دانشکده فنی دانشگاه تهران

دانشکده برق و کامپیوتر

پروژه ۳ پردازش سیگنال‌های زمان گسسته

Auditory Scene Analysis

طراح:
سجاد پاکدامن ساوجی

رایانامه
sj.pakdaman@ut.ac.ir

نیم سال دوم ۹۸-۹۹

دانشجویان عزیز، قبل از پاسخ‌گوئی به سوالات به نکات زیر توجه کنید:

۱. شما باید کدها و گزارش خود را با الگو DSP_CA3_StudentNumber.zip در محل تعیین شده آپلود کنید.

۲. گزارش کار معیار اصلی ارزیابی خواهد بود، در نتیجه زمان کافی برای تکمیل آن اختصاص دهید.

۳. گزارش خود را حتما در قالب قرار گرفته در صفحه درس بنویسید.

۴. پیاده سازی های خود را در محیط MATLAB انجام دهید و کد های خود را همراه با گزارش ارسال کنید.

۵. شما میتوانید سوالات خود را از طریق ایمیل sj.pakdaman@ut.ac.ir بپرسید.

منابع:

۱. A. Bregman, Auditory Scene Analysis, Cambridge, MA: MIT Press, 1990.

۲. R. Duda, “Modeling Head Related Transfer Functions,” IEEE Proceedings of ASILOMAR 1993.

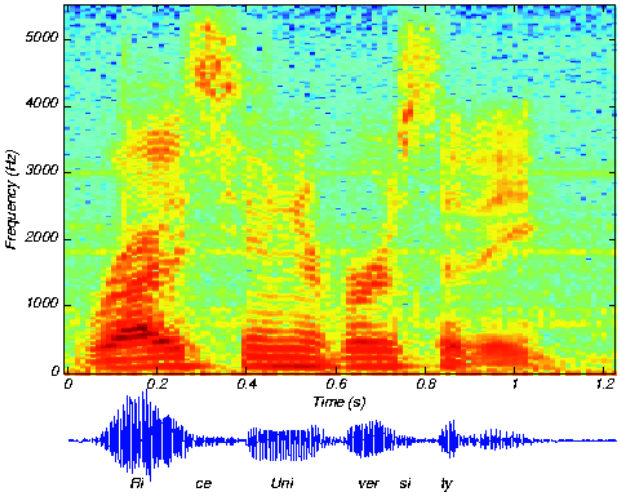
۳. E. Tessier and F. Berthommier, “Speech Enhancement and Segregation Based on the Localisation Cue for Cocktail-Party Processing,”

سیستم شنوایی انسان قادر به انجام اعمال پیچیده‌ای است که تعداد زیادی از آن‌ها در مهندسی کاربرد دارند. یکی از این توانایی ها تقویت انتخابی منابع صوتی است که به شنونده این امکان را می‌دهد که در مکانی با نویز های متنوع به سیگنال دریافتی از یک منبع خاص توجه کند. این پدیده با نام Auditory Scene Analysis شناخته می‌شود.

در این تمرین روشی برای تفکیک صوت یک گوینده دلخواه از طریق binaural recording همزمان از چند گوینده معرفی و پیاده‌سازی می‌شود. بدیهی ترین کاربرد سیستم Auditory Scene Analysis(ASA) در پیش‌پردازش صوت برای speech recognition در unconstrained auditory environments می‌باشد که با استفاده از این سیستم ابتدا صوت یک گوینده تفکیک شده و سپس با سیستم دیگری گوینده بازشناسی می‌شود. [۱]

Spectrogram

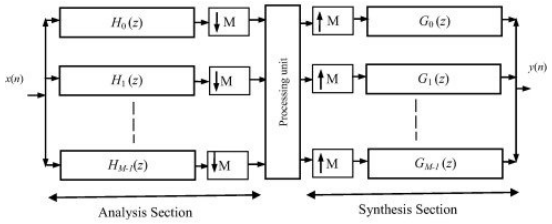
برای تحلیل (و نمایش) زمان–فرکانس سیگنال های صوتی از Spectrogram استفاده می‌شود. این ابزار بر پایه تبدیل Short-Time Fourier Transform استوار است که در آن سیگنال مورد نظر به قطعه‌هایی (chunks) تفکیک می‌شود و تبدیل فوریه هر قطعه جداگانه محاسبه‌می‌شود. با انجام این کار ویژگی‌های فرکانسی یک سیگنال به صورت محلی (در حوزه زمان) بدست خواهد آمد. به این معنی که خروجی هر تبدیل فوریه مشخصات فرکانسی سیگنال را در یک بازه زمانی خاص نشان خواهد داد. برای نمایش تمامی این مشخصه های فرکانسی (این داده دارای ۳ بعد است: ۱ بعد برای زمان، یک بعد برای فرکانس و یک بعد برای اندازه تبدیل فوریه) از نمایش تصویری استفاده می‌شود. برای بدست آوردن اطلاعات بیشتر به اینجا و اینجا مراجعه کنید.



