1– مقدمه

اهداف درس:

آشنایی با مفاهیم پایه ای پردازش سیگنال

2– مبانی سیگنال ها و سیستم ها

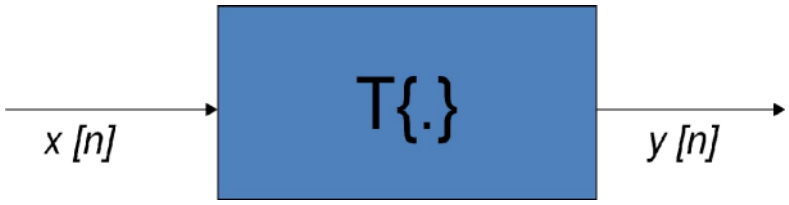
نمایش ریاضی سیگنال ها به صورت توابعی از یک یا چند متغییر مستقل می باشد.

پردازش سیگنال دیجیتال با تبدیل های سیگنال هایی که هم در زمان و هم در دامنه گسسته هستند سروکار دارد.

نمایش ریاضی سیگنال های زمان–گسسته به صورت دنباله ای از اعداد می باشد.

نمایش ریاضی یک سیستم زمان–گسسته به صورت یک تبدیل یا عملگر می باشد:

- $y[n] = T\{ x[n] \}$



تصویر 1 – یک سیستم زمان–گسسته

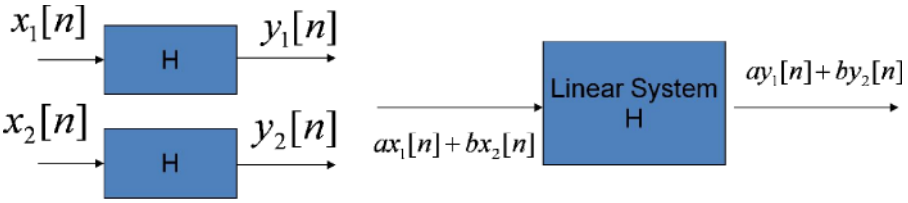
دسته «سیستم های خطی» بوسیله دو اصل زیر تعریف می شود:

- اصل جمع پذیری : $T\{x_1[n]+x_2[n]\} = T\{x_1[n]\} + T\{x_2[n]\} = y_1[n] + y_2[n]$
- اصل همگنی: (a یک مقدار ثابت است) $T\{ax[n]\} = aT\{x[n]\} = ay[n]$

دو اصل بالا را می توان در اصل superposition جمع کرد:

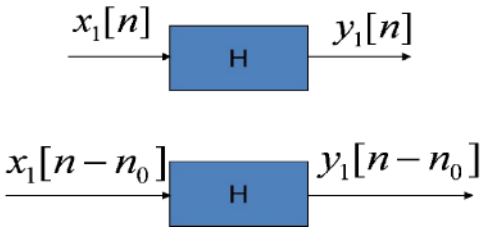
$$T\{ax_1[n]+bx_2[n]\} = aT\{x_1[n]\} + bT\{x_2[n]\}$$

به عبارتی اگر یک سیستم H داشته باشیم و دو بار مختلف ورودی های مختلف به آن بدهیم، اگر ورودی ها را ضرب در عدد ثابتی کرده و با هم جمع کنیم، خروجی هم ضرب در همان عدد ثابت و با هم جمع می شود (تصویر 2).



یک سیستم « نباله ورودی،

همان شیفت



تصویر 3 – سیستم تغییرناپذیر با زمان

یک دسته مهم سیستم ها، سیستم های «خطی و تغییرناپذیر با زمان» هستند (LTI).

سیستم های LTI را می توان کاملاً بوسیله پاسخ ضربه آن ها توصیف کرد: $y[n]=T\left\{\sum_{k=-\infty}^{\infty}x[k]\delta[n-k]\right\}$

$y[n]=\sum_{k=-\infty}^{\infty}x[k]T\{\delta[n-k]\}$

$x[n]*h[n]=h[n]*x[n]$

- اصل superposition: سیگما و x[k] از داخل بیرون می آیند.
- اصل تغییرناپذیر بودن با زمان: پاسخ ضربه در هر زمانی یکسان است.

$y[n]=\sum_{k=-\infty}^{\infty}x[k]h[n-k]$

به فرمول به دست آمده فرمول کانولوشن گفته می شود: $y[n]=\sum_{k=-\infty}^{\infty}x[k]h[n-k]$

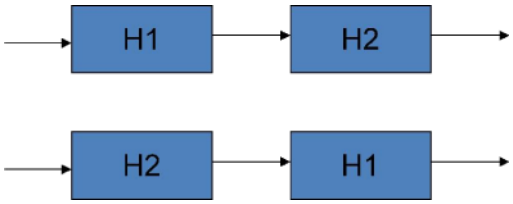
کانولوشن معمولاً به صورت روبرو نشان داده می شود:

کانولوشن دو سیگنال خواص زیر را دارد:

- جابجایی: $(h_1[n]*h_2[n])*h_3[n]=h_1[n]*(h_2[n]*h_3[n])$
- شرکت پذیری: $y[n]=x[n]*h[n]$
- پخشی: $h[n]*(ax_1[n]+bx_2[n])=a(h[n]*x_1[n])+b(h[n]*x_2[n])$
- معکوس زمان: $y[-n]=x[-n]*h[-n]$

- مستقل بودن از ترتیب: در صورتی که دو سیستم LTI پشت سر هم باشند، پاسخ ضربه کل سیستم برابر کانولوشن دو پاسخ ضربه می باشد. همچنین سیستم مستقل از ترتیب سیستم ها می باشد (تصویر 4).

$$h[n] = h_1[n] * h_2[n]$$



تصویر 4 –دو سیستم پشت سر هم LTI

یک ویژگی دیگر سیستم هاخاصیت پایدار (stable) بودن است:

یک سیستم پایدار است اگر یک ورودی کران دار (bounded) $|x[n]| < M$ | یک خروجی کران دار تولید می کند.

یک سیستم LTI پایدار است اگر $\sum_k |h[k]| < \infty$

با توجه به اندازه طول پاسخ ضربه، سیستم ها به دو دسته زیر تقسیم می شوند:

- پاسخ ضربه با طول محدود (FIR): $y[n] = b_0 x[n] + b_1 x[n-1] + \ldots + b_q x[n-q]$ که $h[n] = b_n$
- پاسخ ضربه با طول بینهایت (IIR)

تبدیل ها روش موثری برای ساده سازی تحلیل سیگنال ها و سیستم های خطی می باشند.

تبدیل های زیر را در نظر بگیرید:

- تبدیل خطی: $T[ax + by] = aT[x] + bT[y]$
- کانولوشن دو سیگنال به صورت زیر ساده می شود: $T[x * y] = T[x]T[y]$

بیشترین تبدیل های استفاده شده در مهندسی مخابرات عبارت است از:

- تبدیل لاپلاس (پیوسته در زمان و فرکانس)
- تبدیل فوریه پیوسته (پیوسته در زمان)
- تبدیل فوریه گسسته (گسسته در زمان)
- تبدیل z (گسسته در زمان و فرکانس)

3– تبدیل Z

تبدیل Z به صورت روبرو تعریف می شود:

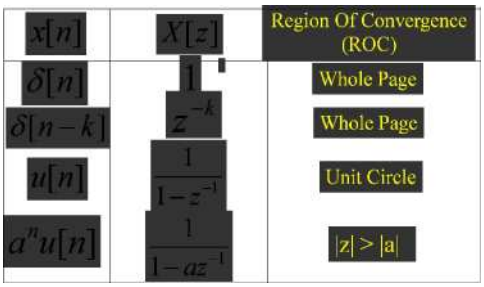
$$X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n}$$

مبحث منطقه همگرایی (Region of Convergence یا ROC) خیلی در این تحلیل مهم است.

برخی توابع پایه و تبدیل Z آن ها عبارتند از:

- ضربه واحد: $x[n]=\delta[n]$ $X(z)=\sum_{n=-\infty}^{\infty}\delta[n]z^{-n}=1$ $ROC:\ z\neq 0$
- ضربه واحد با تاخیر: $x[n]=\delta[n-k]$ $X(z)=\sum_{n=-\infty}^{\infty}\delta[n-k]z^{-n}=z^{-k}$ $ROC:\ z\neq 0$
- پله واحد: $u[n]=\begin{cases}1, & n\geq 0\\ 0, \text{ otherwise}\end{cases}$ $X(z)=\sum_{n=-0}^{\infty}z^{-n}=\frac{1}{1-z^{-1}}$ $ROC:\ z>1$
- نمایی: $x[n]=a^n u[n]$ $X(z)=\sum_{n=-0}^{\infty}a^n z^{-n}=\frac{1}{1-az^{-1}}$ $ROC:\ z>|a|$

در تصویر 5 این توابع مهم را مشاهده می کنید:



تصویر 5 – تبدیل z توابع مهم

ویژگی های مهم تبدیل Z عبارتند از:

- خطی بودن: $Z\{ax[n]+by[n]\}=aX(z)+bY(z)$
- کانولوشن: $w[n]=x[n]*y[n]\rightarrow W(z)=X(z)Y(z)$
- شیفت: $Z\{x[n-k]\}=z^{-k}X(z)$
- مشتق جلو: $\Delta x[n]=x[n+1]-x[n]$
- مشتق عقب: $\nabla x[n]=x[n]-x[n-1]$

چون $\Delta x[n]=x[n]*(\delta[n+1]-\delta[n])$ از ویژگی شیفت نتیجه می شود:

$$Z\{\Delta x[n]\}=(z-1)X(z)$$
$$Z\{\nabla x[n]\}=(1-z^{-1})X(z)$$

تعریف منطقه همگرایی (ROC): ROC یک حلقه و یا صفحه دیسک مانند در صفحه z می باشد که روی مبدا واقع شده است.

تبدیل فوریه x[n] همگرا می شود (وجود دارد) فقط و فقط اگر ROC تبدیل z آن سیگنال شامل دایره واحد شود.

- ROC باید یک منطقه پیوسته باشد.

- اگر x[n] دنباله ای با طول محدود باشد، ROC تمام صفحه z می باشد (شاید به غیر از z=0 و z=∞)
- اگر x[n] دنباله سمت راستی باشد، ROC از بیرونی ترین قطب تا z=∞ ادامه خواهد داشت.
- اگر x[n] دنباله سمت چپی باشد، ROC از درونی ترین قطب تا z=0 ادامه خواهد داشت.
- یک دنباله دو سمتی یک دنباله تا بینهایت است که نه سمت چپی است و نه سمت راستی.
- اگر x[n] یک دنباله دو سمتی باشد، ROC شامل یک حلقه در صفحه z خواهد بود که بوسیله بیرونی ترین و درونی ترین قطب محدود خواهد بود.

دیدیم که برای یک سیستم LTI با پاسخ ضربه h[n]: $y[n] = x[n] * h[n]$

با خاصیت کانولوشن تبدیل Z: $Y(z) = X(z)H(z)$

4– تبدیل فوریه گسسته (Discrete Fourier Transform)

تبدیل فوریه را با توجه به پیوسته یا گسسته بودن زمان و فرکانس به چهار دسته تقسیم می کنند (تصویر 6).

Time	Frequency	Transform Type
Continuous	Continuous	Fourier Transform
Discrete	Continuous	Discrete Time Continuous FFT
Continuous	Discrete	Fourier Series
Discrete	Discrete	Discrete Time Discrete FFT

تصویر 6 – چهار دسته بندی تبدیل فوریه بر اساس پیوسته گسسته بودن زمان و فرکانس

تعریف تبدیل فوریه گسسته: $X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi kn}{N}}$

معمولاً به جای z عبارت $W_N = e^{\frac{j2\pi}{N}}$ را می گذارند: $X[k] = \sum_{n=0}^{N-1} x[n] W_N^{-nk}$

معکوس تبدیل فوریه گسسته: $x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] W^{nk}$

توابع مهم و تبدیل فوریه آن ها:

- ضربه واحد: $x[n] = \delta[n]$ $\quad X[k] = 1$
- ضربه واحد تاخیر دار: $x[n] = \delta[n-p]$ $\quad X[k] = W^{-kp}$
- ثابت: $x[n] = 1$ $\quad X[k] = N\delta[k]$
- نمایی مختلط: $x[n] = e^{j\alpha n}$ $\quad X[k] = N\delta\left(k - \frac{N\alpha}{2\pi}\right)$

- کسینوس: $x[n]=\cos 2\pi f_0 n$ $X[k]=\frac{N}{2}\left(\delta[k-Nf_0]+\delta[N-k+Nf_0]\right)$

ویژگی های DFT عبارتند از:

- تقارن: اگر $f[n]\leftrightarrow F[k]$ آنگاه $f[k]\leftrightarrow NF[-n]$
- خطی بودن: اگر $x[n]\leftrightarrow X[k]$ و $y[n]\leftrightarrow Y[k]$ آنگاه $ax[n]+by[n]\leftrightarrow aX[k]+bY[k]$
- شیفت: چون DFT ذاتاً فرض پریودیک بودن می کند، شیفت مانند چرخش است.

  اگر $x[n]\leftrightarrow X[k]$ آنگاه $x[n-p]\leftrightarrow W^{-kp}X[k]$
- معکوس زمان: اگر $x[n]\leftrightarrow X[k]$ آنگاه $x[-n]\leftrightarrow X[-k]$
- کانولوشن چرخشی: کانولوشن یک عملیان شیفت، ضرب و جمع است. چون همه شیفت های DFT چرخشی است، کانولوشن DFT این شیفت ها را هم در نظر می گیرد. $x[n]*y[n]=\sum_{p=0}^{N-1}x[p]y[n-p]$

5 – خلاصه و نتیجه گیری

در این فصل با چند معیار فاصله آشنا شدیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

آشنایی با مفاهیم پایه ای تئوری احتمال

خیلی از تکنیک های پردازش گفتار نیاز به کار با تئوری احتمال و آمار دارد.

دو دو کاربرد اصلی که برخورد خواهیم کرد عبارتند از:

- بازشناسی الگوری آماری
- مدل کردن سیستم های خطی

2– رخداد ها (Events)

معمول است که به یک احتمال، رخداد گفته شود.

یک رخداد، یک مجموعه مشخص از خروجی (outcome) های یک آزمایش می باشد.

فرض می شود خروجی ها دوبدو با هم اشتراک ندارند و اجتماع آن ها کل حالات را پوشش می دهد.

به هر رخداد A می توان عددی P(A) اختصاص داد که از قواعد زیر پیروی می کند:

- $P(A) \geq 0$
- $P(S)=1$
- اگر A و B دوبدو غیرمشترک باشند آنگاه $P(A+B)=P(A)+P(B)$
- عدد P(A) را احتمال A می گویند.

از قواعد بالا، قضایای زیر به دست می آید:

- اگر $\overline{A}$ مکمل A باشد آنگاه
  - $(A+\overline{A})=S$
  - $P(\overline{A})=1-P(A)$
- P(0) احتمال رخداد غیرممکن صفر است.
- $P(A) \leq 1$.

- اگر دو رخداد اشتراک داشته باشند، می توان نشان داد که *P(A+B)=P(A)+P(B)-P(AB).*

احتمال شرطی

احتمال شرطی یک رخداد A با دانستن اینکه رخداد B رخ می دهد به صورت زیر تعریف می شود:

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

می توان P(B|A) را بوسیله قانون بیزین استنباط کرد:

$$P(B \mid A) = P(A \mid B)\frac{P(B)}{P(A)}$$

استقلال

اگر رخدادهای A و B هیچ ربطی به هم نداشته باشند، می توان گفت که آن ها مستقل اند.

دو رخداد مستقل اند از هم اگر *P(AB)=P(A)P(B).*

از تعریف استقلال قواعد زیر به دست می آید:

- $P(A \mid B) = P(A)$
- $P(B \mid A) = P(B)$
- $P(A + B) = P(A) + P(B) - P(A)P(B)$

سه رخداد A و B و C مستقل اند اگر و فقط اگر:

$$\begin{cases} P(AB) = P(A)P(B) \\ P(AC) = P(A)P(C) \\ P(BC) = P(B)P(C) \\ P(ABC) = P(A)P(B)P(C) \end{cases}$$

3– متغییرهای تصادفی

یک متغییر تصادفی عددی است که به صورت تصادفی به عنوان خروجی آزمایش انتخاب شده است.

متغییرهای تصادفی ممکن است حقیقی و یا مختلط باشند و ممکن است گسسته و یا پیوسته باشند.

البته معمولاً متغییرهای تصادفی گسسته و خقیقی هستند.

می توان یک متغییر تصادفی را با توزیع احتمال آن یا با تابع توزیع احتمال (probability distribution function) آن توصیف کرد.

تابع توزیع یک متغییر تصادفی y احتمال این است که y از یک مقدار u بیشتر نشود:

$$F_y(u) = P(y \le u)$$

همچنین داریم:

$$P(u < y \le v) = F_y(v) - F_y(u)$$

$$f_y(u) = \frac{d}{du}F_y(u)$$

تابع چگالی احتمال مشتق تابع توزیع احتمال می باشد:

- همچنین: $P(u < y \le v) = \int_u^v f_y(y)dy$
- $F_y(\infty) = 1$
- $\int_{-\infty}^{+\infty} f_y(y)dy = 1$

امید ریاضی

می توان یک متغیر تصادفی را علاوه بر تابع توزیع احتمالش با شاخص های آماریش نیز توصیف کرد.

یکی از این شاخص های آماری امید ریاضی (Expected Value) می باشد.

امید ریاضی برای g(x) به صورت E{g(x)} یا <g(x)> نمایش داده می شود و به صورت زیر تعریف می شود:

- متغیر تصادفی پیوسته: $< g(x) >= \int_{-\infty}^{+\infty} g(x) f(x)dx$
- متغیر تصادفی گسسته: $< g(x) >= \sum_x g(x) p(x)$

ممان های متغیر تصادفی

یکی از شاخص های آماری مهم ممان ها (moments) p(x) می باشد.

K امین ممان p(x) برابر امید ریاضی x^k می باشد.

برای یک متغیر تصادفی گسسته: $m_k =< x^k >= \sum_x x^k p(x)$

میانگین و واریانس

ممان اول m_1، همان میانگین (mean) متغیر تصادفی x می باشد.

- پیوسته: $\overline{x} = \int_{-\infty}^{+\infty} xf(x)dx$
- گسسته: $\mu = \overline{x} =< x >= \sum_x xp(x)$

ممان مرکزی دوم، همان واریانس p(x) می باشد:

$$\sigma^2 = \sum_x (x - \overline{x})^2 p(x) = m_2 - \overline{x}^2$$

برای تخمین شاخص ها آماری یک متغییر تصادفی، آزمایشات بسیار زیادی را انجام می دهیم که متغییر را برای دفعات زیادی تولید کند.

در صورتی که آزمایش را Nبار انجام دهیم، هر مقدار x ، Np(x) بار اتفاق می افتد:

- تخمین ممان kام: $\hat{m}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k$
- تخمین میانگین: $\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i$

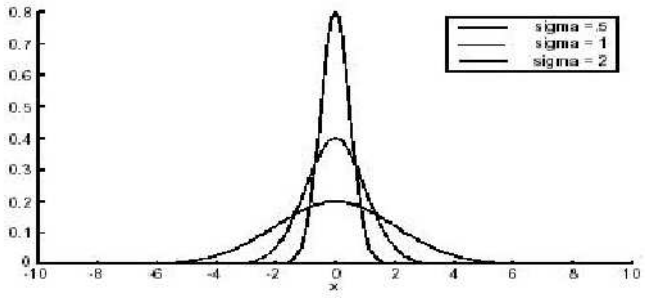
در زیر دو چگالی احتمال مهم را بررسی می کنیم:

- **چگالی احتمال یکنواخت:** یک متغییر تصادفی چگالی یکنواختی روی بازه (a, b) دارد اگر:

$$F_x(x)=\begin{cases}0, & x<a\\ (x-a)/(b-a), & a\le x\le b\\ 1, & x>b\end{cases} \quad f_x(x)=\begin{cases}1/(b-a), & a\le x\le b\\ 0, & \text{otherwise}\end{cases} \quad \sigma^2=\frac{1}{12}(b-a)^2$$

- **چگالی گووسی:** تابع چگالی گووسی یا نرمال به صورت زیر تعریف می شود (تصویر 1):

$$n(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$



تصویر 1 – تابع چگالی گووسی

تابع توزیع احتمال یک متغییر گووسی:

$$N(x;\mu,\sigma)=\int_{-\infty}^{x}n(u;\mu,\sigma)du$$

اگر تابع خطا به صورت روبرو تعریف شود:

$$erf(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-u^2/2}du$$

به فرمول زیر می رسیم:

$$N(x;\mu,\sigma)=\frac{1}{\sigma}erf\left(\frac{x-\mu}{\sigma}\right)$$

4– دو متغییر تصادفی

اگر دو متغییر تصادفی X و Y با هم در نظر گرفته شوند، می توان تابع چگالی احتمال مشترک f(x,y) برای متغییرهای پیوسته یا p(x,y) برای متغییرهای گسسته.

دو متغییر تصادفی مستقلند اگر:

$$p(x,y)=p(x)p(y)$$

با داشتن یک تابع g(x,y) امید ریاضی آن به صورت زیر تعریف می شود:

$$<g(x,y)>=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)f(x,y)dxdy$$

- پیوسته:
- گسسته: $< g(x,y) >= \sum_{x,y} g(x,y)p(x,y)$

ممان مشترک دو متغییر تصادفی x و y به صورت زیر محاسبه می شود: $m_{ij} = \sum_{x,y} x^i y^j p(x,y)$

ممان ها در عمل بوسیله میانگین گیری پشت سر هم آزمایش ها تخمین زده می شوند:

$$\hat{m}_{ij} = \frac{1}{N}\sum_{\delta=1}^{N} x_\delta^i y_\delta^j$$

ممان مرکزی دوم مشترک x و y کواریانس آن ها می باشد: $\sigma_{xy} =< (x-\bar{x})(y-\bar{y}) >= m_{11} - \overline{xy}$

- اگر x و y مستقل باشند کواریانس آن ها صفر است.
- ضرایب همبستگی x و y کواریانس آن ها است که به انحراف معیار نرمال شده باشد: $r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

دو متغییر تصادفی x و y مشترکاً گووسی می باشند اگر تابع چگالی آن ها مانند روبرو باشد:

$$n(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \exp\left[-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)\right]$$

که $r_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$

امید ریاضی جمع دو متغییر تصادفی: $< x+y >=< x > + < y >$

فرمول بالا هم برای متغییرهای مستقل و هم وابسته صادق است.

همچنین داریم: $\left\langle \sum_i x_i \right\rangle = \sum_i < x_i >$ و $< cx >= c < x >$

واریانس جمع دو متغییر تصادفی مستقل عبارت است از: $\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2$

اگر دو متغییر تصادفی به هم وابسته باشند، چگالی احتمال جمع آن ها کانولوشن چگالی تک تک متغییرهاست.

- پیوسته: $f_{x+y}(z) = \int_{-\infty}^{\infty} f_x(u) f_y(z-u) du$
- گسسته: $p_{x+y}(z) = \sum_{u=-\infty}^{\infty} p_x(u) p_y(z-u)$

5– مباحث دیگر

- تئوری حد مرکزی (Central Limit Theorem)

تعریف غیررسمی: اگر تعداد زیادی متغییر تصادفی مستقل با هم جمع شوند، تابع چگالی احتمال جمع آن ها مستقل از چگالی های متغییرها به سمت یک چگالی گووسی میل می کند.

- تابع چگالی گووسی چندمتغییره

تابع چگالی نرمال را می توان به هر تعداد متغییر تصادفی عمومیت داد.

اگر X یک بردار تصادفی باشد، $N(x) = (2\pi)^{-n/2} \mid R \mid^{-1} \exp\left[-\frac{1}{2} Q(x-\bar{x})\right]$

که $Q(x-\bar{x}) = (x-\bar{x})^T R^{-1} (x-\bar{x})$

ماتریس کوواریانس X با نام R: $R =< (x-\bar{x})(x-\bar{x})^T >$

- توابع تصادفی

یک تابع تصادفی تابعی است که به صورت خروجی یک آزمایش ناشی شود.

تابع تصادفی لزوماً توابعی از زمان نیستند، ولی در مطالعه ما معمولاً تابعی از زمانند.

یک فرآیند تصادفی گسسته بوسیله تعداد زیادی چگالی احتمال توصیف می شود.

$$p(x_1, x_2, x_3, ..., x_n, t_1, t_2, t_3, ..., t_n)$$

در صورتی که سیگنال های تصادفی از هم مستقل باشند،

$$p(x_1, x_2, ..., x_n, t_1, t_2, ..., t_n) = p(x_1, t_1) p(x_2, t_2) ... p(x_n, t_n)$$

اگر همه این چگالی های احتمال یکسان باشند، نتیجتاً دنباله ای از نمونه های مستقل و یکسان توزیع شده (i.i.d) خواهیم داشت.

- میانگین و خودهمبستگی

میانگین امید ریاضی x(t) می باشد: $\bar{x}(t) =< x(t) >= \sum_x x p(x,t)$

تابع خود همبستگی امید ریاضی x(t1)x(t2) می باشد: $r(t_1, t_2) =< x(t_1) x(t_2) >= \sum_{x_1, x_2} x_1 x_2 p(x_1, x_2, t_1, t_2)$

- میانگین زمانی و میانگین کلی (ensemble)

میانگین و خودهمبستگی را می توان به دو صورت مشخص کرد:

1. آزمایش می تواند به تعداد خیلی زیاد انجام شود و میانگین روی همه این توابع گرفته شود. به این میانگین، «میانگین ensemble» می گویند.
2. یکی از توابع را در نظر گرفته و آن را نمایانگر کل در نظر بگیریم. میانگین را از یک سری نمونه های این تابع محاسبه نماییم. به این مورد، «میانگین زمانی» می گویند.

- Ergodic بودن و ایستا بودن

اگر میانگین زمانی و کلی یک تابع تصادفی یکسان باشد، به آن ergodic می گویند.

یک تابع تصادفی ایستا است اگر شاخص آماری آن با تغییر زمان تغییر نکند.

همه توابع ergodic، ایستا هستند.

در یک سیگنال ایستا داریم: $\bar{x}(t) \equiv \bar{x}$ $\quad p(x_1, x_2, t_1, t_2) \equiv p(x_1, x_2, \tau)$ که $\tau = t_2 - t_1$

تابع خودهمبستگی عبارت است از: $r(\tau) = \sum_{x_1, x_2} x_1 x_2 p(x_1, x_2, \tau)$

یک x(t)ergodic است که میانگین و خودهمبستگی آن:

$$r(\tau) =< x(t)x(t-\tau) >= \lim_{N\to\infty} \frac{1}{N} \sum_{t=-N}^{N} x(t)x(t-\tau) \qquad \bar{x} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=-N}^{N} x(t)$$

- همبستگی تقاطعی

همبستگی تقاطعی دو تابع تصادفی ergodicعبارت است از:

$$r_{xy}(\tau) =< x(t)y(t-\tau) >= \lim_{N\to\infty} \frac{1}{N} \sum_{t=-N}^{N} x(t)y(t-\tau)$$

اندیس xy نشان دهنده تقاطعی بودن است.

- چگالی توانی: تبدیل فوریه $r(\tau)$ را چگالی توانی طیف x(t) می گویند $\quad S(\omega) = \sum_{\tau=-\infty}^{\infty} r(\tau) e^{-j\omega\tau}$
- چگالی طیفی تقاطعی دو تابع تصادفی عبارت است از:

$$S_{xy}(\omega) = \sum_{\tau=-\infty}^{\infty} r_{xy}(\tau) e^{-j\omega\tau}$$

برای سیگنال های ergodic می توان به صورت زیر نوشت: $r(\tau) = x(\tau) * x(-\tau)$

و بر اساس تبدیل فوریه:

$$\begin{aligned} S(\omega) &= X(\omega)X(-\omega) \\ &= X(\omega)X^*(\omega) \\ &= |X(\omega)|^2 \end{aligned}$$

در صورتی که همه مقادیر یک سیگنال تصادفی غیرهمبسته باشند (بعنی سیگنال نویز سفید باشد)

$$r(\tau) = \sigma^2 \delta(\tau)$$

طیف توانی نویز سفید مقدار ثابت است: $S(\omega) = \sigma^2$

نویز سفید مخلوطی از تمامی فرکانس ها می باشد.

اگر T[] عملیات خطی باشد، $< T[x(t)] >= T[< x(t) >]$

با داشتن یک پاسخ فرکانسی h(n): $< y(n) >=< x(n) * h(n) >=< x(n) > * h(n)$

یک سیگنال ایستا که به یک سیستم خطی اعمال می شود یک خر.جی ایستا می دهد.

$r_{yy}(\tau) = r_{xx}(\tau) * h(\tau) * h(-\tau)$

$S_{yy}(\omega) = S_{xx}(\omega) | H(\omega) |^2$

6 – خلاصه و نتیجه گیری

در این فصل مروری بر بحث احتمال انجام دادیم.

7 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

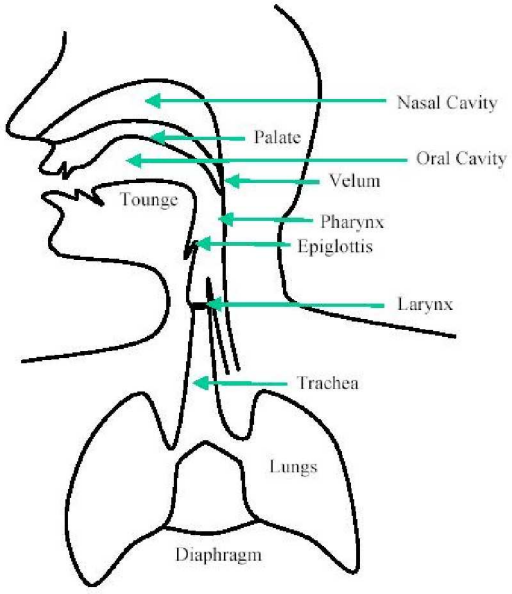
1– مقدمه:

اهداف درس:

آشنایی با نحوه تولید گفتار: مطالعه آناتومی اندام های گفتار پیش نیازی برای مطالعه آواشناسی (هم از لحاظ صوتی و هم از لحاظ مفصلی) می باشد.

2– اندام های تولید گفتار

در تصویر 1 اندام های تولید صوت انسان را مشاهده می کنید. در ادامه این اندام ها را از پایین به بالا شرح می دهیم و وظایف آن را بیان خواهیم کرد.

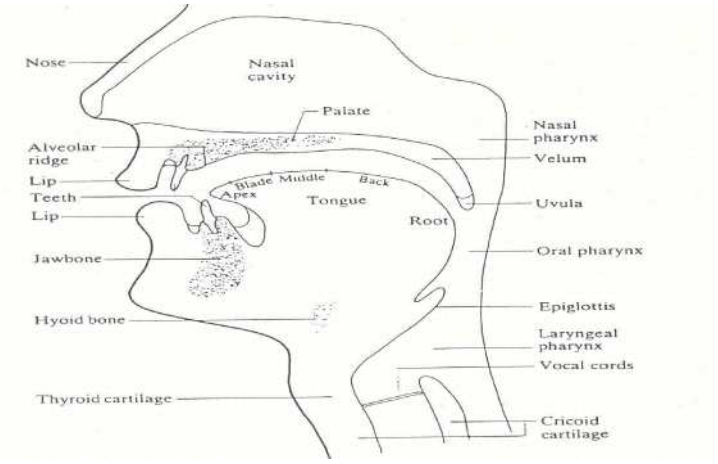


تصویر 1 – اندام های تولید صوت انسان

- شش ها و نای (Lungs and Trachea)
  - این قسمت به عنوان منبع هوا در حین تولید گفتار عمل می کند.
  - اندام های صوتی بوسیله هوای فشرده شده کار می کنند. این هوای فشرده توسط شش ها فراهم شده و توسط نای به سیستم صوتی منتقل می شود.
  - این اندام همچنین وظیفه کنترل بلندی گفتار تولید شده را بر عهده دارند.
  - به شش و نای با هم «مجرای ریوی» گفته می شود.
- حنجره (Larynx)

- این عضو یک سیستم پیچیده است ساخته شده از غضروف و ماهیچه است که شامل و کنترل کننده تارهای صوتی می باشد.
- قسمت های اصلی حنجره عبارتند از:
  - ✓ غضروف حلقه ای (Cricoid Cartilage)
  - ✓ غضروف سپر مانند (Thyroid Cartilage)
  - ✓ غضروف آرتنوید (Arytenoid Cartilage)
  - ✓ تار های صوتی (Vocal Cords)
- جاییکه تارهای صوتی به هم می رسند چاکنای (glottis) نام دارد.

- مسیر صوتی (Vocal Tract)
  - قسمت های مسیر صوتی را می توانید به صورت دقیق تر در تصویر 2 مشاهده کنید.
  - حلق حنجره (Laryngeal pharynx): زیر نای بند (epiglottis) واقع شده است.
  - حلق دهانی (Oral pharynx): پشت زبان، بین نای بند و velum واقع است.
  - حلق دماغی (Nasal pharynx): بالای velum، انتهای حفره دماغی واقع شده است.
  - حفره دهانی (Oral Cavity): جلوی velum واقع شده است و بوسیله لب ها، زبان و سقف دهان بسته شده است.
  - حفره دماغی (Nasal Cavity): بالای سقف دهان واقع شده است و از حلق تا سوراخ بینی را شامل می شود.

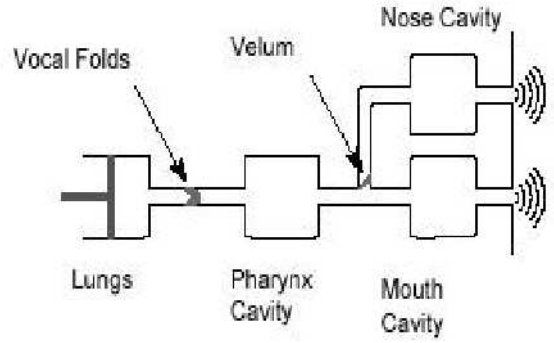


تصویر 2 – قسمت های مسیر صوتی

3– مدل سازی مسیر صوتی

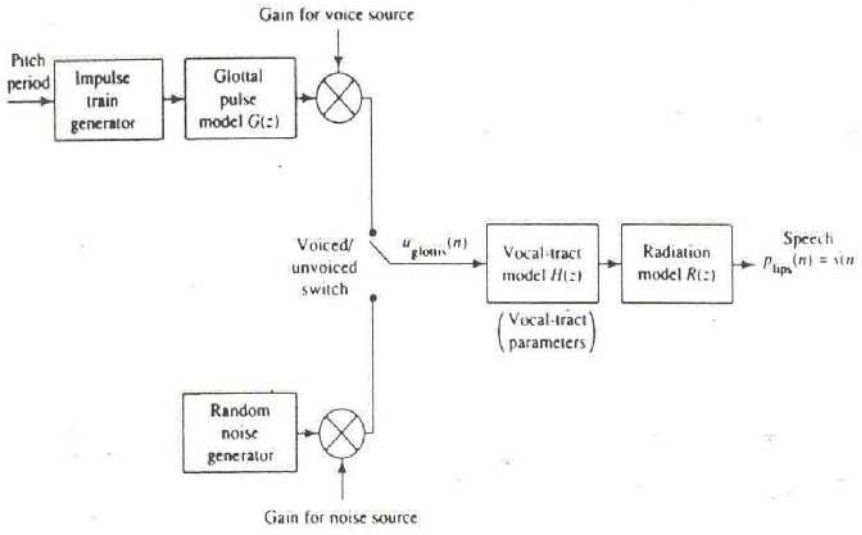
- سعی می شود که در کاربردهای پردازش و بازشناسی گفتار، مسیر صوتی مدل شود.

- مدل های زیادی برای مسیر صوتی انسان ارائه شده است.
- در تصویر 3 یک مدل فیزیکی از تولید صوت انسان (شامل مدل مسیر صوتی) مشاهده می کنید.



تصویر 3 – مدل تولید صوت انسان که شامل مدل مسیر صوتی نیز می باشد

- مدل های گسسته زیادی برای مسیر صوتی ارائه شده است.
- هدف از این مدل ها، مدل کردن مسیر صوتی بر روی تجهیزات دیجیتالی (از جمله کامپیوتر ها و موبایل ها) می باشد.
- در تصویر 4 یک مدل کلی گسسته برای تولید صوت مشاهده می کنید.



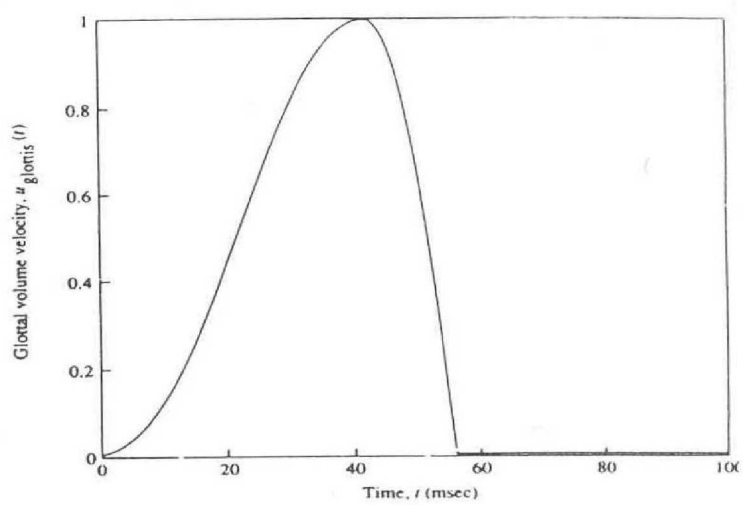
تصویر 4 –یک مدل کلی گسسته برای تولید صوت انسان

- اجزای این تصویر در فصول بعد توضیح داده خواهند شد.

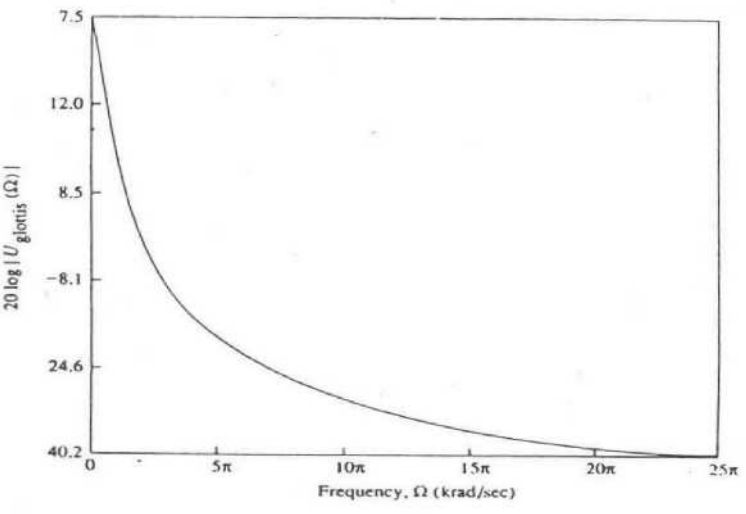
4– پالس تارهای صوتی

- در تصویر 4 پالس های پریودیک به عنوان هوای فشرده رد شده از تارهای صوتی عمل می کنند.

- H(z) هم مدل مسیر صوتی می باشد.
- با رد شدن پالس (هوا) از مدل (مسیر صوتی) گفتار تولید می شود.
- این فرآیند مشابه نحوه تولید گفتار در انسان است.
- یک پریود پالس تولید شده توسط تار صوتی انسان را در حوزه زمان در تصویر 5 مشاهده می کنید.
- یک پریود پالس تولید شده توسط تار صوتی انسان را در خوزه فرکانس در تصویر 6 مشاهده می کنید.



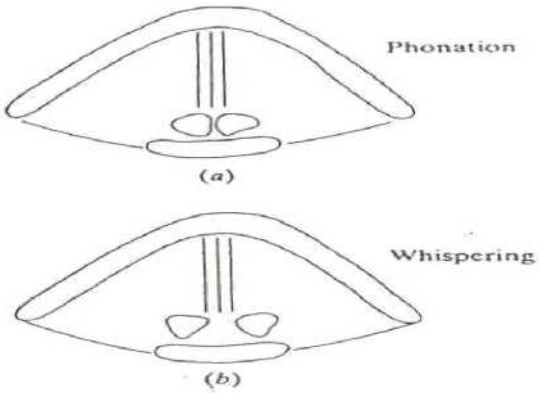
تصویر 5 – یک پریود از پالس تولید شده توسط تار صوتی انسان در حوزه زمان



تصویر 6 – یک پریود از پالس تولید شده توسط تار صوتی انسان در حوزه فرکانس

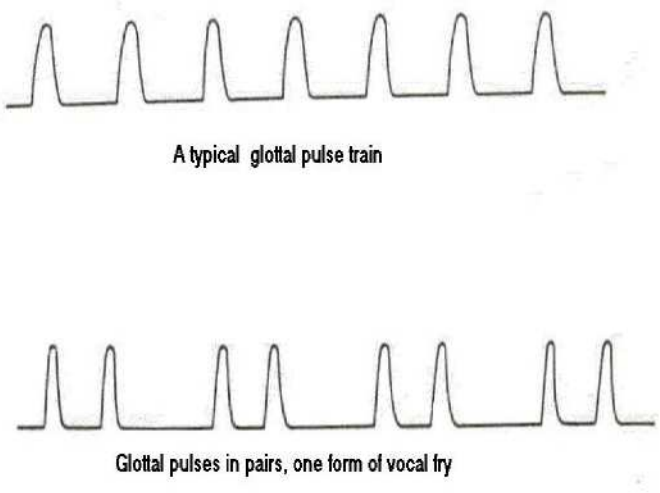
- همان طور که مشاهده می کنید پالس تولید شده توسط تار صوتی انسان بسیار شبیه ضربه (impulse) می باشد.

- یعنی در یک جا از زمان فشردگی زیادی به وجود می آید (مانند ضربه سیگنال به سمت بینهایت می رود) و در بقیه جاها صفر است.
- در حوزه فرکانس این سیگنال در فرکانس های پایین زیاد و در فرکانس های بالا کم است (تصویر 6).
- تارهای صوتی و غضروف ها برای هر واج یک شکل می گیرند.
- مثلاً واج ها ـَ که واکه است تارهای صوتی می لرزند (امتحان کنید) ولی برای واج ش نمی لرزند.
- در تصویر 7 تفاوت قرار گرفتن غضروف ها و تارهای صوتی برای این دو حالت را مشاهده می کنید.



تصویر 7 –نحوه قرار گرفتن تارهای صوتی و غضروف ها برای دو واج ـَ و ش

- پالس نهایی تولید شده پریودیک است (مانند تصویر 8 بالا).
- برخی پالس های تولید شده یک ضربه نیستند (مانند صداهای خش دار) (مانند تصویر 8 پایین).



تصویر 8 – پالس تولید شده توسط تارهای صوتی سالم (بالا) و خش دار (پایین)

4– تولید گفتار

- کل عملیات تولید گفتار به دو قسمت تقسیم می شود:

- ✓ تحریک (Excitation): همان پالس تولید شده توسط تارهای صوتی است.
- ✓ مدولاسیون یا فرکانس گذاری: همان اعمال مسیر صوتی است.
  - این دو مرحله به صورت پشت سر هم هستند و در تصویر 9 نشان داده شده اند.

شکل 9 –مراحل کلی تولید گفتار

- **تحریک**به چند روش انجام می شود.

1. آواگری (Phonation):

- تولید صدای واکدار می باشد. به عبارتی تارهای صوتی به لرزه در می آیند. در این هنگام غضروف های arytenoids بسته می شوند و تارهای صوتی را می کشند. هنگامی که هوا از درون تارهای صوتی عبور می کند، آن ها را می لرزاند. باز و بسته شدن تارها جریان هوا را به یک سری پالس می شکند. نمونه پالس تولید شده را می توانید در شکل 8 مشاهده نمایید.
- به میزان تناوب این پالس فرکانس گام (pitch) گفته می شود.
- فرکانس گام زیری و بمی (کلفتی و نازکی) گفتار را تعیین می کنند.
- صداهای گفتار که بوسیله آواگری تولید می شوند (تار صوتی در حین تولید آن ها می لرزد) را واکدار (voiced) می گویند.
- به بقیه صدا ها بدون واک یا مصوت (unvoiced or mute) گفته می شود.
- مثال: ژ، ـَ

6 – خلاصه و نتیجه گیری:

در این فصل یاد گرفتیم که:

- اندام های تولید گفتار انسان کدامند.
- فرآید تولید صوت در انسان شامل چه مراحلی است.
- پالس های تحریک چگونه توسط تارهای صوتی تولید می شوند.
- چگونه می توان تولید صوت در انسان را مدل کرد که بتوان بر روی دستگاه های دیجیتال از آن استفاده نمود.

7 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1. پچ پچ کردن (Whispering):
   - در این حالت تارهای صورت به سمت هم رانده می شوند ولی یک دهانه مثلثی کوچک بین غضروف های آرتنوید باقی می ماند.
2. سایش (Frication):
   - سایش ممکن است با/بدون آواگری رخ دهد.
   - معمولاً انسدادی نسبی در مسیر صوتی رخ می دهد که باعث صدای سایشی می شود.
   - مثال: ش، س، ف
3. فشرده سازی (Compression):
   - در صورتی که رها شدن ناگهانی باشد، صدا انسدادی می باشد.
   - مثال: ت
   - در صورتی که رها شدن تدریجی و متلاطم باشد، صدا شبیه سایشی ها می شود. به این نوع صداها شبه سایشی گفته می شود.
   - مثال: چ.
4. لرزشی (Vibration):
   - هوا از درون یک دریچه غیر از تارهای صوتی عبور می کند که باعث لرزش می شود.
   - مثال: ر.

- **فرکانس گذاری** در این مرحله است که اطلاعات گفتاری بر روی پالس تارهای صوتی گذاشته می شود.

✓ آواشناسی مفصلی: اندام های گفتاری چگونه قرار می گیرندتا هر صدای گفتار تولید شود.

✓ آواشناسی صوتی: ویژگی های صوتی قابل محاسبه مطالعه می شوند. همچنین همبستگی این ویژگی ها با ویژگی های واجی و مفصلی بررسی می شود.

1– مقدمه:

اهداف درس:

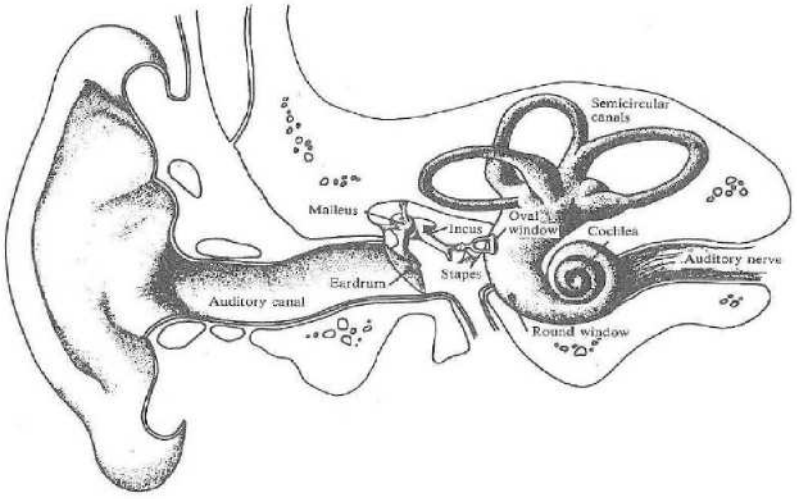
آشنایی با نحوه شنوایی و ادراک انسان: مطالعه شنوایی و ادراک صوتی انسان در زمینه سنتز گفتار و بهبود گفتار نیاز است. همچنین این اطلاعات در زمینه بازشناسی گفتار کاربردی هستند.

تعریف شنوایی: شنوایی فرآیندی است که در آن صدا دریافت شده و تبدیل به سیگنال های عصبی می شوند.

تعریف ادراک: پردازش بعدی درون مغز که در آن صداهای شنیده شده تفسیر شده و دارای معنی می شوند.

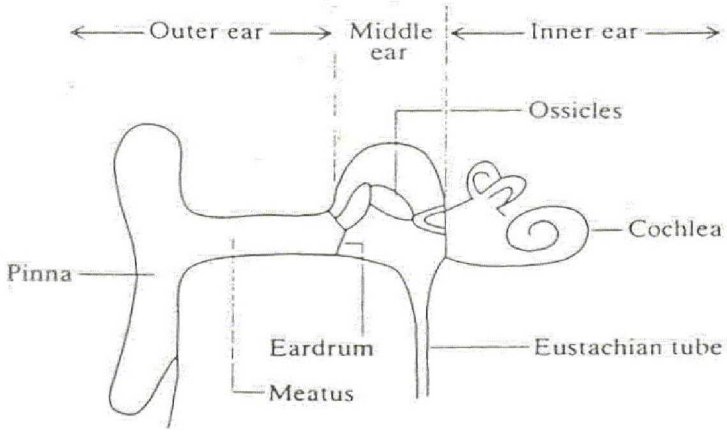
2– ساختار کلی گوش

در تصویر 1 ساختار یک گوش انسان را مشاهده می کنید.



تصویر 1 –ساختار گوش انسان

در تصویر 2 شکل نمادین از یک سطح مقطع گوش انسان را مشاهده می کنید.



تصویر 2– تصویر سطح مقطع گوش انسان

گوش انسان به سه قسمت کلی تقسیم شده است:

- گوش بیرونی
- گوش میانی

- گوش درونی

3– گوش بیرونی

گوش بیرونی، شامل:

- لالهگوش (غضروفیپیچپیچوقابلمشاهده): پیچپیچبودنآنباعثبهخاطریکسریجهتدهیهامیباشد.
- کانالخارجی (مجرایصوتیخارجی): لولهاییکنواختبا 2.7 سانتیمترطولکهیکسریبسامدهمنوادرحدود 3 کیلوهرتزدارد.
- پردهگوش: غشاییبهشکلطبلاستوشکلمخروطیوسفتیدارد. درانتهایمجرایصوتیخارجیواقعشدهاست.

هنگامبرخوردصوتباآنمیلرزد.

4– گوش میانی

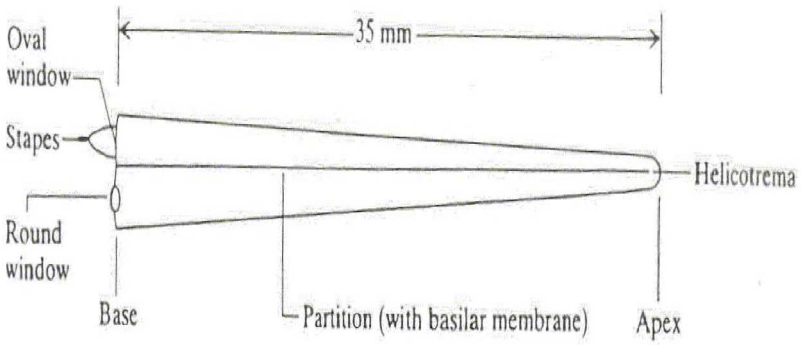
- گوشمیانییکحفرهاستکهدرونآنپرازهوااست.
- بوسیلهغشاییبهشکلطبلازگوشخارجیجدامیشود.
- بهگوشدرونیبوسیلهیکپنجرهبیضییادایرویمتصلاست.
- بوسیله لوله استاخی (Eustachian) به دنیای بیرون متصل است.
- این لوله باعث تعادل فشار هوا بین گوش میانی و اتمسفر اطراف می شود.
- گوش میانی شامل سه استخوان کوچک است.
  - استخوان چکشی
  - استخوان سندانی
  - استخوان رکابی
- وظیفه این استخوانچه ها
  - انتقال امپدانس
  - محدود کردن دامنه نوسان

5– گوش درونی

گوش درونی شامل:

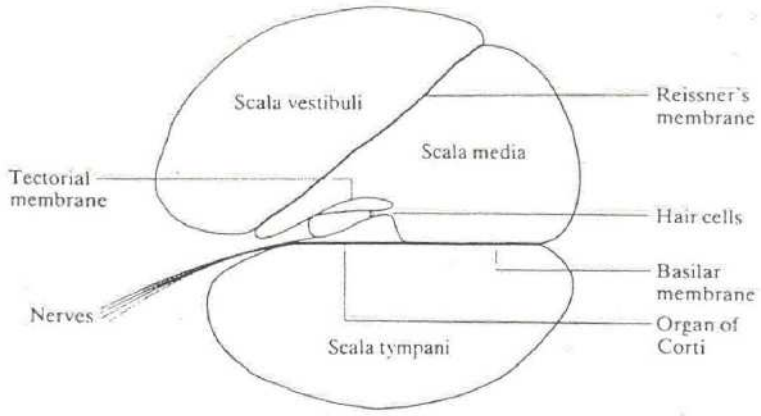
- Vestibulat apparatus: که وظیفه تعادل و حس جهت یابی را بر عهده دارد.
- پنجره بیضوی و دایروی
- حلزونی گوش
  - یک مسیر حلزونی مانند است.

- بوسیله پنجره بیضوی دایروی با گوش میانی در ارتباط است.
- تبدیل کننده هایی دارد که لرزش های صوتی را به سیگنال های عصبی تبدیل می کند.
- در تصویر 3 یک حلزونی پهن شده را مشاهده می کنید.



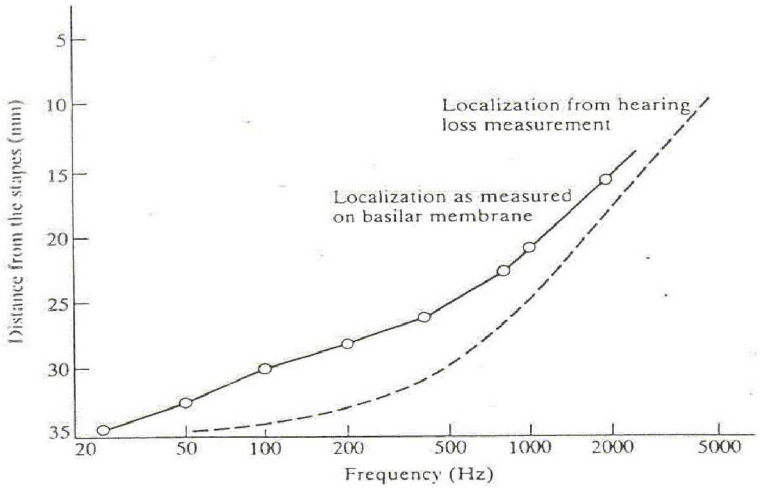
تصویر 3 – حلزونی پهن شده

- سطح مقطع حلزونی گوش را در تصویر 4 مشاهده می کنید.



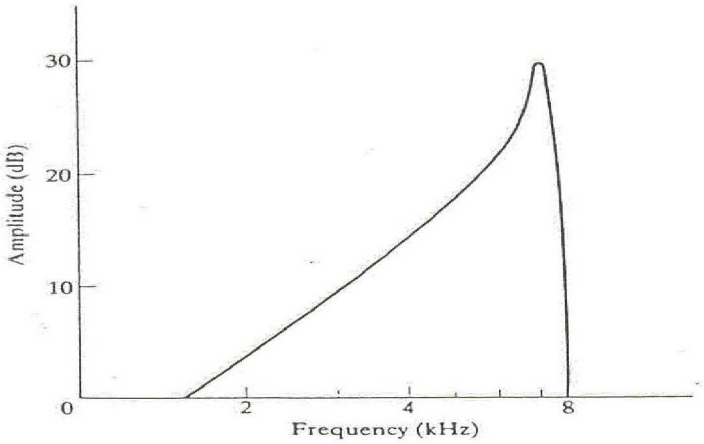
تصویر 4 – سطح مقطع حلزونی گوش

- در تصویر 5 نمودار مکان غشا پایه را نسبت به فرکانس مشاهده می کنید.



تصویر 5 – مکان غشا پایه نسبت به فرکانس

- در تصویر 6 واکنش فرکانس یک نقطه از غشا پایه را مشاهده می کنید.



8 – خلاصه و نتیجه گیری:

در این فصل ساختار گوش انسان را مطالعه کردیم.

متوجه شدیم که گوش انسان را می توان به سه قسمت تقسیم کرد:

- گوش بیرونی
- گوش میانی
- گوش درونی

هر کدام از این قسمت ها شامل قسمت ها دیگری هستند که در این فصل به صورت کلی بدان پرداختیم.

9 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

آشنایی آوا شناسی (phonetics): هدف آوا شناسی مطالعه کلی واج ها است. این مطالعه مستقل از زبان انجام می گیرد.

آشنایی با واج اشناسی (phonemics): هدف واج شناسی مطالعه کلی واج ها در یک زبان خاص است.

2– آواشناسی

مطالعه آواشناسی به دو دسته آواشناسی مفصلی و آواشناسی صوتی تقسیم می شود.

2–1– آواشناسی مفصلی

یک صوت موجود در گفتار چگونه تولید می شود. به عبارتی تاکید بر روی نحوه قرارگیری اندام های تولید صوت مانند زبان و ... است.

- تحریک (Excitation): همان طور که در فصل قبل گفته شد، تحریک پنج نوع است:
  1. آواگری (Phonation): مانند ـَ
  2. پچ پچ کردن
  3. فشرده سازی مانند ت
  4. سایش مانند ش
  5. لرزش مانند ر
- **صامت ها (Consonants)**: بررسی صامت ها از دید نحوه قرارگیری اندام های صوتی آسان است.
- در هنگام تولید صامت ها، سه نکته ماهیت صامت خروجی را تعیین می کند.
  1. مکان تولید (point of articulation): مکان انسداد اصلی که در مسیر صوتی ایجاد می کنیم.
     - دو لبی(Bilabial) مانند ب
     - لب و دندانی(Labiodental) مانند think
     - Apicodental
     - Apicogingival
     - Apicoalveolar
     - Apicodomal
     - Laminoalveolar
     - Laminodomal
     - Centrodomal
     - Dorsovelar

- حلقی (Pharyngeal)
- حنجره ای (Glottal)

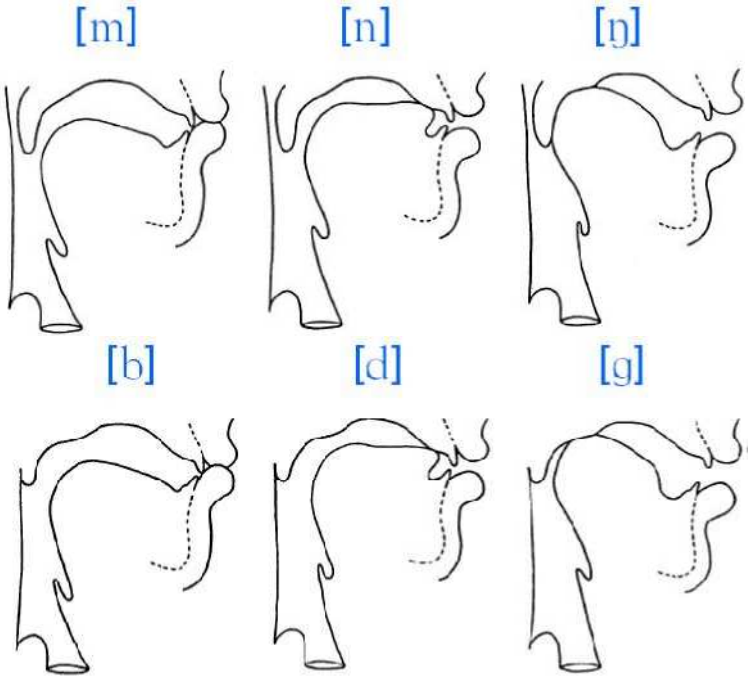
2. نحوه تولید (manner of articulation): درجه و قدرت انسداد و نحوه آزادسازی انسداد

- انفجاری (Plosive) مانند ت
- دمشی (Aspirated) مانند ه
- انفجاری-سایشی (Affricative) مانند چ
- سایشی (Fricative) مانند س
- جانبی (Lateral) مانند ل
- نیمه واکه (Semivowel) مانند ی

- دماغی (Nasal) مانند ن
- لرزشی (Trill) مانند ر

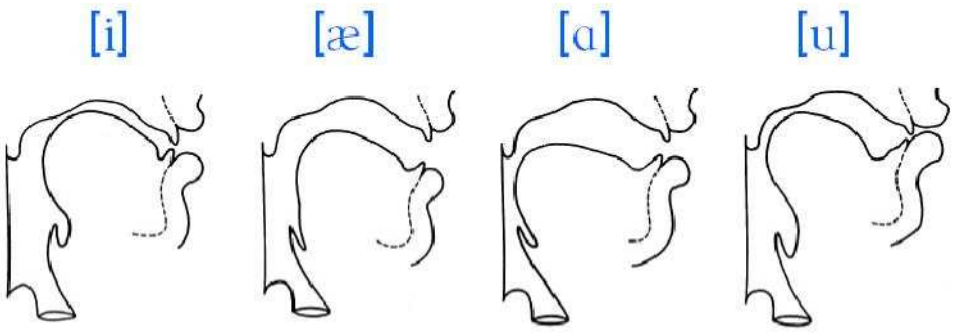
۱. واکداری (voicing)

- صدادار (voiced) مانند گ
- بی صدا (unvoiced) مانند ک



تصویر ۱ – اندام های گفتاری انسان در حین تلفظ تعدادی از صامت ها

- **واکه ها (vowels)**:تعریف واکه ها بر مبنای اندام های صوتی سخت تر است. این به این خاطر است معمولاً زبان هیچوقت اندام دیگری را لمس نمی کند.
- واکه ها با موارد زیر تعریف می شوند:
  ۱. بالا–پایین بودن زبان
  ۲. جلو–عقب بودن زبان
  ۳. گرد بودن–نبودن لب
  ٤. تودماغی بودن–نبودن



تصویر ۲ – اندام های گفتاری انسان در حین تلفظ تعدادی از مصوت ها

- **واکه دوگانه (diphthongs)**: دو صدای واکه در یک سیلاب ترکیب می شوند.
- برای تولید اینگونه واج ها، زبان از یک نقطه به نقطه دیگر تغییر مکان می دهد.
- **تولید باهم (Coarticulation)**:هیچ صدایی در گفتار در جوار صداهای دیگر یکجور تولید نمی شود.
- همپوشانی ویژگی های آوایی از یک آوا به آوای دیگر را تولید با هم می گویند.

۲-۲- آوا شناسی صوتی

در این زمینه، تاکید بر روی ویژگی های قابل مشاهده و قابل اندازه گیری در شکل موج گفتار است

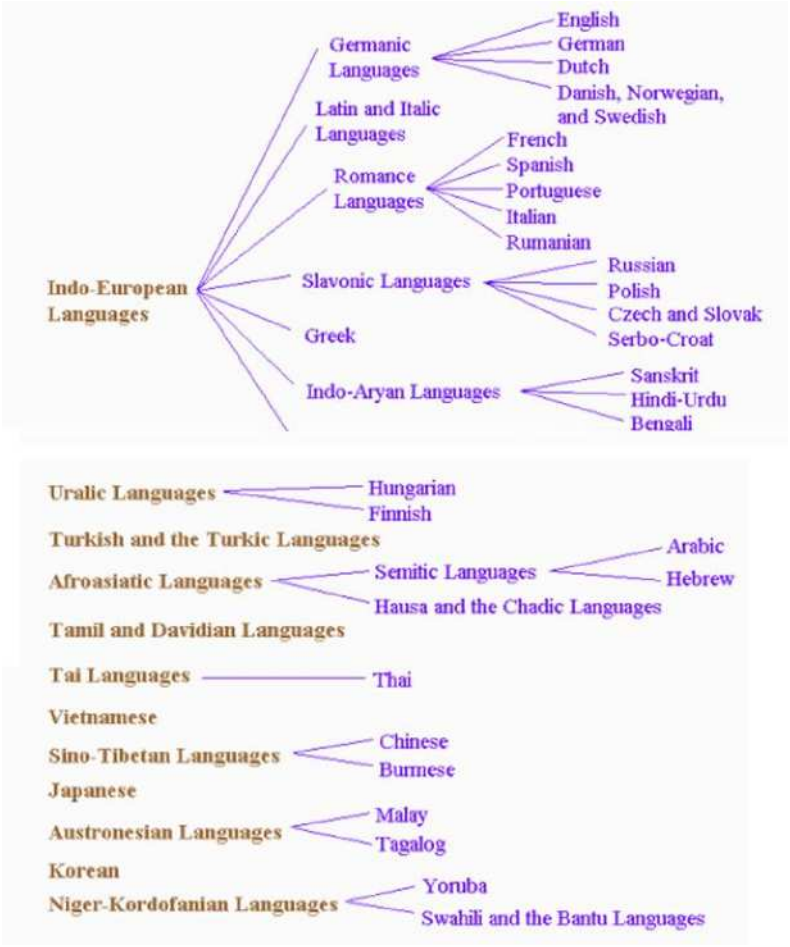
این مطالعات، پیش زمینه های تئوری و عملی برای بازشناسی و سنتز گفتار بوسیله سخت افزار الکترونیکی ارائه می دهد.

۳- واج شناسی

- بحث آواشناسی مطالعه صداها مستقل از زبان است.
- بحث واج شناسی مطالعه صداها در یک زبان خاص می باشد.
- واج: در بحث آواشناسی هر صدا یک «آوا» به حساب می آید. در واج شناسی هر صدا یک «واج» گفته می شود.
- در بحث واج شناسی، کوچک ترین واحد «واج» است.
- تعریف دقیق واج: یک واج کوچکترین واحد آوایی در یک زبان است که کافی است تا یک کلمه را از یک کلمه دیگر تفاوت دهیم.
- مثال: در زبان انگلیسی، ویژگی واکداری بین دو واج باعث تمایز می شود.
- مثال: buck و bug در انگلیسی و کل و گل در فارسی
- در برخی زبان ها مانند آلمانی واکداری یک واج زبان حساب می شود.
- مثال: 'tag' در آلمانی هم [ta:g] و هم [ta:k] تلفظ می شود.

۳- بررسی برخی زبان ها

در تصویر ۳ برخی زبان های معمول و ریشه های آن ها را مشاهده می کنید.



تصویر ۳ – زبان های معمول و ریشه های آن ها

- بیشترین تعداد واج موجود در یک زبان، ٤٥ واج است که در زبان Chipewyan (زبان بومی های آمریکا) موجود است.
- کمترین تعداد واج موجود در یک زبان ۱۳ واج است مربوط به زبان هاوایی Hawaiian.
- انگلیسی بین ۳۱ تا ٦٤ واج دارد (بستگی به این دارد که چگونه تحلیل شوند).
- فارسی ۲۹ تا ٤٥ واج دارد (بستگی به این دارد که چگونه تحلیل شوند).
- **چندصدایی ها (allophones)**: یک واج در حقیقت یک «مجموعه ای» از آواهای شبیه هم است که بوسیله یک گوینده های یک زبان به عنوان یک «صدا» تلقی می شوند.
- به اعضای این مجموعه allophone اطلاق می شود.
- مثال: واج /k/ در kin و cup.

