
بسم الله الرحمن الرحيم

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی برق
گروه مخابرات سیستم

پایان نامه کارشناسی ارشد

بهبود عملکرد مدل ماشین‌های بردار پشتیبان در دیکدر **ATP** گفتار پیوسته فارسی

دانشجو : ملیحه قیدی

استاد راهنما: آقای دکتر صیادیان

بهمن ۸۵

بسمه تعالي

تاريخ : .....
شماره : .....

فرم اطلاعات پايان نامه
كارشناسي ارشد و دكترا

دانشگاه صنعتي اميركبير
(پلي تكنيک تهران)

معاونت پژوهشي
فرم پروژه تحصيلات تكميلي 7

مشخصات دانشجو

نام و نام خانوادگي : مليحه قيدي | √ دانشجوي آزاد | بورسيه | معادل

شماره دانشجويي: 83123119 | دانشكده : برق | رشته تحصيلي: مخابرات سيستم

نام و نام خانوادگي استاد راهنما : ابوالقاسم صيادیان

عنوان به فارسي: بهبود مدل ماشين هاي بردار پشتيبان در ديكدرATP گفتار پيوسته فارسي

عنوان به انگليسي: Improvement in Support Vector Machines Model in Persian Continuous Speech ATP Decoder

نوع پروژه: كارشناسي ارشد √ دكترا | **كاربردي √** | **بنيادي** | **توسعه اي √** | **نظري**

تاريخ شروع: 1384/4/27 | **تاريخ خاتمه :1385/11/15** | **تعداد واحد: 6 واحد**

سازمان تامين كننده اعتبار : مركز تحقيقات مخابرات

واژه هاي كليدي به فارسي : ماشين های بردار پشتيبان، نيم هجا، گفتار پيوسته،واكه

واژه هاي كليدي به انگليسي : Demi-syllable, coetaneous speech, support vector machine, vowel

---

نظرها و پيشنهادها به منظور بهبود فعاليت هاي پژوهشي دانشگاه:

استاد راهنما:

دانشجو:

امضاء استاد راهنما : **تاريخ:**

نسخه 1: معاونت پژوهشي
نسخه 2 : كتابخانه و به انضمام دو جلد پايان نامه به منظور تسويه حساب با كتابخانه و مركز اسنادومدارک علمي

چکیده

درسیستم‌های بازشناسی گفتار انتخاب واحد گفتاری مناسب، از اهمیت ویژه ای برخوردار است. جهت انتخاب واحد آکوستیکی مناسب، در نظر گرفتن ساختار و ویژگی‌های زبان مورد استفاده، بسیار مهم است. با توجه به این که ساختار هجایی زبان فارسی تقریبا همانند زبان‌های هندی، چینی و ژاپنی ساده و محسوس است، در این تحقیق، واحد زیرکلمه نیم هجا به عنوان واحد آکوستیکی مناسب در زبان فارسی مورد توجه ما قرار گرفته است. با توجه به اینکه پایگاه داده گفتاری مبتنی بر نیم هجاها در زبان فارسی موجود نمی‌باشد، تلاش‌های زیادی جهت طراحی و پیاده سازی پایگاه داده گفتاری مناسب مبتنی بر نیم هجاها در طی انجام این تحقیق، صورت گرفته است. برای ارزیابی مدل‌ها، داده‌های گفتاری مربوط به دو گوینده زن و دو گوینده مرد ضبط شده و به صورت با سرپرستی در سطح واکه و نیم هجا برچسب زده شده است.

در این پایان نامه، به عنوان اولین قدم جهت تشخیص واحدهای نیم هجا در سیگنال گفتار، به آشکارسازی و بازشناسی واکه‌ها پرداخته شده است. سعی شده با ترکیب روش ماشین‌های بردار پشتیبان و روش استفاده از ویژگی‌های آکوستیکی ، کارایی سیستم در این بخش تا حد ممکن بهبود داده شود. از پارامترهای آکوستیکی نظیر انرژی میان گذر به دلیل ویژگی‌های مناسبی چون سادگی محاسبات و ناوابسته بودن به گوینده، به منظور تشخیص اولیه محل واکه‌ها استفاده شده است. سپس از قدرت متمایز سازی خوب ماشین‌های بردار پشتیبان جهت طبقه بندی واکه‌ها و تعیین مرز دقیق تر آنها، بهره مند شدیم و به نتایج بسیار مناسبی دست یافتیم. در این تحقیق، برای دادگان گفتار گسسته در صد خطای کل ۱/۶۸٪ و برای گفتار پیوسته در صد خطای کل ۴/۸۶٪ حاصل شد. در مقایسه با نتایج حاصل از دو روش دیگر یعنی مدل مارکوف پنهان و مدل قطعه بندی نرم بر روی همین پایگاه داده، در می‌یابیم که ماشین‌های بردار پشتیبان در کاربردهای طبقه بندی بسیار کارآمد هستند. البته دست یابی به دقت بالا با استفاده از این روش، مستلزم صرف هزینه محاسباتی بیشتر و زمان آموزش طولانی تر خواهد بود.

