به نام خدا

درس: پردازش سیگنال‌های دیجیتال

استاد: دکتر آرش امینی

گزارش پروژه نهایی درس

سیّدمحمّدامین منصوری طهرانی

۹۴۱۰۵۱۷۴

بخش اول

قسمت اول:

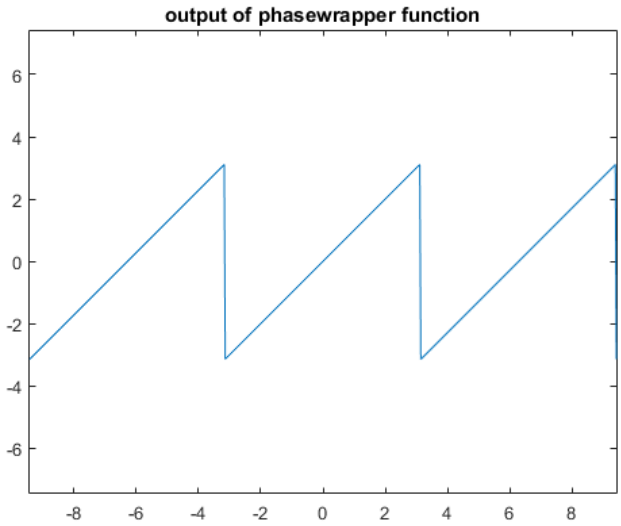
1. این عملکرد با دستور باقی‌مانده پیاده‌سازی می‌شود. در واقع اگر باقی‌مانده با 2π گرفته شود همه زوایا به بازه صفر تا 2π می‌رود. حال برای رفتن به بازه منفی پی تا پی، ابتدا هر زاویه‌ای با π جمع شده، باقی‌مانده آن با 2π گرفته می‌شود و نتیجه منهای π می‌شود. شکل خروجی تابع نیز در ۴,۰ آورده می‌شود.

۲. تبدیل فوریه به سادگی با fft گرفته شده و سپس فرکانس صفر آن در مرکز بردار قرار می‌گیرد.(به کمک fftshift) برای تولید زاویه تصادفی نیز خروجی‌های رندم که بین صفر و یک هستند را در 2π ضرب کرده و نتیجه‌ها را منهای π می‌کنیم و به این ترتیب به بازه مورد نظر نگاشته می‌شود.

توجه: در قسمت ifft گرفتن و بازسازی نتیجه، به اشتباه اول ifft گرفته شده بود و بعد ifftshift در حالی‌که ابتدا باید ifftshift گرفته شود تا فرکانس‌هایی که wrap شده‌اند به مکان اصلی بازگردند و سپس از ifft استفاده شود. نهایتاً برای پخش نیز از قسمت real آن استفاده می‌شود. این تغییرات در کد ایجاد شده است.

۳. مانند قسمت قبل تبدیل فوریه به سادگی محاسبه شده و فقط فاز آن را صفر می‌گذاریم. تبدیل فوریه وارون آن‌را گرفته و از قسمت حقیقی نتیجه استفاده می‌کنیم.(سیگنال صوت حقیقی است.)

۴,۰.



۴,۱. خروجی برای سیگنالی کوتاه یعنی basket ضمیمه شده‌است تا حجم فایل ارسالی زیاد نباشد. علت خش‌دار شدن صدا تاخیر گروه است. با اضافه شدن فاز رندم تاخیر گروه که مشتق فاز است تغییرات نسبتاً زیادی(بسته به ضریب hoarsening) می‌کند و باعث جابجایی ترتیب فرکانس‌های مختلف می‌شود و اگر

کم باشد این تاخیر چند نمونه اندک خواهد بود و نتیجه را فقط اندکی نامطلوب کرده و اگر زیادتر شود بین فرکانس‌های مختلف حروف تفاوت در پخش شدن بیشتر شده و احساس خش‌دار شدن بیشتری پیدا می‌کنیم. نتیجه به ازای ضریب ۰,۳ و ۰,۸ در خروجی ضمیمه شده‌است.