شکل ۱: یک نمونه از spectogram

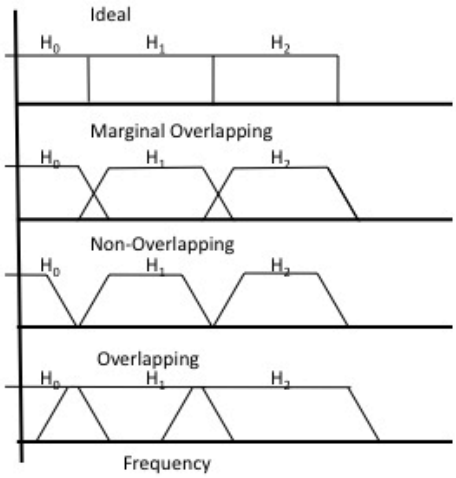
Filter Bank

به یک مجموعه M تایی از فیلترهای Analysis که به همراه یک دیگر M سیگنال خروجی تولید می‌کنند، Analysis Filter Bank گفته می‌شود. همچنین به یک مجموعه K تایی از فیلترهای Synthesis که به همراه یکدیگر ۱ سیگنال خروجی تولید می‌کنند، Synthesis Filter Bank گفته می‌شود. معمولا هر فیلتر در بانک–فیلتر به گونه ای طراحی می‌شود که خروجی آن یک سیگنال band limited باشد که سیگنال های narrow band حالت خاص آن‌ها می‌باشند. برای اطلاعات بیشتر به اینجا مراجعه کنید.



شکل ۲: تجزیه وتحلیل سیگنال گسسته با استفاده از filterbank

ساده ترین **filter bank** آن است که سیگنال را به بازه های فرکانسی مساوی تقسیم می‌کند. این **filter bank** معادل انجام موازی ۱ فیلتر پایین‌گذر، ۱ فیلتر بالاگذر و چندین فیلتر میان‌گذر است که این فیلتر ها پهنای باند گذر مساوی دارند. فیلتر های یاد شده می‌توانند در حوزه فرکانس اشتراک داشته باشند و یا می‌توانند اشتراک نداشته باشند (در صورتی که فیلتر ها در حوزه فرکانس اشتراک دارند باید فیلتر های **synthesis** به گونه‌ای طراحی شود که سیگنال بازیابی شده حداقل خطا را داشته باشد.) شکل زیر نمونه ای از این نوع فیلتر است.



شکل ۳: نمونه‌ای ساده از بانک-فیلتر تجزیه‌گر فرکانسی

The HRTF

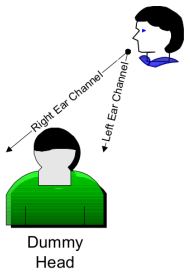
بنای تئوری استفاده شده برای تفکیک منبع صوتی بر این نکته استوار است که موج انتشار یافته از منبع مسیر های متفاوتی برای رسیدن به گوش سمت و چپ و سمت راست طی می‌کند و در نتیجه فیلترینگ های مختلفی بر روی هر کانال اتفاق می‌افتد. این حالت را با معادلات زیر مدل می‌کنیم،

$$Y_{left}(w,t) = H_{left}(w,\phi,\theta)\, X(w,t)$$

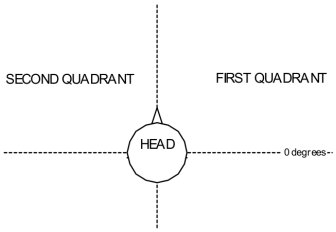
$$Y_{right}(w,t) = H_{right}(w,\phi,\theta)\, X(w,t)$$

که در آن $X(w,t)$ سیگنال اولیه، $Y_{left}(w,t)$ و $Y_{right}(w,t)$ سیگنال های دریافتی از سمت چپ و سمت راست هستند. $H_{left}(w,\phi,\theta)$ و $H_{right}(w,\phi,\theta)$ نیز پاسخ کانال های سمت چپ و سمت راست هستند.