فهرست مطالب

فصل اول – مقدمه

مقدمه ۱
۱-۱-دو مبحث مهم در طراحی سیستم‌های بازشناسی گفتار ۱
۱-۲-کارهای انجام شده در زمینه تقطیع و برچسب زنی اتوماتیک گفتار ۷
۱-۲-۱-استفاده از مدل مارکوف پنهان ۸
۱-۲-۲-استفاده از شبکه‌های عصبی ۹
۱-۲-۳-استفاده از ماشین‌های بردار پشتیبان ۱۰
۱-۲-۴-مدل قطعه بندی نرم ۱۰
۱-۲-۵-استفاده از ویژگی‌های آکوستیکی ۱۱
۱-۲-۶-روش‌های ترکیبی ۱۱
۱-۳-هدف و ساختار پایان نامه ۱۲

فصل دوم- ماشین‌های بردار پشتیبان

مقدمه ۱۵
۲-۱- ماشین های بردار پشتیبان ۱۵
۲-۱-۱- اصل مینیمم سازی ریسک ساختاری ۱۶
۲-۱-۲- ماشین های بردار پشتیبان خطی ۲۰
۲-۱-۲-۱- حالت جدایی پذیر خطی ۲۰
۲-۱-۲-۲- حالت جدایی ناپذیر ۲۴
۲-۱-۳- ماشین های بردار پشتیبان غیر خطی ۲۵
۲-۱-۳-۱- شرایط Mercer ۲۷
۲-۱-۴- ماشین های بردار پشتیبان در حالت چند کلاسه ۲۸
۲-۱-۴-۱- روش یکی در مقابل همه ۲۹
۲-۱-۴-۲- روش یکی در مقابل یکی ۲۹
۲-۲-مقایسه با شبکه های عصبی و سایر طبقه بندی کننده های غیر خطی ۳۰
۲-۳- مقایسه عملکرد SVM با طبقه بندی کننده RBF کلاسیک ۳۱
۲-۴- نکات عملی پیاده سازی SVM ها ۳۳
خلاصه فصل ۳۷

فصل سوم- انتخاب واحد گفتاری

مقدمه ۳۸
۳-۱- واحدهای گفتاری ۳۸
۳-۱-۱- واج ۳۸

۳-۱-۱-۱- واکه ۳۹
۳-۱-۱-۲- همخوان ها ۳۹
۳-۱-۲- هجا ۴۰
۳-۱-۳- نیم هجا ۴۲
۳-۲- دادگان مناسب برای بازشناسی گفتار پیوسته فارسی ۴۳
۳-۲-۱- ملاحظات طراحی متن دادگان ۴۳
۳-۲-۱-۱- کلمه‌های با ساختار CVC ۴۳
۳-۲-۱-۲- کلمه‌های با ساختار CVCC ۴۴
۳-۲-۱-۳- کلمه‌های با ساختار CV ۴۴
۳-۲-۱-۴- کلمه‌های با ساختار CV تحت متن ۴۵
۳-۲-۱-۵-تعداد کل کلمه‌های طراحی شده ۴۵
خلاصه فصل ۴۶

فصل چهارم- نتایج پیاده سازی روش پیشنهادی

مقدمه ۴۷
۴-۱-انرژی ۴۷
۴-۱-۱- ملایم سازی منحنی انرژی ۴۸
۴-۱-۲- انرژی میان گذر ۴۸
۴-۲-نتایج پیاده سازی روش پیشنهادی بر روی دادگان گفتار گسسته ۵۱
۴-۲-۱- مشخصات دادگان مورد استفاده ۵۱
۴-۲-۲- استخراج ویژگی ۵۱
۴-۲-۳- الگوریتم استفاده از انرژی میان گذر ۵۱
۴-۲-۴- ترکیب مدل SVM و روش استفاده از ویژگیهای آکوستیکی ۵۵
۴-۲-۴-۱- مرحله آموزش ۵۶
۴-۲-۴-۱-۱-نرمالیزه کردن داده ها ۵۶
۴-۲-۴-۱-۲- انتخاب نوع طبقه بندی کننده چند کلاسه ۵۶
۴-۲-۴-۱-۳- انتخاب نوع تابع هسته ۵۷
۴-۲-۴-۱-۴-انتخاب پارامترهای تابع هسته و پارامتر C ۵۷
۴-۲-۴-۱-۵-نحوه انتخاب داده های آموزش ۵۸
۴-۲-۴-۲- مرحله تست ۵۹
۴-۲-۴-۲-۱-نحوه ارزیابی خروجی SVM ۵۹
۴-۲-۴-۲-۲-نحوه انتخاب مجموعه دادگان تست ۶۰
۴-۲-۴-۲- بهبود نتایج حاصل از انرژی میان گذر با استفاده از SVM ها ۶۰
۴-۲-۵- مقایسه نتایج روش SVM با روش های HMM و قطعه بندی نرم در گفتار گسسته ۶۳
۴-۲-۵-۱-تحلیل نتایج مدل ترکیبی ۶۳

۴-۲-۶- تعیین مرز واکه ها تنها با استفاده از مدل SVM ۶۴
۴-۳-نتایج پیاده سازی روش پیشنهادی در دادگان گفتار پیوسته فارسی ۶۵
۴-۳-۱- مشخصات دادگان مورد استفاده ۶۵
۴-۳-۲- استخراج ویژگی ۶۶
۴-۳-۳- الگوریتم مورد استفاده ۶۶
۴-۳-۵-ترکیب مدل SVM و روش استفاده از اطلاعات آکوستیکی ۶۹
۴-۳-۵-۱- بهبود نتایج حاصل از انرژی میان گذر با استفاده از SVM ها ۷۰
۴-۳-۵-۲- مقایسه نتایج روش SVM با روش های HMM و قطعه بندی نرم در گفتار پیوسته ۷۱
۴-۳-۵-۳-تحلیل نتایج مدل ترکیبی ۷۱

