《主 題》

저 전송속도 음성부호화를 위한 디지탈 음성처리 신호처리기술

이 황 수
(한국과학기술원 정보및 통신공학과)

■ 차 례 ■

Ⅰ. 서 론

음성은 인간과 인간사이의 가장 자연스러운 의사전달, 즉 통신수단이다. 따라서 어떻게 하면 멀리 떨어져 있는 두 지점사이의 음성통신을 효율적으로 할 수 있느냐가 전화가 발명된 이래 음성 신호처리에 있어서 주된 관심사로 연구되고 있다. 이러한 연구는 최근들어 급속한 발전을 보이고 있는 반도체 및 컴퓨터 기술에 힘입어 양자화된 음성신호, 즉 디지탈 음성신호처리 분야에 집중되고 있으며, 인간과 인간사이만이 아닌 인간과 컴퓨터, 다시 말하면 인간과 기계사이의 통신에서도 음성을 이용하려는 방향으로 진전되고 있다.

디지탈 음성신호처리를 위하여 우선 수행되어야 할 과제는 연속적인 음성신호를 어떻게 효율적으로 부호화하여 디지탈 형태로 변환하느냐하는 음성정보 부호화 및 음성정보 압축에 대한 문제이다.

음성정보 압축에 대한 요구는 음성의 디지탈 전송을 시작하면서 부터 생겨난 것으로 대역폭이 제한된 통신 채널을 효율적으로 사용하기 위하여 필수적이라 할 수 있다. 상용전화통신을 위한 디지탈 음성부호화 방식의 표준화 추세를 보더라도 이와 같은 경향을 알 수 있는데 현재까지 국제 표준으로 가장 널리 사용되고 있는 음성부호화 방식은 대수적으로 신호를 변환하여 양자화하는 μ-law 또는 A-law PCM 방식으로 전송속도는 64kbps이다. 그러나 디지탈 음성처리 기술의 발달로 ADPCM 방식이 개발되어 국제표준방식으로 채택이 되면서 전송속도를 32kbps로 줄일 수 있어서 기존 PCM 채널의 효율을 높일 수 있게 되었다. 현재에는 그보다 더 전송속도가 낮은 16kbps에서 음성을 부호화하여 전송하려는 연구가 활발히 행하여지고 있으며 SBC, ATC, APC, RELP 등의 여러가지 부호화 방식들에 대한 성능이 비교되고 있으며 곧 국제 표준화가 이루어지리라 생각된다.

그러나 앞에서 언급한 국제간 전화통화나 국내 전화통신등과 같은 상용전화통신이외의 경우에는 음질은 좀 낮지만 16kbps 이하의 낮은 전송속도로 효율적인 음성통신을 할 수 있는 부호화 방식에 대한 필요성이 증대되고 있다. 따라서 음성의 해석 및 합성기술에 의한 부호화 방식이 활발히 연구가 높아지고 있다. 이와같은 목적으로 9.6, 4.8, 2.4kbps의 디지탈 음성부호화 방식이 연구되고 있으며 만약 2.4kbps 음성 부호화기를 사용할 경우 한 9.6kbps 회선을 통신량에 따라 음성, FAX, 데이타등의 전송에 복합적으로 분배 사용할 수 있어 회선 사용의 효율을 높일 수 있게 된다. 이러한 저 전송속도의 음성부호화 기술로는 선형예측

부호화(LPC)기술이 대표적이라 할 수 있다.

최근에 들어와서 음성부호화에 대한 향상된 성능의 알고리즘들이 개발되고 또한 VLSI 기술의 발달로 고성능 디지탈 신호처리기가 출현함에 따라 종래에는 실시간 실현이 불가능하다고 여겨졌던 음성부호화방식들이 구현 가능해 졌다. 이러한 저 전송속도 디지탈 음성부호화방식의 주요 응용분야로는 디지탈 이동통신 (8kbps 이하의 VSELP 또는 QCELP 등), 전화선을 이용한 비화통신 (DOD의 4.8kbps CELP 및 4kbps LPC 표준)등을 들 수 있으며 디지탈 통신의 확대와 더불어 그 응용분야는 더욱 넓어질 전망이다.