۴,۲. نتیجه به ازای ضریب طول پنجره ۴ و ۱۶ آورده شده‌است. از ضریب طول پنجره ۱ که شروع کنیم، طول پنجره نسبتاً کوچک است و اثر صفر کردن فاز یا معادلاً صفر کردن تاخیر گروه این خواهد بود که فرکانس‌های متفاوت که باید در زمان‌های متفاوت پخش شوند، در بازه‌های کوتاه به طول پنجره، همگی از یک زمان شروع می‌شوند و صدا مانند صداهای زیادی است که با هم شنیده می‌شوند یا در واقع رباتیک است. هرچه طول این پنجره بیشتر می‌شود، فرکانس‌های متفاوتی که مربوط به بخش بلندتری از صوت هستند با هم پخش می‌شوند و اثر این به هم ریختگی ترتیب بیشتر و بیشتر مشاهده می‌شود تا جایی‌که باعث می‌شود صوت نتیجه قابل فهم نباشد.

قسمت دوم:

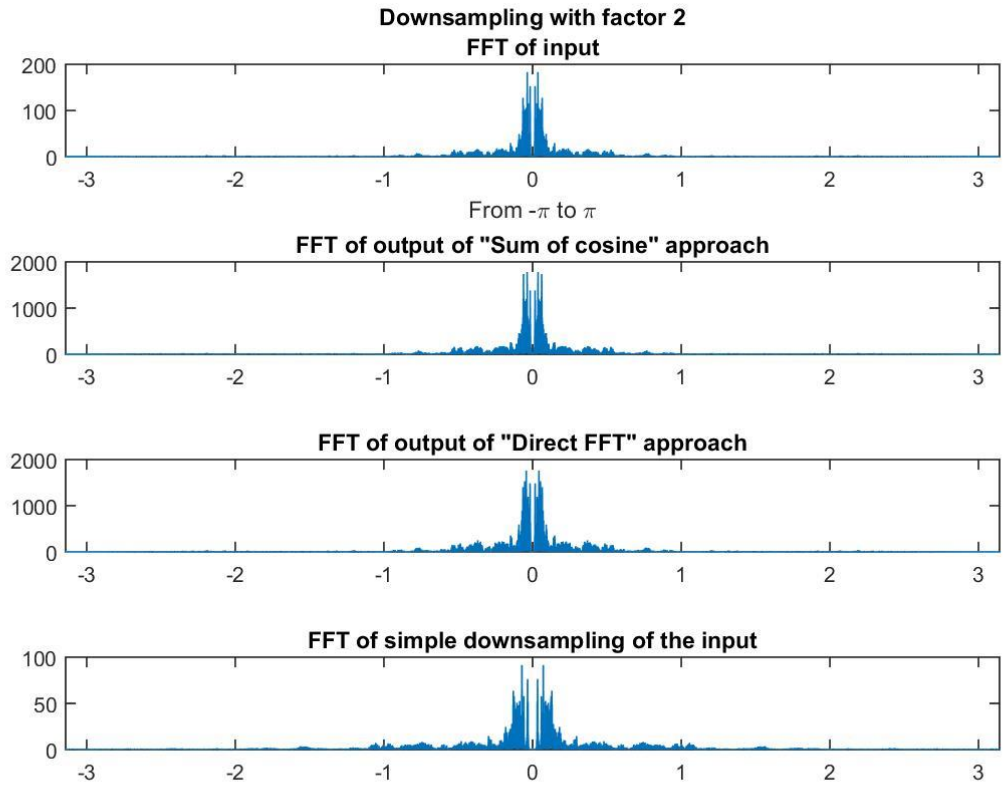
۵. چیزی برای پاسخ دادن وجود ندارد. کد این قسمت به سادگی در تابع پیوست نوشته شده‌است.

۶. چیزی برای پاسخ دادن وجود ندارد.