نتیجه گیری و پیشنهادات ۷۳

مراجع ۷۶

فهرست شکل‌ها

شکل ۱-۱- بلوک دیاگرام کلی سیستم بازشناسی گفتار بر اساس واحدهای آکوستیکی..... ۳
شکل ۱-۲ - مثالی از مسئله تفکیک دو کلاس توسط تخمین ML که سطح آستانه تصمیم گیری ناشی از آن ایده آل نیست و در ناحیه خاکستری رنگ احتمال خطا وجود دارد. ۶
شکل ۱-۳- ساختار سلسله مراتبی HMM ها برای بازنمایی واج ها، کلمات و جملات..... ۹
شکل ۱-۴- بلوک دیاگرام کلی سیستم ترکیبی تشخیص واکه ۱۲
شکل ۲-۱- تفاوت طبقه بندی کننده حاصل از ریسک تجربی و ریسک ساختاری، هر سه ابرصفحه C_0, C_1, C_2 ریسک تجربی را حداقل میکنند، اما تنها ابرصفحه C_0 بهینه است و ریسک ساختاری را حداقل میکند. ۱۸
شکل۲-۲- اصل مینیمم سازی ریسک ساختاری در واقع تلاش برای پیدا کردن نقطه بهینه در منحنی باند ریسک واقعی میباشد..... ۱۸
شکل ۲-۳- حداکثر سه نقطه در فضای دو بعدی توسط خطوط جهت دار مستقیم، شکسته میشود..... ۱۹
شکل ۲-۴- ابر صفحه جدا کننده دو کلاس در حالت داده های جدایی پذیر خطی، بردارهای پشتیبان با حلقه ای دور نقاط مشخص شده اند..... ۲۰
شکل ۲-۵ – ابر صفحه جدا کننده در حالت داده های جدایی ناپذیر، بردارهای پشتیبان با حلقه ای دور نقاط مشخص شده اند..... ۲۵
شکل ۲-۶ – ماشین های بردار پشتیبان در حالت غیر خطی، قدرت SVM ها در این است که میتوانند داده ها را به فضای با دیمانسیون بالا منتقل کنند و در آن فضا یک طبقه بندی کننده خطی بسازند..... ۲۷
شکل ۲-۷ – بردارهای پشتیبان(مراکز خوشه ها) که به طور خوکار توسط SVM ها بدست آمده، با دایره اضافی روی نمونه ها مشخص شده است. مراکز خوشه ها که با استفاده از روش RBF کلاسیک حاصل شده، با ضربدر مشخص شده است. ۳۲
شکل ۲-۸- اثر استفاده از مجموعه ارزیابی بر عملکرد SVM..... ۳۵
شکل ۲-۹- هیستوگرام خروجی SVM برای نمونه های مثبت و منفی با استفاده از روش cross-validation..... ۳۶
شکل ۲-۱۰- یک نمونه از تخمین با استفاده از تابع توزیع سیگموئیدی برای یک طبقه بندی کننده خاص ۳۶
شکل ۴-۱- منحنی انرژی برای جمله " دستی از غیب او را یاری داد. "، این منحنی با پنجره همینگ به طول ۱۳ هموار شده است. ۵۰
شکل ۴-۲- منحنی انرژی میان گذر برای جمله " دستی از غیب او را یاری داد. "، این منحنی با پنجره همینگ به طول ۱۳ هموار شده است..... ۵۰
شکل ۴-۳- پنجره اعمال شده برای محاسبه انرژی میان گذر... .....۵۲
شکل ۴-۴- منحنی انرژی میان گذر برای کلمه "جت"، به همراه مرز واکه حاصل از تقطیع دستی و مرز واکه حاصل از انرژی میان گذر..... ۵۵
شکل ۴-۵- منحنی انرژی میان گذر برای کلمه "نیم"، به همراه مرز واکه حاصل از تقطیع دستی و مرز واکه حاصل از انرژی میان گذر، در این حالت استفاده از انرژی میان گذر منجر به خطای درج شده است..... ۶۲
شکل ۴-۶- بهبود حاصل از اعمال مدل SVM برای کلمه "نیم"، خطای درج حاصل از روش انرژی میان گذر از بین رفته است. ۶۲

شکل ۴-۷- منحنی انرژی میان گذر برای جمله " دستی از غیب او را یاری داد "، به همراه مرز واکه حاصل از تقطیع دستی و مرز واکه حاصل از انرژی میان گذر..... ۶۹
شکل ۴-۸- بلوک دیاگرام کلی روش پیشنهادی..... ۷۲

