ارائه روش SVQ در بازشناسی گفتار گسسته

ابوالقاسم صیادیان
استادیار

دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

چکیده

فرآیند VQ به عنوان یك مرحله کمکی در کاهش بار محاسباتی اغلب روش های بازشناسی گفتار مورد استفاده قرار می گیرد. در این تحقیق از فرآیند SVQ به عنوان یك روش مستقل در بازشناسی گفتار گسسته استفاده نمودیم. در واقع روش SVQ را به عنوان یك نوع مدلسازی غیرپارامتریك آماری برای بازشناسی گفتار گسسته با بار محاسباتی و حجم حافظه کم مطرح نمودیم. نتایج عملی نشان می دهد که بازدهی این روش در بازشناسی گفتار گسسته بسیار قابل توجه می باشد.

Presentation of SVQ Method for Isolated Word Recognition

A. Sayadian
Assistant Professor

Department of Electrical Engineering ,
Amirkabir University of Technology

Abstract

The VQ Process is often used as computational reduction step in speech recognition systems. We use SVQ as an independent and new method for Isolated word recognition. The SVQ has been introduced as a nonparametric statistical modelling requiring a limited storage and computing load in comparison with the other methods (such as DTW). The results of this research show that the performance of SVQ is comparable with the other classic methods.

۱ـ مقدمه

روش‌های کلاسیک در بازشناسی گفتار، روش‌های DTW[۲] و HMM[۳] و ANN[۴] می‌باشند [۶]، ...، [۱]. از میان روش‌های مطرح شده، روش HMM به عنوان یک روش پارامتریک آماری و روش DTW به عنوان یک روش غیر پارامتریک آماری مطرح هستند. در تمامی این روش‌ها از فرآیند VQ به عنوان یک مرحله کمکی برای کاهش هزینه محاسباتی و حجم حافظه مورد نیاز سیستم می‌توان استفاده نمود [۷] و [۶]. هدف این تحقیق بررسی یک روش غیر پارامتریک آماری با بار محاسباتی و حجم حافظه پایین بوده است، به قسمی که بازشناسی ۲۰۰ کلمه وابسته به گوینده بر روی سیستم‌های PC فعلی (پنتیوم) بدون امکانات اضافی در مد چند وظیفه‌ای[۵] به راحتی امکان پذیر باشد. تأمین اهداف فوق توسط روش DTW به علت بالا بودن بار محاسباتی امکان پذیر نگردید، در نتیجه این تحقیق بر روی روش بازشناسی SVQ به عنوان روش جایگزین DTW با بار محاسباتی بسیار پایین متمرکز گردیده است. در بخش دو به روش مدلسازی SVQ می‌پردازیم. در بخش سه به ارائه روش TCP برای تولید الگوهای مرجع برای سیستم SVQ خواهیم پرداخت. در بخش ۴ نتایج شبیه سازی مورد تجزیه و تحلیل قرار می‌گیرد. بخش ۵ بر روی تحلیل بار محاسباتی مورد نیاز روش SVQ متمرکز شده است. در بخش ۶ نیز به جمع بندی کار انجام شده و نتایج حاصل می‌پردازیم.

۲ـ مدلسازی SVQ برای بازشناسی گفتار گسسته

مدل SVQ برای هر کلمه از یک مجموعه کتاب کد مرتب $\lambda = (CB_1, ..., CB_N)$ تشکیل شده است. N تعداد قطعات هر کلمه است $(3 \le N \le 10)$. CB_i $(i = 1$ تا $N)$ کتاب کد متناظر با قطعه iام از کلمه بوده و دارای N_i بردار مرجع می‌باشد. به منظور مدلسازی مؤثر روش SVQ برای جبران تغییرات طول کلمه و همچنین جهت کاهش خطا در اثر تغییرات غیر خطی طول فونمها[۶] در داخل هرکلمه، از روش OS[۷] به شرح زیر استفاده می‌نماییم. البته در روش کلاسیک از روش DS[۸] یا NOS[۹] استفاده می‌شود [۸] و [۷].