۷,۱. نتایج به پیوست ضمیمه شده‌اند. در ۳ حالت اول هدف مورد نظر به طرز بسیار بدی محقق شده‌است. در واقع صدای گوینده به طور کامل تغییر کرده است. هدف ما تغییر سرعت بدون تغییر فرکانس‌هایی است که حاوی محتوا هستند. در حالت استفاده از Direct fft این هدف با ضرب کردن اختلاف فازهای پنجره‌های مجاور در ضرایبی باعث ضرب شدن تاخیر گروه در ضریب‌هایی شده و به این ترتیب بر طبق خواسته ما پنجره‌های متوالی یا با فاصله کمتری نسبت به هم یا فاصله بیشتر پخش می‌شوند. اما فرکانس‌ها تغییری نمی‌کند. در حالت downsample همان‌طور که از درس می‌دانیم محدوده فرکانسی بزرگ می‌شود و این معادل زیرتر شدن صدا است که در صداهای پیوست شده مشاهده می‌شود. در ادامه همان صوت نتیجه upsample هم موجود است که می‌دانیم بازه فرکانسی را کوچکتر کرده که معادل بم شدن صدا است. باز هم این در نتیجه مشاهده می‌شود و صدا تغییر فاحشی کرده است. اگر از up2 استفاده کنیم به دلیل کپی کردن مقادیر ورودی به جای صفر گذاشتن، خروجی دارای انرژی بیشتری بوده و می‌توان با گوش دادن به صوت مشاهده کرد که کمی نسبت به صوت دومی که پخش می‌شود بلندتر است. اما از لحاظ فرکانسی همان اتفاق بد می‌افتد.(تغییر صدای شخص)

۷,۲. چیزی برای گزارش کردن وجود ندارد. فایل پیوست شده‌است.

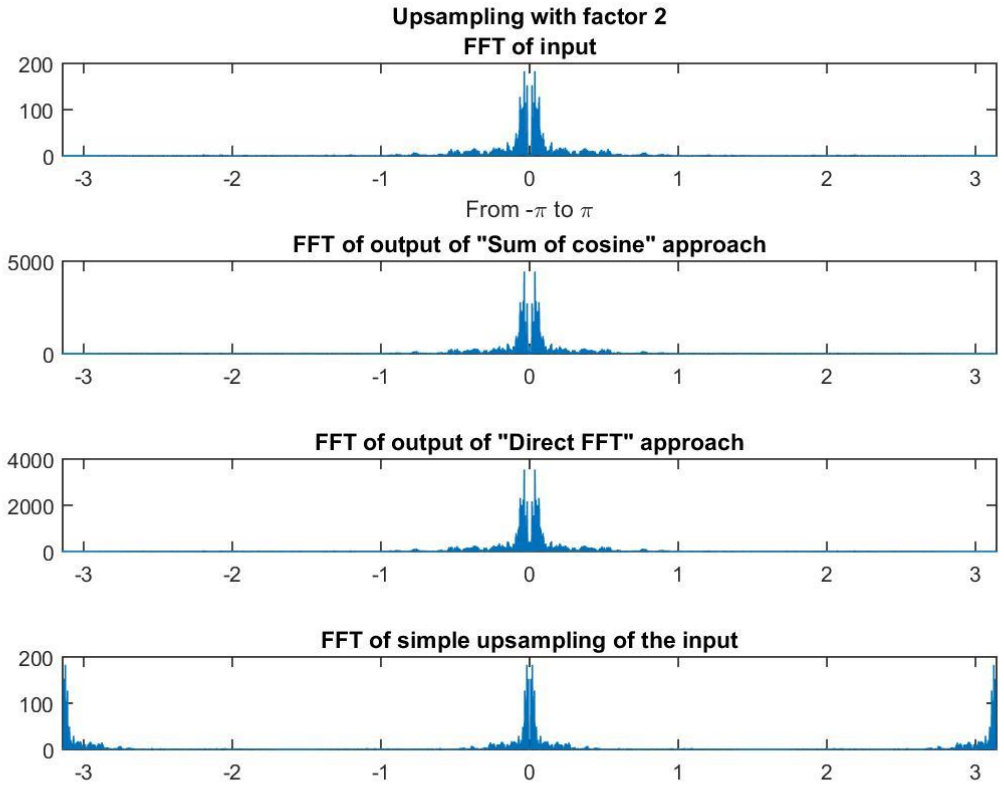
۷,۳. نتیجه در نمودارهای زیر گذاشته شده‌است.