این توابع با نام Head Related Transfer Functions شناخته می‌شود. در معادلات بالا فرض شده است که کانال‌های مختلف با استفاده از زوایای ϕ و θ در دستگاه مختصات استوانه‌ای از یک دیگر تفکیک می‌شوند. [۲]

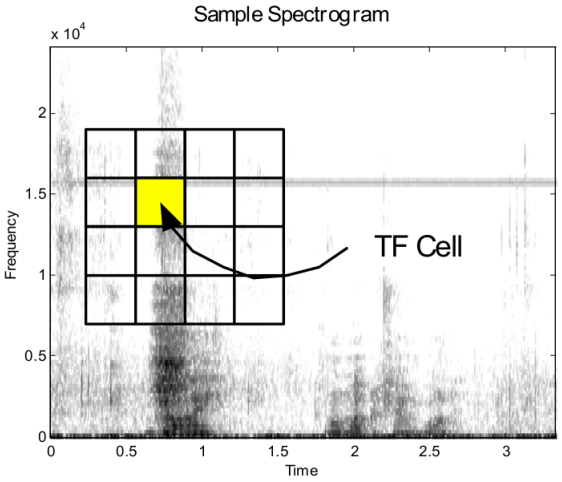


(آ) تفاوت کانال های چپ و راست



(ب) دستگاه مختصات استوانه‌ای در معادلات

فرضیه ای که در این تمرین آن را بررسی می کنیم آن است که با تجزیه سیگنال حاوی چند منبع صوتی به Time-Frequency Cell ها و با اعمال وزن مناسب بر هر یک از این سلول ها می‌توان سیگنال حاوی یک منبع صوتی را بازیابی کرد [۳]. در ادامه روشی بر مبنای ویژگی های خاص سیگنال موجود برای تعیین وزن های هر سلول ارائه می‌شود.



شکل ۵: یک سلول زمان-فرکانس در spectogram

The Time Delay of Arrival

تفاوت مکانی که بین مسیر انتقال سیگنال پیام تا دو گوش (دو گیرنده در حالت کلی) وجود دارد، باعث تفاوت فازی در دریافت های چپ و راست می‌شود. این واقعه می‌تواند سرنخی برای تعیین وزن های سلول‌های زمان-فرکانس باشد. به صورت تئوری اختلاف فاز ناشی از دو دریافت از طریق رابطه زیر بدست می‌آید.

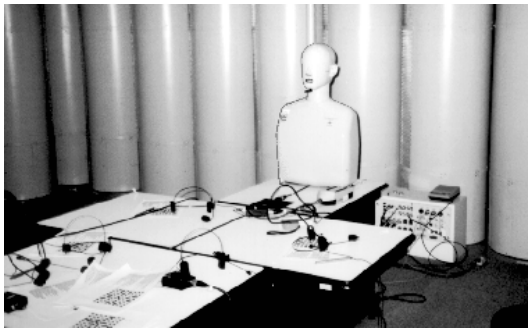
$$itd = arg\{Y_{left} \, . \, Y^*_{right}\}$$

به علت غیر خطی بودن تابع فاز، از استفاده از آن صرف‌نظر می‌کنیم و از روش جایگزینی بر مبنای **correlation** استفاده می‌کنیم. در این روش تاخیر میان هر **chunk** از داده با استفاده از **correlation** محاسبه می‌شود و ضرایبی بر مبنای این **lag index** به سلول زمان-فرکانس آن اختصاص داده می‌شود. **lag index** را معادل اندیس بزرگ‌ترین المان در **correlation** میان سیگنال دریافتی از کانال چپ و سیگنال دریافتی از کانال راست در نظر می‌گیریم.

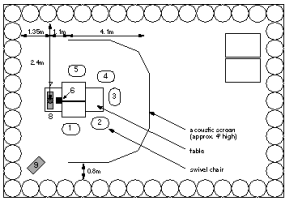
The ShATR Corpus

مجموع داده‌ای که در این تمرین از آن استفاده میکنیم **The ShATR Multiple Simultaneous Speaker Corpus** نام دارد. در این مجموعه داده ۵ گوینده حضور دارند و مکالمه آن ها از طریق ۸ گیرنده ضبط شده است. ۵ گیرنده در جلوی هر یک از گویندگان وجود داشته است، ۲ گیرنده همانند گوش چپ و گوش راست عمل میکنند و یک گیرنده در وسط قرار گرفته است. برای اطلاعات بیشتر در مورد این مجموعه داده به اینجا مراجعه کنید.

برای پیاده سازی ها دو فایل از معرفی گوینده شماره ۳ و دو فایل از مکالمه کلی قرار داده شده است. این فایل ها از طریق گیرنده های شماره ۷ و ۸ ضبط شده اند. همچنین ۱ فایل از مکالمه اصلی که از گیرنده ۳ ضبط شده است قرار داده شده تا برای مقایسه نتیجه نهایی استفاده شود.