- مثال: واج /k/ در cope و scope
- مثال: واج /k/ در کاهو و کلم
- واج های زبان انگلیسی را در تصویر ٤ مشاهده می کنید.

Vowels	uw ux uh ah ax ah-h aa ao ae eh ih ix ey iy ay ow aw oy er axr el
Semi-vowels	y r l el w
Fricatives	jh ch s sh z zh f th v dh
Nasals	m n ng em en eng nx
Stops	b d g p t k dx q bcl dcl gcl pcl tcl kcl
Aspiration	~~hv hh~~

- واج های زبان ف

واکه ها	
یه، ي	*i*
به،	*e*
ـَ	*a*
و	*u*
وه	*o*
ا،آ	*σ*

انفجاري ها	
ب	*b*
پ	*p*
ت، ط	*t*
د	*d*
ك	*c*
ك	*(k)*
گ	*τ*
گ	*(g)*
ق،غ	*G*
ء	*ι*

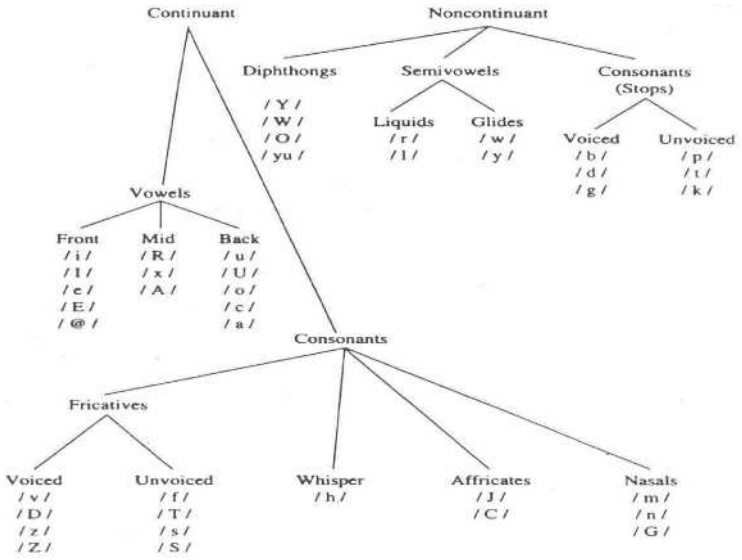
سایشي ها	
ف	f
و	v
ث، س، ص	s
ز، ذ، ض، ظ	Z
ش	$\int$
ژ	$\Im$
خ	χ
هه، ه، ح	h

انفجاري سايشي ها	
ج	$d\Im$
چ	$t\int$

شبه واكه ها	
ل	l
ر	r
م	m
ن	n
يه، ي	j

تصویر ۵ – واج های زبان فارسی به همراه دسته بندیشان

- دسته بندی واج های زبان انگلیسی را در تصویر ٦ مشاهده می کنید.



تصویر ٦ – دسته بندی واج های انگلیسی

۴– الفبای آوا شناسی

آواشناس ها برای راحتی و استاندارد شدن مطالعات، سیستم نمادگذاری به صورت الفبا طرح کرده اند تا با استفاده از آن بتوانند آوا ها را معرفی نمایند.

دو نمونه از این آوا ها عبارتند از:

- IPA
- ARPAbet

نمونه ای از الفبای IPA را برای زبان انگلیسی در تصویر ۷ مشاهده می کنید.

IPA symbol	Arpabet		Examples	IPA symbol	Arpabet		Examples
i	i	IY	heed	v	v	V	verve
ɪ	I	IH	hid	θ	T	TH	thick
e	e	EY	hayed	ð	D	DH	those
ɛ	E	EH	head	s	s	S	cease
æ	ɑ̃	AE	had	z	z	Z	pizzaz
ɑ	a	AA	hod	ʃ	S	SH	mesh
ɔ	c	AO	hawed	ʒ	Z	ZH	measure
o	o	OW	hoed	h	h	HH	heat
ʊ	U	UH	hood	m	m	M	mom
u	u	UW	who'd	n	n	N	noon
ɝ	R	ER	heard	ŋ	G	NX	ringing
ə	x	AX	*a*head	l	l	L	lulu
ʌ	A	AH	bud	l̩	L	EL	bat*tle*†
aɪ	Y	AY	hide	m̩	M	EM	bot*tom*†
aʊ	W	AW	how'd	n̩	N	EN	but*ton*†
ɔɪ	O	OY	boy	ɾ	F	DX	ba*tter*‡
ɨ	X	IX	roses	ʔ	Q	Q	§
p	p	P	pop	w	w	W	wow
b	b	B	bob	j	y	Y	yoyo
t	t	T	tug	ɹ	r	R	roar
d	d	D	dug	ʧ	C	CH	church
k	k	K	kick	ʤ	J	JH	judge
g	g	G	gig	ʍ	H	WH	where
f	f	F	fife				

† Vocalic l, m, n ‡ Flapped t § Glottal stop

تصویر ۷ – نمادهای IPA برای واج های زبان انگلیسی

۵ – خلاصه و نتیجه گیری:

در این فصل با بحث آواشناسی و واج شناسی آشنا شدیم.

دیدیم که آواشناسی مستقل از زبان است و واج شناسی وابسته به زبان است.

همچنین نحوه تولید صامت ها و واکه ها را توضیح دادیم.

برای تولید صامت ها ویژگی های زیر مهم است:

- نقطه تولید
- نحوه تولید
- واکداری

برای تولید واکه های موارد زیر مهم است:

- بالا–پایین بودن زبان

- جلو–عقب بودن زبان
- گرد بودن–نبودن لب
- تودماغی بودن

۶ – منابع درس:

۱- Rabiner, “Fundumentals of Speech Recognition”
۲- Huang, Acero, “Spoken Language Processing”
۳- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

آشنایی با مفهوم اسپکتروگرام ها

آشنایی با نحوه خواندن اسپکتروگرام ها

2– مفاهیم اولیه

اسپکتروگرام ها شکل موج در حوزه زمان را در دو بعد زمان فرکانس نمایش می دهند.

با این کار خوانایی سیگنال خیلی بیشتر می شود.

چون در هر زمان (محور افقی) طیف سیگنال در آن زمان (محور عمودی) نشان داده می شود (تصویر 1).

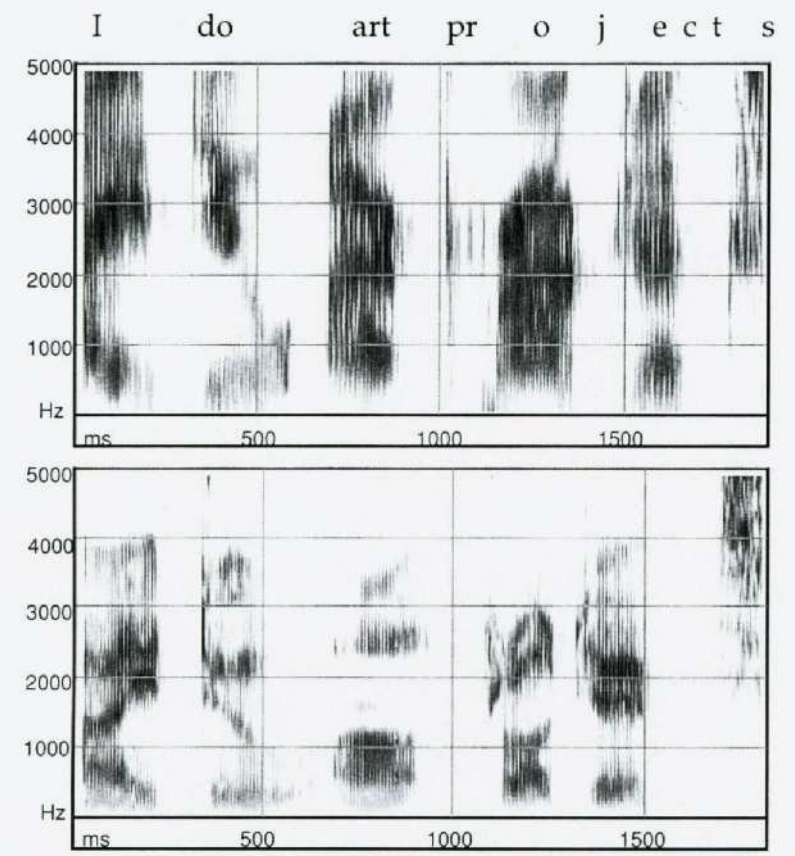
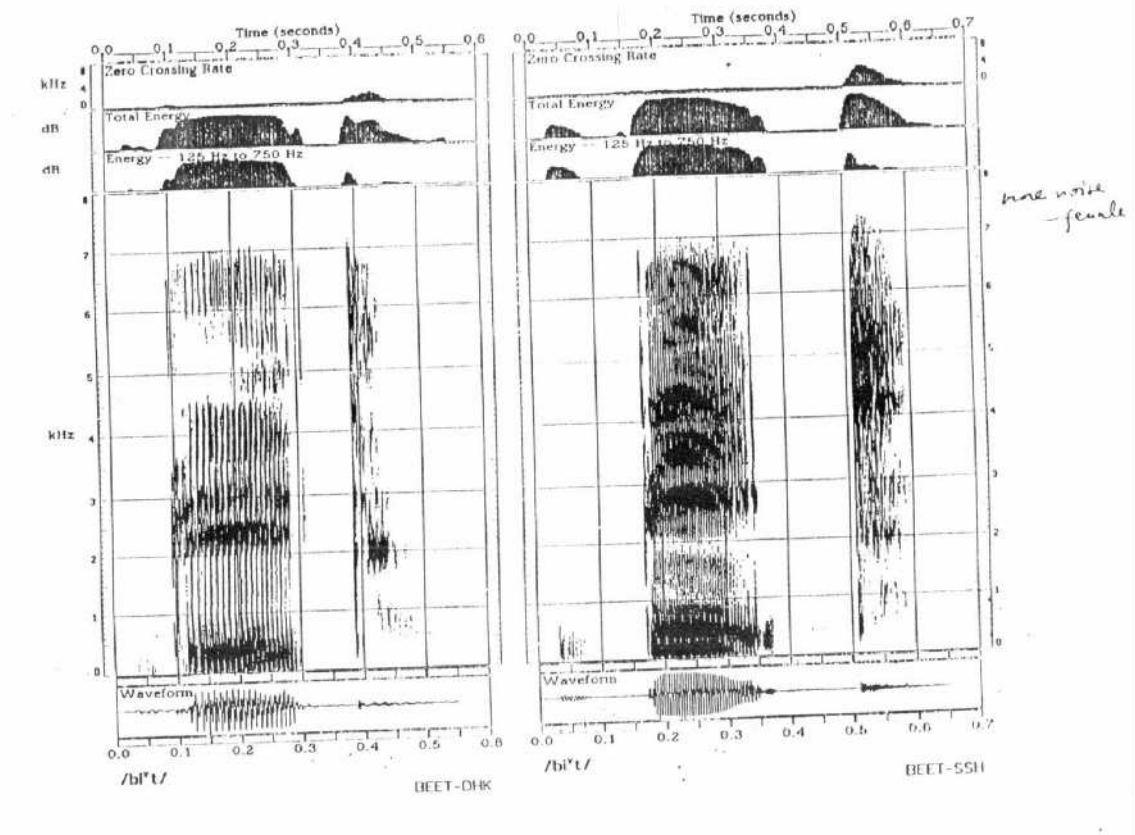


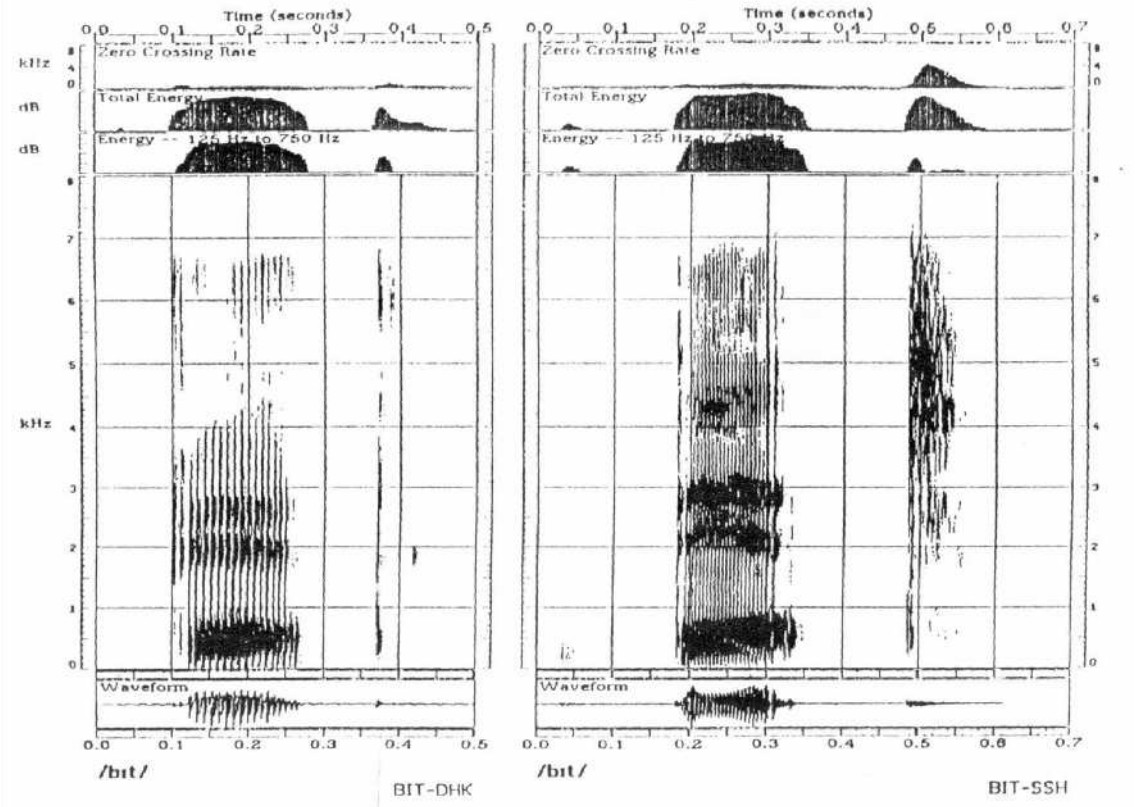
Figure 9.7 Upper part: spectrogram of a child saying *I do art projects*. Lower part: my repetition of the same phrase.

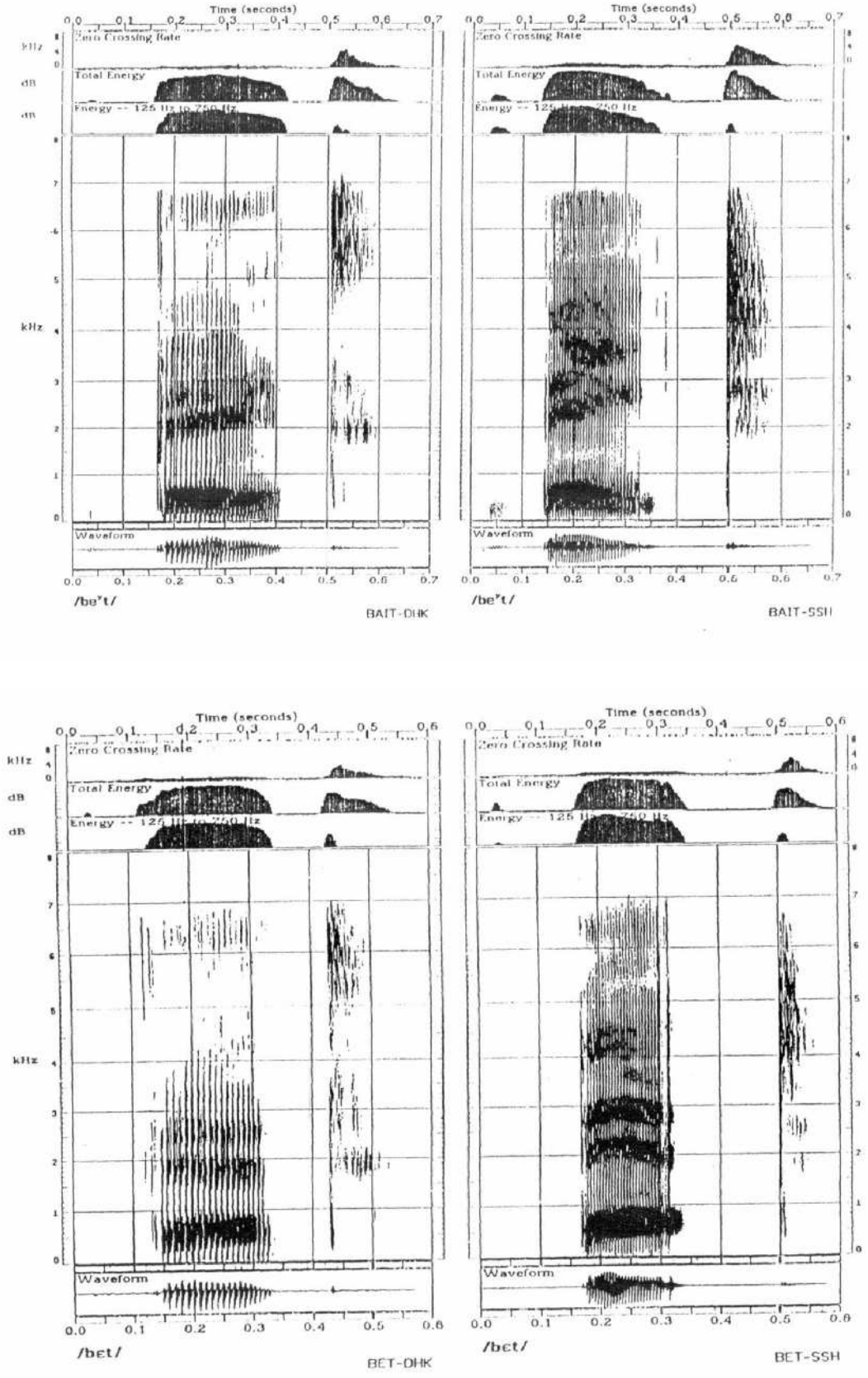
تصویر 1 – نمونه از یک اسپکتروگرام

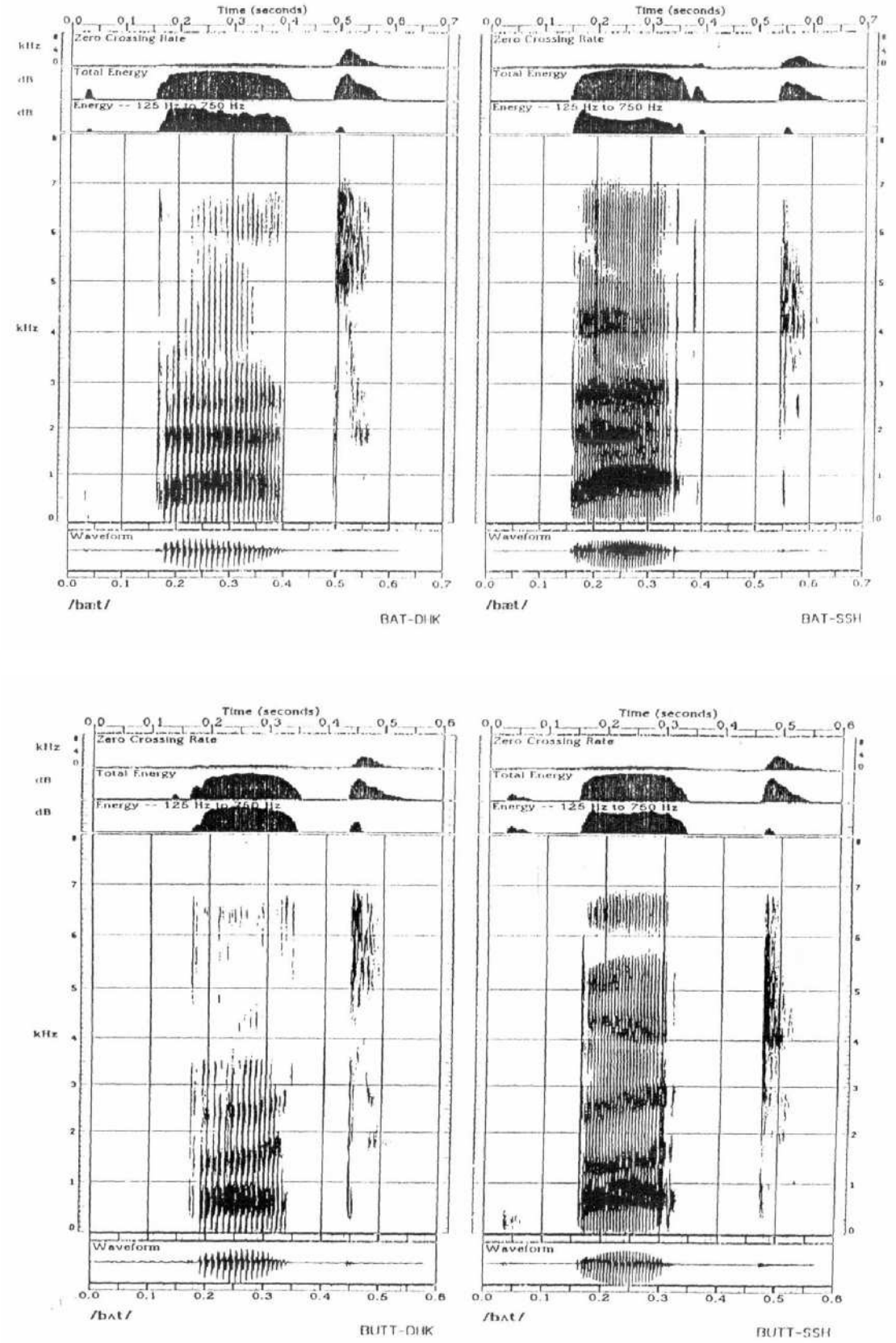
3– اسپکتروگرام ها

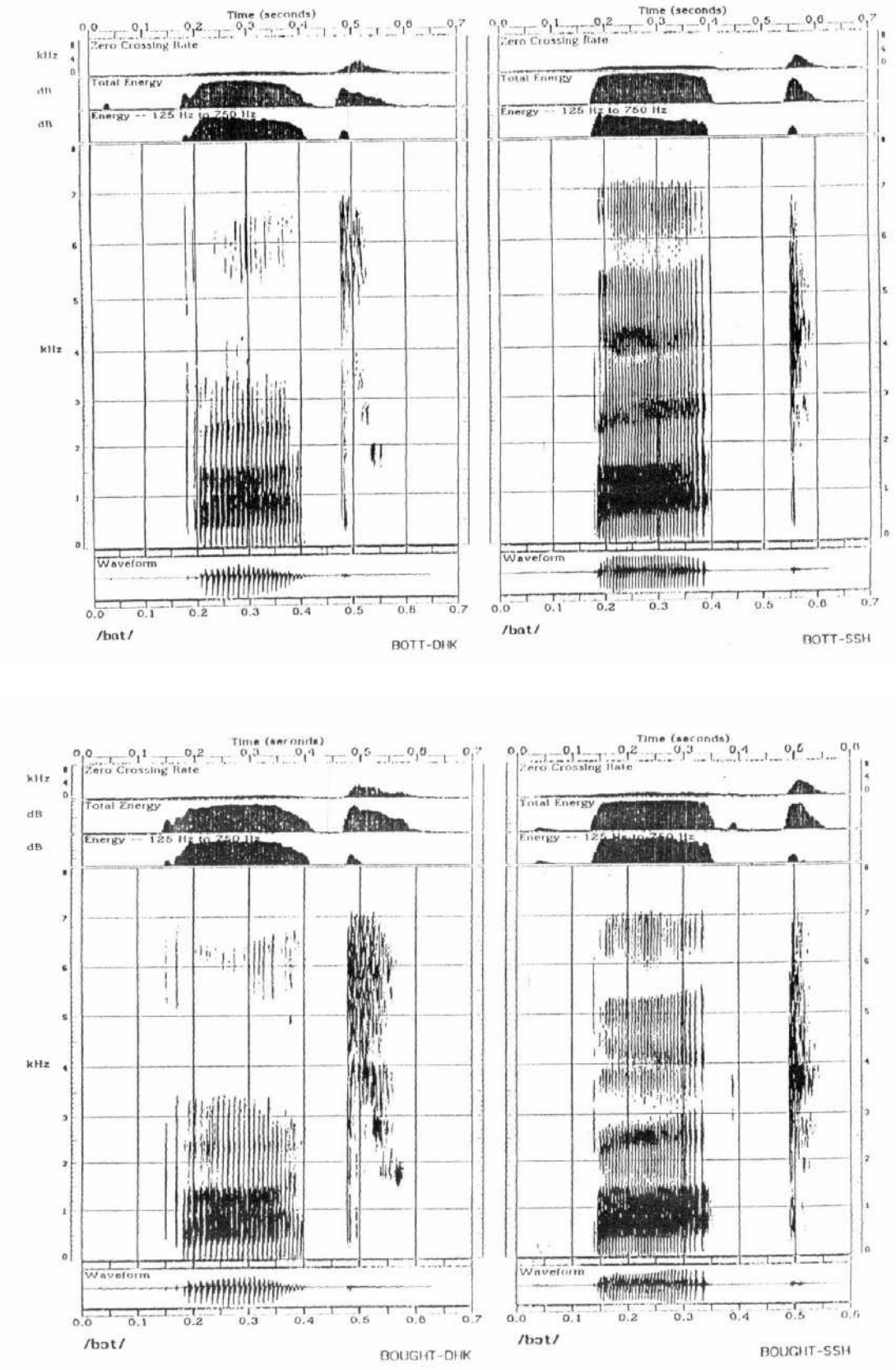
- اسپکتروگرام

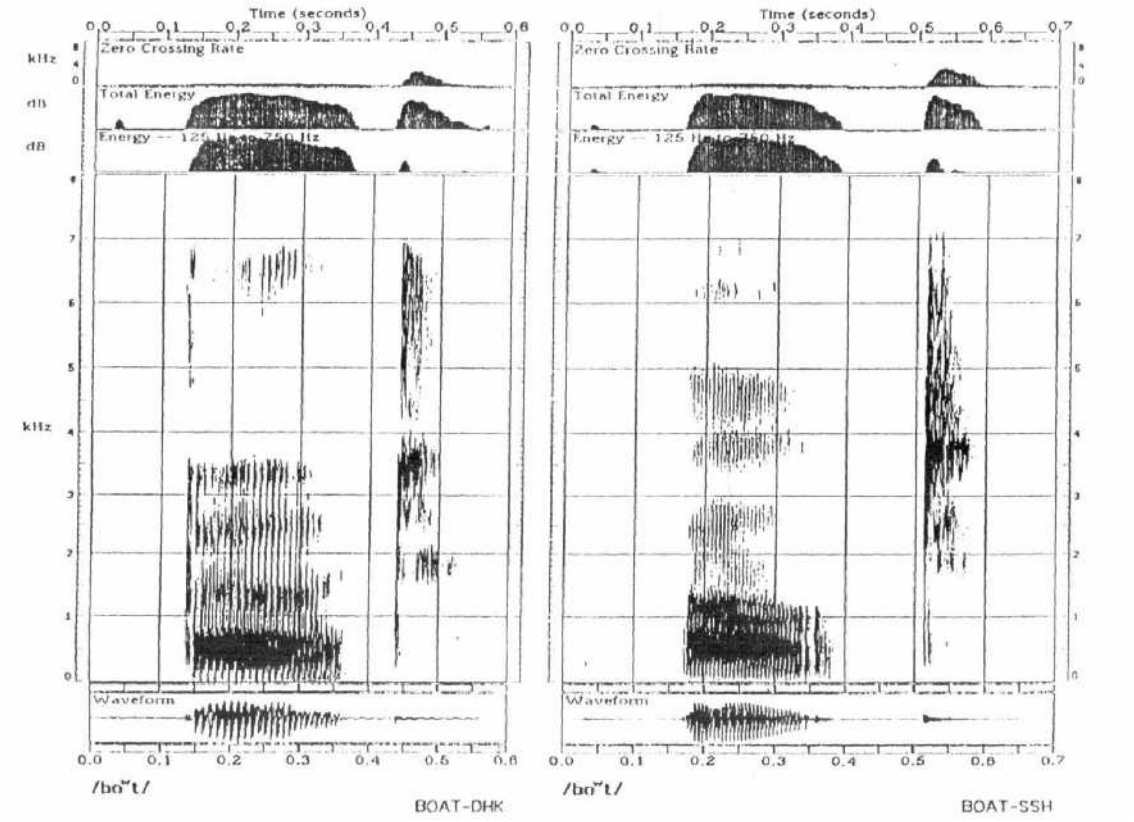

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
/bi't/
BEET-DHK
/bi't/
BEET-SSH

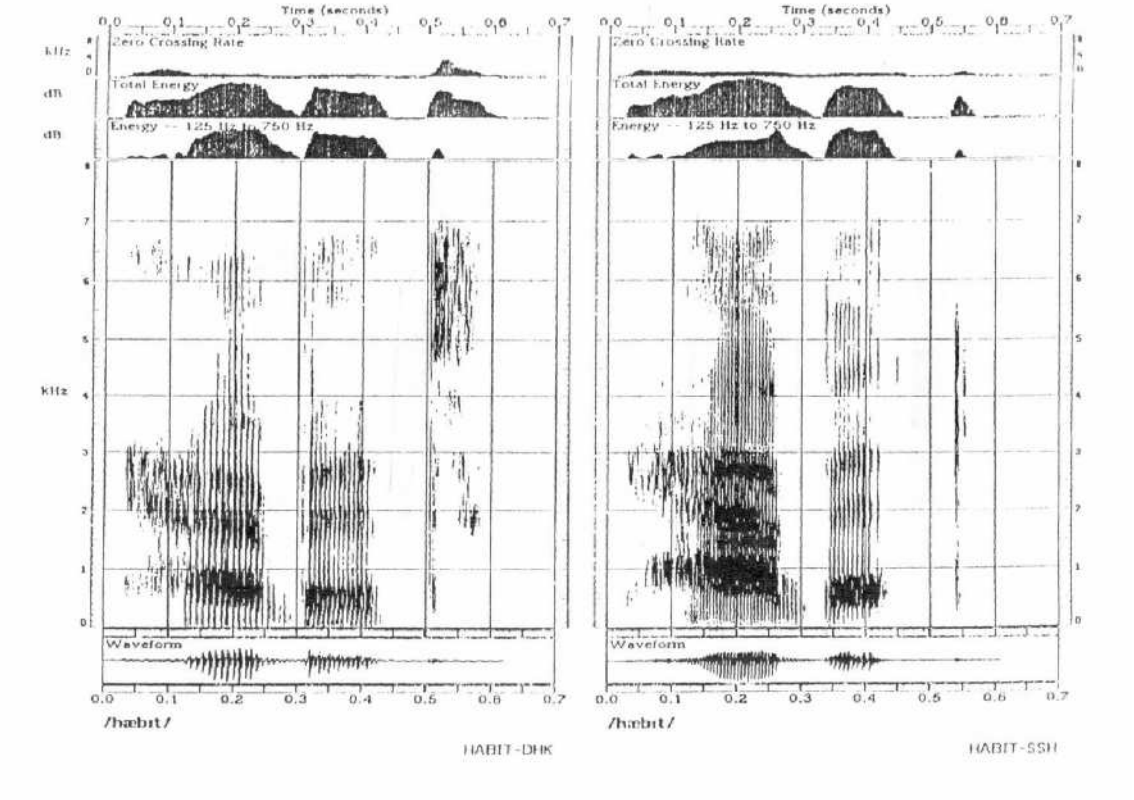

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
/bɪt/
BIT-DHK
/bɪt/
BIT-SSH

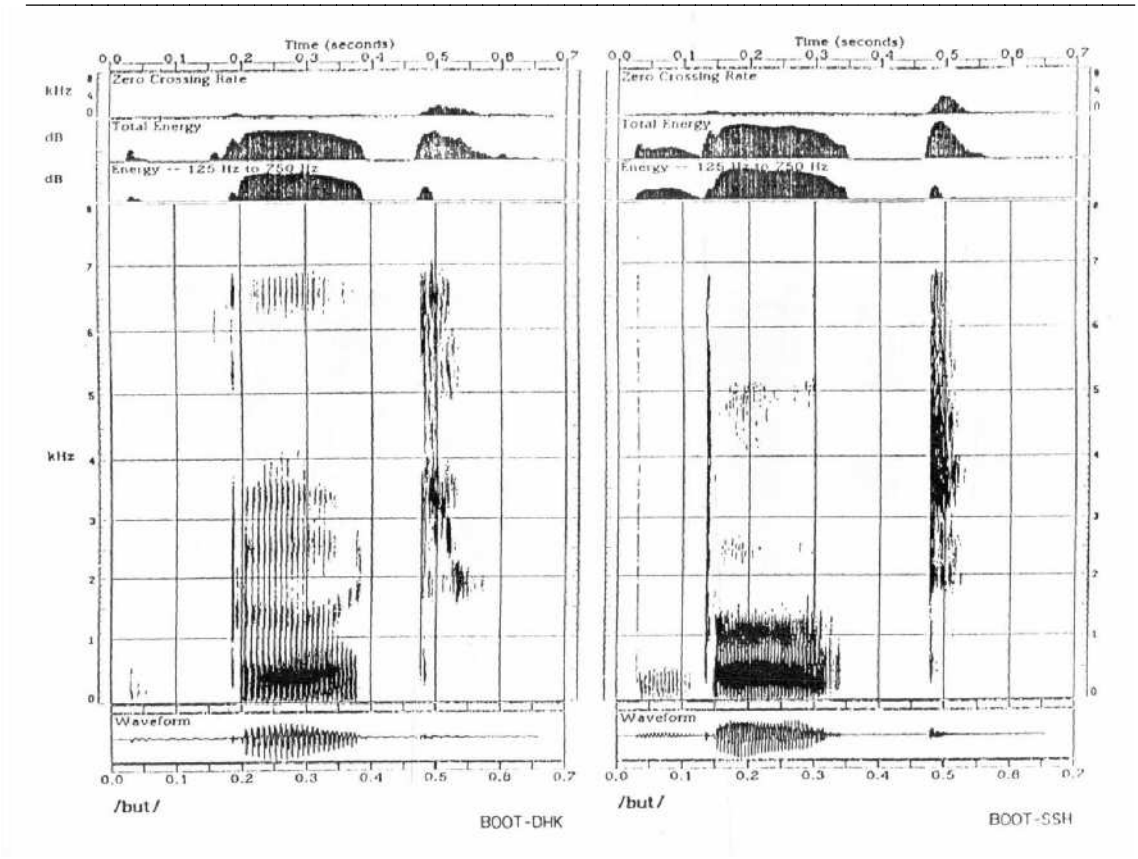

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
/beᵊt/
BAIT-OHK
/beᵊt/
BAIT-SSH
/bɛt/
BET-OHK
/bɛt/
BET-SSH

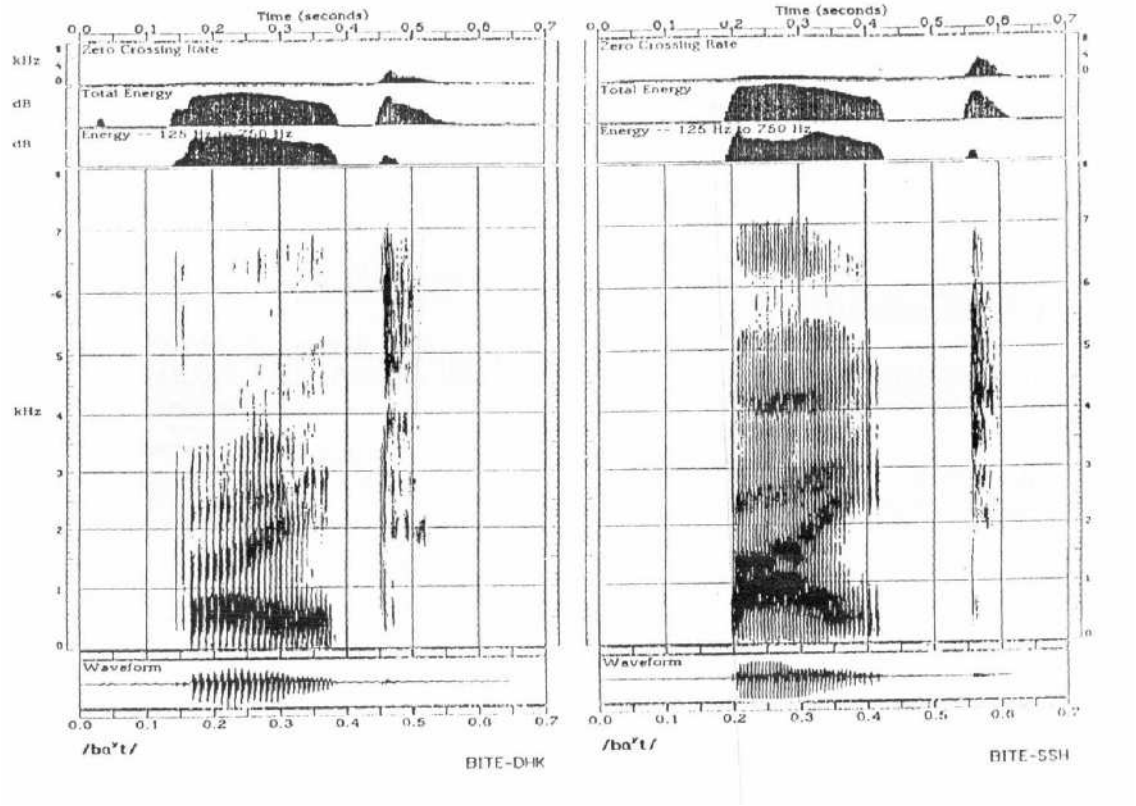

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
/bat/
BAT-DIK
/bæt/
BAT-SSH
/bʌt/
BUTT-DIK
/bʌt/
BUTT-SSH

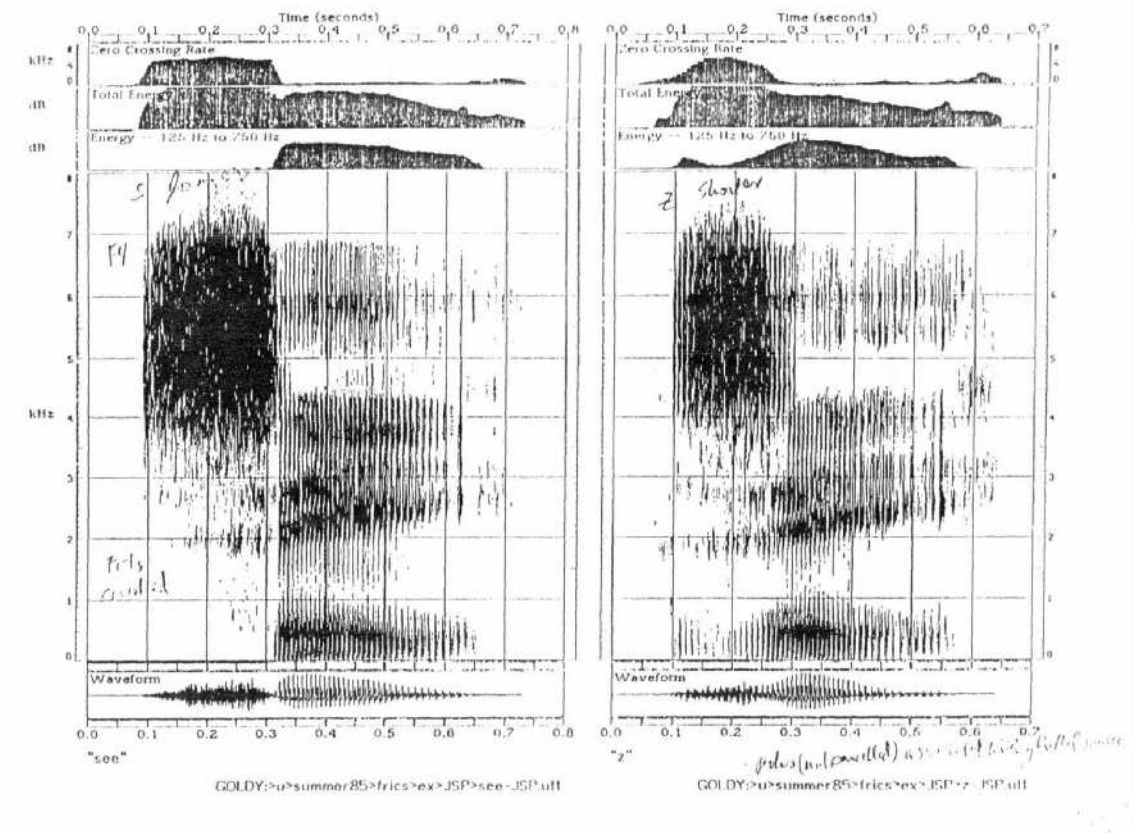

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
/bat/
BOTT-DHK
BOTT-SSH
/bɔt/
BOUGHT-DHK
BOUGHT-SSH

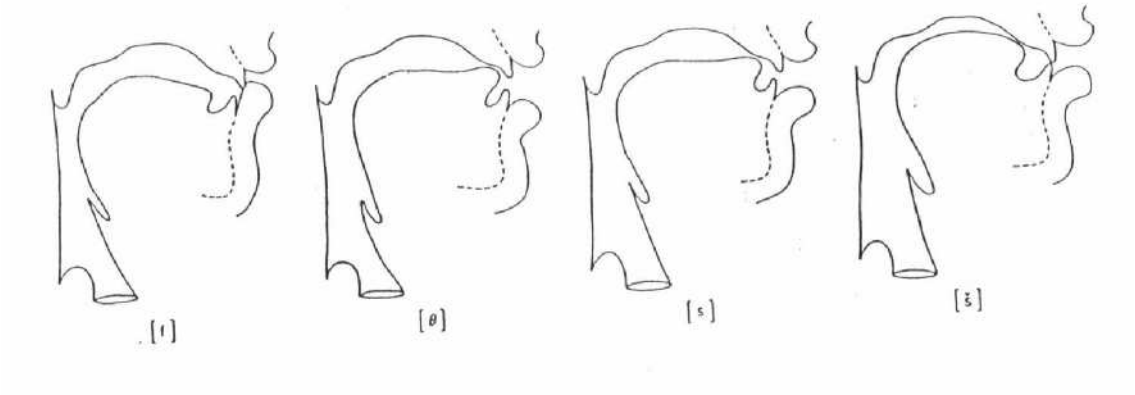

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
/boʷt/
BOAT-DHK
BOAT-SSH

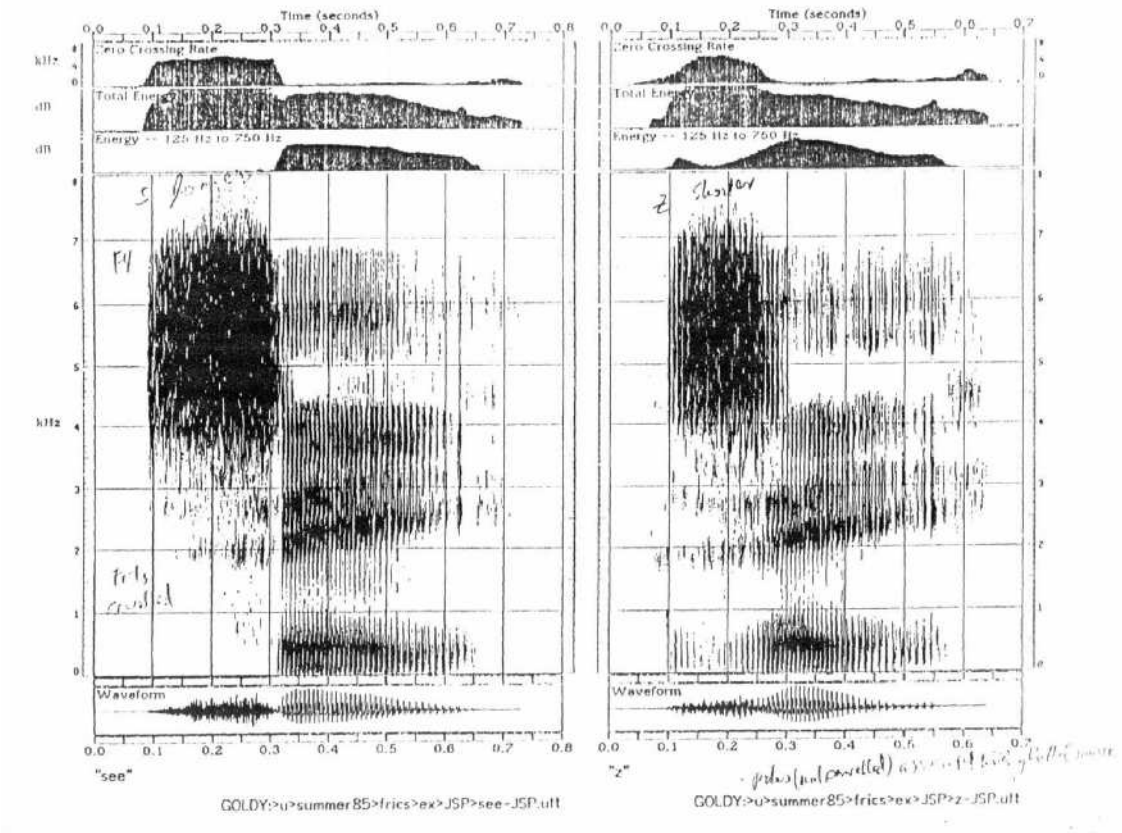

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
/hæbɪt/
HABIT-DHK
HABIT-SSH


Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
/but/
BOOT-DHK
/but/
BOOT-SSH


Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
/baˣt/
BITE-DHK
/baˣt/
BITE-SSH


Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
"see"
"z"
GOLDY:>u>summer85>frics>ex>JSP>see-JSP.ulf
GOLDY:>u>summer85>frics>ex>JSP>z-JSP.ulf


[f]
[θ]
[s]
[ʒ]


Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
"see"
"z"
GOLDY:>u>summer85>frics>ex>JSP>see-JSP.ult
GOLDY:>u>summer85>frics>ex>JSP>z-JSP.ult

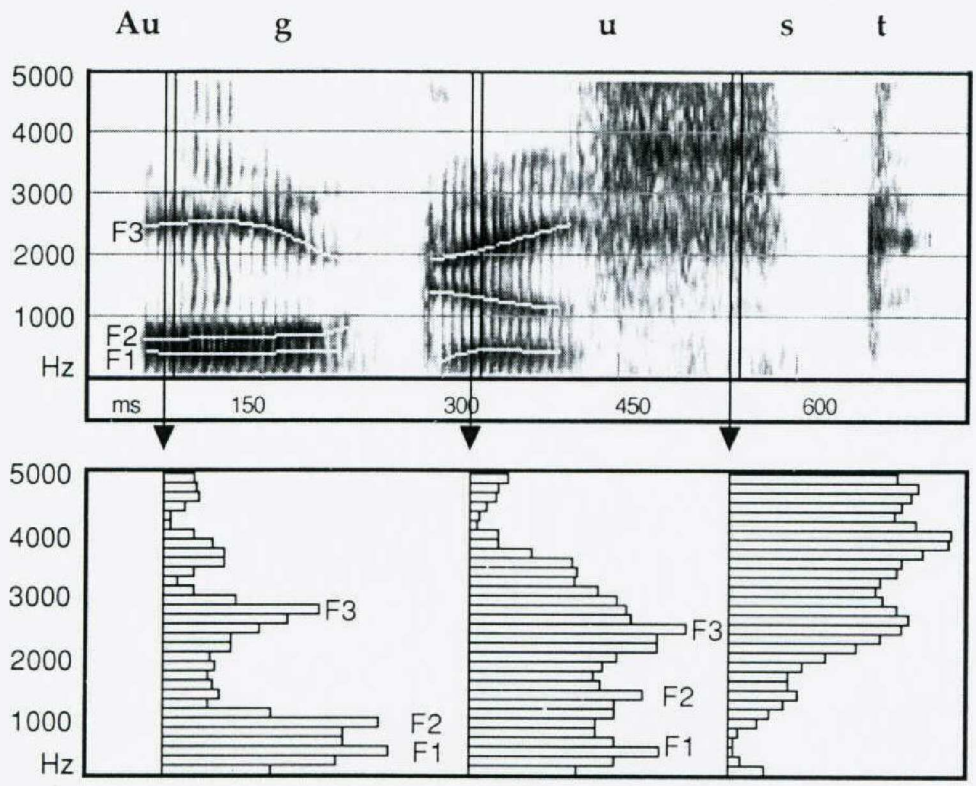


Figure 9.1 Upper part: a spectrogram of the word *August*. Lower part: three out of a possible 75 graphs representing the energy in a one-hundredth of a second section of this word.

خودآزمایی

یکی از روزهای هفته در شکل زیر تلفظ شده است:

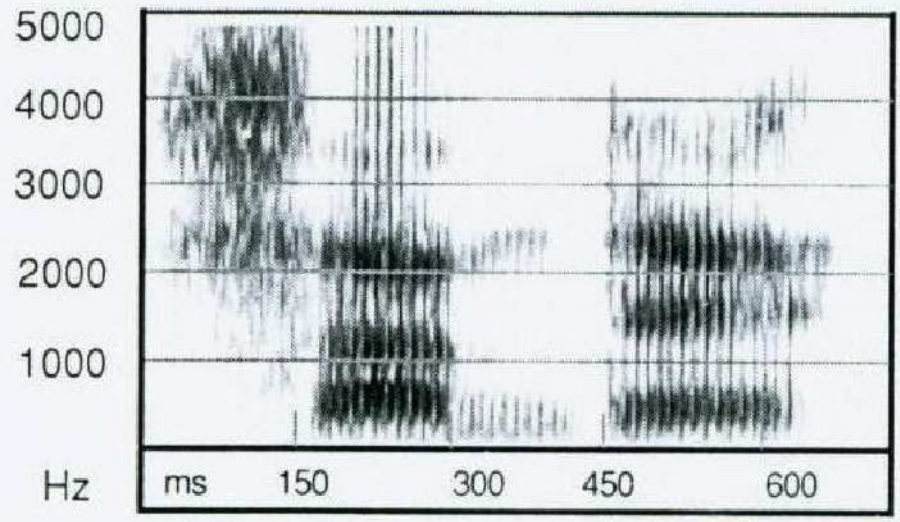


Figure 9.2 A spectrogram of one of the days of the week.

در اسپکتروگرام زیر جمله با It's شروع شده است و بیان کننده زمان است.

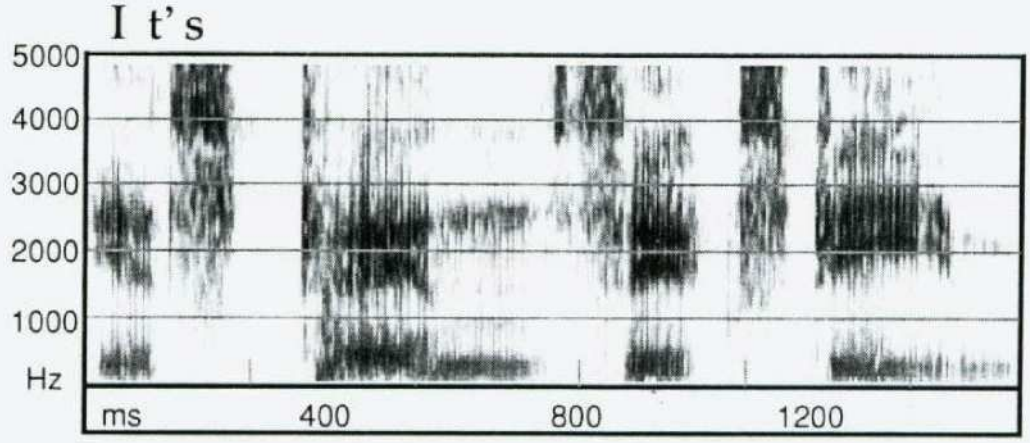


Figure 9.3 A spectrogram beginning with *It's*, announcing the time.

4– خلاصه و نتیجه گیری:

در این فصل با چند نمونه اسپکتروگرام آشنا شدیم.

10 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”

2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

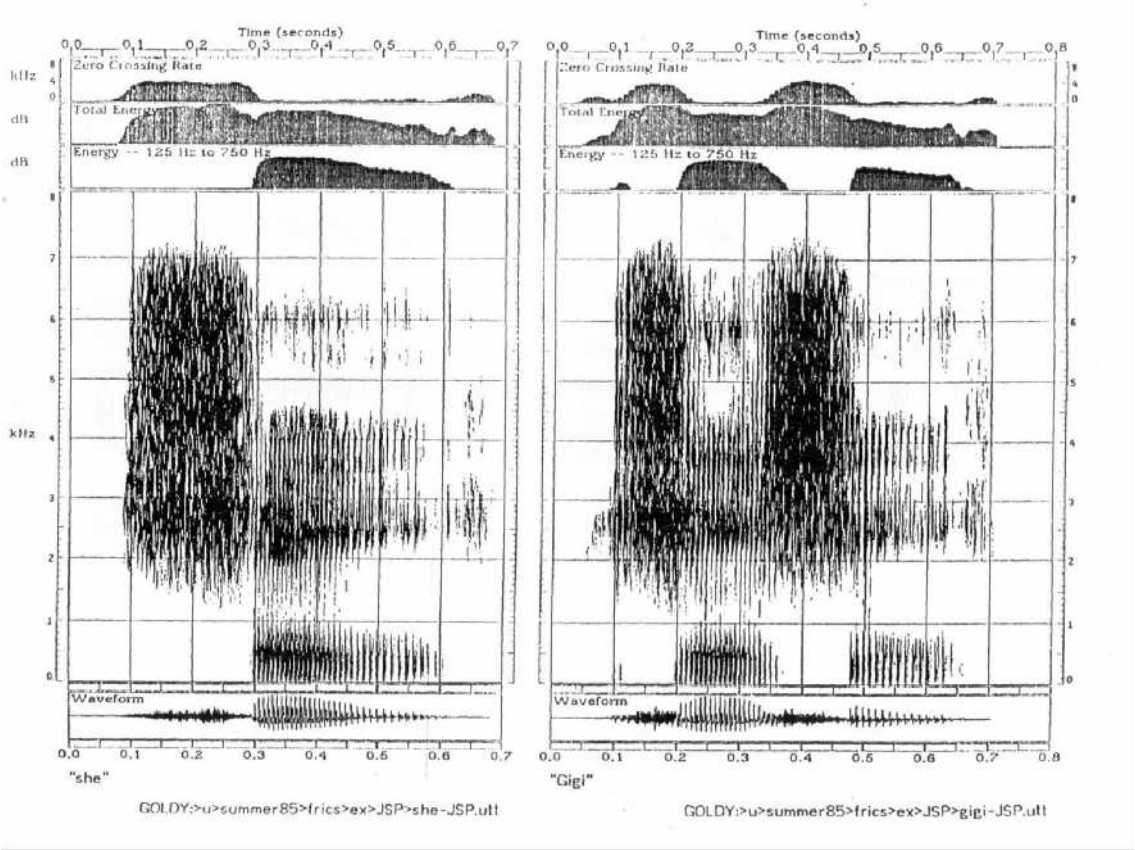
1– مقدمه

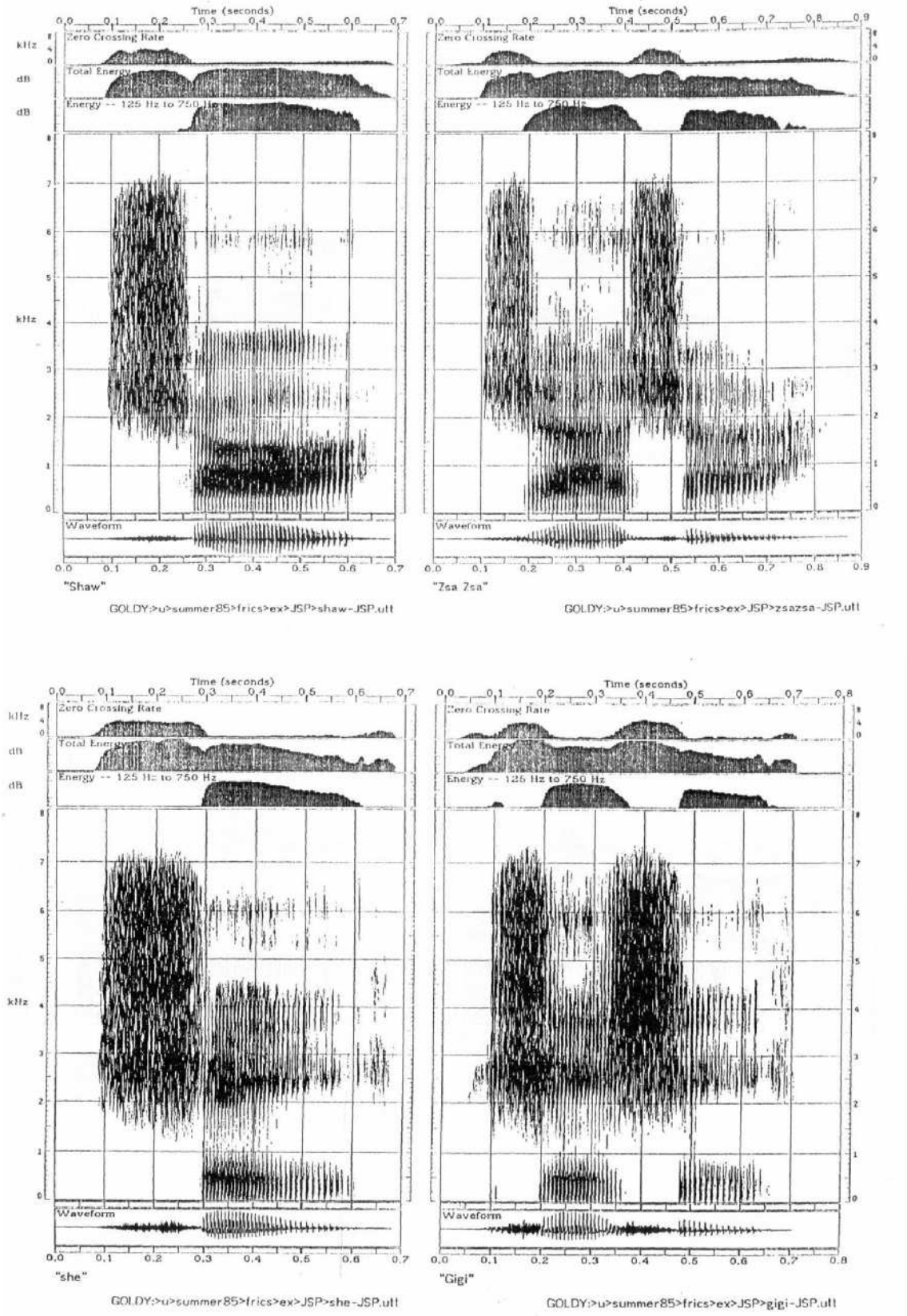
اهداف درس:

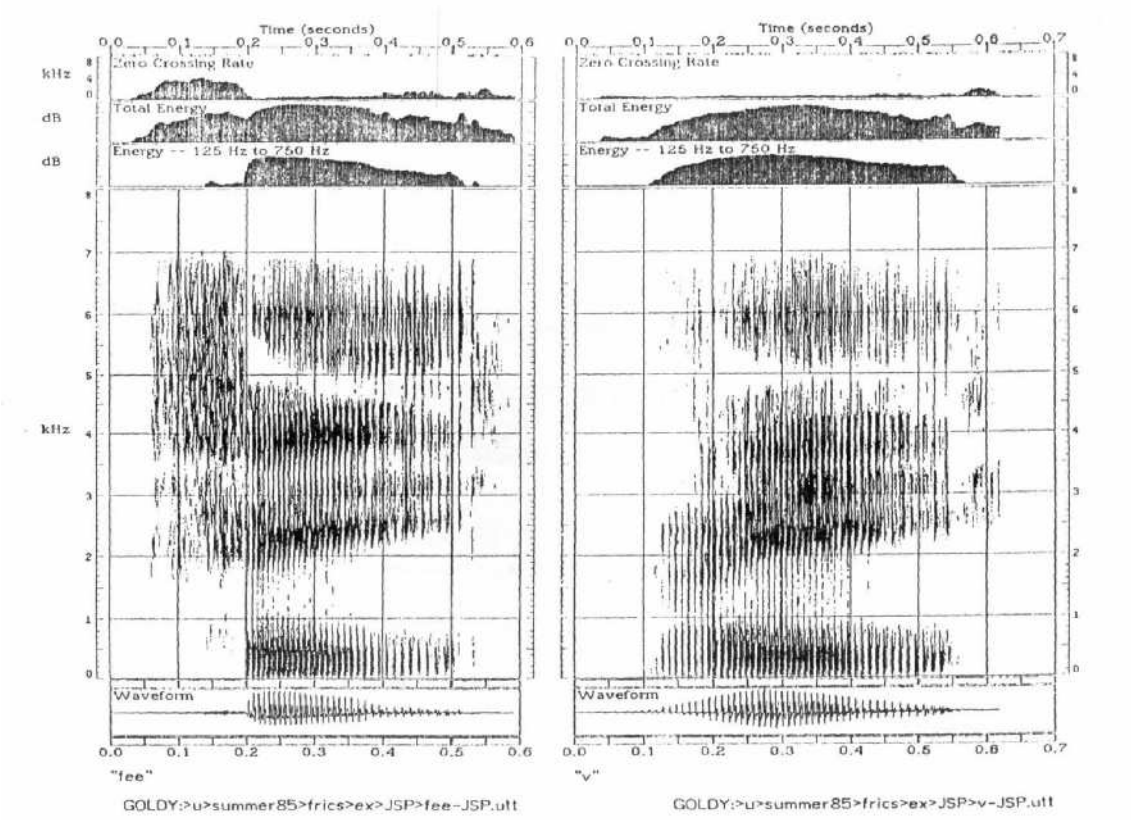
آشنایی با مفهوم اسپکتروگرام ها

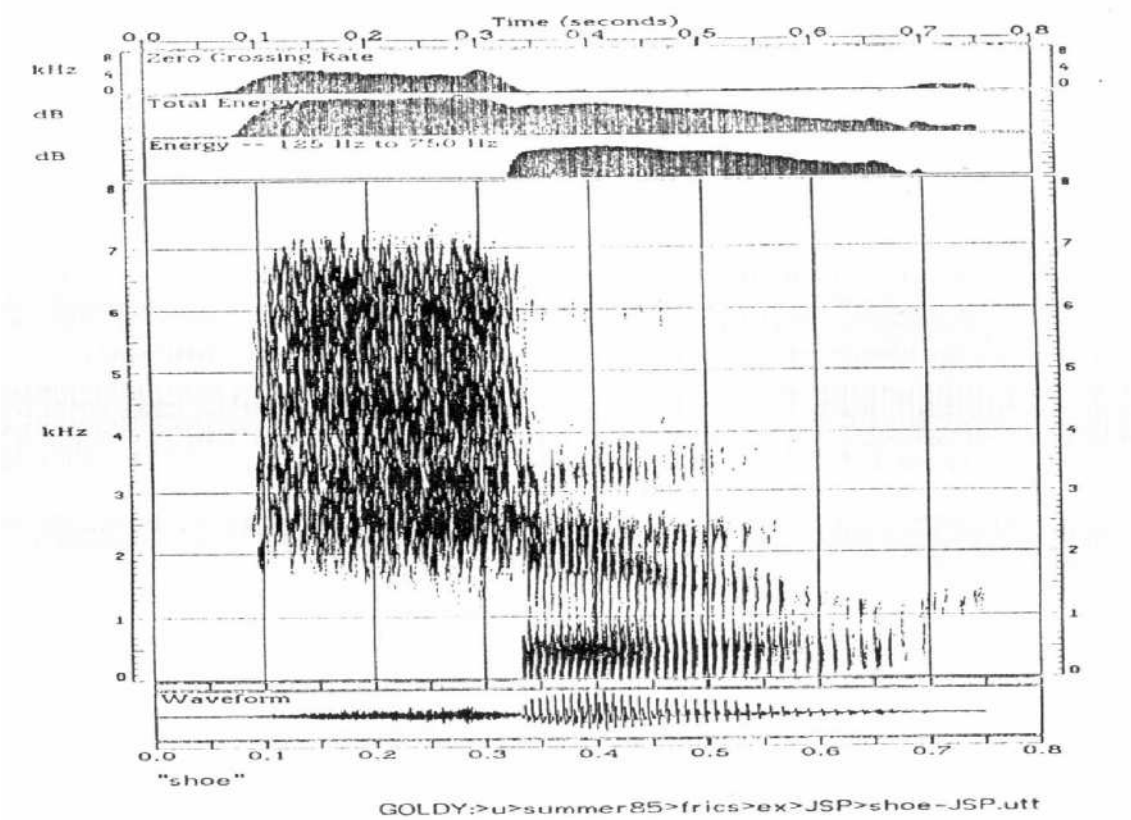
آشنایی با نحوه خواندن اسپکتروگرام ها

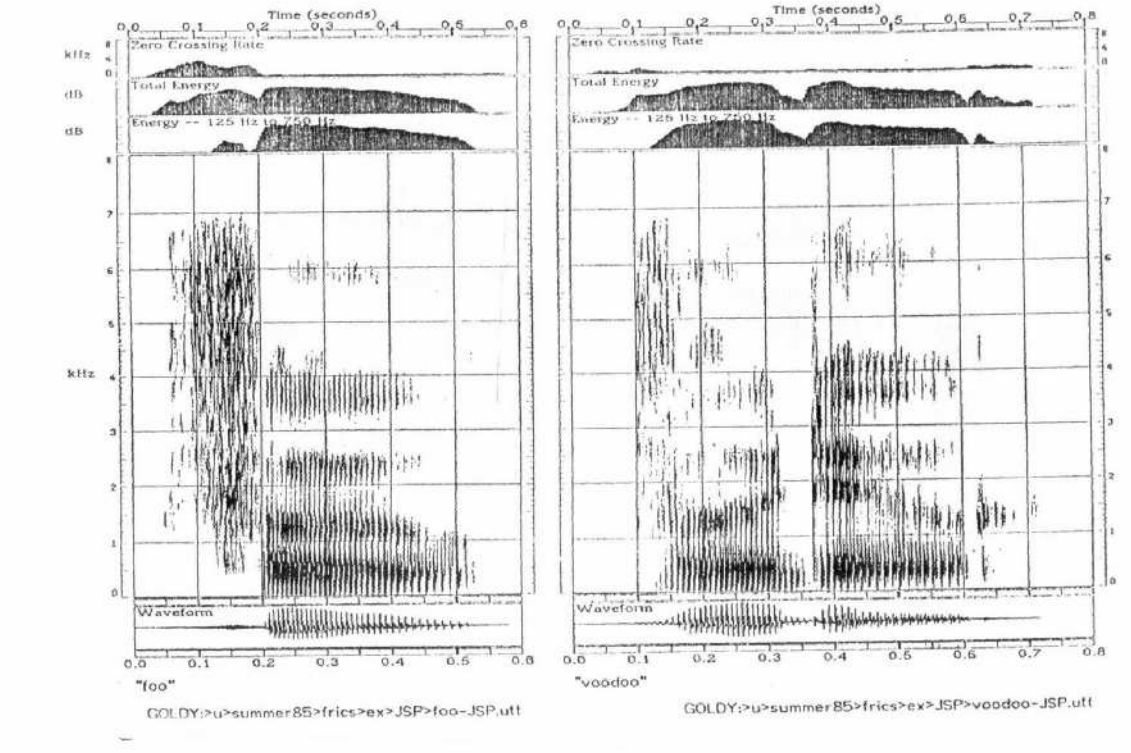
2– اسپکتروگرام ها

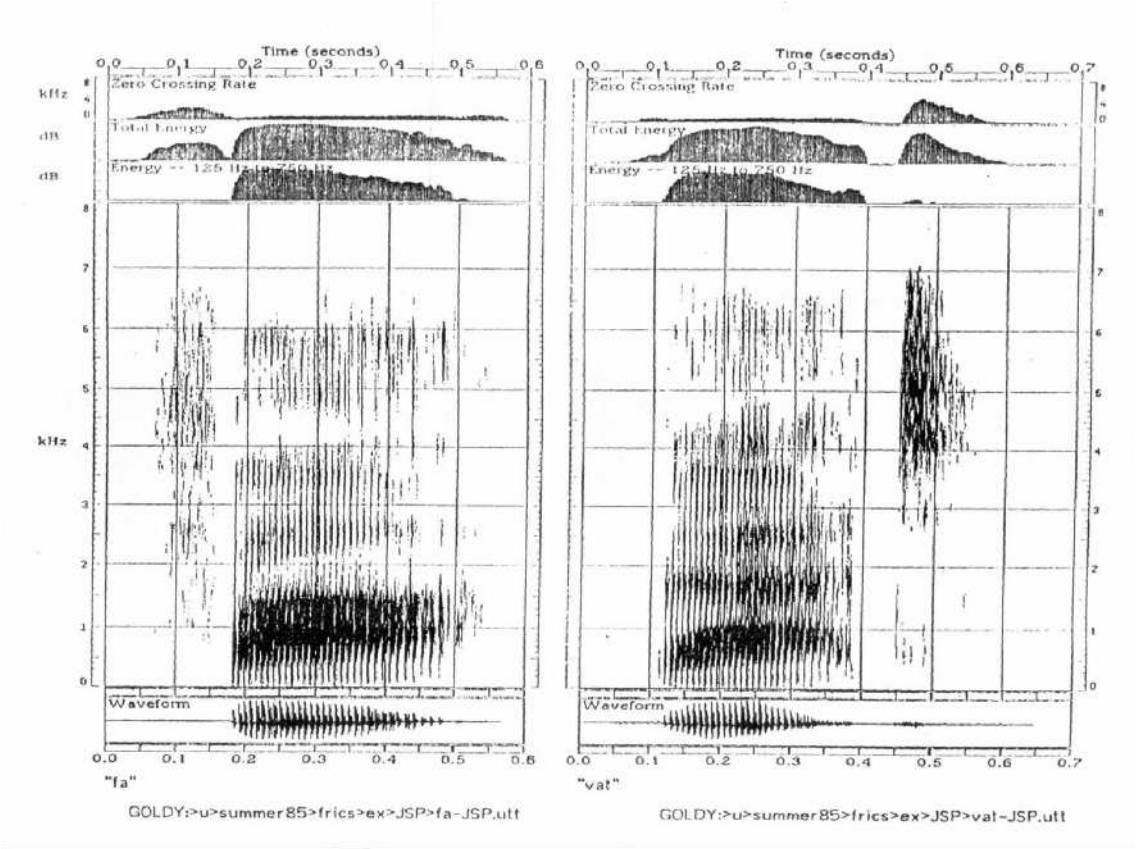


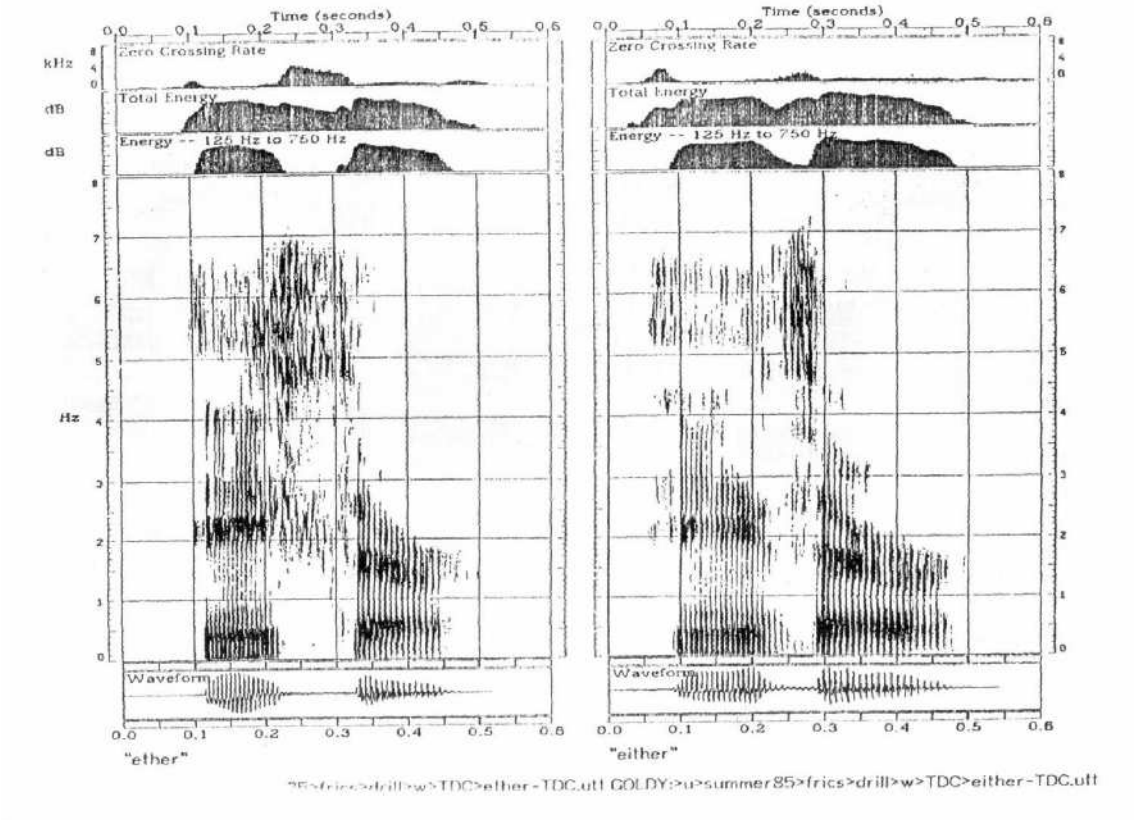

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
"Shaw"
GOLDY:>u>summer85>frics>ex>JSP>shaw-JSP.utt
"Zsa Zsa"
GOLDY:>u>summer85>frics>ex>JSP>zsazsa-JSP.utt
"she"
GOLDY:>u>summer85>frics>ex>JSP>she-JSP.utt
"Gigi"
GOLDY:>u>summer85>frics>ex>JSP>gigi-JSP.utt

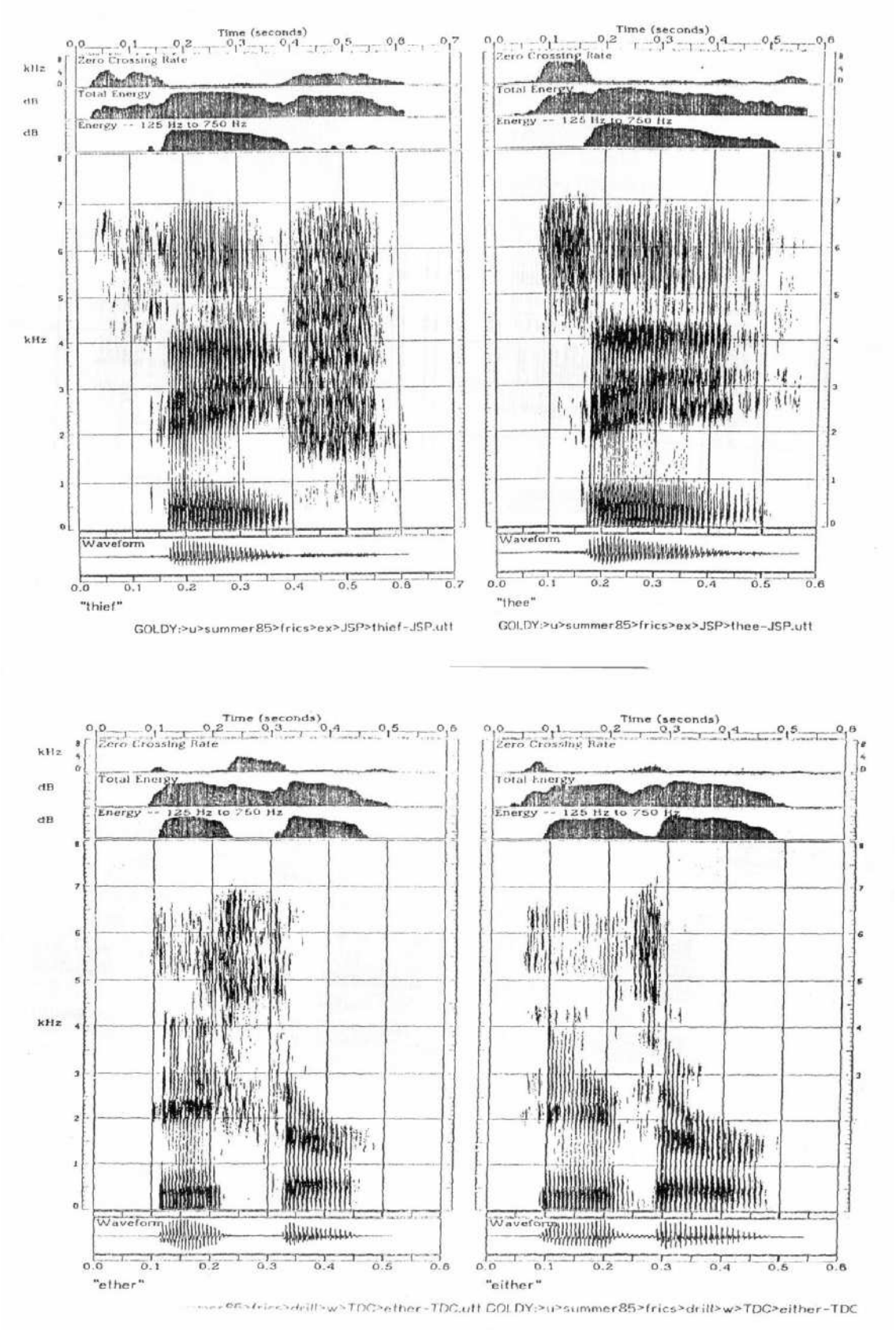

Time (seconds)
Zero Crossing Rate
kHz
Total Energy
dB
Energy -- 125 Hz to 750 Hz
dB
kHz
Waveform
"fee"
GOLDY:>u>summer85>frics>ex>JSP>fee-JSP.utt
Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
"v"
GOLDY:>u>summer85>frics>ex>JSP>v-JSP.utt

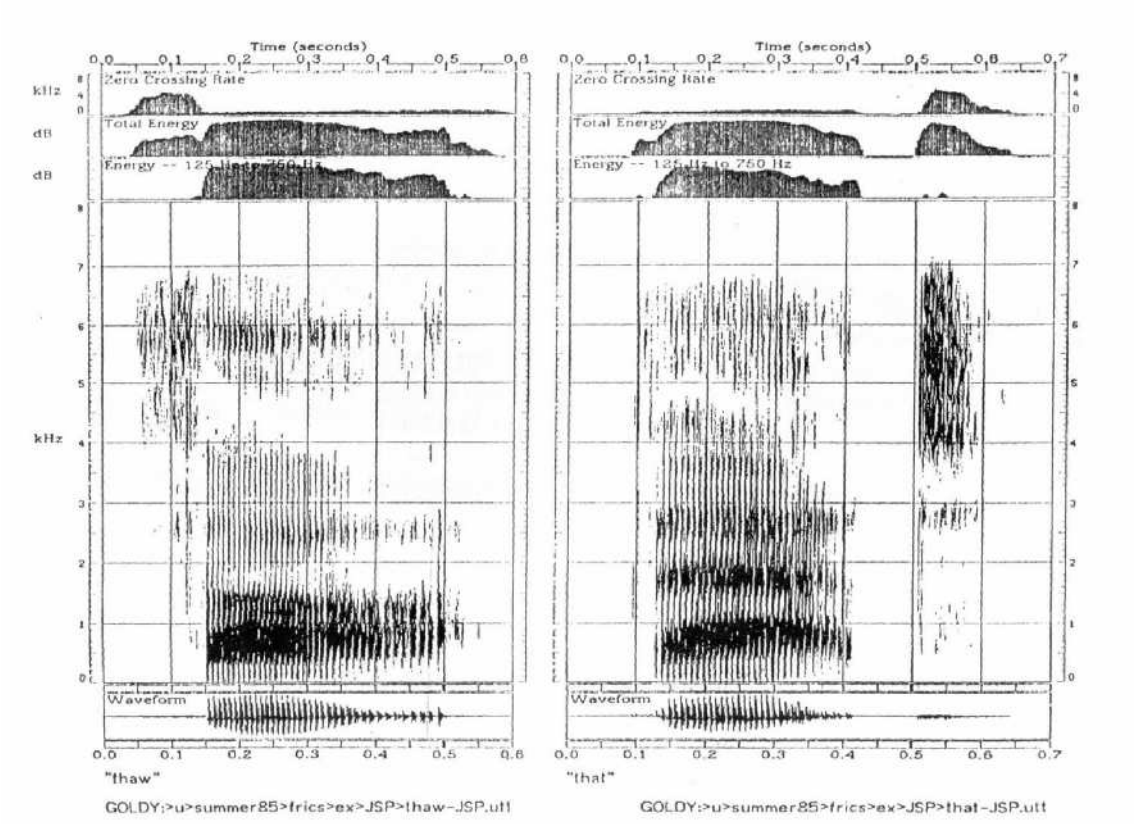

Time (seconds)
Zero Crossing Rate
kHz
Total Energy
dB
Energy -- 125 Hz to 7500 Hz
dB
kHz
Waveform
"shoe"
GOLDY:>u>summer85>frics>ex>JSP>shoe-JSP.utt

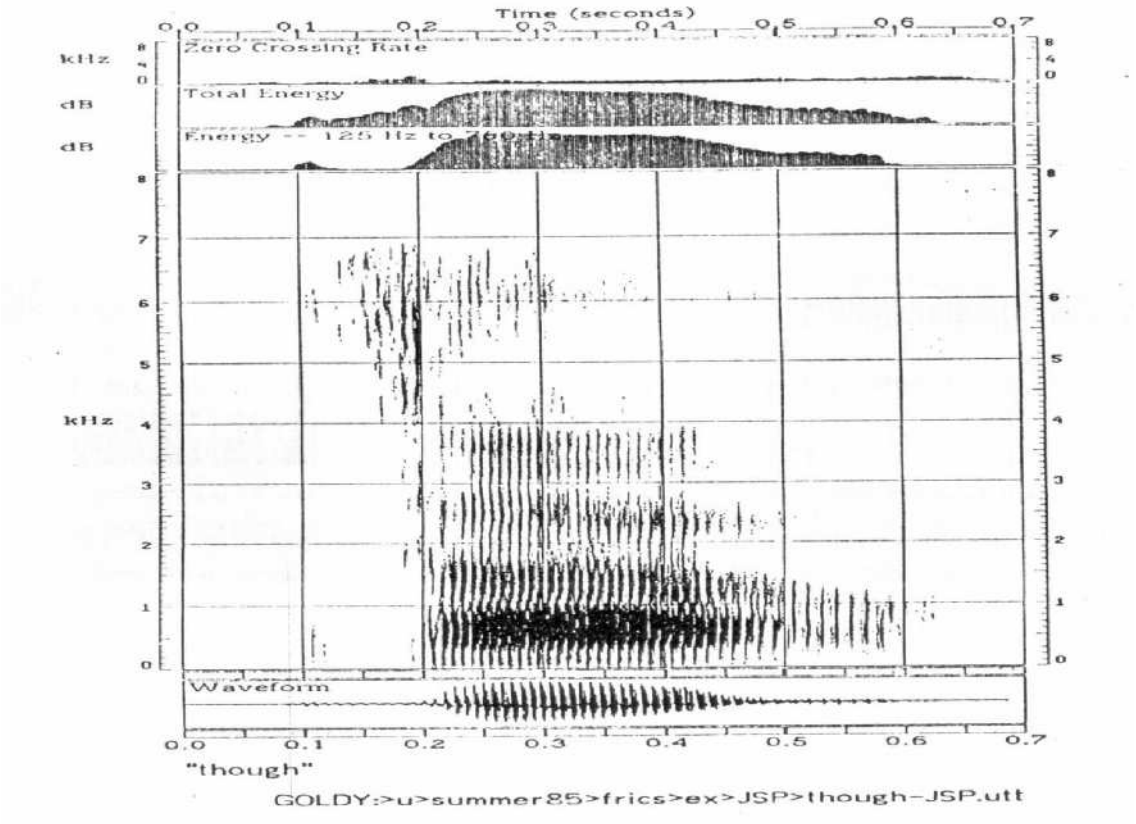

Time (seconds)
Zero Crossing Rate
kHz
Total Energy
dB
Energy -- 125 Hz to 7500 Hz
dB
kHz
Waveform
"foo"
GOLDY:>u>summer85>frics>ex>JSP>foo-JSP.utt
Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 7500 Hz
Waveform
"voodoo"
GOLDY:>u>summer85>frics>ex>JSP>voodoo-JSP.utt

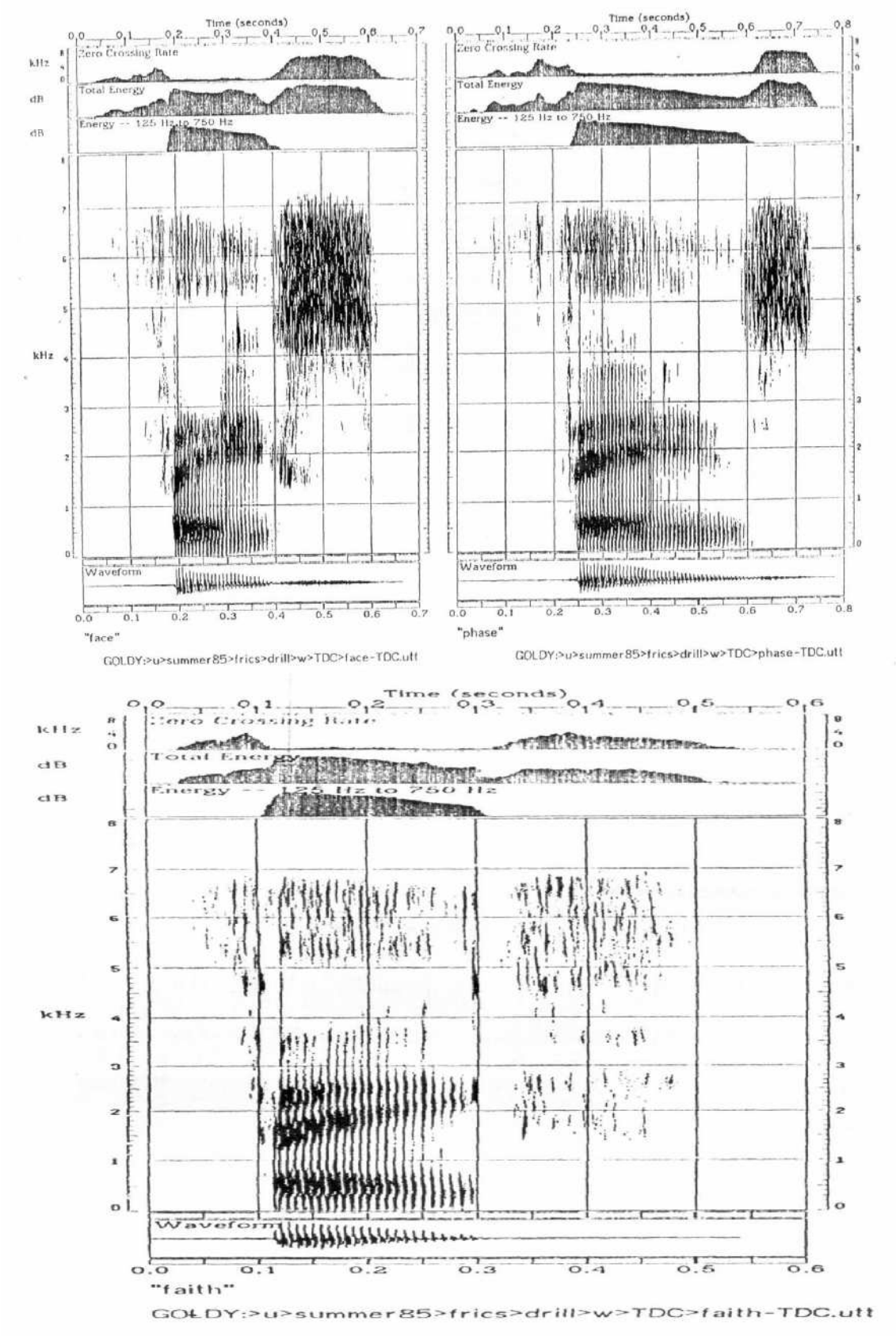

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
"fa"
"vat"
GOLDY:>u>summer85>frics>ex>JSP>fa-JSP.utt
GOLDY:>u>summer85>frics>ex>JSP>vat-JSP.utt

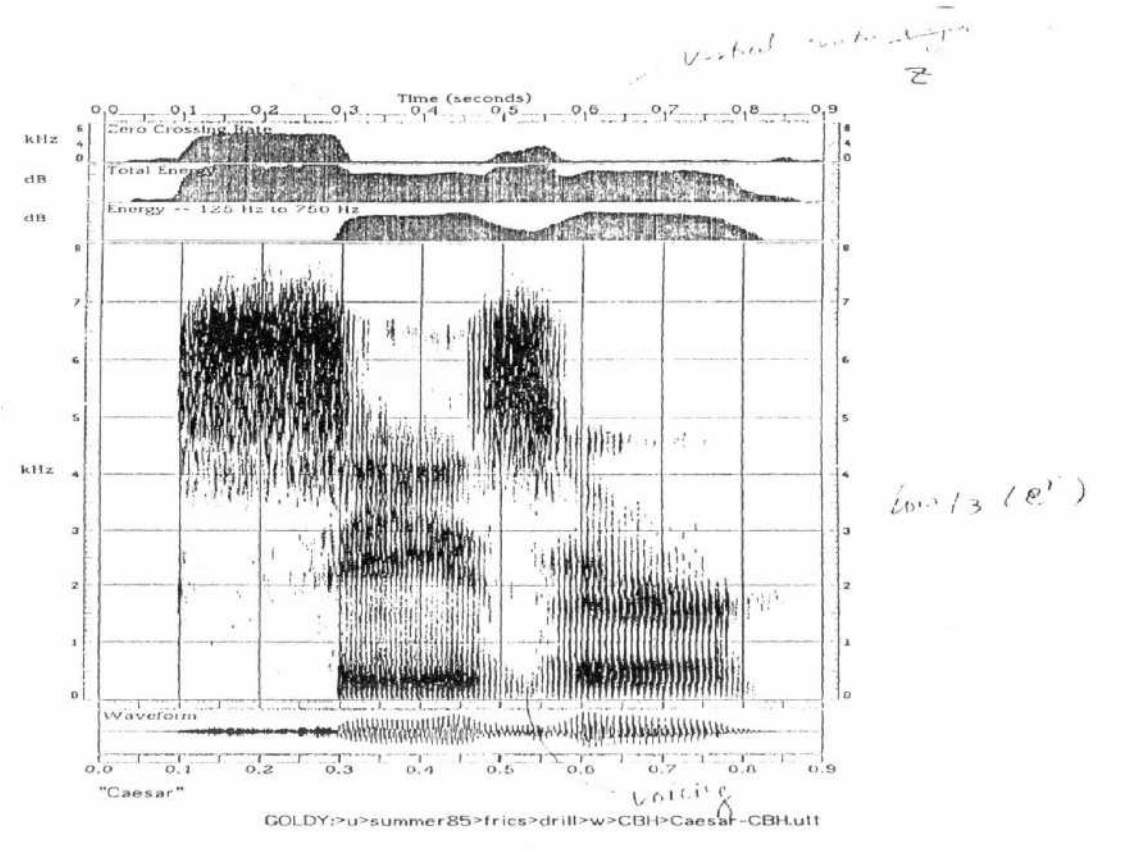

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
"ether"
"either"
GOLDY:>u>summer85>frics>drill>w>TDC>either-TDC.utt

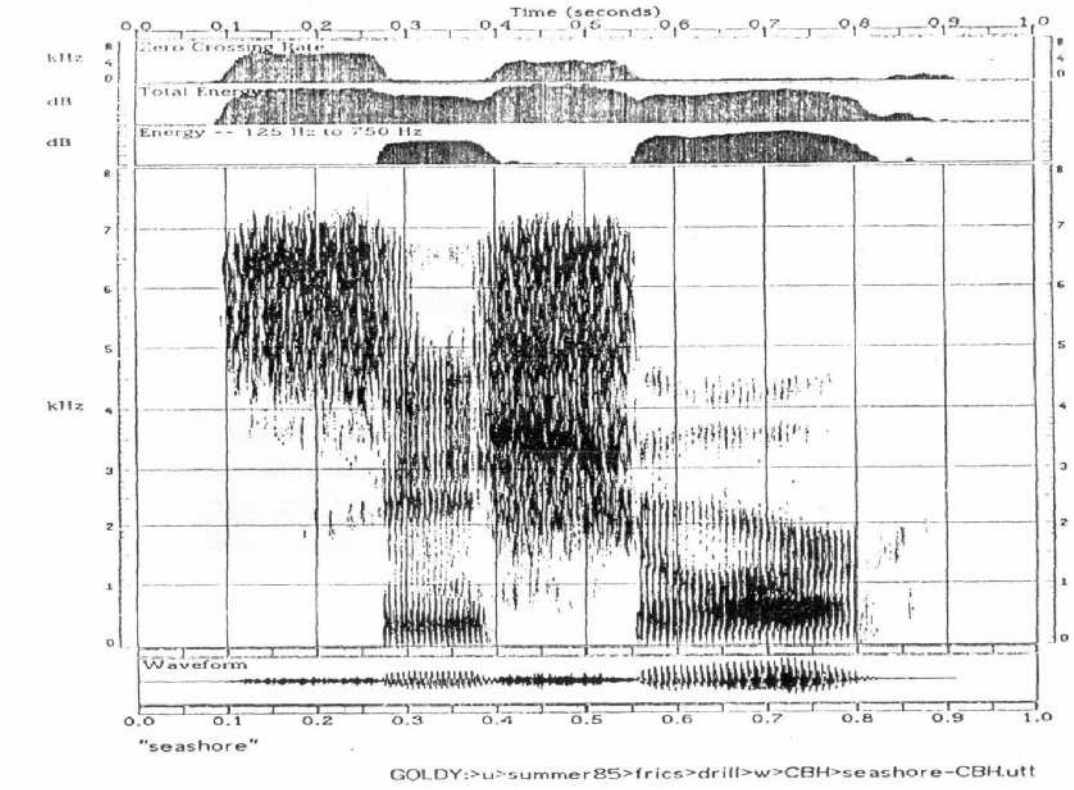

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
"thief"
GOLDY:>u>summer85>frics>ex>JSP>thief-JSP.utt
"thee"
GOLDY:>u>summer85>frics>ex>JSP>thee-JSP.utt
"ether"
"either"
GOLDY:>u>summer85>frics>drill>w>TDC>either-TDC

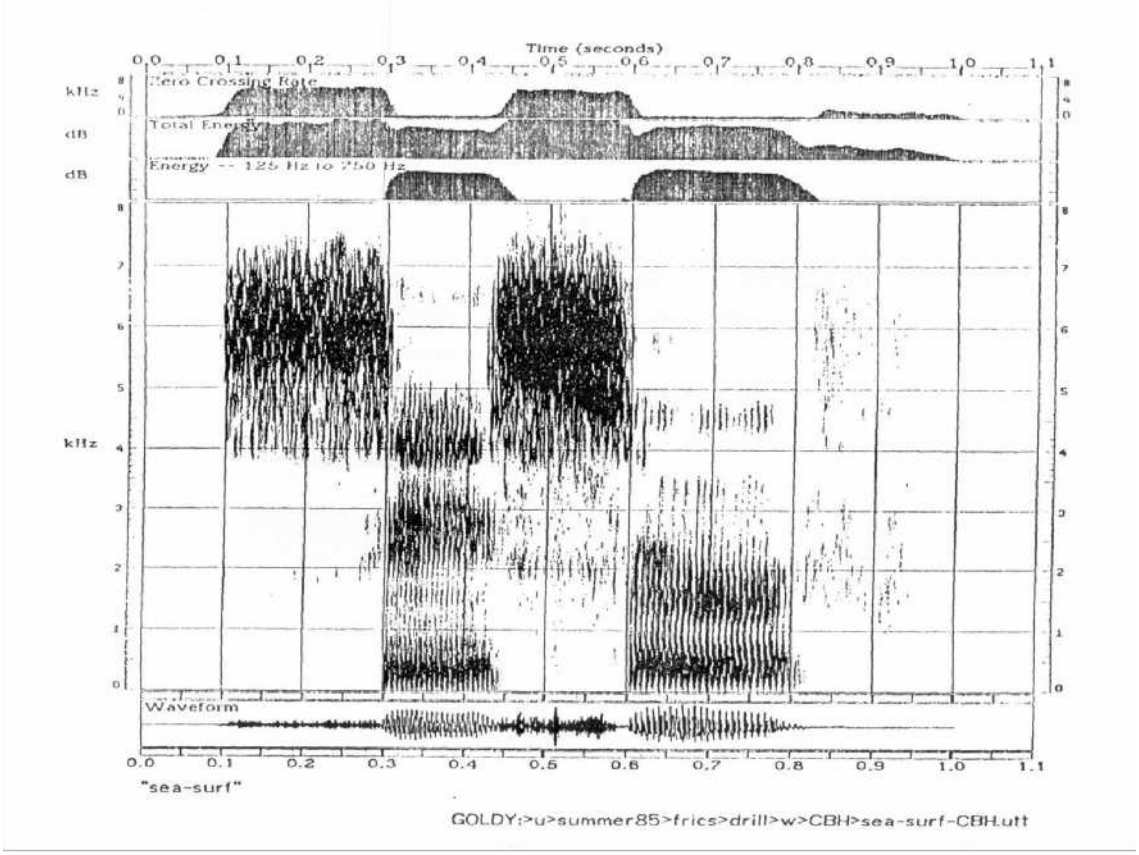

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
"thaw"
GOLDY:>u>summer85>frics>ex>JSP>thaw-JSP.utt
"that"
GOLDY:>u>summer85>frics>ex>JSP>that-JSP.utt

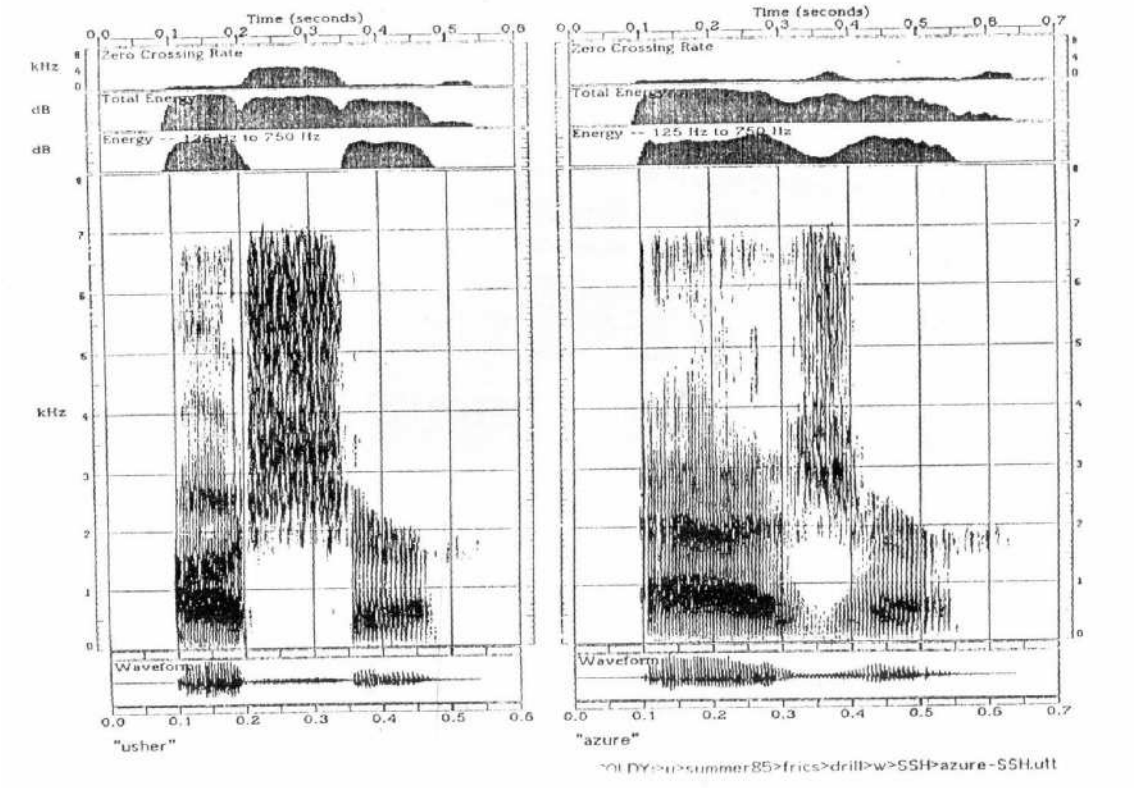

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
Waveform
kHz
dB
"though"
GOLDY:>u>summer85>frics>ex>JSP>though-JSP.utt

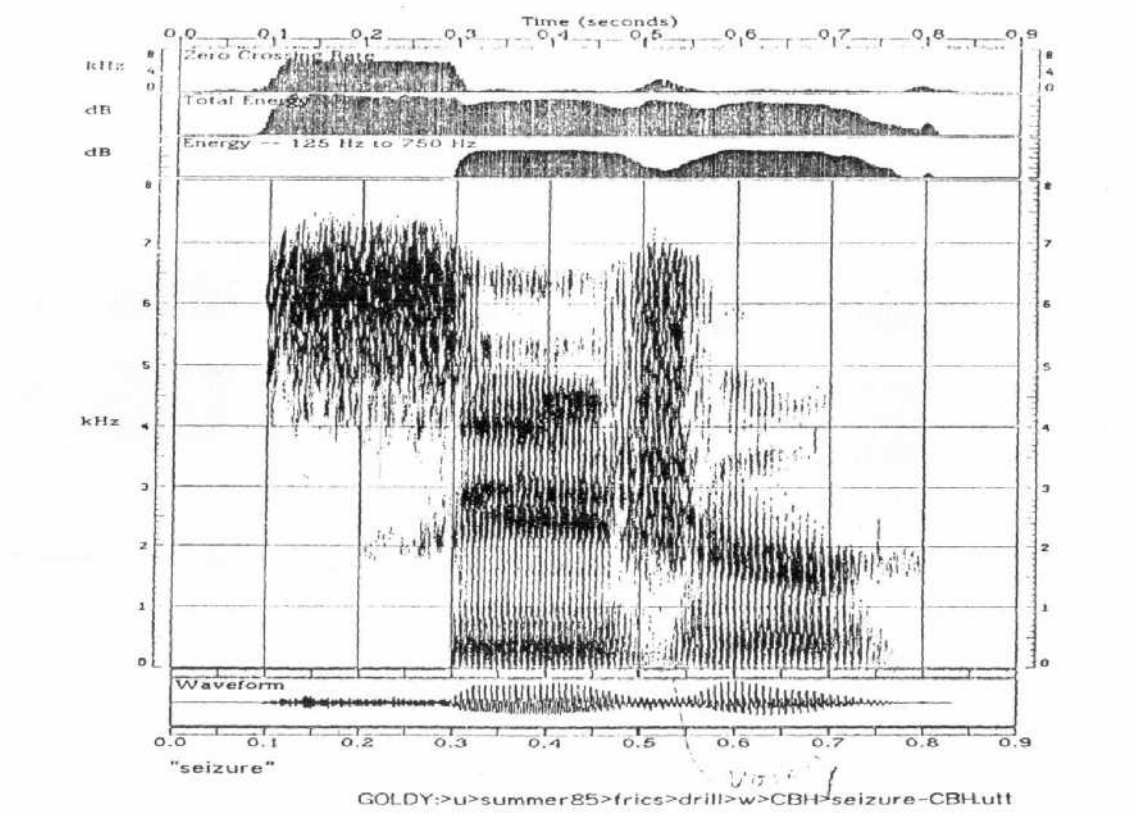

Time (seconds)
Zero Crossing Rate
Total Energy
Energy -- 125 Hz to 750 Hz
kHz
dB
Waveform
"face"
GOLDY:>u>summer85>frics>drill>w>TDC>face-TDC.utt
"phase"
GOLDY:>u>summer85>frics>drill>w>TDC>phase-TDC.utt
"faith"
GOLDY:>u>summer85>frics>drill>w>TDC>faith-TDC.utt


Time (seconds)
Zero Crossing Rate
kHz
Total Energy
dB
Energy -- 125 Hz to 750 Hz
dB
kHz
Waveform
"Caesar"
GOLDY:>u>summer85>frics>drill>w>CBH>Caesar-CBH.utt


Time (seconds)
Zero Crossing Rate
kHz
Total Energy
dB
Energy -- 125 Hz to 750 Hz
dB
kHz
Waveform
"seashore"
GOLDY:>u>summer85>frics>drill>w>CBH>seashore-CBH.utt


Time (seconds)
Zero Crossing Rate
kHz
Total Energy
dB
Energy -- 125 Hz to 750 Hz
dB
kHz
Waveform
"sea-surf"
GOLDY:>u>summer85>frics>drill>w>CBH>sea-surf-CBH.utt


Time (seconds)
Zero Crossing Rate
kHz
dB
dB
kHz
Waveform
"usher"
Time (seconds)
Zero Crossing Rate
Energy -- 125 Hz to 750 Hz
Waveform
"azure"
GOLDY:>u>summer85>frics>drill>w>SSH>azure-SSH.utt



3 – خلاصه و نتیجه گیری:

در این فصل با مثال هایی از چند اسپکتروگرام آشنا شدیم.

10 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

آشنایی با واج های زبان انگلیسی:

- واکه ها
- سایشی ها

2– واج های موجود در انگلیسی لهجه آمریکایی

بیش از 40 صدا در گفتار لهجه انگلیسی آمریکایی وجود دارد.

می توان آن ها را بوسیله نحوه تولید (manner of articulation) تقسیم بندی کرد:

- واکه (18 واج)
- سایشی (8 واج)
- انفجاری (6 واج)
- دماغی (3 واج)
- نیمه واکه ای (4 واج)
- انفجاری–سایشی (2 واج)
- دمشی (1 واج)

یک دسته بندی دیگر

- واکه ها،
- glide ها
- صامت ها

می باشد.

این دسته ها در درجه انسداد با هم تفاوت دارند.

صامت های پرصدا (sonorant) هیچ فشار هنگام انسداد ایجاد نمی شود.

صامت های دماغی با پایین آوردن velum باعث عبور جریان هوا از حفره دماغی می شوند.

صامت های پیوسته جریان هوا را در حفره دهانی مسدود نمی کنند.

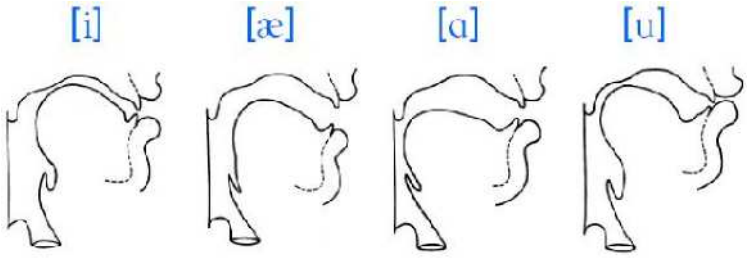
3– واکه ها

انسداد قابل توجهی در مسیر صوتی ایجاد نمی شود.

معمولاً با تحریک پریودیک تولید می شوند.

ویژگی های صوتی کاملاً بستگی به مکان قرارگیری فک، زبان و لب ها دارد.

در تصویر 1 نحوه قرارگیری زبانی برای چهار واج انگلیسی را مشاهده می کنید.



تصویر 1 – نحوه قرارگیری زبان و لب ها برای تلفظ واج های انگلیسی

حدوداً 18 واکه در انگلیسی لهجه آمریکایی وجود دارد.

این واج ها شامل واکه های معمولی، واکه های مرکب (diphthong) و واکه های کاهیده مانند schwa می باشد.

معمولاً بوسیله ویژگی های تولیدی آن ها را توصیف می کنند:

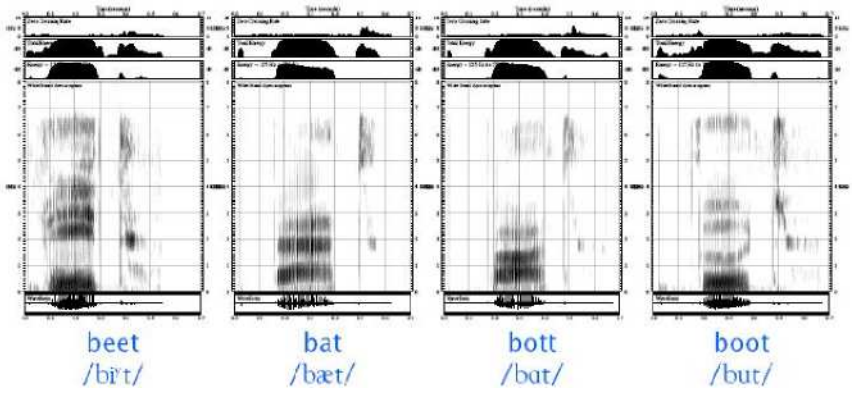
- High/Low (در مورد زبان)،
- Front/Back (در مورد زبان)،
- Retroflexed،
- Rounded (لب)
- Tense/Lax

در جدول تصویر 2 نمونه هایی از واکه های زبان انگلیسی را مشاهده می کنید.

/iʸ/	iy	beat	/ɔ/	ao	bought	/ɑʸ/	ay	bite
/ɪ/	ih	bit	/ʌ/	ah	but	/ɔʸ/	oy	Boyd
/eʸ/	ey	bait	/oʷ/	ow	boat	/ɑʷ/	aw	bout
/ɛ/	eh	bet	/ʊ/	uh	book	[ə]	ax	about
/æ/	ae	bat	/u/	uw	boot	[ɨ]	ix	roses
/ɑ/	aa	Bob	/ɝ/	er	Bert	[ɚ]	axr	butter

تصویر 2 – نمونه هایی از واکه های انگلیسی

در تصویر 3 اسپکتروگرام برخی از واکه های اصلی را مشاهده می کنید.



تصویر 3 – اسپکتروگرام واکه های اصلی

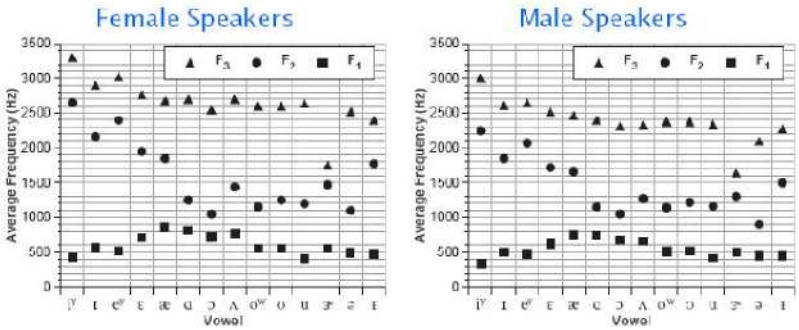
همان طور که در فصول قبل یاد گرفتیم، به خطوط تیره افقی فرکانس های فرمنت می گویند.

واکه ها را می توان فقط با دانستن این فرکانس ها با دقت بالایی تمییز داد.

معمولاً فقط سه فرمنت اول برای توصیف واکه ها کفایت می کند:

- F_1 فرکانس اول – با ویژگی High/Low همبستگی بالایی دارد.
- F_2 فرکانس دوم – با ویژگی Front/Back همبستگی بالایی دارد.
- F_3 با ویژگی Retroflexion همبستگی بالایی دارد.

در تصویر 4 سه فرمنت اول را برای واکه های مختلف خانم ها و آقایان مشاهده می کنید.

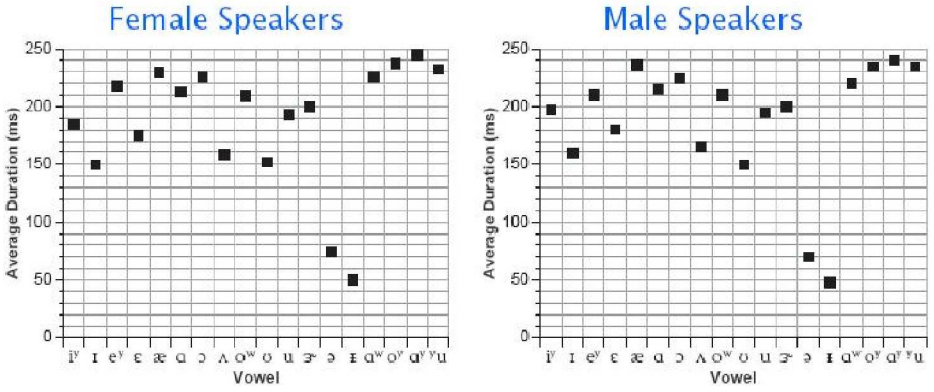


تصویر 4 – سه فرمنت اول برای واکه های مختلف خانم ها و آقایان

طول واکه ها

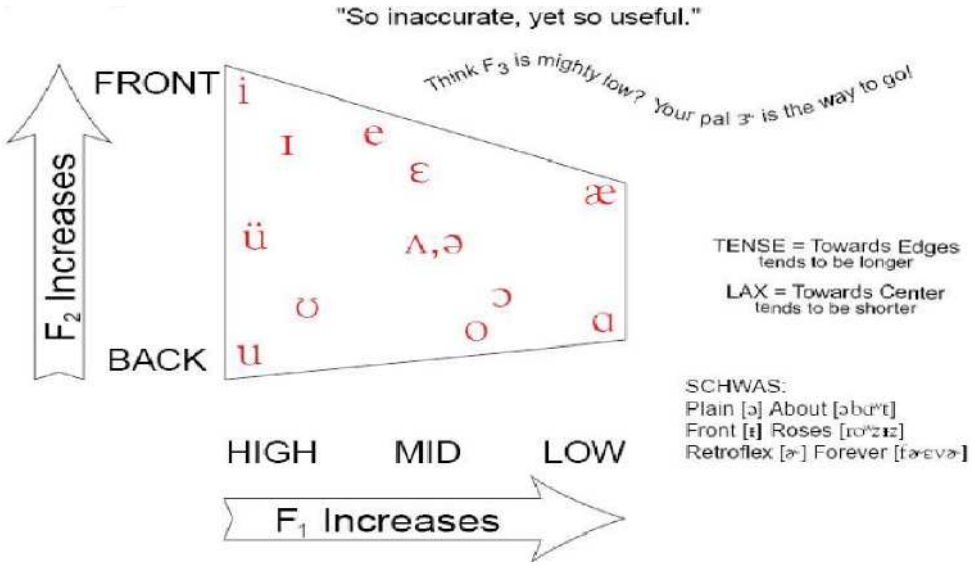
- هر واکه طول مدت ذاتی دارد.
- Schwa ها به صورت بسیار واضحی طول کوچکتری دارند (حدود 50 میلی ثانیه).
- واکه های /I, ɛ, Λ, Ω/ کوچکترین واکه ها هستند.
- متن واج ها (واج قبلی و بعدی) می تواند تاثیر بسیار زیادی بر روی مدت واکه داشته باشد.

در تصویر 5 متوسط مدت واکه های مختلف را در زبان انگلیسی مشاهده می کنید.



تصویر 5 – متوسط مدت واکه های مختلف را در زبان انگلیسی برای گوینده های خانم و آقا

در تصویر 6 نمودار جالبی از واکه ها بر مبنای فرمنت اول و دوم مشاهده می کنید.



تصویر 6 – نمودار جالبی از واکه ها بر مبنای فرمنت اول و دوم

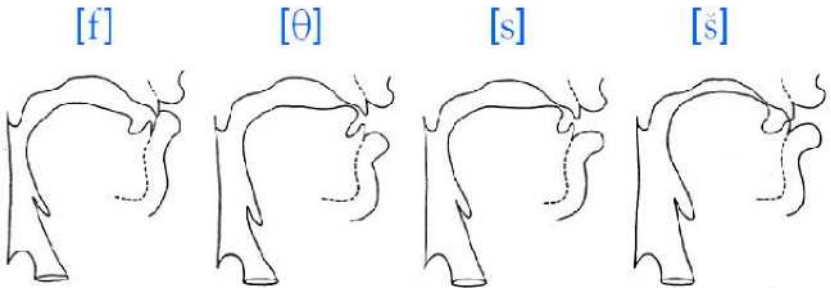
4– سایشی ها

این صدا از انسداد خیلی باریک تولید می شود.

مکان انسداد کاملاً ویژگی های صوتی را تعیین می کند.

می توان بوسیله تحریک پریودیک تولید شود.

در تصویر 7 چند نمونه از چهار واج سایشی و نحوه قرار گرفتن زبان، لب ها و دندان را مشاهده می کنید.



تصویر 7 – نمونه از چهار واج سایشی و نحوه قرار گرفتن زبان، لب ها و دندان

در کل 8 سایشی در زبان انگلیسی وجود دارد.

چهار مکان تولید (place of articulation) وجود دارد:

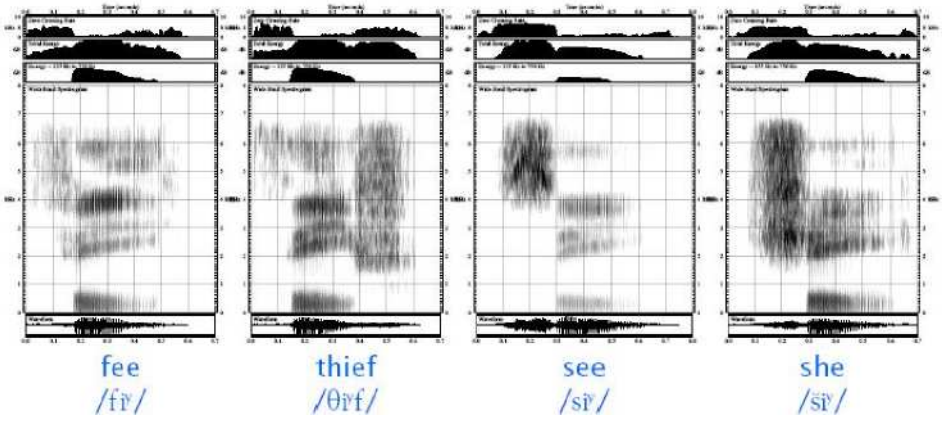
- لب–دندان (labio-dental) یا لبی
- دندانی
- Alvevolar
- Palate-alveolar

در تصویر 8 نمونه ای از واج های تولید شده توسط هر مکان تولید را مشاهده می کنید:

Type	Unvoiced			Voiced		
Labial	/f/	f	fee	/v/	v	v
Dental	/θ/	th	thief	/ð/	dh	thee
Alveolar	/s/	s	see	/z/	z	z
Palatal	/š/	sh	she	/ž/	zh	Gigi

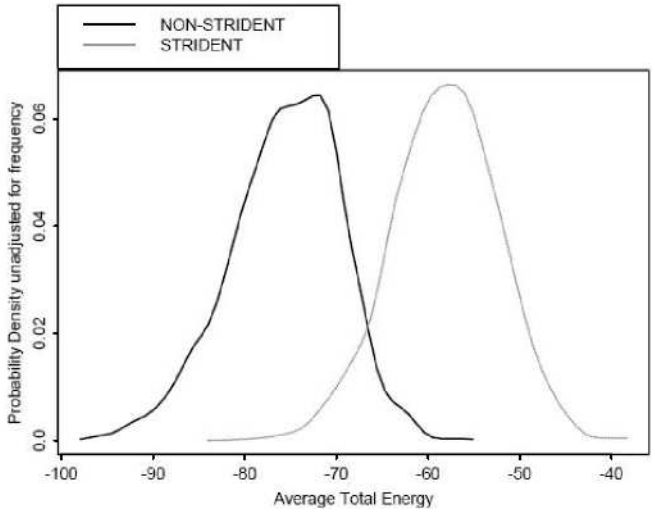
تصویر 8 – نمونه ای از واج های تولید شده توسط هر مکان تولید

در تصویر 9 اسپکتروگرام برخی از واج های سایشی بدون صدا را مشاهده می کنید.



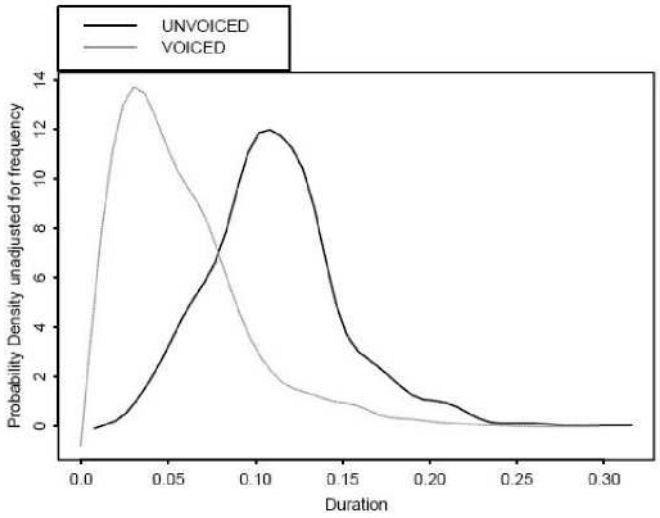
تصویر 9 – اسپکتروگرام برخی از واج های سایشی بدون صدا

در تصویر 10 مشاهده می کنید که سایشی های strident پر انرژی تر از بقیه سایشی ها هستند.



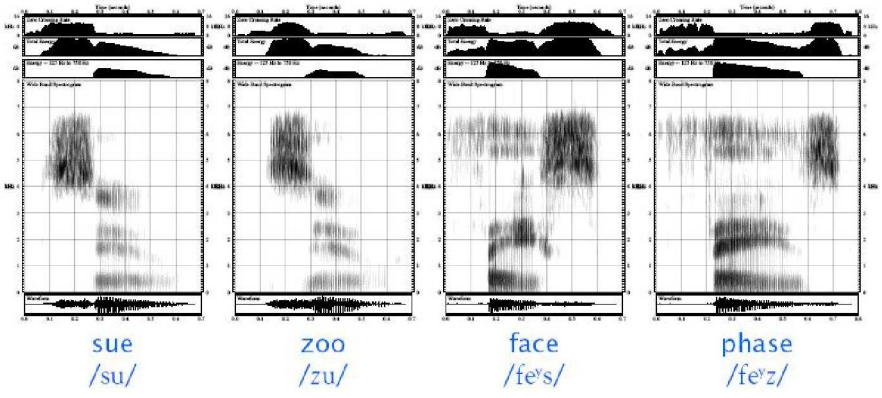
تصویر 10 – سایشی های strident پرانرژی تر از سایشی های غیر strident

در تصویر 11 مشاهده می کنید که معمولاً سایشی های صدادار کوتاه تر از سایشی های بدون صدا هستند.



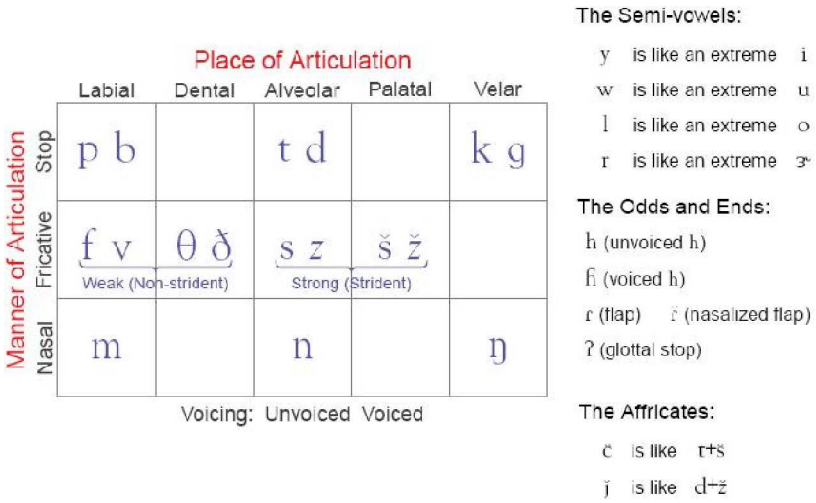
تصویر 11 – سایشی های صدادار کوتاه تر از سایشی های بدون صدا هستند

در تصویر 12 اسپکتروگرام چند سایشی صدادار را مشاهده می کنید.



تصویر 12 – اسپکتروگرام چند سایشی صدادار

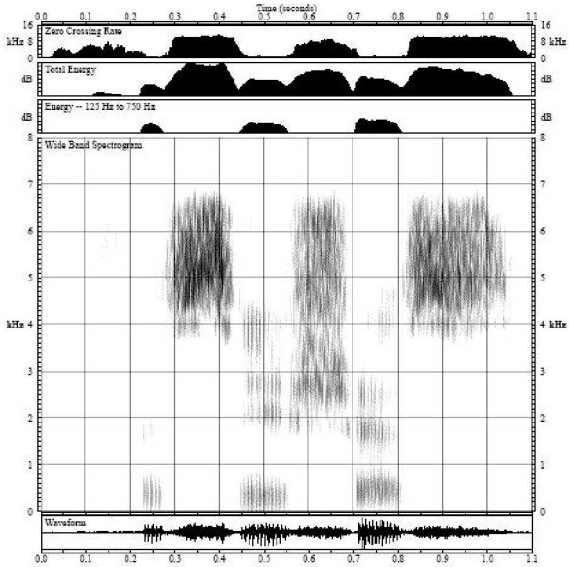
در تصویر 13 نمودار سایشی ها را مشاهده می کنید.



تصویر 13 – نمودار سایشی ها

5– خودآزمایی

خودآزمایی: کلمه زیر را تشخیص دهید.



6 – خلاصه و نتیجه گیری:

در این فصل با

- واکه ها
- سایشی های

زبان انگلیسی آشنا شدیم.

10 – منابع درس:

آواشناسی زبان فارسی، دکتر یدالله ثمره، 1371

آواشناسی و دستور زبان کُردی، لهجه سقزی، مصطفی کاوه، ۱۳۸٦

آواشناسی و دستور زبان کُردی، دکتر علی رخزادی، 1379

1– مقدمه

اهداف درس:

- آشنایی با واج های انفجاری
- آشنایی با واج های تودماغی

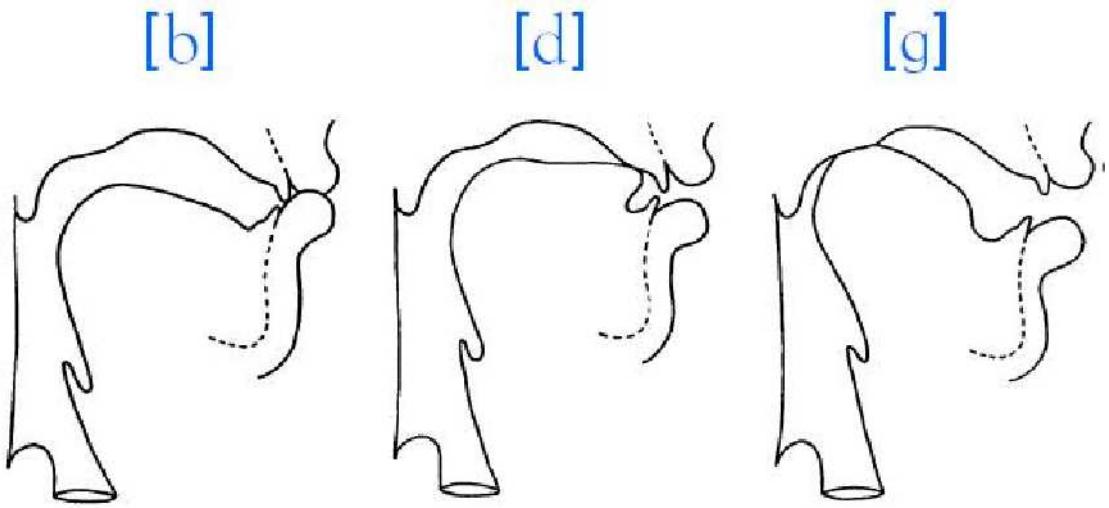
2– واج های انفجاری

برای ایجاد واج های انفجاری یک انسداد کامل در مسیر صوتی ایجاد می شود.

این انسداد به صورت ناگهانی باز می شود که منجر به یک صدای turbulent می شود.

این واج ها می توانند در حین بسته بودن تحریک پریودیک داشته باشند.

در شکل 1 نحوه تولید ب، گ و د را مشاهده می کنید.



شکل 1 – نحوه تولید ب، گ، د

در کل 6 صامت انفجاری در زبان انگلیسی وجود دارد.

سه مکان گفتار وجود دارد:

- لبی،
- alveolar و
- velar

هر مکان تولید دارای یک انفجاری صدادار و یک بدون صدا می باشد.

انفجاری های بدون صدا معمولاً خاصیت دمشی دارند.

انفجاری های صدادار معمولاً در اسپکتروگرام باعث یک نوار صدادار بودن می شوند (به جلسات اسپکتروگرام مراجعه کنید).

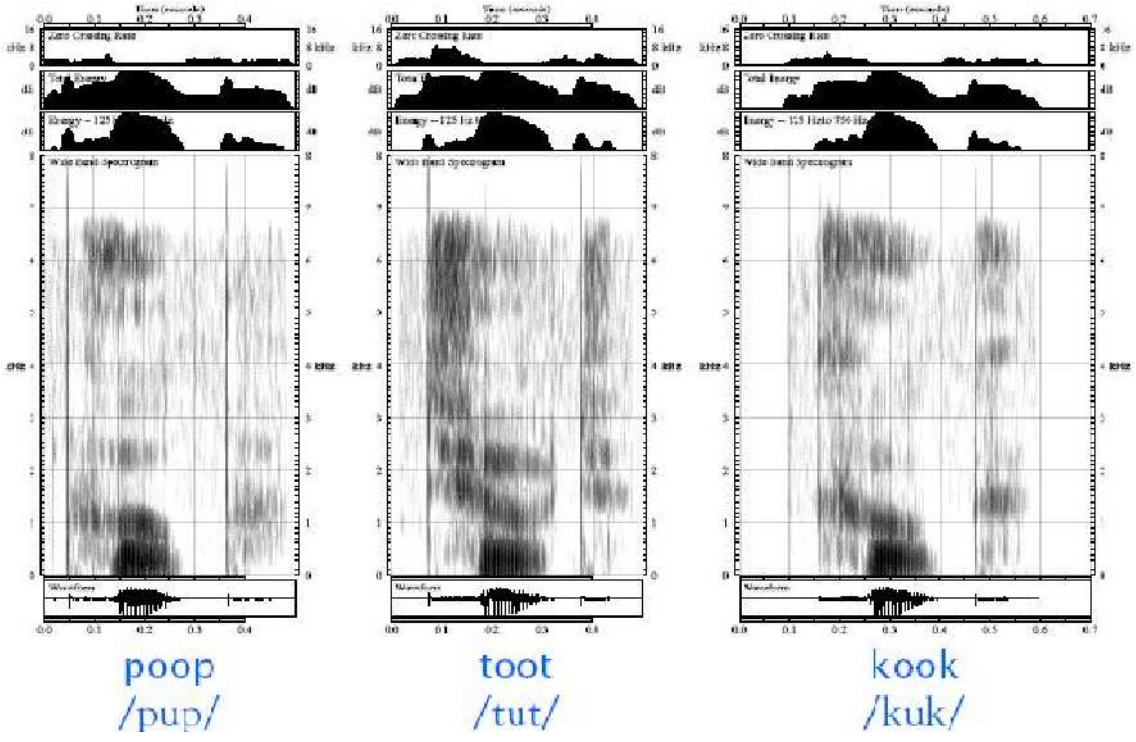
اطلاعات مربوط به مکان و گذر فرمنت ها در جوار این نوع واج ها معمولاً برای دسته بندی آن ها خیلی موثر است.