فهرست جداول

جدول ۳-۱- واکه های زبان فارسی ۳۹
جدول ۴-۱- نتایج حاصل از پیاده سازی روش استفاده از انرژی میان گذر بر روی دادگان گفتار گسسته ۵۳
جدول ۴-۲- نتایج پیاده سازی مدل ترکیبی در گفتار گسسته بر مجموعه آموزش ۶۱
جدول۴-۳- نتایج پیاده سازی مدل ترکیبی در گفتار گسسته بر مجموعه آزمون..... ۶۱
جدول۴-۴- مقایسه نتایج روش SVM با روش های HMM و قطعه بندی نرم در گفتار گسسته بر مجموعه آزمون ۶۳
جدول ۴-۵- نتایج پیاده سازی مدل SVM در گفتار گسسته بر مجموعه آموزش ۶۵
جدول ۴-۶- نتایج پیاده سازی مدل SVM در گفتار گسسته بر مجموعه آزمون.....۶۵
جدول ۴-۷- نتایج حاصل از پیاده سازی روش استفاده از ویژگیهای آکوستیکی بر روی دادگان گفتارپیوسته ۶۸
جدول ۴-۸- نتایج پیاده سازی مدل ترکیبی در گفتار پیوسته بر مجموعه آموزش.....۷۰
جدول ۴-۹- نتایج پیاده سازی مدل ترکیبی در گفتار پیوسته بر مجموعه آزمون ۷۰
جدول۴-۱۰- مقایسه نتایج روش SVM با روش های HMM و قطعه بندی نرم در گفتار پیوسته بر مجموعه آزمون۷۱

فصل اول

مقدمه

مقدمه

سیستم‌های بازشناسی گفتار پیوسته ، یکی از پرکاربرد ترین سیستم‌های کاربردی جهت ارتباط طبیعی انسان و ماشین می‌باشند. هدف اصلی این سیستم‌ها، تبدیل سیگنال صحبت به متن می‌باشد. اغلب سیستم های بازشناسی گفتار، شامل دو زیر مرحله کلی می‌باشند: ۱) تبدیل سیگنال صحبت به سمبل یا در واقع تبدیل آکوستیک به فونتیک[1] ۲) تبدیل سمبل به متن. در مرحله اول سیگنال صحبت به قطعات کوچکتر (واحدهای آکوستیکی) تقسیم می‌شود و به هر یک از این واحدهای آکوستیکی، یک سمبل یا فونتیک اختصاص داده می‌شود. در این پایان نامه هدف، ایجاد بهبود در این مرحله از بازشناسی گفتار می‌باشد. در مرحله دوم، این رشته سمبل‌ها با استفاده از فرهنگ لغات و در نظر گرفتن قواعد دستوری و معنایی زبان، به متن تبدیل می‌شوند.

۱-۱- دو مبحث مهم در طراحی سیستم‌های بازشناسی گفتار

هر سیستم بازشناسی گفتار پیوسته مبتنی بر واحدهای آکوستیکی، از قسمت‌های اساسی نشان داده شده در شکل ۱-۱ تشکیل شده است. در بلوک اول، ویژگی‌های مناسب برای توصیف خواص سیگنال گفتار استخراج می‌شود. در مرحله بعد واحد گفتاری مناسب برای سیستم انتخاب می‌شود، سپس مدل آکوستیکی- آماری مناسب برای بازشناسی واحدهای آکوستیکی، آموزش داده می‌شود. در بلوک سوم با استفاده از این مدل‌ها، واحدهای آکوستیکی بازشناسی می‌شوند. پس از بازشناسی واحدهای آکوستیکی، با استفاده از فرهنگ لغات بر مبنای تلفظ-های واقعی، تطبیق در سطح کلمه انجام می‌شود. در نهایت با در نظر گرفتن مدل‌های دستوری و معنائی زبان، جمله متناسب با سیگنال گفتار ورودی، بازشناسی می‌شود.

در این پروژه صرفا" بهبود عملکرد بلوک‌های (۲) و (۳) مورد توجه و تحقیق می‌باشد. برای انجام وظایف بلوک شماره (۱) از الگوریتم‌های آماده استفاده می شود.

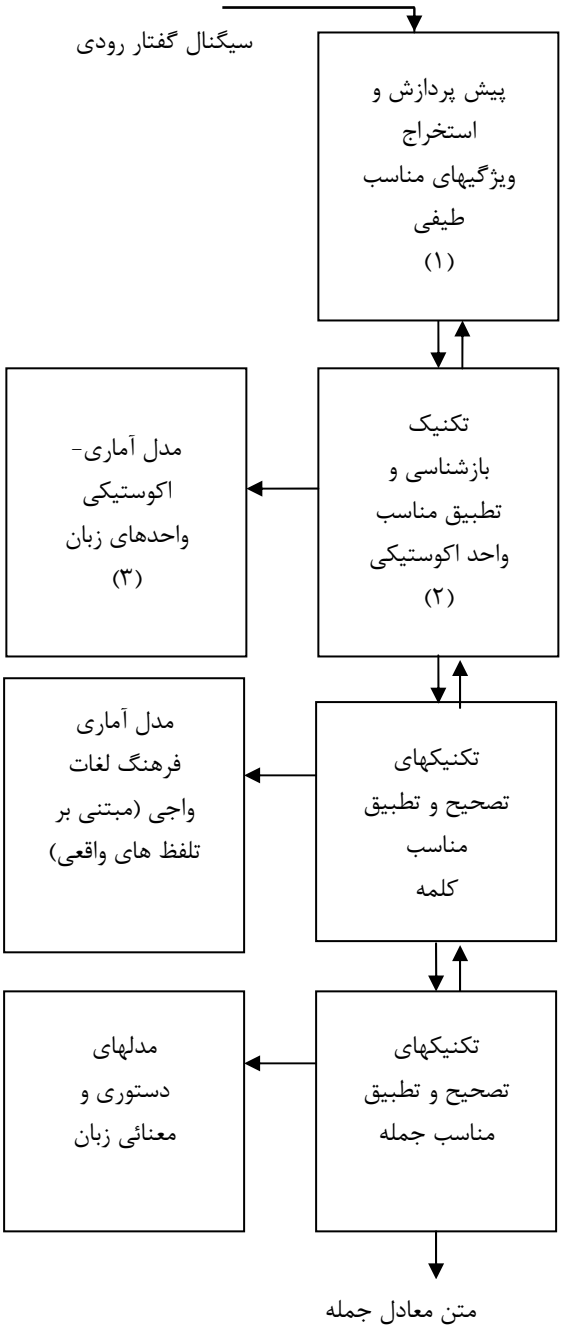
[1] Acoustic to Phonetic (ATP)