۱ـ۲ـ مرحله آموزش (فاز یادگیری)

الف) طول هر کلمه را به ۲N قطعه متساوی الطول تقسیم می‌نماییم. چنانچه در تقسیم طول کلمه به ۲N باقیمانده داشتیم، فریم‌های باقیمانده را به طور مساوی به قطعات ابتدا و انتها نسبت می‌دهیم. حال از روی ۲N قطعه متوالی، تعداد N قطعه متوالی به شرح زیر انتخاب می‌نماییم:

شکل (۱) نحوه تقسیم طول کلمه به ۲N قطعه

$$S'_1 = \{S_1, S_2, S_3, S_4\} \ , \ S'_N = \{S_{2N-3}, S_{2N-2}, S_{2N-1}, S_{2N}\}$$

$$S'_i = \{S_{2i-2}, S_{2i-1}, S_{2i}, S_{2i+1}\} \quad \text{for } i = 2 \text{ تا } N-1 \qquad (۱)$$

در واقع هر قطعه انتخاب شده به اندازه نصف طول قطعات با قطعه ماقبل و مابعد همپوشانی[۱۰] دارند (به غیر از قطعه شروع و پایانی که به اندازه طول قطعات همپوشانی دارند).

ب) برای هر قطعه iام S'_i یک کتاب کد متمایز به نام CB_i با تعداد الگوهای N_i طراحی می‌نماییم. چنانچه تعداد نمونه‌های آموزشی برای هر کلمه بیش از یکی باشد، کتاب کد CB_i از روی قطعات iام متناظر کلیه نمونه‌های آموزشی مربوط به همان کلمه طراحی می‌گردد.

ج) مراحل الف و ب را برای کلیه کلمات سیستم بازشناسی به طور مجزا انجام می‌دهیم.

۲ـ۲ـ مرحله بازشناسی

الف) همانند مرحله الف فاز یادگیری، هر کلمه تست ورودی را به N قطعه با مقدار همپوشانی بیان شده تقسیم می‌نماییم.

ب) تابع درست نمایی[۱۱] بین کلمه تست X و مدل کلمه W_i را به شرح زیر محاسبه می‌نماییم:

$$W_i \equiv (CB^i_1 , CB^i_2 , \dots CB^i_N) \quad (۲)$$

$$X \equiv (X_1 , X_2 , \dots, X_N) \quad (۳)$$

$$X_k \equiv (X^1_k , X^2_k , \dots, X^{pk}_k) \quad (۴)$$

$$d_{min} (Y , CB^i_k) \equiv \min \{ d\left(Y , CB^i_k (1)\right), \dots , d\left(Y , CB^i_k (N_i)\right)\} \quad (۵)$$

$$d (X , Y) = \text{Distance Measure Between Vectors X And Y} \quad (۶)$$

$$P_k = \text{Number of Frames in Segment K} \quad (۷)$$

$$d (X_k , CB^i_k) \equiv \sum_{m=1}^{P_K} d_{min} (X^m_k , CB^i_k) \quad (۸)$$

$$D (X , W_i) \equiv \sum_{k=1}^{N} d (X_k , CB^i_k) \quad (۹)$$

$$D (X , W_i) = \sum_{k=1}^{N} \sum_{m=1}^{P_k} d_{min} (X^m_k , CB^i_k) \quad (۱۰)$$

مقدار تابع $D (X, W_i)$ را برای L تا $i = 1$ به دست می‌آوریم (L برابر تعداد کلمات سیستم می‌باشد). خروجی فاز بازشناسی، کلمه‌ای است که مقدار $D (X, W_i)$ متناظر آن از بقیه کلمات کمتر باشد.

۳ ـ ارائه روش TCP [۱۲] برای طراحی الگوهای مرجع قطعات

برای طراحی بردارهای مرجع از روش‌های کلاسیک فرآیند VQ همانند الگوریتم LBG[۱۳] [۱۰] [۹] و FSL[۱۴] [۱۲] می‌توان استفاده نمود. باتوجه به دقت مورد نیاز در روش SVQ و وابستگی روش‌های ذکر شده به انتخاب شرایط اولیه، روش TCP را برای این منظور پیشنهاد و مورد استفاده قرار دادیم. این روش از نظر پیاده‌سازی بسیار ساده و سریع بوده و نتیجه عملکرد آن با انتخاب یک پارامتر کمکی به نام حداکثر خطای نسبی ε به راحتی قابل تنظیم می‌باشد. مراحل عملی طراحی بردارهای مرجع به روش TCP برای هر قطعه به شرح زیر می‌باشد:

الف) فرض می‌کنیم D_{max}، مقدار حداکثر فاصله بین دو بردار در جمیع بردارهای سیستم بازشناسی (در فاز آموزش) باشد. در این صورت تابع فاصله نسبی را به صورت زیر تعریف و مورد استفاده قرار می‌دهیم.

$$dr (X, Y) \equiv d (X , Y) / D_{max} \quad (۱۱)$$

بعد از این در تمامی مراحل فاز آموزش و بازشناسی از تابع فاصله نسبی ( )dr استفاده می‌نماییم. فرض می‌کنیم ε حداقل فاصله نسبی لازم برای متمایز کردن دو بردار مشاهده، باشد. همچنین فرض می کنیم Y_F .. و Y_1 و Y_2 مجموعه بردارهای آموزشی مربوط به قطعه iام باشند ($F > N_j$). می‌خواهیم N_i بردار مرجع از روی این بردارهای آموزشی طراحی و تولید نماییم.

ب) مراحل زیر را برای تولید N_i کد انجام می‌دهیم:

۱ ـ $C_1 \Leftarrow Y_1 \quad , \quad N_i \Leftarrow 1 \quad , \quad K \Leftarrow 2$

۲ ـ $dr_{min} = \min \{dr (Y_k , C_1) , \dots dr (Y_k , C_{Ni})\}$ را به دست می‌آوریم.

۳ ـ چنانچه $dr_{min} > \varepsilon$ باشد، یک کد جدید تولید می نماییم.

$$C_{Ni+1} \Leftarrow Y_k \quad , \quad N_i \Leftarrow N_i + 1 \quad , \quad K \Leftarrow K + 1 \quad (۱۲)$$

اگر $dr_{min} \leq \varepsilon$ باشد، کدی تولید نخواهد شد و فقط $K \Leftarrow K + 1$ جایگزین می‌شود.

۴ ـ مراحل ۲ و ۳ را تا اتمام بردارهای آموزشی (K< F) تکرار می‌نماییم.

با تنظیم ε می‌توان تعداد کدهای تولید شده را تنظیم نمود. εهای کوچک کدهای بیشتری را تولید نموده و دقت سیستم را افزایش می‌دهند.

٤ ـ بررسی نتایج شبیه‌سازی

به منظور ارزیابی عملکرد روش SVQ، تحت شرایط $\frac{S}{N} \geq 25db$, $F_s = 8$ KHz و دیجیتایزر ۱۶ بیتی مجموعه

آزمون های زیر انجام پذیرفته است. در تمامی این آزمون ها درجه فیلتر مدل LPC برای بازنمایی پوش طیف سیگنال برابر ۱۰ و تعداد ضرایب کپسترم LPC برابر ۱۶ در نظر گرفته شده است. البته در تعدادی از مراجع [۱۳] و [۱۲] درجه فیلتر را ۸ و تعداد ضرایب را ۱۲ پیشنهاد می نمایند، ولی تحقیقات ما نشان می دهد که با تمهیداتی درجه ۱۰ بهتر عمل می نماید. ضرایب کپسترم ابتدا در لیفترجانگ [۱۲] ضرب شده و سپس تبدیل لازم برای بردن آنها در مقیاس MEL[۱۴] انجام پذیرفت. علاوه بر ۱۶ ضریب کپسترم در مقیاس MEL، ۱۶ ضریب دلتا کپسترم و ۱۶ ضریب دلتا دلتاکپسترم و گین نرمالیزه شده (نسبت به گین ماکزیمم) و دلتاگین و دلتادلتاگین، به عنوان مجموعه ویژگی ها ، مورد استفاده قرار گرفتند. سیستم برای بازشناسی ۲۰۰ کلمه وابسته به گوینده [۱۵] طراحی و پیاده سازی گردید. نمونه های آموزشی ضبط شده از ۵ مرد و ۵ زن و ۵ بچه مورد استفاده قرار گرفتند . هر گوینده هر کلمه را ۳۰ بار بیان نموده است (طی روزهای متفاوت). ۲۰ نمونه (به صورت انتخاب تصادفی) برای آموزش و ۱۰ نمونه باقیمانده برای تست و ارزیابی سیستم مورد استفاده قرار گرفت. در بیان کلمه توسط گویندگان سعی شده که تنوع حالات و تأکید بیان، حتی الامکان رعایت گردد. در تمامی آزمون ها برای تولید الگوهای مرجع از روش TCP استفاده شده است.

