1– مقدمه

تبدیل متن به گفتار

- انتخاب واحد

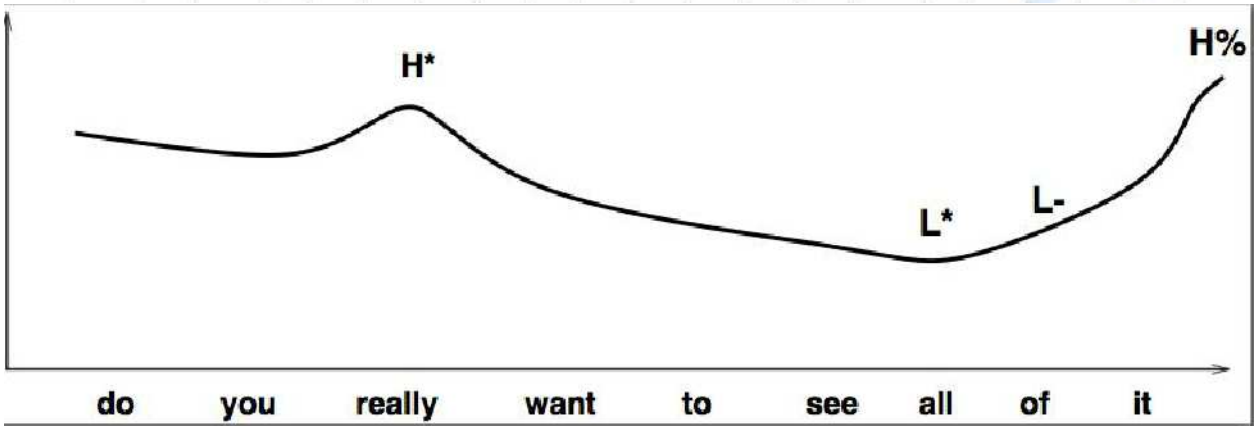
2– روش انتخاب واحد (unit selection)

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.

				H*										L*		L- H%		
do		you		really				want				to		see		all		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

