اما در HMM این روش بر روی پارامترهای مدل اعمال می‌شود.
علت اینکه انسان می‌تواند از چنین الگوریتمی استفاده کند این است که حالات ممکن تعداد کمی از کل پیام‌های قابل ارسال را تشکیل می‌دهند. بدین ترتیب رسیدن به یک حالت ممکن خود دلیل بر درستی تشخیص است.
چون سیستم ما مبتنی بر نظریه امکان است، اگر نمونه دیده شده دارای شباهت کافی به یک کلمه یا جمله نباشد، سیستم متوجه می‌شود. سپس فرآیندی را اجرا می‌کند که در آن سیستم سعی می‌کند با تغییر ضرایب حذف نویز که در ضرایب بانک فیلتر ضرب می‌شوند به سیگنال شبیه‌تری برسد. نکته مهم دیگر این است که در اینجا مساله تشخیص به مساله کنترل تبدیل شده است. در حقیقت مساله تنظیم ضرایب حذف نویز یک مساله کنترلی است.

۷-۶-۲) روش حذف اثر دامنه سیگنال

یکی دیگر از مشکلات امروزی سیستم‌های تشخیص صحبت مساله دامنه سیگنال است. این درحالی است که ما در محاسبه ضرایب بانک فیلتر از تابع لگاریتم استفاده کردیم که منجر به کم شدن تفاوت بین دامنه‌های بالا می‌شود و تفاوت بین ویژگی‌هایی که از سیگنال‌های یکسان و با دامنه‌های مختلف استخراج شده‌اند را از بین می‌برد.
اگر ویژگی‌های ما به مقدار انرژی حساس باشند، مجبوریم به طریقی دامنه سیگنال را نرمال کنیم. در حقیقت می‌توان ضریب دیگری را نیز اضافه کرد که در تمام بانک‌های فیلتر ضرب می‌شود و اثر آن تنظیم دامنه است. هدف، یادگیری یک فرآیند کنترلی است که در آن دامنه سیگنال را طوری تغییر می‌دهیم که سیگنال شنیده شده به سیگنال صحبت شبیه‌تر باشد. ورودی این سیستم کنترلی میزان اعتقاد به عبارت تشخیص داده شده و نیز ویژگی‌هایی مانند متوسط انرژی است که از سیگنال صحبت استخراج شده‌اند.

۷-۷) امتیازدهی

برای امتیازدهی به سیستم از ابزار HResults در HTK استفاده شده است. همچنین واج سکوت معمولا در امتیازدهی در نظر گرفته نشده است. برای کار با این نرم‌افزار باید خروجی سیستم بازشناسی به فرمت MLF باشد و همچنین در یک فایل MLF دیگر، دنباله صحیح در هر فایل نوشته شده باشد.

فصل ۸
آزمایش‌ها

بررسی توانایی OBSFE **در سیستمهای مبتنی بر تشخیص کلمه**
آزمایش OBSFE **بر روی پایگاه داده** Aurora2 **با استفاده از ابزار** HTK
نتایج به دست آمده از ویژگیهای MFCC_0_D_A **با پیاده‌سازی نگارنده**
بررسی توانایی OBSFE **در سیستمهای مبتنی بر تشخیص واج**
بررسی توانایی سیستم امکانی در مقابل درج اشیاء ناشناخته
بررسی توانایی سیستم بازشناسی امکانی بر روی ویژگی‌های MFCC

۸-۱) بررسی توانایی *OBSFE* در سیستم‌های مبتنی بر تشخیص کلمه

۸-۱-۱) آزمایش OBSFE بر روی پایگاه داده Aurora2 با استفاده از ابزار HTK

سیستم HTK دارای قابلیتی است که می‌تواند بدون استفاده از تقطیع دستی و تنها بر اساس دنباله کلمات آموزش ببیند. برای این منظور ابتدا با یک محاسبه آماری مدل تمام کلمات را یکسان قرار می‌دهد. این باعث می‌شود که در مرحله بعدی آموزش HMM، سیگنال صحبت به بخش‌های مساوی تقسیم شود و بر اساس آن مدل اولیه‌ای برای کلمات بدست آید. با تکرار بیشتر سیستم می‌تواند مدل کلمات را کاملا یاد بگیرد.

در مرحله استخراج ویژگی، دنباله ویژگی‌هایی را که از روش OBSFE به دست می‌آید را تولید کردیم. چون تعداد اشیاء کمتر از تعداد حالات مدل HMM بود هر شیء را سه بار تکرار کردیم. سپس با HTK مدل کلمات (اعداد) را آموزش دادیم. نتایج با سیستم مبنا که مبتنی بر استخراج ویژگی به روش MFCC_0_D_A[1] است مقایسه شد.

جدول ۵: دقت OBSFE در مقابل MFCC_0_D_A بر روی زیرمجموعه‌ای از پایگاه داده Aurora2. آموزش بر روی داده تمیز انجام شده است. نویزهای مترو، نمایشگاه و خیابان حذف شده‌اند.

	A			B						WER
	Babble	Car	Average	Restaurant	Airport	Station	Average	Overall	Overal (BL)	Improvement
15dB	85.40	88.04	***86.72***	81.52	86.34	86.08	***84.65***	85.48	80.43	25.77%
10dB	69.92	77.75	***73.84***	60.70	74.53	72.48	***69.24***	71.08	57.26	32.32%
5dB	35.70	49.51	***42.61***	26.16	42.02	44.55	***37.58***	39.59	30.10	13.57%
Average	63.67	71.77	***67.72***	56.13	67.63	67.70	***63.82***	65.38	55.93	21.44%
Average(BL)	50.00	63.71	***56.86***	54.01	53.73	57.29	***55.01***	55.93	**21.44%**	
WER Improvement	27.35%	22.19%	25.18%	4.61%	30.04%	24.39%	19.58%	21.44%		

جدول ۶: دقت و صحت متوسط OBSFE و MFCC_0_D_A بر روی دو گروه از نویزها

[1] این یک اصطلاح در HTK است و به معنای ویژگی ضرایب کپسترال فرکانس Mel به همراه c_0 و مشتق اول و دوم است.