الف) در آزمون اول متوسط درصد بازشناسی برای مردها ، زنها و بچه ها به صورت متمایز محاسبه گردید. نتایج حاصل در جدول (۱) درج شده است. این آزمون با ٪۲ = ε و ۶ = N انجام پذیرفته است.

جدول (۱) نتایج متوسط صحت بازشناسی روش SVQ در بازشناسی ۲۰۰ کلمه مجزا.

	مردها	زنها	بچه ها	متوسط افراد
مقدار متوسط	۹۹/۲۸	۹۶/۷۲	۹۴/۹۵	۹۷/۳۲
انحراف معیار	۰/۷۳	۲/۱۱	۲/۹۵	۱/۹۶

جدول (۲) نتایج تأثیر ε در عملکرد روش SVQ.

ε	٪۱	٪۲	٪۳	٪۴	٪۵
متوسط صحت بازشناسی	۹۷/۴۱	۹۷/۳۲	۹۶/۵۳	۹۵/۱۱	۹۳/۱۸

ملاحظه می شود که نتایج حاصل کاملاً قابل مقایسه با نتایج منتشر شده روش های DTW و DDHMM می باشد [۱۴] [۱۳] [۶] [۴] [۱].

ب) هدف آزمون دوم بررسی تأثیر مقدار ε در صحت بازشناسی بوده است. جدول (۲) متوسط نتایج صحت بازشناسی را برای ۵ مقدار ε نشان می دهد.

همانطوری که از جدول (۲) مشاهد می گردد، با افزایش ε خطای بازشناسی افزایش می یابد. $\varepsilon < 3$ رنج مناسبی به نظر می رسد. بهترین حالت برای $\varepsilon = 1\%$ حاصل شده است. کاهش نامناسب ε موجب افزایش تعداد الگوهای مرجع شده و در نتیجه موجب افزایش بار محاسباتی و حافظه مورد نیاز سیستم می گردد. به نظر می رسد که $\varepsilon = 2\%$ به علت ایجاد توازن مناسب در نتیجه بازشناسی و تولید تعداد الگوهای مرجع مقدار مناسبی باشد.

ج) هدف آزمون سوم بررسی تأثیر مقدار N (تعداد قطعات) در نتیجه عملکرد سیستم بوده است. نتایج حاصل در جدول ۳ درج شده است. آزمون در این حالت با $\varepsilon = 2\%$ انجام پذیرفته است.

جدول (۳) تأثیر مقدار N در نتیجه بازشناسی روش SVQ.

	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
متوسط صحت بازشناسی	۸۵/۸۳	۹۰/۷۵	۹۴/۱۷	۹۵/۹۱	۹۷/۱۲	۹۷/۳۲	۹۶/۵۲	۹۵/۱۳	۹۴/۴۷	۹۳/۱۹

همانطوری که در جدول (۳) مشاهده می‌شود، $N = 5, 6$ تعداد قطعات مناسبی برای عملکرد بهینه سیستم می‌باشند. برای $N=1$ (یعنی کل طول کلمه به عنوان یک قطعه در نظر گرفته شود) هنوز سیستم دارای حدود ۸۵٪ صحت بازشناسی است. ملاحظه می‌گردد که با افزایش تعداد قطعات از ۱ به ۲، حدود ۵٪ نتیجه صحت بازشناسی اضافه شده است. در جمع‌بندی کلی، به نظر می‌رسد که انتخاب $N = 6$ تعداد قطعات مناسبی برای عملکرد بهینه روش SVQ خواهد بود (حداقل برای بازشناسی کلمات گسسته زبان فارسی).

