
بسم الله الرحمن الرحيم

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی برق

پایان نامه کارشناسی ارشد مهندسی برق گرایش الکترونیک

استفاده از تکنیک خوشه بندی گوینده در تطبیق گوینده در سیستم بازشناسی گفتار

نگارش:

اولدوز حضرتی یادکوری

استاد راهنما:

دکتر سید محمد احدی

بهمن ماه ۸۶

بسمه تعالي

تاريخ : .....
شماره : .....

فرم اطلاعات پايان نامه
كارشناسي ارشد و دكترا

دانشگاه صنعتي اميركبير
(پلي تكنيک تهران)

معاونت پژوهشي
فرم پروژه تحصيلات تكميلي ۷

مشخصات دانشجو

نام و نام خانوادگي : **اولدوز حضرتی یادکوری** ■ دانشجوي آزاد ☐ بورسيه ☐ معادل ☐

شماره دانشجويي: **۸۴۱۳۰۷۳** دانشكده : **مهندسی برق** رشته تحصيلي: **الکترونیک**

نام و نام خانودگي استاد راهنما : سيد محمد احدی

عنوان به فارسي: استفاده از تکنیک خوشه بندی گوینده در تطبیق گوینده در سیستم بازشناسی گفتار

عنوان به انگليسي: Speaker Adaptation Using Speaker Clustering in Speech Recognition Systems

نوع پروژه: كارشناسي ارشد ☐ **كاربردي** ☐ **بنيادي** ■ **توسعه اي** ☐ **نظري**

تاريخ شروع: ۸۵/۸ **تاريخ خاتمه :** ۸۶/۱۱ **تعداد واحد:** ۶

سازمان تامين كننده اعتبار :

واژه هاي كليدي به فارسي : تطبیق گوینده، خوشه بندی گوینده، تابع کرنل، ماشین های بردار پشتیبان تک کلاسی، K-means ملایم

واژه هاي كليدي به انگليسي : Speaker adaptation, speaker clustering, kernel method, one class support vector machine (OCSVM), soft K-means

---

نظرها و پيشنهادها به منظور بهبود فعاليت هاي پژوهشي دانشگاه:

استاد راهنما: دکتر سید محمد احدی

دانشجو: اولدوز حضرتی یادکوری

امضاء استاد راهنما : **تاريخ:**

نسخه ۱ : معاونت پژوهشي
نسخه ۲ : كتابخانه و به انضمام دو جلد پايان نامه به منظور تسويه حساب با كتابخانه و مركز اسنادومدارک علمي

تقدیم به مادر فداکارو پدر اندیشمندم.

با تشکر فراوان از زحمات استاد گرامی، جناب آقای دکتر احدی که مرا در پیشبرد و تکمیل این پایان‌نامه یاری رساندند. همچنین از آقای مهندس سجادی که مرا در انجام این پروژه مساعدت نمودند، کمال تشکّر را دارم.

چکیده

بطور کلی در بازشناسی گفتار، مدل وابسته به گوینده (SD) عملکرد بهتری نسبت به مدل ناوابسته به گوینده (SI) در تشخیص گفتار یک گوینده خاص دارد. یکی از روش های عملی تر برای دستیابی به عملکردی نزدیک به سیستم SD استفاده از تکنیک های تطبیق گوینده است.

خوشه بندی گوینده یکی از تکنیک های اصلی در تطبیق گوینده است. روش خوشه بندی می تواند به دلیل راحتی ترکیب با تکنیک های رایج تطبیق نظیر MAP و MLLR مورد استفاده قرار بگیرد.

در این پروژه، روشی مبتنی بر کرنل تنها در فاز خوشه بندی مورد استفاده قرار می گیرد. پیاده سازی این روش ساده و حجم محاسباتی مطلوبی دارد. ما این روش خوشه بندی بر مبنای توابع کرنل را که الهام گرفته از روش متداول K-means و بر مبنای ماشین های بردار پشتیبان تک کلاسی (OCSVM) است به عنوان مرحله پیش تطبیق قبل از تکنیک های رایج تطبیق نظیر MAP و MLLR در تطبیق گوینده با نظارت سریع مورد استفاده قرار می دهیم.

در بخشی از کار الگوریتم های خوشه بندی مختلفی نظیر خوشه بندی های جنسیت، خوشه بندی K-means، Soft K-means و روشی بر مبنای بردار های پشتیبان تک کلاسی جهت تطبیق مورد استفاده قرار گرفته اند و با روش پیشنهادی به دقتی در حدود ٪۳ بهتر از مدل مبنا با دقت ٪۹۰/۳۷ (دقت مدل SI) دست یافتیم.