본 고에서는 앞에서 살펴본 음성부호화 방식을 저 전송속도(9.6kbps 이하)와 중, 고 전송속도(16kbps 이상)로 분류하고 저 전송속도 음성부호화방식의 기본이 되는 신호처리 방식에 대하여 설명하고자 한다. 일반적으로 높은 전송속도의 음성부호화 방식에는 음성의 파형을 충실히 재현시키고자하는 파형부호화(Waveform Coding) 방식을 주로 사용하고 있다. 반면에 낮은 전송속도의 음성부호화 방식으로는 음성의 발성 모델을 가정하고 그 발성 모델을 특징지워주는 계수들을 부호화하는 파원 부호화(Source Coding) 방식이 주로 사용된다. 이와같은 저 전송속도의 음성부호화 방식을 일반적으로 vocoding 방식이라하며 vocoding 방식에 의한 음성부호화기를 vocoder라 한다.

Ⅱ. Vocoding에 의한 음성 부호화 방식

전화와 같은 정도의 음질을 갖는 음성신호의 디지탈 부호화 방식으로는 파형 부호화 방식인 PCM과 ADPCM을 들 수 있는데 이 경우 전송속도는 64kbps 또는 32kbps가 된다. 이것은 음성신호에 대한 이론적인 고찰을 통하여 볼 때 전달되는 정보량에 비하여 너무 많은 데이타가 전송되는 것이다. 따라서 음성신호의 정보량과 사람의 청각특성을 이용한 저 전송속도의 디지탈 음성전송방식에 대한 연구가 음성처리 연구의 중요한 분야가 되어있다. 특히 시간에 따라 변화하는 음성의 주파수 스펙트럼을 분석, 전송하고 수신측에서는 이를 이용하여 음성을 합성하게 되면 전송속도를 수천 bps 이하로 낮출 수가 있다.

Vocoding에 의한 음성부호화의 연구에 있어서 가장 중요한 과제로는 음성신호에서 부터 주파수 스펙트럼과, 여기 신호의 주기를 정확히 추출하는 것을 들 수 있다. 그러한 목적으로 1960년대 이전에는 아나로그 대역 필터를 사용하여 주파수 스펙트럼을 분석하고 비선형 아나로그 처리방법으로 피치 주기를 추출하려는 연구가 행하여졌다. 그러나 이와같은 아나로그처리 방법으로는 정확한 음성의 특징 추출이 어려웠으나 60년대 이후 디지탈 신호처리 기술의 발전으로 analysis-by-synthesis, Formant 해석, Cepstrum 해석 방식들이 도입되어 피치주기 추출 및 스펙트럼의 envelope 추출 방법 등이 음성의해석, 합성시스템에 이용되기 시작하였다. 그 중에서도 특히 주목할 만한 것은 1966년 음성스펙트럼을 AR process로 가정한 all-pole 모델을 기초로한 maximum likelihood 스펙트럼 추정 방식으로 현재 선형예측 부호화(LPC, Linear Predictive Coding)방식이라고 불리고 있으며 저 전송속도의 음성부호화, 음성인식 및 합성에 널리 응용되고 있다.

2.1 Vocoding에 의한 음성 부호화의 원리

시간에 대한 음성 신호파형을 관찰해 보면 신호의 성질이 시간에 따라 서서히 변화하는 것을 알 수 있다. 예를 들면 10ms 정도의 짧은 시간 구간의 음성신호는 그 특성의 변화가 적어 stationary한 신호로 볼 수 있다. 따라서 짧은 구간의 신호는 그 pitch 주기와 평균전력 및 스펙트럼 envelope의 평균치로 특징 지워질 수 있다. 이와같은 짧은 구간에서의 음성을 해석하여 그 특성 계수들을 추출한 후 이를 부호화하여 전송하고, 수신측에서는 전송된 pitch주기와 에너지 정보를 이용하여 펄스를 발생시켜 음성의 여기신호로 사용한다. 그러나 입력음성의 주기성이 존재하지

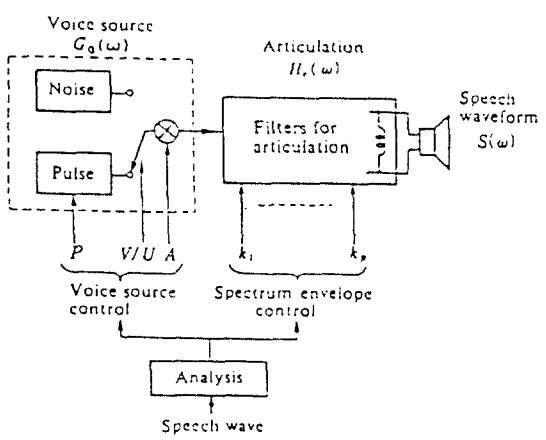


그림 1. 음성발생 모델

않을 경우(무성음의 경우)에는 백색잡음을 여기신호로 사용하게 된다. 이 여기 신호들은 스펙트럼 envelope를 형성하여 주는 필터를 지나게 되고 그 결과 원래 입력음성신호의 특성이 보위된 출력 음성 신호를 내게 된다. 이와같은 음성발생 모델이 그림 1에 그려져 있다.