در تصویر 2 واج ها را با دسته بندی مکان تولید و صدادار بودن/نبودن مشاهده می کنید

Type	Voiced			Unvoiced		
Labial	/b/	b	bought	/p/	p	pot
Alveolar	/d/	d	dot	/t/	t	tot
Velar	/g/	g	got	/k/	k	cot

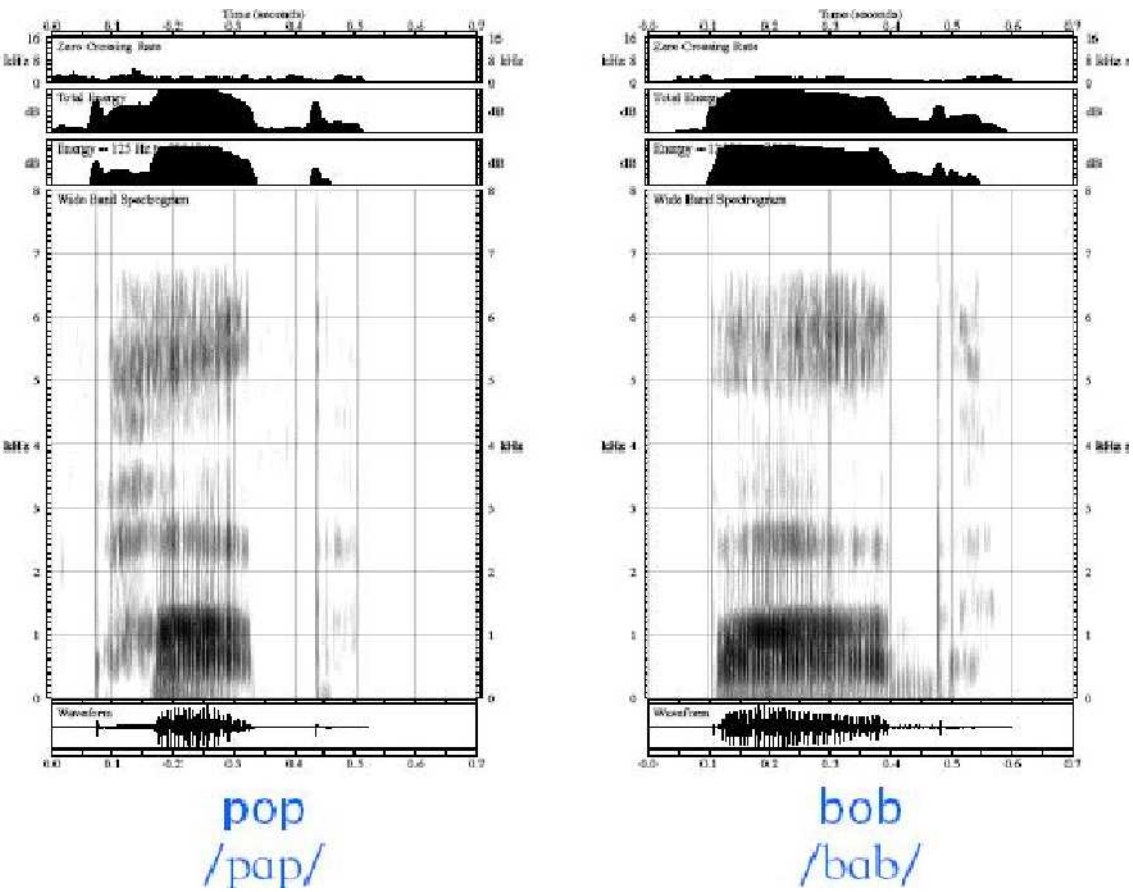
تصویر 2 – دسته بندی واج ها با توجه به مکان تولید و صدادار بودن/نبودن

در تصویر 3 اسپکتروگرام سه واج انفجاری را مشاهده می کنید.



تصویر 3 – سه اسپکتروگرام سه واج انفجاری

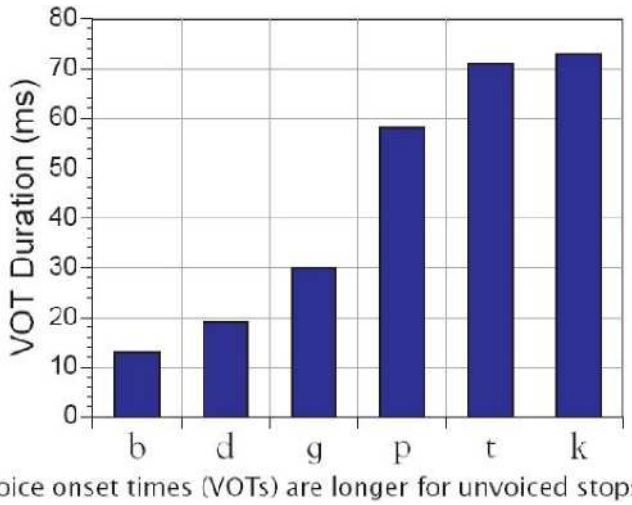
تاثیر این واج ها را بر روی نحوه گذر از واکه به صامت انفجاری مشاهده می کنید.



تصویر 4 – اسپکتروگرام دو واج ب و پ که در صدادار بدون تفاوت دارند

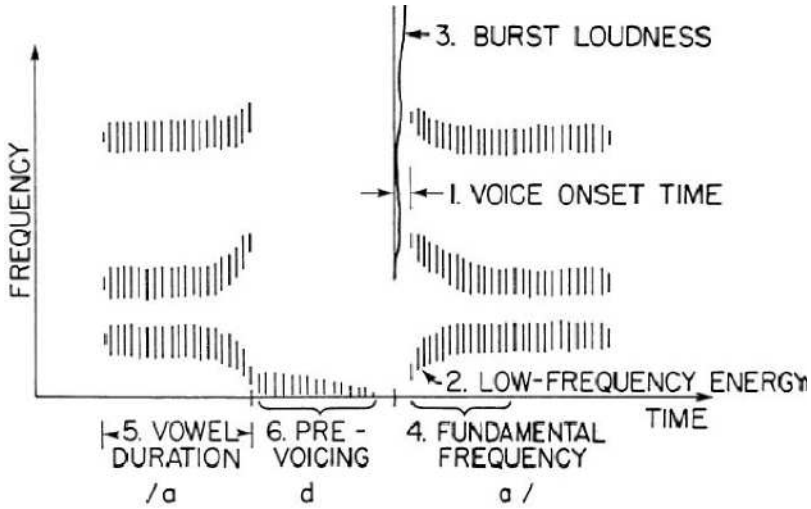
در تصویر 4اسپکتروگرام دو واج ب و پ را مشاهده می کنید که فقط در صدادار بدون تفاوت دارند.

در تصویر 5 متوسط مدت زمان مکث بعد از انسداد را برای انفجاری های مختلف مشاهده می کنید.



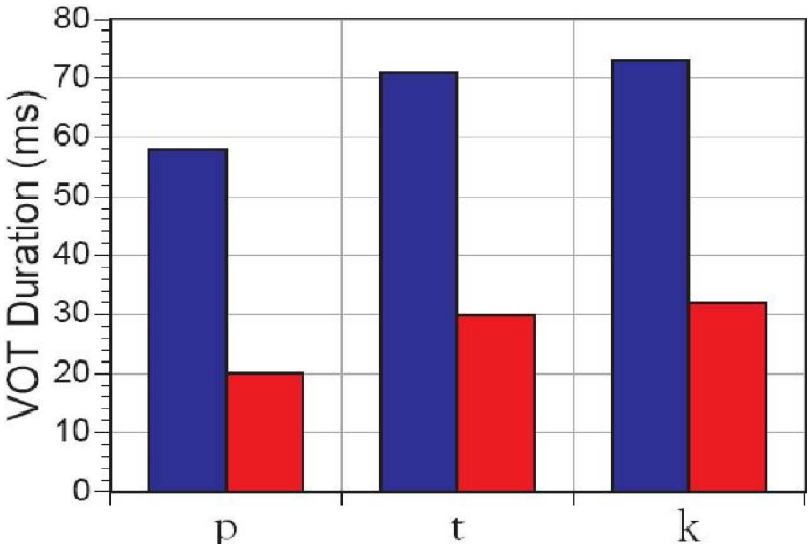
تصویر 5– متوسط مدت زمان مکث بعد از انسداد برای انفجاری های مختلف

نشانه های زیادی برای صدادار بدون یک انفجاری وجود دارد (تصویر 6)



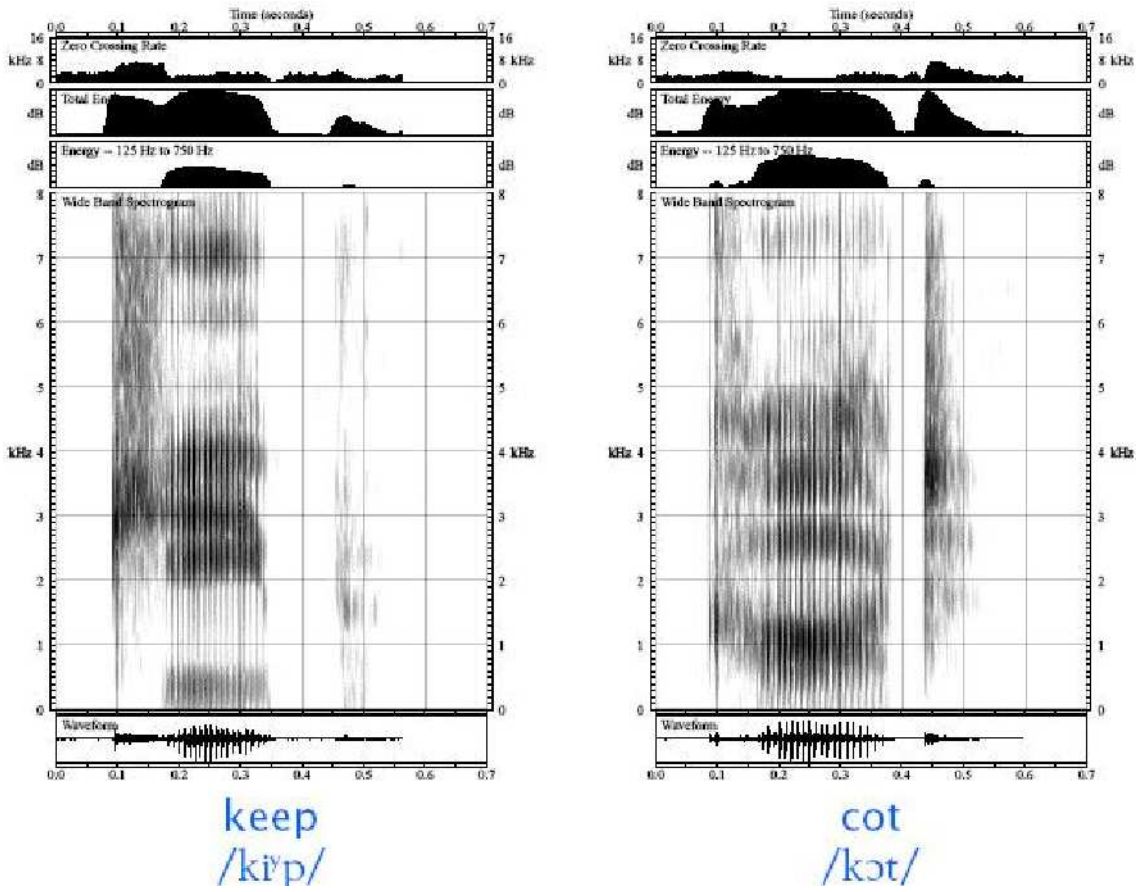
تصویر 6 – نشانه های صدادار بدون یک انفجاری روی یک اسپکتروگرام

در تصویر هفت مدت زمان را مشاهده می کنید.



تصویر 7 – مدت زمان انفجاری ها

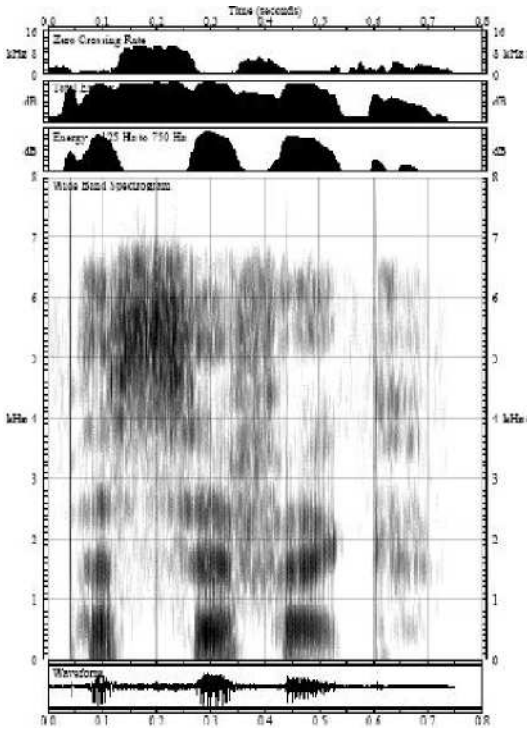
در تصویر 8 نمونه های از مکان توید velar عقب و velar جلو را مشاهده می کنید.



تصویر 8 – مکان تولید velar جلویی و velar عقبی

خود آزمایی 1

کلمه تلفظ شده در تصویر 9 را حدس بزنید.



تصویر 9 – خودآزمایی 1

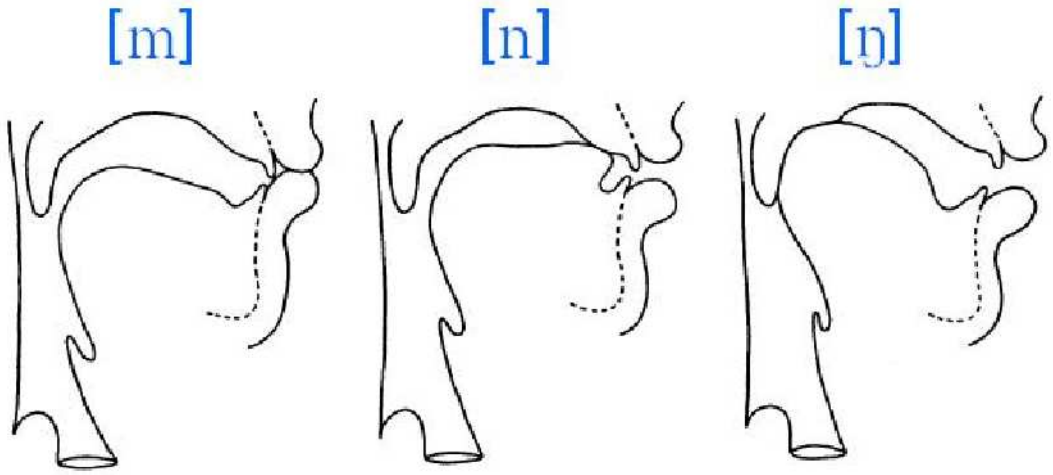
2– **واج های** nasal

با پایین آوردن velum، جریان هوا از درون حفره بینی می گذرد.

همان طور که گفتیم صامت ها با ایجاد یک انسداد در حفره دهانی تولید می شوند.

Nasal ها شکل طیفی خیلی شبیه هم دارند.

در تصویر 10 نحوه تولید nasal ها را مشاهده می کنید.



تصویر 10 – نحوه تولید nasal ها

سه مکان تولید:

- لبی
- Alveolar
- Velar

صامت های nasal همیشه به یک واکه متصلند

واج /ng/ همیشه در انگلیسی به صورت post-vocalic تلفظ می شود.

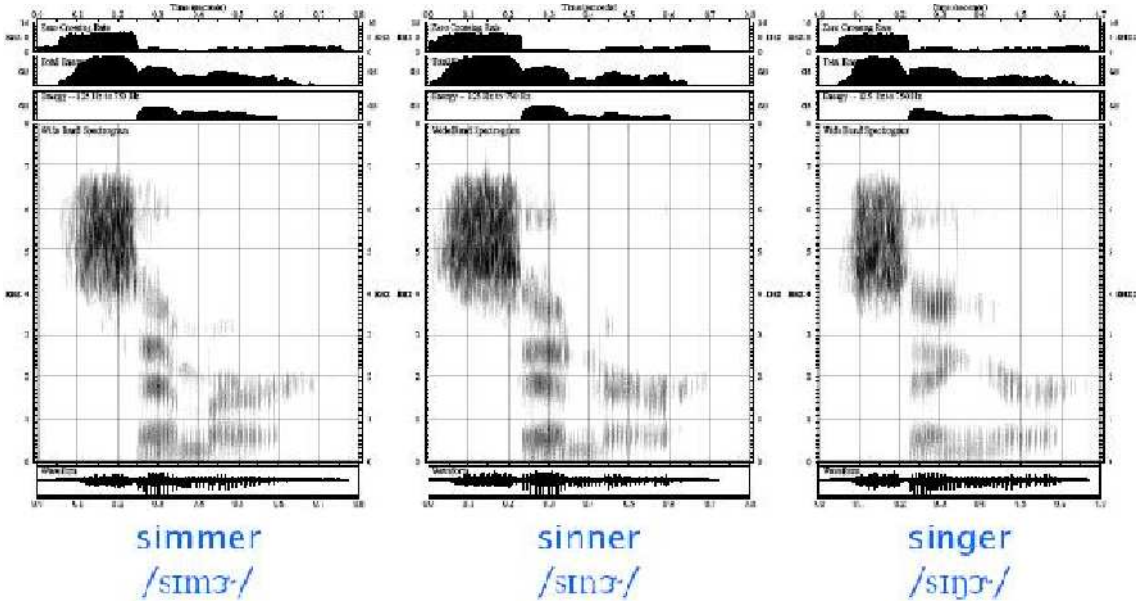
مکان تولید بوسیله گذرهای فرمنت های همسایه تعیین می شود.

در تصویر 11 مکان تولید nasalها را مشاهده می کنید.

Type	Nasal		
Labial	/m/	m	me
Alveolar	/n/	n	knee
Velar	/ŋ/	ng	sing

تصویر 11 – مکان تولید nasalها

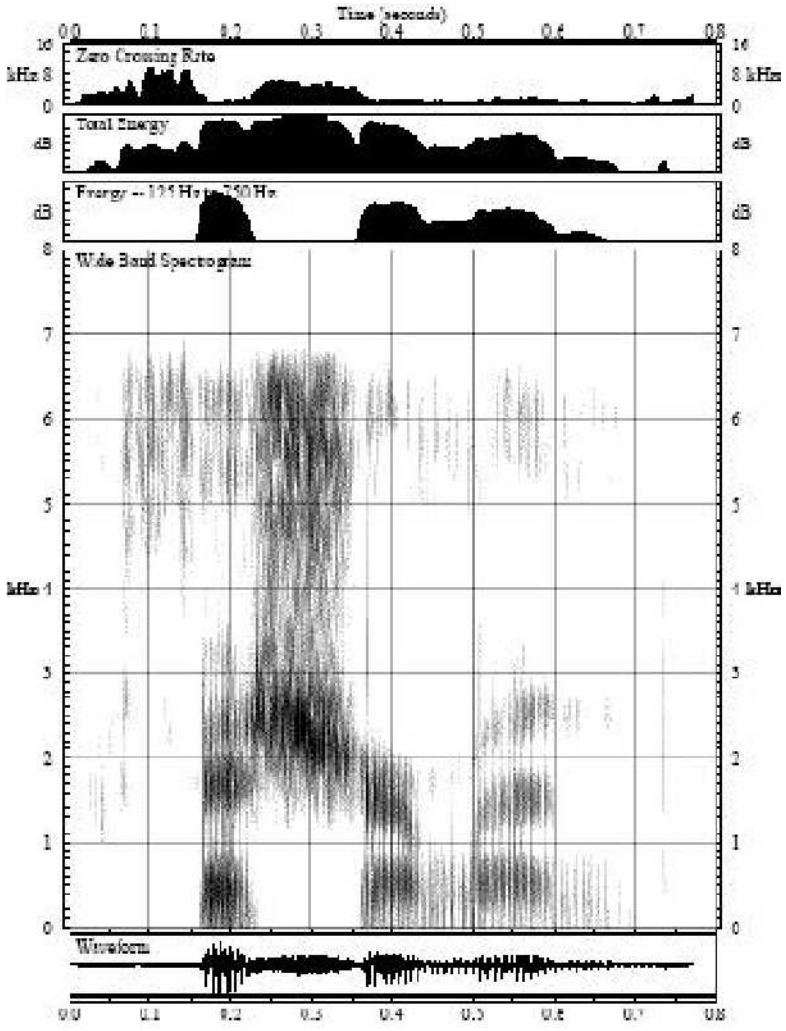
اسپکتروگرام nasal ها را در تصویر 12 مشاهده می کنید.



تصویر 12 – اسپکتروگرام nasal ها

خودآزمایی 2

کلمه تلفظ شده در تصویر 12 چیست؟



تصویر 12 – خودآزمایی 2

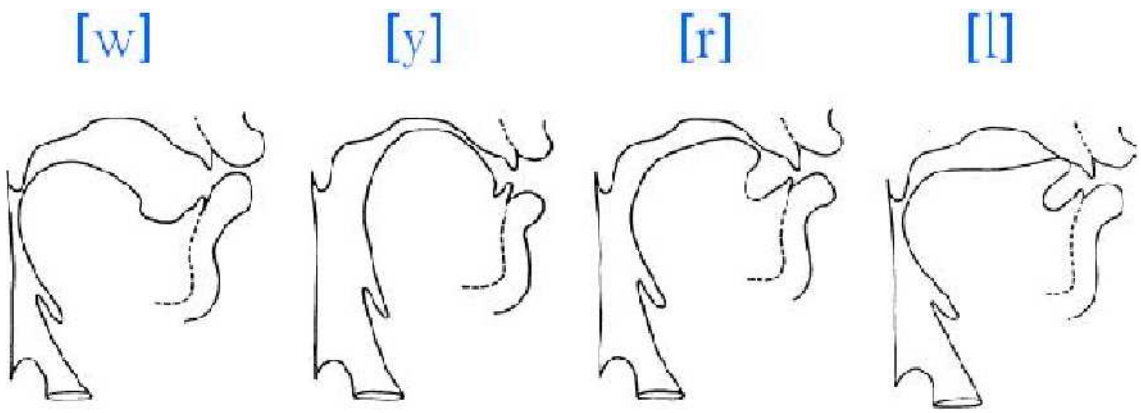
– واج های نیمه واکه

انسداد در مسیر صوتی ایجاد می شود ولی هیچ turbulence رخ نمی دهد.

نسبت به صامت های دیگر حرکات مفصلی کمتری دارند.

برای واج l و r، بوسیله جلوی زبان انسداد کاملی رخ می دهد و جریان هوا از کناره های انسداد رد می شود.

در تصویر 1 نحوه قرار گرفتن اجزای دهان را در هنگام ادای آن ها مشاهده می کنید.



تصویر 1 – نحوه قرارگرفتن اجزای دهان در هنگام تلفظ نیمه واکه ها

در زبان انگلیسی 4 نیمه واکه وجود دارد.

برخی موارد به آن ها Liquids یا Glides گفته می شود.

Glide ها تولید با شدت بیشتری از واکه های هم ارزشان هستند.

- فرمنت ها شبیه هستند ولی شدت بیشتری دارند.
- معمولاً به دلیل انسداد باریک تر ضعیف ترند.

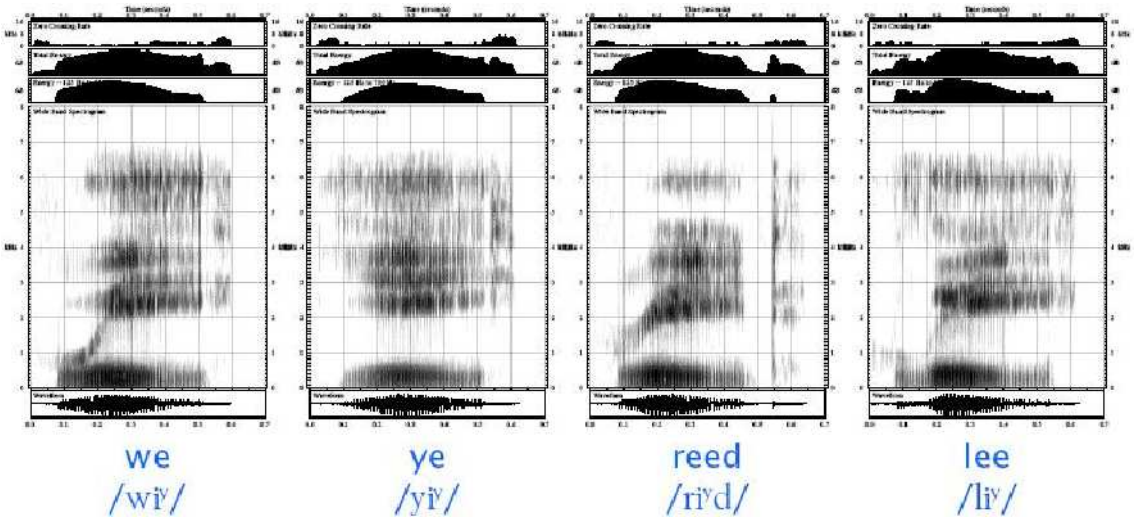
نیمه واکه ها همیشه به یک واکه متصلند.

در تصویر 2 نیمه واکه ها همراه به واکه های هم ارز (نزدیک) به آن ها را مشاهده می کنید.

Type	Semivowel			Nearest Vowel
Glides	/w/	w	wet	/u/
	/y/	y	yet	/i/
Liquids	/ɹ/	r	red	/ɝ/
	/l/	l	let	/o/

تصویر 2 – نیمه واکه ها همراه واکه های نزدیک به آن ها

در تصویر 3 اسپکتروگرام نیمه واکه ها را مشاهده می کنیم.



تصویر 3 – اسپکتروگرام نیمه واکه ها

برخی از ویژگی های صوتی نیمه واکه ها عبارتند از

- W و l خیلی با هم اشتباه می شوند.
- w دارای فرمنت اول و دوم خیلی پایین است.

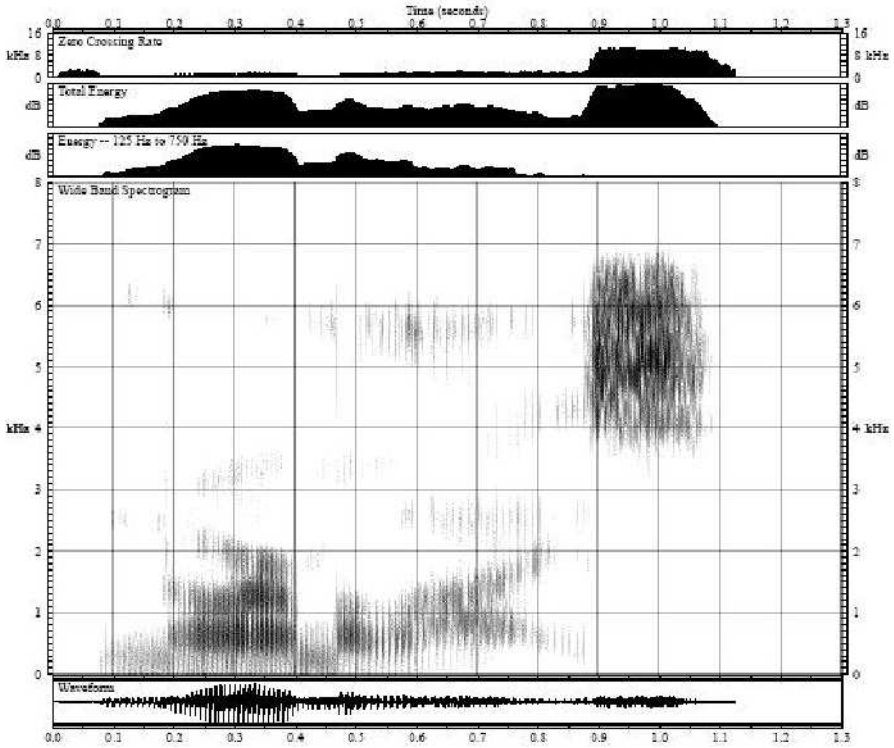
  در فرکانس های بالای فرمنت دوم افت شیب زیاد است.
- l بوسیله فرمنت اول و دوم پایین مشخص می شود.

  معمولاً انرژی در فرکانس های بالا وجود دارد.

  گذر فرمنت ها خیلی پیوسته می باشد.
- Y بوسیله فرمنت اول خیلی پایین و فرمنت دوم خیلی بالا مشخص می شود.
- R بوسیله فرمنت سوم خیلی پایین مشخص می شود.
-

خودآزمایی 1

کلمه تلفظ شده در تصویر 4 چیست؟



تصویر 4 – خودآزمایی 1

4 – خلاصه و نتیجه گیری:

در این فصل با بحث واج های انفجاری و nasal آشنا شدیم.

5 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

- آشنایی با واج های نیمه واکه
- آشنایی با واج های انفجاری–سایشی و دمشی
- آشنایی با سیلاب ها

3– واج های انفجاری–سایشی

در کل دو واج انفجاری–سایشی در زبان انگلیسی وجود دارد. ج و چ (تصویر 5)

Voiced			Unvoiced		
/ǰ/	jh	judge	/č/	ch	church

تصویر 5 – دو واج انفجاری–سایشی زبان انگلیسی

این دو واج، شرایط زیر را دارند:

- انفجاری alveolar به علاوه
- سایشی palatal

یعنی هم باز شدن ناگهانی انسداد دارند و هم نویز turbulence دارند.

می تواند در هنگام بسته شدن، تحریک پریودیک داشته باشد.

4– واج دمشی

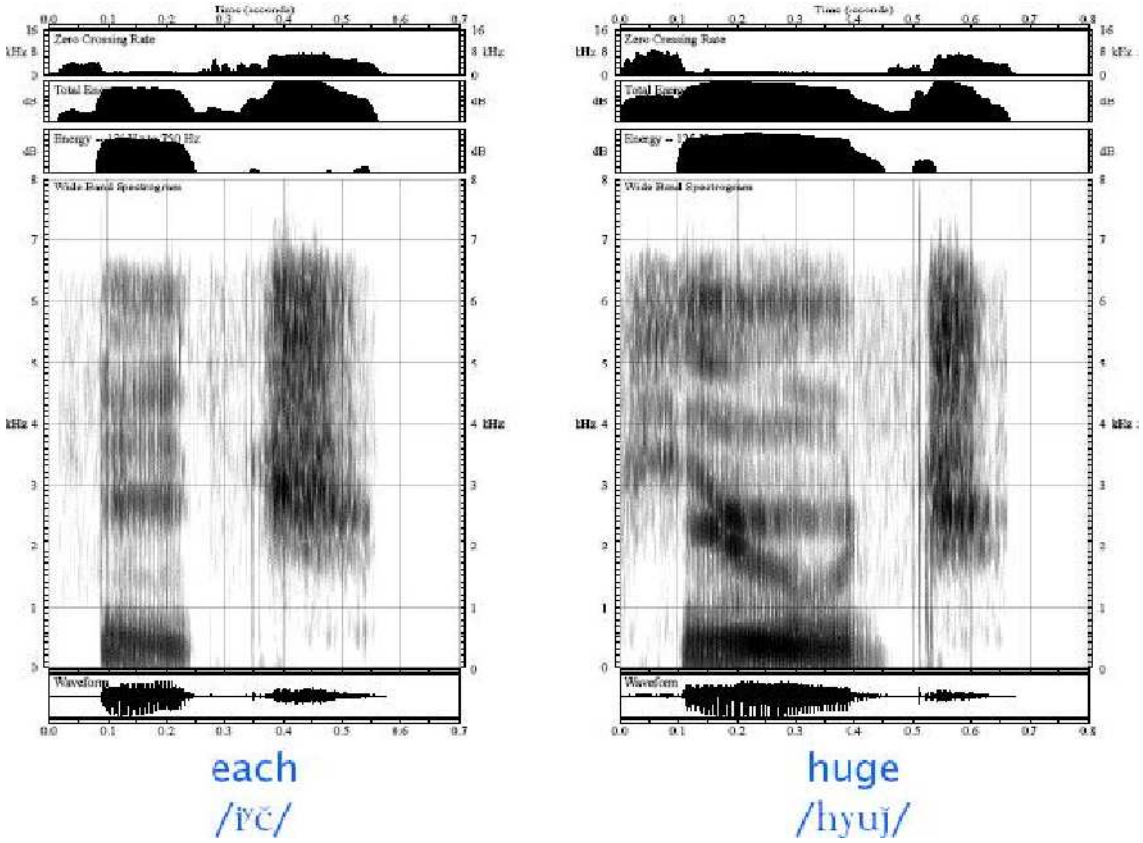
تنها واج دمشی در زبان انگلیسی: /h/ مانند hat است.

بوسیله تحریک turbulence در حنجره تولید می شود.

هیچ انسدادی در مسیر صوتی وجود ندارد. تولید فرمنت آن معمولی است.

معمولاً در مکان فرمنت اول انرژی دارد.

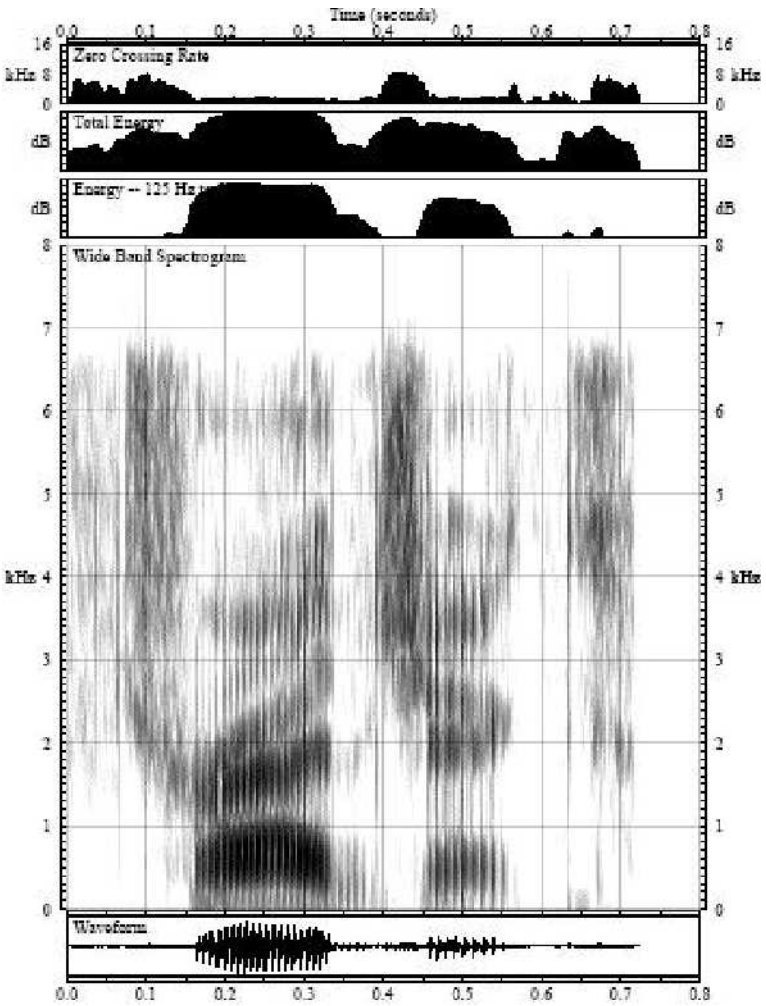
در تصویر 6 نمونه ای مشاهده می کنید.



نصویر 6 – اسپکتروگرام واج دمشی

خودآزمایی 1

کلمه تصویر 7 چیست؟



تصویر 7 – خودآزمایی 1

5– محدودیت های واجی

Phonotactics مطالعه دنباله های موجود صدا ها می باشد.

- بررسی 73 خوشه شروع مجزا وجود دارد.
- 208 خوشه پایان وجود دارد.

می توان از این محدودیت ها برای حذف دنباله های غیرممکن استفاده کرد.

- /tk/ نمی تواند یک کلمه را پایان دهد.
- /kt/ نمی تواند یک کلمه را شروع کند.

در تصویر 6 صامت های شروع کننده کننده را مشاهده می کنید.

-	of	hy	human	sf	sphere	tr	true
b	be	ǰ	just	sk	school	ts	tsunami
bl	black	k	can	skl	sclerosis	tw	twenty
br	bring	kl	class	skr	screen	ty	tuesday
by	beauty	kr	cross	skw	square	θ	thief
č	child	kw	quite	sky	skewer	θr	through
d	do	ky	curious	sl	slow	θw	thwart
dr	drive	l	like	sm	small	ð	the
dw	dwell	m	more	sn	snake	v	very
f	for	mw	moire	sp	special	vw	voyager
fl	floor	my	music	spl	split	vy	view
fr	from	n	not	spr	spring	w	was
fy	few	p	people	spy	spurious	y	you
g	good	pl	place	st	state	z	zero
gl	glass	pr	price	str	street	zl	zloty
gr	great	pw	pueblo	sw	sweet	zw	zweiback
gw	guava	py	pure	š	she	ž	genre
h	he	r	right	šr	shrewd		
hw	which	s	so	t	to		

تصویر 6 – صامت های شروع کننده (برگرفته از دیکشنری MWP)

6– سیلاب

ساختار سیلاب عمومیت های خیلی زیادی به خود می گیرد.

- معمولاً ادراک واج وابسته به سیلاب بندی است.
- تعداد زیادی قانون صوت شناسی وابسته به ساختار سیلاب ها می باشد.

ساختار سیلاب بر این اساس پیش بینی می شود که صداها گفتار را بر حسب مقدار sonority آن ها رده بندی می کنند (تصویر 7).

Sounds	Sonority Values	Examples
Low Vowels	10	/ɑ, ɔ/
Mid Vowels	9	/e, o/
High Vowels	8	/i, u/
Flaps	7	/ɾ/
Laterals	6	/l/
Nasals	5	/m, n, ŋ/
Voiced Fricatives	4	/v, ð, z/
Unvoiced Fricatives	3	/f, θ, s/
Voiced Stops	2	/b, d, g/
Unvoiced Stops	1	/p, t, k/

تصویر 7 – رده بندی صدا ها بر اساس مقدار sonority آن ها

کل تلفظ یک کلمه را می توان به سیلاب های هم ارز آن تجزیه کرد.

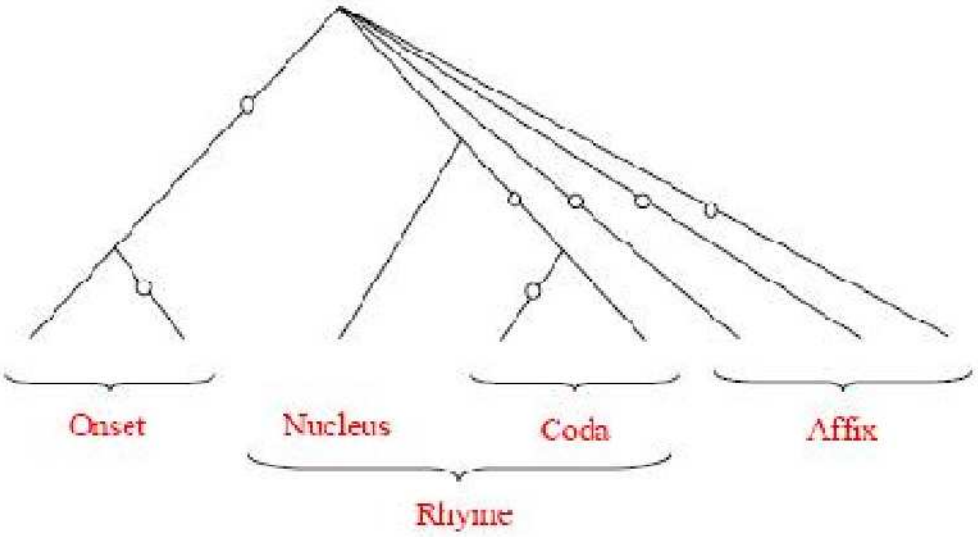
تعداد سیلاب ها برابر تعداد قله های sonority می باشد.

در هر سیلاب، یک قطعه شامل یک قله sonority می باشد که قبل یا پس از آن قطعات با مقدار کمتر sonority وجود دارد (تصویر 8).

suprasegmental													
s	u	p	ɹ	ʌ	s	ɛ	g	m	ɛ	n	t	ə	l
3	8	1	7	9	3	9	2	5	9	5	1	9	6
minimization													
	m	ɪ	n	ɪ	m	ɑʸ	z	e	š	ə	n		
	5	8	5	8	5	10	4	9	3	9	5		
fire													
					f	ɑʸ		ɚ					
					3	10	(8)	9					

تصویر 8 – چند کلمه همراه با مقادیر sonority

قالب یک سیلاب را در تصویر 9 مشاهده می کنید.



تصویر 9- قالب یک سیلاب

شاخه هایی که بوسیله دایره مشخص شده اند دلخواه هستند.

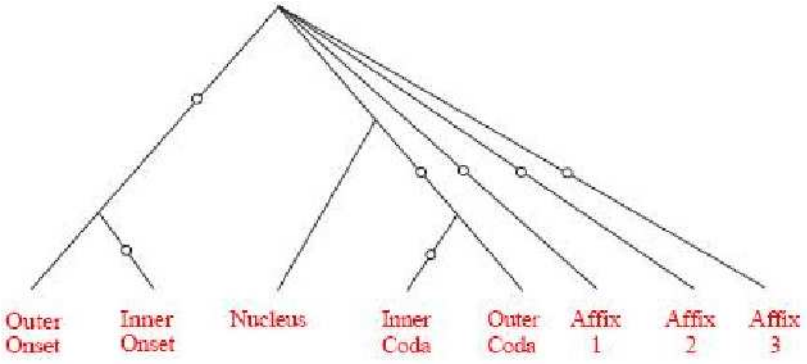
هسته باید حتماً شامل یم non-obstruent باشد.

Sonority با دورشدن از هسته، کم می شود.

آخرین سیلاب در کلمه می تواند affix داشته باشد.

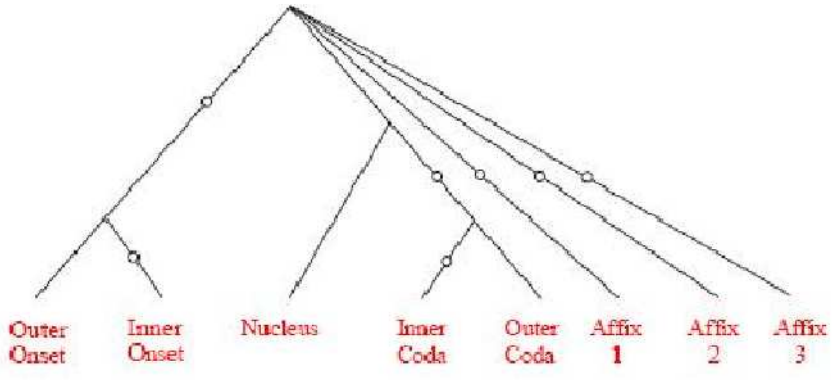
با /sp/، /st/ و /sk/ مانند obstruent های واحد رفتار می شود.

در تصویر 10 و 11 چند مثال را مشاهده می کنید.



crown	k	r	ɑ	w	n			
fledged	f	l	ɛ		j	d		
links	l		ɪ	ŋ	k	s		
dwarves	d	w	ɑ	r	v	z		
stick	st		ɪ		k			
sixths	s		ɪ		k	s	θ	s

تصویر 10 – چند مثال از سیلاب

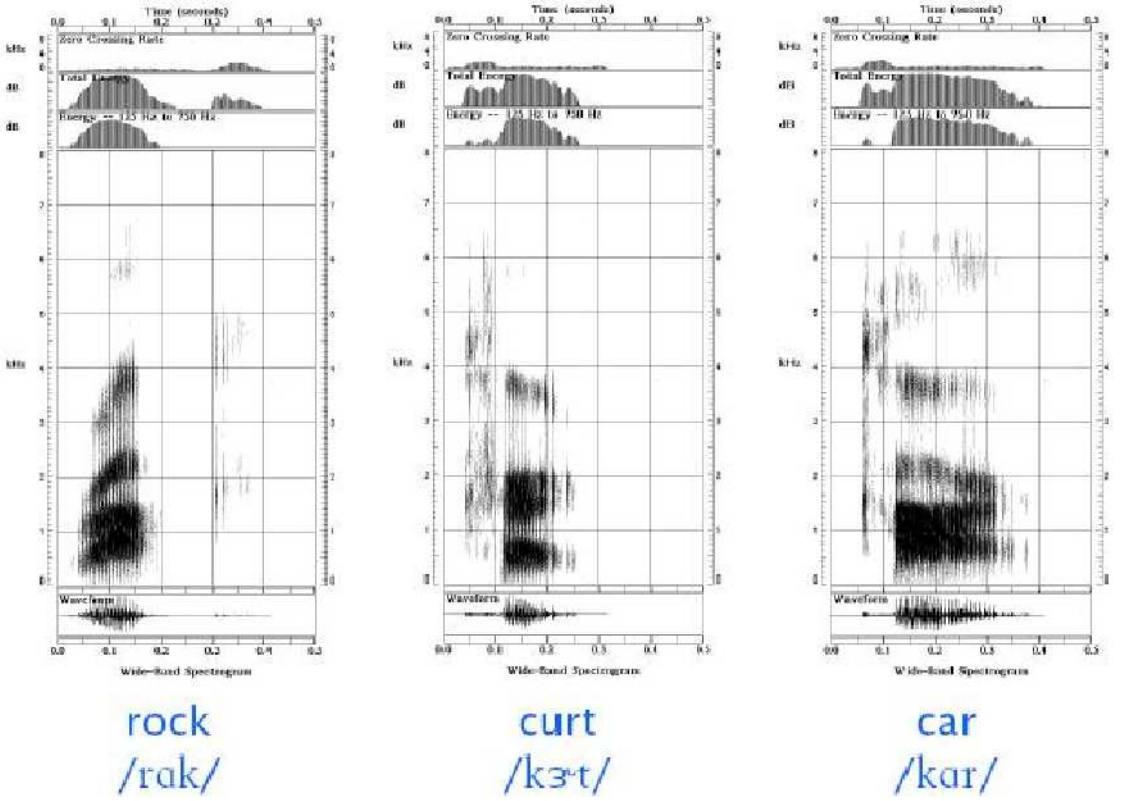


rock	r		ɑ		k	
crock	k	r	ɑ		k	
curt	k		ɜ˞		t	
cart	k		ɑ	r	t	
car	k		ɑ		r	
lick	l		ɪ		k	
bottle	b		ɑ,l̩		t	
kill	k		ɪ		l	

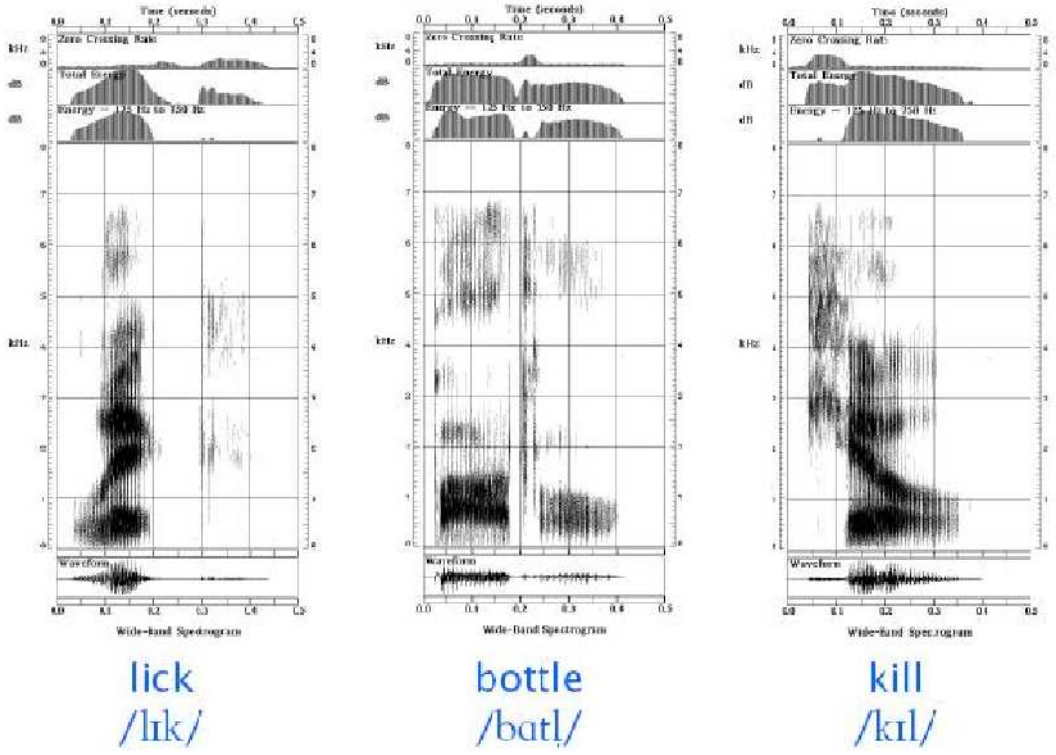
تصویر 11 – چند مثال از سیلاب ها

در تصویر 12 چند اسپکتروگرام از /r/ را مشاهده می کنید.

در تصویر 13 چند اسپکتروگرام از /l/ را مشاهده می کنید.

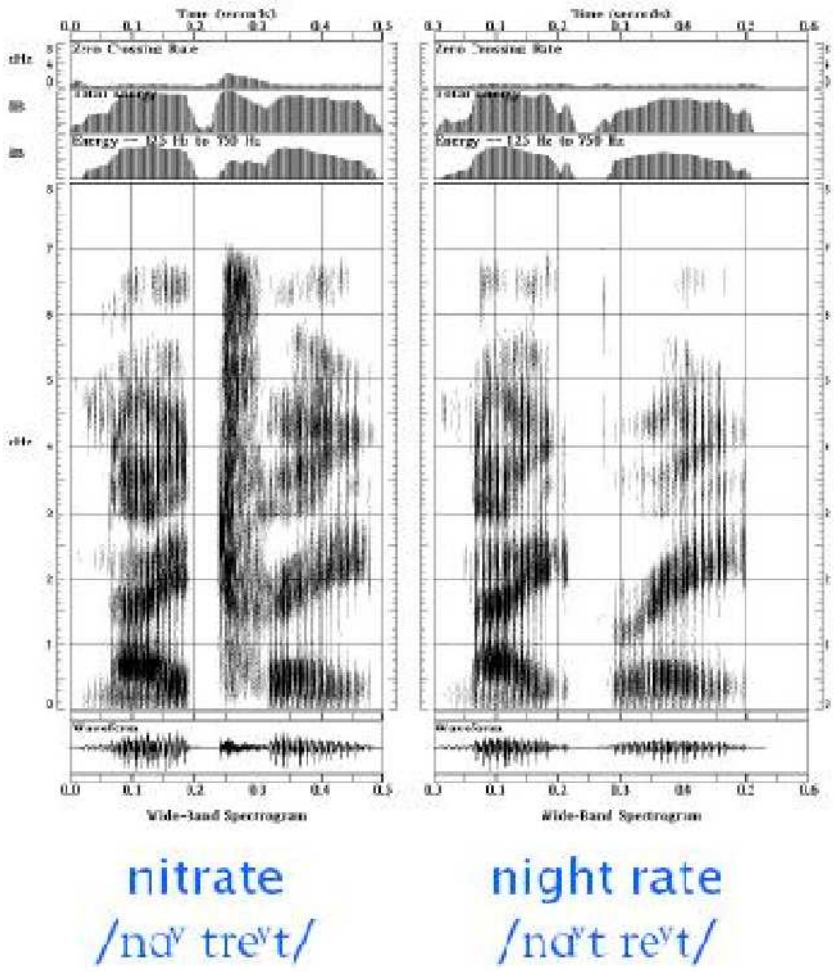


تصویر 12 – چند اسپکتروگرام از /r/



تصویر 13 – چند اسپکتروگرام از /l/

تغییرات allophonic در مرزهای سیلاب را در تصویر 14 مشاهده می کنید.



تصویر 14 – تغییرات allophonic در مرزهای سیلاب

7 – خلاصه و نتیجه گیری:

در این فصل بحث های زیر آشنا شدیم:

- با واج های نیمه واکه
- با واج های انفجاری-سایشی و دمشی
- با سیلاب ها

8 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

آشنایی با ضرایب پیشگویی خطی

آشنایی با نحوه محاسبه ضرایب پیشگویی خطی

آشنایی با نحوه محاسبه ضرایب کپسترال از ضرایب پیشگویی خطی

2– ضرایب پیشگویی خطی

هدف اصلی پیشگویی خطی «تخمین دنباله خروجی است از یک ترکیب خطی از نمونه ها ورودی و خروجی های گذشته».

$$\hat{y}(n) = \sum_{j=0}^{q} b(j)x(n-j) - \sum_{i=1}^{p} a(i)y(n-i)$$

به a(i) و b(j) ضرایب پیش بینی کننده گفته می شود.

اغلب سیستم هایی که برای ما جالب توجه اند را می توان بوسیله معادلات دیفرانسیلی خطی با ضرایب ثابت توصیف کرد.

$$\sum_{i=0}^{p} a(i)y(n-i) = \sum_{j=0}^{q} b(j)x(n-j)$$

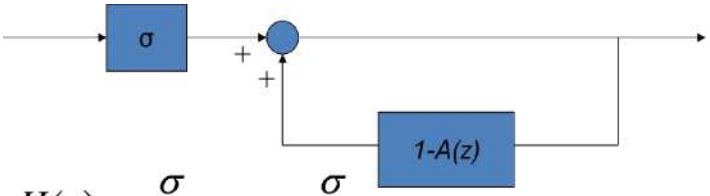
اگر Y(z)/X(z)=H(z) باشد که H(z) نسبت چندجمله ای های N(z)/D(z) باشد،

$$N(z) = \sum_{j=0}^{q} b(j)z^{-j} \text{ and } D(z) = \sum_{i=0}^{p} a(i)z^{-i}$$

می توان گفت که با داشتن ضرایب پیش بینی کننده (a و b) می توان صفرها و قطب های H(z) را در اختیار داشت.

دو نوع مهم از پیشبینی کننده ها موجود است:

- مدل های تمام قطب (AllPole): در آمار به این مدل ها مدلهای autoregressive (AR) گفته می شود.در این مدل ها N(z) یک عدد ثابت است. در تصویر 1 یک نمونه از مدل AR مشاهده می کنید.
- مدل های تمام صفر (All-zero): به این مدل ها، مدل های moving average (MA) گفته می شود. در این مدل ها مخرج D(z) عدد «یک» است.
- ترکیب دو مدل بالا را مدل autoregressive moving average (ARMA) می گویند.



$$H(z)=\frac{\sigma}{A(z)}=\frac{\sigma}{1-\sum_{i=1}^{p}a_i z^{-i}}$$

تصویر 1 – مدل پیش بینی کننده AR

با داشتن یک سیگنال y(n) با میانگین صفر در و با در نظر گرفتن مدل AR داریم:

$$\hat{y}(n)=-\sum_{i=1}^{p}a(i)y(n-i)$$

خطا به صورت زیر محاسبه می شود:

$$e(n)=y(n)-\hat{y}(n)$$
$$=\sum_{i=0}^{p}a(i)y(n-i)$$

برای به دست آوردن پیش بینی کننده، از *اصل تعامد* استفاده می کنیم. این اصل بیان می کند که ضرایب مورد نظر ضرایبی هستند که باعث می شوند خطا متعاند بر نمونه ها y(n-1)، y(n-2)... y(n-p) شود.

پس نتیجه می گیریم که:

$$< y(n-j)e(n) >= 0 \quad \text{for j} = 1, 2, ..., \text{p}$$

با به عبارت دیگر:

$$\left\langle y(n-j)\sum_{i=0}^{p}a(i)y(n-i)\right\rangle=0$$

با تعویض عمل متوسط گیری و جمع کردن ، به فرمول زیر می رسیم: $\sum_{i=0}^{p}a(i)\sum_{n}y(n-i)y(n-j)=0,\ \text{j}=1,...,\text{p}$

پیش بینی کننده های مورد نیاز بوسیله حل این معادلات به دست می آیند.

اصل تعامد همچنین بیان می کند که کمینه خطا بوسیله فرمول زیر به دست می آید: $E=\left\langle e^2(n)\right\rangle=\left\langle y(n)e(n)\right\rangle$

$$\sum_{i=0}^{p}a(i)\sum_{n}y(n-i)y(n)=E$$

می توان خطا را برای همه زمان ها کمینه کرد:

$$\sum_{i=0}^{p}a(i)r_{i-j}=0,\ j=1,2,...,p$$

$$\sum_{i=0}^{p}a(i)r_i=E \qquad r_i=\sum_{n=-\infty}^{\infty}y(n)y(n-i)$$

حال سئوال این است که چگونه مقدار مناسبی برای p انتخاب کنیم.

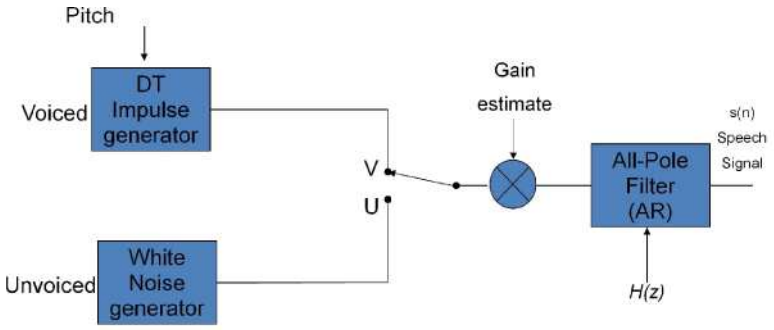
برای انتخاب تعداد ضریب مناسب (p) روش های سرانگشتی موجود است.

تعداد p بستگی مستقیمی به پهنای باند سیگنال گفتار دارد.

$$p = \frac{2BW}{1000} + c$$

- برای یک مسیر صوتی معمولی، به صور متوسط یک فرمنت در هر کیلوهرتز از پهنای باند وجود دارد.
- هر فرمنت به دو قطب complex conjucate نیاز دارد.
- در نتیجه برای هر فرمنت دو ضریب پیش بینی کننده نیاز است و یا دو ضریب برای هر کیلوهرتز از پهنای باند نیاز است.

در تصویر 2 نحوه مدل کردن گفتار به وسیله پیشگویی خطی را مشاهده می کنید.



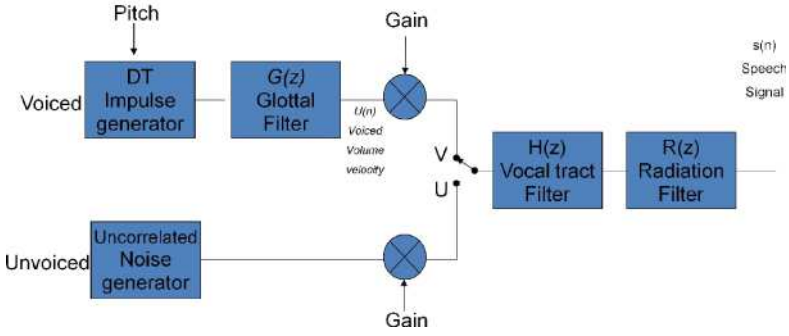
همان طور که در فص

- تحریک: ک
  از نویز به
- مسیر صوت

های بدون صدا هم

مدل ارائه شده در تص
کرد. چنین مدلی را ب

بیه تر به واقعیت



تصویر 3– مدل دقیق تر گفتار

بلوک های زیر اضافه شده اند:

- Gain: نشان دهنده بلندی و کمی صدا است و رابطه مستقیمی با انرژی دارد.
- Glottal Filter: تارهای صوتی انسان را مدل می کند.یعنی از ضربه ساده استفاده نمی کند بلکه به پالس تار صوتی انسان شکل طبیعی تر از ضربه می دهد.
- Radiation Filter: سعی می کند لب انسان را مدل کند. معمولاً این فیلتر به صورت مشتق گیر می باشد.

این بلوک ها به طبیعی تر شدن گفتار تولید شده کمک می کنند.

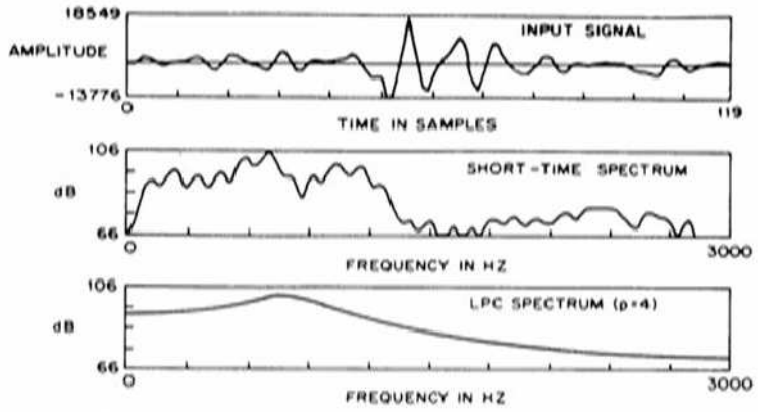
3– ضرایب پیشگویی خطی برای پردازش گفتار

مثالی از آنالیز LPC با 4 ضریب بر روی سیگنال گفتار را در تصویر 4 مشاهده می کنید.

همان طور از تصویر 4 وسط و پایین مشخص است، LPC پوش فرکانس را به دست می دهد.

به عبارتی روند کلی طیف سیگنال را مدل می کند ولی با تعداد پارامترهای خیلی کمتر (4 ضریب) در برابر نمونه 120 ورودی.

پس هم مدل سازی انجام داده ایم و هم فشرده سازی.



تصویر 4– نمونه ای از سیگنال در حوزه زمان (بالا) طیف فرکانسی (وسط) و پوش LPC (پایین)

در تصویر 5 سیگنال تصویر 4 با تعداد ضرایب مختلف مدل شده است.

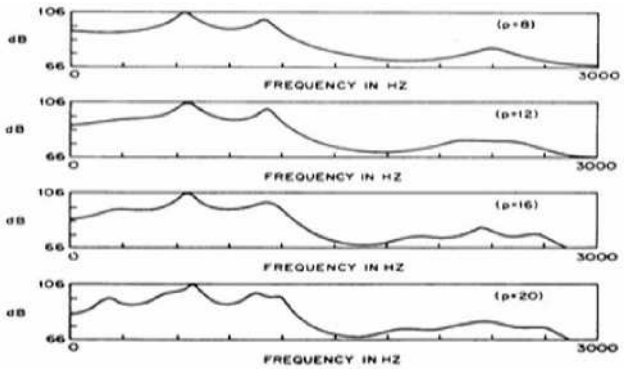
همان طور که مشخص است با تعداد ضرایب کم سیگنال خیلی صاف است و روند کلی را مدل می کند

ولی هرچه p افزایش می یابد جزئیات طیف سیگنال افزایش می یابد.

این طبیعی است زیرا هر چه p بیشتر می شود تعداد پارمترهای مدل بیشتر می شود و نتیجتاً مدل تصویر دقیق تری از واقعیت می شود.

البته افزایش تعداد p تا یک جایی مناسب است. زیرا بیش از یک حدی (حدود 20) تعداد پارامترها خیلی زیاد می شود و این برای پردازش های بعدی نامناسب است.

زیرا اطلاعات اضافی (بالاپایین شدن های ریز) هم وارد مدل می شود که طبیعتاً اطلاعات جدیدی به مدل اضافه نمی کنند و فقط پیچیدگی مدل را افزایش می دهند.



تصویر 5 – اثر افزایش تعداد ضریب پیش بینی کننده (p)

5 – خلاصه و نتیجه گیری

در انی فصل با مفهوم ضرایب پیشگویی خطی آشنا شدیم.

در این فصل با نحوه محاسبه ضرایب پیشگویی خطی آشنا شدیم.

در این فصل با نحوه محاسبه ضرایب کپسترال پیشگویی خطی آشنا شدیم.

6– منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

۱– مقدمه

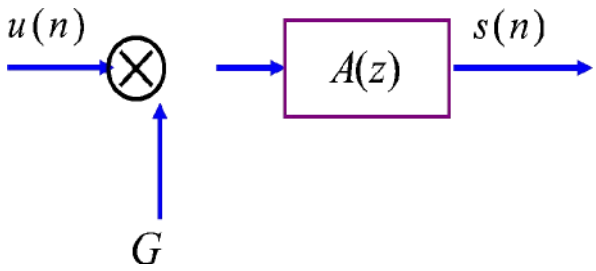
آشنایی با نحوه محاسبه ضرایب LPC

آشنایی با روش محاسبه ضرایب کپسترال LPC

۲– محاسبه ضرایب پیشگویی خطی

همان طور که در بخش قبل گفته شد از این ضرایب برای مدل کردن گفتار استفاده می شود.

به این صورت که مدل سیگنال تحریک (قطار ضربه) را به مدل مسیر صوتی (مدل AR) اعمال کرده و خروجی صوتی را می گیریم. این امر به صورت خیلی خلاصه در تصویر ۱ نشان داده شده است.



تصویر ۱– مدل سازی گفتار بوسیله مدل AR به صورت خیلی ساده

همان طور که گفته شد، مدل AR فرض می کند که سیگنال در زمان n از ترکیب خطی p نمونه قبلی سیگنال محاسبه می شود.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p),$$

با خلاصه کردن و اضافه کردن عبارت تحریک (Gu(n)):

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n),$$

با تبدیل z گرفتن از طرفین:

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z)$$

که در نهایت پاسخ سیستم به صورت روبرو به دست می آید:

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)}.$$

سیگنال اصلی برابر روبر است:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n).$$

سیگنال تخمین زده شده برابر روبرو است:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k).$$

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

در نتیجه خطا به صورت روبرو به دست می آید:

و A(z) هم به صورت روبرو می باشد:

معادلات آنالیز LPC

در صورتی که دو فرض روبرو را بکنیم:

$$S_n(m) = s(n+m)$$
$$e_n(m) = e(n+m)$$

هدف کمینه کردن سیگنال متوسط مربعات خطا می باشد:

$$E_n = \sum_m e_n^2(m)$$

یا به عبارتی:

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2.$$

برای کمینه شدن خطا En باید مشتق آن نسبت به پارامترها (ak) صفر باشد:

$$\frac{\partial E_n}{\partial a_k} = 0, \qquad k = 1,2,...,p$$

که به دست می آید:

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_m S_n(m-i)S_n(m-k)$$

در صورتی که از این نماد استفاده شود:

$$\phi_n(i,k) = \sum_m S_n(m-i)S_n(m-k)$$

خواهیم داشت:

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \, \phi_n(i,k) \qquad i = 1,2,...,p$$

که مجموعه ای از p معادله و p مجهول می باشد.

