به نام خدا

دانشکده علوم و فنون نوین

پردازش گفتار (۱۴۸-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

تاریخ تحویل: ۱۴۰۳/۰۲/۲۱

تمرین شماره ۳

---

۱. (۱۰٪) [**پژوهش**] امروزه هوش مصنوعی مولد (Generative AI) در حوزه‌های مختلفی قادر به تولید داده با کیفیت مطلوب (مانند تولید با ChatGPT و تولید تصویر با Dall-E)، شده است. در این تمرین مروری بر **دو** نمونه از سامانه‌ها و روش‌های علمی تولید گفتار یا تولید موسیقی داشته باشید و نتیجه بررسی خود را همراه با ذکر منابع گزارش کنید.

---

۲. (۲۰٪) [**تخمین**]

۲-۱ فرض کنید یک سری زمانی اسکالر به صورت $y(0), y(1), \ldots$ وجود دارد که توسط مدل زیر نمایش داده می‌شود:

$$y(t) = a_0 w(t) + a_1 w(t-1) + \cdots + a_N w(t-N)$$

که در آن $w(t)$ نمایانگر نویز سفید گاوسی با توزیع مستقل و همگن $N(0,1)$ است. ضرایب $a_0, a_1, \cdots, a_N$ شناخته شده هستند.

الف) پیش‌بینی‌کننده مبتنی بر روش تخمین $MMSE$ برای $y(t+1)$ بر اساس $y(t)$ را بیابید.

ب) برای $t > 1$، پیش‌بینی‌کننده $MMSE$ برای $y(t)$ را بر اساس $y(t-1)$ و $y(t+1)$ برای $t \geq 1$ بیابید

۲-۲ تخمین بیشینه شباهت (Maximum Likelihood) متغیر λ در تابع توزیع پواسون را برای تعداد N نمونه از داده‌ها بدست آورید.

توزیع پواسون $\quad p_x(k) = P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad , k = 0,1,2, \ldots$

---

۳. (۲۰٪) [**پیاده‌سازی: پنجره‌گذاری**] کار شما در این سوال پر کردن دو فایل ex3_main و ex3_windowing است که همراه با تمرین ارائه شده است. تابع فایل ex3_windowing را کامل کنید که این تابع چهار ورودی را می‌گیرد:

خود صدا :data

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

پردازش گفتار (۱۴۸-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

تاریخ تحویل: ۱۴۰۳/۰۲/۲۱

تمرین شماره ۳

frame_length: طول یک فریم (تعداد نمونه‌ها)

hop_size: (frame_length – overlap_size)

windowing_function: نوع پنجره که یکی از چهار مورد زیر است

```
('rect','hann','cosine','hamming')
```

در ضمن منظور از overlap_size طول میزان اشتراک یک فریم با فریم بعدی (تعداد نمونه‌ها) است. خروجی این تابع یک ماتریس N در M است. که N برابر frame_length است و M برابر number_of_frames است.

number_of_frames = 1 + [(data_length – frame_length) / overlap_size]

حال در فایل ex3_main ابتدا فایل صدای موجود در پوشه‌ی Sounds به نام SX83 را بخوانید (frame_rate = 16000). با فرض اینکه مدت زمان هر فریم برابر ۲۵ میلی ثانیه باشد و overlap برابر ۵۰% باشد و از پنجره‌ی hamming استفاده کنیم. خروجی تابع ex3_windowing را بدست آورید. حال یک plot با سه subplot متفاوت به صورت عمودی در نظر بگیرید. در subplot1 باید سیگنال صدای اصلی را رسم کنید که محور x را محور زمان (برحسب ثانیه) و محور y را برابر با دامنه (amplitude) در نظر بگیرید. در subplot2 باید یک فریم صدادار (voiced) به طور دلخواه را رسم کنید، محور x برابر با زمان (برحسب میلی ثانیه) و محور y برابر با دامنه آن فریم صدادار است. در subplot3 باید همان فریم صدادار که در نظر گرفته‌اید را در دامنه فرکانس رسم کنید. بدین صورت که محور x برابر با فرکانس و محور y برابر با دامنه باشد. حال در یک plot جداگانه نصفه اول magnitude spectrum همان فریم صدادار را رسم کنید که محورx برابر با frame_number و محور y برابر با فرکانس است. کل این تمرین را با عوض کردن پنجره هر بار به یکی از سه پنجره باقی مانده دوباره گزارش کنید.