위와 같은 음성부호화과정에서 저 전송속도의 부호화가 가능한 이유는 다음과 같다. 첫째로, 음성의 해석단계에서 음성파형의 위상에 대한 정보를 무시한다. 이는 음성파형의 위상정보가 실제로 음성을 청취하여 이해하는데 큰 영향을 미치지 못하기 때문이다. 둘째로는, 추출하려는 특성계수의 시간적 변화량이 작다고 가정하고 10-20ms 정도의 간격으로 표본화를 하는데 있다. 셋째로는, 스펙트럼 envelope 정보의 부호화에는 10개 정도의 계수로 충분하고 각 계수당 부호화율도 평균 4 bit 정도로 적기 때문이다.

2.2 음성 발생 모델

음성은 성대에서의 여기 신호가 성도를 거치면서 발생되는 것으로 음성의 종류에 따라 다음 두가지로 나눌 수 있다.

첫째로는 유성음 발생에 대한 것으로 여기신호는 성대의 진동으로 변조된 공기의 흐름으로 볼 수 있다. 이 여기신호는 pitch주기에 따라 주기적이므로 그 스펙트럼은 신호의 주기성에 의한 harmonics를 보이게 된다.

두번째는 무성음 발생에 대한 것으로 성도의 특정 부분에서 공기의 흐름에 대한 constriction이 일어나고 이로인한 air turbulence가 여기 신호가 되며 이 여기 신호는 잡음과 같은 성질을 갖게 된다.

성도는 공진기와 같은 역할을 함으로써 여기신호의 스펙트럼을 변형시키게 되는데 유성음의 경우 대략 4개 정도의 공진점을 찾을 수 있고 이들을 formant라 부른다. 이 formant들의 음을 특성지워주는 parameter로서 그림 2에 대표적인 유성음과 무성음의 스펙트럼을 보였다.

유성음의 스펙트럼을 관찰하면 주기적인 여기신호의 스펙트럼과 formant 정보를 볼 수 있다. 만약 nasal tract가 유성음 발생시에 일어나게 되면 전체 전달함수에는 antiformant로 작용하는 zero점이 생기게 된다.

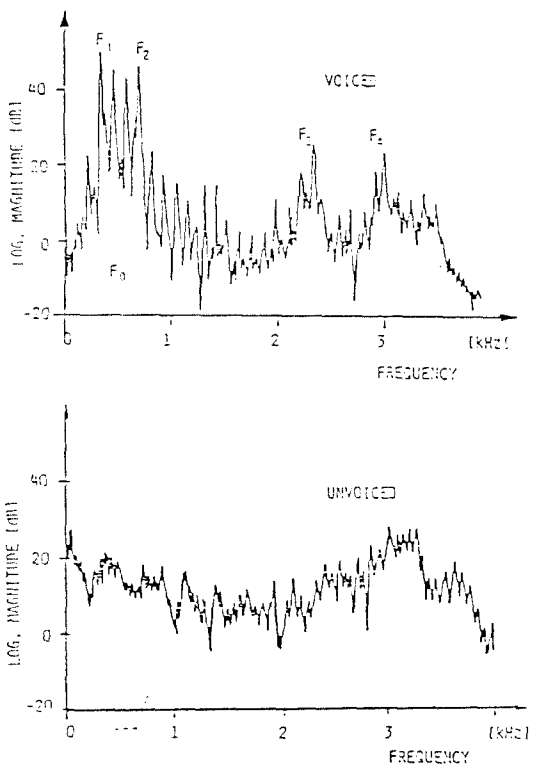


그림 2. 유성음과 무성음의 스펙트럼 비교

III. 음성의 해석및 합성

음성발생 모델에 기초한 vocoder의 기본구성이 그림 3에 나타나 있다.

이 그림에서 보듯이 음성해석의 기능은 기본 pitch 주기 P_0의 검출과 스펙트럼 envelope의 추출로 나눌 수 있고 음성 합성의 기능은 여기신호의 발생과 이 신호를 이용 스펙트럼 envelope를 다시 발음이내는 것으로 생각할 수 있다.