	Subway,Exhibition, Street				Babble, Car, Resturant, Airport, Station			
	SSFE		MFCC_0_D_A		SSFE		MFCC_0_D_A	
	Corr	Acc	Corr	Acc	Corr	Acc	Corr	Acc
Clean	*92.74*	*91.34*	*99.42*	*99.04*	*92.80*	*91.45*	*99.39*	*99.01*
20dB	*91.95*	*89.20*	*97.89*	*96.39*	*91.57*	*89.87*	*98.43*	*92.58*
15dB	*87.13*	*81.65*	*95.08*	*91.33*	*89.38*	*85.48*	*95.36*	*80.14*
10dB	*78.04*	*64.10*	*84.22*	*73.83*	*83.35*	*71.08*	*83.26*	*57.07*
5dB	*58.29*	*32.94*	*59.11*	*45.15*	*67.16*	*39.59*	*58.03*	*30.03*

در مرحله استخراج ویژگی از ۱۸ باند فیلتر استفاده شد و از هر باند فیلتر ۴ ویژگی استخراج شد. ویژگی‌های استخراج شده از هر باند عبارتند از انرژی، تحدب/تقعر، شیب و مرکز ثقل. ما از DCT برای کاهش تعداد هر نوع ویژگی از ۱۸ به ۹ استفاده کردیم. در مجموع ۹ ویژگی انرژی، ۹ ویژگی تحدب/تقعر، ۹ ویژگی شیب، ۹ ویژگی مرکز ثقل و یک ویژگی طول زمانی از هر بخش استخراج شدند. بدین ترتیب تعداد ویژگی‌های ما (۳۷) از ویژگی‌های MFDCC_0_D_A (۳۹) کمتر نیز هست.در جدول ۵ ما نتایج بازشناسی را در شرایطی که منجر به افزایش نرخ تشخیص شده است آورده‌ایم.

ما به طور متوسط درصد بازشناسی را در نویزهای همهمه، اتومبیل، رستوران، فرودگاه و ایستگاه قطار که ما به آنها گروه اول نویزها می‌گوییم، ۲۱.۴٪ افزایش داده‌ایم. ما نویزهای مترو، نمایشگاه و خیابان را که در آن به نتایج پایین‌تری دست یافته‌ایم، گروه دوم نویزها می‌نامیم. در جدول ۶ ما متوسط درصد بازشناسی را در این دو گروه نویز مقایسه کرده‌ایم.

مقایسه OBSFE و MFCC_0_D_A حداقل دو مطلب را نشان می‌دهد:

۱- OBSFE در دسیبل‌های 15dB, 10dB, 5dB و در نویزهای همهمه، اتومبیل، رستوران، فرودگاه و ایستگاه قطار بسیار بهتر عمل می‌کند.

۲- OBSFE در دسیبل‌های 15dB, 10dB, 5dB در نویزهای مترو، نمایشگاه و خیابان بدتر عمل می‌کند. در این شرایط با وجود اینکه دقت کاهش یافته است، اما صحت همچنان بالا است.

این نشان می‌دهد که استحکام روش ما مستقل از نوع نویز نیست. تفاوت نویزهای گروه ۱ با نویزهای گروه ۲، مثلا نویز همهمه با نویز مترو، در چیست؟ نویز همهمه اشیاء جدید تولید نمی‌کند و در نتیجه باعث افزایش تعداد بخش‌ها نمی‌شود. ولی نویز مترو صداهایی را که متناظر با مراحل مختلف حرکت قطار هستند تولید می‌کند. این نویز منجر به درج اشیائی می‌شود که مشابهتی با صدای صحبت ندارند. از آنجا که سیستم HTK فرض می‌کند که همه بردارهای ویژگی از سیگنال صحبت استخراج شده‌اند، این اشیاء اضافی به اشتباه به کلماتی نسبت داده می‌شوند و درنتیجه دقت کاهش می‌یابد (البته درستی همچنان بالا می‌ماند).

۸-۱-۲) نتایج به دست آمده از ویژگی‌های MFCC_0_D_A با پیاده‌سازی نگارنده

در ادامه بحث می‌خواهیم خطایی را که هر یک از مراحل استخراج ویژگی ایجاد می‌کنند بررسی کنیم. نتایجی که سیستم استخراج ویژگی Aurora2 می‌دهد مرهون برخی پیش‌پردازش‌ها مانند حذف دامنه ثابت[1] و پیش‌تاکید[2] است. اگر این دو خاصیت را حذف کنیم دقت به ۹۸.۷۴٪ (از ۹۸.۹۳٪) و درستی به ۹۹.۳۲٪ (از ۹۹.۳۹٪) می‌رسد. چون سیستم استخراج ویژگی ما این بخش‌ها را ندارد، بررسی خطای ایجاد شده توسط مراحل مختلف استخراج ویژگی باید با پیاده‌سازی ما از MFCC_0_D_A انجام شود. به همین دلیل ما از خروجی بانک فیلتر سیستم استخراج ویژگی خود برای به دست آوردن MFCC_0_D_A استفاده کردیم. بدین ترتیب ما به دقت ۹۷.۱۱٪ و درستی ۹۸.۲۵٪ رسیدیم.

۸-۱-۳) بررسی خطای ناشی از تقریب زدن با خط

در این بخش ابتدا سیگنال را با خط تقریب زدیم و سپس دوباره از روی آن اطلاعات قاب‌ها را به دست آوردیم و ویژگی‌های MFCC_0_D_A را تولید کردیم. بدین ترتیب دقت سیستم از ۹۷.۱۱٪ به ۹۵.۷۶٪ و درستی آن از ۹۸.۲۵٪ به ۹۶.۸۴٪ رسید.

۸-۱-٤) بررسی خطای ناشی از کوانته کردن به ۱۰۰ مقدار

در این آزمایش مقدار ویژگی‌های MFCC_0_D_A که توسط سیستم همراه Aurora2 استخراج شده‌اند را بر حسب صدک به اعداد صحیح بین ۰ تا ۱۰۰ می‌نگاریم. در نتیجه آن دقت سیستم از ۹۸.۹۳ به ۹۸.۲۲ و صحت آن از ۹۹.۳۱ به ۹۸.۹۶ رسید. این نتیجه به خوبی ادعای عدم نیاز به دقت بالا برای ساختن یک سیستم تشخیص صحبت را تایید می‌کند. این مطلب همچنین نشان می‌دهد که در سیستم ما نیز این بخش دقت را پایین نیاورده است.

۸-۱-۵) نتیجه‌گیری

در این بخش نشان داده شد که روش OBSFE از مقاومت خوبی در مقابل نویزهایی که شیء اضافه نمی‌کنند برخوردار است. همچنین نتایجی که با این روش در سطح کلمه به دست آمد امیدوار کننده است[3].

---

[1] DC offset removal
[2] Pre-emphasis

[3] نگارنده با استفاده از روش OBSFE2 به دقت ۹۵٪ در داده تمیز نیز دست یافته است.