درسیستم‌های بازشناسی گفتار، انتخاب واحد آکوستیکی در بلوک (۲) ، از اهمیت ویژه ای بر خوردار است. برای سیستم‌های بازشناسی گفتار گسسته با تعداد لغات کم، معمولاً از واحد کلمه استفاده می‌شود. با افزایش دایره لغات، استفاده از کلمه به عنوان واحد آکوستیکی، سبب ایجاد محدودیت‌هایی در آموزش، ذخیره و جستجوی واحد کلمه می‌شود. برای حل این مشکل به واحدهای زیر کلمه مانند واج[۱]، هجا[۲] و نیم هجا[۳] روی آورده شده است [1],[2],[3],[4]. در فصل سوم راجع به این واحدهای آکوستیکی بیشتر توضیح داده خواهد شد. واحد آکوستیکی مرسوم در بلوک (۲) معمولا" واج می‌باشد. واج‌ها تحت متن‌های مختلف از نظر مشخصه آکوستیکی به شدت تغییر می کنند و در واقع باید اثرات واج قبل و بعد از آنها نیز لحاظ شود. نشان داده شده است که مدلسازی وابسته به متن[۴]، بازدهی بازشناسی بسیار بهتری دارد [5],[6],[7],[8]. به این ترتیب روش‌های دو آوایی[۵] و سه آوایی[۶] مطرح شدند و تحقیقات بسیاری در این زمینه انجام شد [9],[10],[11]. اما یکی از مشکلات این روش‌ها این است که تعداد حالت‌های آنها زیاد است. بنابراین حجم جستجو و پردازش بالا می‌رود.

هدف ما مدلسازی مناسب واج‌ها تحت متن می باشد. یکی دیگر از مدل‌های وابسته به متن، استفاده از واحد آکوستیکی هجا می‌باشد. با توجه به اینکه ساختار هجایی زبان فارسی همانند زبان های هندی، چینی و ژاپنی نسبتاً ساده و محسوس است، استفاده از هجا به عنوان واحد زیر کلمه برای این نوع زبان ها مناسب می‌باشد. اما یکی از مشکلات استفاده از این واحد آکوستیکی نیز تعداد حالت‌های زیاد آن می‌باشد. یک روش برای حل این مشکل، می‌تواند استفاده از واحدهای نیم هجا باشد. بنابراین با توجه به پوشش مناسب و تعداد حالت‌های کم نیم هجا، استفاده از این واحد آکوستیکی کارآیی سیستم‌های بازشناسی گفتار پیوسته را بالا می‌برد. لذا طی دو دهه اخیر تحقیقاتی در زمینه بازشناسی گفتار مبتنی بر نیم هجاها درزبانهای دیگر انجام گرفته است [6] ,[7], [12] ,[13],[14]. در زبان فارسی در این قسمت، تحقیق و توجه کمتری شده است. این امر انگیزه ای برای انجام تحقیق در زمینه بازشناسی گفتار مبتنی بر نیم هجاها در زبان فارسی شد.

---

[1] Phoneme
[2] Syllable
[3] Demi-syllable
[4] Context-dependent
[5] Diphone
[6] Triphone



شکل ۱-۱- بلوک دیاگرام کلی سیستم بازشناسی گفتار بر اساس واحدهای آکوستیکی

با توجه به این که واکه‌ها[1] از نظر مشخصه فرکانسی و زمانی نسبت به سایر واج ها پایدارتر هستند، در بازشناسی گفتار بسیار پراهمیت هستند. ثابت شده است که در یک سیگنال گفتار واکه ها برای بازشناسی خودکار

[1] Vowels

گفتار کاملا مناسب می‌باشند [1] ,[15], [16], زیرا معمولا به اندازه کافی طولانی بوده و انرژی کافی برای تشخیص در سیگنال صوتی را دارند. یکی دیگر از ویژگی واکه‌ها پریودیک بودن آنها می‌باشد. با استفاده از این ویژگی‌ها می‌توان با محاسبات نسبتا ساده به محل تقریبی واکه‌ها در سیگنال گفتار دست یافت. بررسی‌ها نشان می‌دهد که این ویژگی واکه‌ها تقریبا مستقل از گوینده و نیز مقاوم در برابر شرایط محیطی (از جمله نویز) می-باشد. لذا با استفاده از آنها می‌توان نقاط قابل اعتمادی در سیگنال گفتار پیوسته پیدا کرد.