۵ـ تحلیل بار محاسباتی روش SVQ

همانطوری که در مقدمه بیان گردید، روش SVQ یک روش بازشناسی آماری غیر پارامتریک برای بازشناسی گفتار گسسته می‌باشد (در مقابل روش آماری پارامتریک HMM). روش کلاسیک مشهور غیر پارامتریک برای بازشناسی گفتار روش DTW می‌باشد. بنابراین منطقی است که مقایسه‌ای از نظر بار محاسباتی مورد نیاز بین این دو روش انجام پذیرد. (با توجه به عملکرد تقریباً یکسان در نتیجه بازشناسی). روش DTWبا ساختارهای متفاوتی پیاده‌سازی می‌گردد، معمولترین آن دارای رنج مقایسه طول (۲ و ۱/۲) می‌باشد (یعنی طول هر کلمه تست در مقایسه با کلمه مرجع حداکثر ۲ برابر، انبساط یا انقباض می‌یابد). فرض می‌کنیم L تعداد فریم‌های متوسط کلمات و M تعداد کل کلمات سیستم و M_1 تعداد مدل‌های هر کلمه باشد ($M_1 \geq 1$). در این صورت بار محاسباتی لازم در فاز بازشناسی برای روش DTW متناسب با $M_1 . M. L^2$(محاسبه یک تابع فاصله و عملیات مقایسه) می‌باشد. در روش SVQ به علت گسسته بودن فضای بردارهای طیفی، امکان استفاده از اطلاعات پیشین [۱۶] برای محاسبه تابع درست نمایی (به منظور کلاس بندی کلمه تست ورودی) به سهولت وجود دارد. فرض می‌کنیم F تعداد کل بردارهای مرجع سیستم SVQ باشد (نوعاً در رنج ۱۰۲۴ – ۲۵۶ قرار دارد). عملیات فاز کدینگ فرآیند VQ را می‌توان متناسب با تعداد $\text{Log}_2 F$ انجام داد (با ایجاد ساختارها سریع بر روی الگوهای مرجع). اگر از اطلاعات پیشین برای بازشناسی کلمات استفاده نماییم، هزینه محاسباتی مرحله بازشناسی این روش متناسب با $L.\text{Log}_2 F.\text{Log}_2 M$ خواهد بود. به منظور مقایسه نسبی و درک قابلیت روش SVQ در کاهش هزینه محاسباتی فرض می‌کنیم، $L = ۱۰۰$ و $M = ۲۰۰$ و $M_1 = ۳$ (سه مدل بازاء هر کلمه در روش DTW) و $F = ۱۰۲۴$ باشند. در این صورت نسبت هزینه محاسباتی روش DTW به روش SVQ در فاز بازشناسی عبارت است از:

$$r = \frac{O\,(DTW)}{O\,(SVQ)} \approx \frac{3\times 200\times 100\times 100}{100\times 10\times 8} \approx 750$$

نسبت فوق دلیل جذابیت روش SVQ در کاهش شدید بار محاسباتی بازشناسی گفتار گسسته در مقایسه با روش DTW می‌باشد. حال اگر بار محاسباتی پیش‌بینی نشده‌ای را برای روش SVQ هم در نظر بگیریم به قسمی که هزینه محاسباتی آن را سه برابر نماید، باز هم نسبت فوق بالای ۲۵۰ برابر خواهد بود که در کاربردهای عملی بسیار مناسب و جذاب می‌باشد. خوشبختانه به علت استفاده از روش TCP در تولید بردارهای مرجع، بار محاسباتی فاز آموزش روش SVQ نیز بسیار کمتر از روش DTW می‌باشد. تنها بار محاسباتی اضافی روش SVQ پس از فاز آموزش، ایجاد ساختار مناسب (برای جستجوی سریع در کتاب کد) بر روی الگوهای مرجع می‌باشد. با توجه به اینکه فرآیند فوق یک بار و آنهم پس از طراحی الگوهای مرجع انجام می‌پذیرد، مشکلی در استفاده از روش SVQ ایجاد نمی‌نماید (بار محاسباتی آن در مقایسه قابل اغماض می‌باشد).