۸-۲) بررسی توانایی OBSFE در سیستم‌های مبتنی بر تشخیص واج

۸-۲-۱) آزمایش OBSFE بر روی دادگان فارس‌دات با سیستم بازشناسی امکانی

در این بخش نتایج سیستم بازشناسی خود را بر روی بخش تهرانی دادگان فارس‌دات می‌آوریم. توجه داریم که داده آموزشی و تست کاملا مجزا هستند[۱]. پارامترهای سیستم در شکل ۵۳ نشان داده شده است. ابتدا برای هر واج ۸ مخلوط در نظر گرفته شده است که توزیع امکان آنها با استفاده از VQ به دست می‌آید (مجموعا ۲۴۰ مخلوط). سپس سیستم یک‌بار دیگر به اصلاح نام‌دهی خود می‌پردازد و مخلوط‌های کم نمونه را حذف می‌کند و از ۲۴۰ مخلوط ۱۵۷ مورد آنها باقی می‌مانند. در این حالت دقت سیستم ۳۴.۹۷٪ و درستی آن ۳۸.۳۱٪ است. این نتایج نسبتا پایین است. در ادامه به دنبال بررسی علت پایین بودن نتایج برای پیشنهاد روش مناسبی برای ادامه کار هستیم.

```
NEWARUCRITERIA          MINCOUNT
NEWARUMINCOUNT          70
MODELNOISE              FALSE

LEASTPOSSIBILITYRATIO1  0.25
LEASTPOSSIBILITY1       5

LEASTPOSSIBILITYRATIO2  0.75
LEASTPOSSIBILITY2       30

SHORTPHONEMES           -1
NEWPHONEMEPENALTY       10
COST_OF_LENGTH          0 1 2 3 4 5 6 7 8 9 10 -1
FAVOURHANDYLABELS       TRUE
LEAST_X_POSSIBILITY     10
NEGATIVE_MODEL          OTHER_PHONEMES
WEIGHT_FEATURES         TRUE
XOR_LEARN               FALSE
PHONEME_ONLY            TRUE
```

شکل ۵۳: پارامترهای سیستم ما هنگام آموزش و بازشناسی بخش تهرانی از دادگان فارس‌دات.

۸-۲-۲) آزمایش OBSFE بر روی دادگان فارس‌دات با سیستم HTK

در این بخش می‌خواهیم توانایی مدل HMM را در کار با ویژگی OBSFE بررسی کنیم. هدف از این آزمایش تعیین توقعی است که از سیستم بازشناسی امکانی می‌خواهیم داشته باشیم. توجه داریم که

[۱] جالب است که اگر داده آموزشی و تست یکسان باشد نتایج حدود ۶۰٪ است.

ویژگی‌های استخراج شده در روش OBSFE برای کلمه خوب عمل کرده بودند. نتایجی که در این قسمت خواهد آمد نشان می‌دهد که ویژگی‌های OBSFE برای سطح واج مناسب نیستند. برای آزمایش از سیستم HTK استفاده کردیم. مدل هر واج شامل ۱ حالت و ۸ مخلوط است (چون تعداد اشیاء کم است نمی‌توان از مدل‌های پیچیده‌تر استفاده کرد). همچنین داده آموزشی شامل تمام فایل‌های فارس‌دات و داده تست نیز همان داده آموزشی است. اگر سکوت را نیز در امتیازدهی شریک کنیم دقت سیستم ۲۶.۶۷٪ و درستی آن ۵۰.۸۷٪ است و اگر در امتیازدهی سکوت را در نظر نگیریم دقت سیستم ۲۲.۳۶٪ و درستی آن ۴۹.۲۵٪ می‌باشد. نکته‌ای که در اینجا جلب توجه می‌کند فاصله زیاد دقت و صحت است.

۸-۲-۳) بررسی خطای ناشی از دیده نشدن واج‌ها در بخش‌بندی

در اینجا می‌خواهیم ببینیم که اگر بتوانیم تمام واج‌های آموزشی را درست تشخیص دهیم به چه دقتی خواهیم رسید. توجه داریم که در اینجا برای هر واج یک شیء تولید نمی‌شود و واج‌هایی هستند که برای آنها هیچ شیئی وجود ندارد. آزمایش نشان می‌دهد که حتی اگر تمام بخش‌ها را درست (بر اساس بخش‌بندی دستی) تشخیص دهیم دقت سیستم ۸۴٪ خواهد بود.

۸-۲-٤) آزمایش MFCC بر روی دادگان فارس‌دات با سیستم شرکت عصر گویش

در این بخش اثر کوانته کردن به ۱۰۰ مقدار را بر روی سیستم مبتنی بر HMM شرکت عصر گویش بررسی می‌کنیم. آموزش بر روی زیر مجموعه تهرانی از دادگان فارس‌دات انجام شده است. ۳۶ ویژگی MFCC استخراج می‌شود و مدل HMM دارای ۳ حالت و ۱۶ مخلوط است. سیستم اولیه دارای دقت ۶۶.۷۸ و درستی ۶۷.۲۷ است.

۸-۲-۵) بررسی خطای ناشی از تقریب زدن با خط

در این بخش اثر تقریب زدن خط سیر انرژی با خط را در بانک‌های فیلتر بررسی می‌کنیم. در این آزمایش از سیستم HTK برای تشخیص واج استفاده می‌کنیم. در این آزمایش از یک HMM با ۵ حالت و ۸ مخلوط برای تشخیص واج در دادگان فارس‌دات استفاده شده است. نتایج بر روی ویژگی‌های MFCC برابر دقت ۶۱.۱۶٪ و درستی ۷۷.۶۵٪ است. پس از تقریب زدن داده تست با خط، دوباره ویژگی‌های MFCC را به دست آوردیم. بدین ترتیب سیستمی را که بر روی ویژگی‌های تقریب‌زده نشده آموزش دیده بود را بر روی ویژگی‌هایی که پس از تقریب خطی به دست آمدند آزمایش کردیم. بدین ترتیب دقت به ۴۳.۳۹٪ و درستی به ۶۶.۴۹٪ کاهش یافت. در نهایت سیستم را با ویژگی‌های تقریب زده شده آموزش دادیم و بر روی ویژگی‌های تقریب زده شده تست کردیم. در این حالت دقت ۶۰.۰۶٪ و درستی ۷۵.۶۶٪ است. واقعا جالب است که تقریب خطی تنها ۱٪ نتایج را پایین آورده است.