در این پایان نامه، نیم هجاها و واکه‌ها به عنوان بهترین واحد زیر کلمه برای گفتار پیوسته فارسی انتخاب شده است. به این ترتیب که برای هر واکه دو بخش ایستان و گذرا در نظر گرفته می‌شود. بخش ایستان واکه تقریباً نشان دهنده مشخصات خود واکه می‌باشد و اثر آواهای قبل و بعد از آن نسبتاً ناچیز است. بنابراین ابتدا قسمت ایستان واکه ها که تشخیص آنها ساده تر و مطمئن تر است، در سیگنال گفتار شناسایی می‌شوند، سپس همخوان[1] (یا همخوان ها) و قسمت گذرای واکه به عنوان نیم هجا مدل سازی می‌شود. به این ترتیب اثرات متقابل واکه ها و همخوان ها در نظر گرفته می‌شود. البته در این پروژه فقط قسمت اول یعنی آشکارسازی محل واکه ها و طبقه بندی آنها انجام شده است. سعی شده با استفاده از ترکیب روش استفاده از ویژگی‌های آکوستیکی و روش آماری، کارایی سیستم در این بخش، تا حد ممکن بهبود داده شود. انجام قسمت دوم جزء کارهای آتی می‌باشد.

یکی دیگر از مباحث مهم در طراحی سیستم‌های بازشناسی گفتار، نحوه مدلسازی آکوستیکی-آماری سیگنال گفتار در بلوک (۳) می‌باشد. مدل های مارکوف پنهان[2] با چگالی مشاهدات مخلوط گوسی[3] روش غالب در سیستم های بازشناسی گفتار می‌باشد. این سیستم‌ها معمولا از قدرت بازنمایی[4] خوب برای مدل سازی استفاده می‌کنند[17]. یکی از علل مهم موفقیت HMM ها در بازشناسی گفتار، توانایی آنها در مدل کردن حالت های زمانی سیگنال گفتار، توسط فرایند مارکوف می‌باشد. تابع توزیع احتمالی که برای هر حالت در مدل HMM در نظر گرفته می شود، تغییرات سیگنال گفتار ناشی از گوینده های مختلف و یا ناشی از متون مختلف آوایی را مدل می‌کند. این تابع توزیع معمولا مخلوطی از توابع گوسی در نظر گرفته می‌شود.

---

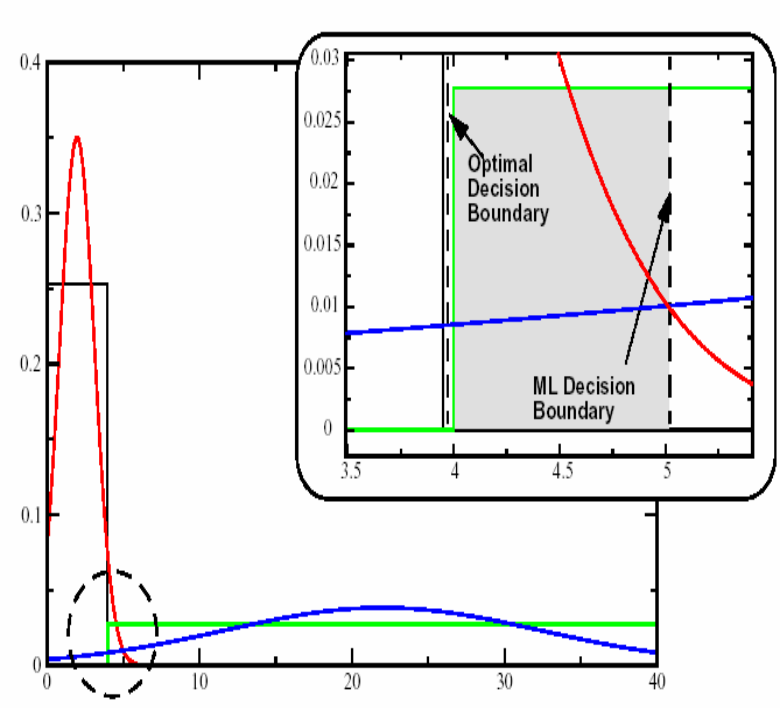
[1] Consonant
[2] Hidden Markov Model (HMM)
[3] Gaussian Mixture Model (GMM)
[4] Representation

برای تخمین پارامترهای HMM ، تخمین بر اساس ماکزیمم کردن درستنمایی[1]، یکی از مؤثرترین روش-هاست. به هر حال مشکلاتی در رابطه با تخمین ML برای کاربردهایی همچون بازشناسی گفتار وجود دارد. تخمین ML سعی در بازنمایی بهتر یک کلاس دارد ولی از نظر ایجاد تمایز میان دو کلاس ضعیف می‌باشد، در حالی که قدرت تمایز بالا برای سیستم های بازشناسی گفتار، یک نیاز اساسی می‌باشد[17] ,[18] ,[19] ,[20]. در تخمین ML پارامترهای مدل تنها براساس داده های مربوط به کلاس مورد نظر[2] و بدون در نظر گرفتن داده های خارج از کلاس مورد نظر[3]، تخمین زده می‌شوند [18]. شکل ۱-۲ یک مثال ساده از این مشکلات را نشان می‌دهد. دو کلاس با تابع توزیع یکنواخت به صورت کاملا جدا از هم نشان داده شده است. از تخمین ML برای بازنمایی این دو کلاس به صورت گوسی و از قاعده بیز[4] برای طبقه بندی داده ها استفاده شده است. همان طور که مشاهده می‌شود سطح آستانه تصمیم گیری که از تخمین ML به دست آمده است، با سطح تصمیم گیری بهینه فاصله دارد و در این محدوده احتمال خطای قابل ملاحظه ای وجود دارد. این مساله برای تمامی داده هایی همانند سیگنال گفتار که در فضای ویژگی ها همپوشانی دارند و مرز کلاس ها در همسایگی یکدیگر قرار دارد ، وجود دارد [18].