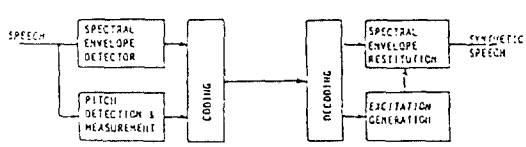


그림 3. 음성해석및 합성의 기본 구성

3.1 Short-time 음성 해석

음성신호의 특성은 시간에 따라 서서히 변화하기 때문에 음성신호를 작은 구간의 frame들로 나누어 각 frame내에서는 동일한 특성을 갖는 것으로 간주하여 해석하는 short-time 음성해석 방법등이 사용된다. 음성의 frame들은 window 함수 w(n)을 입력신호에 곱하여 얻게 된다. 그러므로 시간 n=i 부근 frame에서의 음성신호는

$$x_i(n) = x(n) \cdot w(i - n)$$

로 표시된다.

Short-time 특성계수의 예로 short-time 영교차율은 다음과 같이 정의된다.

$$Z(i) = \sum_{n=-\infty}^{+\infty} c(n)w(i - n)$$

$$c(n) = \begin{cases} 1 & \text{if } sign(x(n)) \neq sign(n-1)) \\ 0 & otherwise \end{cases}$$

3.2 Short-time Fourier 해석

Short-time Fourier 변환은 재한된 시간축상의 sample sequence에 대한 일반적인 Fourier 변환으로 볼 수 있다. Window w(n)에 있어 $n = 1,2,\cdots,N-1$ 까지만 값을 갖고 그 이외에서는 0인 경우 short-time Fourier 변환은

$$X_i(k) = \sum_{n=-\infty}^{+\infty} w(i-n)x(n)exp(-j2\pi kn/N),$$
$$k = 0, 1, \ldots, N-1$$

으로 표시된다. 여기에서 N은 window w(k)의 최대길이이며 위 변환은 discrete 주파수 k와 window의 위치 i에 대한 함수이다. 위 식은 신호 x(n)을 시간 i에 위치한 impulse 응답이 w(n)인 window를 통하여 관찰한 sample sequence w(i - n)x(n)의 DFT로 볼 수 있다.

다른 해석 방법으로는

$$X_i(k) = [x(n)exp(-j2\pi kn/N)] * w(n)$$

으로 나타낼 수 있으며 이것은 신호 x(n)을 $exp(-j2\pi kn/N)$으로 변조한 후 impulse 응답이 w(n)인 filter를 통과시킨 것으로 볼 수 있다.

위의 식과 동일한 결과를 얻도록 위의 식을 재배치하면

$$X(k) = exp(-j2\pi ki/N) \sum_{n=-\infty}^{+\infty} x(i-n)w(n)exp(j2\pi kn/N)$$

가 된다. 이것은 입력신호 x(n)을 impulse 응답이 $w(n)exp(j2\pi kn/N)$인 complex bandpass filter를 통과시킨 후 그 출력을 $exp(-j2\pi kn/N)$으로 변조한 것으로 그 의미는 입력신호 x(n)의 k번째 주파수 성분 $X_i(k)$는 x(n)을 bandpass filtering한 후 modulator를 거치면 얻을 수 있다는 것이다. 다음과 위 식을 이용하여 vocoder를 구성할 경우 k번째 filter의 출력의 크기는

$$|X_i(k)| = |\sum_{n=-\infty}^{+\infty} x(i-n)w(n)exp(j2\pi kn/N)|$$

으로 주어진다. 즉 $|X_i(k)|$는 impulse 응답이 $w(n)exp(j2\pi kn/N)$인 bandpass filter와 RMS 회로가 연결된 시스템에 x(n)을 입력시켰을 때의 출력으로 구할 수 있다. 또한 RMS 회로 대신 envelope 검출기를 연결하였을 때에도 마찬가지 결과를 얻을 수 있다.

Short-time Fourier 변환의 위 세가지 해석방법에 따라 각 방법마다의 window의 역할을 살펴보면 우선 첫째로는 본래 그대로 단순한 window 함수 w(i - n)으로 볼 수 있다. 둘째로는 low-pass impulse 응답으로서의 w(n)의 의미가 포함되어 있고, 셋째로는 $w(n)exp(j2\pi kn/N)$인 band-pass impulse 응답으로 해석할 수 있다.