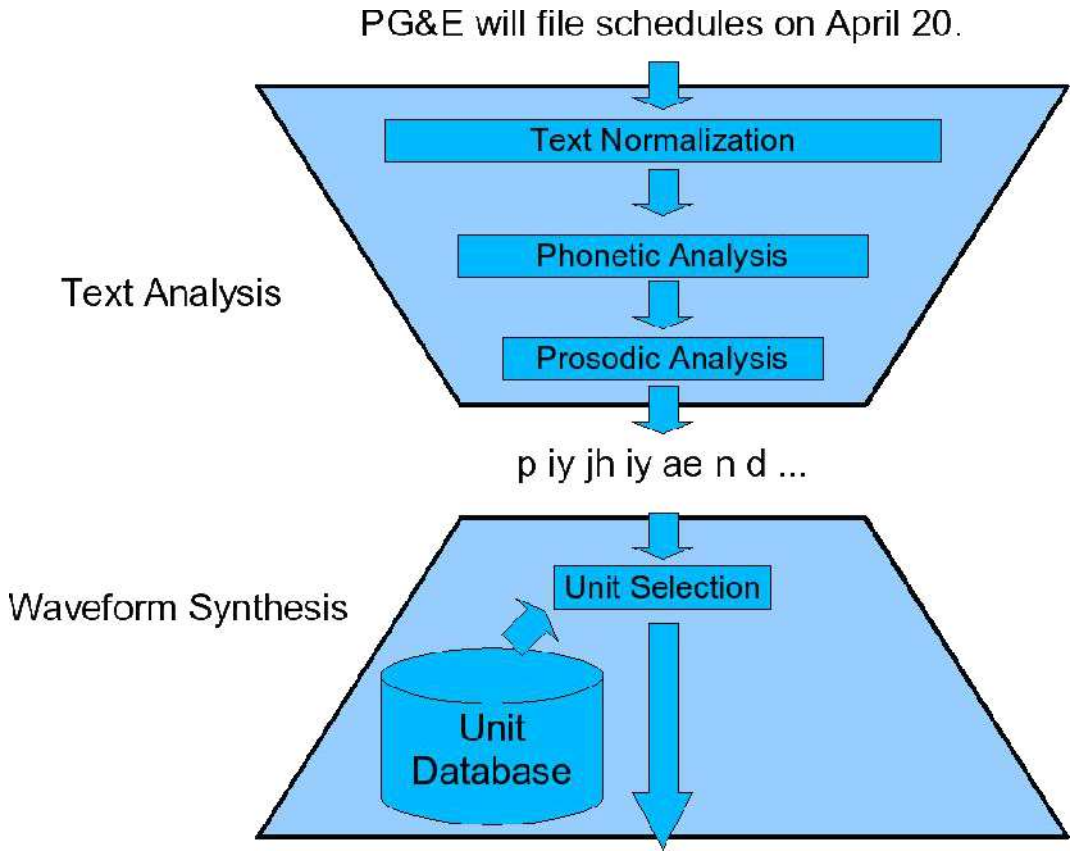


تصویر 1 – ورودی برای تولید شکل موج

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 انختاب واحد
  - هزینه هدف
  - هزینه الحاق
- الحاق شکل موج ها
  - ساده
  - PSOLA

ساختار کلی روش الحاقی را در تصویر 2 مشاهده می کنید.



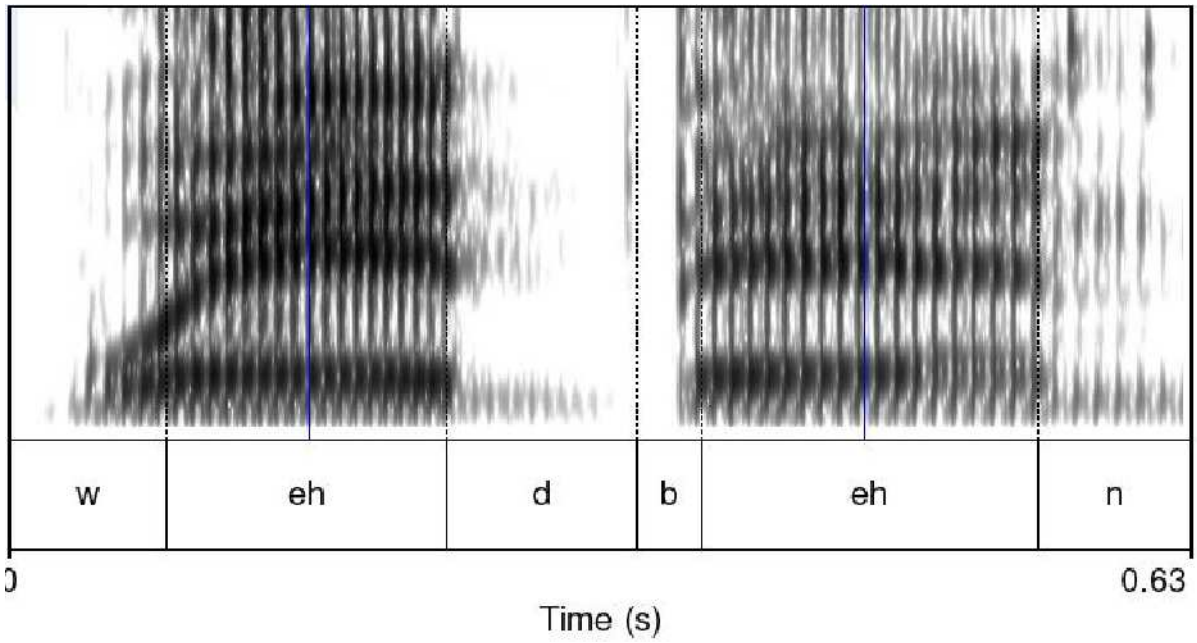
تصویر 2 – ساختار کلی سیستم الحاقی

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.

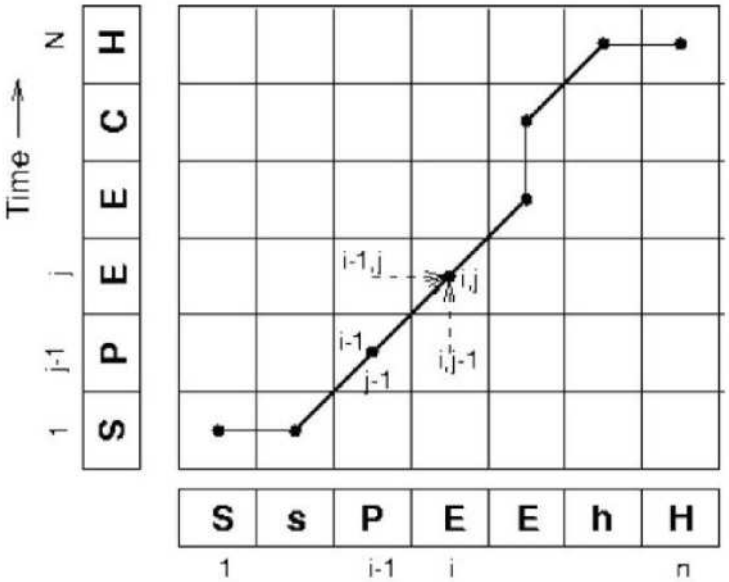
سئوال این است مه آیا می توان این قسمت را به صورت خودکار یافت؟