در این مثال دیدیم که آموزش ML یک مدل گوسی، هیچگاه به طبقه بندی کامل دست نمی‌یابد. اما نکته مهم این است که مدل های گوسی لزوما انتخاب نامناسبی نیستند، بلکه این مطلب نشان دهنده نیاز به روش های با قدرت تمایز[5] بالا برای ایجاد مدل های مقاوم تر و دقیق تر می‌باشد[18].

---

[1] Maximum Likelihood (ML)
[2] In-class data
[3] Out-of-class data
[4] Bayes' rules
[5] Discriminative



شکل ۱-۲ - مثالی از مسئله تفکیک دو کلاس توسط تخمین ML که سطح آستانه تصمیم گیری ناشی از آن ایده آل نیست و در ناحیه خاکستری رنگ احتمال خطا وجود دارد[18].

یکی دیگر از موارد مطرح دربحث بازشناسی ، قدرت تعمیم دهی[1] سیستم است. عملکرد سیستم‌های HMM در یک سیستم حلقه بسته[2] بسیار خوب است اما در یک سیستم حلقه باز[3] کارایی آن به طور قابل ملاحظه ای کاهش می‌یابد. منظور از یک سیستم حلقه بسته این است که شرایط آموزش[4] و آزمون[5] یکسان باشد و همین طور، منظور از یک سیستم حلقه باز این است که شرایط آموزش و آزمون متفاوت باشد. همچنین سیستم های HMM در یک سیستم وابسته به گوینده به مراتب بهتر از سیستم ناوابسته به گوینده عمل می‌کنند [17].

[1] Generalization
[2] Closed-loop
[3] Open-loop
[4] Training
[5] Test

دراین پایان نامه یک روش نسبتا جدید بر اساس اصل مینمم سازی ریسک ساختاری[1] بررسی می-شود که از یک قالب با قدرت تمایز بالا بر اساس ماشین‌های بردار پشتیبان[2] برای بازشناسی گفتار استفاده می‌کند. ماشین‌های بردار پشتیبان توانایی همزمان قدرت تعمیم دهی و قدرت تمایز مدل‌های آکوستیکی را دارند [17] ,[18] ,[19] ,[20].

ماشین‌های بردار پشتیبان در چند سال گذشته در بسیاری از کاربردهای طبقه بندی موفقیت‌های چشمگیری داشته اند [18]. مجموعه ای که باعث شد، ماشین های بردار پشتیبان در اوایل دهه ۹۰ بسیار برجسته شوند، داده های رقمی سرویس پستی آمریکا بود که SVM ها بهترین ارقام را گزارش دادند [21]. ابتدا تلاش‌های زیادی برای به کارگیری SVM ها برای بازشناسی گوینده شد [22] ,[23]. اما این تلاش‌ها به دلیل کمبود روش‌های پیاده سازی موثر برای تخمین SVM ها در آن زمان، موفقیت‌های محدودی داشتند.

بعدها با گسترش روش های بهینه سازی، ماشین‌های بردار پشتیبان برای طبقه بندی های وسیع تر مثل طبقه بندی متون، تشخیص چهره و غیره استفاده شدند. موفقیت های چشمگیر ماشین های بردار پشتیبان در بسیاری از کاربردهای طبقه بندی موجب شد که از آن ها در سیستم های بازشناسی گفتار استفاده شود. اما همچنان که ملاحظه می‌شود، تمامی کاربرد هایی که ذکر شد، یک ویژگی مشترک دارند، این که همه این طبقه بندی ها استاتیک هستند. SVM ها برای ساختار زمانی داده ها طراحی نشده اند و قادر نیستند حالت دینامیک گفتار را مدل کنند، برای رفع این مشکل معمولا از SVM ها به صورت ترکیب با HMM در بازشناسی گفتار استفاده می‌شود [18] ,[19] ,[24].

۱-۲- کارهای انجام شده در زمینه تقطیع و برچسب زنی اتوماتیک گفتار

به طور کلی هر مبدل آکوستیک به فونتیک می‌تواند شامل دو قسمت زیر باشد:

[1] Structural-Risk Minimization
[2] Support Vector Machines

۱) تقطیع[1] : در این مرحله سیگنال گفتار به نواحی گسسته در زمان تقسیم می‌شود که در این نواحی، خواص آوایی سیگنال، نمایش دهنده یک واج (یا چندین واج) می‌باشد. خطا در تعیین محل واج ها نه فقط به خاطر محدودیت الگوریتم ها بلکه به دلیل ابهام ذاتی درآنالیز سیگنال گفتار نیز ایجاد می‌شود و به طور کلی حدود دقیق یک قطعه را نمی‌توان تعیین کرد.