۸-۲-۶) بررسی خطای ناشی از کوانته کردن به ۱۰۰ مقدار

در این بخش اثر کوانته کردن به ۱۰۰ مقدار را بر روی سیستم مبتنی بر HMM شرکت عصر گویش بررسی می‌کنیم. آموزش بر روی زیر مجموعه تهرانی از دادگان فارس‌دات انجام شده است. ۳۶ ویژگی MFCC استخراج می‌شود و مدل HMM دارای ۳ حالت و ۱۶ مخلوط است. سیستم اولیه دارای دقت ۶۶.۷۸ و درستی ۶۷.۲۷ است. پس از کوانته شدن به مقادیر صحیح بین ۰ تا ۱۰۰ بر حسب صدک‌ها دقت به ۵۲.۷ و درستی به ۵۳.۰۵ رسید. اما ما آزمایشی ترتیب دادیم تا نشان دهیم که این کاهش دقت بخاطر وابستگی سیستم HMM به مقدار دقیق ویژگی‌ها[1] است و نه کوانته شدن به ۱۰۰ مقدار. در این آزمایش ما مقادیر ویژگی‌ها را به ۱۰۰ مقدار صدک‌ها کوانته کردیم ولی بجای مقادیر ۰ تا ۱۰۰ به عنوان ویژگی خود مقدار صدک‌ها را قرار دادیم. بدین ترتیب داده ما به داده اولیه شبیه‌تر خواهد بود. بدین ترتیب دقت به ۶۴.۹۵ و درستی به ۶۵.۲۵ رسید. جدول ۷ خلاصه نتایج را نشان می‌دهد.

جدول ۷: نتایج آزمایش کوانته کردن ویژگی‌های MFCC استاندارد سیستم مبتنی بر HMM به ۱۰۰ سطح.

شماره آزمایش	نوع ویژگی‌ها	دقت	درستی
۱	MFCC معمولی	۶۶.۷۸	۶۷.۲۷
۲	MFCC کوانته شده به مقادیر صحیح بین ۰ تا ۱۰۰	۵۲.۷	۵۳.۰۵
۳	MFCC کوانته شده به ۱۰۰ مقدار صدک	۶۴.۹۵	۶۵.۲۵

۸-۲-۷) نتیجه‌گیری

آزمایش‌های این بخش نشان می‌دهند که روش OBSFE در سطح واج مناسب نیست. همچنین کوانته کردن به ۱۰۰ مقدار و نیز تقریب زدن با خط در این عدم موفقیت تاثیری ندارد. مشخصا تعداد کم اشیاء و حذف شدن بسیاری از آنها تاثیر بسزایی در پایین بودن نتایج دارد. به نظر می‌رسد می‌توان با مدل‌سازی بهتر اشیاء نتایج را بالا برد.

۸-۳) بررسی توانایی سیستم امکانی در مقابل اشیاء ناشناخته

خواننده به‌خاطر دارد که در مورد نتایجی که با HTK و ویژگی‌های OBSFE برروی Aurora2 به دست آوردیم در مواردی که سیستم جواب پایین‌تری داده بود مشکل را به سیستم بازشناسی نسبت دادیم. حال می‌خواهیم ببینیم که آیا سیستم بازشناسی ما توانایی کافی برای نادیده گرفتن اشیاء جدید را دارد یا خیر.

[1] برای مثال در HMM پارامتری وجود دارد که حداقل واریانس را مشخص می‌کند. این قبیل پارامترها به ما اجازه تغییر در مقدار ویژگی‌ها را نمی‌دهند. هرچند می‌توان نشان داد که اگر مقادیر ویژگی‌ها همگی در عدد ثابتی ضرب شوند نتیجه باید ثابت باشد، ولی در عمل مشکلاتی از قبیل مقدار آغازی مدل باعث می‌شود که حتی ضرب تمام ویژگی‌ها در عددی ثابت نیز منجر به تغییر شدید در نتیجه شود.

بدین منظور این آزمایش را ترتیب دادیم. فرض کنید برای یک سیگنال صحبت n شیء استخراج شده است. n بردار ویژگی را به شکل کاملا تصادفی تولید می‌کنیم و در میان بردارهای ویژگی اولیه قرار می‌دهیم. توقع ما از سیستم بازشناسی امکانی این است که متوجه جدید بودن این اشیاء بشود و آنها را نادیده بگیرد. جدول ۸ نتایج اولیه سیستم ما و سیستم HTK را بر روی ویژگی‌های OBSFE و نیز نتایجی را که پس از اضافه شدن اشیاء ناشناخته به دست آمده است نشان می‌دهد. سیستم HTK با تمام فارس‌دات آموزش دیده است و سپس بر روی همان داده آموزشی (که اکنون نویز به آن اضافه شده است) آزمایش شده است. همانطور که دیده می‌شود ما به هدف خود کاملا دست یافته‌ایم. درحالی که نتایج در سیستم ما کمترین افتی نداشته است، نتایج سیستم HMM به ٪۲۷۲.۶۱- رسیده است[1].

جدول ۸: مقایسه بین سیستم بازشناسی احتمالی بدون قابلیت حذف نویز و بازشناسی امکانی از نظر مقاومت نسبت به درج اشیاء جدید

شماره آزمایش	نوع ویژگی‌ها	سیستم بازشناسی	دقت٪	درستی٪
۱	OBSFE	HMM	۲۲.۳۶	۴۹.۲۵
۲	OBSFE	سیستم ما	۳۴.۹۷	۳۸.۳۱
۳	OBSFE به همراه نویز	HMM	-۲۷۲.۶۱	۵۰.۰۸
۴	OBSFE به همراه نویز	سیستم ما	۳۴.۹۷	۳۸.۳۱

۸-٤) بررسی توانایی سیستم بازشناسی امکانی بر روی ویژگی‌های *MFCC*

آزمایش‌هایی که توسط سیستم امکانی بر روی ویژگی‌های OBSFE انجام شد نشان داد که بیشتر خطا مربوط به خطای حذف است. بدین ترتیب به نظر می‌رسد که بالا بردن تراکم بردارهای ویژگی در ثانیه باید به بالا رفتن نتیجه بیانجامد. از آنجا که با این کار خطای درج بالا می‌رود، سیستم بازشناسی امکانی را تغییر دادیم تا یک واج تشخیص‌داده شده را تنها در صورتی که حداقل سه قاب مجاور آن را تایید کنند، بپذیرد. این کار مشابه استفاده از مفهوم حالت در مدل مخفی مارکوف است.

ما در مجموع ۷۷۷ مخلوط برای واج‌ها ساختیم. در این آزمایش حداقل تعداد نمونه لازم برای تشکیل مخلوطی از یک واج برابر ۱۰۰ نمونه است. همچنین مخلوط‌های اولیه با استفاده از VQ به دست آمدند. برای آموزش و تست از داده‌های آموزش و تست مجموعه تهرانی از دادگان فارس‌دات استفاده کردیم. بدین ترتیب ما به دقت ۴۹.۳۹٪ و درستی ۶۴.۵٪ در تشخیص واج رسیدیم. جالب است که دقت سیستم بر روی داده آموزشی ۸۰.۸۳٪ و درستی آن برابر ۹۰٪ است.