در شکل فوق که به افزایش سرعت پخش مربوط است مشاهده می‌شود در بدترین روش که کم نمونه‌برداری است، از هر دو نمونه یک نمونه برای پخش انتخاب می‌شود و به نظر روش ساده‌ای برای زیاد کردن سرعت است اما با دقت بیشتر متوجه می‌شویم طیف خروجی حاوی فرکانس‌های بیشتر از ورودی است و در حوزه فرکانس گسترده‌تر شده(چیزی که از downsample انتظار داریم.) که طبق استدلال‌های قبلی که گفته شد باعث می‌شود صدای شخص را عوض کند و صدای زیری بشنویم(خیلی بچه‌گانه)(جمله همان شنیده می‌شود.) در روش Direct fft همان‌طور که قبل‌تر گفته شد تاخیر گروه دچار تغییر شده و به این روش صداها زودتر یا دیرتر در فاصله پنجره‌های متوالی پخش می‌شوند و نکته مهم تغییر نکردن فرکانس‌های اصلی یه بازه بزرگتر یا کوچکتر است که باعث می‌شود صدای شخص حفظ شود. روش جمع کسینوس‌ها نیز که خواسته نشده‌است ولی در حد روش direct fft خوب است و مطابق تصویر بالا شکل طیف و فرکانس‌ها را حفظ می‌کند و تنها تفاوت آن شنیده شدن سوت‌های کم در بعضی صداها است.

در تصویر زیر نیز که به بیش نمونه‌برداری مربوط است مشاهده می‌شود که با بیش نمونه‌برداری ساده هم محتوای فرکانسی جمع شده(صدا بم‌تر می‌شود) و هم صداهایی در فرکانس‌های بالا به وجود آمده‌اند.(البته چون در حدود fs/2 هستند شنیده نمی‌شوند و مشکلی ایجاد نمی‌کنند.) در واقع در نگاه اول بین نمونه‌ها فاصله گذاشتن به نظر می‌رسد سرعت پخش را کند می‌کند اما این صفر بودن نمونه‌های میانی در تبدیل فوریه اثرات مخربی دارد که ذکر شد و در تصویر مشاهده می‌شود. باز هم مثل قبل در روش direct fft اندازه طیف

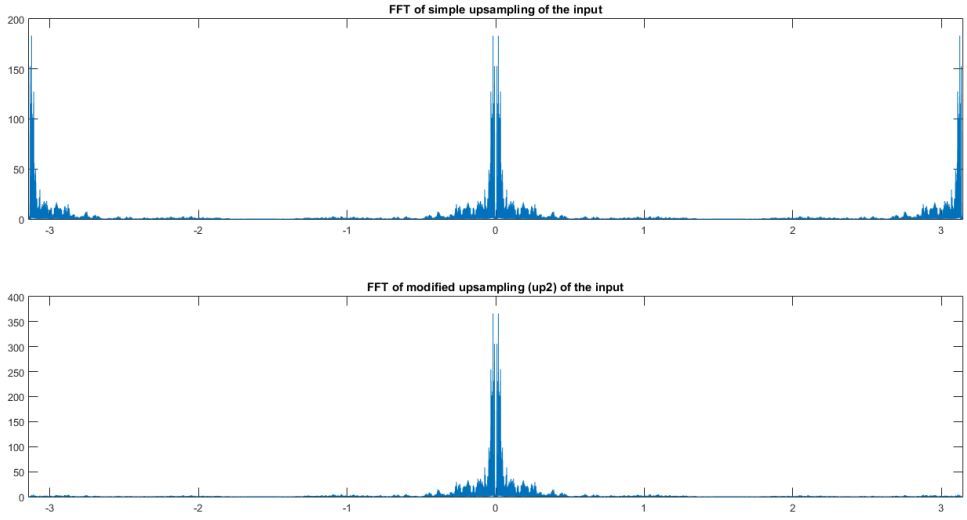
در فرکانس‌های خودش تغییر زیادی نکرده و تاخیرهای گروه باعث تغییر در زمان پخش بین فریم‌های متوالی و به تبع آن تغییر سرعت پخش می‌شوند و تفاوت کیفیت‌ها از همین حفظ کردن اندازه‌ها در فرکانس‌های واقعی ناشی می‌شود. در روش جمع کسینوس نیز نمونه‌های بین دو نمونه‌ای که مثلاً برای کند کردن سرعت تولید می‌شوند به جای این‌که صفر باشند یا کپی مقدار قبل باشند، به نحوی از تبدیل فوریه وارون بدست می‌آیند(تغییرات بین فریم‌ها به جای ناگهانی شدن پله پله صورت می‌گیرد و تبدیل وارون آن در تخمین نمونه‌های میانی اثر خود را می‌گذارد.)