Ⅳ. Channel Vocoder

입력신호의 주파수 스펙트럼을 여러개의 contiguous한 bandpass filter를 통과시킨 후 그 각각의 출력을 다시 합하면 음성신호를 재생시킬 수 있다. 실제적으로 filter bank를 사용하거나 DFT를 이용하게 되는 데 스펙트럼을 어떻게 나누느냐가 중요한 문제이다. 스펙트럼을 나누는 데 있어서 균일하게 나누는 방법과 불균일하게 나누는 방법을 고려할 수 있는데 일반적으로 인간의 청각특성에 맞도록 주파수가 높아짐에 따라 대역폭을 늘려 잡는다. 이와같은 음성의 해석·합성 시스템을 channel vocoder라 하며 그림 4에 그 원리가 표시되어 있다.

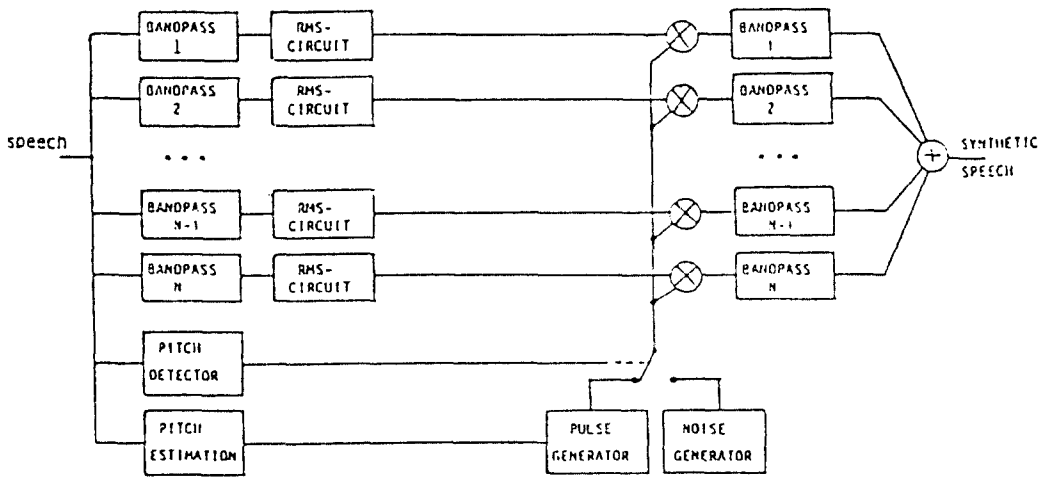


그림 4. Channel vocoder

V. Formant Vocoder

Channel vocoder에서와 같이 음성신호의 주파수 대역을 고정된 contiguous bandpass 영역으로 나누는 대신에 이 방법은 스펙트럼 envelope을 formant들의 위치, 크기, 대역폭으로 나타내려는 것이다. 이로 인하여 음성 data 양을 더욱 줄일 수 있으면서도 그 음질을 높일 수 있다.

우선 formant vocoding 방법의 음성합성 model은 스펙트럼 envelope을 나타내는 time varying filter로 여러 개의 공진기를 연결하여 사용한다. 그러므로 입력 이에 맞는 model 계수를 구하여야하는 데 이 계수들로 얻을 수 있는 스펙트럼과 DFT로 얻은 입력음성의 스펙트럼 envelope이 best match가 되도록 정한다. 예를 들면 유성음의 경우 5 pole 디지탈 필터를 사용하고 1 pole 1 zero 디지탈 필터로 무성음을 modeling할 수 있다. Variable filter를 정하여 주는 parameter로는 유성음의 F_1, F_2, F_3에 해당하는 pole 위치와 무성음 부분의 pole과 zero의 위치 F_p와 F_z이다.

음성해석의 방법으로는 short time 스펙트럼에서 peak점 추출방법으로 처음 3개의 최대치를 직접 찾아내는 방법이 있다. 이 경우 spectral resolution을 실용적으로 되도록 하려면 30-50개 이상의 bandpass channel이 요구되며 스펙트럼에서 pitch 주기의 structure를 제거하기 위한 smoothing이 필요하다. Smoothing의 방법으로 cepstrum을 이용한 pitch신호의 제거 방법이

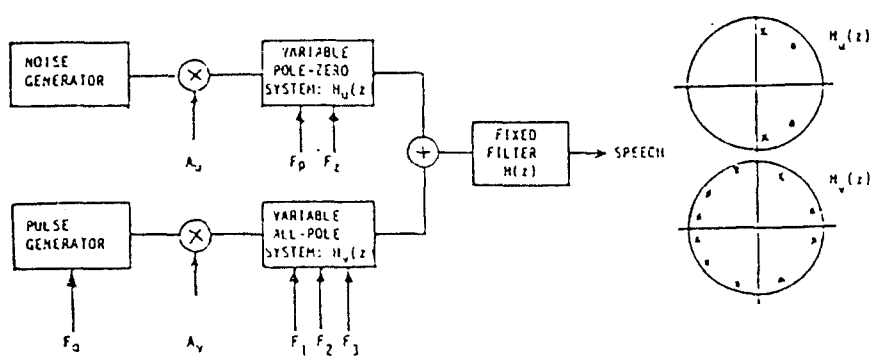


그림 5. Formant vocoder와 pole-zero plot

사용되기도 한다.

