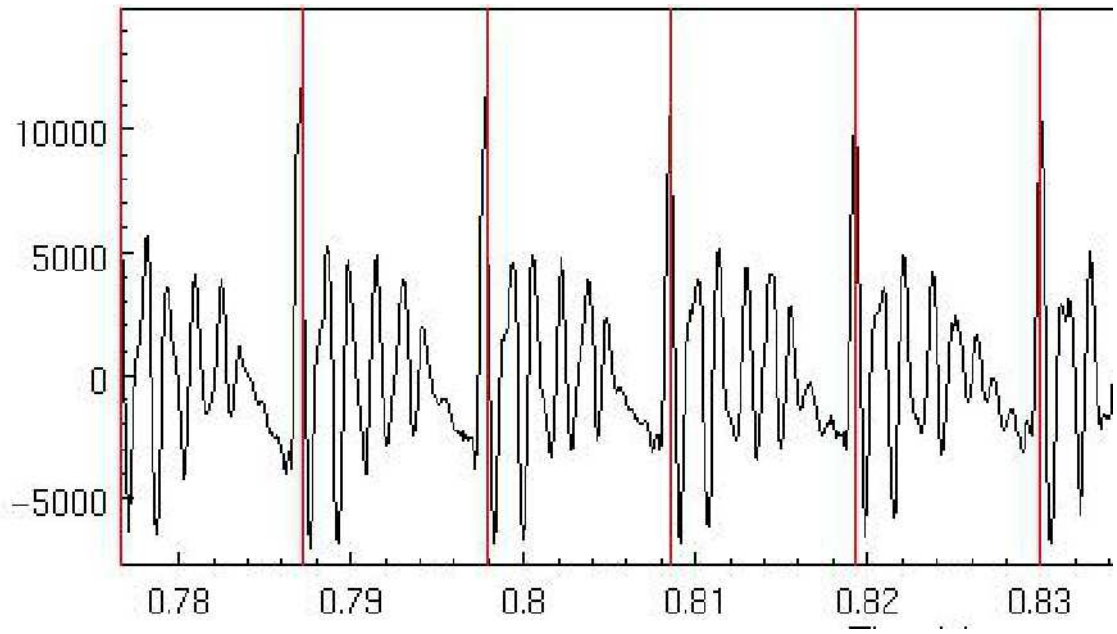
برای زیر و بم کردن یک نمونه گفتار باید فرکانس گام تغییر کند ولی مدت زمان ثابت بماند.

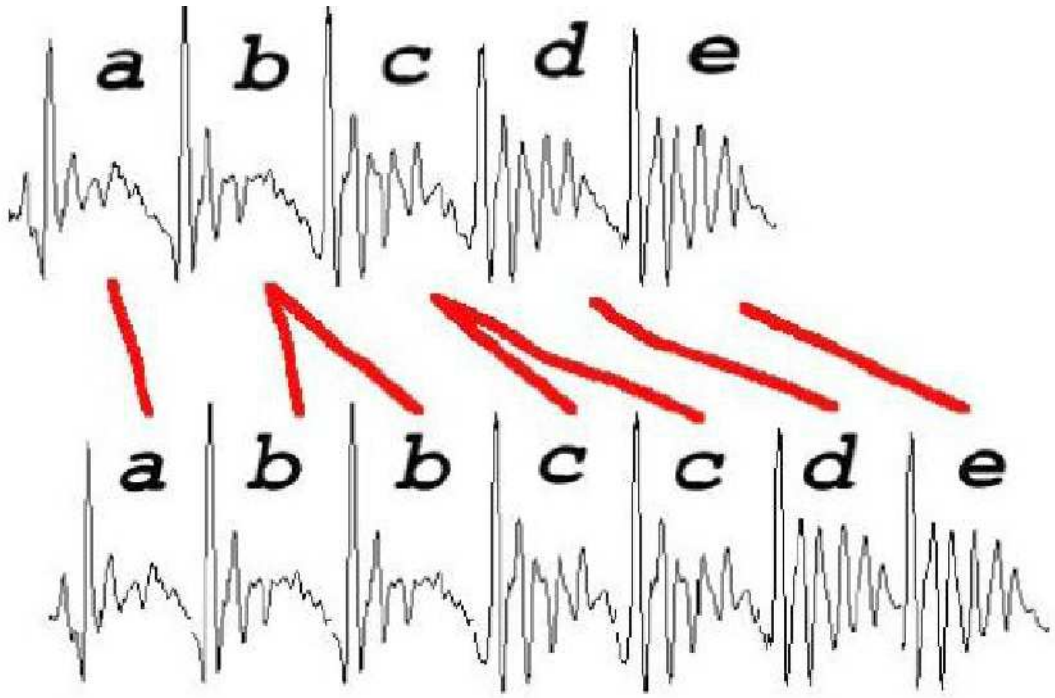
زیاد کردن فرکانس نمونه برداری باعث گفتار زیر تر می شود ولی مدت زمان سیگنال نیز کم می شود.