۷,۴. نتیجه در تصویر زیر مشاهده می‌شود. تفاوت در فرکانس‌های بالای آن‌ها است که اگر از up2 استفاده کنیم در فرکانس‌های نزدیک $\pm\pi$ خروجی صفر می‌شود و در فرکانس‌های نزدیک صفر دامنه دو برابر حالت upsample معمولی است. این اتفاق به سادگی با روابط زیر قابل توجیه است.

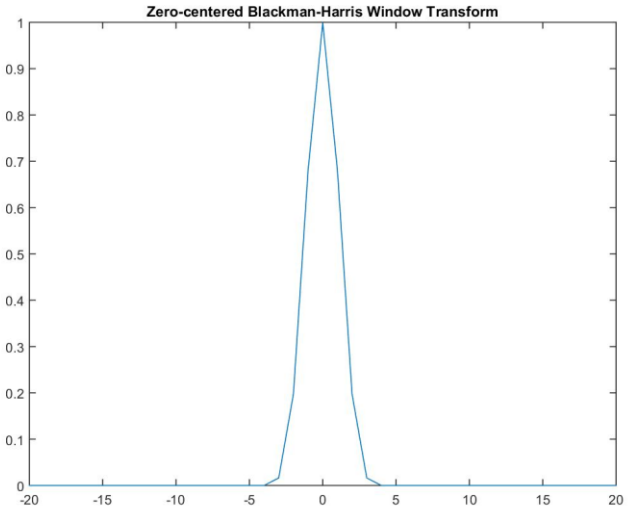
$$up2(n) = y(n) + y(n-1), y(n) = upsample(input, 2), input(n) = x(n)$$
$$UP2(z) = Y(z) + Y(z)z^{-1}, z = e^{j\omega} \Rightarrow UP2(z) = \left(1 + e^{-j\omega}\right)Y(z)$$

که مشخص است رابطه همان رابطه تبدیل فوریه upsample ساده است که در فرکانس‌های بالا($e^{\pm j\pi} = -1$) خروجی را صفر می‌کند. در فرکانس‌های نزدیک صفر نیز($e^{j0} = 1$) خروجی را تقریباً دو برابر می‌کند که این هم در شکل نمایان است.



بخش دوم

۱. در دستور کار ضرایب برخلاف ضرایب blackmanharris در متلب همه علامت مثبت دارند و احتمالاً اشتباه تایپی بوده و به همین دلیل ما از ضرایب متلب که یکی در میان مثبت و منفی‌اند استفاده می‌کنیم. برای جدا کردن آن بخشی از طیف که معنا دار است(به وسیله x یا bin position) ابتدا با fftshift فرکانس صفر را به مرکز می‌آوریم تا zero-centered شود و bin position را با N/2+1 جمع می‌کنیم که [-20,20] به وسط طیفی که از fftshift بدست آمده نگاشته شود. به این ترتیب می‌توانیم قسمت معنادار آن را جدا کنیم. دقت می‌کنیم که اندازه آن را برداریم و فاز ناشی از این جابجایی مبدأ را در نظر نگیریم چون blackman harris زوج و حقیقی است و تبدیل فوریه آن نیز زوج و حقیقی است.



۲. در این قسمت کار ما فقط تولید کردن دامنه‌ها است و به سادگی با تبدیل مقادیر داده شده برحسب دسیبل به مقادیر اصلی و ضرب آن‌ها در تبدیل فوریه پنجره طیف پنجره ضرب شده در طیف سینوسی‌ها را می‌سازیم. اضافه کردن فاز در ادامه کد نوشته‌شده توسط ما انجام می‌شود. تغییر کوچکی نیز در کد داده شده تا سرعت اجرا زیاد شود. به جای call شدن با دفعات زیاد bh92transform داخل حلقه، یک بار بیرون حلقه فراخوانی می‌شود و در دفعه‌های بعد استفاده می‌شود.

۳. کد این قسمت در کامنت‌ها توضیح داده شده‌است. تنها نکته‌ای که لازم است اشاره شود محاسبه d است. در حالتی که از ضرب معمولی برای توان ۲ استفاده شود برای صدای مسخ این زمان اجرا به ۱۲۰ ثانیه نیز می‌رسد و برای کاهش آن از کانولوشن استفاده می‌کنیم.