در بخش دیگری از این پروژه استفاده از روش های خوشه بندی ذکر شده به عنوان یک مرحله پیشین در تطبیق به روش های MAP و MLLR مورد استفاده قرار گرفته است. در این حالت نیز به ٪۶/۵ بهبودی نسبت به مدل مبنا رسیدیم.

در ادامه به جهت بررسی کارایی این روش در امر تطبیق، از چندین روش مختلف برای انتخاب HMM ها استفاده شده است.

در بخش دیگری از کار به منظور مقایسه کارایی HMM و GMM در روش استفاده شده جهت امر خوشه بندی در این پروژه، به جای HMM های بکار رفته در مرحلهٔ پیش از خوشه بندی، از GMM استفاده شده است و نتایج بدست آمده مورد بررسی قرار گرفته است.

تمامی آزمایش های انجام شده در این پروژه بر روی دادگان TIDIGITS صورت گرفته و هیچگونه همپوشانی ای بین گوینده های آموزش و تست وجود ندارد.

در انتها نتایج حاصل از تطبیق گوینده به روش eigenvoice و eigenvoice مقاوم آورده شده است که متأسفانه نسبت به مدل مبنا عملکرد ضعیفتری دارند.

کلمات کلیدی: تطبیق گوینده، خوشه بندی گوینده، تابع کرنل، ماشین های بردار پشتیبان تک کلاسی، K-means ملایم.

فهرست مطالب

فهرست اشکال

فهرست جداول

فهرست علایم اختصاری

DFT	*Discrete Fourier Transform*
DTW	*Dynamic Time Warping*
eKEV	*Embedded Kernel Eigenvoice*
EMAP	*Extended MAP*
EV	*EigenVoice*
FFT	*Fast Fourier Transform*
GMM	*Gaussian Mixture Model*
HMM	*Hidden Markov Models*
KMLLR	*Kernel Maximum Likelihood Linear Regression*
MAP	*Maximum a Posteriori*
MFCC	*Mel-Frequency Cepstral Coefficients*
ML	*Maximum Likelihood*
MLE	*Maximum Likelihood Estimation*
MLLR	*Maximum Likelihood Linear Regression*
OCSVM	*One Class Support Vector Machine*
SD	*Speaker Dependent*
SI	*Speaker Independent*
SVM	*Support Vector Machine*
VQ	*Vector Quantization*

فهرست نمادها

تعداد بردارهای مشاهده	T
بعد بردارهای مشاهده	n
بردارهای مشاهده	o
حالت های مدل مارکوف پنهان	$s_1,...,s_N$
تعداد عناصر مخلوط	M
وزن عنصر مخلوط	c_{jk}
بردارمیانگین	μ
ماتریس کواریانس	Σ
واریانس	σ
مدل گوینده	λ
Eigenvoice	e
مرکز خوشه	z_j
مرکز کره	A

فصل اول

مقدمه

فصل اول

مقدمه

۱-۱- کلیات

همانگونه که استفاده از گفتار و کلام یکی از راه های ابتدایی برقراری ارتباط بین انسانهاست، بکارگیری آن در ادارهٔ امور مختلف به جای دست و پا می تواند تا حد زیادی باعث سهولت گردد. بکارگیری این امردر تسهیل انجام کارهای کنترلی و استفاده از ابزارهایی که برای ادارهٔ آنها به حرکت دست و پا نیاز است، برای همه به ویژه افراد ناتوان بسیار حائز اهمیت است.

این امر مستلزم بازشناسی گفتار[1] افراد مختلف است که یکی از دشوارترین مسائل موجود در مبحث شناسایی الگوست. تمهیدات فراوانی به منظور کاهش نرخ خطای[2] بازشناسی گفتار صورت گرفته است که منجر به سیستم های بازشناسی مقاومتری شده اند.