فرض کنید سیگنال تصویر 7 را داریم.



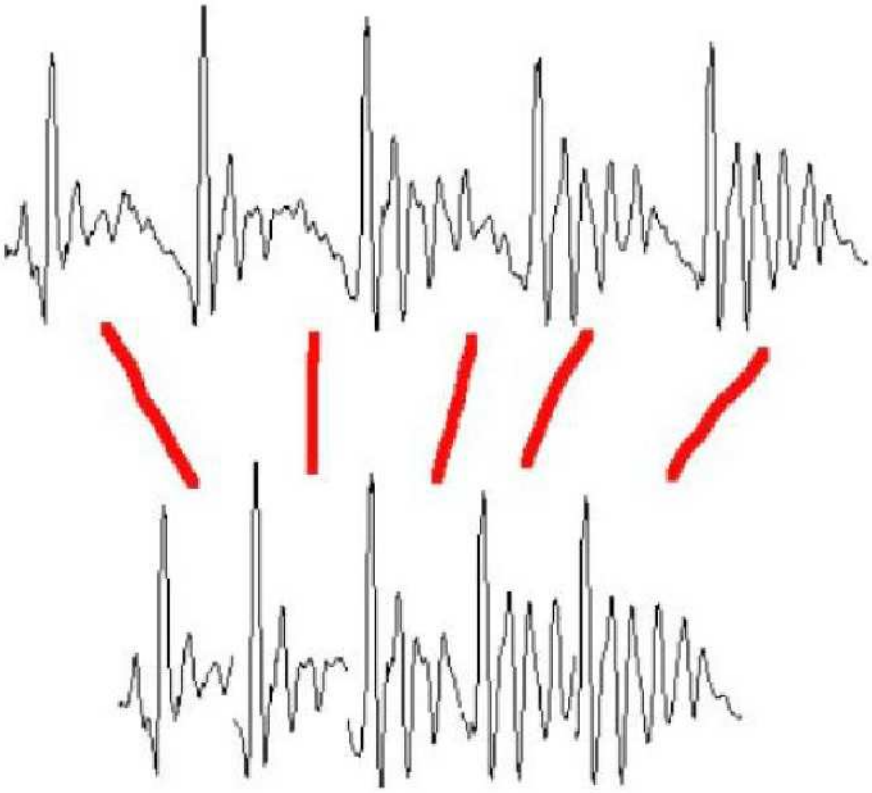
تصویر 7 – سیگنال صدادار نمونه

در تصویر 8 چگونگی تغییر طول مدت زمان تلفظ را مشاهده می کنید.



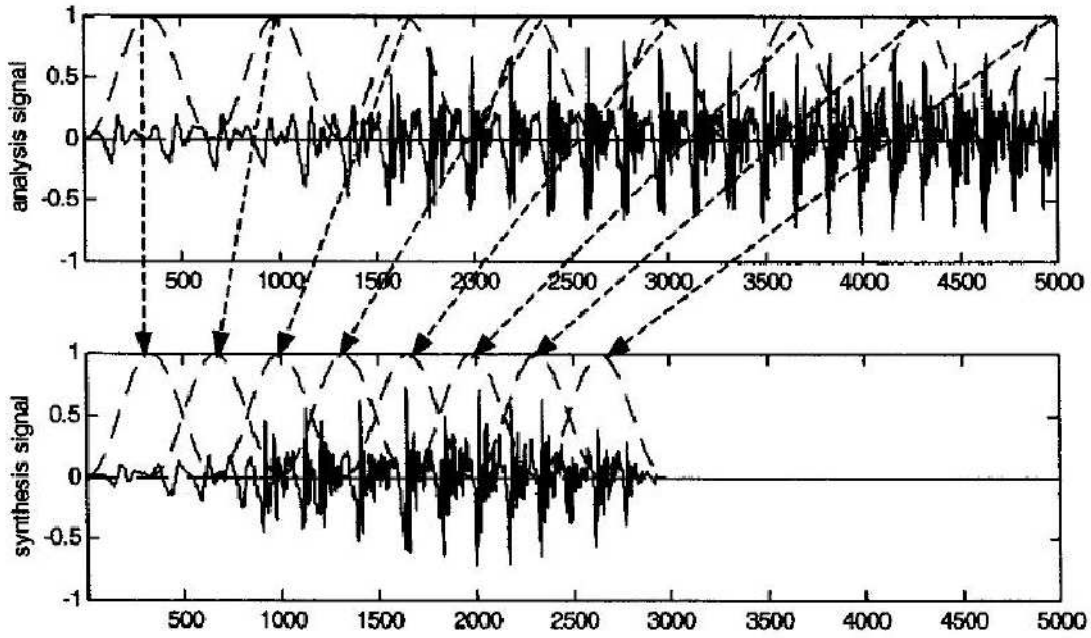
تصویر 8 – تغییر مدت زمان تلفظ

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.