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

پردازش گفتار (۱۴۸-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۳-۱۴۰۲

تاریخ تحویل: ۱۴۰۳/۰۲/۲۱

تمرین شماره ۳

---

۴. (۵۰٪) [**پیاده‌سازی: تشخیص اعداد انگلیسی با ویژگی‌های MFCC و LPC**] در این تمرین می‌خواهیم از دو روش استخراج ویژگی MFCC و LPC برای تشخیص اعداد انگلیسی صفر تا نه استفاده کنیم. برای این کار از دادگان صوتی همراه تمرین استفاده کنید. برای اینکه کار تشخیص را انجام دهید لازم است برای هر عدد تعداد ۱۹۰ نمونه آن عدد را از مجموعه Train به عنوان داده بخش آموزش خود استفاده کنید. سپس برای اینکه کار مقایسه آسان‌تر شود، همه داده‌ها را با افزودن صفر به اول و آخر فایل‌ها با بزرگترین نمونه (در کل داده‌های آموزش و آزمون) هم طول کنید.

برای انجام کار تشخیص، بردارهای ویژگی MFCC یا LPC هر نمونه را با روش بیان شده در قسمت‌های زیر استخراج کنید و هر فایل آزمون را با همه نمونه‌های آموزش همه اعداد مقایسه کنید. برای این کار، فاصله اقلیدسی بردارهای ویژگی آنها را محاسبه کنید. سپس، برای هر عدد آزمون، تعداد K نمونه (مثلا ۵ نمونه) از داده‌های آموزش که فاصله کمتری با آن عدد دارند را انتخاب کنید. حال از بین این K نمونه آموزشی (که می‌دانید چه عددی است)، عددی را که بیشتر از بقیه تکرار شده است، به عنوان پاسخ انتخاب کنید. مثلا اگر برای فایل text1.wav تعداد ۵ نمونه آموزشی نزدیک به آن بیانگر اعداد ۲، ۳، ۶، ۹ و ۳ باشند، عدد ۳ را به دلیل اینکه بیشتر از بقیه تکرار شده است، انتخاب کنید. به این روش دسته‌بندی K نزدیک‌ترین همسایه (KNN) گفته می‌شود!

به عنوان نتیجه، درصدی از تعداد نمونه‌های درست تشخیص داده شده از دادگان آزمون (Test Set) توسط این روش به نسبت تعداد کل نمونه‌های آزمون را حساب کنید و گزارش کنید (معیار Accuracy).

الف) از روش MFCC با طول فریم ۲۰ میلی‌ثانیه، ۲۴ فیلتر مل و تعداد ۱۲ ویژگی به همراه مشتق‌های اول و دوم استفاده کنید و مقدار Accuracy هر عدد و میانگین دقت کلی را به ازای K برابر با ۷، ۱۱ و ۱۵ و ۲۰ گزارش کنید.

ب) قسمت الف را تکرار کنید با این تفاوت که از روش LPC با تعداد ۱۴ ویژگی استفاده کنید.

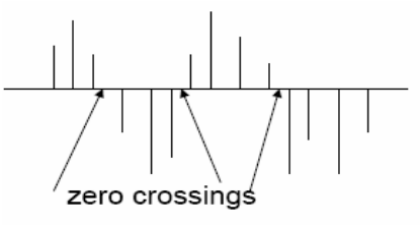
ج) به جاری روش KNN، از روش SVM با دو هسته (کرنل) خطی (linear) و غیرخطی (poly) استفاده کنید و کارایی دو روش استخراج ویژگی را با هم مقایسه کنید. حال، دو بردار ویژگی LPC و MFCC را با هم متصل (Concatenate) کرده و از آن به عنوان یک بردار برای هر فریم استفاده

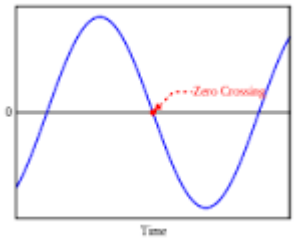
h.veisi@ut.ac.ir

به نام خدا

کنید و نتیجه را دوباره بدست آورید. به بردارهای [mfcc;lpc] ویژگی نرخ عبور از صفر (ZCR: zero-crossing rate) را اضافه کنید و نتایج را دوباره بدست آورید. ZCR بیانگر تعداد بارهایی است که سیگنال تغییر علامت داده است (مقدار نمونه‌های آن از مثبت به منفی یا برعکس تغییر کرده است) (شکل‌های زیر). نحوه محاسبه این معیار برای سیگنال x[n] با طول N به صورت زیر است.

$$ZCR = \sum_{n=1}^{N} |Sign(x[n]) - Sign(x[n-1])|; \qquad Sign(x[n]) = \begin{cases} 1 & if\ x[n] \geq 0 \\ -1 & if\ x[n] < 0 \end{cases}$$





لذا در نهایت برای این سوال جدول زیر را تکمیل کنید که در آن مقدار هر سلول بیانگر درستی (Accuracy) روی داده آزمون است. برداشت (تحلیل) خود را در مورد نتایج و کارایی روش‌های مورد استفاده بیان کنید.

ویژگی	KNN (K=7)	SVM- Linear	SVM-Poly
MFCC			
LPC			
[MFCC;LPC]			
[MFCC;LPC; ZCR]			

h.veisi@ut.ac.ir