[1] ممکن است خواننده علاقمند به HMM بگوید که مقایسه شما نابرابر بوده است. ما نیز با این گفته موافقیم. حقیقت این است که ما در اینجا در سدد مقایسه دو روش نیستیم و تنها می‌خواهیم نشان دهیم که به هدف خود در ساختن سیستمی که اشیاء جدید را رد می‌کند دست یافته‌ایم.

نتیجه گیری

در این پایان‌نامه ابتدا روش انسان در تشخیص صحبت بررسی شد. می‌توان نتایجی را که از این بررسی به دست آمده است چنین خلاصه کرد:

۱- ویژگی‌هایی که انسان استخراج می‌کند صرفا فرکانسی نیستند و دارای ماهیت زمانی-فرکانسی هستند.

۲- ویژگی‌هایی که انسان به آنها توجه می‌کند دارای انرژی کافی هستند. اگر ویژگی با انرژی مناسب در خود واج نباشد، انسان از اطلاعات میان‌واجی و یا واج‌های دیگر استفاده می‌کند.

۳- دقت انسان از بیشترین دقت قابل حصول پایین‌تر است. برای مثال درحالی که ماشین بر روی پایگاه داده Aurora2 ۹۹٪ پاسخ می‌دهد، نگارنده حدس می‌کند دقت خودش حداکثر ۹۰٪ باشد. این که انسان سیستمی بهینه نباشد ولی بسیار خوب کار کند با برداشت ما از فازی سازگار است.

سپس تلاش شد تا بر اساس تغییرات انرژی فرکانس‌ها در بین قاب‌ها مرز بین واج‌ها بدست آید. بدین ترتیب موفق شدیم حدود ۴۰٪ از لبه‌ها را بدون اشتباه تشخیص دهیم. به هر حال به این نتیجه رسیدیم که برخی تغییرات درون‌واجی از گذر بین برخی واج‌ها شدیدتر است. بدین ترتیب با الهام از برخی مقالات به سراغ بخش‌بندی شنوایی رفتیم که در آن رابطه بین بخش‌ها و واج‌ها (و اصولا بخش‌های آوایی) اهمیتی ندارد. بدین ترتیب روش OBSFE به عنوان یک روش بخش‌بندی و استخراج ویژگی ارائه شد. ویژگی‌های این روش عبارتند از:

۱- ویژگی‌ها در دامنه زمان-فرکانس استخراج می‌شوند.

۲- ویژگی‌ها نسبت به نویز مقاوم‌تر هستند.

۳- بخش‌ها دارای همپوشانی هستند.

۴- ویژگی‌ها برای انسان قابل درک و تفسیر هستند.

آزمایش این روش بر روی پایگاه داده Aurora2 نشان داد که این روش استخراج ویژگی نسبت به MFCC بسیار مقاوم‌تر است و نرخ خطا را در نویزهای همهمه، اتومبیل، رستوران، فرودگاه و ایستگاه قطار (که آنها را نویزهای نوع اول می نامیم) با شدت‌های 5dB، 10dB و 20dB به میزان ۲۱.۴۴٪ افزایش می‌دهد. اما این روش در نویزهای مترو، نمایشگاه و خیابان (که آنها را نویزهای نوع دوم می نامیم) بدتر عمل می‌کند. با توجه به اینکه پایین‌تر بودن دقت روش OBSFE نسبت به MFCC در نویزهای نوع دوم به علت خطای درج بوده است و نیز با توجه به ماهیت این نویزها به این نتیجه رسیدم که علت پایین‌تر بودن نتایج در این نویزها حضور صداهایی است که به صدای انسان شباهت ندارد (مانند صدای چرخ قطار). می‌توان گفت که تقریب زدن خط سیر انرژی در بانک‌های فیلتر باعث افزایش

استحکام سیستم و در عین حال کاهش دقت بازشناسی در حالت تمیز می‌شود. این آزمایش‌ها با سیستم بازشناسی HTK که مبتنی بر مدل مخفی مارکوف است انجام شده‌اند.

بدین ترتیب ما تصمیم گرفتیم که یک سیستم بازشناسی مشابه انسان بسازیم. این سیستم بازشناسی باید از تشخیص خود مطمئن باشد. در چنین سیستمی مساله فقط انتساب یک نمونه به یکی از گروه‌های شناخته شده نیست. مساله، انتساب یک نمونه به یکی از گروه‌های شناخته شده و یا گروه اشیاء ناشناخته است. بدین ترتیب می‌توان به مرور اشیاء ناشناخته را نیز طبقه‌بندی کرد و گروه‌های شناخته شده‌ای را به گروه‌های موجود اضافه کرد. از آنجا که نظریه امکان اجازه مدل‌سازی جهل را می‌دهد، تصمیم گرفتیم یک سیستم بازشناسی مبتنی بر نظریه امکان بسازیم. در این سیستم به‌جای عملگر جمع از عملگر max و به‌جای عملگر ضرب از عملگر min استفاده می‌شود. آزمایش بر روی این سیستم حدس ما در مورد توانایی سیستم مبتنی بر نظریه امکان در تشخیص نویز را تایید کرد. در حالی که دقت سیستم HTK به ۲۷۲.۶۱- رسیده بود، دقت سیستم مبتنی بر نظریه امکان ثابت مانده بود. اما مشکل اینجا بود که دقت سیستم مبتنی بر نظریه امکان و ویژگی‌های OBSFE در تشخیص واج تنها ۳۶٪ است که نسبت به HMM با ویژگی‌های MFCC که دقتی حدود ۶۰٪[1] دارد پایین‌تر است. تلاش نگارنده برای بالاتر بردن نتایج به بیش از ۳۶٪ موفقیت آمیز نبود. به همین دلیل حدس زدیم که احتمالا ویژگی‌های OBSFE برای تشخیص واج مناسب نیستند. برای آزمایش این حدس، از یکHMM با یک حالت و ۸ مخلوط برای تشخیص واج بر اساس ویژگی‌های OBSFE استفاده کردیم. آزمایش نشان داد که دقت این سیستم تنها ۲۲٪ است. بدین ترتیب به نظر می‌رسد که برای استفاده از روش OBSFE در تشخیص واج باید تغییراتی در آن صورت گیرد. برای تعیین علت پایین‌تر بودن تشخیص سیستم مبتنی بر OBSFE نسبت به MFCC آزمایش‌هایی ترتیب دادیم تا میزان خطایی که هر یک از گام‌های الگوریتم استخراج ویژگی OBSFE بوجود می‌آید را بدست آوریم. نتایج این آزمایش‌ها چنین است:

۱- خطایی که نوشتن ویژگی‌ها بر حسب صدک در تشخیص کلمه ایجاد می‌کند حدود ۰.۵٪ است. دقت سیستم اولیه ۹۸.۹۳٪ است.