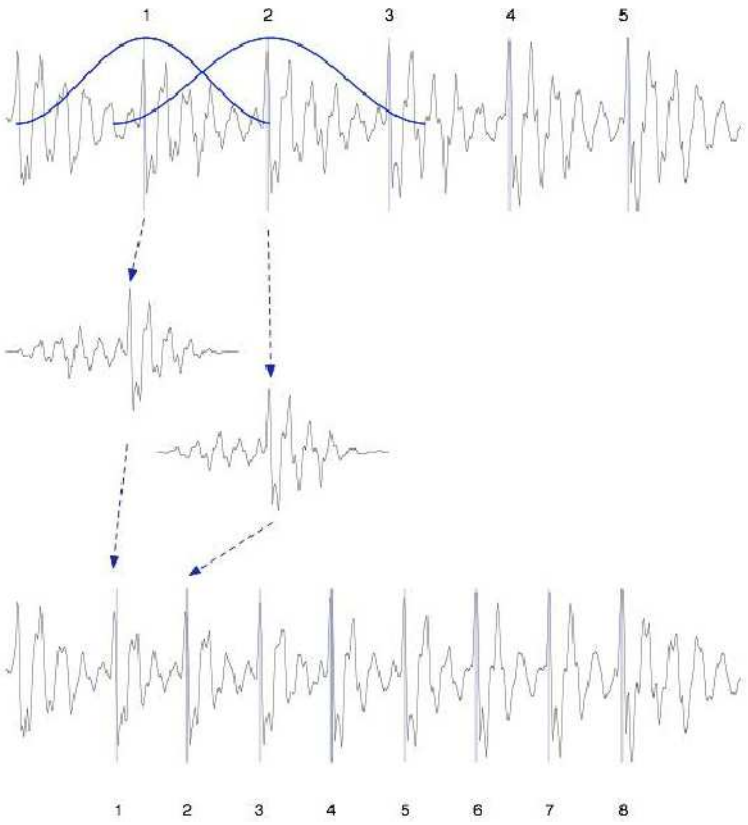
تصویر 10 – روش overlap-add

روش PSOLA به این صورت است که این عمل را بر روی بسته شدن های حنجره انجام می دهد.

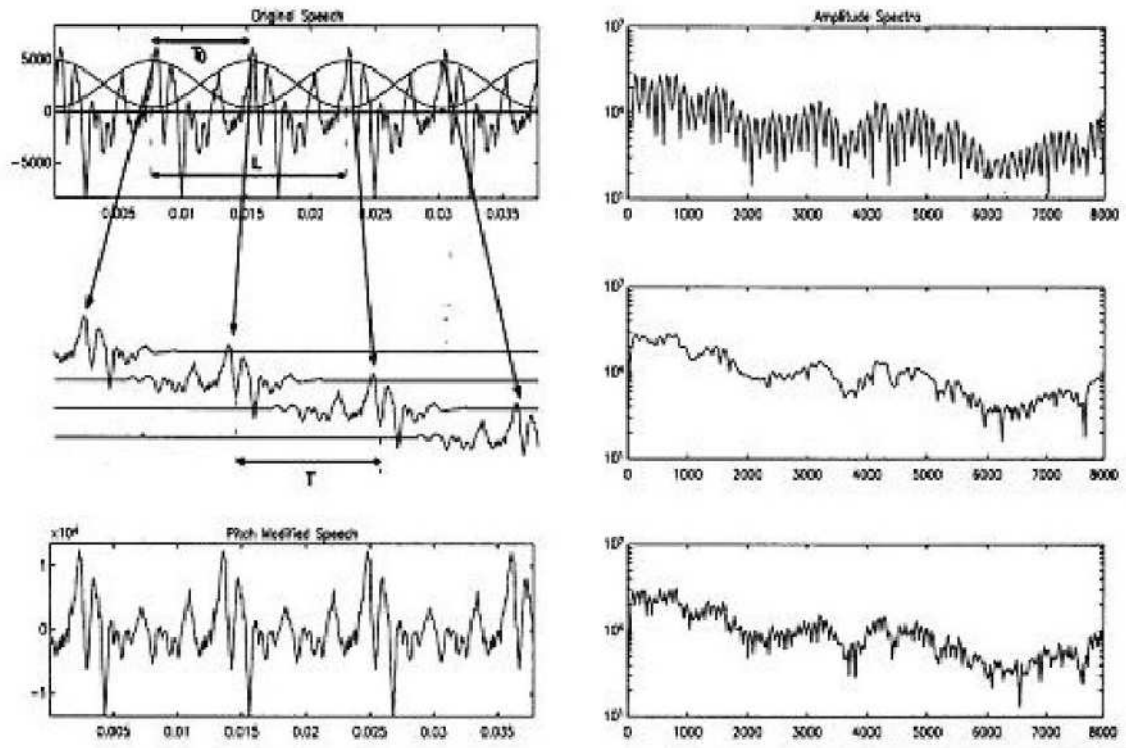
یعنی فرکانس گام را محاسبه می کند و روی رخداد های گام پنجره می زند.