کمترین مقدار متوسط مربعات خطا به صورت روبرو به دست می آید:

$$\hat{E}_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m)s_n(m-k)$$

$$= \phi_n(0,0) - \sum_{k=1}^{p} \hat{a}_k \, \phi_n(0,k).$$

دو روش مهم برای حل این معادله ارائه شده است:

۱. **روش خودهمبستگی (Autocorrelation):**

در صورتی که تعریف روبرو را در نظر بگیریم:

$$s_n(m)=\begin{cases}s(m+n).w(m), & 0\le m\le N-1\\ 0, & otherwise.\end{cases}$$

متوسط مربعات خطا برابر است با:

$$E_n=\sum_{m=0}^{N-1+p}e_n^2(m)$$

$$\phi_n(i,k)=\sum_{m=0}^{N-1+p}s_n(m-i)s_n(m-k),\qquad \begin{matrix}1\le i\le p\\ 0\le k\le p\end{matrix}$$

$$\phi_n(i,k)=\sum_{m=0}^{N-1-(i-k)}s_n(m)s_n(m+i-k),\qquad \begin{matrix}1\le i\le p\\ 0\le k\le p\end{matrix}.$$

به این دلیل که $\phi_n(i,k)$ تنها تابعی از i-k می باشد، تابع کواریانس به یک تابع ساده خودهمبستگی کاهش می یابد:

$$\phi_n(i,k)=r_n(i-k)$$

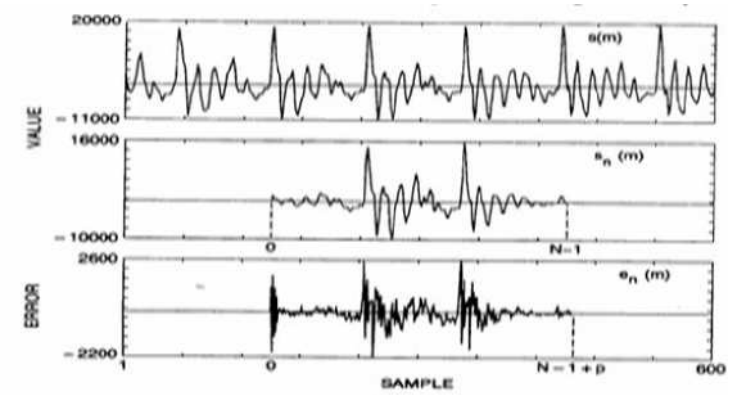
چون تابع خودهمبستگی متقارن است:

$$r_n(-k)=r_n(k)\rightarrow\sum_{k=1}^{p}r_n(|i-k|)\hat{a}_k=r_n(i),\qquad 1\le i\le p$$

$$\begin{bmatrix}r_n(0) & r_n(1) & r_n(2) & \dots r_n(p-1)\\ r_n(1) & r_n(0) & r_n(1) & \dots r_n(p-2)\\ r_n(2) & r_n(1) & r_n(0) & \dots r_n(p-3)\\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \dots\ r_n(0)\end{bmatrix}\begin{bmatrix}\hat{a}_1\\ \hat{a}_2\\ \hat{a}_p\end{bmatrix}=\begin{bmatrix}r_n(1)\\ r_n(2)\\ r_n(3)\\ r_n(p)\end{bmatrix}.$$

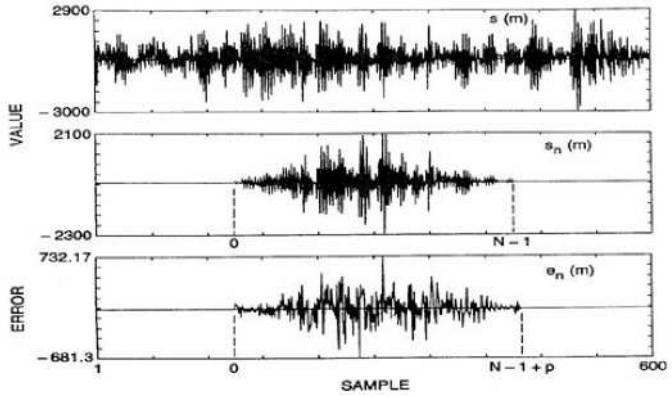
با حل معادله بالا که p معادله و p مجهول است (p مجهول همان ak ها هستند) ضرایب به دست می آیند.

در تصویر ۲ نمونه ای از یک سیگنال صدادار و خطای پیشگویی آن را مشاهده می کنید.



تصویر ۲ – نمونه ای از یک سیگنال صدادار و خطای پیشگویی آن

در تصویر ۳ نمونه ای از یک سیگنال بدون صدا و خطای پیشگویی آن را مشاهده می کنید.



تصویر ۳ – نمونه ای از یک سیگنال بدون صدا و خطای پیشگویی آن

۲. روش کوواریانس:

در این روش بازه محاسبه خطا را ۰<=m<=N-۱ در نظر می گیریم.

$$E_n = \sum_{m=0}^{N-1} e_n^2(m)$$

سپس از گفتار بدون وزن دهی مستقیماً استفاده می کنیم.

و $\phi_n(i,k)$ به یکی از دوصورت روبرو تعریف می شود:

$$\phi_n(i,k) = \sum_{m=0}^{N-1} s_n(m-i)s_n(m-k), \quad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix}$$

$$\phi_n(i,k) = \sum_{m=-i}^{N-i-1} s_n(m)s_n(m+i-k), \quad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix}.$$

نهایتاً به فرمول زیر می رسیم:

$$\begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \phi_n(1,3) & \cdots & \phi_n(1,p) \\ \phi_n(2,1) & \phi_n(2,2) & \phi_n(2,3) & \cdots & \phi_n(2,p) \\ \phi_n(3,1) & \phi_n(3,2) & \phi_n(3,3) & \cdots & \phi_n(3,p) \\ \vdots & \vdots & \vdots & & \vdots \\ \phi_n(p,1) & \phi_n(p,2) & \phi_n(p,3) & \cdots & \phi_n(p,p) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \phi_n(1,0) \\ \phi_n(2,0) \\ \phi_n(3,0) \\ \vdots \\ \phi_n(p,0) \end{bmatrix}.$$

ماتریس کوواریانس نتیجه شده متقارن است ولی Toeplitz نیست.

می توان این معادله را بوسیله یک سری روش ها به نام تجزیه Cholesky حل کرد.

۳. روش لوینسون–دوربین

فرمول های زیر به صورت جلو رونده محاسبه می شود:

$$E^{(0)} = r(0)$$

$$k_i = \left\{ r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|) \right\} \Big/ E^{(i-1)}, \quad (*) \qquad 1 \le i \le p$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$$

$$E^{(i)} = (1-k_i^2) E^{(i-1)},$$

note: the summation in (*) is ommitted for i = 1

به ak ها ضرایب پیشگویی، به kiها ضرایب PARCOR و به gm که در زیر تعریف می شود ضرایب لگاریتم گفته می شود.

$$g_m = \log \text{area ratio coefficients} = \log\left(\frac{1-k_m}{1+k_m}\right).$$

این ضرایب قابل تبدیل شدن به هم هستند. یعنی با داشتن یک سری از آن ها می توان سری دیگر را یافت.

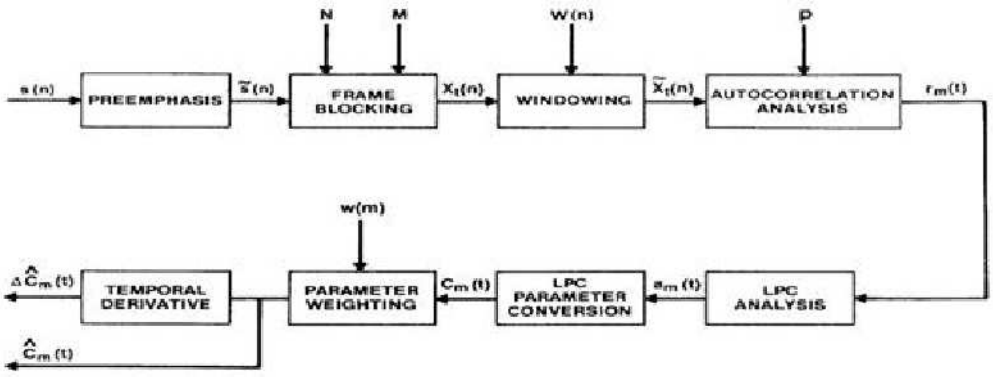
در نتیجه می توان مدل را با هر کدام از این ضرایب به عنوان پارامتر مدل کرد.

۴– ضرایب کپسترال پیشگویی خطی

برای کاربردهای بازشناسی گفتار از پارامترهای LPC ضرایب کپسترال استخراج می شود.

همان طور که در فصل پیش دیدیم با بردن ویژگی ها به حوزه مپستروم، ویژگی های نهایی تقریباً غیرهمبسته می شوند.

خلاصه این کار در تصویر ٤ نشان داده شده است.



تصویر ٤– خلاصه استخراج ضرایب کپسترال پیشگویی خطی

شرح بلوک های تصویر ٤ به صورت زیر است:

- **پیش تاکید:** معمولاً یک فیلتر مشتق گیر می باشد اثر فرکانس های بالا را بیشتر می کند تا بتواند با فیلتر محیط که فرکانس های بالا را تضعیف می کند مقابله کند.

$$H(z) = 1 - \tilde{a} z^{-1}, \qquad 0.9 \le a \le 1.0.$$

$$\tilde{s}(n) = s(n) - \tilde{a} s(n-1).$$

تصویر پاسخ فرکانسی یک فیلتر پیش تاکید در تصویر ٥ آمده است.

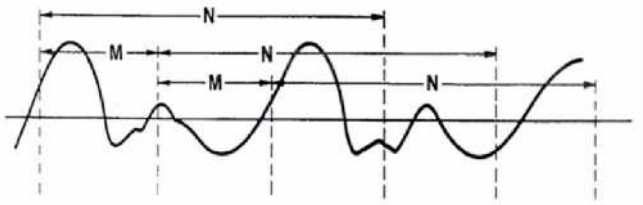
تصویر ٥–فیلتر پیش تاکید با ضریب ۰.۹۵

- **فریم بندی:** مانند دیگر کاربردهای پردازش گفتار ابتدا سیگنال به فریم های کوتاهی تقسیم می شود.

اگر بخواهیم این عمل را فرموله کنیم به صورت روبرو به دست می آید.

$$x_\ell(n) = \tilde{s}(M\ell + n), \qquad \begin{matrix} n = 0,1,\ldots,N-1 \\ \ell = 0,1,\ldots,L-1. \end{matrix}$$

خلاصه این عمل را در تصویر ٦ مشاهده می کنید.



تصویر ٦– خلاصه عمل فریم بندی

- **پنجره گذاری**: پس از فریم بندی روی هر فریم یک پنجره ضرب می کنیم تا مرزهای فریم را به سمت صفر سوق دهد تا گسستگی در سیگنال کم شود.

$$\tilde{x_\ell}(n) = x_\ell(n)w(n), \qquad 0 \le n \le N-1.$$

پنجره هایی مثل همینگ و هنینگ استفاده می شوند که در مرز ها به صفر متمایل می شوند.

- **تحلیل خودهمبستگی**: همان طور که گفته شد برای محاسبه ضرایب از تحلیل خودهمبستگی استفاده می شود.
- **تحلیل پیشگویی خطی**: به روش لوینسون–دوربین ضرایب پیشگویی محاسبه می شوند.
- **انتقال به حوزه کپستروم**: بوسیله فرمول زیر و به صورت جلو رونده ضرایب کپسترال LPC محاسبه می شوند.

$$c_0 = \ln \sigma^2 \qquad \sigma^2 \text{ is the gain term in LPC model}$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k,} \qquad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k,} \qquad m > p,$$

- **وزن دهی پارامترها**: ضرایب کپسترال درجه پایین به شیب کلی طیف وابسته اند.

  درجات بالای ضرایب کپسترال بیشتر به نویز حساس اند.

  وزن دهی طوری انجام می شود که این حساسیت ها برطرف شود.
- **پردازش زمانی**: همان طور که در فصل گذشته توضیح داده شد، ضرایب دلتا و دلتادلتا به بردار ویژگی اضافه می شوند.

در تصویر ۷ مقادیر معمول تحلیل LPC را مشاهده می کنید.

فرض کنید که:

- N تعداد نمونه های هر فریم
- M تعداد شیفت های بین فریم ها
- P درجه تحلیل LPC
- Q بعد بردار کپسترال LPC
- K تعداد فریم هایی که مشتق های زمانی بر روی آن ها محاسبه می شوند

Parameter	Fs=6.67kHz	Fs=8kHz	Fs=10kHz
N	300 (45 msec)	240 (30 msec)	300 (30 msec)
M	100 (15 msec)	80 (10 msec)	100 (10 msec)
p	8	10	10
Q	12	12	12
K	3	3	3

تصویر ۷–مقادیر معمول تحلیل LPC برای بازشناسی گفتار با توجه به فرکانس نمونه برداری

۵– خلاصه و نتیجه گیری

در انی فصل با مفهوم ضرایب پیشگویی خطی آشنا شدیم.

در این فصل با نحوه محاسبه ضرایب پیشگویی خطی آشنا شدیم.

در این فصل با نحوه محاسبه ضرایب کپسترال پیشگویی خطی آشنا شدیم.

۶ – منابع درس

۱- Rabiner, “Fundumentals of Speech Recognition”
۲- Huang, Acero, “Spoken Language Processing”
۳- Deller, “Discrete-time processing of speech signals”

۱– مقدمه

آشنایی با کدینگ شکل موج

مادولاسیون کدینگ پالس (پی سی ام)

۲– مادولاسیون کدینگ پالس یکنواخت

در کدینگ PCM هر نمونه از سیگنال در حوزه زمان به ۲^B سطح دامنه کوانتایز می شود.

این ۲^B به صورت B بیت نشان داده می شود.

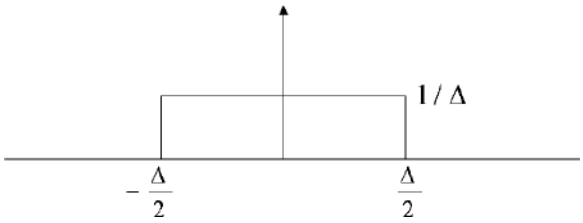
نرخ بیت منبع BF_s b بیت بر ثانیه می باشد.

شکل موج کوانتایز شده به صورت زیر مدل می شود: $\tilde{s}(n) = s(n) + q(n)$

q(n) نشان دهنده خطای کوانتیزیشن می باشد. این نویز را به صورت نویز جمع شونده در نظر می گیریم.

خطای کوانتیزیشن به صورت یک فرآیند تصادفی ایستا q درک می شود. هر متغییر تصادفی q(n) یک تابع چگالی احتمال یکنواخت دارد. $-\frac{\Delta}{2} \le q \le \frac{\Delta}{2}$

اندازه گام کوانتایزر $\Delta = 2^{-B}$ می باشد.



تصویر ۱ – تابع چگالی احتمال

در صورتی که دامنه سیگنال A_{max} باشد، اندازه گام $\Delta \approx \frac{A_{\max}}{2^B}$ می باشد.

میانگین مقدار مربعات خطای کوانتیزیشن برابر فرمول ۱ است.

فرمول (۱)

$$< q^2(n) > = \int_{-\Delta/2}^{\Delta/2} \frac{1}{\Delta} q^2(n) dq$$
$$= \frac{1}{3\Delta} q^3(n) \Big|_{-\Delta/2}^{\Delta/2} = \frac{\Delta^2}{12} = \frac{A_{\max}^2}{2^{2B} \times 12}$$

اندازه در مقیاس دسیبل، مقدار متوسط مربعات نویز برابر است با فرمول ۲.

فرمول ۲
$$10\log_{10}\frac{\Delta^2}{12}=10\log_{10}\frac{2^{-2B}}{12}=-6B-10.8\,\text{dB}.$$

می توان گفت که به ازای اضافه شدن هر بیت، خطای کوانتیزیشن ٦ دسیبل کم می شود.

در صورتی که فاکتور headroom را با h نشان دهیم، انرژی سیگنال را با فرمول ۳ نشان می دهیم.

فرمول ۳
$$X_{rms}=\frac{A_{\max}}{h}=\frac{2^B\Delta}{h}$$

نسبت سیگنال به نویز این کدینگ (با فرض A_{max}=۱) در فرمول ٤ نشان داده شده است.

فرمول ٤
$$\text{SNR}=\frac{S}{N}=\frac{X_{rms}^2}{\Delta^2/12}=12\frac{2^{2B}}{h^2}$$

در دسیبل در فرمول ۵ نشان داده شده است.

فرمول ۵
$$\text{SNR}_{\text{dB}}=10\log_{10}\frac{12\times 2^{2B}}{h^2}=6B+10.8-20\log_{10}h$$

مثال ۱: فرض کنید یک نسبت سیگنال به نویز ٦۰ دسیبل از کدر می خواهیم. فرض کنید فاکتور headroom ٤ مورد نظرمان است.

تعداد بیت های مورد نظر برابر است با $60=10.8+6B-20\log_{10}4$ که $B=\lceil 10.2\rceil=11\,\text{bit}$

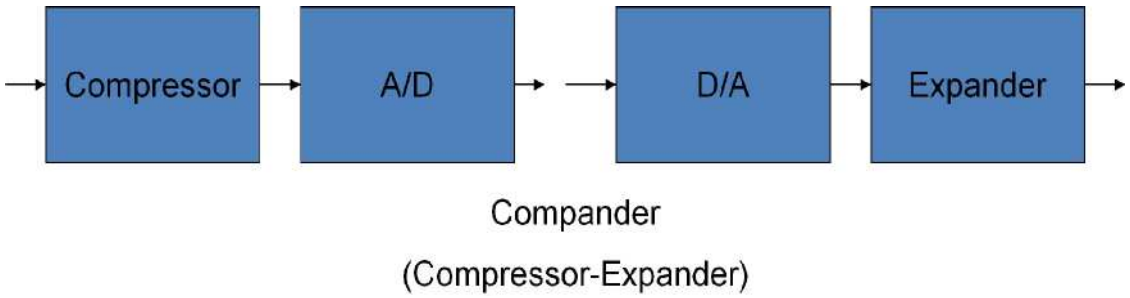
در صورتی که فرکانس نمونه بردار ۸ کیلوهرتز باشد، نرخ ارسال بیت $8k\times 11=88000\,\text{bit/s}.$ می باشد.

۳– مادولاسیون کدینگ پالس غیریکنواخت

یک کوانتایزر غیرخطی معمولاً به این صورت به دست می آید که دامنه سیگنال را از یک دستگاه غیرخطی عبود می دهند تا دامنه سیگنال فشرده شود.

سپس دامنه سیگنال را بوسیله یک کوانتایزر یکنواخت کد می کنند.

در سمت گیرنده عکس این عمل (گذراندنش از تابع معکوس فشرده ساز) انجام می شود (تصویر ۲).



تصویر ۲ – کدینگ پی سی ام غیرخطی

فشرده ساز لگاریتمی

یک فشرده سازی لگاریتمی در سیستم های مخابراتی آمریکای شمالی استفاده می شود.

ویژگی های اندازه ورودی–خروجی به شکل زیر است: $$|y| = \frac{\log(1+\mu|s|)}{\log(1+\mu)}$$

μ پارامتری است که میزان فشرده سازی را مشخص می کند.

به فشرده ساز لگاریتمی به کار رفته در مخابرات اروپا A-law گفته می شود وبه صورت زیر تعریف می شود:

$$|y| = \frac{\log(1+A|s|)}{1+\log A}$$

۶ – خلاصه و نتیجه گیری

در این فصل بحثکدینگ شکل موج را شروع نمودیم.

اولین کدینگ به نام پی سی ام را بیان کردیم.

۷ – منابع درس

۱- Rabiner, “Fundumentals of Speech Recognition”
۲- Huang, Acero, “Spoken Language Processing”
۳- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با کدینگ شکل موج

دی پی سی ام

2– پی سی ام تفریقی (دی پی سی ام)

فرض کنید که یک دنباله u(m) داریم که *u(m),* m=0 to m=n-1

فرض کنید که مقدار دنباله بازتولید شده به صورت روبرو نمایش دهند: $\tilde{u}(n-1),\tilde{u}(n-2),...$

وقتی m=n باشد و u(n) دریافت می شود، یک تخمین $\bar{\tilde{u}}(n)$ از u(n) ، از نمونه های پیشین $\tilde{u}(n-1),\tilde{u}(n-2),...$ تخمین زده می شود (فرمول 1).

فرمول 1 $$\bar{\tilde{u}}(n)=\psi(\tilde{u}(n-1),\tilde{u}(n-2),...);$$

که تابع $\psi(.)$: تابع پیشگویی است.

خطای پیشگویی با فرمول 2 مشخص می شود.

فرمول 2 $$e(n)=u(n)-\bar{\tilde{u}}(n)$$

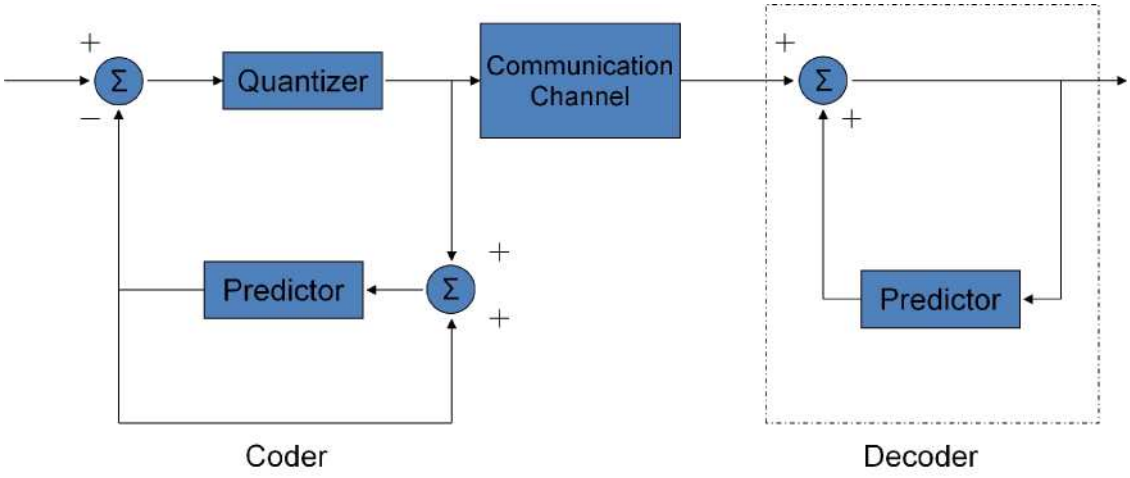
اگر $\tilde{e}(n)$ نشان دهنده کوانتایز شده e(n) باشد، مقدار بازتولید شده برابرا است با فرمول 3.

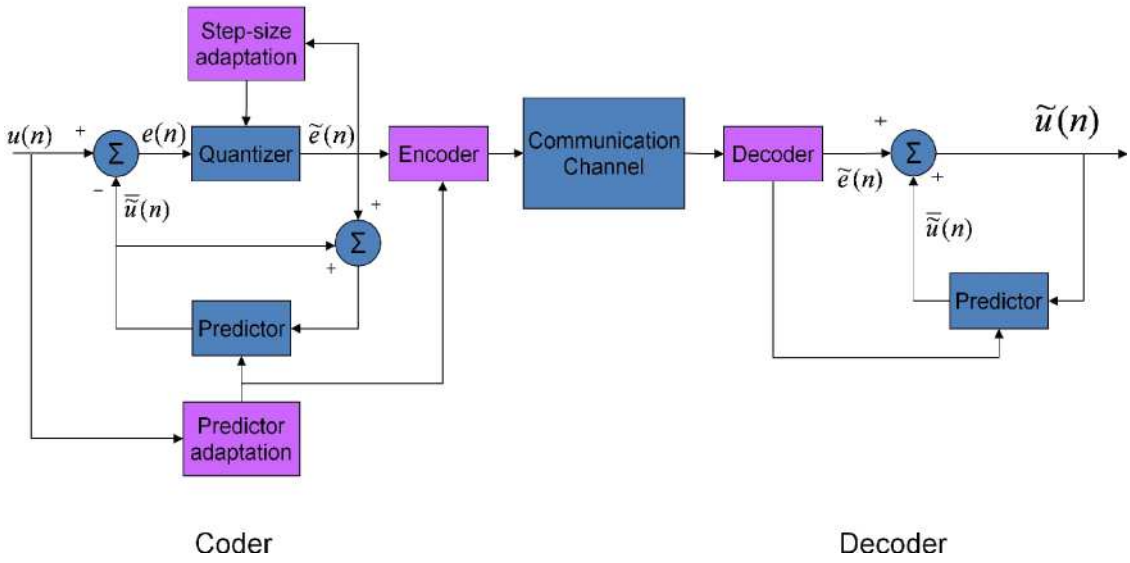
فرمول 3 $$\tilde{u}(n)=\bar{\tilde{u}}(n)+\tilde{e}(n)$$

خطای کوانتیزیشن برابر فرمول 4 است.

$$\begin{aligned} u(n)&=\bar{\tilde{u}}(n)+e(n)\\ u(n)-\tilde{u}(n)&=(\bar{\tilde{u}}(n)+e(n))-(\bar{\tilde{u}}(n)+\tilde{e}(n))\\ &=e(n)-\tilde{e}(n)\\ &=q(n):\text{The Quantization error in } e(n)\end{aligned}$$

در تصویر 1 کدر دی پی سی ام را مشاهده می کنید.





تصویر 2 – دی پی سی ام با پیش بینی کننده خطی

3– پی سی ام تطبیقی و پی سی ام تفریقی تطبیقی (APCM و ADPCM)

سیگنال های گفتار نیمه ایستا هستند. به عبارتی واریانس و تابع خودهمبستگی خروجی منبع به آهستگی با زمان تغییر می کند.

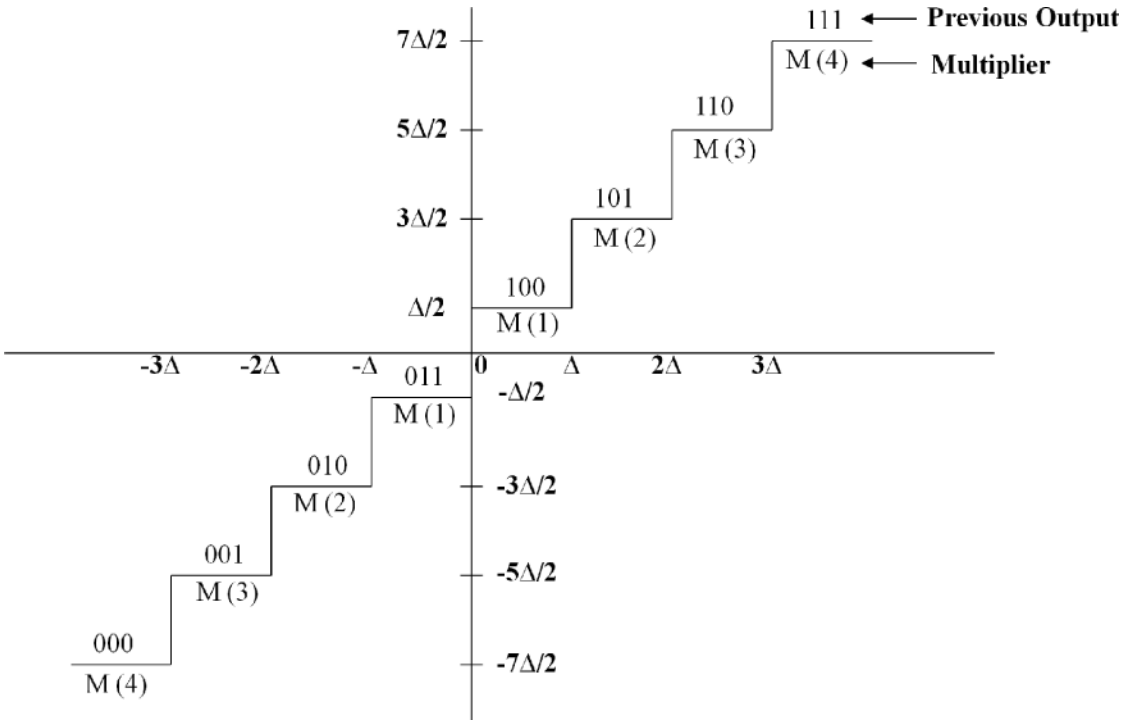
پی سی ام و دی پی سی ام فرض می کنند که خروجی منبع ایستا می باشد.

کارایی این کدرها را می توان با تطبیق به ویژگی های آماری وابسته به زمان بهبود داد.

کوانتایزر تطبیقی به دو صورت می باشند:

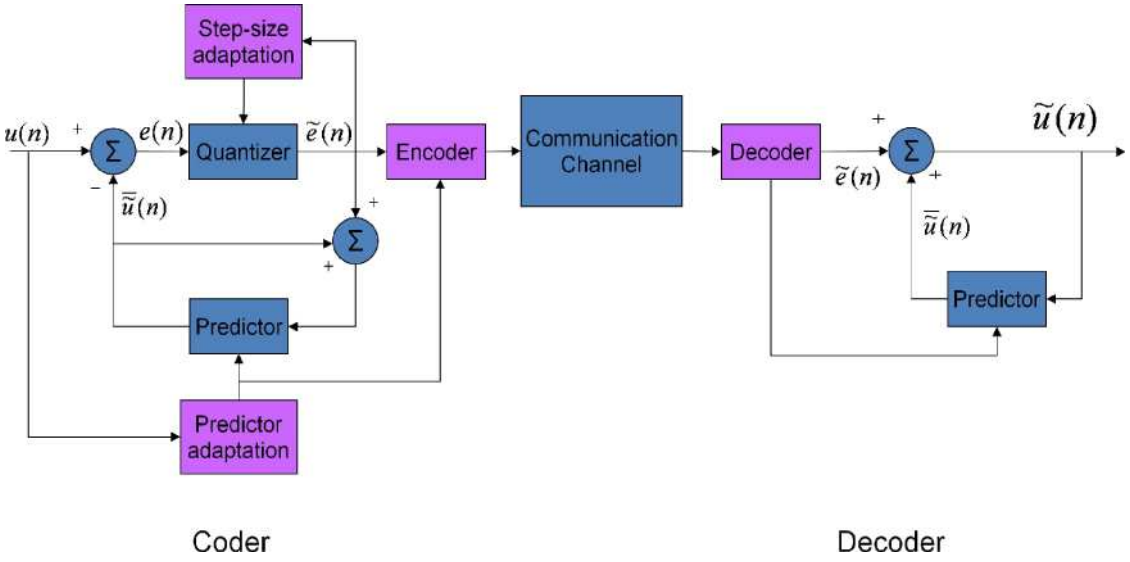
- به سمت جلو (feedforward)
- به سمت عقب (feedbackward)

مثالی از یک کوانتایزر پی سی ام با یک اندازه گام تطبیقی را در تصویر 3 مشاهده می کنید.



تصویر 3 – مثالی از یک کوانتایزر پی سی ام با یک اندازه گام تطبیقی

در تصویر 4 مثالی از یک کوانتایزر دی پی سی ام با پیش بینی کننده تطبیقی مشاهده می کنید.



تصویر 4 – مثالی از یک کوانتایزر دی پی سی ام با پیش بینی کننده تطبیقی

6 – خلاصه و نتیجه گیری

در این فصل بحث کدینگ شکل موج را ادامه دادیم.

کدینگ پی سی ام تفریقی (دی پی سی ام) را بیان کردیم.

کدینگ پی سی ام و دی پی سی ام تطبیقی را نیز بیان کردیم.

7 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با کدینگ شکل موج

دی پی سی ام

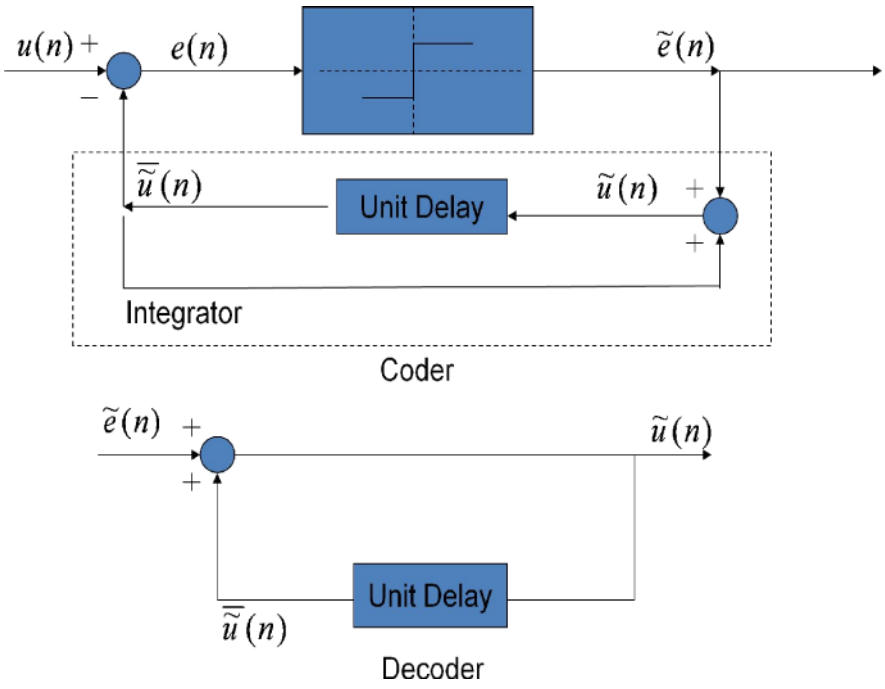
2– دلتا مادولاسیون (دی ام)

نوع خاصی از دی پی سی ام می باشد:

- تابع پیش بینی کننده: تابع تاخیر یک واحدی
- کوانتایزر: یک بیتی

$$\tilde{\tilde{u}}(n) = \tilde{u}(n-1)$$
$$e(n) = u(n) - \tilde{u}(n-1)$$

در تصویر 1 کدر و دیکدر دی ام را مشاهده می کنید

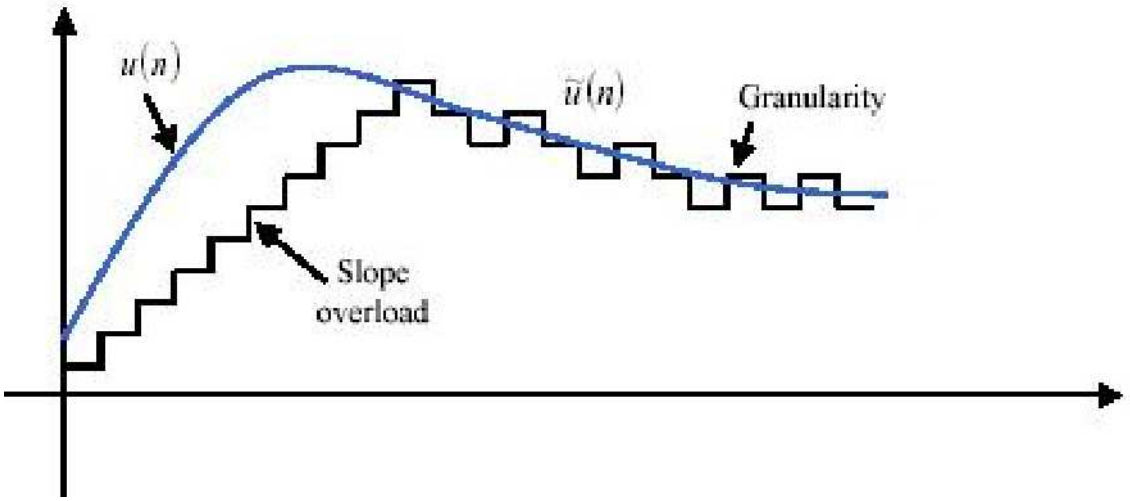


تصویر 1 – کدر و دیکدر دی ام

محدودیت های اصلی در ام عبارتند از:

- سربار شیب (slope overload): در جاهایی که پرش های بزرگ داریم رخ می دهد
  - Max.slope=(step size)X(Sampling Freq)
- نویز دانه دانه بودن (granularity noise): در مکان های با مقدار ثابت رخ می دهد
- نویز عدم ایستایی در کانال

در تصویر 2 این خطاها را مشاهده می کنید.



تصویر 2 – خطاهای شیب و دانه دانه بودن

تاثیر اندازه گام:

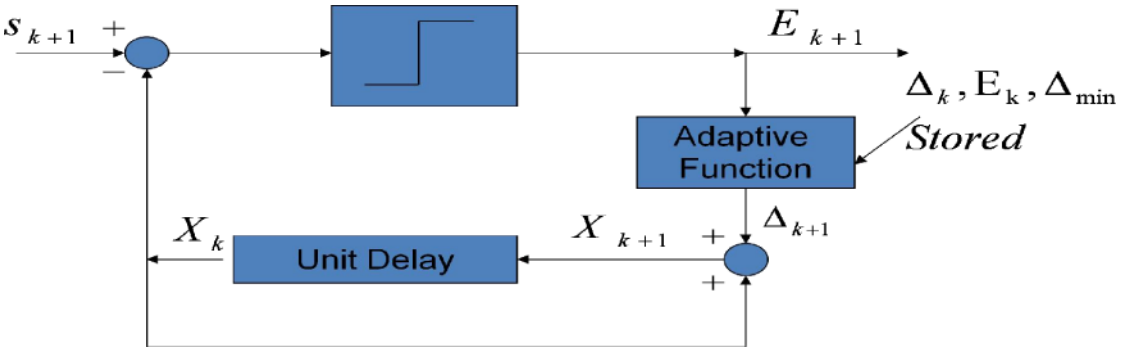
افزایش اندازه گام –> خطای شیب کمتر

افزایش فرکانس نمونه برداری –> نویز دانه دانه شدن کمتر

می توان دی ام را هم مانند دی پی سی ام تطبیقی کرد:

- به این صورت که جاهایی که تغییرات زیاد است اندازه گام را افزایش داد تا خطای شیب کمتر شود و
- در جاهایی که سیگنال خیلی تغییر نمی کند اندازه گام را کم کرد تا خطای دانه دانه شدن کمتر شود.

در تصویر 3 بلوک دیاگرام دی ام تطبیقی را مشاهده می کنید.



تصویر 3 – بلوک دیاگرام دی ام تطبیقی

سیستم نشان داده شده در تصویر 3 به صورت تطبیقی آثار خطای شیب و دانه دانه شدن را کاهش می دهد.

$$E_{k+1} = \text{sgn}[S_{K+1} - X_k]$$

$$\Delta_{k+1} = \begin{cases} |\Delta_k|[E_{k+1} + \frac{E_k}{2}] & \text{if } |\Delta_k| \geq \Delta_{min} \\ \Delta_{min} E_{k+1} & \text{if } |\Delta_k| < \Delta_{min} \end{cases}$$

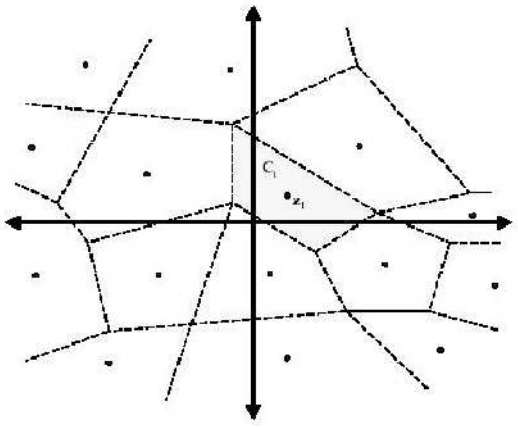
$$X_{k+1} = X_k + \Delta_{k+1}$$

فرمول 1

3– کوانتیزیشین بردار (VQ)

کوانتیزیشین فرآیند نشان دادن دامنه های پیوسته سیگنال بوسیله نماد های گسسته می باشد.

در تصویر 4 یک قسمت بندی از فضای دو بعدی به 16 سلول را مشاهده می کنید.



تصویر 4 –تقسیم بندی فضای دو بعدی به 16 سلول

الگوریتم LBG به این صورت عمل می کند که ابتدا یک کدبوک تک برداره محاسبه می کند. سپس بوسیله الگوریتم تقسیم کننده آن را به دو بردار تقسیم می کند و این فرآیند را تا آنجا ادامه می دهد که کدبوک Mبرداره به دست آید.

الگوریتم LBG به صورت زیر عمل می کند:

- گام اول: مقدار M (تعداد قسمت ها) را برابر 1 بگذار. مرکز همه داده های آموزشی را بیاب.
- گام دوم: M را به 2M قسمت تقسیم کن به این صورت که در هر قسمت دو نقطه را که از هم بیشترین فاصله را دارند بیاب و از آن نقاط جدید برای ساختن 2M کدبوک استفاده کن. حال M را برابر 2M قرار دهید.
- گام سوم: بوسیله یک الگوریتم تکرارشوند، به بهترین مجموعه مراکز برسید.
- در صورتی که M برابر اندازه کدبوک مورد نیاز است، متوقف شو، در غیر این صورت برو به گام دوم.

5 – خلاصه و نتیجه گیری

در این فصل بحث کدینگ شکل موج را ادامه دادیم.

کدینگ دلتا مادولاسیوم (دی ام) بررسی شد.

نسخه تطبیقی دی ام بررسی شد.

بحث کوانتیزیشن بردار مطرح شد.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با وکودر ها

وکدر کانال و وکدر فاز

2– مفاهیم اولیه

هدف از بحث وکدرها طراحی سیستم هایی برای انتقال گفتار انسان است.

معمولاً وکودر ها از دو بخش کدر (رمزکننده) و دیکدر (رمزگشا) استفاده می کنند.

هدف این است که با توجه به خواص گفتار انسان، داده ها را طوری فشرده کنیم که گفتار با کیفیت مطلوبی از سمت فرستنده به سمت گیرنده ارسال شود.

انواع زیادی وکودر ارائه شده است. در این چند جلسه به این مباحث می پردازیم.

3– وکودر کانال

وکودر کانال از یک فیلتربانک استفاده می کنید.

این فیلتر بانک شامل مجموعه ای از فیلترها می باشد.

هر فیلتر پهنای باندی بین 100 تا 300 هرتز دارد.

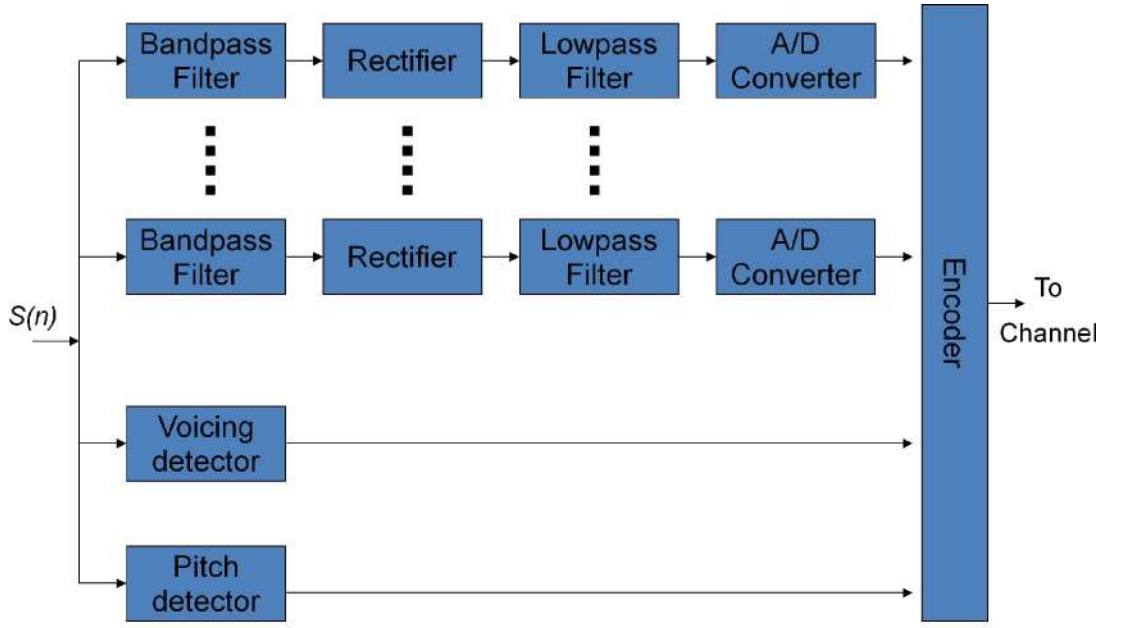
معمولاً از 16 تا 20 فیلتر FIR استفاده می شود.

خروجی خر فیلتر یکسو می شود و سپس از فیلتر پایین گذر عبور داده می شود.

پهنای باند فیلتر پایین گذر طوری انتخاب می شود که به نوسان های زمانی موجود در مسیر صوتی جور شود.

برای اندازه گیری اندازه طیف، یک شناسایی کننده صدادار بودن و یک تخمین زننده فرکانس گام در آنالیز گفتار قرار داده می شود.

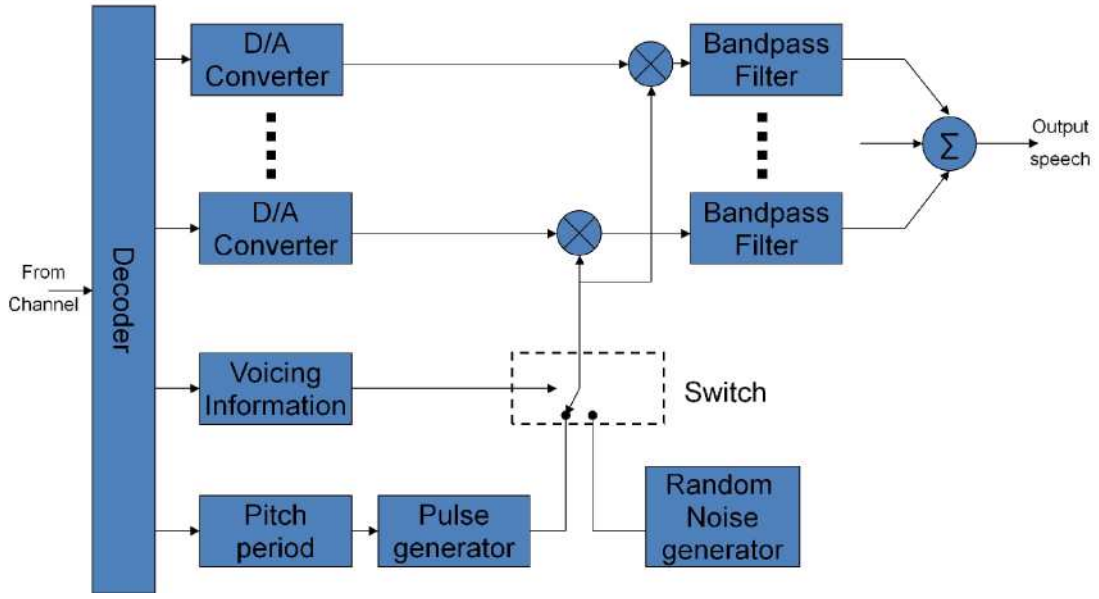
در تصویر 1، قسمت کدر وکودر کانال را مشاهده می کنید.



تص

قسـ

در



تصویر 2– کدگشای وکودر کانال

کدگشا خواص زیر را دارد:

- 16–20 فیلتر FIR فاز خطی

- 0 تا 4 کیلوهرتز را پوشش می دهد
- هر فیلتر پهنای باندی بین 100 تا 300 هرتز دارد
- فریم های 20 میلی ثانیه یا به عبارتی تغییر اندازه طیف به صورت 50 هرتز
- پهنای باند فیلتر پایین گذر 20–20 هرتز
- نرخ نمونه برداری خروجی فیلترها 50 هرتز

نرخ ارسال بیت به صورت زیر محاسبه می شود:

- 1 بیت به ازای شناسایی صدادار بودن
- 6 بیت برای پریود گام
- برای 16 کانال، که هر کدام با 3–4 بیت کد شده اند، هر ثانیه 50 بار آپدیت می شود.
- نرخ ارسال بیت 2400 تا 3200 بیت بر ثانیه می باشد.
- می توان با استفاده از همبستگی های بین اندازه طیف در باندهای مجاور نرخ ارسال بیت را به 1200 بیت بر ثانیه کاهش داد.

در قسمت دریافت کننده، نمونه های سیگنال از یک مبدل دیجیتال به آنالوگ عبور داده می شوند.

خروجی D/A در منابع صدادار بودن یا بدون صدا بودن ضرب می کند.

خروجی های فیلترهای میان گذر با هم جمع می شوند تا سیگنال گفتار سنتز شده را شکل دهند.

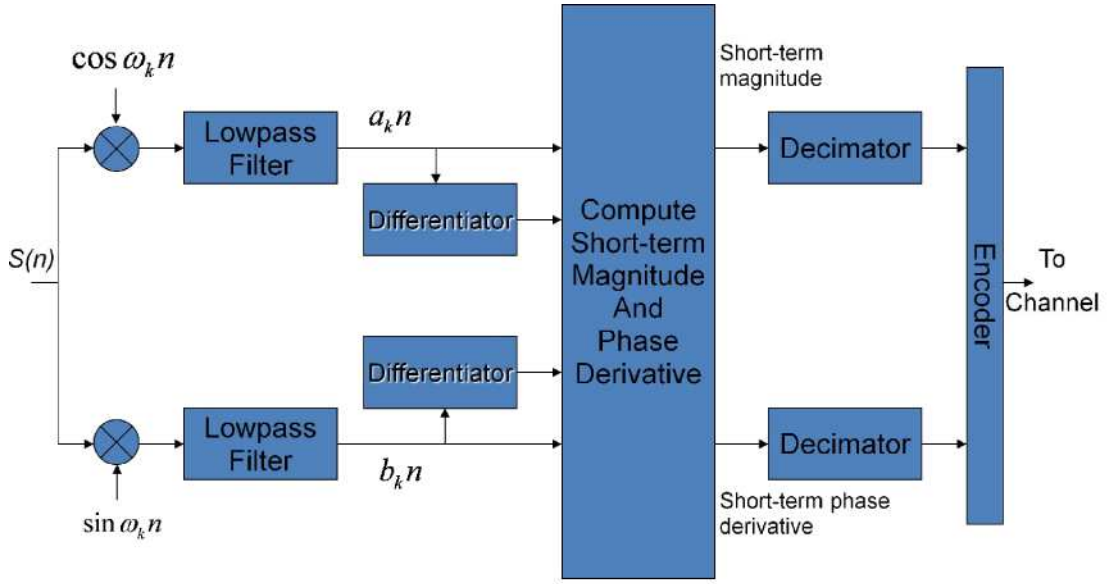
4– وکودر فاز

مشابه وکود کانال می باشد.

ولی به جای تخمین زدن فرکانس گام، مشتق فاز خروجی هر فیلتر تخمین زده می شود.

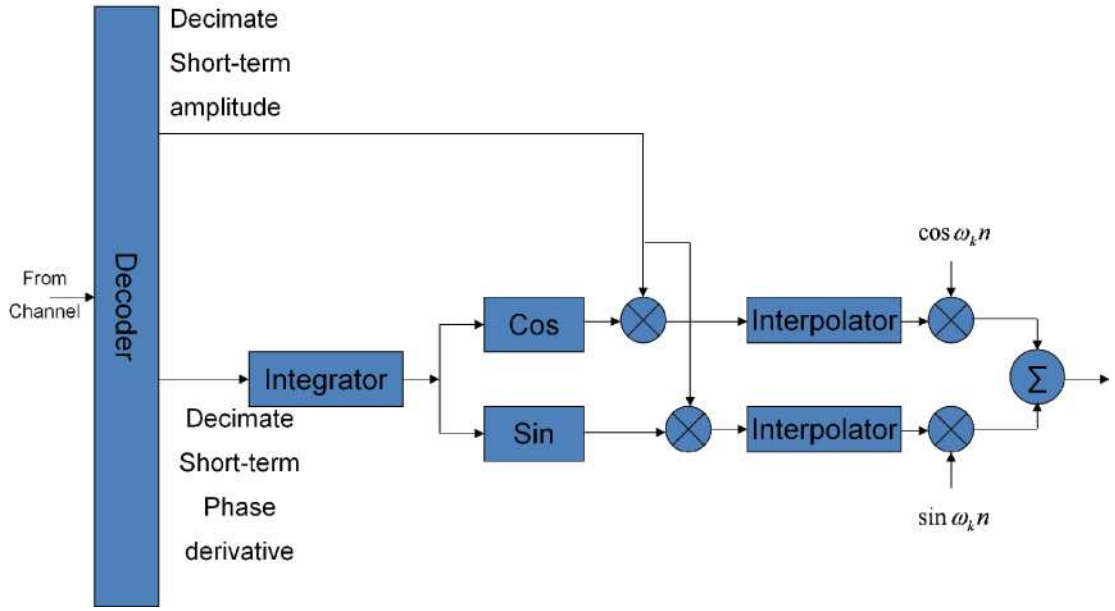
با کد کردن و ارسال مشتق فاز، این وکودر اطلاعات فاز را از بین می برد.

در تصویر 3 کدکننده این وکودر را مشاهده می کنید.



ته

د



تصویر 4 – کدگشای وکودر فاز

ویژگی های معمول این وکودر در زیر آمده است:

- پهنای باند فیلتر پایین گذر: 50 هرتز
- تعداد فیلترها: 25–30

- نرخ نمونه برداری اندازه طیف و مشتق فاز: 50 تا 60 نمونه در هر ثانیه
- اندازه طیف بوسیله PCM یا DPCM کد می شود.
- مشتق فاز به صورت خطی توسط 2–3 بیت کد می شود.
- نرخ ارسال بیت حاصله 7200 بیت بر ثانیه است.

5 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را آغاز نمودیم.

وکودر کانال را بیان کردیم.

وکودر فاز را نیز توضیح دادیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با وکودر ها

وکدر فرمنت و وکدر LPC

2– وکودر فرمنت

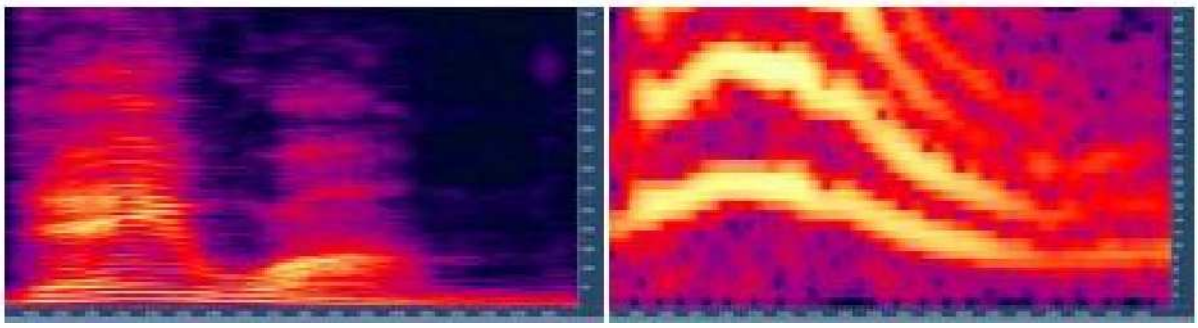
وکودر فرمنت را می توان نوعی وکودر کانال دانست که سه فرمنت اول را برای یک قطعه از گفتار را تخمین می زند.

این اطلاعات به همراه پریود گام کد می شود و به سمت گیرنده ارسال می شود.

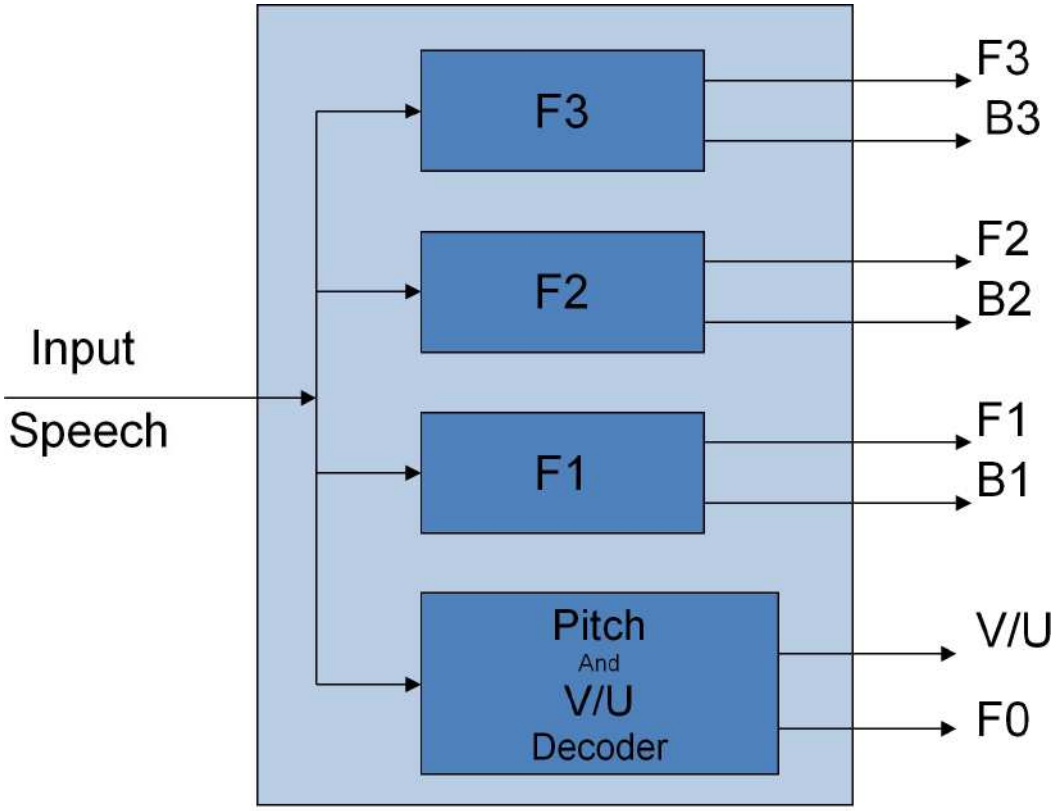
مثال: تصویر 1:

تصویر چپ: اسپکتروگرام تلفظ "day one" که گام و ساختار هارمونیک گفتار را نشان می دهد.

تصویر راست: یک اسپکتروگرام بزرگ نمایی شده از فرکانس گام و هارمونیک دوم.

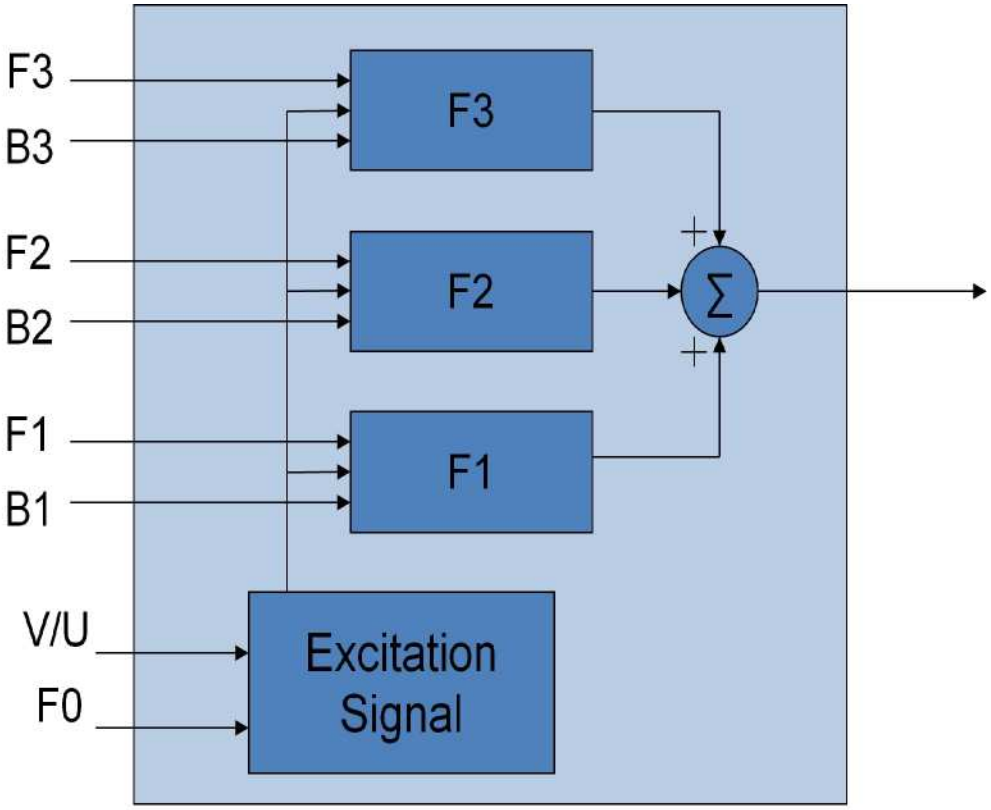


تصویر 1 – اسپکتروگرام



تصـ

Fk :The frequency of the kth formant
Bk :The bandwidth of the kth formant

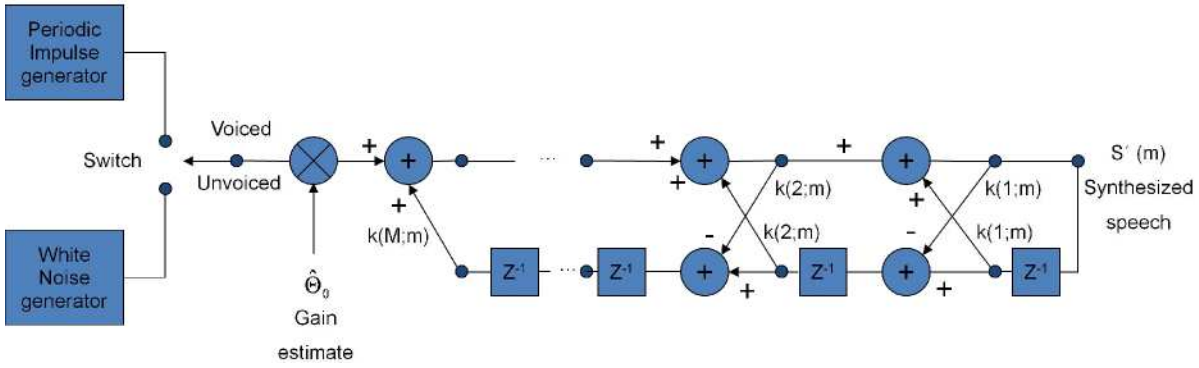


تع

3

ه

چ



تصویر 4- ساختار سنتز به روش پیشگویی خطی

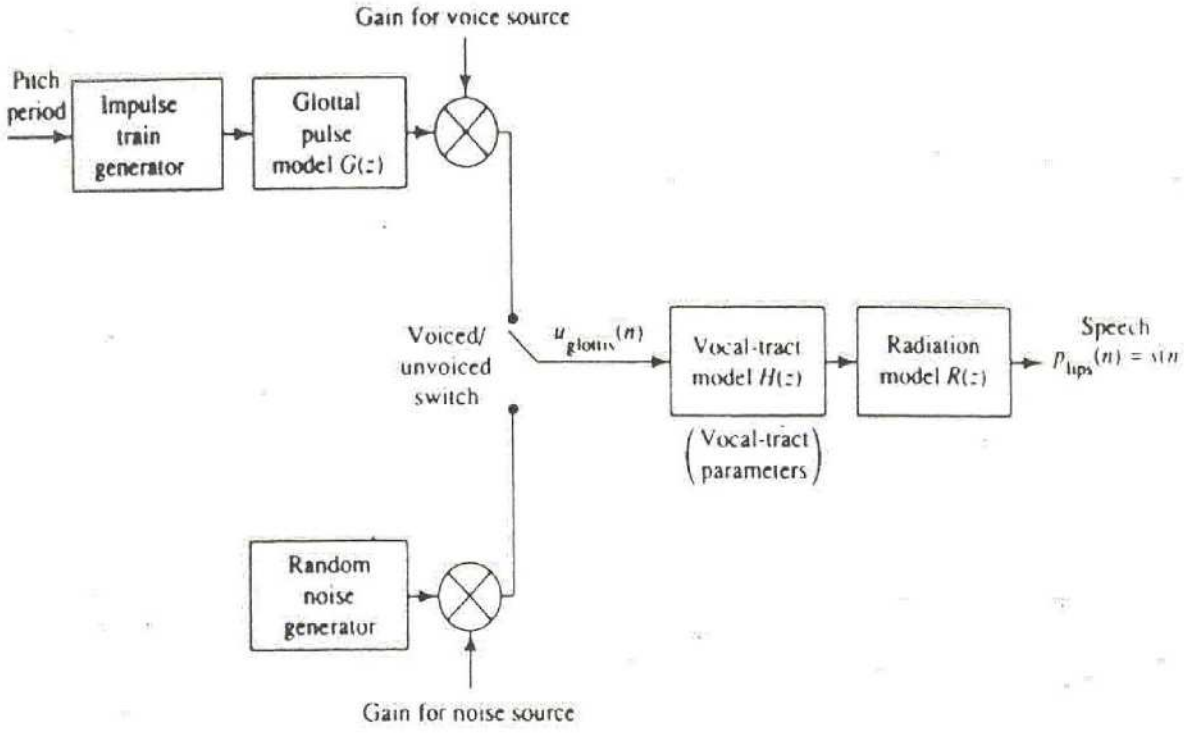
روش LPC10

این روش به این دلیل این نام را دارد که از 10 ضریب برای مدل کردن استفاده می شود.

LPC10 گفتار را به فریم های 180 نمونه ای تقسیم می کند.

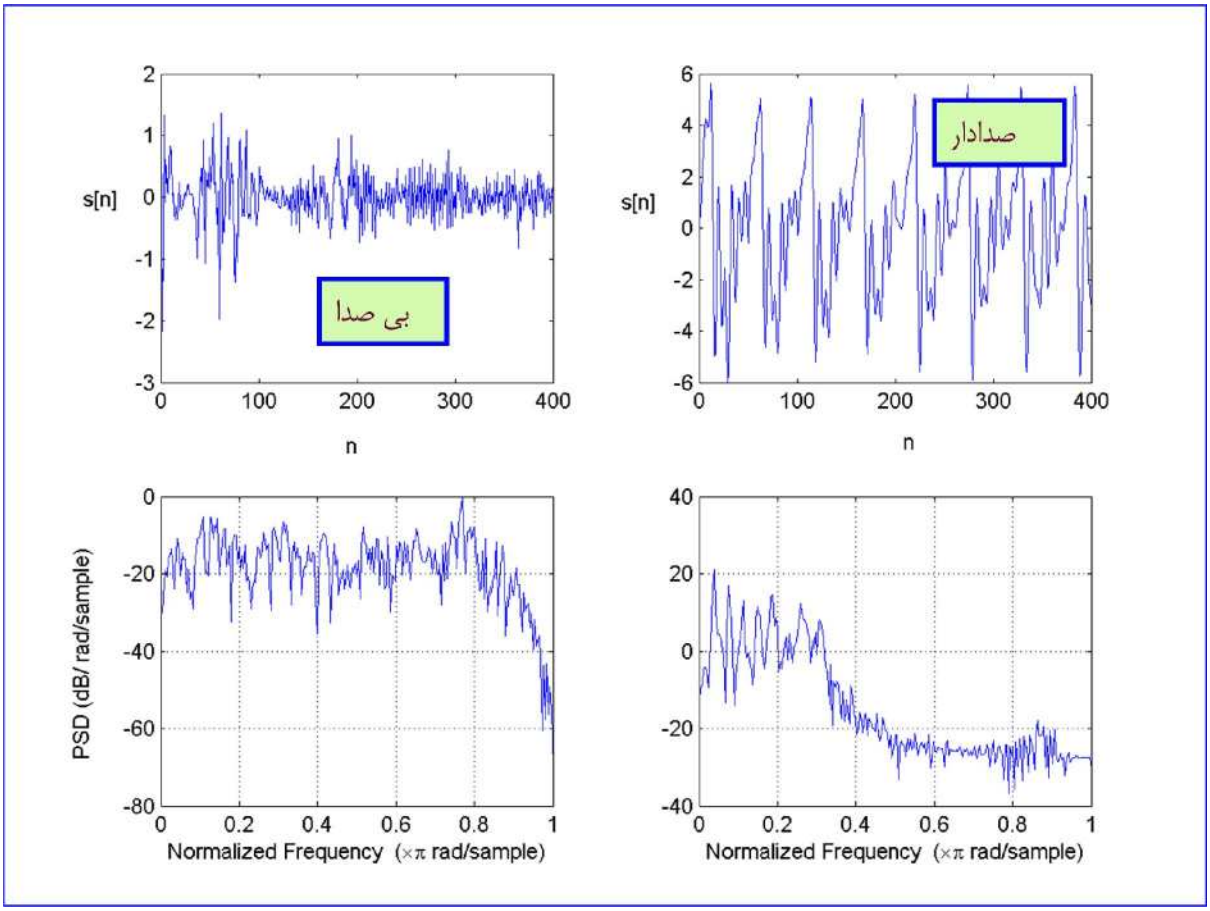
گام و شناسایی صدادار بودن بوسیله روش AMDF و اندازه گیری های نرخ عبور از صفر محاسبه می شوند.