۲- خطایی که نوشتن ویژگی‌ها بر حسب صدک در تشخیص واج ایجاد می‌کند حدود ۲٪ است. دقت سیستم اولیه ۶۶.۷۸٪ است.

۳- خطایی که تقریب زدن با خط و سپس نوشتن ویژگی‌ها بر حسب صدک در تشخیص کلمه ایجاد می‌کنند کمتر از ۲٪ است.

۴- خطایی که تقریب زدن با خط بدون نوشتن ویژگی‌ها بر حسب صدک در تشخیص واج ایجاد

---

[1] چون نحوه ارزیابی دو سیستم متفاوت است، نگارنده حدس خود را در مورد نتایجی که پس از هماهنگ شدن سیستم‌ها به‌دست می‌آید در اینجا نوشته است. سیستم ما واج سکوت را در امتیازدهی لحاظ نمی‌کند که حدس زده می‌شود حدود ۵٪ نتایج سیستم HMM بخاطر تشخیص این واج است.

می‌کنند حدود ۱٪ است.

۵- خطایی که تعداد کم اشیاء استخراج شده ایجاد می‌کند حداقل برابر ۱۶٪ خطای حذف است. یعنی اگر تمام اشیاء را نیز درست تشخیص دهیم دقت برابر ۸۴٪ خواهد بود.

۶- از آنجا که دقت سیستم بر روی داده آموزشی ۶۰٪ و بر روی داده تست ۳۶٪ است، نتیجه می‌شود که بخش بندی قطعی خطای سیستم را افزایش می‌دهد. یکی از حسن‌های HMM این بود که بخش‌بندی را پس از تشخیص به دست می‌آورد. اما ما ترجیح می‌دهیم این مشکل را با افزایش تعداد بخش‌ها و در نتیجه داشتن تعداد زیادی بخش با همپوشانی زیاد حل کنیم. علت این رجحان این است که در عوض سیستمی به دست می‌آید که قابل تفسیر است.

پیشنهادات: محور های مطالعه و گسترش بیشتر

با توجه به نتایجی که از این پایان‌نامه به دست آمد موارد زیر به عنوان ادامه کار پیشنهاد می‌شوند:

۱- کدر ۲۰۰ بیت در ثانیه

بررسی روش‌های کد کردن صحبت [53][38][63] نشان می‌دهد که روش‌های متداول کد کردن صحبت از اطلاعات مجاورتی بین قاب‌ها بهره کافی را نمی‌برند. همچنین تلاش‌هایی که جدیدا در این راستا صورت گرفته است فشردگی را به شدت افزایش داده است. در حقیقت ما با تقریب زدن خط سیر انرژی در بانک‌های مختلف فیلتر نشان دادیم که می‌توان اطلاعات بانک‌های فیلتر را به شدت خلاصه کرد. تنها اطلاع دیگری که مورد نیاز است اطلاعات مربوط به pitch و منبع انرژی (پالس/نویز) است.
روش دیگر فشرده کردن مبتنی بر کد کردن اشیاء است. با توجه به اینکه تغییرات در اشیاء رخ می‌دهند، می‌توان امیدوارتر بود که اطلاعات صحبت منتقل شده است.

۲- منطق خط سیر

همانطور که در بخش ۷-۴-۷ و ۷-۴-۸ ذکر شد، می‌توان به‌جای محاسبه میانگین و واریانس خط سیرهای موجود در فضا را به دست آورد. پیدا کردن یک مبنای ریاضی مناسب برای کار با این خطوط سیر یکی از مسیرهای ادامه پروژه است.

۳- یک سیستم بازشناسی کامل مبتنی بر نظریه امکان

ما در این پروژه یک پیاده‌سازی مختصر (در مقایسه با HMM) از یک سیستم بازشناسی مبتنی بر نظریه امکان ارائه دادیم. یکی از عیب‌های این روش مدل‌سازی، عدم وجود مفهومی مشابه مفهوم حالت در HMM است. یکی از محورهای ادامه پروژه، تولید یک سیستم بازشناسی کامل مبتنی بر نظریه امکان است.

٤-اصلاح ویژگی‌های *OBSFE* برای استفاده در سطح واج

آزمایش‌های ما در تشخیص صحبت مبتنی بر واج موفقیت آمیز نبود. به نظر می‌رسد که متهم اصلی در این عدم موفقیت، روش استخراج ویژگی ما است. این روش مزایای فراوانی دارد که در متن پایان‌نامه به آن اشاره شده است. یکی از کارهای بعدی اصلاح این روش برای استفاده شدن در تشخیص واج است.

۵- استفاده از گراف مفهومی برای تولید سیستمی که به مرور زبان را یاد می‌گیرد

گراف مفهومی [51] یکی از روش‌های نمایش دانش است که از انسان الهام گرفته شده است. مبنای آن ارتباط بین اشیاء در مغز است. در حقیقت همه چیز در مغز شیء است و اشیاء جدید با برقراری ارتباط

بین چند شیء شناخته شده به دست می‌آیند. بدین ترتیب می‌توان از گراف مفهومی برای کشف کلمات و گرامر زبان توسط سیستم بازشناسی گفتار استفاده کرد.

۶- ساختن یک سیستم *HMM* بسیار سریع بدون کاهش نتیجه

آزمایش‌های ما در کوانته کردن مقدار ویژگی‌ها به صد مقدار نشان دادند که این کوانته کردن تاثیر ناچیزی در دقت سیستم دارد. از طرف دیگر می‌توان نشان داد که این روش مدل‌سازی منجر به حذف مفاهیم مخلوط و تابع نرمال از HMM می‌شود. همچنین محاسبه لگاریتم که یکی از مهم‌ترین دلایل کند شدن بازشناسی است از روش محاسبه حذف می‌شود. همچنین بسیاری از محاسبات را می‌توان با اعداد صحیح انجام داد که باعث افزایش سرعت سیستم می‌شود. در مجموع می‌توان توقع داشت که سرعت بازشناسی متحول شود.

۷- بررسی ترکیب سیستم تشخیص و تولید صحبت برای یادگیری *CoEvolutive*.

واقعیت این است که انسان شنیدن و صحبت کردن را با هم یاد می‌گیرد[1]. به نظر می‌رسد که صحبت کردن نقش مهمی در یادگیری با معلم واج‌ها و کلمات دارد. بدین ترتیب به نظر می‌رسد که برای تولید یک سیستم تشخیص صحبت مشابه انسان باید فرآیند صحبت کردن را در کنار فرآیند تشخیص صحبت بررسی کرد.

[1] نگارنده تاکنون شخصی را که لال باشد ولی کر نباشد ندیده است.