۲) برچسب زنی[2] : در این مرحله به هریک از قطعه های به دست آمده از قسمت قبل، یک برچسب نسبت داده می‌شود.

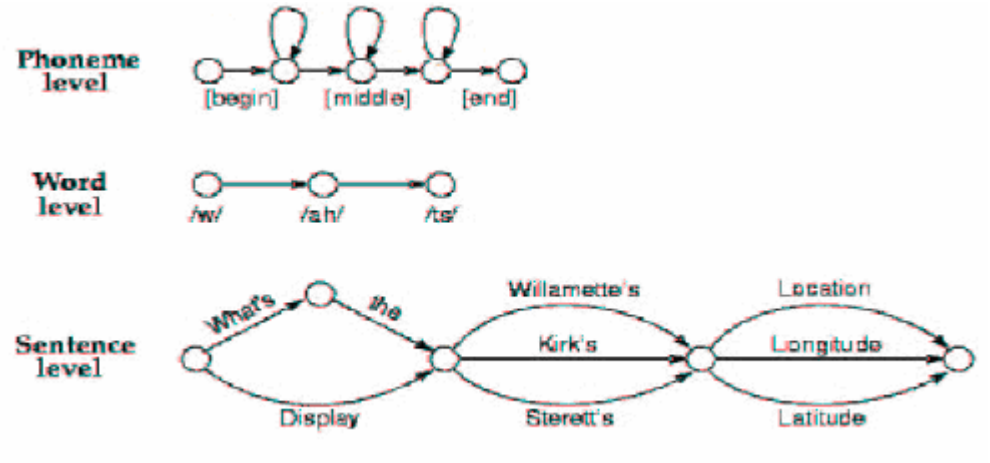
از آنجاییکه هدف این پایان نامه ایجاد بهبود در مبدل آکوستیک به فونتیک گفتار پیوسته فارسی می‌باشد، در این بخش، به بررسی کارهای انجام شده در زمینه تقطیع و برچسب زنی اتوماتیک گفتار در مقالات می‌پردازیم.

۱-۲-۱- استفاده از مدل مارکوف پنهان

در این دسته از روش ها سیگنال گفتار به صورت آماری مدلسازی می‌شود. مدل های مارکوف پنهان، حالت های زمانی سیگنال گفتار را با استفاده از زنجیره مارکوف[3]، مدل می‌کنند. این مدل ها، احتمال مشاهده یک رشته زمانی[4] را بدون دانستن رشته حالت هایی[5] که موجب تولید این رشته مشاهدات[6] شده اند، بدست می‌آورند. تابع توزیع احتمالی که برای هر حالت در مدل HMM در نظر گرفته می‌شود، تغییرات سیگنال گفتار ناشی از گوینده های مختلف و یا ناشی از متون مختلف آوایی را مدل می‌کند. این تابع توزیع معمولا مخلوطی از توابع گوسی در نظر گرفته می‌شود. شکل ۳-۱ نشان می‌دهد که چگونه در نظر گرفتن حالت ها و انتقال حالات[7] در یک مدل HMM می‌تواند به صورت سلسله مراتبی[8]، واج‌ها، کلمات و جملات را نمایش دهد. از مدل های HMM به طور وسیعی برای مدلسازی واحد های گفتاری در سطوح مختلف استفاده شده است [12] ,[25] [26] ,[27]. همان-

[1] Segmentation
[2] Labeling
[3] Markov chain
[4] Temporal sequence
[5] State sequence
[6] Observation sequence
[7] State transition
[8] Hierarchically

طور که در بخش قبل توضیح داده شد، مدل HMM متداول‌ترین روش در سیستم های بازشناسی گفتار می باشد. این مدل دارای قدرت بازنمایی بالایی است. اما برای ایجاد مدل های مقاوم تر و دقیق تر نیاز به روش های با قدرت تمایز بالا برای سیستم‌های بازشناسی گفتار وجود دارد.



شکل ۱-۳- ساختار سلسله مراتبی HMM ها برای بازنمایی واج ها، کلمات و جملات [6]

۱-۲-۲- استفاده از شبکه‌های عصبی

در روش های تقطیع بر اساس شبکه های عصبی، معمولا شبکه های عصبی برای تخمین احتمال وقوع واحد آکوستیکی مورد نظر، آموزش می‌بینند. در بخش برچسب زنی نیز، از خاصیت طبقه بندی و متمایز سازی خوب شبکه های های عصبی استفاده می‌شود. برای مثال در مرجع [28] روشی برای تخمین محل شروع هجاها ارائه شده است. ابتدا ویژگی های مناسبی از سیگنال گفتار استخراج می‌شود و سپس طبقه بندی کننده از این ویژگی ها برای اندازه گیری احتمال وقوع هجا استفاده می‌کند. در مرجع [6] نیز برای بازشناسی واحدهای زیر کلمه با ساختار CV در زبان هندی، از شبکه‌های عصبی استفاده شده است. در این مرجع اساس کار پیدا کردن نقطه شروع واکه ها[1] در سیگنال گفتار می‌باشد. پس از پیدا کردن نقطه شروع واکه، یک فاصله زمانی ثابت ۲۰۰ میلی ثانیه در اطراف این نقطه، به منظور بازشناسی واحد زیرکلمه CV در سیگنال گفتار مورد پردازش قرار می‌گیرد.

[1] Vowel onset point (VOP)