در تصویر 5 یک مدل گسسته در زمان برای تولید گفتار مشاهده می کنید.



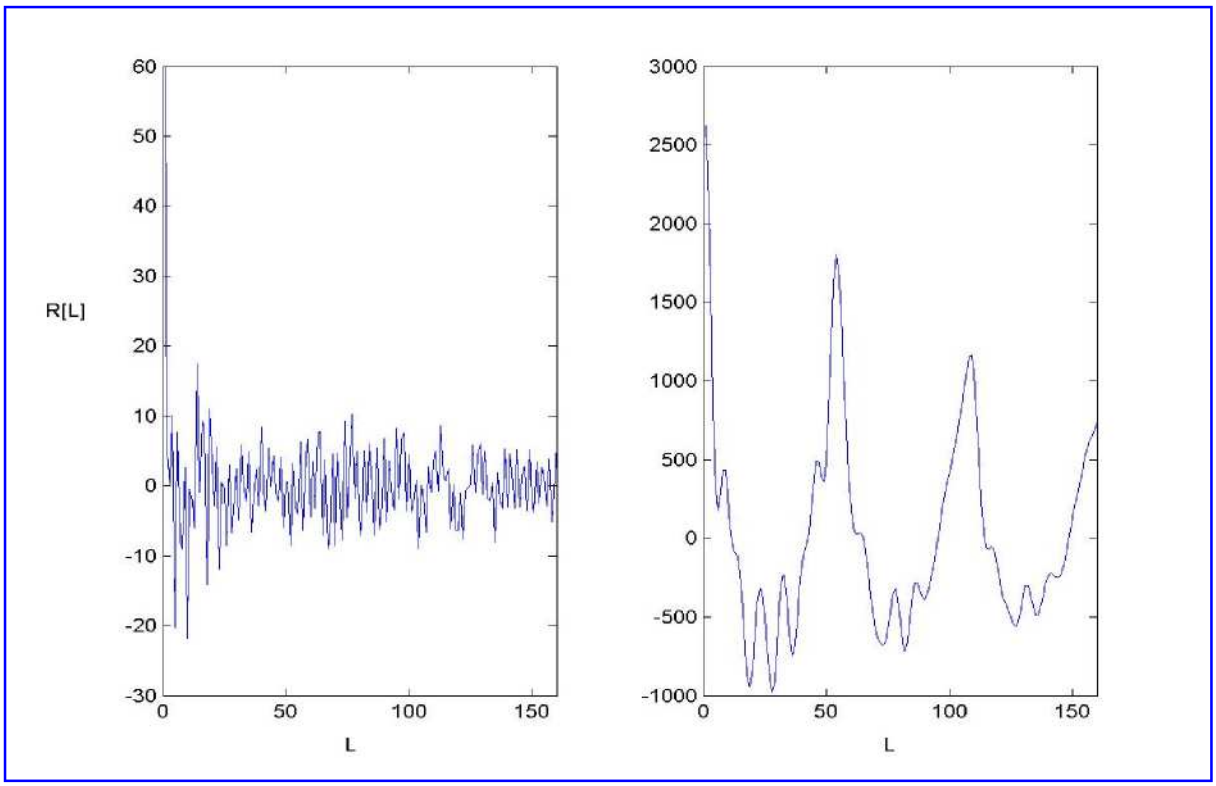
تصویر 5 – مدل گسسته در زمان برای تولید گفتار

در تصویر 6 دو قطعه گفتار بی صدا و صدادار و طیف آن ها را مشاهده می کنید.



تصویر 6 – دو قطعه گفتار بی صدا و صدادار و طیف آن ها

در تصویر 7 خودهمبستگی دو قطعه گفتار تصویر 6 را مشاهده می کنید.

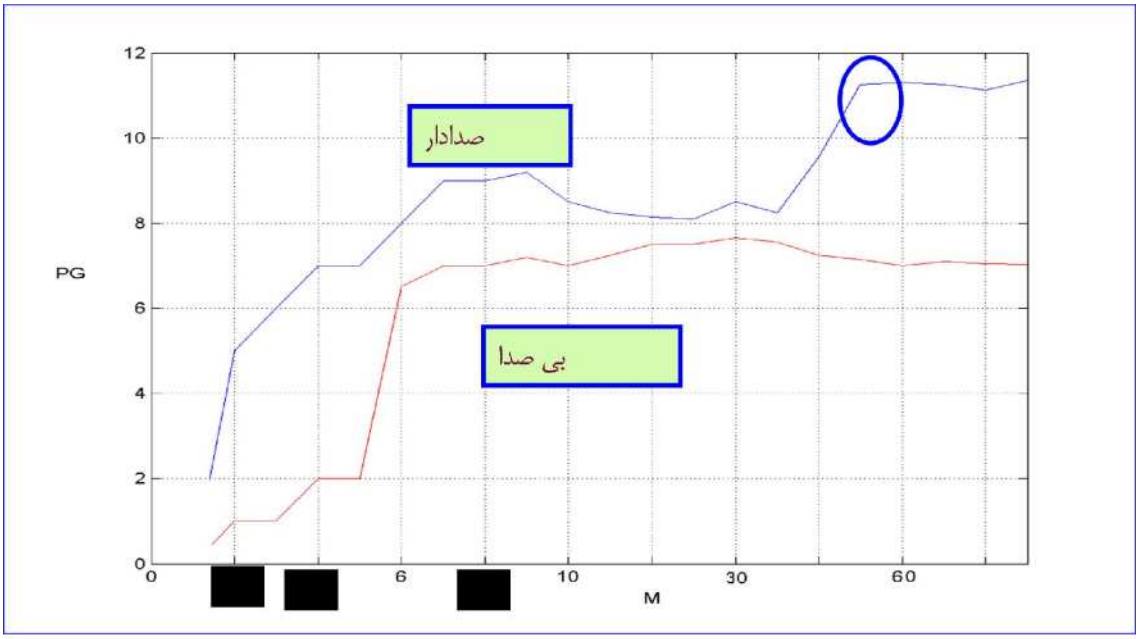


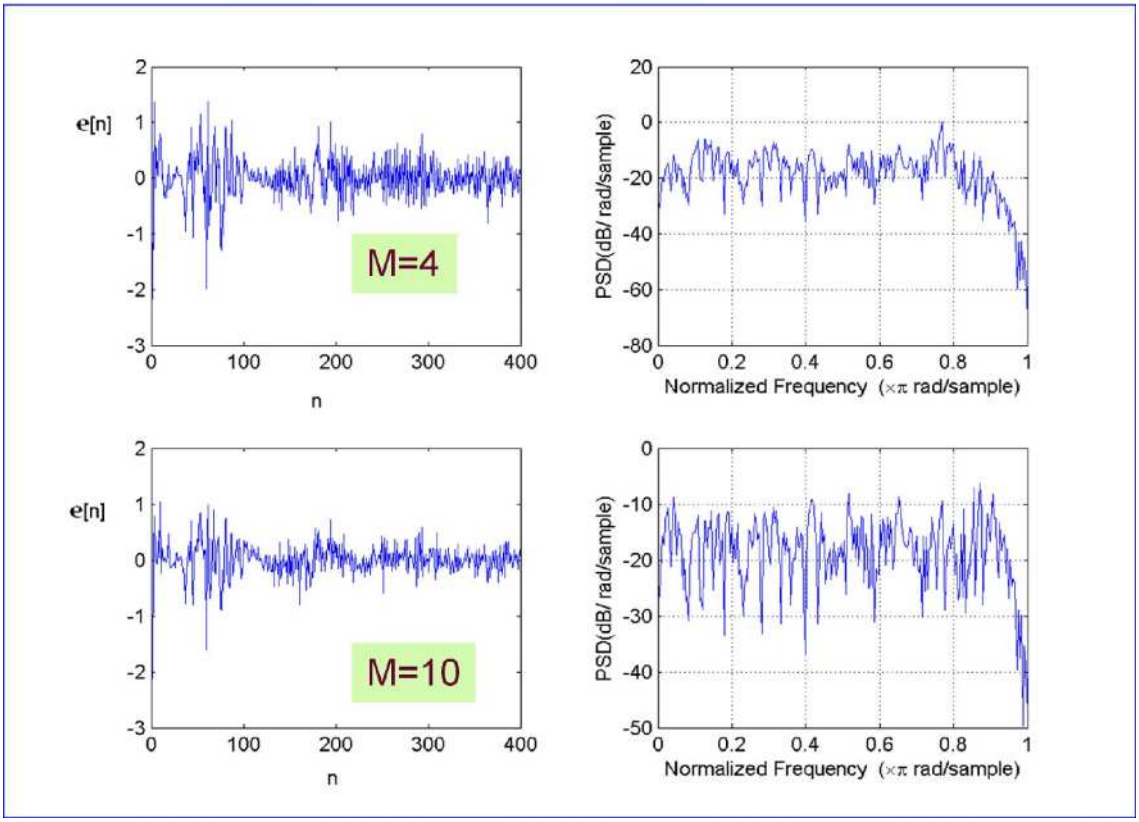
تصویر 6 – خودهمبستگی تصویر 5

بوسیله فرمول 1 مرتبه پیشگویی محاسبه می شود.

فرمول 1

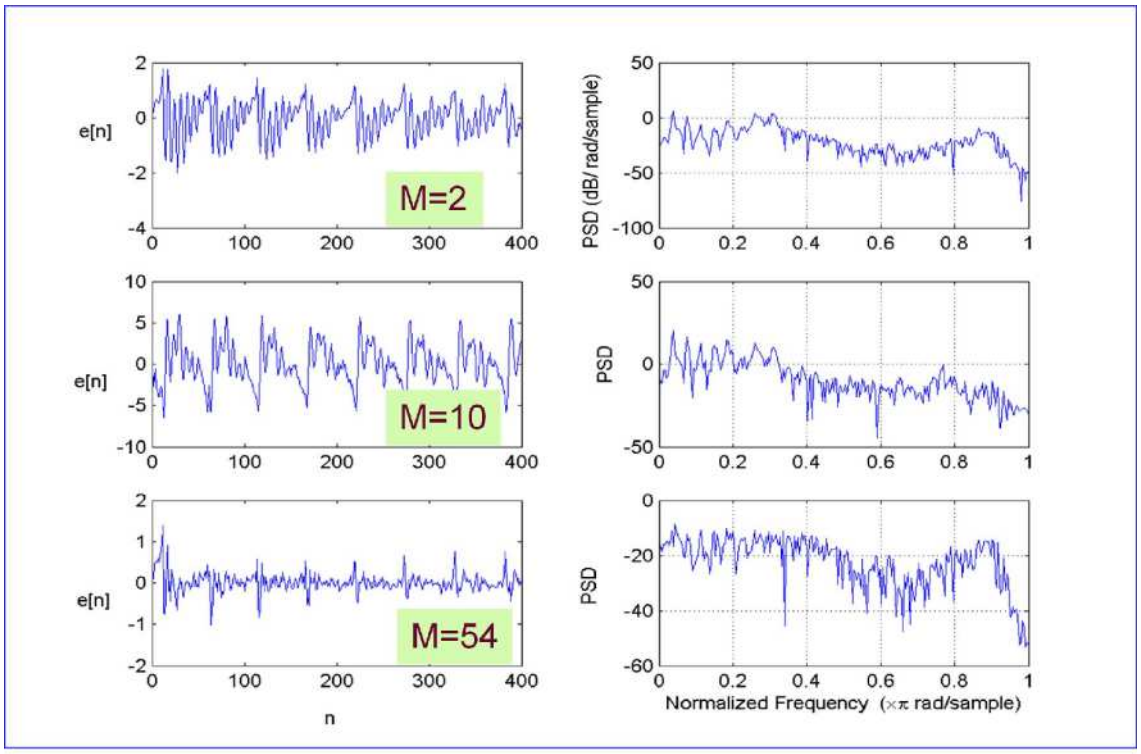
$$PG = 10\log\left(\frac{\sum_{n=m-M+1}^{m} s^2[n]}{\sum_{n=m-M+1}^{m} e^2[n]}\right)$$

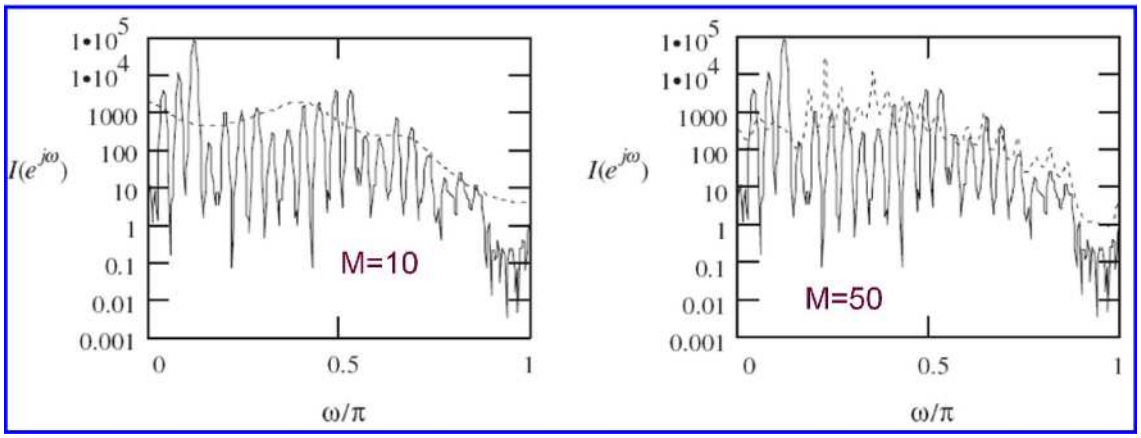

صدادار
بی صدا
PG
M
0
2
4
6
8
10
12
6
10
30
60


e[n]
M=4
n
PSD(dB/ rad/sample)
Normalized Frequency (×π rad/sample)
e[n]
M=10
n
PSD(dB/ rad/sample)
Normalized Frequency (×π rad/sample)

تصویر 8 – خطای پیشگویی را برای تعداد ضرایب مختلف

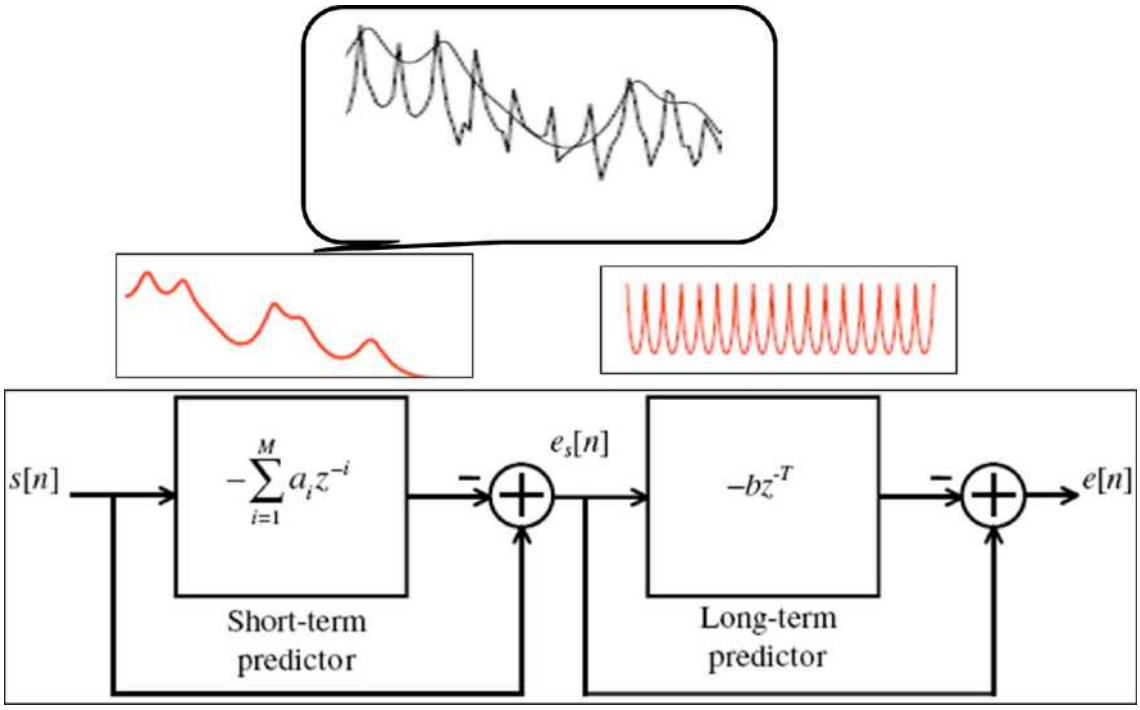
در تصویر 9 خطای پیشگویی برای قطعه صدادار را مشاهده می کنید.





تصویر 10 – پوش طیف برای ضرایب مختلف

ایده پیشگویی خطی بلندمدت را در تصویر 11 مشاهده می کنید.

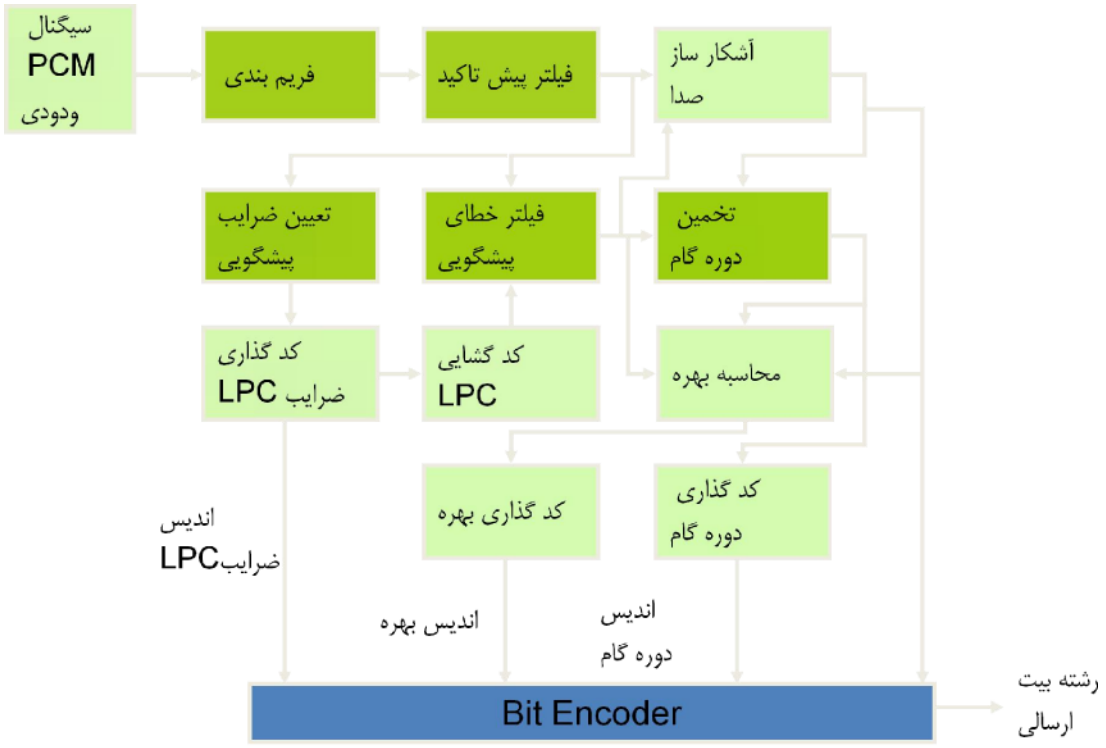


تصویر 11 – ایده پیشگویی خطی بلندمدت

مشخصات عمومی LPC10 را در زیر می بینید:

- بخاطر ارسال 10 ضریب پیشگویی خطی به LPC10 معروف است.
- نرخ ارسال برابر 2400 بیت بر ثانیه می‌باشد.
- تعداد نمونه‌ها در هر فریم برابر 180 نمونه در نظر گرفته شده است.
- تعداد 54 بیت به ازای هر فریم ارسال می‌شود.
- سیگنال آنالوگ ورودی آن با نرخ 8000 هرتز نمونه برداری شده و با 16 بیت کوانتایز می‌شود.

در تصویر 12 قسمت کدکننده LPC10 را مشاهده می کنید.



تصو

برای

$$R[l,m] = \sum_{n=m-N+1}^{m} s[n]s[n-l]$$

$$MDF[l,m] = \sum_{n=m-N+1}^{m} |s[n]-s[n-l]|$$

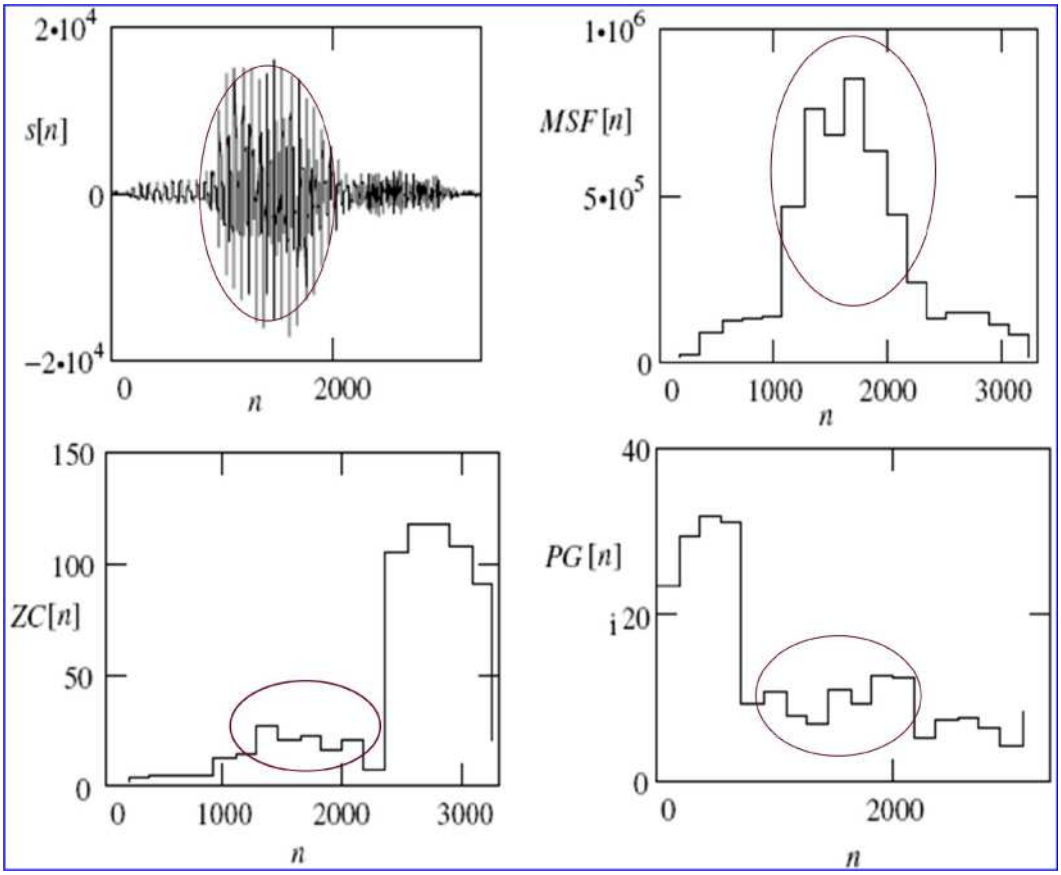
$$s[n] = b.s[n-N]+e[n], \quad m-N+1 \le m$$

برای آشکارسازی صدا از ویژگی های زیر استفاده می شود (تصویر 13):

1- محاسبه انرژی (باند پایین)

2- محاسبه نرخ عبور از صفر

3- محاسبه بهره پیشگویی



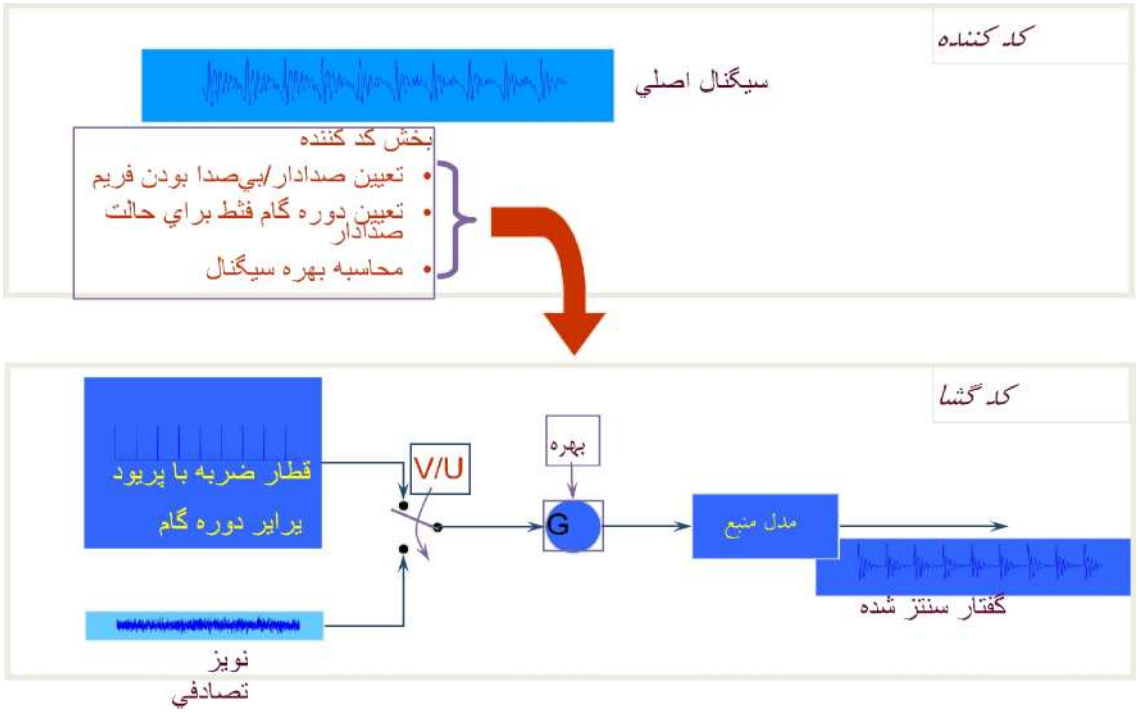
تصویر 13 – آشکارسازی صدادار بودن بوسیله ویژگی های گفته شده

ضرایب LPC به صورت تصویر 14 کوانتایز می شوند.

Parameter	Resolution	
	Voiced	Unvoiced
g_1 (LAR)	5	5
g_2 (LAR)	5	5
k_3 (RC)	5	5
k_4 (RC)	5	5
k_5 (RC)	4	—
k_6 (RC)	4	—
k_7 (RC)	4	—
k_8 (RC)	4	—
k_9 (RC)	3	—
k_{10} (RC)	2	—

تصویر 14 – کوانتیزیشن ضرایب پیشگویی خطی در LPC10

در وکودر LPC10 به صورت تصویر 15 عمل سنتز انجام می شود.



تصویر 15 – سنتز LPC10

محدودیت های LPC10:

1- تقسیم بندی به دو قسمت صدادار و بی‌صدا

2- استفاده از نویز تصادفی و قطار ضربه پریودیک جهت تحریک(قطار ضربه تنها نمی‌تواند تمامی صوتهای واکدار را ایجاد کند.)

3- حفظ نشدن فاز سیگنال اصلی

4- استفاده از قطار ضربه یک تخطی از مدل AR است.

5 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را ادامه دادیم

وکودر فرمنت را بیان کردیم.

وکودر LPC را نیز توضیح دادیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با وکودر ها

وکدر Residual Excited LP

وکدر Multipulse LP

2– وکودر RELP

همان طور که از اسم این وکودر مشخص است از سیگنال residual به عنوان سیگنال تحریک استفاده می کندک.

کیفیت سیگنال را می توان با فرستادن تعداد بیشتری بیت افزایش داد. به این صورت که خطای residual در قسمت کدکننده محاسبه شده و ارسال می شود (همانند DPCM).

یک روش به این صورت است که مدل LPC و پارامترهای تحریک از یک فریم گفتار تخمین زده شوند.

گفتار در قسمت کدکننده سنتز می شود و از سیگنال اصلی تفریق می شود که خطای residual نتیجه این عمل است.

خطای residual کوانتایز می شود، سپس کد شده و به سمت گیرنده ارسال می شود.

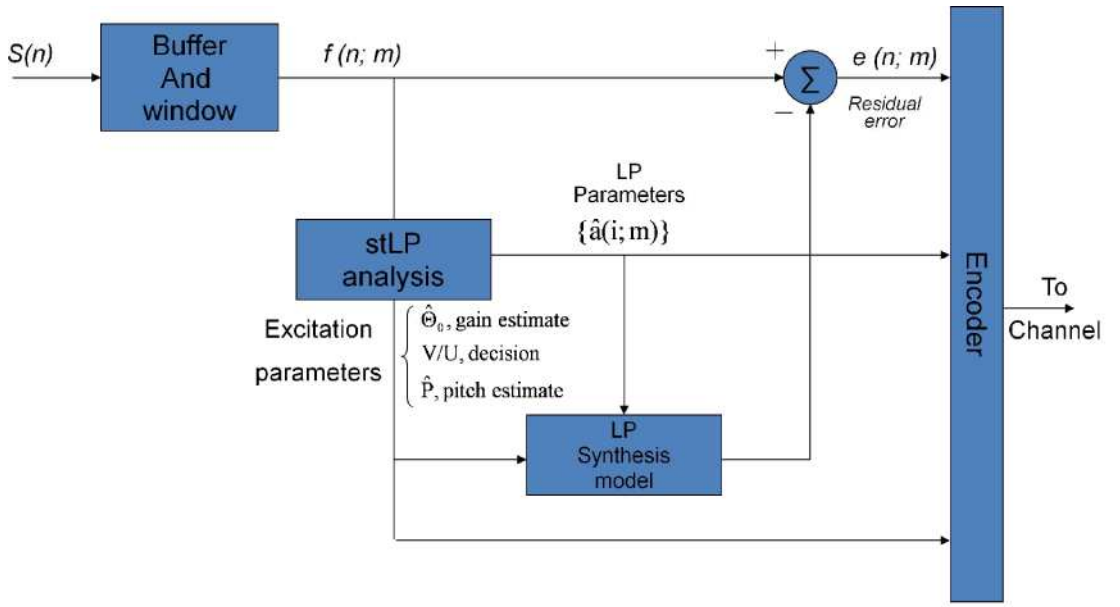
در قسمت گیرنده سیگنال سنتز شده با جمع کردن خطای residual به سیگنال تولید شده از مدل حاصل می شود.

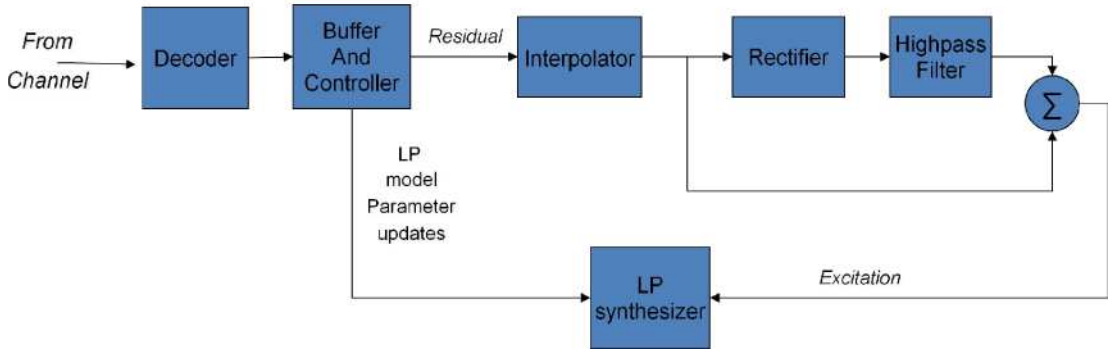
سیگنال residual از فیلتر پایین گذر 1000 هرتز عبور داده می شود تا نرخ بیت کاهش پیدا کند.

در سنتزکننده، این سیگنال یکسو می شود و طیف آن بوسیله فیلتر بالاگذر صاف می شود، سپس سیگنال های پایین گذر و بالاگذر جمع می شوند و سیگنال خطای residual به دست آمده برای تحریک مدل LPC استفاده می شوند.

وکودر های RELP یک کیفیت مناسب با نرخ ارسال بیت 9600 بیت بر ثانیه فراهم می کنند.

در تصویر 1 کدکننده RELP را مشاهده می کنید.





تصویر 2– کدگشای وکودر RELP

2– **وکودر** Multipulse Excitation

RELP باید قسمت های فرکانس بالا را در قسمت کدگشا بازتولید کند.

وکودر Multipulse LPC یک روش در حوزه زمان آنالیز–با–سنتز می باشد که بوسیله آن می توان سیگنال های تحریک بهتری یافت.

اطلاعاتی که دنباله تحریک را شامل می شود عبارتند از:

- مکان پالس ها
- یک فاکتور مقیاس کلی به بزرگترین دامنه پالس
- دامنه های پالس به نسبت آن فاکتور مقیاس

فاکتور مقیاس به صورت لگاریتمی به 6 بیت کوانتایز می شود.

مکان پالس ها بوسیله یک روش کدینگ تفریقی کد می شوند.

پارامترهای تحریک هر 5 میلی ثانیه به روز می شوند.

پارامترهای پیشگویی خطی مسیر صوتی و پریود گام هر 20 میلی ثانیه به روز می شوند.

نرخ ارسال بیت 9600 بیت بر ثانیه می باشد.

یک دنباله ذخیره شده (که به صورت کدبوکی از تحریک های گووسی می باشد) مقیاس می شود و به یک فیلتر سنتز گام و سپس فیلتر سنتز LPCاعمال می شود.

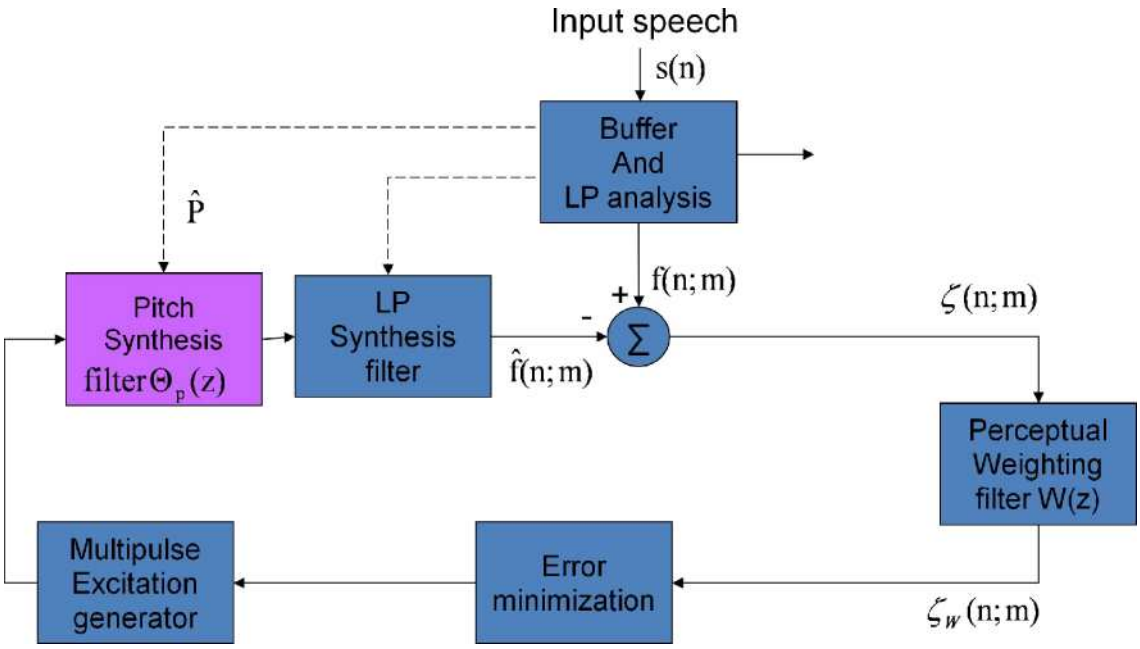
گفتار سنتز شده با گفتار اصلی مقایسه می شود.

سیگنال خطای residual بوسیله فیلتر فرمول 1 وزن دهی می شود.

فرمول 1

$$W(z) = \frac{\hat{\theta}(z/c)}{\hat{\theta}(z)} = \frac{\hat{A}(z)}{\hat{A}(z/c)}$$

در تصویر 3 کدکننده MultipulseLPC را مشاهده می کنید.



تصویر 3 – کدکننده وکودر Multipulse LPC

5 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را ادامه دادیم

وکودر RELP را بیان کردیم.

وکودر MultipulseLPC را نیز توضیح دادیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با وکودر ها

وکدر Code Excited LP

وکدر low-delay CELP

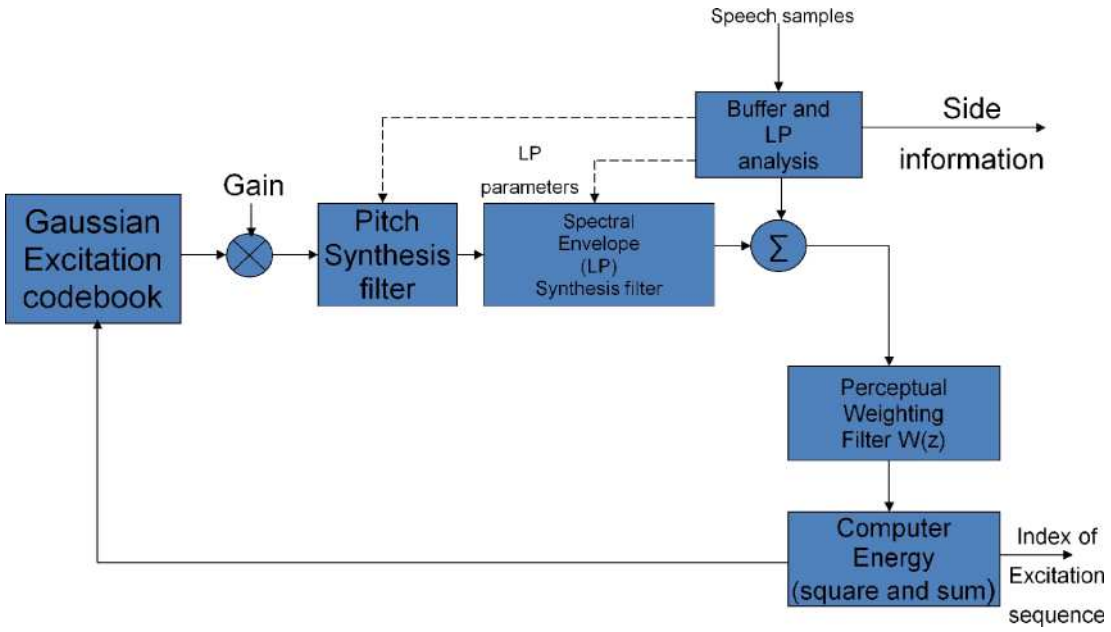
2– وکودر CELP

همان طور که از اسم این وکودر مشخص است از یک مجموعه کدبوک برای ایجاد سیگنال تحریک استفاده می کند.

CELP یک روش آنالیز–با–سنتز می باشد. به این صورت که دنباله تحریک از یک کدبوکی از دنباله های گووسی با میانگین صفر انتخاب می شود.

نرخ ارسال بیت CELP برابر 4800 بیت بر ثانیه می باشد.

در تصویر 1 کدکننده CELP را مشاهده می کنید.



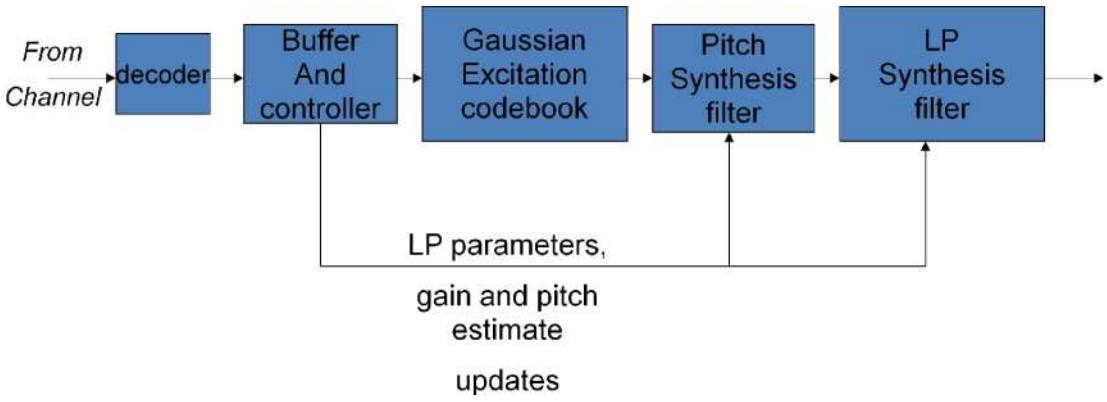
تصویر 1 – کدکننده وکودر CELP

خطای وزن داده شده به توان 2 می رسد و بر روی زیرفریم های یک بلوک جمع می شود تا انرژی خطا را بدهد.

با انجام یک جستجوری کامل درون کدبوک، دنباله تحریکی که انرژی خطا را کمینه می کند را می یابیم.

فاکتور gain برای مقیاس کردن دنباله تحریک برای هر عضو کدبوک محاسبه می شود. این کار با کمینه کردن انرژی خطا برای بلوکی از نمونه ها انجام می شود.

در تصویر 2 کدگشای CELP را مشاهده می کنید.



تصویر 2 – کدگشای وکودر CELP.

اتصال دو فیلتر تمام قطب را مشاهده می کنید. پارامترهای این دو فیلتر به صورت پریودیک آپدیت می شوند.

فیلتر اول یک فیلتر تاخیر بالای گام می باشد که برای تولید گام پریودیک در گفتار صدادار استفاده می شود.

این فیلتر به صورت فرمول 1 می باشد.

$$\theta_p(z) = \frac{\theta_p}{1 - bz^{-p}}$$

پارامتر فیلترها را می توان با کمینه کردن انرژی خطای پیش بینی پس از تخمین گام، بر روی یک فریم 5 میلی ثانیه ای انجام داد.

فیلتر دوم یک فیلتر تاخیر کوتاه تمام قطب می باشد (برای مدل کردن مسیر صوتی) و 10 تا 12 ضریب دارد که هر 10 تا 20 میلی ثانیه محاسبه می شوند.

مثال:

فرض کنید فرکانس نمونه برداری 8 کیلوهرتز باشد.

مدت زمان زیرفریم ها برای تخمین گام و دنباله تحریک 5 میلی ثانیه باشد

در هر 5 میلی ثانیه 40 نمونه داریم

دنباله تحریک شامل 40 نمونه خواهد بود.

یک کدبوک 1024 دنباله ای منجر به یک گفتار کیفیت خوب می شود.

برای این اندازه کدبوک ها نیاز به 10 بیت برای ارسال اندیس کدبوک می باشد.

در نتیجه نرخ ارسال بیت 4 برابر کاهش می یابد.

با در نظر گرفتن ارسال پارامترهای پیش بینی کننده گام و طیف، نرخ ارسال بیت 4800 بیت بر ثانیه می شود.

3– **وکودر** low-delay CELP

از CELP برای رسیدن به گفتار با کیفیت بالا با نرخ ارسال در حدود 16000 بیت بر ثانیه استفاده شده است.

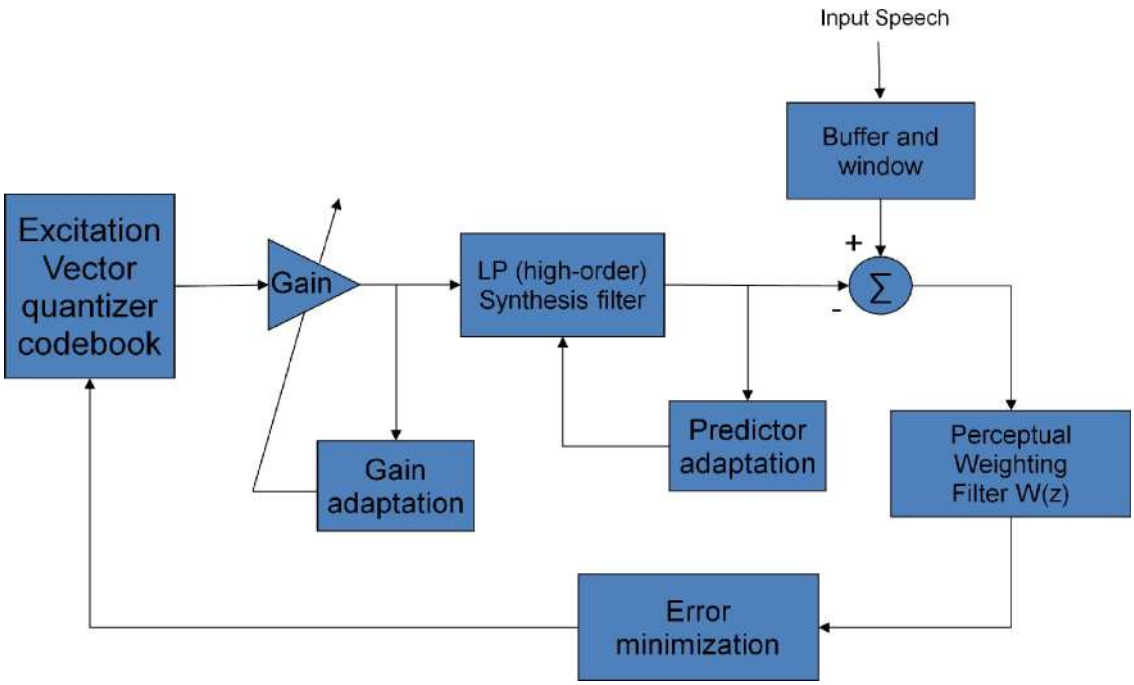
اگرچه وکودرهای نوع دیگر نیز می توانند با نرخ 16000 بیت بر ثانیه به کیفیت خیلی بالا برسند ولی این وکودر ها 10–20 میلی ثانیه از گفتار را بافر می کنند.

کل تاخیر بین 20 تا 40 میلی ثانیه می باشد.

با تغییراتی در CELP می توان کل تاخیر را تا 2 میلی ثانیه کاهش داد.

CELP تاخیر کوتاه به این صورت تاخیر را کاهش می دهد که از یک پیش بینی کننده رو به عقب استفاده می کند و پارامتر gain و دنباله تحریک را هر 5 نمونه محاسبه می کند.

در تصویر 1 کدکننده low-delay CELP را مشاهده می کنید.



تصویر 1 – کدکننده low-delay CELP

تخمین زننده گام حذف شده است.

برای خنثی کردن اثر این حذف، تعداد ضرایب پیش بینی کننده LPC افزایش می یابد (به حدود 50 ضریب).

ضرایب LPC بیشتر از بقیه ضرایب به روز می شوند (هر 2.5 میلی ثانیه)

بردارهای تحریک 5 نمونه ای مانند یک بلوک تحریک 0.625 میلی ثانیه می باشند (در نرخ نمونه برداری 8 کیلوهرتز) .

5 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را ادامه دادیم

وکودر CELP را بیان کردیم.

وکودر low-delay CELP را نیز توضیح دادیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

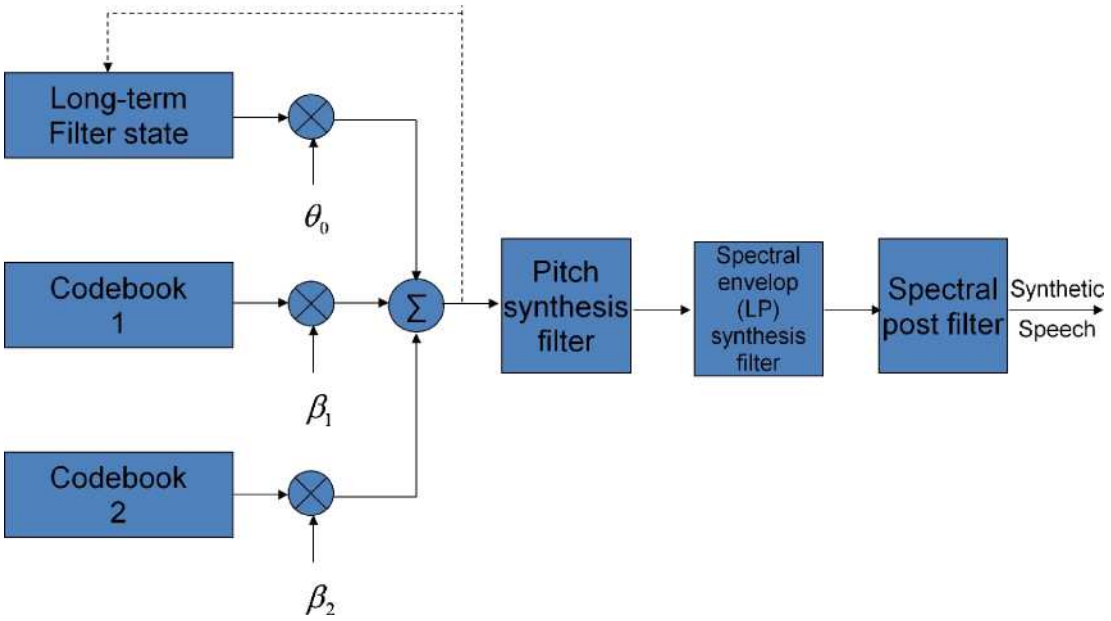
آشنایی با وکودر ها

وکدر Vector Sum Excited LP

2– وکودر VSELP

کدکننده VSELP و کدگشای آن در نحوه ایجاد دنباله تحریک تفاوت دارند.

در بلوک دیاگرام تصویر 1 مشاهده می کنید که VSELP سه منبع تحریک دارد.



تصویر 1 – کدکننده وکود VSELP

یکی از سه تحریک از پریود گام به دست می آید.

دو منبع تحریک دیگر از دو کدبوک به دست می آیند.

فیلتر سنتز LPC بوسیله یک فیلتر تمام قطب با 10 ضریب پیاده سازی می شود و ضرایب آن هر 20 میلی ثانیه کد شده و ارسال می شوند.

ضرایب هر 5 میلی ثانیه به روزرسانی می شوند

پارامترهای تحریک نیز هر 5 میلی ثانیه به روز می شوند

128 عضو در هر کدبوک وجود دارد.

این اعضا از دو مجموعه از 7 عضو پایه (به صورت ترکیب خطی) تشکیل شده اند.

فیلتر تاخیر طولانی نیز یک کدبوک با اعضای 128 تایی می باشد.

در هر فریم 5 میلی ثانیه ای، اعضای این کدبوک از فیلتر گفتار عبور دادن می شوند.

اعضای فیلتر شده استفاده می شوند تا تاریخچه انتقال ها را حفظ کنند.

این به روزرسانی به این صورت انجام می شود که بهترین عضو فیلتر شده به کدبوک تاریخچه اضافه می شود و قدیمی ترین نمونه از کدبوک حذف می شود.

نتیجتاً به یک کدبوک تطبیقی می رسیم.

سه دنباله تحریک به صورت ترتیبی از هر سه کدبوک انتخاب می شوند.

هر جستجوی کدبوک سعی می کند که عضوی را بیابد که که انرژی کل خطای وزن دهی شده را کمینه می کند.

هنگامی که اعضا انتخاب شدند، سه پارامتر gain بهینه می شوند.

بهینه سازی همزمان در صورتی به بهینه سازی ترتیبی تبدیل می شود که همه اضای کدبوک وزن دار شده با هم متعامد شده باشند (قبل از شروع جستجو).

نرخ ارسال بیت VSELP 8000 بیت بر ثانیه می باشد (تصویر 2).

Parameters	Bits/5-ms Frame	Bits/20ms
10 LPC coefficients	-	38
Average speech energy	-	5
Excitation codewords from two VSELP codebooks	14	56
Gain parameters	8	32
Lag of pitch filter	7	28
Total	29	159

تصویر 2 – نرخ ارسال بیت VSELP

4 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را ادامه دادیم

وکودر VSELP را بیان کردیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

آشنایی با وکودر ها

وکدر Mixed Excitation LP

2– وکودر MELP

وکودرهای معمولی LPC که دارای تحریک بر اساس گام می باشند، یا از یک قطار ضربه پریودیک و یا از یک نویز سفید به عنوان تحریک استفاده می کنند.

با این کار با نرخ ارسال بیت پایین به گفتار قابل درک می رسیم.

ولی با اینکه گفتار حاصله قابل درک توسط انسان می باشد، ولی از کیفیت پایینی برخوردار است.

به خصوص اینکه گفتار در برخی موارد حالت وزوز داشتن به خود می گیرد.

این مشکلات به این دلیل رخ می دهند که:

- قطار ضربه ساده قادر به بازتولید همه نوع گفتارهای صدادار نمی باشد.

وکودرهای MELP از یک مدل تحریک مخلوط استفاده می کنند و نتیجتاً قادر به بازنمایی بازه بزرگتری از گفتار هستند.

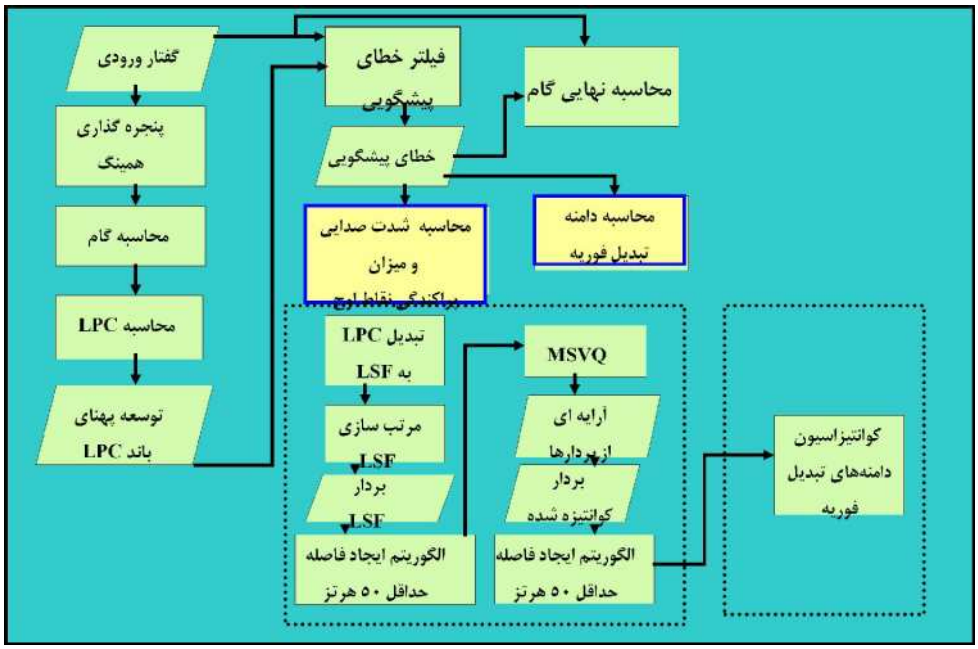
با این کار به گفتار طبیعی تری می رسیم.

این روش در محیط های با نویز پس زمینه مقاوم است.

اساس آن مبتنی برمدل LPC می باشد. یک سری ویژگی ها برای تحریک مخلوط به مدل اضافه شده اند:

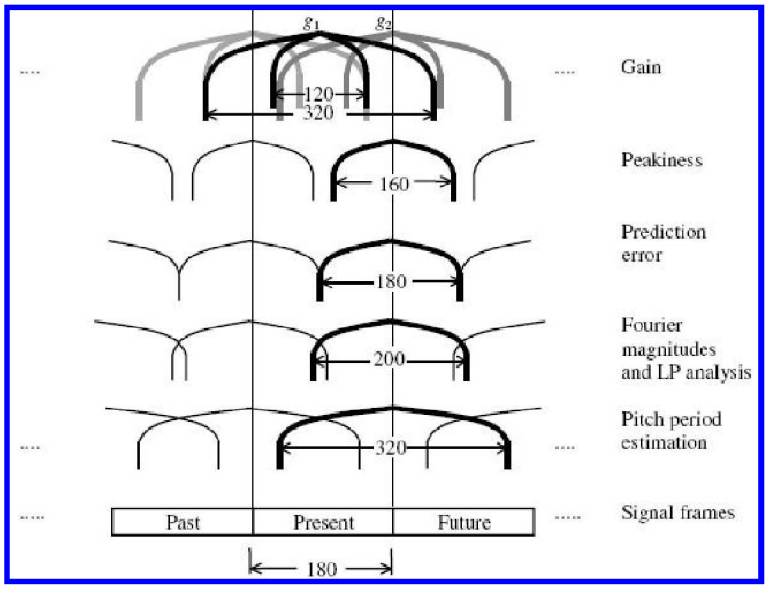
- تحریک مخلوط
- پالس های غیرپریودیک
- بهبود طیف پالس به صورت تطبیقی

در تصویر 1 کدکننده MELP را مشاهده می کنید.



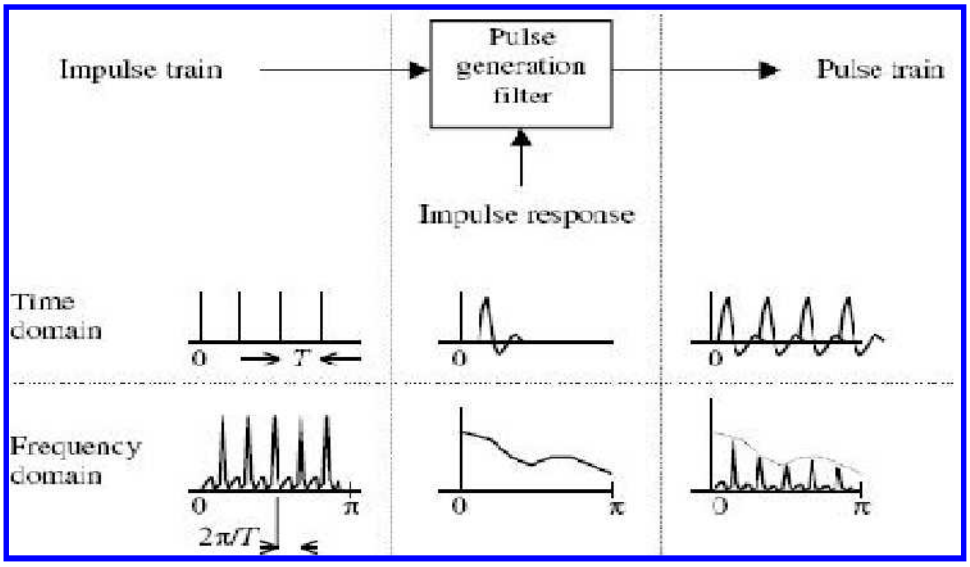
تصویر 1 – کدکننده MELP

در تصویر 2 موقعیت پنجره های آنالیز را مشاهده می کنید.



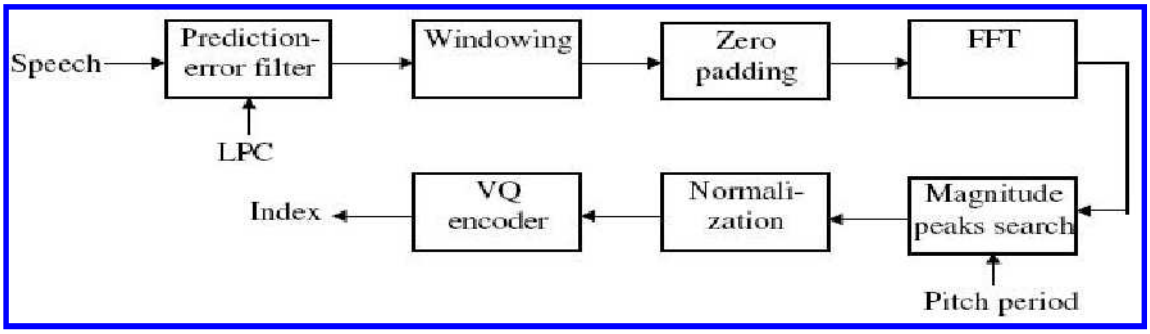
تصویر 2 – پنجره های آنالیز

در تصویر 3 نحوه محاسبه دامنه های تبدیل فوریه را مشاهده می کنید.



تصویر 3 – محاسبه دامنه های تبدیل فوریه

در تصویر 4 بلوک دیاگرام مراحل تصویر 3 را مشاهده می کنید.



تصویر 4 – بلوک دیاگرام مراحل تصویر 3

در تصویر 5 جدول تخصیص بیت را مشاهده می کنید.

حالت بی‌صدا	حالت صدادار	پارامتر
25	25	**ضرایب LSF**
-	8	**دامنه‌های تبدیل فوریه**
8	8	**بهره (2بار به ازای هر فریم)**
7	7	**دوره گام + VS1**
-	4	**شدت‌های صدایی**
-	1	**پرچم غیر پریودیک**
13	-	**محافظت از خطا**
1	1	**بیت سنکرونیزاسیون**
54	**54**	**کل بیت‌های اختصاصی**

تصویر 5 – جدول تخصیص بیت MELP

تحریک مخلوط به روش کدل ترکیب چند باندی پیاده سازی می شود.

این مدل قادر است صدادار بودن را در هر فرکانس به صورت مستقل محاسبه کند.

در نهایت ترکیبی از پالس های پریودیک/غیرپریودیک و همچنین نویز سفید را به عنوان تحریک نهایی استفاده می کند.

وقتی گفتار صدادار است، MELP با استفاده از پالس های پریودیک یا غیرپریودیک تحریک را مدل می کند.

4 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را ادامه دادیم

وکودر MELP را بیان کردیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

پایان بحث وکودرها

وکدر Multi-Band Excitation LP

2– وکودر MBE

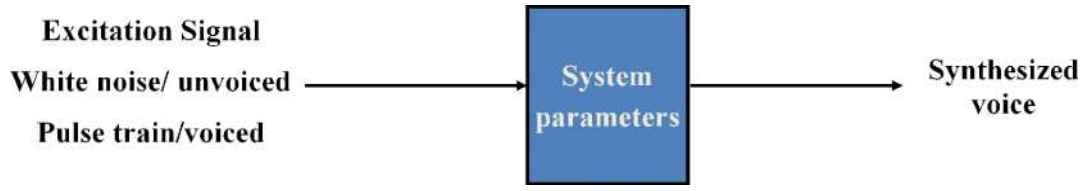
همان طور که می دانیم در وکودرها، گفتار قبل از هر پردازشی به پنجره هایی تقسیم می شود.

سپس پارامترهای تحریک و پارامترهای سیستم برای آن قسمت ها محاسبه می شوند.

- پارامترهای تحریک شامل: شناسایی صدادار بودن/نبودن و پریود گام
- پارامترهای سیستم شامل: پوش طیفی یا همان پاسخ ضربه سیستم

سپس این اطلاعات به دست آ»ده ارسال می شوند.

در تصویر 1 مشاهده می کنید که در سمت گیرنده، سیگنال تحریک از پارامترها ساخته می شود و سپس با عبور آن از فیلتر، گفتار سنتز شده حاصل می شود.



تصویر 1 – فرآیند تولید گفتار سنتز شده

معمولاً این وکودرهای ساده کیفیت پایینی دارند:

- مدل ها گفتار محدودیت های اساسی دارد.
- تخمین پارامترها ممکن است دقیق نباشد.
- عدم قادر بودن قطار ضربه /نویز سفید برای تولید همه صداها: گفتار سنتز شده با پالس پریودیک تقریباً حالت وزوز دارد و تحریک کاملاً نویز کیفیت بدی دارد.

برای رفع این خاصیت وزوز داشتن صدا استفاده از تحریک هایی است که مخلوطی از پالس پریودیک و نویز هستند.

در این وکودرها پالس های پریودیک و نویز ها با یک نسبت خاصی با هم ترکیب می شوند و این نسبت است که به سمت گیرنده ارسال می شود تا سیگنال تحریک ساخته شود.

به دلیل ایستا بودن سیگنال گفتار در یک فریم، یک پنجره به سیگنال اعمال می شود $s_w(n) = w(n)s(n)$

تبدیل فوریه قطعه پنجره شده $s_w(\omega)$ را می توان به صورت ضرب پوش طیف $H_w(\omega)$ و طیف تحریک $|E_w(\omega)|$ در نظر گرفت (فرمول 1).

فرمول 1 $\hat{s}_w(\omega) = H_w(\omega)\,|E_w(\omega)|$

در اغلب مدل ها $H_w(\omega)$ نسخه صاف شده از طیف اصلی $s_w(\omega)$ می باشد.

پوش طیف باید به صورت دقیق بازنمایی شود.

می توان با اضافه کردن تصمیم های صدادار بودن/نبودن به صورت وابسته به فرکانس کیفیت را افزایش داد.

در مدل های ساده پیشین، طیف تحریک تماماً بوسیله فرکانس گام و تصمیم صدادار بدون برای کل طیف ساخته می شود.