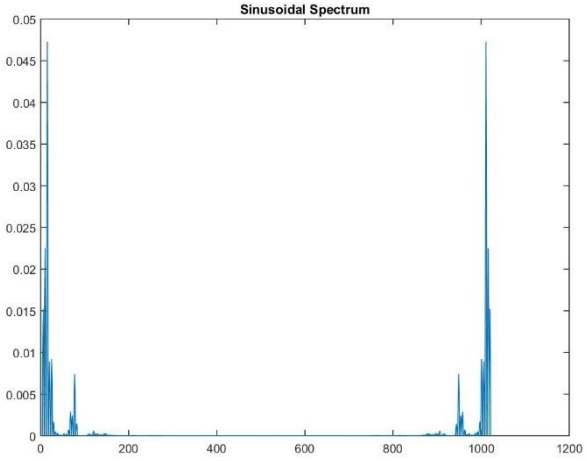
$$\sum\big(x(j)-x(j+\tau)\big)^2=\sum x(j)^2+\sum x(j+\tau)^2-2\sum x(j)x(j+\tau)$$

جمله اول توسط summation constant در کد اجرا شده. جمله دوم به وسیله کانولوشن با برداری از یک‌ها به طول ws با ورودی از اندیس ۱ تا maxlag+ws انجام می‌شود. با دادن پارامتر سوم به مقدار 'val' فقط خروجی در نقاطی حساب می‌شود که دو بردار با هم اشتراک دارند.(قسمت‌هایی که با zero pad حساب می‌شوند در خروجی داده نمی‌شوند.) به این ترتیب مجموع مربعات پنجره‌هایی به طول ws که هر دو تای آن‌ها فقط در یک مولفه با هم تفاوت دارند و بقیه المان‌هایشان overlap دارد در بردار حاصل از این کانولوشن ذخیره می‌شود. جمله سوم جمع بالا نیز کانولوشن با قرینه ورودی است که با دستور flip و پارامتر val دوباره همین کار انجام شده‌است. در ادامه کد نیز d2 محاسبه شده و محدودیت‌ها و شرط‌هایی که در صورت پروژه و کامنت‌های ابتدایی تابع گفته شده اعمال شده‌اند.(هر شرط با کامنت مربوط به خودش توضیح داده شده.)

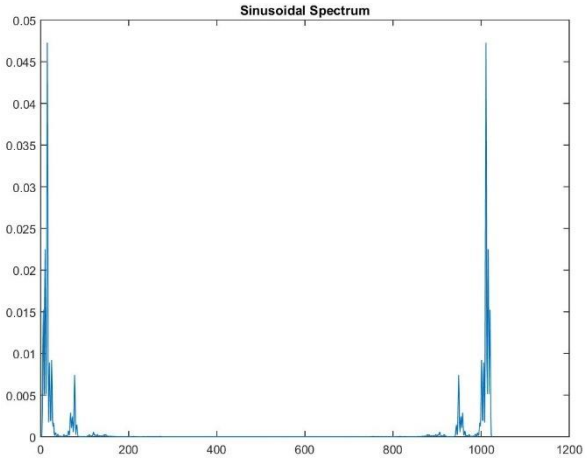
۴.

الف) همان‌طور که در سوال ۱ همین بخش مشاهده می‌شود قسمت اعظم انرژی آن حول صفر متمرکز است. تا ۴ اندیس قبل و بعد از صفر تبدیل فوریه غیر صفر است که معادل $4\frac{2\pi}{N}$است که معادل حدود ۱۷۰ هرتز است.(N=1024) البته محتوا در این فرکانس کم است.

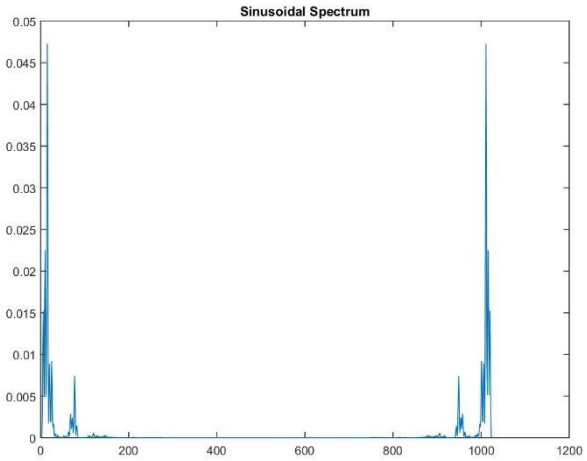
ب) نتیجه به ازای $bn = 1$:



نتیجه به ازای $bn = 4$:



نتیجه به ازای $bn = 10$:



با اضافه شدن طول bin هایی که در محاسبه وارد می‌شوند تعداد محاسبات بیشتر شده ولی در این اجرا تغییر محسوسی مشاهده نمی‌شود ولی برای صوت‌های با طول زمان بیشتر ۴ را انتخاب می‌کنیم چون بعد از آن مقدار تبدیل صفر است و نتیجه بهتری هم در طیف خروجی نمی‌دهد و این امر از مقایسه دو تصویر آخر در تصاویر بالا تایید می‌شود. پس برای کاهش دادن محاسبات بیهوده ۴ را برمی‌گزینیم. مشکل ۱ این است که خیلی تقریب مناسبی نیست و تغییرات شدید دارد(بین صفر و پیک‌ها تغییر شدید دارد.) پس اجرای بهینه با bn=4 منطقی به نظر می‌رسد.

پ) فرکانس‌های بدست آمده به شرح زیراند:

طول پنجره	فرکانس پایه(هرتز)
0.0125	240.482677
0.0100	240.458094
0.005	240.476499
0.003	0.000000

با کاهش طول پنجره‌ای که d در آن حساب می‌شود مقدار lag ای که باید داشته باشیم تا به یک دوره تناوب و تناوب‌های بعد برسیم زیاد می‌شود و به همین علت چون تناوب پیدا نمی‌شود خروجی را صفر می‌دهد. در سه حالت اول دوره تناوب پایه حدود ۴ میلی‌ثانیه بوده و طول پنجره ما چند برابر آن است و بنابراین با lag می‌توان تغییرات d2 را مشاهده کرد ولی وقتی طول پنجره از تناوب کمتر است تناوب اصلی یافت نمی‌شود چون تعداد شیفت‌ها برای رسیدن به تناوب زیاد است و از حد تعیین شده بیشتر می‌شود(maxlag) پس نتیجه صفر می‌شود.

توجه: تغییرات خواسته شده را به خوبی با کلمات نمی‌توان توصیف کرد! با اجرای کد می‌توانید به صحت آن‌ها پی ببرید.

ت) صوت نتیجه ضمیمه شده‌است. با تغییر fscale به مقادیر بیشتر صدا زیرتر شده(به سمت بچه‌گانه‌تر شدن می‌رود) و با کم کردن آن صدا بم‌تر و مردانه‌تر می‌شود. البته اگر خیلی کم شود صدا به صدای مرد هم شبیه نیست.

با زیاد کردن محدوده timbre برای خروجی(رسیدن به ۶۰۰۰)، صدا به فرکانس‌های زیرتر نگاشته شده و بچه‌گانه‌تر می‌شود. با کاهش آن (حدود ۳۰۰۰) صدا نسبتاً مردانه است و حالت تو دماغی(!!) دارد.

ث) صوت نتیجه ضمیمه شده است. با کمتر کردن fscale صدا خش‌دار می‌شود. با زیاد شدن fscale صدا به صدای اصلی زن باز می‌گردد. با زیادتر کردن fscale حتی از صدای اصلی زن نیز زیرتر شده و اندکی تو دماغی(!!) می‌شود.

اگر محدوده نگاشت خروجی را کم کنیم(حدود ۳۰۰۰) صدا بم‌تر شده و حالت گرفته و خسته پیدا می‌کند. اگر محدوده فوق زیاد شود(۶۰۰۰) صدا زیر شده و به شدت تو دماغی می‌شود.

ج) صوت نتیجه ضمیمه شده است. با زیاد کردن fscale صدا زیرتر و بچه‌گانه‌تر می‌شود. طبیعتاً تغییر در جهت برعکس نتیجه برعکس می‌دهد و صدا نسبت به صدای اصلی بچه‌تر است ولی خیلی زیر نیست. با زیاد کردن محدوده نگاشت خروجی(۷۰۰۰) صدا بسیار بچه‌گانه می‌شود. با کم کردن آن(۳۰۰۰) مطابق انتظار صدا به مردانه شدن نزدیک می‌شود.

صداها توصیف‌کردنشان سخت است! لطفاً به این مورد توجه داشته باشید!