۶ـ جمع‌بندی و نتیجه‌گیری

روشی به نام SVQ برای بازشناسی کلمات گسسته طی این نوشتار ارائه گردید. مبانی اولیه این روش قبلاً وجودداشته و مؤلف با توسعه مبانی اولیه و بهینه سازی

مدل، قادر گردید که به نتایج ارزشمندی در بازشناسی گفتار گسسته دست یابد. مزیت عمده روش SVQ در مقایسه با روش DTW (روش کلاسیک غیر پارامتریک دیگر) کاهش شدید بار محاسباتی در فاز بازشناسی می‌باشد (با عملکرد تقریباً یکسان در صحت بازشناسی). به نظر می‌رسد که این روش با اصلاحات انجام شده در طی این تحقیق دارای پتانسیل مناسبی برای بازشناسی گفتار گسسته باشد. برای بررسی کارائی روش SVQ در مقابل روش‌های آماری پارامتریک مطرح (روش HMM) تحقیقاتی توسط مؤلف شروع گردیده و هم اکنون ادامه دارد.

زیرنویس‌ها

1 - Segmental Vector Quantization
2 - Dynamic Time Warping
3 - Hidden Markov Model
4 - Artificial Neural Net
5 - Multi Task Mode
6 - Phoneme
7 - Overlap Segment
8 - Non Overlap Segment
9 - Distinct Segment
10 - Overlapping
11 - Maximum Likelihood
12 - Therthold Code Production
13 - Linde, Bazo, Gray
14 - Frequency Sensteive Learning
15 - Modified Equal Loadness Scale
16 - Speaker Dependent
17 - Priori Information

مراجع

[1] C. Myers, L. R. Rabiner, A. E. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Recognition", IEEE Trans. on ASSP, ASSP-28 (6): 623-635, December 1980.

[2] H. Sakoe, S. Chiba, "Dynamic Programming Optimization for Spoken Word Recognition", IEEE Trans. on ASSP, ASSP-26 91): 43-49, February 1978.

[3] S. E. Levinson, L. R. Rabiner, M. M. Sondhi, "An Introduction to the Application of the Theory of Probahilistic Functions of a Markov Process to Autoatic Speech Recognition", Bell System Tech. J. 62 94): 1035-1074, April 1983.

[4] L.R. Rabiner, B. H. Juang, S.E. Levinson, M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", AT & T Tech. J., 64 (6): 1211-1234, July 1985.

[5] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", Proc. IEEE, 77 (2): 257-286, February 1989.

[6] L.R. Rabiner, C. K. Pan, F. K. Soong, "On the Performance of Isolated Word Recognizer Using Vector Quantization and Temporal Energy Co-tours", At & T Tech. J., 63 97): 1245-1260, 1990.

[7] K. C. Pan, F.K. Soong, L. R. Rabiner,: A Vector Quatization Based on Preprocessor for Speaker Independent Isolated Word Recognition", IEEE Trans. On ASSP Proc. ASSP-30 93): 546-560, June 1985.

[8] J.E Shore, D. K. Burton, "Discrete Ulterance Speech Recognition Without Time Alignment", IEEE Trans. on I.T., IT-29 (4) : 437 - 491, July 1983.

[9] R. M. "Vector Quantization", IEEE ASSP Mag. April 1984.

[10] Y. Linde, A. Bazo, R.M. Gray, "An Algorithm for Vector Quantization Design", IEEE Trans. on Com 28, 1980.

[11] S.C. A Halt, A. K. Krishnamurthy, P. chen, D. Melton," Vector Quantization Using Frequecy Sensitive Competitive Learning" IEEE Proc. on ICSE, 1989.

[12] B. H. Juang. J. G. Wilpon, and L.R. Rabiner, "On the Use of Band Pass Liftering in Speech Recognition", IEEE Trans. On ASSP , 35 (7): 947-954, July 1987.

[13] L.R. Rabiner, J. G. Wilpon, "Some Performance Benchmark for Isolated Word Recognition Systems", Computer Speech and Language 2: 343-357, 1987.

[14] J. E. Shore, D.K. Burton, "Some New Performance Benchmark for Large and Speaker Independent Isolated Word Recognition", IEEE Proc on ICASSP, 1996.