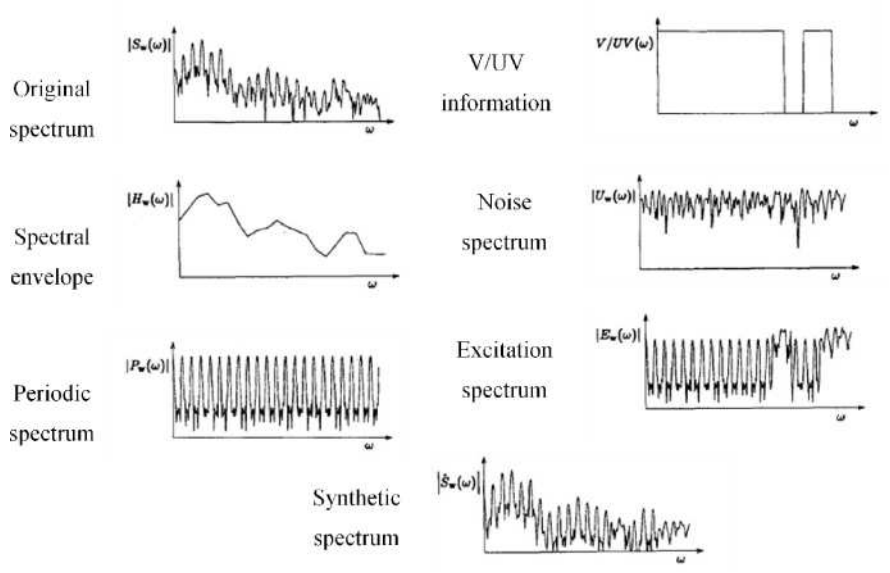
«در مدل MBE طیف تحریک بوسیله فرکانس گام و تصمیم های صدادار‌بودن وابسته به فرکانس ساخته می شود.»

در کل برای یک طیف پیوسته به تعداد خیلی زیادی تصمیم صدادار بودن نیاز داریم.

برای کاهش تعداد بیت های مورد نیازطیف را به چندین باند فرکانسی تقسیم می کنیم و برای هر باند یک پارامتر باینری صدادار بودن در نظر گرفته می شود.

تفاوت MBE با مدل های دیگر این است که تعداد باند ها معمولاً زیاد و در حدود 20 در نظر گرفته می شود.

در تصویر 2 خلاصه این عمل را مشاهده می کنید.



تصویر 2 – خلاصه ایجاد سیگنال تحریک در وکودر MBE

پارامترهای MBE عبارتند از:

- پوش طیف
- فرکانس گام
- اطلاعات صدادار بودن هر باند
- برای فریم های صدادار، فاز آن باند

پارامترهای پوش طیف بوسیله ضرایب پیشگویی خطی محاسبه می شوند.

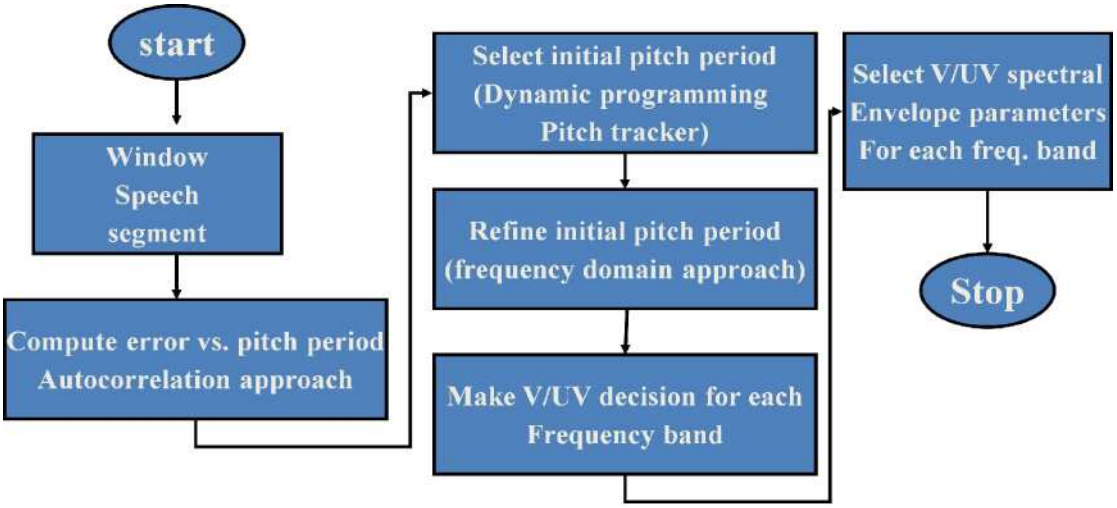
در مدل های ساده معمولاً پارامترهای پوش طیف و پارامترهای تحریک به صورت کاملاً مستقل محاسبه می شوند.

ولی در وکودر MBE این دو به صورت همزمان محاسبه می شوند به این صورت که سعی می کنند سیگنال سنتز شده از لحاظ میانگین مربعات خطا کمترین فاصله را با سیگنال اصلی داشته باشد.

کل مراحل تخمین به دو گام اصلی تقسیم شده اند:

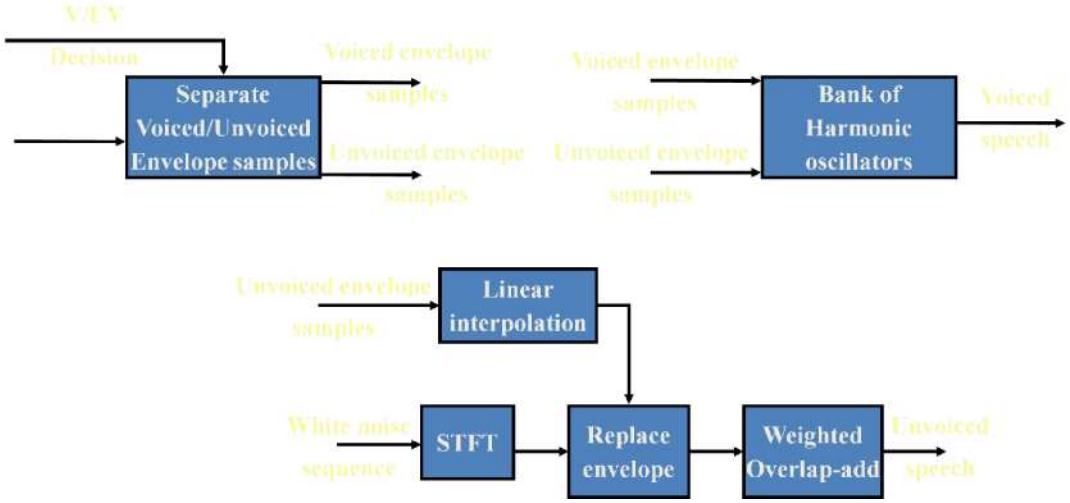
1. در گام اول پریود گام و پارامترهای پوش طیف تخمین زده می شوند به طوری که خطای بین طیف اصلی و طیف سنتز شده کمینه شود.
2. سپس تصمیم گیری های صداداربودن باند ها گرفته می شود.

در تصویر 3 بلوک دیاگرام این وکودر آمده است.



ته

د



تصویر 4 – بلوک دیاگرام کدگشای MBE

در تصویر 5 جدول تخصیص بیت MBE را مشاهده می کنید.

Parameter	Bits
Fundamental Frequency	9

Magnitude Harmonic	139-94
Harmonic Phase	0-45
Voiced/Unvoiced Bits	12
Total	**160**

تصویر 5 – جدول تخصیص بیت MBE

4 – خلاصه و نتیجه گیری

در این فصل بحث وکودر ها را به پایان رساندیم.

وکودر MBE را بیان کردیم.

6 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

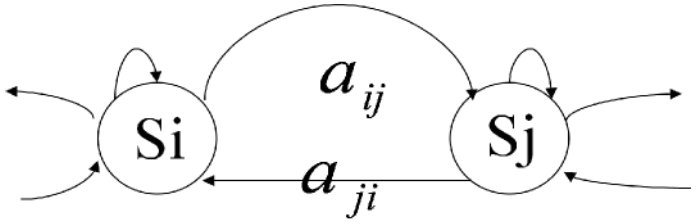
اهداف درس:

در این فصل با مدل مخفی مارکوف آشنا شدیم

همچنین با دو مسئله از سه مسئله مهم HMMها آشنا شدیم و نحوه حل آن را فرا گرفتیم

2– مدل مخفی مارکوف

در تصویر 1 یک مدل مخفی مارکوف را مشاهده می کنید.



مشاهدات: O1، O2، ....

$$O_1, O_2, O_3, \cdots, O_t$$

حالات در زمان: q1، q2، ...

$$q_1, q_2, q_3, \cdots, q_t$$

همه حالات: s1، s2، ...

احتمال های مدل مخفی مارکوف به حالات قبلش بستگی ندارد.

بلکه فقط به آخرین حالت بستگی دارد.

یعنی احتمال شرطی روبرو

$$P(q_t = s_j \mid q_{t-1} = s_i, q_{t-2} = s_k, \cdots, q_1 = s_z)$$

خلاصه می شود به:

$$= P(q_t = s_j \mid q_{t-1} = s_i)$$

به مدل مارکوفی که خاصیت بالا را دارد مارکوف درجه 1 گفته می شود (چون فقط به 1 حالت قبلش بستگی دارد).

$$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i) \qquad 1 \le i, j \le N$$

- a_{ij}: احتمال گذار از حالت Si به Sj ، به عبارتی

مثال: یک نمونه مدل مخفی مارکوف

حالات زیر وجود دارد:

S1 : The weather is rainy

S2 : The weather is cloudy

S3 : The weather is sunny

در زیر احتمال گذرهای بین حالات را مشاهده می کنید:

$$A = \{a_{ij}\} = \begin{matrix} \\ \text{rainy} \\ \text{cloudy} \\ \text{sunny} \end{matrix} \begin{matrix} \begin{matrix} \text{rainy} & \text{cloudy} & \text{sunny} \end{matrix} \\ \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \end{matrix}$$

سئوال 1: احتمال مشاهدات زیر چقدر است؟

؟ oudy

$$q_1 q_2 q_3 q_4 q_5 q_6 q_7 q_8$$
$$\underbrace{s_3 s_3 s_3 s_1 s_1 s_3 s_2 s_2}_{a_{33} a_{33} a_{31} a_{11} a_{13} a_{32} a_{22}} = 1.536 \times 10^{-4}$$

سئوال 2: احتمال ماندن در یک حالت برای d روز اگر در حالت

$$P(\underbrace{s_i s_i \cdots s_i}_{\text{d Days}} s_{j \neq i}) = a_{ii}^{d-1}(1 - a_{ii}) = P_i(d)$$

اجزای یک HMM

یک HMM دارای اجزای زیر می باشد:

- N: تعداد حالات
- M: تعداد خروجی ها
- A: ماتریس احتمال گذر حالت
- B: ماتریس احتمال رخداد خروجی
- π: احتمال رخداد اولیه

مجموعه پارامترهای یک HMM را به صورت روبرو نمایش می دهند: $\lambda = (A, B, \pi)$

سه مسئله اساسی HMM

1. فرض کنید که یک HMM با پارمترهای λ و دنباله ای از مشاهدات O داریم، احتمال $P(O \mid \lambda)$ چقدر است؟
2. فرض کنید یک مدل λ و یک دنباله مشاهدات O داریم، محتمل ترین دنباله حالات مدل که آن مشاهدات را تولید کرده اند کدام است؟
3. فرض کنید یک مدل λ و یک دنباله مشاهدات O داریم، چگونه می توان پارمترهای مدل را تنظیم کرد که $P(O \mid \lambda)$ بیشینه شود (به عبارتی آموزش مدل از روی مشاهدات)؟

3– مسئله اول

- مسئله اول این بود:

فرض کنید که یک HMM با پارمترهای λ و دنباله ای از مشاهدات O داریم، احتمال $P(O \mid \lambda)$ چقدر است؟

- راه حل اول مسئله اول:

یک راه حل برای یافتن احتمال $P(O \mid \lambda)$ این است که همه حالت های ممکن را ردیف کرده و احتمال آن ها را با هم جمع کنیم:

$$P(o \mid \lambda) = \sum_{q} P(o, q \mid \lambda)$$

در واقع هدف یافتن $P(o, q \mid \lambda)$ می باشد. این احتمال را به صورت زیر به دست می آوریم:

$$P(o \mid q, \lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda) = \prod_{t=1}^{T} b_{q_t}(o_t)$$

$$P(q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

می دانیم که: و

$$\Rightarrow P(o,q \mid \lambda) = P(o \mid q,\lambda)P(q \mid \lambda)$$

نتیجه می گیریم که:

$$\Rightarrow P(o,q \mid \lambda) = \Pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

که به دست می آید:

$$\Rightarrow P(o \mid \lambda) = \sum_{q} P(o,q \mid \lambda) =$$

$$\sum_{q_1 q_2 \cdots q_T} \Pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

که فرمول نهایی به صورت زیر می باشد:

مرتبه زمانی الگوریتم بالا $O(2TN^T)$ می باشد. این مقدار بسیار بالا می باشد.

این به این خاصط است که سعی می شود در حین محاسبه همه دنباله حالت ها در نظر گرفته شوند.

در ادامه راه حل هایی برای کاهش مرتبه زمانی ارائه خواهیم کرد.

- راه حل دوم مسئله اول:

نام روش روش Forward است.

این روش نوعی الگوریتم پویاست که در یک متغییر مقادیر میانی ذخیره می شوند برای استفاده آینده.

در این روش متغییری به نام α وجود دارد. $\alpha_t(i) = P(o_1, o_2, \cdots, o_t, q_t = i \mid \lambda)$

مقدار این متغییر احتمال حضور در حالت qt است با دیدن مشاهدات o1,....ot .

کل روند الگوریتم به صورت زیر است:

- مقداردهی اولیه: $\alpha_1(i) = b_i(o_1)\Pi_i, 1 \le i \le N$ احتمال شروع از هر حالت و تولید مشاهده اول.
- حلقه:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1})$$

$$1 \le t \le T-1, 1 \le j \le N$$

احتمال حضور در حالت بعدی با توجه به مشاهدات انجام شده به صورت بالا محاسبه می شود.

یعنی فرض کنید می خواهیم α حالت j را محاسبه کنیم (عامل $[\sum_{i=1}^{N} \alpha_t(i) a_{ij}]$)

باید از همه حالت ها بپریم به j و سپس مشاهده مربوطه را تولید کنیم (عامل $b_j(o_{t+1})$)

- پایان: احتمال مورد نظر با استفاده از جمع α های زمان T محاسبه می شود:

$$P(o \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

مرتبه زمانی این الگوریتم $O(N^2 T)$ می باشد.

- راه حل سوم مسئله اول:

این روش شبیه روش Forward می باشد.

با این تفاوت که از عقب به جلو عمل می شود. به همین خاطر نام متغییر Backward است.

در این روش متغییری به نام β داریم: $\beta_t(i) = P(o_{t+1}, o_{t+2}, \cdots, o_T \mid q_t = i, \lambda)$

کل روند الگوریتم به صورت زیر است:

- مقداردهی اولیه: $\beta_T(i) = 1, 1 \le i \le N$
- حلقه:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

$$t = T-1, T-2, \cdots, 1 And 1 \le j \le N$$

احتمال حضور در حالت قبلی با توجه به مشاهدات انجام شده به صورت بالا محاسبه می شود.

- پایان: احتمال مورد نظر با استفاده از جمع β های زمان 1 محاسبه می شود:

$$P(o \mid \lambda) = \sum_{i=1}^{N} \beta_1(i)$$

مرتبه زمانی این الگوریتم $O(N^2 T)$ می باشد.

4– مسئله دوم

- تعریف مسئله دوم:

فرض کنید یک مدل λ و یک دنباله مشاهدات O داریم، محتمل ترین دنباله حالات مدل که آن مشاهدات را تولید کرده اند کدام است؟

- راه حل اول مسئله دوم:

هدف یافتن محتمل ترین دنباله حالات می باشد.

فرض کنید یک متغییر γ به صورت روبرو تعریف شود: $\gamma_t(i) = P(q_t = i \mid o, \lambda)$

یعنی احتمال وجود در حالت i در زمان t با دیدن مشاهدات. یعنی در هر زمان احتمال وجود در یک حالت را داریم.

مقدار γ به صورت زیر محاسبه می شود:

$$\gamma_t(i) = P(q_t = i \mid o, \lambda) = \frac{P(o, q_t = i \mid \lambda)}{P(o \mid \lambda)}$$

$$= \frac{P(o, q_t = i \mid \lambda)}{\sum_{i=1}^{N} P(o, q_t = i \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

که α و β همان متغییرهای Forward و Backward هستند.

محتمل ترین حالت به این صورت به دست می آید: $q_t^* = \arg\max_i [\gamma_t(i)], 1 \le t \le T, 1 \le t = n \le N$

یعنی در هر زمان ماکزیمم گیری می شود که کدام حالت بیشترین احتمال رخداد را دارد.

6 – خلاصه و نتیجه گیری

در این فصل با مدل مخفی مارکوف آشنا شدیم

همچنین با دو مسئله از سه مسئله مهم HMMها آشنا شدیم و نحوه حل آن را فرا گرفتیم

7 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

- راه حل دوم مسئله دوم:

نام این الگوریتم ویتربی (Viterbi) می باشد.

تعریف زیر را در نظر بگیرید:

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} P[q_1,q_2,\cdots,q_{t-1},q_t=i,o_1,o_2,\cdots,o_t \mid \lambda]$$
$$1 \le i \le N$$

P محتمل ترین دنباله حالات با این شرط است که در زمان t در حالت i باشیم و مشاهدات o1,...ot را دیده باشیم.

از فرمول δ مشخص است که یک ماکزیمم گیری رو دنباله های پیشین انجام می شود.

پس فرمول بازگشتی زیر را داریم:

$$\delta_{t+1}(j) = [\max_i \delta_t(i)a_{ij}].b_j(o_{t+1})$$

کل روند الگوریتم ویتربی به صورت زیر است:

- مقداردهی اولیه:

$$\delta_1(i) = \Pi_i b_i(o_1), 1 \le i \le N$$
$$\psi_1(i) = 0$$

از Ψ برای ذخیره مسیر استفاده می شود.

$\psi_t(i)$ محتمل ترین حالت قبل از حالت i در زمان t می باشد.

- حلقه:

فرمول های بازگشتی زیر مهم ترین قسمت الگوریتم هستند:

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t)$$
$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]$$
$$2 \le t \le T, 1 \le j \le N$$

- پایان:

$$p^* = \max_{1\le i\le N}[\delta_T(i)]$$
$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)]$$

- بازگشت به عقب برای پیمایش معکوس مسیر:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \cdots, 1$$

1– مقدمه

اهداف درس:

در این فصل با آموزش مدل مخفی مارکوف آشنا می شویم

آموزش هم برای حالت گسسته و هم پیوسته بررسی خواهد شد

2– مسئله سوم برای HMMهای گسسته

تعریف مسئله سوم:

فرض کنید یک مدل λ و یک دنباله مشاهدات O داریم، چگونه می توان پارمترهای مدل را تنظیم کرد که $P(O|\lambda)$ بیشینه شود (به عبارتی آموزش مدل از روی مشاهدات)؟

تخمین پارامترها بوسیله الگوریتم بیشینه سازی Expectation انجام می شود.

قبل از شروع بحث تعریف زیر را در نظر بگیرید:

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid o, \lambda)$$
$$= \frac{P(o, q_t = i, q_{t+1} = j \mid \lambda)}{P(o \mid \lambda)}$$
$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}$$

این مقدار احتمال پرش از حالت i به حالت j در زمان t را نشان می دهد.

نتیجه می شود که $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$ یرا اگر همه j ها را در نظر بگیریم به احتمال وجود در حالت i در زمان t می رسیم که همان γ است.

- $\sum_{t=1}^{T-1} \gamma_t(i)$: مقدار expected (از لحاظ آماری) تعداد پرش ها از حالت i
- $\sum_{t=1}^{T} \xi_t(i,j)$: مقدار expected (از لحاظ آماری) تعداد پرش ها از حالت i به حالت j

با تعریف یک مقدار expectation و بیشینه کردن آن به فرمول ها زیر می رسیم (برای دیدن اثبات فرمول های زیر به کتاب درسی مراجعه کنید):

$$\bar{\Pi}_i = \gamma_1(i) \qquad \bar{a}_{ij} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \qquad \bar{b}_j(k) = \frac{\sum_{t=1,\, o_t = V_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

فرمول های بالا برای آپدیت پارامترها (a و b و π) در هر تکرار از آموزش به کار می روند.

$$Q(\lambda' \mid \lambda) = \sum_{q} P(o, q \mid \lambda') \log P(o, q \mid \lambda)$$

تابع کمکی Baum به صورت زیر تعریف می شود:

همچنین $if: Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow P(o \mid \lambda') \geq P(o \mid \lambda)$

با این اوصاف می توان اثبات کرد که فرمول های آپدیت پارمتر به دست آمده همیشه در جهت بهبود مدل روی داده های آموزشی عمل می کنند و این تغییرات باعث بدتر شدن مدل روی داده های آموزشی نمی شود.

یعنی با هر تکرار آموزش مدل بهبود می یابد (اثبات در کتاب درسی).

فرمول های آپدیت پارامتر محدودیت های زیر را دارند:

$$\sum_{i=1}^{N} \overline{\Pi}_i = 1$$

$$\sum_{j=1}^{N} \overline{a}_{ij} = 1, 1 \leq i \leq N$$

$$\sum_{k=1}^{M} \overline{b}_j(k) = 1, 1 \leq j \leq N$$

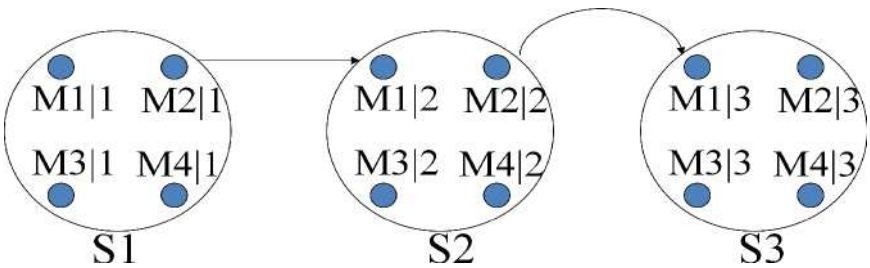
3– مسئله سوم برای HMMهای پیوسته

در صورتی که مشاهدات HMM پیوسته باشند، باید از یک تابع چگالی احتمال به عنوان خروجی مشاهدات استفاده کرد.

یعنی به جای $b_j(k) = P(o_t = V_k \mid q_t = j)$ باید از $b_j(o) = \sum_{k=1}^{M} C_{jk} \mathrm{N}(o, \mu_{jk}, \Sigma_{jk}), \int_{-\infty}^{\infty} b_j(o) do = 1$ استفاده کرد.

Mixture Coefficients — Average — Variance

مخلوط های گووسی به عنوان PDF هر حالت در نظر گرفته می شوند (تصویر 1).



تصویر 1 – مخلوط های گووسی به عنوان PDF حالات HMM

$$b_j(o) = \underset{k}{Max}\, C_{jk} \mathrm{N}(o, \mu_{jk}, \Sigma_{jk})$$

در برخی روش ها فقط از مخلوط غالب استفاده می کنند.

مدل HMM با مشاهدات پیوسته و PDF مخلوط گووسی دارای پارامترهای زیر می باشد:

$$\lambda = (A, \Pi, C, \mu, \Sigma)$$

$N\times N$ $\quad 1\times N$ $\quad N\times M$ $\quad N\times M\times K$ $\quad N\times M\times K\times K$

که N تعداد حالات، M تع

پارامترهای a و π با فرمول

تعریف: $\gamma_t(j,k)$ برابر ا

فرمول آپدیت پارمترهای

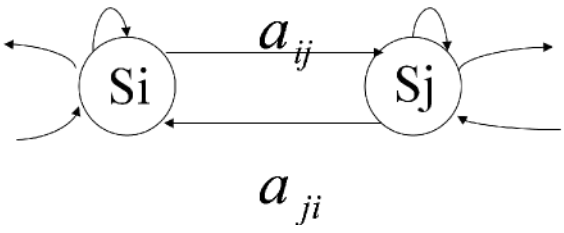
$$\bar{C}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(j,k)} \qquad \underline{\bar{\mu}}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)o_t}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)\cdot(\underline{o}_t - \underline{\mu}_{jk})\cdot(\underline{o}_t - \underline{\mu}_{jk})'}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

4– مدل کردن مدت ح

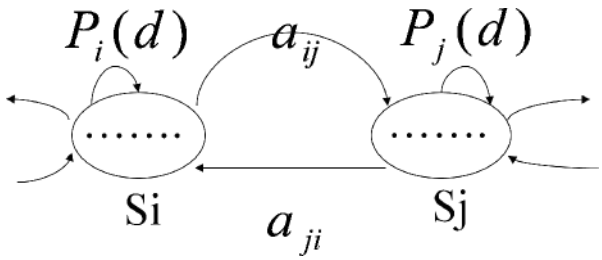
احتمال ماندن d بار در حا

$$P_i(d) = a_{ii}^{d-1}(1-a_{ii})$$



تصویر 2 – دو حالت از یک HMM

در تصویر 3 یک HMM با پارامتری با عنوان مدت زمان مشاهده می کنید.



تصویر HMM – 3 با پارامتر مدت زمان ماندن در یک حالت

با در نظر گرفتن مدت حالت یک سری موارد مطرح می شود:

- انتخاب $q_1 = i$ بوسیله π_i ها
- انتخاب d_1 بوسیله $P_{q_1}(d)$
- انتخاب دنباله مشاهدات $O_1, O_2, \cdots, O_d$

  بوسیله $b_{q_1}(O_1, O_2, \cdots, O_{d_1})$

  در عمل فرض استقلال می کنیم: $b_{q_1}(O_1, O_2, \cdots, O_{d_1}) = \prod_{t=1}^{d_1} b_{q_1}(t, O_t)$
- انتخاب حالت بعد $q_2 = j$

  بوسیله احتمالات گذر $a_{q_1 q_2}$

  یک محدودیت دیگر نیز داریم: $a_{q_1 q_1} = 0$

1– مقدمه

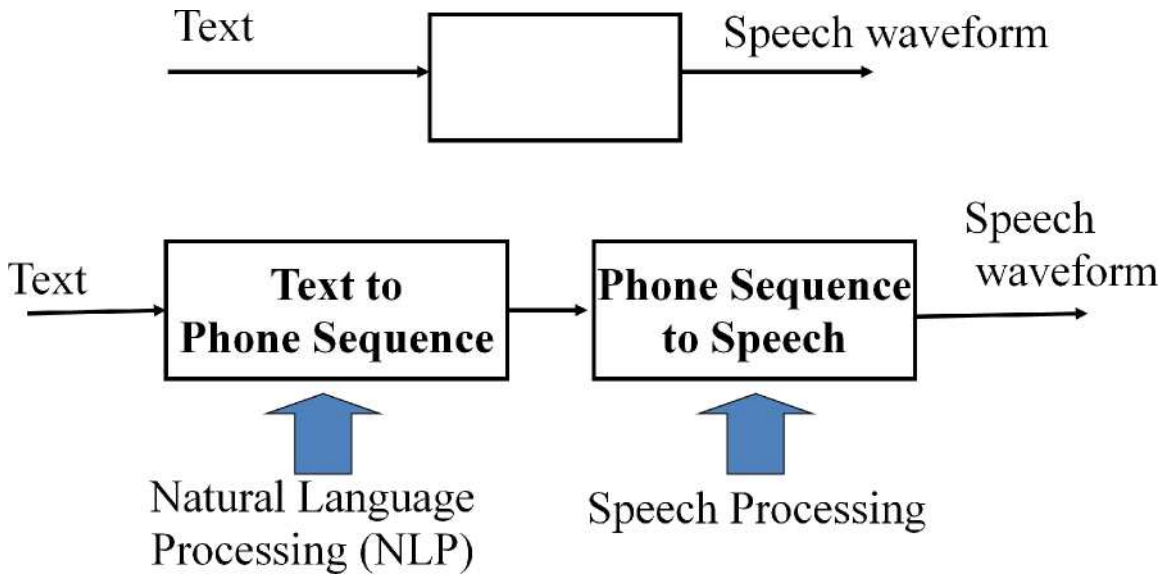
تبدیل متن به گفتار

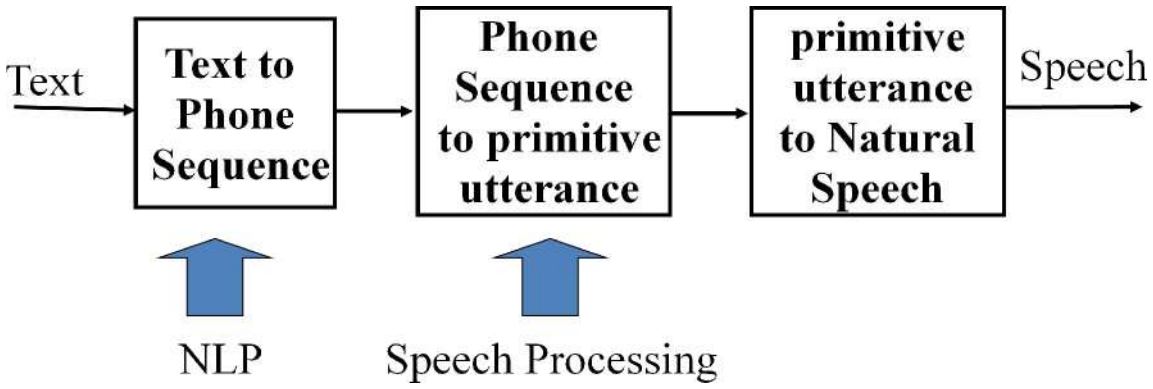
- مبتنی بر قانون
- مفصلی
- چسباندنی (concatenative)
- انتخاب واحد
- آماری (مبتنی بر مدل مخفی مارکوف)

2– مفاهیم اولیه

در این فصل بحث تبدیل دنباله آوایی به شکل موج را بررسی می کنیم (تصویر 1).

یعنی فقط بحث پردازش سیگنال تبدیل متن به گفتار را و نه بحث پردازش زبان طبیعی (NLP) بررسی می کنیم.





تصویر 1 – تبدیل دنباله آوایی به شکل موج

برای طبیعی بودن گفتار باید موارد زیر را در نظر بگیریم:

- انرژی گفتار
- مدت تلفظ واج ها
- گام
- آهنگ
- تاکید

آهتگ و تاکید در طبیعی بودن گفتار بسیار موثر هستند.

تعریف آهنگ: تغییر فرکانس گام در حین صحیت کردن

تعریف تاکید: افزایش فرکانس گام در یک زمان مشخص

آهنگ گفتار بستگی به زمینه دارد.

- معمولاً اطلاعات بیان شده در جواب یک سئوال آهنگ بیشتری دارد.
- در حالی که اطلاعاتی که از قبل می دانیم آهنگ دار نیست.

مثال:

سئوال 1: What types of foods are a good source of vitamins?

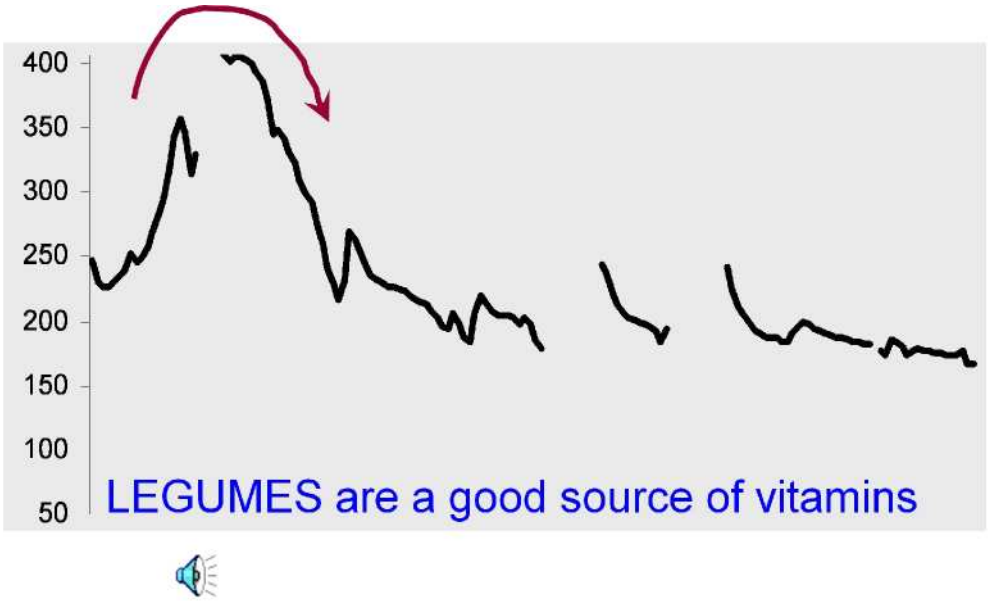
جواب 1: LEGUMES are a good source of vitamins. (تصویر 1)

سئوال 2: Are legumes a source of vitamins?

جواب 2: Legumes are a GOOD source of vitamins. (تصویر 2)

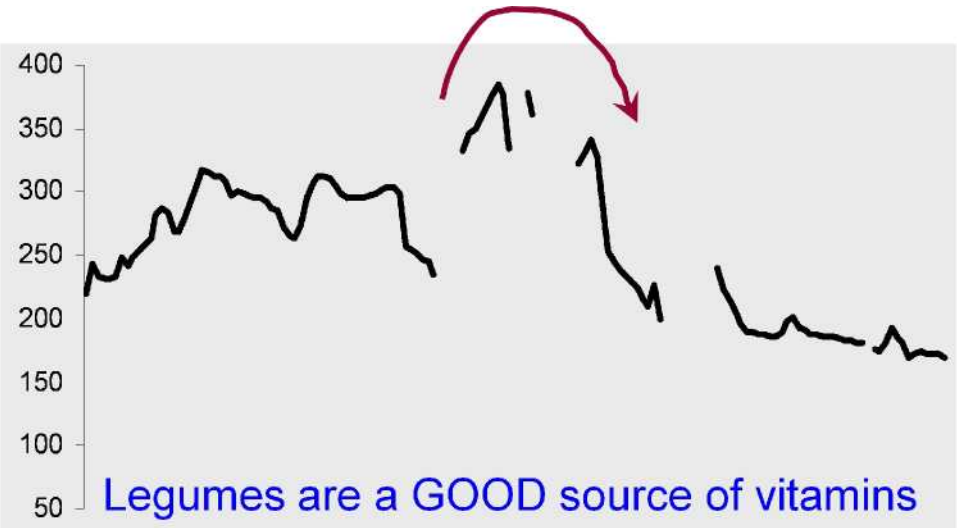
سئوال 3: What are legumes a good source of ?

جواب 3: Legumes are a good source of VITAMINS. (نصویر 3)



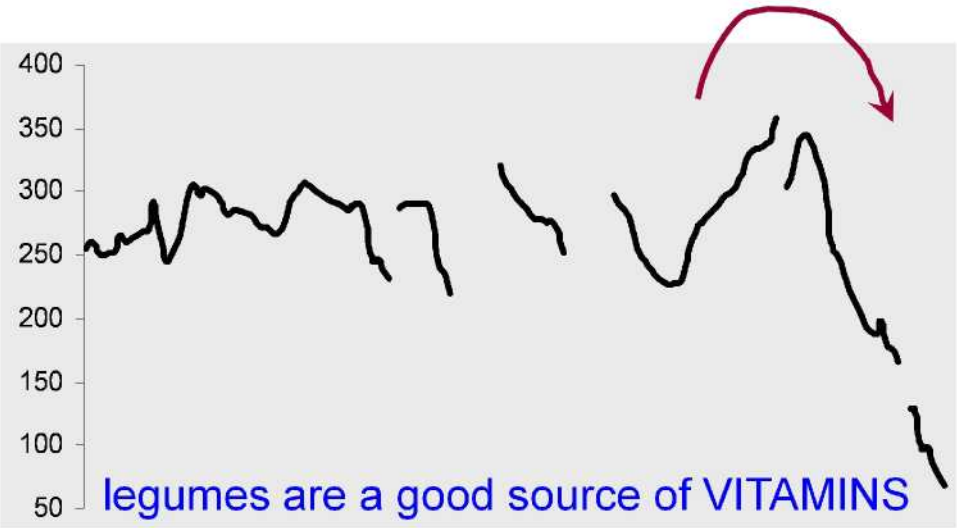
The main **rise-fall** accent (= "I assert this") shifts locations.

تصویر 1 – جواب 1



The main **rise-fall** accent (= “I assert this”) shifts locations.

نصر



The main **rise-fall** accent (= “I assert this”) shifts locations.

تصویر 3 – جواب 3

همان طور که گفتیم، در این فصل روش های تبدیل دنباله واجی (نه متن) به شکل موج را مشاهده می کنید.

روش های بررسی شده در این فصل عبارتند از:

- سنتز مفصلی (articulatory)
  - حرکت مفاصل، اعضا و ویژگی های صوتی مسیر صوتی انسان را مدل می کند.
- سنتز الحاقی (Concatenative)
  - از دیتابیسی از نمونه گفتارهای ذخیره شده برای تولید شکل موج نهایی استفاده می کند.
    - سنتز دایفون (diphone)
    - سنتز انتخاب واحد (unit selection)
- سنتز آماری (مبتنی بر مدل مخفی مارکوف)
  - پارامترهایی را از روی دیتابیس گفتار آموزش می دهد.
- مبتنی بر قانون (rule-based)
  - استفاده از قوانین و فیلترهایی برای تولید شکل موج

2– سنتز مفصلی

شبیه سازی فرآیندهای فیزیکی تولید گفتار انسان

در برخی روش ها از به هم وصل کردن لوله هایی برای ساختن ماشین های مکانیکی سخنگو استفاده شده است.

در روش های جدید تاثیر مکان اعضا، شکل مسیر صوتی و ... بر روی جریان هوا «شبیه سازی» می شود و خروجی نهایی تولید می شود.

3– سنتز الحاقی

دو روش اصلی وجود دارد:

1. الحاق واحدهای واجی (diphone concatenation)
   - مثال: اتصای نمونه های دایفون ها یا سیلاب های ضبط شده

2. انتخاب واحد آوایی (unit selection)

- استفاده از تعداد زیادی نمونه برای هر واحد آوایی و انتخاب بهترین نمونه در هنگام سنتز

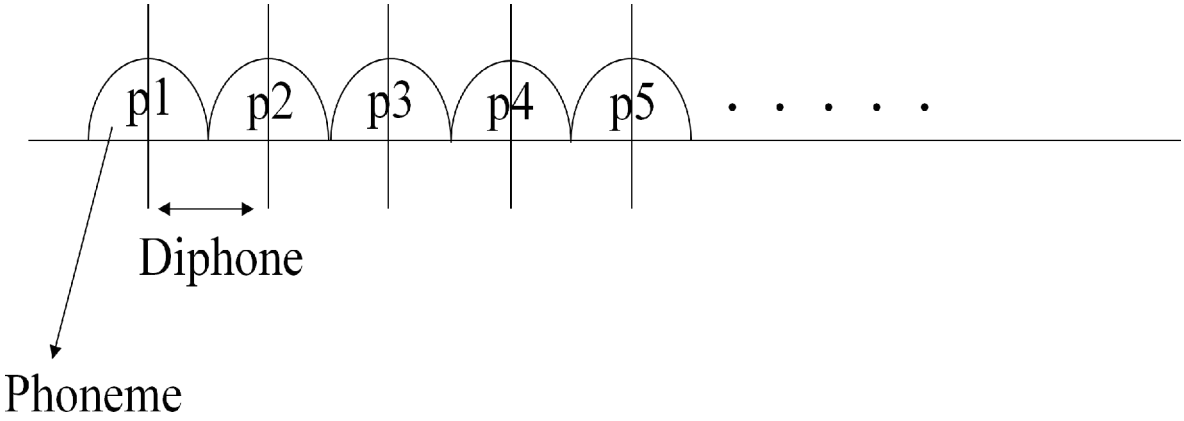
در هر دو روش باید واحد آوایی تعریف شود.

در این بخش واحد های آوایی را بررسی می کنیم:

- پاراگراف
- جمله
- کلمه (بستگی به زبان دارد و تعدا کلمات تا 100 هزار کلمه در زبان نیز می باشد)
- سیلاب
- دایفون و ترایفون
- واج (بین 10 تا 100 وابسته به زبان ها)

دایفون

گذر بین واج ها را بررسی می کنیم.



تصویر 1 – واحد آوایی دایفون

- تعداد واج های فارسی: 30 واج
- تعداد دایفون های فارسی: 30×30=900
- دایفون /zh o/ وجود ندارد
- ترایفون های فارسی: 27 هزار در تئوری
- ولی در عمل همه ترایفون ها استفاده نمی شوند (مثلاً /پ خ ک/)

سیلاب

سیلاب= صامت + ریتم

سیلاب مجموعه ای از واج ها است که دقیقاً یک واکه دارد.

سیلاب های فارسی سه نوعند: CV، CVC و CVCC

حدود 4 هزار سیلاب در فارسی داریک.

سیلاب های انگلیسی بسیار متنوعند: V, CV, CVC, CVCC, CCVCC, CCCVC, CCCVCC,...

تعداد سیلاب های زبان انگلیسی بسیار زیاد است.

همان طور که گفتیم در بحث سنتز الحاقی باید واحد آوایی تعریف شود.

در این بخش این واحد ها تعریف شدند.

همان طور که گفتیم نمونه هایی از این واحدهای آوایی ذخیره می شوند تا بعداً در موقع سنتز به هم الحاق شوند و شکل موج نهایی تولید شود.

می توان واحد آوایی مناسب را از بین یک سری واحد آوایی انتخاب کرد.

می توان به جای شکل موج اصلی از پارامترهای فشرده شده استفاده کرد.

مزایای ذخیره کردن پارمترهای فشرده عبارتند از:

- نیاز به حافظه کمتر دارد.
- می توان هر کلمه و جمله ای را تولید کرد.
- تولید پروزودی ساده تر است.

نحوه ذخیره سازی روش های مختلف سنتز را در تصویر 2 مشاهده می کنید.

Phone Unit	Type of Storing
Paragraph	Main Waveform
Sentence	Main Waveform
Word	Main Waveform
Syllable	Coded/Main Waveform
Diphone	Coded Waveform
Phoneme	Coded Waveform

تصویر 2 – نحوه ذخیره سازی واحدهای آوایی مختلف

در صورتی که به ازای هر واج در دنباله واجی شکل موجی از دیتابیس انتخاب شود، باید به روشی آن ها را به هم بچسبانیم.

روش Pitch Synchronous Overlap-Add-Method روشی مشهور در صاف کردن گذر واج ها می باشد.

روش Ovelap-Add یک روش استاندارد در بحث پردازش سیگنال دیجیتال می باشد.

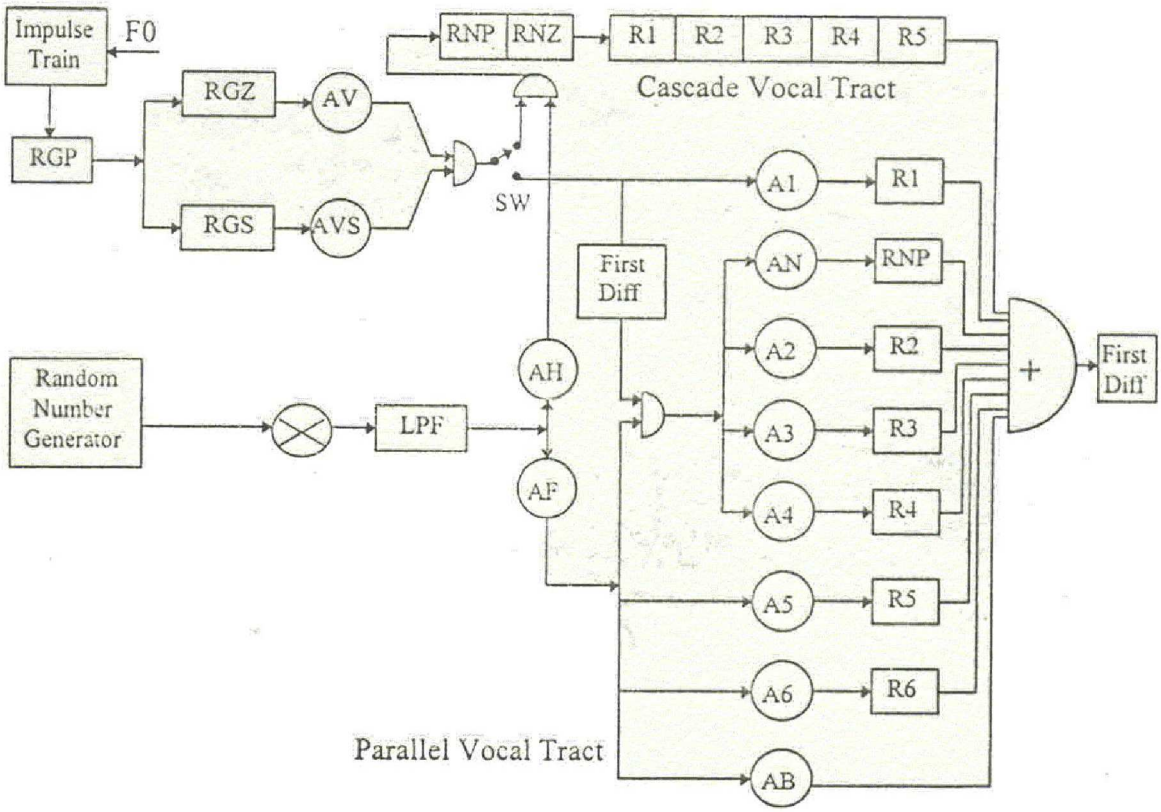
-

4– سنتز مبتنی بر قانون

کل فرآیند مبتنی بر قانون به شرح زیر است:

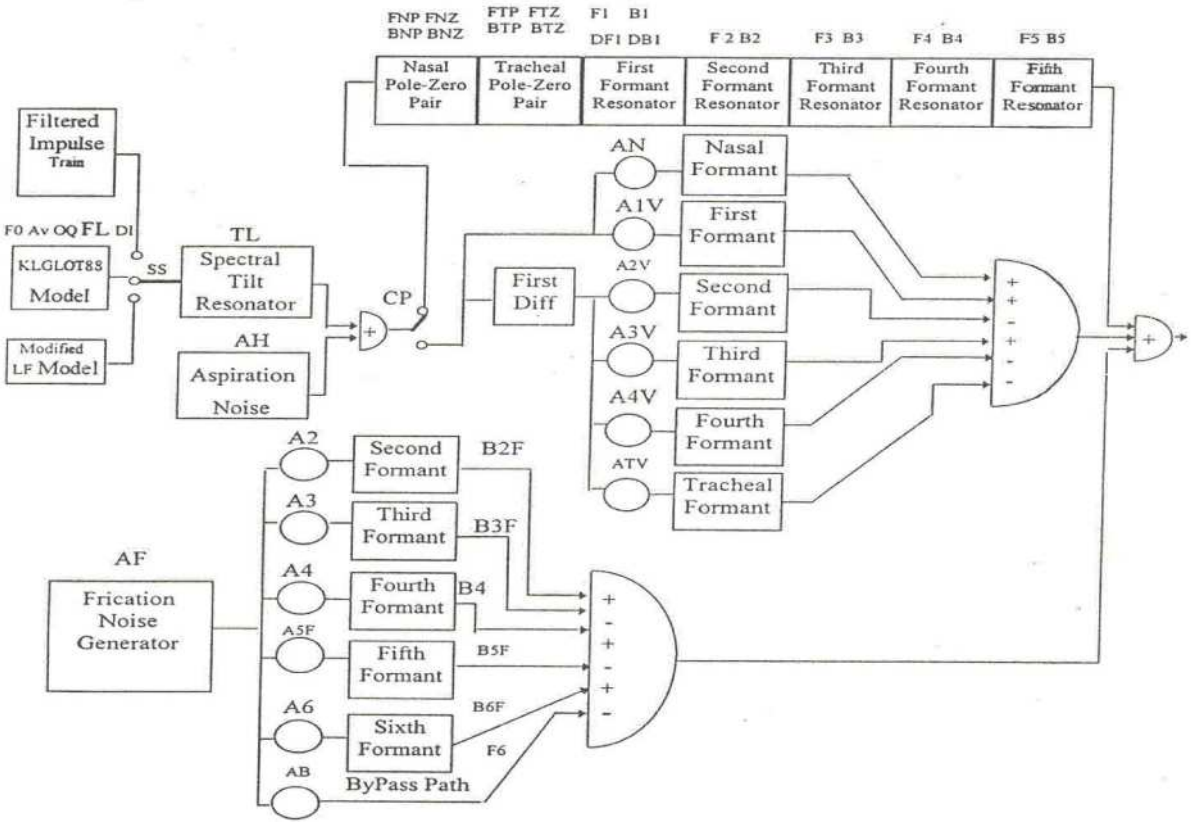
- تعیین مدل گفتار و پارامترهای مدل
- تعیین نوع واحد های آوایی
- تعیین پارامترهای مناسب برای تولید واحد های آوایی مختلف
- جانشین کردن واحدهای آوایی با دنباله پارمتری هم ارز آن
- قرار دادن دنباله پارمترهای درون مدل گفتار طراحی شده

در تصویر 4 یکی از مدل های مبتنی بر قانون مشهور به نام KLAT80 را مشاهده می کنید.



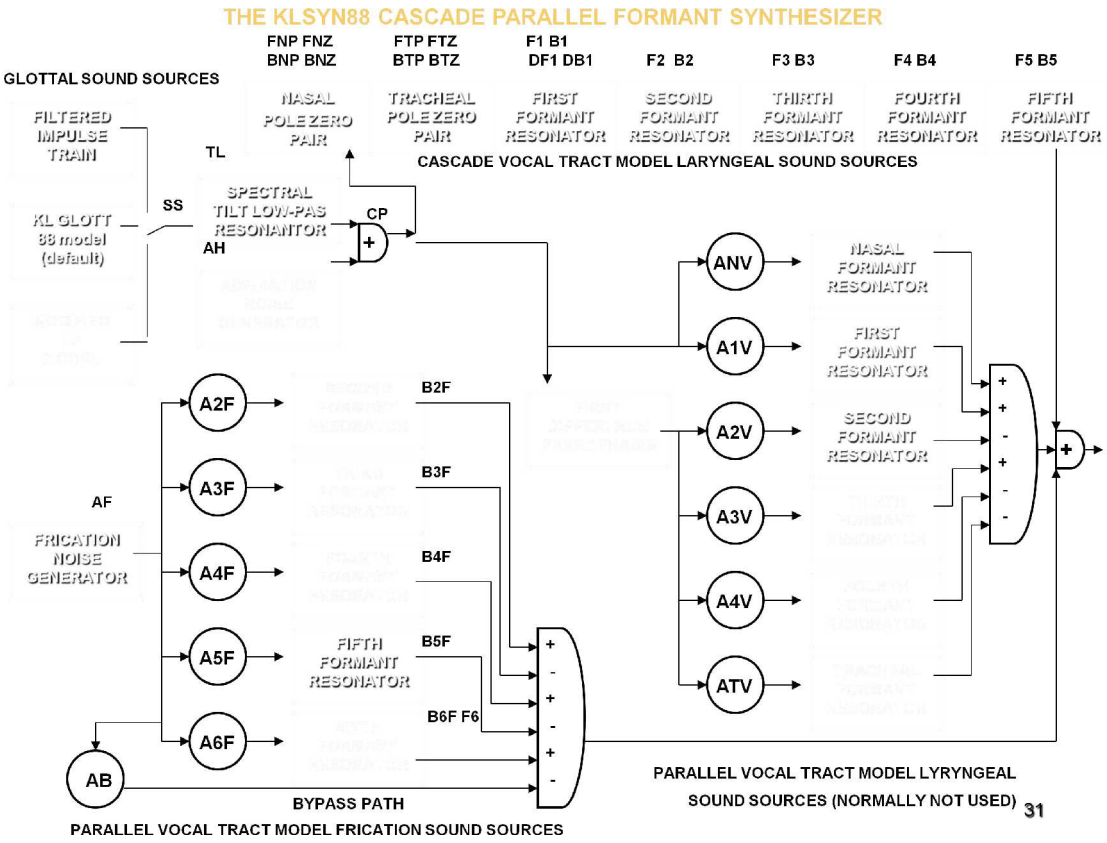
تصویر 4 – KLAT80

در تصویر 5 مدل بهبود یافته آن یعنی KLATT88 را مشاهده می کنید.



تصویر 5 – KLATT88

نمایی دیگر از KATT88 را در تصویر 6 مشاهده می کنید.



تصویر 6 – KLATT88

یکی از نمونه های سنتز مبتنی بر قانون روش سنتز مبتنی بر فرمنت می باشد.

5– سنتز آماری

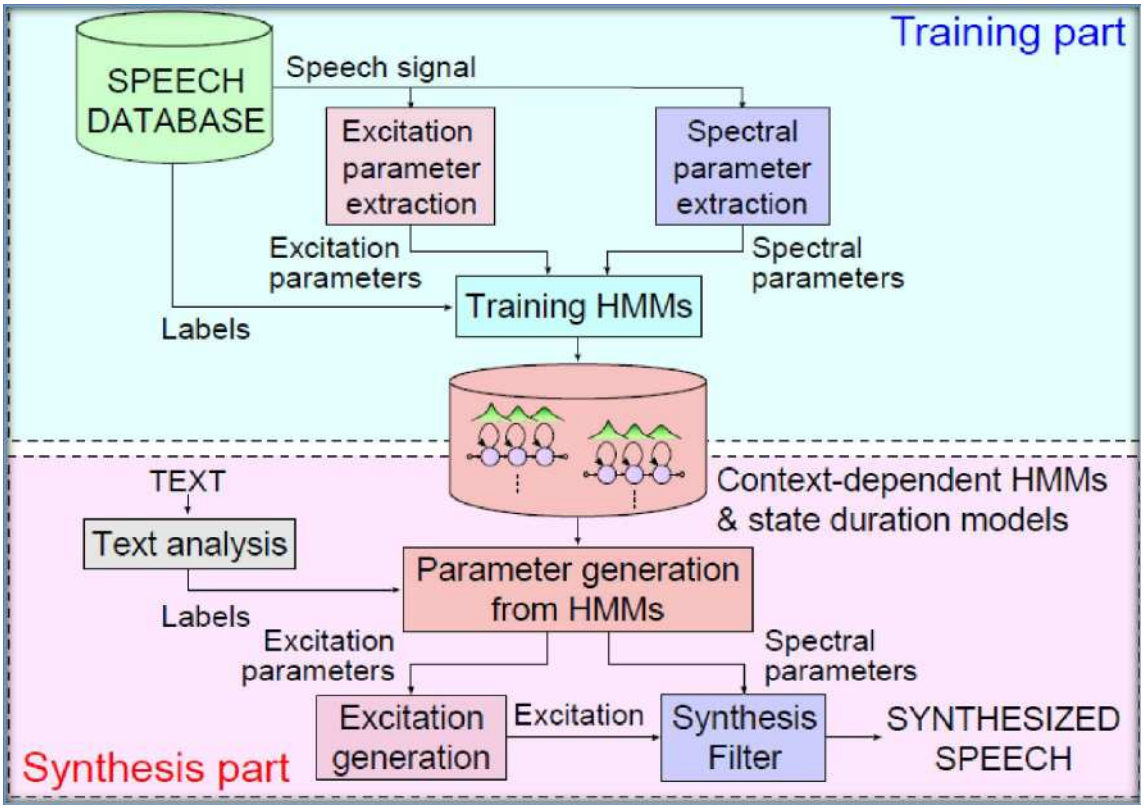
این سنتز مبتنی بر دادگان زیاد می باشد.

پارامترهای زیادی را از روی این دادگان آموزشی، آموزش می دهند.

مدل منبع–فیلتر + مدل صوتی آماری برای این روش مورد نیاز است.

معمولاً از مدلهای مخفی مارکوف به عنوان مدل صوتی آن استفاده می کنیم.

کل فرآیند سنتز مبتنی بر مدل مخفی مارکوف را در تصویر 7 مشاهده می کنید.



تصوی 7 – سنتز مبتنی بر مدل مخفی مارکوف

ابتدا بازنمایی های پارمتری گفتار (شامل پارامترهای طیف و پارامترهای تحریک) را از دیتابیس گفتار به دست می آوریم.

بوسیله مجموعه ای از مدل های تولیدکننده (مانند HMM) آن ها را مدل می کنیم.

$$\hat{\lambda} = \arg\max_{\lambda} p(O \mid \mathcal{W}, \lambda)$$

آموزش: تخمین پارمترها

سنتز: به دست آوردن پارمترهای طیف و تحریک از روی دنباله واجی

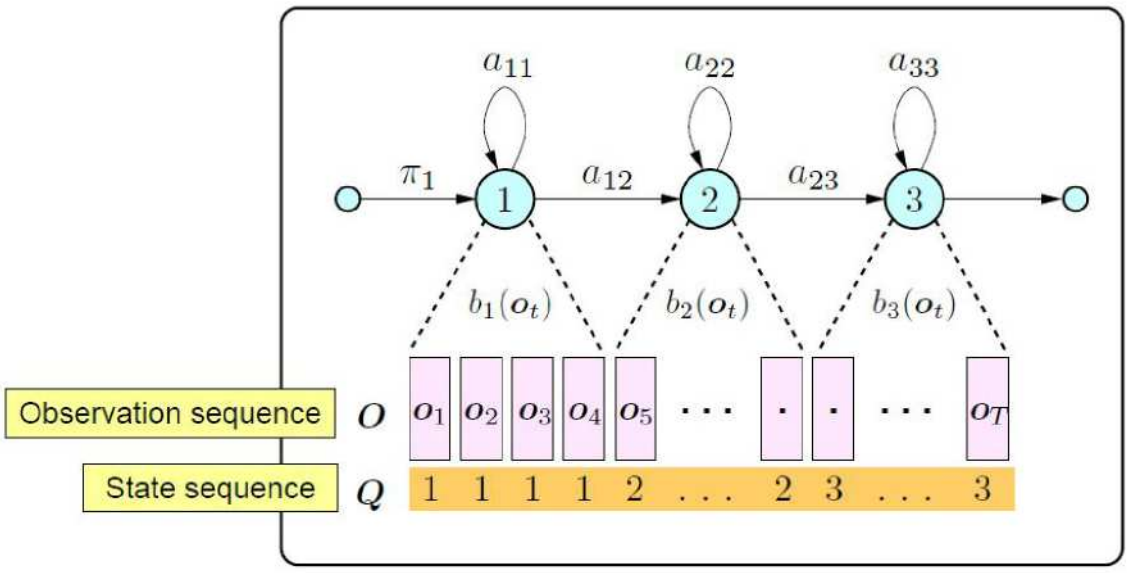
$$\hat{o} = \arg\max_{o} \; p(o \mid w, \hat{\lambda})$$

سه مدل کردن:

- مدلینگ پارامترهای طیف
- مدلینگ پارامترهای تحریک
- مدلینگ مدت زمان حالت ها

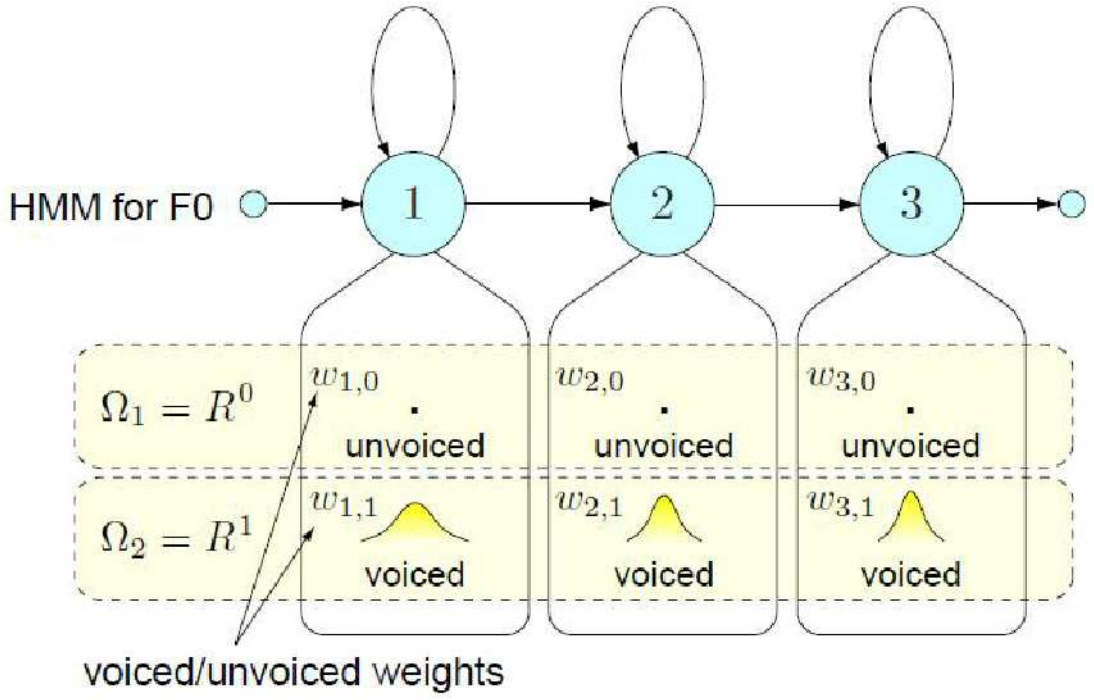
از آنالیز مل-کپسترال برای تخمین طیف استفاده می شود.

از یک HMM با چگالی احتمال پیوسته برای مدل کردن مسیر صوتی استفاده می شود (درست مانند سیستم بازشناسی گفتار) (تصویر 8).



تصویر 8 – استفاده از HMM برای مدل کردن مسیر صوتی

برای مدل کردن فرکانس گام، این مشکل وجود دارد که در قسمت های صدادار فرکانس گام وجود دارد ولی در قسمت های بدون صدا وجود ندارد. می توان فرض کرد که مقدار پیوسته قسمت صدادار از یک فضای یک بعدی و قسمت بدون صدا از یک فضای صفر بعدی آمده است (تصویر 9).



تصویر 9 – در نظر گرفتن فرکانس گام درون HMM

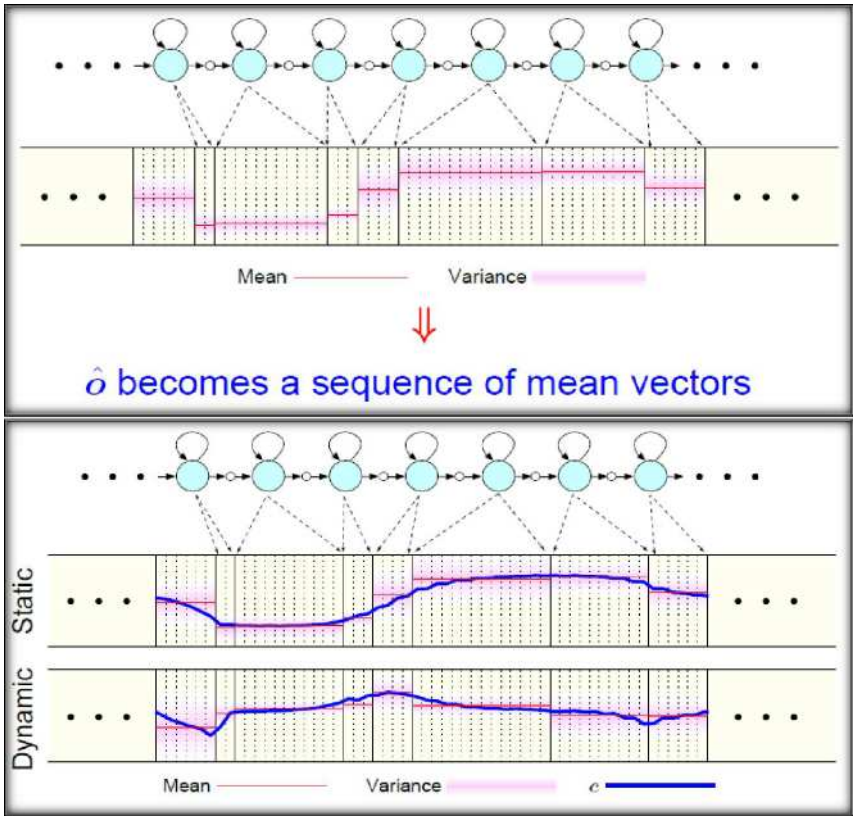
مانند کاربردهای بازشناسی، در اینجا نیز از ویژگی های پویا استفاده می شود.

فرمول 2

$$\Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1})$$

$$\Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1}$$

این ها باعث پیوسته شدن طیف نهایی می شوند (تصویر 10).



تصویر 10 – به کار گیری ضرایب پویا در HMM

فیلترمل-کپستروم استفاده شده فیلتر MLSA نام دارد.

7 – خلاصه و نتیجه گیری

در این فصل بحث سنتز را شروع کردیم.