فهرست منابع

منابع فارسی

[۱] موسوی بلده, سید محسن. حلیة القرآن سطح (۲), چاپ بیست و چهارم. تهران: انتشارات احیاء کتاب, ۱۳۸۰.

[۲] غیاثی شیرازی, سید کمال‌الدین. "جستجوی روش بهینه در دستیابی به نرم‌افزار تشخیص صحبت" پایان‌نامه کارشناسی، دانشگاه شهید بهشتی، شهریور ۱۳۸۱.

English References

[3] Allen J. B., How do humans process and recognize speech? *IEEE Trans on Speech and Audio Processing, pp. 567-577, Vol 2, No. 4, Oct 1994*.

[4] Arai T., Greenberg S., "Speech intelligibility in the presence of cross-channel spectral asynchrony". Proc. IEEE ICASSP, Seattle, pp. 933-936, 1998.

[5] Arai T., Pavel M., Hermanskey H. Avendano C., "Intelligibility of speech with filtered time trajectories of spectral envelopes". Int. Conf. Spoken Lang. Proc., Philadelphia, pp. 2490-2493, 1996.

[6] Avendano, C., and Hermansky, H. "*Study on the dereverberation of speech based on temporal envelope filtering"*. Proceedings of the International Conference on Speech and Language Processing 1996 (October 1996), 889—892, 1996.

[7] Aversano G., Esposito A., and Marinaro M., "A new text-independent method for phoneme segmentation" Proceedings of IEEE Midwest Symposium on Circuits and Systems, Dayton 14-17 August 2001.

[8] Cavicchi T.J., *Digital Signal Processing* : John Wiley & Sons, Inc, 2000.

[9] Chang, S. and Greenberg, S. "Syllable-proximity evaluation in automatic speech recognition using fuzzy measures and a fuzzy integral". Proceedings of the IEEE Fuzzy Systems Conference, St. Louis, 2003.

[10] Charpentier F. and Moulines E., Text to Speech algorithms based on FFT Synthesis, *Proc. ICASSP 88*, pp. 667-670, April 1988.

[11] Cheok A.D., Chevalier S., Kaynak M., Sengupta K., Chung K.C., "Use of a novel generalized fuzzy hidden markov model for speech recognition", Proceedings of the IEEE Fuzzy Systems Conference (IEEE FUZZ 2001), vol. 3, pp. 1207-1210, Melbourne, Australia, Dec. 2001.

[12] Cormen T. H., Leiserson C. E., Rivest R. L., Stein C., *Introduction to algorithms*. 2nd edition, MIT Press, 2001.

[13] Deller, J.R. , Proakis, J.G., Hansen, J.H.L. . *Discrete-Time Processing Of Speech Signals* : Macmillan, 1993.

[14] Dubois D., Prade H., *Possibility theory*. New York, London. 1988.

[15] Enderton H.B., *A Mathematical Introduction to Logic*. 2nd edition, Harcourt/ Academic Press, 2001.

[16] Fosler-Lussier, E., Greenberg, S. and Morgan, N., "Incorporating contextual phonetics into automatic speech recognition." *Proc. XIVth Int. Cong. Phon. Sci.*, 1999.

[17] Gold B., Morgan N., *Speech And Audio Signal Processing* : John Wiley & Sons, Inc ,1999.

[18] Greenberg, S., "Understanding speech understanding - towards a unified theory of speech perception". Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, Keele, England, p. 1-8, 1996.

[19] Greenberg S., Kingsbury B., "The modulation spectrogram: In persuit of an invariant representation of speech". ICASSP97 , pp. 1647-1650, 1997.

[20] Greenberg S., Arai T., Silipo R., "Speech intelligibility derived from exceedingly sparse spectral information". International Conference on Spoken Language Processing, Sydney, pp. 2803-2806, 1998.

[21] Greenberg S., Arai T., "The relation between speech intelligibility and the complex modulation spectrum". Proc. Eurospeech, pp. 473-476, 2001.

[22] Halmos P. R., Measure Theory, *Springer-Verlag*, 1950.

[23] Hermansky H., and Morgan N., RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2 (4), pp. 578-589, October 1984.

[24] Hermansky H., "Perceptual Linear Predictive (PLP) Analysis of Speech", *Journal of Acoust. Soc. Am.*, pp. 1738-1752, April 1990.

[25] Hermansky H., "Speech beyond 10 milliseconds (Temporal filtering in feature domain)". Invited keynote lecture, in Proceedings of the International Workshop on Human Interface Technology 1994, Aizu, Japan, Sept. 1994.

[26] Hermansky H., "Exploring temporal domain for robustness in speech recognition", in *The 15th International Congress on Acoustics*, vol. 2, pp. 61-64, Trondheim, Norway, Jun, 1995.

[27] Hermansky H. ,"Should recognizers have ears?". In Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 1-10, France 1997.

[28] Hermansky H., Sharma S., "TRAPS - Classifiers of Temporal Patterns", in ICSLP'98,

Sydney, Australia, 1998.

[29] Hermansky, H., and Sharma, S., "Temporal Patterns (TRAPS) in ASR of Noisy Speech," in Proc. ICASSP'99, Phoenix. March, 1999.

[30] Hopcroft J.E. and Ullman J.D.. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Publishing Company, Reading, MA, 1979.

[31] Huang X., Acero A., Horn H.W., *Spoken Language Processing*, Prentice Hall, 2000.

[32] Hu Z., Schalkwyk J., Barnard E., Cole R., "Speech recognition using syllable-like units". In Proceedings: International Conference on Spoken Language Processing (ICSLP), pp. 1117-1120, Philadelphia, USA, October 1996.

[33] Jain R., Kasturi R., Schunck B., *Machine vision*, McGraw-Hill, 1995.

[34] Jurafsky D., Martin J.H., *Speech and Language Processing,* Prentice Hall, 2000.

[35] Kanedera N., Arai T., Hermanskey H., Pavel M., "On the importance of various modulation frequencies for speech recognition", in proceedings EUROSPEECH97, Rhodes, Greece 1997.

[36] Larsen R.J., Marx M.L., *An Introduction to Mathematical Statistics and its Applications,* Prentice Hall, 2001.

[37] Liberman, A. M., & I. G. Mattingly. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1-36.

[38] Mao J.S., Chan S.C. and Ho, K.L. *A Mixed Excitation LPC Vocoder Operating at Very Low Bit Rate*, San Diego, CA, IEEE 6th Annual Intern. Conf. on Universal Personal Comm., 1997.

[39] McCree A.V., Barnwell T.P., A mixed excitation LPC Vocoder Model for Low Bit Rate Speech Coding. *IEEE Transactions on Speech and Audio Processing, Vol. 3., No. 4, July 1995.*

[40] Mitchell T. M., *Machine Learning*. McGraw-Hill, 1997.