- 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

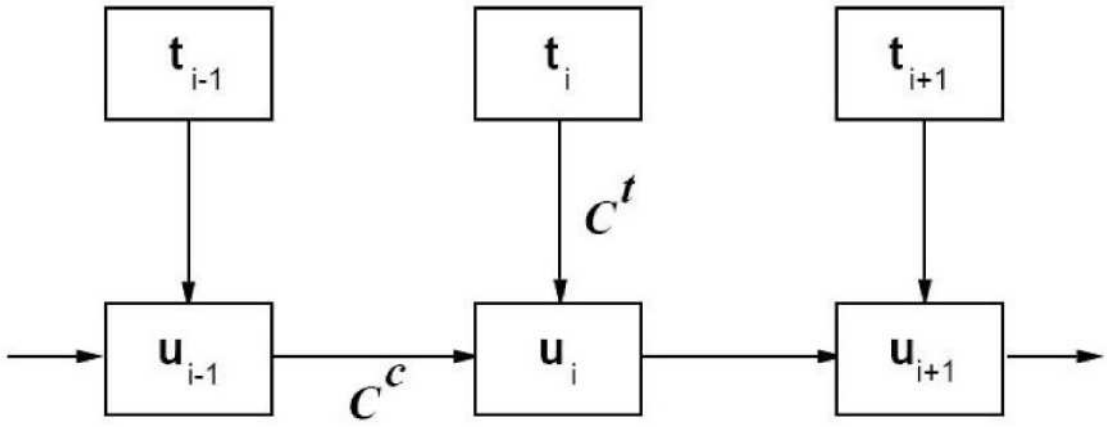
در نهایت هزینه نهایی به صورت جمع دو هزینه هدف و الحاق محاسبه می شود.

فرمول 4

$$C(t_1^n,u_1^n)=\sum_{i=1}^{n} C^{target}(t_i,u_i)+\sum_{i=2}^{n} C^{join}(u_{i-1},u_i)$$

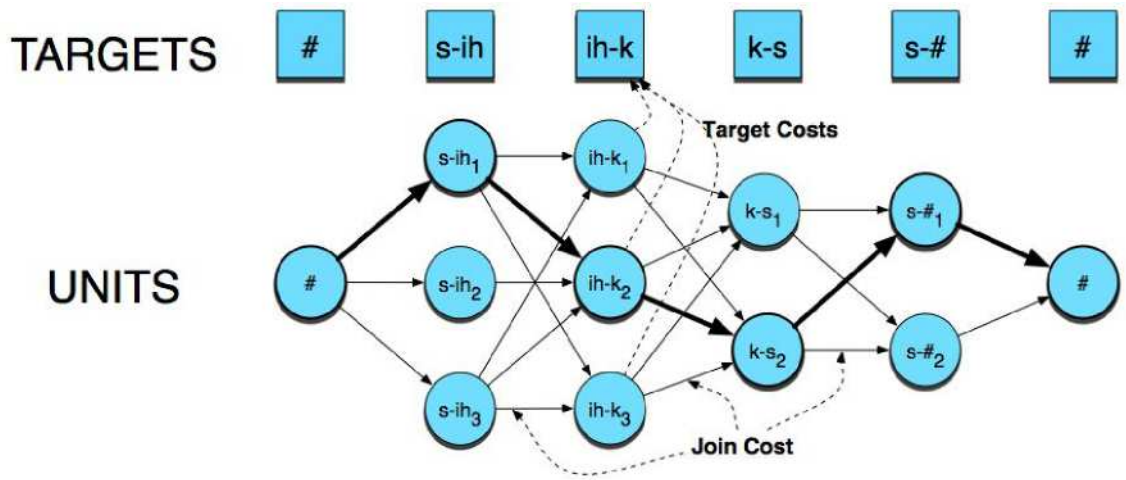
مسئله یافتن مسیری است که فرمول 4 را مینیمم می کند.

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”