(آ) تصویری از چینش گیرنده‌ها



channel	component	equipment type
1 - 5	participant, wearing headset microphone	RAMSA WM-S10
6	omnidirectional (pressure zone) microphone	Crown PZM30
7,8	mannikin	B&K type 4128, with B&K ears, type 4158 (R) and 4159 (L); microphones type 4134 powered by a B&K power source type 2804.
9	video camera	ceiling-mounted

(ب) نحوه قرار گیری گیرنده‌های ۱-۸

۱. فایل های قرار گرفته در پوشه **Intro** را بارگیری کنید و **spectrogram** هر یک را رسم کنید.

۲. یک بانک-فیلتر تجزیه‌گر فرکانسی ایده‌آل طراحی کنید که طیف فرکانسی را به ۴ قسمت مساوی تقسیم کند. سپس هر یک از سیگنال های مربوط به گیرنده۷ و گیرنده۸ را از بانک-فیلتر عبور دهید تا سیگنال های نهایی بدست آیند. در پایان این قسمت باید ۸ سیگنال مجزا داشته باشید. این سیگنال ها شامل ۴ زیرکانال برای ۲ سیگنال هستند. در ادامه از هر یک از ۴ خروجی بانک-فیلتر با نام زیر کانال یاد خواهد شد.

۳. برای هر یک از زیر کانال‌ها سیگنال‌های مربوطه را (دو سیگنال، یکی برای گیرنده۷ و دیگری برای گیرنده۸) به قسمت (chunk) هایی به طول ۲۵۶ قسمت کنید

۴. برای هر یک از زیر کانال‌ها، با استفاده از chunk های متناظر و تابع correlation مقدار **lag index** را بدست آورید.

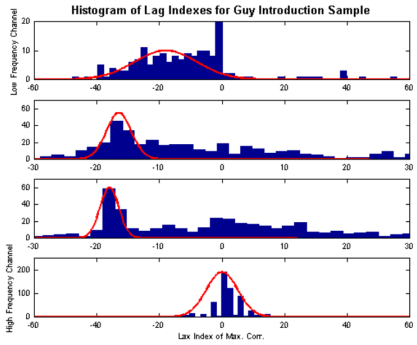
۵. برای هر یک از زیرکانال ها هیستوگرام مربوط به **lag index** را رسم کنید. پس از رسم این هیستوگرام ها مشاهده می‌شود که توزیع **lag index** شبیه نورمال است. برای هر یک از زیر کانال ها μ و σ را دستی محاسبه کنید.

۶. مراحل ۱ -۴ را برای فایل های موجود در پوشه **main** نیز تکرار کنید.

۷. در این مرحله ضریبی برای هر سلول زمان–فرکانس بر اساس **lag index** مطابق رابطه زیر بدست می‌آید.

$$w_{ch}[n] = \exp^{\frac{-(lag-\mu_{ch})^2}{2\sigma^2_{ch}}}$$

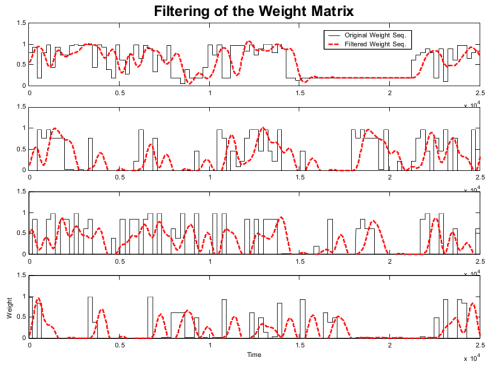
که در آن n معین شماره **chunk** است و ch معین شماره زیر کانال می‌باشد.



شکل ۷: نمونه ای از هیستوگرام تاخیر ها به همراه ضرایب تولیدی

۸. حال پس از اعمال ضریب هر یک از سلول‌های زمان–فرکانس ،سیگنال های چپ و راست را بازیابی کنید. این دو سیگنال را با نام های مناسب ذخیره کنید و در پاسخ خود ارسال کنید.

۹. صوت‌های استخراج شده را گوش دهید و با صوت موجود در پوشه **solo** مقایسه کنید. صوتهای استخراج شده دارای نویزی هستند که علت آن غیر پیوستگی ضرایب $w_{ch}[n]$ می‌باشد. با استفاده از یک فیلتر پایین گذر دلخواه سیگنال‌های $w_{ch}[n]$ را نرم تر کنید.
نمودار $w_{ch}[n]$ را برای یک ch دلخواه پیش و پس از فیلترینگ رسم کنید.



شکل ۸: نمونه ای از تغییر ضرایب با استفاده از فیلترینگ

۱۰. حال فرایند بازیابی را با وزن های جدید مجددا انجام دهید. نتیجه را با نام مناسب ذخیره کنید و به همراه پاسخ خود ارسال کنید. همچنین با گوش دادن به صوت بازیابی شده آن را با صوت **solo** مقایسه کنید.

- موفق باشید