مشکلی که معمولاً در بکارگیری یک سیستم بازشناسی مطرح است، استفاده از یک سیستم آموزش دیده برای گوینده های جدیدی است که به علت نداشتن سیستم وابسته به گویندهٔ مناسب معمولاً دقت بازشناسی مطلوبی ندارند. پس از آموزش یک سیستم با داده آموزشی کافی جمع آوری شده از گوینده های مختلف سیستم حاصله یک سیستم مستقل از گوینده[3] (SI) خواهد بود و چنین سیستمی مسلماً نسبت به سیستمی که تنها با گفتار یک گوینده خاص آموزش دیده و برای همان گوینده نیز تست می

[1] Speech Recognition
[2] Error Rate
[3] Speaker Independent

شود (سیستم وابسته به گوینده[4] (SD))، دقت بازشناسی پایین تری خواهد داشت. از آنجائیکه همیشه جمع آوری داده زیاد از یک گوینده به منظور آموزش سیستم SD او ممکن نیست، بنابراین سعی بر این است که پس از یکبار آموزش سیستم SI، برای گوینده هایی غیر از گوینده های آموزشی نیز بتوان از سیستم SI با دقت خوبی استفاده کرد. به این منظور از روش های تطبیق گوینده استفاده می گردد. در این روش ها با داشتن تنها مقدارمحدودی داده از گویندهٔ جدید، سیستم SI به گونه ای تغییر داده می شود که برای گویندهٔ جدید عملکردی نظیر سیستم SD آن گوینده داشته باشد.

روش هایی نظیر روش های مبتنی بر خوشه بندی گوینده،MAP و MLLR، سالهاست که برای این منظور مورد استفاده قرار می گیرند.

خوشه بندی گوینده یکی از تکنیک های اصلی در تطبیق گوینده است. در این روش، گوینده هایی که خصوصیات آکوستیکی مشابه یا حداقل ویژگی های مشابهی در فضای ویژگی دارند خوشه بندی شده و سپس مدل های وابسته به خوشه برای هر خوشه تشکیل می گردد. در مرحله بازشناسی، خوشه ای که بیشترین مطابقت را با گفتار ورودی داراست، خوشه ای که بیشترین درستنمایی را دارد، انتخاب شده و از مدل های مارکوف پنهان (HMM) این خوشه برای بازشناسی استفاده می شود. این HMM ها همچنین می توانند در تطبیق گوینده به عنوان مدل های ابتدایی به جای HMM های مدل SI مورد استفاده قرار بگیرند. روش خوشه بندی همچنین می تواند به دلیل راحتی ترکیب با تکنیک های رایج تطبیق نظیر MAP و MLLR مورد استفاده قرار بگیرد.

روش های تطبیق گویندهٔ موجود را می توان به دو دسته تقسیم بندی کرد:

روش های مبتنی بر ویژگی و روش های مبتنی بر مدل. نرمالیزاسیون لولهٔ صوتی (VTLN) یکی از روش های تطبیق بر مبنای ویژگی است که تغییرات گوینده ای ایجاد شده به علت طول های مختلف لولهٔ صوتی را اصلاح می کند.

تکنیک های مبتنی بر مدل را می توان به سه مجموعه تقسیم بندی کرد: تکنیک های بیزین، تکنیک های مبتنی بر تبدیل و تکنیک های مبتنی بر فضای ویژه[5]. این روش ها به مقدار مشخصی دادهٔ تطبیق از گویندهٔ جدید احتیاج دارد تا با استفاده از این اطلاعات عملکرد بازشناسی گویندهٔ جدید را بهبود بخشند.

[4] Speaker Dependent
[5] Eigen Space

در تطبیق گوینده، در صورتیکه نسخهٔ آوایی از دادهٔ تطبیق موجود باشد تطبیق با نظارت و در غیر اینصورت بدون نظارت نامیده می شود.

همچنین در این پروژه منظور از تطبیق گویندهٔ سریع، تطبیق گوینده با دادهٔ تطبیقی کمتر از ۱۰ ثانیه می باشد.

۱-۲- طرح کلی پایان نامه

در این پروژه به منظور بهبود بازشناسی گفتار گوینده های جدید، از روش های خوشه بندی استفاده شده است. با استفاده از دادهٔ تطبیق کمتر از ۱۰ ثانیه، با بکارگیری تکنیک خوشه بندی به عنوان یک مرحلهٔ ابتدایی پیش از تطبیق های رایج MAP و MLLR عملکرد سیستم را به میزان قابل توجهی بهبود داده ایم.

در فصل دوم از این پایان نامه ابتدا کلیاتی در مورد ویژگی های گفتار، بازشناسی گفتار، استخراج ویژگی، مدل های مارکوف پنهان و مدل های مخلوط گوسی آورده شده است.

در فصل سوم روش های متداول موجود در تطبیق گوینده شامل روش MAP[6] و MLLR[7] و روش نسبتاً جدید Eigenvoice معرفی شده اند.