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:

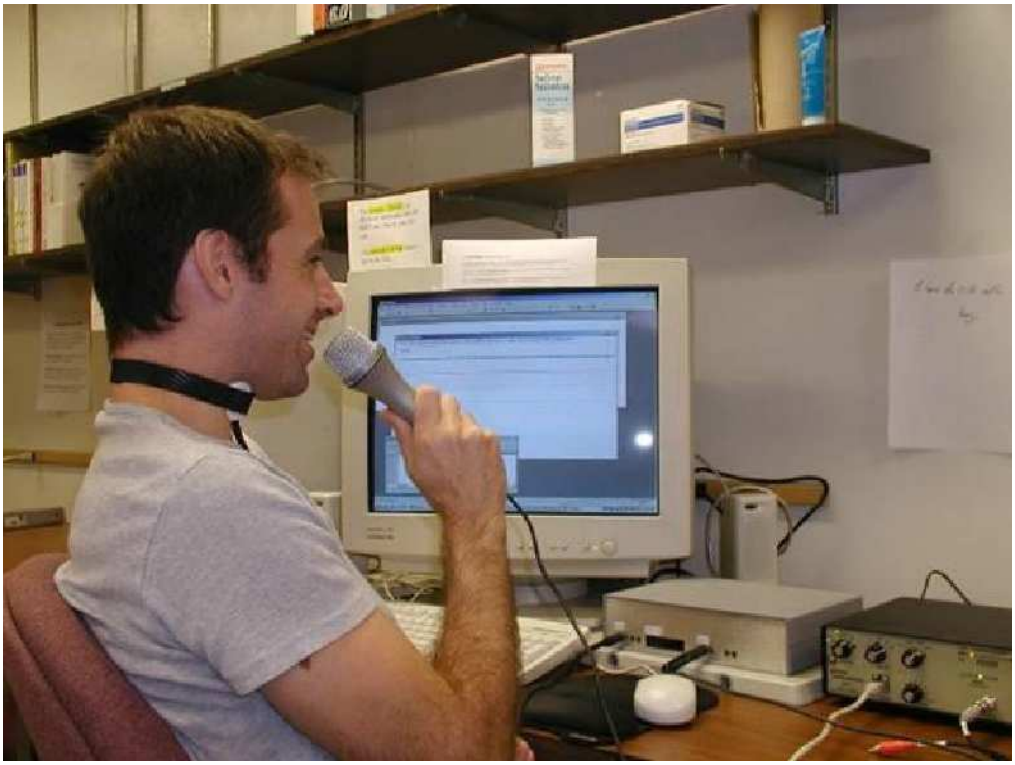
- برای انفجاری ها: یک سوم داخل
- برای واج–سکوت ها: یک چهارم داخل
- برای بقیه دایفون ها: 50 درصد داخل

در هنگام سنتز باید شکل موج ها را به هم بچسبانیم.

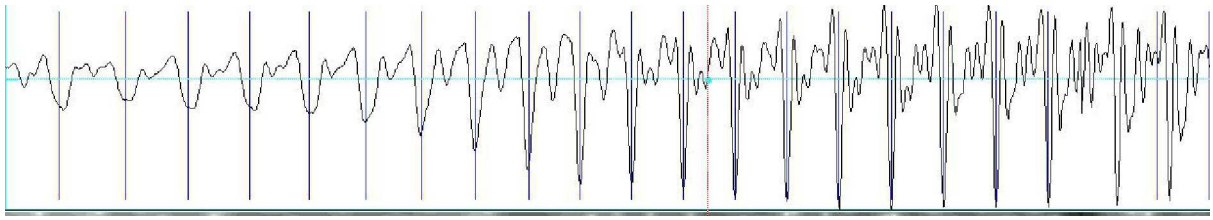
همچنین نیاز است فرکانس گام نمونه های دایفون را تغییر دهیم. برای این کار نیاز است مکان رخداد گام در سیگنال مشخص شود. یعنی باید زمان بسته شده تارهای صوتی مشخص شود.

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