تبدیل متن به گفتار

- مبتنی بر قانون
- مفصلی
- چسباندنی (concatenative)
- انتخاب واحد
- آماری (مبتنی بر مدل مخفی مارکوف)

8 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

تبدیل متن به گفتار

- انتخاب واحد

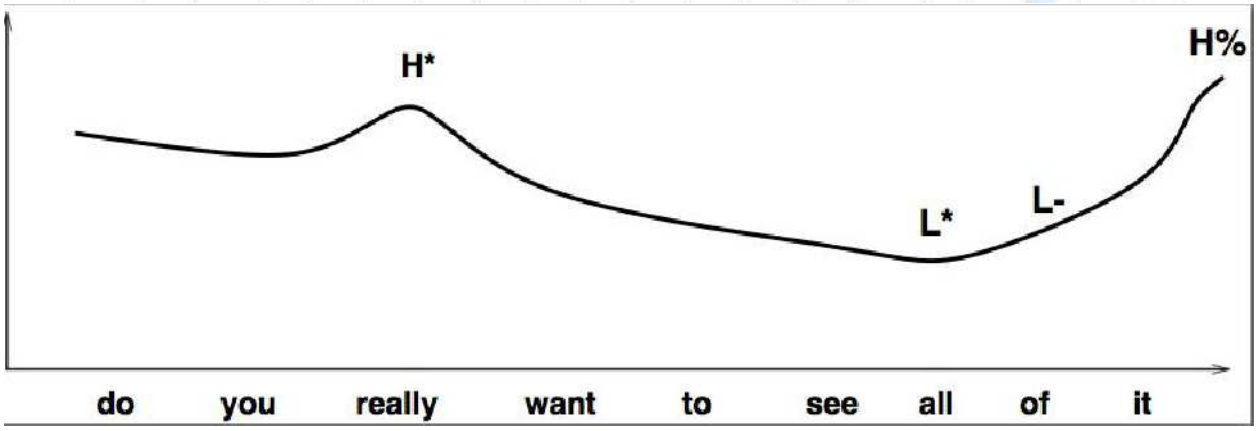
2– روش انتخاب واحد (unit selection)

فرض کنید که اطلاعات زیر را داریم (تصویر 1):

- دنباله واجی
- پروزودی
  - فرکانس گام کل گفتار خروجی
  - مدت زمان هر واج
  - مقدار تاکید هر واج

هدف این است که «شکل موج» خروجی را تولید کنیم.

	do		you		H* really					want				to		see		L* all		L- H% of		it	
	d	uw	y	uw	r	ih	l	iy	w	aa	n	t	t	ax	s	iy	ao	l	ah	v	ih	t
	110	110	50	50	75	64	57	82	57	50	72	41	43	47	54	130	76	90	44	62	46	220

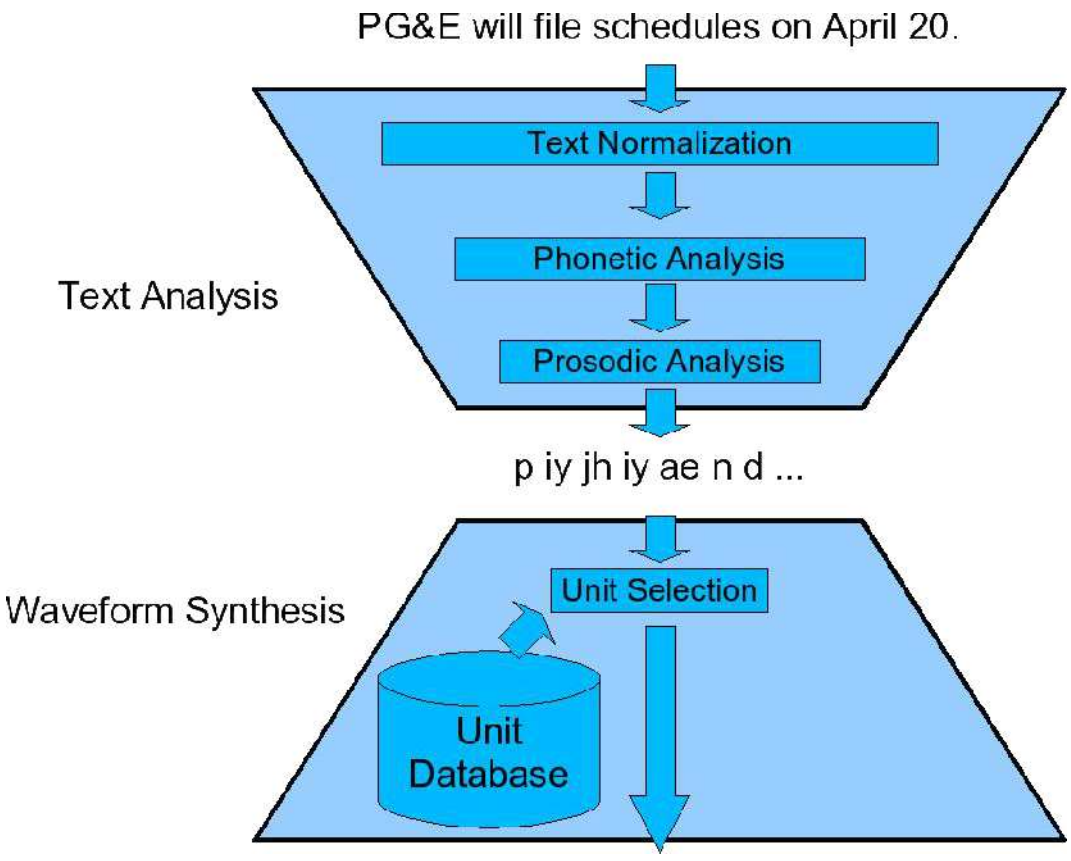


تصویر 1 – ورودی برای تولید شکل موج

در این بخش دو روش دایفون و انتخاب واحد که در جلسات قبل به اختصار توضیح دادیم را به تفصیل توضیح خواهیم داد:

- سنتز دایفون
- سنتز انتخاب واحد
  - هزینه هدف
  - هزینه الحاق
- الحاق شکل موج ها
  - ساده
  - PSOLA

ساختار کلی روش الحاقی را در تصویر 2 مشاهده می کنید.



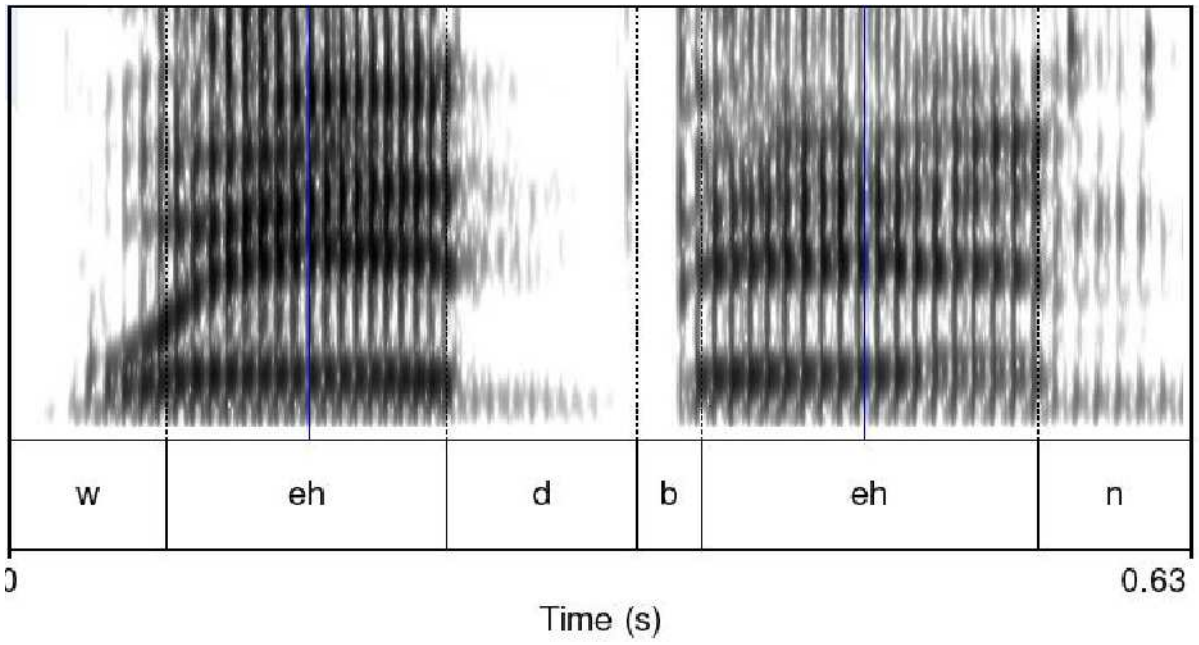
تصویر 2 – ساختار کلی سیستم الحاقی

آموزش

- انتخاب واحد آوایی (دایفون)
- ضبط صدای یک گوینده که هر دایفون را تلفظ می کند
- مرزهای دایفون را مشخص می کنیم

سنتز

- دنباله دایفون مناسب را از دیتابیس استخراج کن.
- دایفون ها را با هم الحاق کن (بوسیله عملیات پردازش سیگنال)
- استفاده از پردازش سیگنال برای تغییر پروزودی (گام، انرژی و مدت) دنباله دایفون ها



تصویر 3 – میانه واج پایدارتر از مرزهای واج

در تصویر 3 مشاهده می کنید که میانه واج ها پایدارتر از لبه ها می باشد.

در کل برای دایفون ها به $O(phone^2)$ واحد آوایی نیاز داریم.

برخی ترکیب ها اصلاً در زبان وجود ندارند.

سیستم ATT دارای 43 واج می باشد.

در کل 1172 دایفون در زبان انگلیسی وجود دارد. (در تئوری 1849 دایفون می تواند وجود داشته باشد).

این سیستم از دیتابیس کوچکی استفاده می کند (8 مگابایت)

برای ساختن دیتابیس دایفون دو روش وجود دارد:

1. استفاده از کلمات بی معنی که شامل دایفون های مورد نظر باشند.

برای مثال:

- pau t aa b aa b aa pau
- pau t aa m aa m aa pau
- pau t aa m iy m aa pau
- pau t aa m iy m aa pau
- pau t aa m ih m aa pau

مزیت:

- به راحتی همه دایفون ضبط می شوند
- دایفون ها به درستی تلفظ می شوند
- به فرهنگ لغت ربط نخواهد داشت

عیب:

- دیتابیس بزرگ
- گوینده در حین تلفظ خسته می شود

2. انتخاب کلمات و جملاتی به صورت دلخواه

مزیت:

- تلفظ ها طبیعی خواهند بود
- برای تلفظ آسان تر خواهند بود
- دیتابیس کوچکتر

عیب:

- ممکن است تلفظ طبیعی باشد ولی اشتباه باشد

نکات زیر در مورد ضبط باید رعایت شود:

- دایفون باید از میانه کلمه انتخاب شود. در این صورت articulation کامل خواهد بود.
- به صورت یکسان تلفظ شود. یعنی گام، انرژی و مدت زمان برابر باشد

برای برچسب گذاری دایفون ها باید به صورت های زیر عمل کرد:

یک بازشناسی گفتار به صورت fore alignment اجرا کرد تا برچسب در همه زمان ها به دست آید.

برای این کار نیاز به :

- سیستم بازشناس گفتار خودکار آموزش داده شده
- فایل صوتی
- کلمات تلفظ شده در فایل صوتی

داریم. به عنوان خروجی واج های تلفظ شده در هر زمان داده می شود.

می توان فقط از بازشناس واج استفاده کرد.

زیرا دنباله واجی را می توان از دنباله کلمات به دست آورد.

سپس با استفاده از رمزگشایی HMM مرز واج ها را به دست آورد.

تنها مشکل تلفظ اشتباه گوینده می باشد.

البته اشتباه در شناسایی مرزها تا 10-+ میلی ثانیه مشکلی ندارد.

ولی قسمت میانی واج ها مهم است. اینکه کدام قسمت، قسمت پایدار واج می باشد.

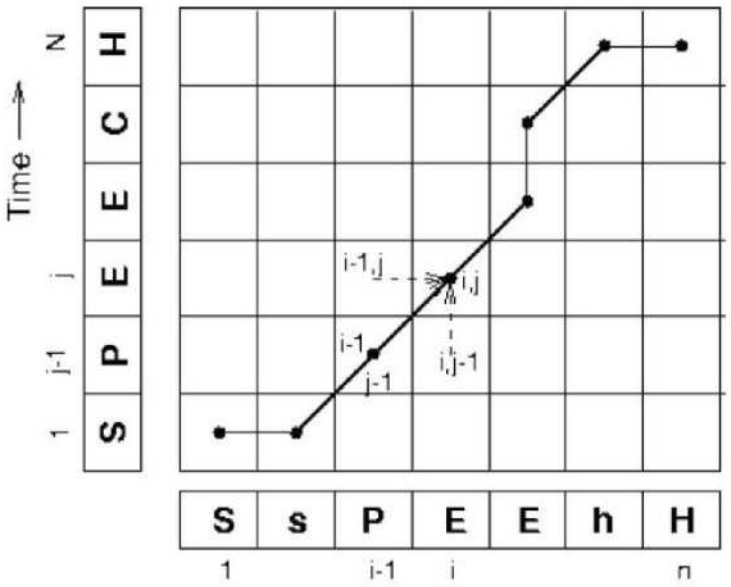
سئوال این است مه آیا می توان این قسمت را به صورت خودکار یافت؟

روش دیگر برای برچسب گذاری دایفون ها استفاده ازتطبیق زمانی پویا می باشد (به جلسات بحث بازشناسی گفتار مراجعه شود).

فرض می کنیم داریم:

- تلفظ انسانی جمله
- تلفظ سنتز شده جمله

بوسیله تطبیق زمانی پویا یک تطبیق بین آن ها انجام بده. (فاصله اقلیدسی استفاده می کنیم) (تصویر 4).



تصویر 4 – اجرای DTW بر روی دو تلفظ از یک کلمه

برای شناسایی قسمت های پایدار واج ها به صورت زیر عمل می شود:

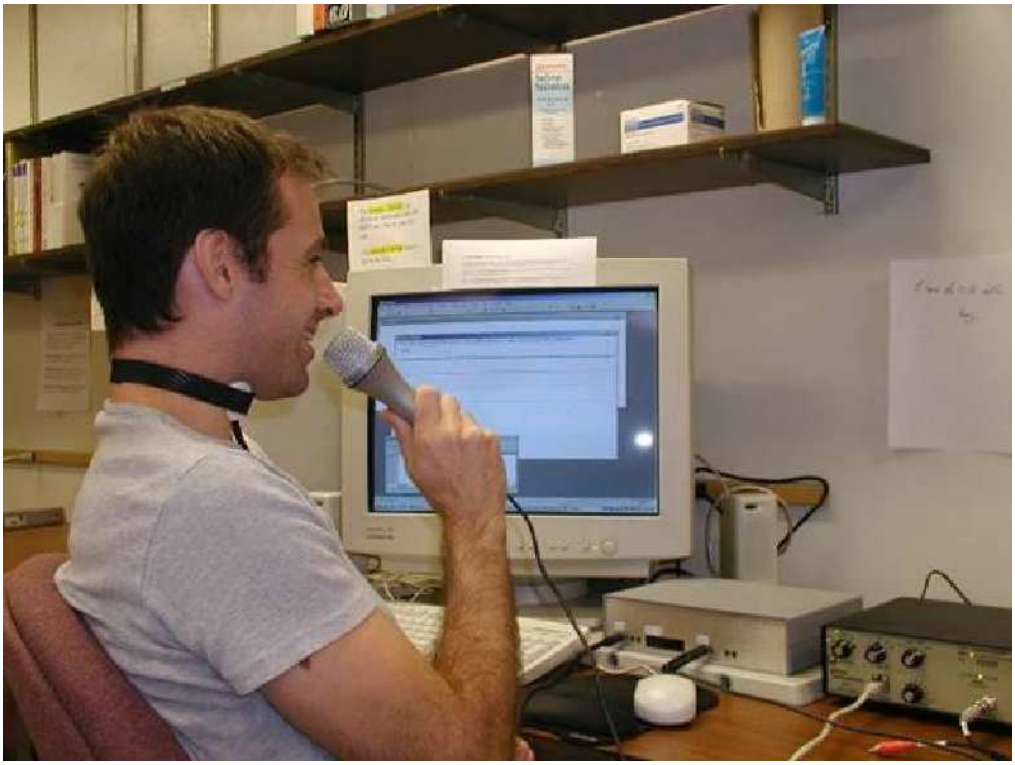
- برای انفجاری ها: یک سوم داخل
- برای واج–سکوت ها: یک چهارم داخل
- برای بقیه دایفون ها: 50 درصد داخل

در هنگام سنتز باید شکل موج ها را به هم بچسبانیم.

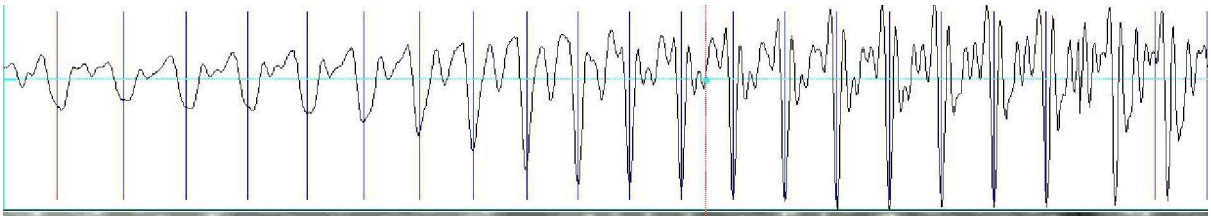
همچنین نیاز است فرکانس گام نمونه های دایفون را تغییر دهیم. برای این کار نیاز است مکان رخداد گام در سیگنال مشخص شود. یعنی باید زمان بسته شده تارهای صوتی مشخص شود.

برای این کار دو روش استفاده می شود:

- بوسیله دستگاه EGG و در هنگام ضبط. این دستگاه روی گلو بسته می شود (تصویر 5)
- به روش های پردازش سیگنال (تصویر 6)



تصویر 5 – استفاده از دستگاه EGG

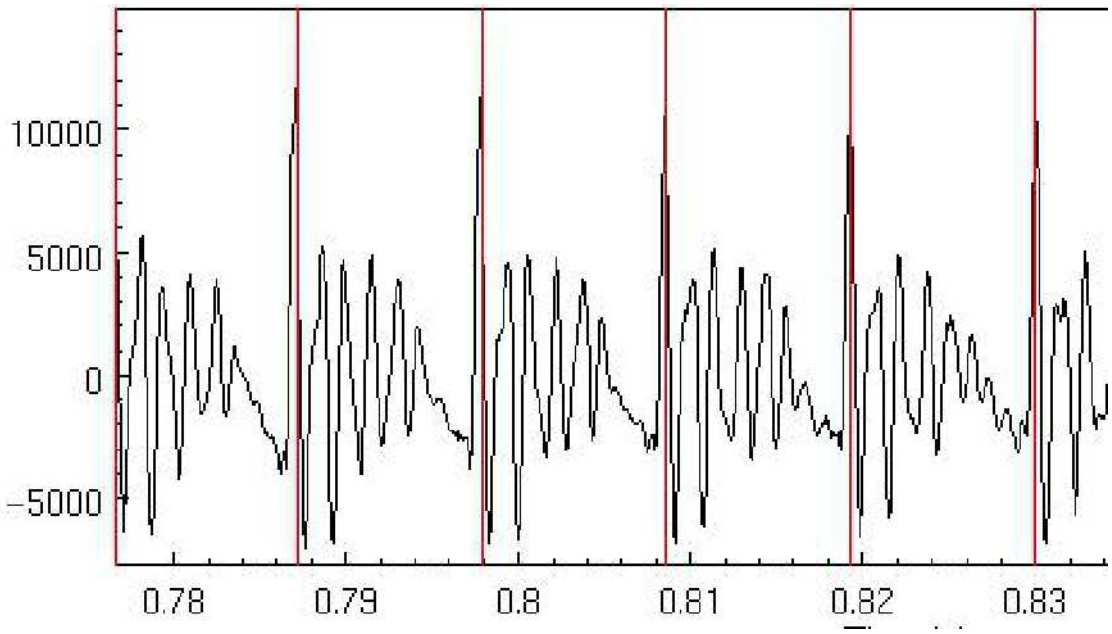


تصویر 6 – استخراج نقاط بسته شدن حنجره در نرم افزار Pratt

برای زیر و بم کردن یک نمونه گفتار باید فرکانس گام تغییر کند ولی مدت زمان ثابت بماند.

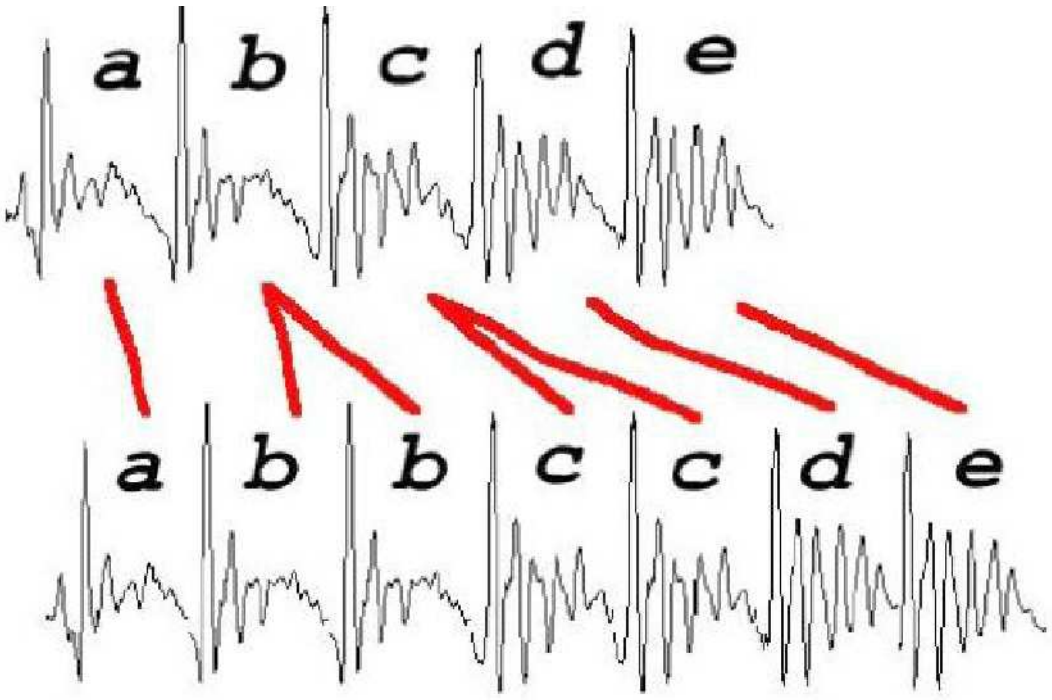
زیاد کردن فرکانس نمونه برداری باعث گفتار زیر تر می شود ولی مدت زمان سیگنال نیز کم می شود.

فرض کنید سیگنال تصویر 7 را داریم.



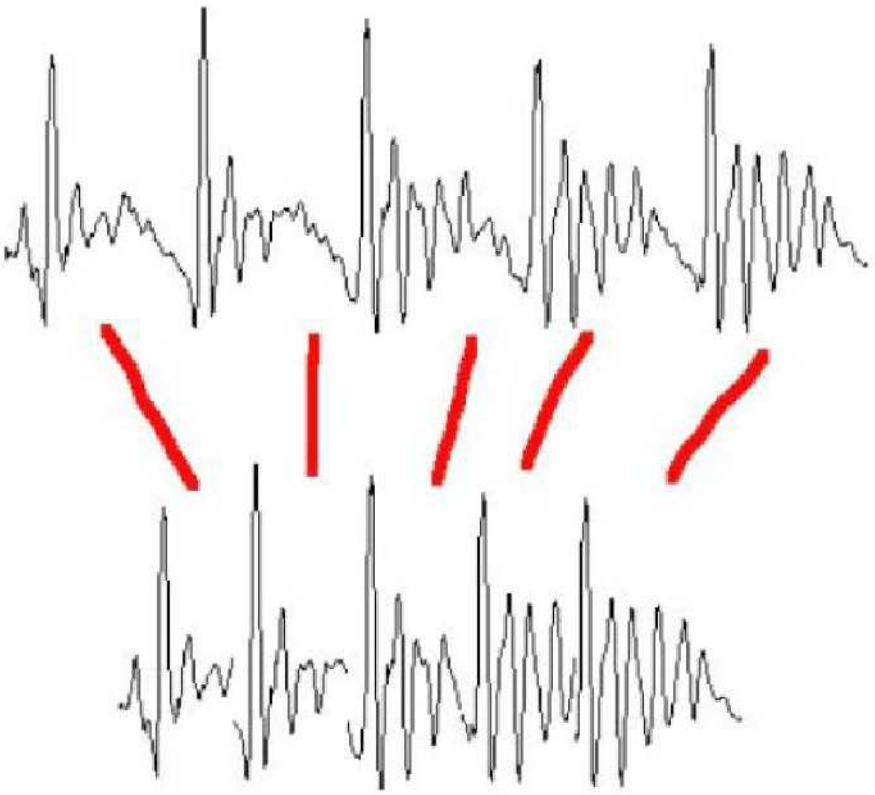
تصویر 7 – سیگنال صدادار نمونه

در تصویر 8 چگونگی تغییر طول مدت زمان تلفظ را مشاهده می کنید.



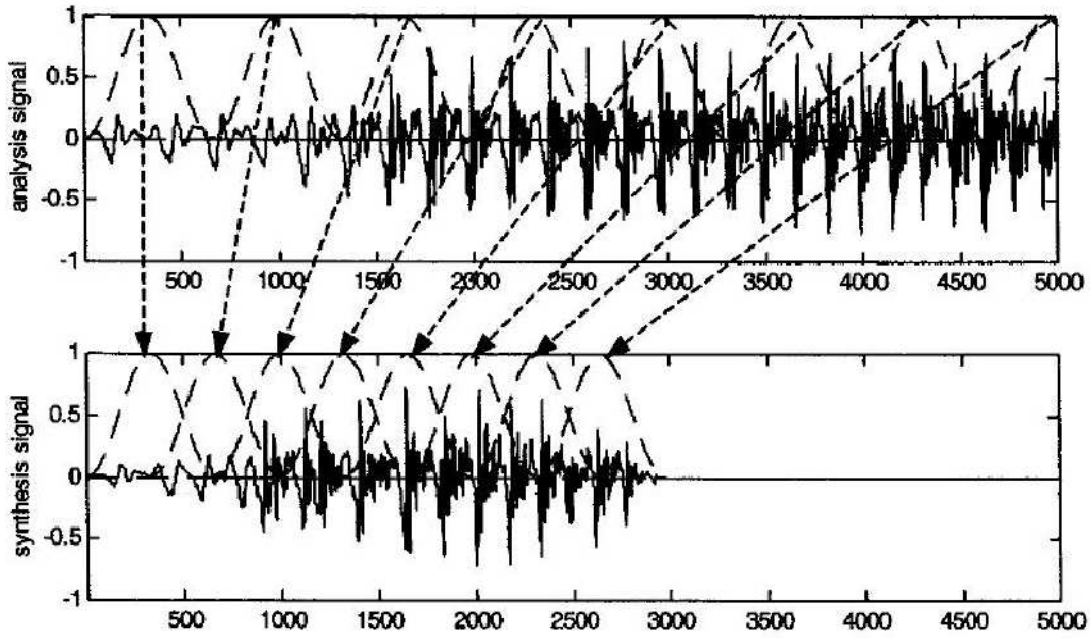
تصویر 8 – تغییر مدت زمان تلفظ

در تصویر 9 مثالی از تغییر فرکانس گام را مشاهده می کنید.



تصویر 9 – تغییر (کم کردن) فرکانس گام با کپی کردن و قطع قسمت های اضافی

روش overlap-add این تفاوت را دارد که سیگنال را به جای قطع کردن پنجره می زند و با هم همپوشانی داده و قطع می کند.



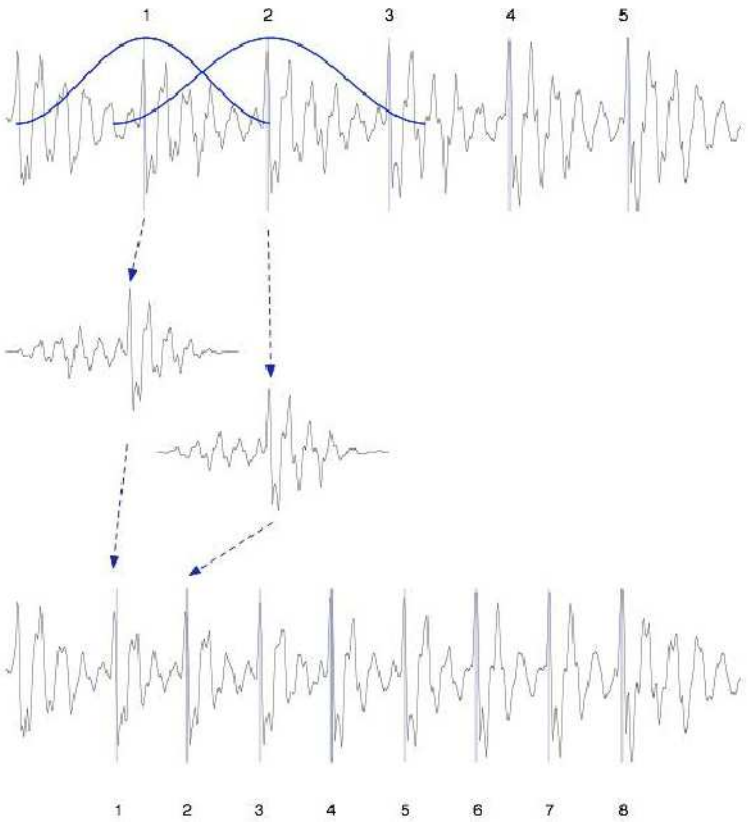
تصویر 10 – روش overlap-add

روش PSOLA به این صورت است که این عمل را بر روی بسته شدن های حنجره انجام می دهد.

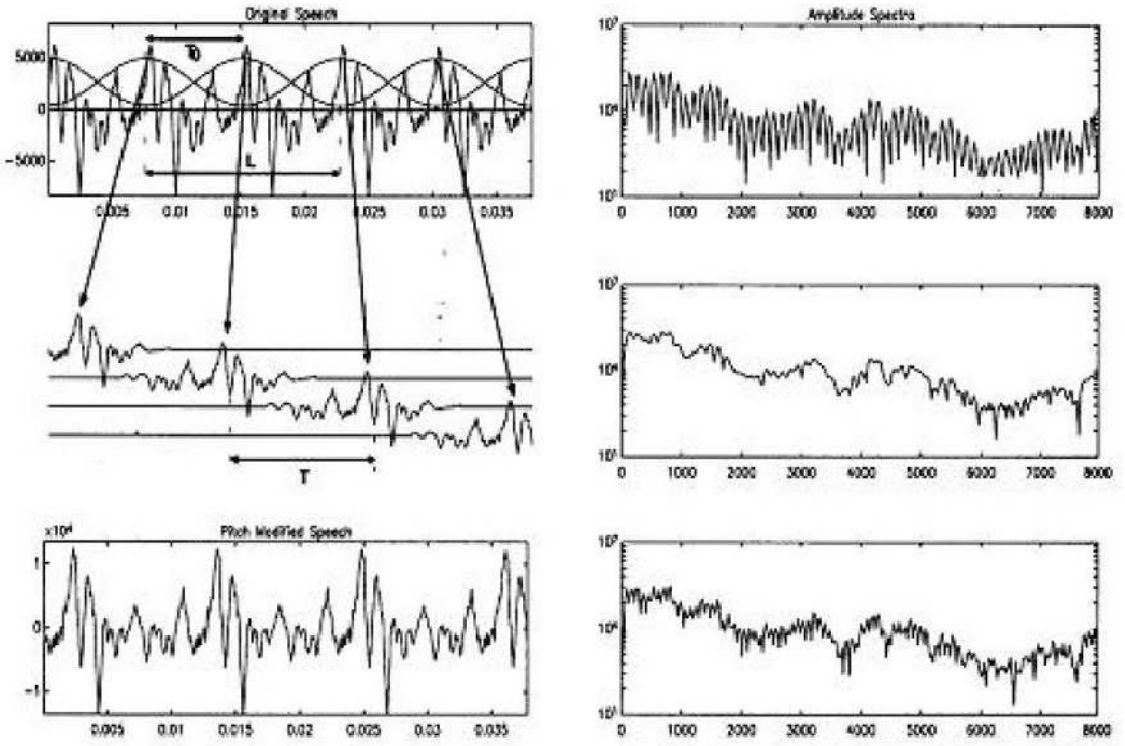
یعنی فرکانس گام را محاسبه می کند و روی رخداد های گام پنجره می زند.

- برای افزایش فرکانس (زیر کردن صدا): آن پنجره را نصف می کند و تعداد آن را دو برابر می کند.
- برای کاهش فرکانس (بم کردن صدا): پنجره را دو برابر می کند (با interpolation) و تعداد پنجره ها را نصف می کند.

همان طور که مشخص است، روش PSOLA فقط قادر است فرکانس سیگنال را یا نصف کند و یا دو برابر کند (تصویر 11 و 12)



تصویر 11 – دو برابر کردن فرکانس گام یک شکل موج به روش PSOLA



تصویر 12 – نصف کردن فرکانس گام یک شکل موج به روش PSOLA

حال که روش الحاق شکل موج ها را مشاهده کردیم، در ادامه روش انتخاب شکل موج مناسب را بیان می کنیم.

داده طبیعی مشکلات سنتز دایفون را حل می کند. زیرا اغلب این مشکلات به کم بودن و غیر طبیعی بودن دیتا مربوط می شود.

فرض کنید دیتابیس بزرگی از واحدها داریم.

برای هر دایفون که قصد سنتز آن را داریم:

- واحدی را در دیتابیس پیدا کن که «بهترین» برای زمینه مورد نظر است.

حال سئوال این است که بهترین چه معنایی دارد؟ برای تعریف بهتر بودن و بدتر بودن دو هزینه تعریف می شود:

- هزینه هدف (Target Cost): نزدیک ترین مطابقت با توصیف هدف، با در نظر گرفتن:
  - زمینه آوایی

- گام، تاکید و مکان عبارت
- هزینه الحاق (Join Cost):
  - تطبیق فرمنت + دیگر ویژگی های طیفی
  - مطابقت انرژی
  - مطابقت فرکانس گام

کل فرمول هزینه در فرمول 1 آمده است.

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n} C^{target}(t_i,u_i)+\sum_{i=2}^{n} C^{join}(u_{i-1},u_i)$$

فرمول 1

هزینه هدف

این هزینه نشان می دهد که یک واحد آوایی در دیتابیس چه مقدار به واحد آوایی مورد نظر نزدیک است.

برای محاسبه این مقدار نیاز به ویژگی ها، هزینه ها و وزن ها می باشد.

شامل k زیر هزینه می باشد:

- تاکید
- مکان عبارت
- فرکانس گام
- مدت واج
- شناسه فرهنگ لغت

$$C^{target}(t_1^n,u_1^n)=\sum_{k=1}^{p} w_k^t C_k^t(t_i,u_i)$$

فرمول 2

روش های خیلی زیادی برای تنظیم وزن وجود دارد. ساده ترین روش این است که از وزن ثابت استفاده کنیم.

هزینه الحاق

هزینه میزان صاف بودن الحاق

بین دو واحد آوایی دیتابیس محاسبه می شود (هدف ربطی به این مورد ندارد).

از k زیر هزینه تشکیل شده است:

- ویژگی های طیفی
- فرکانس گام
- انرژی

فرمول 3

$$C^{join}(u_{i-1},u_i)=\sum_{k=1}^{p} w_k^j C_k^j(u_{i-1},u_i)$$

در یکی از روش ها به صورت زیر استفاده شده است:

- طیف: ویژگی های ضرایب کل-کپسترال
- فرکانس گام محلی
- انرژی کل محلی
- وزن های به صورت دستی مقداردهی شده

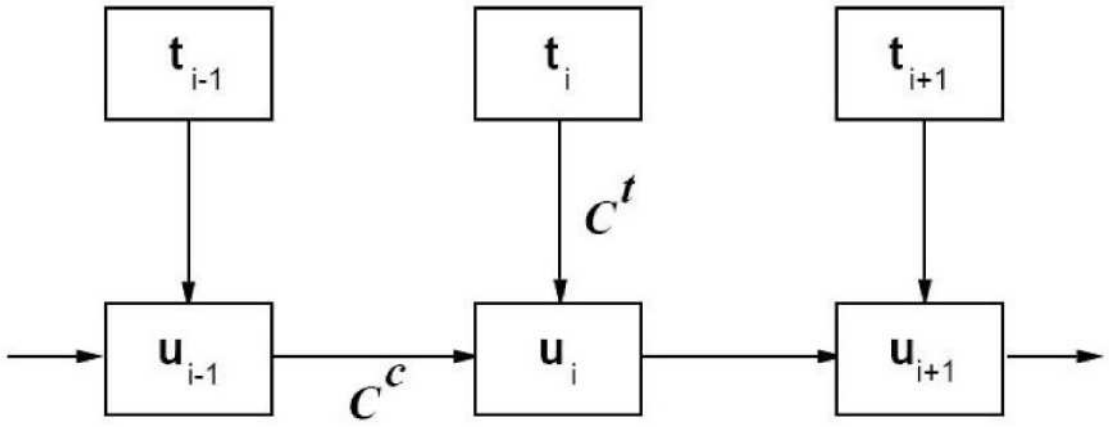
در نهایت هزینه نهایی به صورت جمع دو هزینه هدف و الحاق محاسبه می شود.

فرمول 4

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n} C^{target}(t_i,u_i)+\sum_{i=2}^{n} C^{join}(u_{i-1},u_i)$$

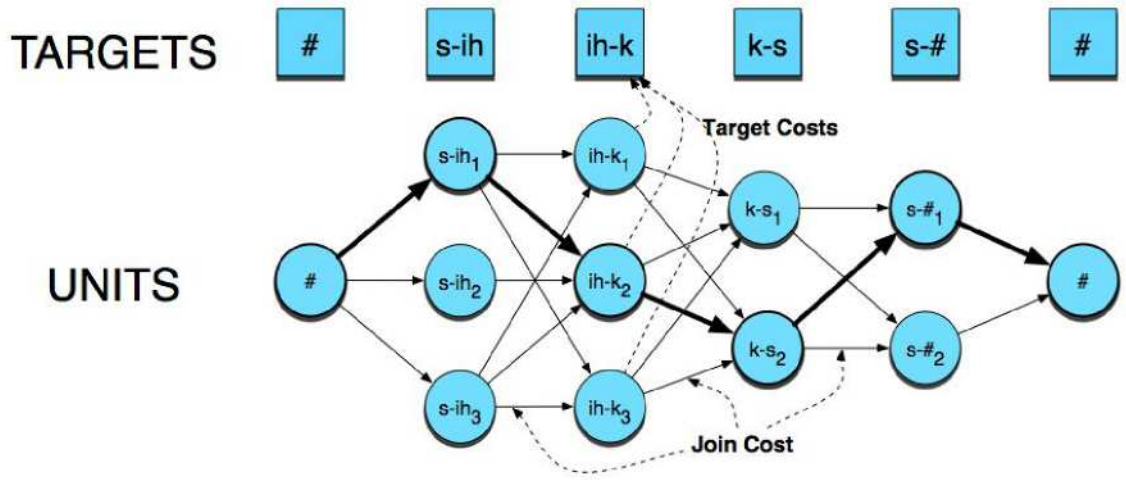
مسئله یافتن مسیری است که فرمول 4 را مینیمم می کند.

می توان بوسیله یک جستجوی ویتربی این مسئله را حل کرد (یافتن مسیر).



تصویر 13 – جستجوی انتخاب واحد

خلاصه جستجوی ویتربی انتخاب واحد را در تصویر 14 مشاهده می کنید.



تصویر 14– خلاصه جستجوی ویتربی انتخاب واحد

3 – خلاصه و نتیجه گیری

در این فصل بحث سنتز را بیان کردیم.

تبدیل متن به گفتار

- انتخاب واحد

8 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

ارزیابی کیفیت

- ارزیابی انسانی (subjective)
- ارزیابی کامپیوتری (objective)

نوع ارزیابی

- ارزیابی طبیعی بودن سیگنال (Naturalness)
- ارزیابی قابلیت درک انسان (Intelligibility)

2– مفاهیم اولیه

کیفیت های زیر در بحث مخابرات مطرحند:

- کیفیت پایین (زیر 4.8 کیلوهرتز)
- کیفیت متوسط (4.8 تا 13 کیلوهرتز)
- کیفیت خوب (13 تا 64 کیلوهرتز)
- کیفیت عالی (بالای 64 کیلوهرتز)

در تصویر 1 جدول روش های بررسی شده در این درس را مشاهده می کنید.

	Intelligibility	Naturalness
Subjective	DRT, MRT	MOS, DAM

AI, Global SNR, Seg. SNR, FW-Seg. SNR, Itakura Measure, WSSM	None. Future ASR systems	Objective

تصویر 1 – ارزیابی های بررسی شده در این درس

3– ارزیابی انسانی قابلیت درک

تست Diagnostic Rhyme (DRT)

- انتخاب بین دو CVC با Cهای متفاوت
- مثال: hop-fop و than-dan
- DRT بسیار معتبر و پرکاربرد است.
- در این آزمایش کاربر فقط یکبار گفتار را می شوند.

فرمول 1

$$DRT\% = \frac{N_{Correct} - N_{Incorrect}}{N_{Tests}} \times 100$$

تست Modified Rhyme (MRT)

- انتخاب CVC های با C های متفاوت
- مثال: Cat, bat, rat, mat, fat, sat

4– ارزیابی انسانی طبیعی بودن

امتیاز متوسط عقیده (Mean Opinion Score MOS)

- MOS بسیار پرکاربرد و بسیار معتبر است
- در این تست کاربر می تواند به دفعات گفتار را بشنود
- نمونه امتیاز ها را در تصویر 2 مشاهده می کنید.

Score	Speech Quality
1	Not Acceptable
2	Weak
3	Medium
4	Good
5	Excellent

تصویر 2 – نمونه امتیازهای MOS

تست Diagnostic Acceptability Measure (DAM)

این تست بسیار پیچیده است.

در این تست 19 پارامتر مختلف برای امتیازدهی وجود دارد. این پارامترها به 3 گروه اصلی تقسیم می شوند:

- کیفیت سیگنال
- کیفیت پس زمینه
- کیفیت کل

5– ارزیابی کامپیوتری طبیعی بودن

نمی توان از این تست ها برای ارزیابی قابل درک بودن استفاده کرد.

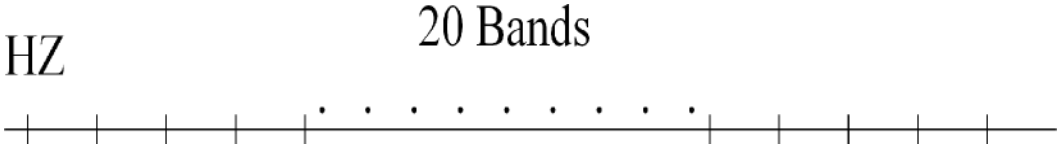
البته استفاده از سیستم های بازشناس گفتار را می توان نوعی ارزیابی برای قابل درک بودن دانست.

تست های کامپیوتری ارائه شده در این درس فقط برای ارزیابی طبیعی بودن گفتار می باشد.

Articulation Index (AI)

AI فرض می کند که باندهای مختلف فرکانسی مستقلند و کیفیت سیگنال را در باندهای مختلف محاسبه می کند.

در هر باند درصد سیگنال شنیده شده توسط شنونده محاسبه می شود (تصویر 3).



تصویر 3 – باندهای مختلف

شرایط قابل شنیدن بودن توسط شنونده:

- بالاتر بودن از سطح آستانه شنوایی انسان
- زیر آستانه درد اسنان
- بیشتر بودن از سطح ماسک کردن نویز

در AI معیار SNR در هر باند محاسبه می شود.

$$AI = \frac{1}{20}\sum_{j=1}^{20}\frac{Min(SNR,30)}{30}$$

فرمول 1

Signal to Noise Ratio (SNR)

همان طور که از اسم این روش بر می آید (نسبت سیگنال به نویز) انرژی سیگنال به نویز را محاسبه می کند.

$$\varepsilon_{(n)} = s_{(n)} - \hat{s}_{(n)}$$

$$E_\varepsilon = \sum_{n=-\infty}^{\infty} \varepsilon_{(n)}{}^2 = \sum_{n=-\infty}^{\infty} [s_{(n)} - \hat{s}_{(n)}]^2 \qquad E_s = \sum_{n=-\infty}^{\infty} s_{(n)}{}^2$$

فرمول 2

$$SNR_{(global)} = 10\log\frac{E_s}{E_\varepsilon} = 10\log\frac{\sum_{n=-\infty}^{\infty} s_{(n)}{}^2}{\sum_{n=-\infty}^{\infty} [s_{(n)} - \hat{s}_{(n)}]^2}$$

Segmental SNR

به این صورت است که سیگنال را به فریم هایی تقسیم می کند و SNR را بر روی آن انجام می دهد و میانگین گیری می کند.

$$SNR_{(seg)} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log\left[\frac{\sum_{n=m_j-N+1}^{m_j} s_{(n)}{}^2}{\sum_{n=m_j-N+1}^{m_j} [s_{(n)} - \hat{s}_{(n)}]^2}\right]$$

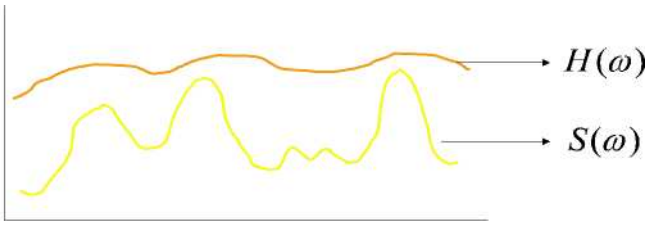
Frequency Weighted Segmental SNR

به هر باند فرکانسی وزنی داده می شود (مثلاً بر اساس مقیاس مل)

فرمول 4

$$SNR_{(fw-seg)} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log\left[\frac{\sum_{k=1}^{K} W_{j,k}[E_{s,k}(m_j)/E_{\varepsilon,k}(m_j)]}{\sum_{k=1}^{K} W_{j,k}}\right]$$

Itakura



$H(\omega)$ Is the envelope spectrum

$$S(\omega) = F\{R(\tau)\} \Rightarrow S(\omega) = |X(\omega)|^2$$

Use from All-Pole (AR) Model

تصویر4 – معیار itakura

$$H(\omega) = \frac{1}{1 - \sum_{i=1}^{p} a_i e^{-j\omega}}$$

معیار itakura به صورت زیر است.

فرمول 5

$$d(g_s(m), g_{\hat{s}}(m)) = \sqrt{\frac{1}{M}\sum_{l=1}^{M}[g_s(l,m) - g_{\hat{s}}(l,m)]^2}$$

6 – خلاصه و نتیجه گیری

در این فصل بحث ارزیابی کیفیت را بیان کردیم.

7 – منابع درس

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”

1– مقدمه

اهداف درس:

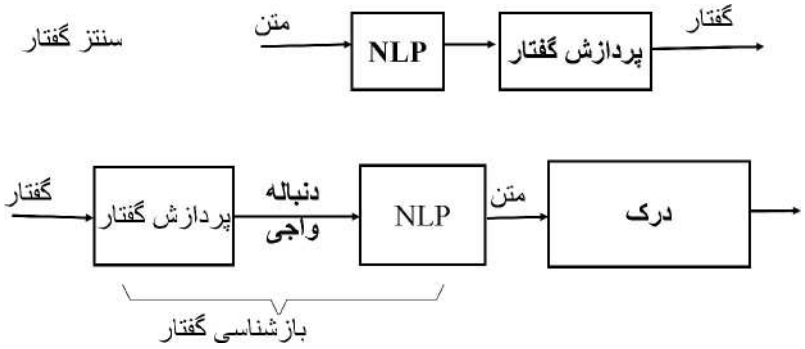
آشنایی با مفهوم اولیه بازشناسی گفتار.

آشنایی کلی با روش های حل مسئله بازشناسی گفتار

2– مفاهیم اولیه

مسئله بازشناسی گفتار مسئله تبدیل ورودی صوتی به متن است.

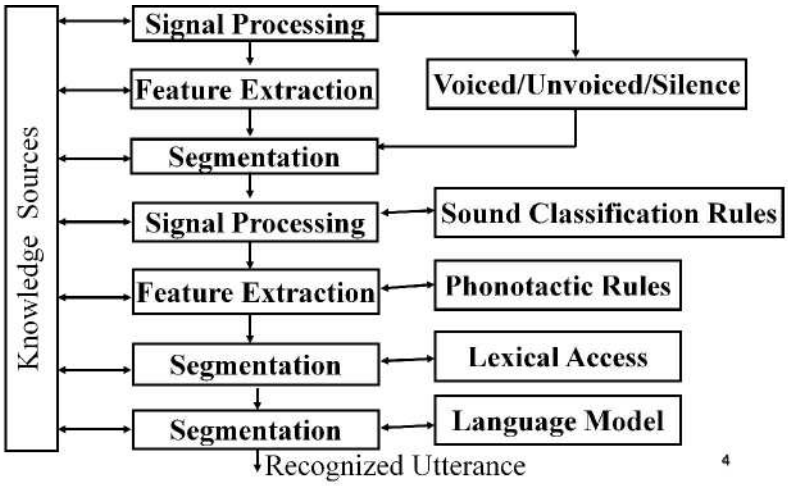
مسئله سنتز گفتار درست عکس این مسئله است. یعنی تبدیل متن به خروجی صوتی.



تصویر 1 – سنتز گفتار در مقابل بازشناسی گفتار

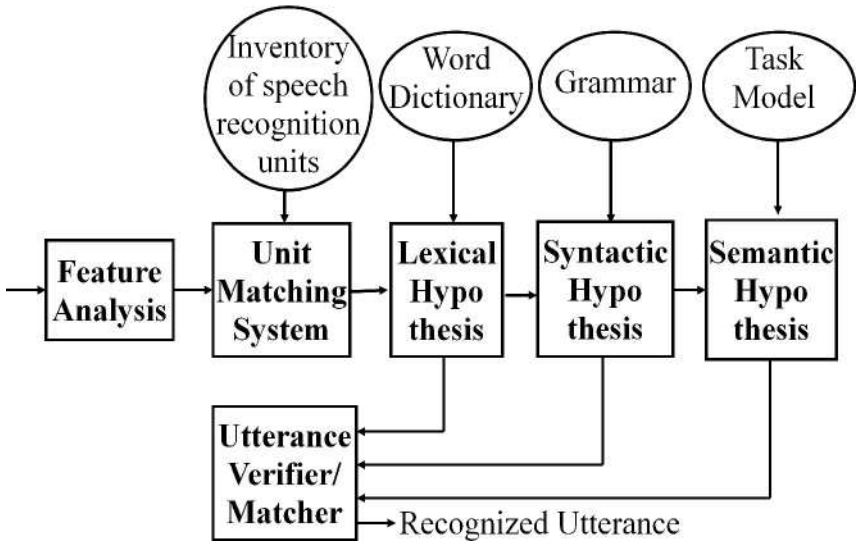
سه روش اصلی برای حل مسئله بازشناسی گفتار ارائه شده است.

- پایین به بالا (bottom-up): پردازش از سیگنال شروع می شود و تا به دست آورن نتایج نهایی له ترتیب ادامه می یابد. در تصویر 2 مراحل این کار شرح داده شده است.



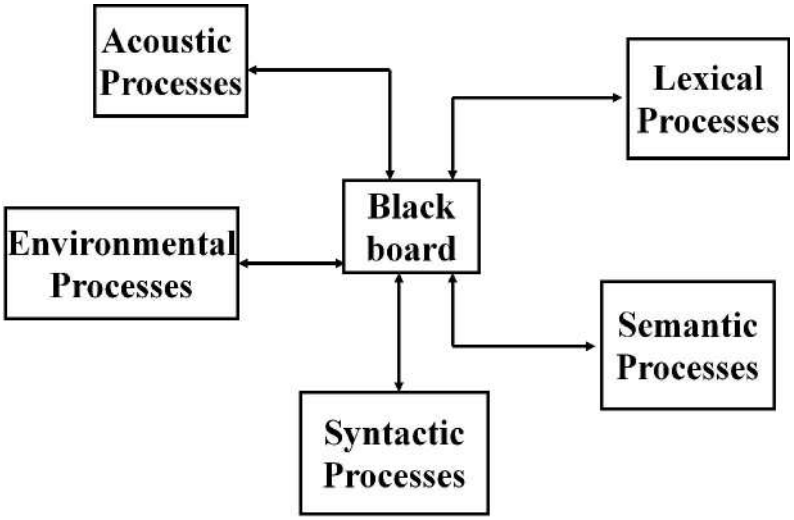
- بالا به پایین (n
  روش را در تصو

  لاصه ای از این

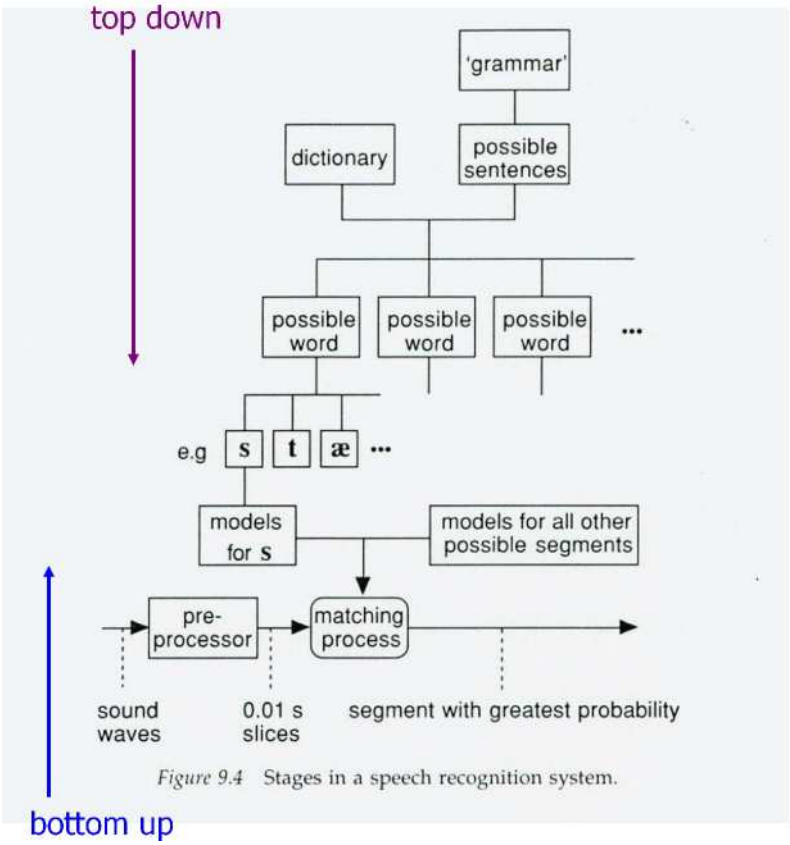


تصویر 3 – روش بالا به پایین برای حل مسئله بازشناسی

- تخته سیاه (blackboard): در این روش بلوک های پردازشی به صورت ترتیبی با هم متصل نیستند بلکه به یک کنترل کننده مرکزی متصل هستند و اطلاعات را به آن می دهند. خلاصه این روش را در تصویر 4 مشاهده می کنید.



- سیستم کلی: یک س



تصویر 5 – سیستم کلی بازشناسی گفتار

چهار تئوری کلی بازشناسی گفتار وجود دارد.

- بازشناسی مبتنی بر مفصل (articulatory-based): استفاده از سیستم مفصلی برای بازشناسی. این تئوری موفق ترین روش تا به حال بوده است.
- بازشناسی مبتنی بر شنوایی: استفاده از سیستم شنوایی برای بازشناسی.
- بازشناسی ترکیبی: ترکیبی از دو روش بالا.
- تئوری موتور: سعی در مدل کردن هدف گوینده دارد.

2– مسئله بازشناسی

مسئله بازشناسی به این صورت است که دنباله ای از نمادهای صوتی داریم و می خواهیم کلماتی که گوینده تلفظ مرده است را از آن ها استخراج کنیم.

یافتن مهتمل ترین دنباله کلمات با داشتن نمادهای صوتی.

- A: دنباله های صوتی
- W: دنباله کلمات
- یافتن که فرمول زیر را بیشینه کند.

$$P(\underline{w}) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1} w_{i-2})$$

$$P(\hat{w} \mid A) = \max_{\underline{w}} P(\underline{w} \mid \underline{A})$$

- قانوین بیزین:

$$P(x \mid y)P(y) = P(x, y)$$

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$$

$$\Rightarrow P(\underline{w} \mid \underline{A}) = \frac{P(\underline{A} \mid \underline{w})P(\underline{w})}{P(\underline{A})}$$

$$P(\hat{\underline{w}} \mid \underline{A}) = \max_{\underline{w}} P(\underline{w} \mid A) \quad = \max_{\underline{w}} \frac{P(\underline{A} \mid \underline{w})P(\underline{w})}{P(\underline{A})}$$

- فرمول نهایی:

$$\hat{w} = Arg \max_{\underline{w}} P(\underline{w} \mid \underline{A})$$

$$= Arg \max_{\underline{w}} P(\underline{A} \mid \underline{w})P(\underline{w})$$

با توجه به فرمول بالا، برای محاسبه محتمل ترین دنباله کلمات، باید مقدار

- P(A|w) و

- P(w)

را محاسبه نماییم.

- **محاسبه P(w)**
  - این مقدار بوسیله «مدل زبانی» محاسبه می شود.
  - مفهوم این مقدار این است که احتمال رخداد دنباله کلمات w در زبان مورد نظر چقدر است. یعنی در زبان مورد نظر به چه احتمالی دنباله کلمات w ظاهر می شود.
  - یک مدل زبانی ساده این است که محاسبه کنیم احتمال رخداد دنباله w=w1....wn چقدر است.
  - به عبارتی

$$\underline{w} = w_1 w_2 w_3 \cdots w_n$$

  - محاسبه احتمال بالا بسیار پیچیده و نیازمند اندازه خیلی زیادی داده است. به همین دلیل از فرم های ساده تر مدل زبانی به نام bigram و trigram استفاده می شود.
  - Monogram:

$$P(\underline{w}) \approx \prod_{i=1}^{n} P(w_i)$$

  - Bigram:

$$P(\underline{w}) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1})$$

  - Trigram:

$$P(\underline{w}) = \prod_{i=1} P(w_i \mid w_{i-1} w_{i-2} \cdots w_1)$$

  - روش محاسبه P(w3|w2w1):

$$P(w_3 \mid w_2 w_1) = \frac{\text{Number of happening W3 after W1W2}}{\text{Total number of happening W1W2}}$$

  - روش محاسبه adHoc:

$$P(w_3 \mid w_2 w_1) \approx \lambda_1 f(w_3 \mid w_2 w_1) + \lambda_2 f(w_3 \mid w_2) + \lambda_3 f(w_3)$$

- **محاسبه P(A|w)**
- چهار روش DTW، مدل مخفی مارکوف، شبکه های عصبی و سیستم های ترکیبی

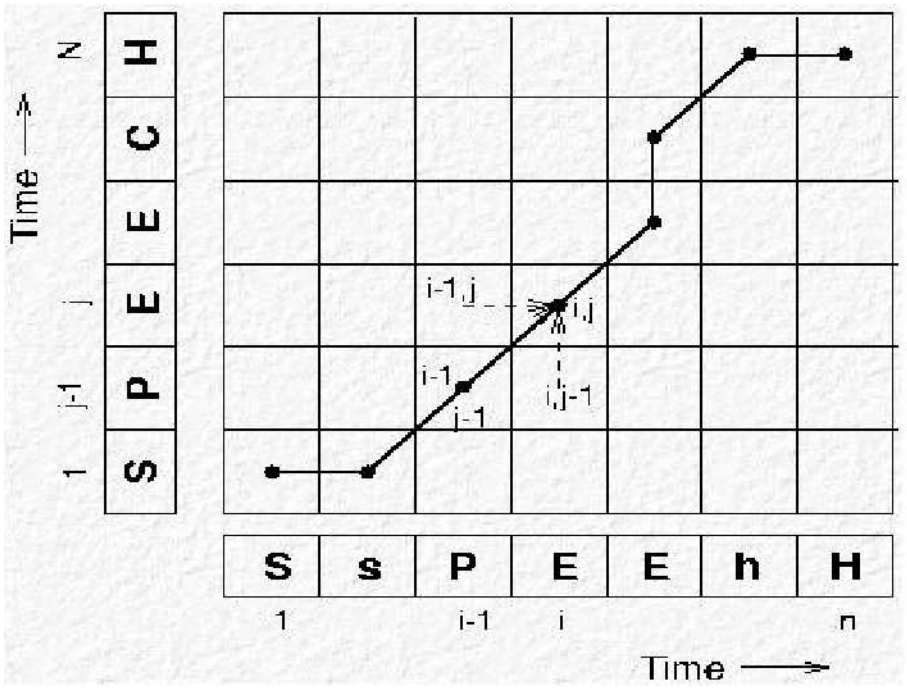
1. Dynamic Time Warping

در این روش سعی می شود که فاصله کلی بین دو نمونه گفتار محاسبه شود.

به عبارتی یک فایل صوتی چه فاصله ای با یک فایل صوتی دیگر دارد.

برای فایل های صوتی نیاز به یک تطبیق زمانی است.

مثال: تطبیق زمانی بین دو کلمه تلفظ شده speech به روش DTW در تصویر 6 نشان داده شده است.

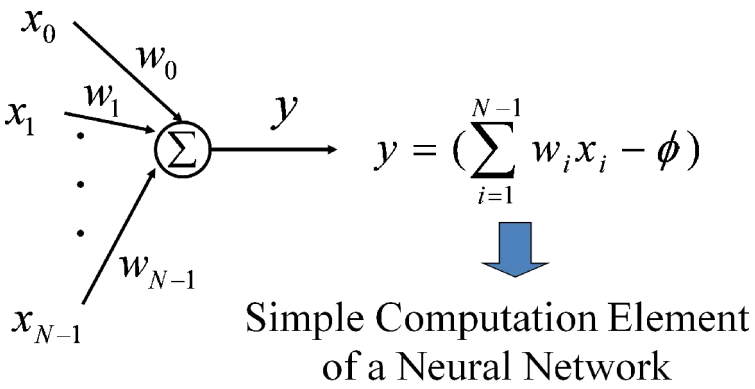


تصویر 6 – انجام تطبیق زمانی بین دو تلفظ speech

2. Artificial Neural Network

یکی از روش های محاسبه P(A|w) استفاده از شبکه عصبی است.

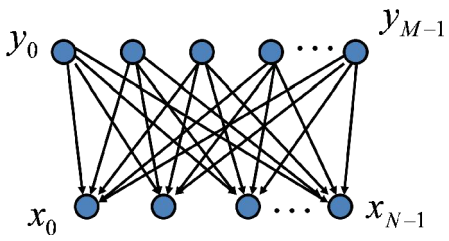
پایه ای ترین واحد شبکه عصبی نرون گفته می شود. یک نرون در شکل 7 نشان داده شده است.



شبکه عصبی نوع های ،

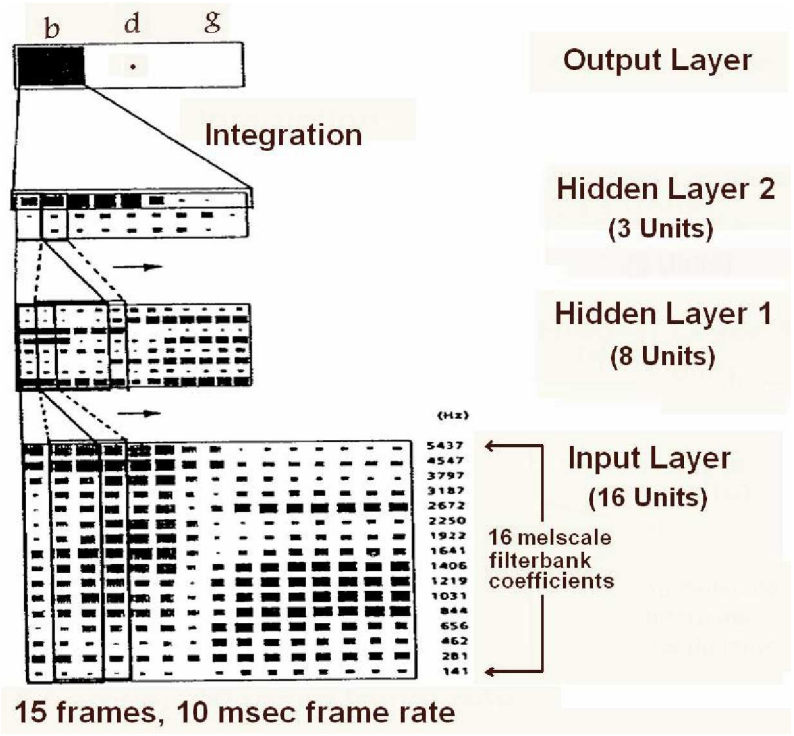
- پرسپترون تک

Single Layer Perceptron



تصویر 8 – پرسپترون تک لایه

- شبکه عصبی با تاخیر زمانی (TDNN): در تصویر 9 یک شبکه عصبی تاخیر دار بازشناسی واج های ب، د و گ ارائه شده است.



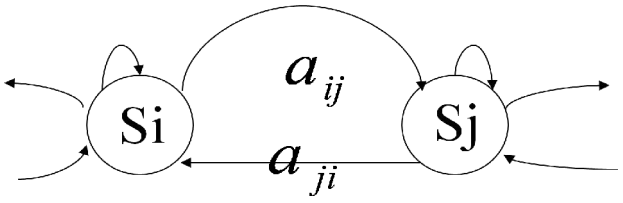
تصویر 9 – یک شبکه عصبی TDNN برای بازشناسی ب، د و گ

3. Hidden Markov Models

مدل های مخفی مارکوف بیشترین استفاده را برای محاسبه P(A|w) استفاده می شود.

این مباحث در فصول آینده توضیح داده شده اند.

یک مدل مخفی ساده را در تصویر 10 مشاهده می کنید.



تصویر 10 – نمونه ای از مدل مخفی مارکوف

مدل مخفی مارکوف دارای یک سری پارامترها است.

- مشاهدات (observation)
- حالات (states)
- پرش بین حالات
- احتمال تولید یک مشاهد در یک حالت

3– حالات مختلف بازشناسی

- بازشناسی کلمات **گسسته** (Isolated Word Recognition) در مقابل بازشناسی گفتار **پیوسته** (Continous Speech Recognition)
- وابسته به گوینده و مستقل از گوینده
- اندازه فرهنگ لغت
  - کوچک (کمتر از 100 لغت)
  - متوسط (بین 100 تا 1000 لغت)
  - بزرگ (بین 1000 تا 10000 لغت)
  - خیلی بزرگ (بزرگتر از 10000 لغت)
- فاکتورهای تولید خطا
  - پروزودی (بازشناسی باید مستقل از پروزودی باشد)
  - نویز (باید از نویز جلوگیری کند)
  - باید قابلیت بازشناسی گفتار محاوره ای را داشته باشد.

4– خودآزمایی

شماره سوال	نوع سوال	صورت سؤال	متن گزینه ها	پاسخ درست	مهلت پاسخگویی	اجازه عبور به قسمتهای بعدی در صورت اشتباه بودن پاسخ
1	■ تستی □ جاخالی	**P(w) نشان دهنده** ..... **است.**		مدل زبانی	1 دقیقه	■ بله □ خیر

5 – خلاصه و نتیجه گیری:

در این فصل با بحث اولیه بازشناسی گفتار آشنا شدیم.

ازجمله با مسئله بازشناسی گفتار به صورت احتمالی آشنا شدیم.

سپس یادگرفتیم که هر قسمت از احتمال ها (P(w) و P(A|w)) را به روش هایی می توان محاسبه کرد.

- P(w) را می توان بوسیله مدل زمانی محاسبه کرد.
- P(A|w) را می توان بوسیله DTW، HMM، شبکه عصبی و روش های ترکیبی این ها محاسبه نمود.

6 – منابع درس:

1- Rabiner, “Fundumentals of Speech Recognition”
2- Huang, Acero, “Spoken Language Processing”
3- Deller, “Discrete-time processing of speech signals”