[41] Mondragon A., Herrera A., "Speech recognition techniques using acoustic segmentation". IASTED International Conference on Signal and Image Processing, Orlando, Florida, November 11-14, 1996.

[42] Oppizzi O., Fournier D., Gilles P., Meloni H. "A fuzzy acoustic-phonetic decoder for speech recognition". Proc. ICSLP 1996.

[43] Rabiner L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE,* Vol. 77, No. 2, pp. 257-286, February 1989

[44] Rich E. and Knight K., *Artificial Intelligence*. McGraw-Hill, 1992.

[45] Russel S. and Norvig P., *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.

[46] Schmid P., Explicit N-Best Formant Features for Segment-Based Speech Recognition. PhD Thesis. *Oregon graduate Institute of Science and Technology*. 1996.

[47] Segura J. C., Benitez M. C., Torre A., Rubio A. J.. "Feature Extraction from Time-Frequency matrices for Robust Speech Recognition". *Euorospeech*, Scandinavia 2001.

[48] Shafer G., *A Mathematical Theory Of Evidence*, Princeton University Press, Princeton, NJ, 1976.

[49] Shastri L., Chang S. and Greenberg S., "Syllable Detection and Segmentation Using Temporal Flow Neural Networks", *Int. Cong. of Phonetic Sciences*, San Francisco, 3:1721-1724, August 1999.

[50] Silipo R., Greenberg S., Arai T., "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations". Proc. Eurospeech, pp. 2687-2690, 1999.

[51] Sowa J.F., *Information Processing in Mind and Machine Reading*, MA: Addison-Wesley Publ., 1984.

[52] Su M.C., Hsieh C.T., Chin C.C., *A neuro-fuzzy approach to speech recognition without time allignment.* Fuzzy Sets and Systems, pp. 33-41, 1998.

[53] Supplee L.M., Cohn R.P., Collura J.S., McCree A.V., "MELP: The New Federal Standard at 2400 bps," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997.

[54] Tibrewala S., Hermansky H., "Sub-Band based recognition of noisy speech". ICASSP'97, vol. 2, pp. 1255-1258, IEEE, Munich, Germany, 1997.

[55] Tran D., Wagner M., Zheng T., *State mixture modeling applied to speech recognition*. Pattern Recognition Letters 20, pp. 1449-1356, 1999.

[56] Yang H., Vuuren S.V., Sharma S. and Hermansky H, "Relevancy Of Time Frequency Features for Phonetic Classification Measured by Mutual Information". *Proc. ICASSP 99*, pp. 225-228. Phoenix, Arizona, USA, March 1999.

.

[57] Yang H., Vuuren S.V., Sharma S. and Hermansky H., Relevance of Time-Frequence Features for Phonetic and Speaker-Channel Classification. *Speech Communication*, Vol. 31(1) pp 35-50, 2001.

[58] Yeldener S., De Martin J.C., Viswanathan V., "A Mixed Sinusoidally Excited Linear Prediction Coder at 4 kb/s and Below" , Proceedings of IEEE ICASSP, Seattle, Washington, Vol. 2, pp. 589-592. , May 1998.

[59] Yoneyama K., "Segmentation Strategies For Spoken Language Recognition: Evidence From Semi-Bilingual Japanese Speakers Of English". Proc. ICSLP, 1996.

[60] Young S., Odell J., Ollason D., Valtchev V., Woodland P., *The HTK Book*, Cambridge University, 1997.

[61] YU H.J., OH Y.H., Fuzzy Expert System for Continuous Speech Recognition. Expert Systems With Applications, Vol. 9. No. 1, pp. 81-89, 1995.

[62] Wang K. and Shamma S., Self-Normalization and Noise Robustness in Early auditory processing, *IEEE Trans. Aud. and Speech, 2(3), 421-435, 1994*.

.

[63] Wang T., Koishida K., Cuperman V., "A 1200 BPS Speech coder based on MELP", Proceedings of IEEE ICASSP, Istanbul, Turkey, Vol. 3, pp. 1375-1378. , 2000.

[64] Zadeh L. A., Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems, 1(1)*, 3-28. 1978.

[65] Zhan P. and Waibel A., "*Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition"*. Technical Report CMU-CS-97-148, School of Computer Science, Carnegie Mellon University, 1997.

[66] Zhan P., "*Speaker Normalization and Speaker Adaptation - a Combination for Conversational Speech Recognition*". Proceedings of Eurospeech Conference, Greece, 1997.

[67] Ziemer R.E., Transter W.H., Fannin , D. Ronald. *Signals And Systems: Continuous And Discrete,* 2nd edition, Macmillan , 1989.

[68] Zimmermann H.J., *Fuzzy Set Theory And Its Applications,* 3rd edition, Kluwer Academic Publishers, 1996.

Abstract

In this dissertation, human method of speech recognition was studied. To study the human method in speech recognition, a tool was made allowing us to manipulate speech signal and hear the changed one. We reached to the conclusion that humans pay more attention to more energetic features. Based on this and other results, a segmentation and feature extraction method called OBSFE was developed which is not sensible to the subtle changes in energies of features. OBSFE has several benefits over MFCC. First of all, it it not model based, i.e. it does not use linguistic information in segmentation. This property is vital for a system aiming to learn the language like humans. The second benefit is that it extracts features in time-frequency space and not only in frequency space. Works of other researchers show that this method is similar to human method of feature extraction and more robust to noise. Our experiments also assert the robustness of our method to noise. OBSFE has improved word error rate at a rate of 21.44% in babble, car, restaurant, airport, and train station and in 5, 15, and 20 decibels. The third benefit of OBSFE is that segments have overlap. We are aware of no other segmentation algorithm generating overlapping segments. Experiments have shown that this method works well in word recognition but is not yet suitable for phoneme recognition. In addition, possibility theory was suggested as a substitution for probability theory and it was shown that theoretically this theory has the ability of learning new objects. A simple implementation of a possibilistic recognizer showed that possibilistic recognizers work robustly in noisy environments. In addition, a new possibilistic measurement based on quantizing features to 100 values according to percentiles was proposed and implemented. It was shown that this measure does not deteriorate the results of word and phoneme recognition. Finally, we could achieve to an accuracy of 49.39% and a correctness of 64.5% in phoneme recognition in Tehrani selection of Farsdat database.

Keywords:

1- Speech recognition
2- Fuzzy logic
3- Possibility theory
4- Human
5- Feature extraction

Sharif University of Technology
Faculty of Computer Engineering

M.S.C. Thesis

Studying and simulating human abilities in speech recognition

Sayed Kamal-Aldin Ghiathi Shirazi

Supervisor:
Dr. Saeed Bagheri Shouraki

Winter 2004