در فصل چهارم چند روش خوشه بندی شامل خوشه بندی جنسیت، باینری، K-means، C-means و Soft K-means توضیح داده شده اند. همچنین در انتهای این فصل روش خوشه بندی ای بر مبنای بردارهای پشتیبان تک کلاسی[8](OCSVM) معرفی شده است.

در فصل پنجم آزمایش های انجام شده بر روی دادگان TIDIGITS، نحوه انجام هر یک و نتایج حاصله آورده شده است.

در فصل ششم نیز به جمع بندی، نتیجه گیری و پیشنهاد پرداخته ایم.

[6] Maximum a Posteriori
[7] Maximum Likelihood Linear Regression
[8] One Class Support Vector Machines (OCSVM)

فصل دوم :

کلیاتی بر بازشناسی گفتار

فصل دوم

کلیاتی بر بازشناسی گفتار

۲-۱- مقدمه

در این بخش توضیحاتی در رابطه با بازشناسی گفتار آورده شده که به صورت زیر سازماندهی شده است: ابتدا ویژگی های سیگنال گفتار توضیح داده شده است، سپس بازشناسی گفتار در سیستم های بازشناسی گفتار کنونی توضیح داده شده است. همچنین مدل های مارکوف پنهان[1] (HMM) و مدل های مخلوط گوسی[2] (GMM) مورد بررسی قرار گرفته است و در انتها روش استخراج ویژگی سیگنال گفتار توضیح داده شده است.

۲-۲- ویژگی های گفتار

در بازشناسی گفتار سعی بر استخراج اطلاعات موجود در سیگنال و تشخیص آن گفتار است. درانتقال سیگنال از فرستنده به گیرنده ممکن است سیگنال دچار اغتشاشاتی نظیر اغتشاش کانال، صدای دستگاه های اطراف و یا گفتار سایر گوینده ها شود. هر تبدیلی سیگنال را تغییر داده و بازشناسی آن را مشکل تر می سازد.

سیگنال گفتار دارای تغییر پذیری های ویژه ای است که بیان ریاضی آن را مشکل می سازد. برای مفید بودن، یک بازشناس گفتار باید کاری کند که این اثرات حذف شوند و یا ساختاری را متضمن شود که تاثیر این تغییرات را در نظر بگیرد[1].

[1] Hidden Markov Model
[2] Gaussian Mixture Model

در این بخش منبع چنین تغییراتی را بیان می کنیم. برای سادگی این تغییرات را به چندین دسته تقسیم بندی می کنند، البته باید در نظر داشت که این بخش ها ممکن است همپوشانی هم داشته باشند [2]:

۱) تغییرات شیوه[3] صحبت کردن: این تغییرات توسط گوینده قابل کنترل هستند که شامل مواردی از قبیل: دقت، وضوح ، شمرده سخن گفتن و ... می باشد.

۲) زمینه: شرایطی که در آن تولید سیگنال صورت می گیرد تاثیراتی بر گفتار دارد. این شرایط ممکن است بر سرعت صحبت، تغییرات شیوه صحبت و تاکید تاثیر داشته باشد.

۳) استرس: شامل فاکتورهای احساسی و تغییرات تحمیلی محیط می باشد. نمونه های متداول آن ترس و واکنش لمبارد[4] است.

۴) کیفیت صدا: این بخش شامل تاثیراتی نظیر صدای ناراحت و عصبی، سوت زدن و ... است.

۵) آهنگ صحبت: آهنگی که صحبت با آن ادا شده است که در قابلیت فهم نیز موثر است.

علاوه بر این موارد، تفاوت های فیزیولوژیکی نظیر جنسیت نیز بر گفتار تاثیر می گذارد. به عنوان مثال خانم ها لوله صوتی کوتاهترو فرکانس گام[5] بالاتری نسبت به آقایان دارند. همچنین احتمال اینکه خانم ها حجم[6] صدای پایین تری نسبت به آقایان داشته باشند بیشتر است.

۲-۳- بازشناسی گفتار

در سیستم های بازشناسی متداول دو مرحله وجود دارد:

۱) فاز آموزش: در این مرحله سیستم بازشناسی راه اندازی می شود.

۲) فاز بازشناسی: در آن از سیستم بازشناسی برای تشخیص صحبت بیان شده استفاده می شود.

[3] Style
[4] Lombard reflex
[5] Pitch
[6] volume