다음의 해석방법으로는 음성대역을 3개의 channel로 나누어 bandpass filtering을 하여 parameter들을 구할 수 있다. 여기에서 각 channel은 F_1, F_2, F_3의 각 한개씩의 formant를 찾아내도록 대역을 정한다. 각 formant들의 정확한 위치와 power는 각 channel에 대한 평균 영교차율과 평균 전력을 구하여 사용한다.

세번째 방법은 analysis-by-synthesis 방법으로 successive approximation에 의하여 계수를 추출하는 방법으로 model 계수들을 변화시켜 입력신호와 잘 맞는 approximation이 되도록 한다. 이와 같은 formant vocoder에 대한 블럭도가 그림 5에 나타나 있다.

Ⅵ. 선형 예측 부호화기(Linear Prediction Vocoder)

Formant vocoder의 성능을 개선하기 위하여 2차 filter들이 cascade된 합성 filter model을 고차의 선형 시스템으로 바꿀 수 있다. 이와 같은 시스템은 여기 신호의 pulse shape과 성도등의 여러가지 효과들을 동시에 modeling 할 수 있다. 시스템의 전달함수는

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

로 표시되며 여기에서 M은 filter의 차수이고 대략 10-12 사이의 값을 갖는다. $\{a_i,\ i=1,\cdots,M\}$는 model parameter로서 예측계수라 불리우고 이를 구하는 방법으로 선형예측해석, PARCOR해석 또는 선스펙트럼 해석방법이 주로 사용된다.

6.1 선형예측 해석 방법

음성신호의 파형을 관찰하여보면 음성파형의 이웃한 sample간에 상관관계가 높음을 알 수 있다. 이와같은 관계를 간단한 선형예측 형태로 표시하면 다음과 같다.

$$x_n = \alpha_1 x_{n-1} + \alpha_2 x_{n-2} + \cdots + \alpha_p x_{n-p}$$

위 관계식은 음성파형의 한 sample을 과거의 p개의 sample들의 선형결합으로 예측할 수 있다는 것을 가정하고 있는데, 이때 각 sample들에 곱하여 지는 가중치 $\{\alpha_i,\ i=1,2,\cdots,p\}$를 선형예측계수라 한다. 선형예측 해석에서는, 이 선형예측계수들을 예측오차신호의 평균자승차가 최소가 되도록 정한다.

먼저 이 선형예측계수들을 정하는 방법에 대하여 고찰하여 보기로 한다. 우선 입력 음성 신호 sample x_n에 대한 예측치를 $\hat{x}_n$이라 하면

$$\hat{x}_n = \alpha_1 x_{n-1} + \alpha_2 x_{n-2} + \cdots + \alpha_p x_{n-p}$$

로 표시할 수 있다. 그러므로 예측오차 e_n는

$$e_n = x_n - \hat{x}_n = x_n - \sum_{i=1}^{p} \alpha_i x_{n-i}$$

가 된다. 여기에서 α의 부호를 바꾸어 주면 예측오차는 다음과 같이 나타낼 수 있다.

$$e_n = x_n + \sum_{i=1}^{p} \alpha_i x_{n-i} = \sum_{i=0}^{p} \alpha_i x_{n-i}, \quad \alpha_0 = 1$$

예측오차 신호의 평균자승치는

$$\overline{e_n^2} = \overline{(\sum_{i=0}^{p} \alpha_i x_{n-i})^2} = \overline{(x_n + \alpha_1 x_{n-1} + \cdots + \alpha_p x_{n-p})^2}$$

로 표시된다. 예측오차신호의 평균자승치를 최소로 하는 $\{\alpha_i\}$를 구하기 위하여 $\overline{e_n^2}$의 식을 α_i에 대하여 편미분하면 다음과 같은 식을 얻게 된다.

$$R_p \alpha = -r$$

여기서 $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_p]^t$, $r = [r_{01}, r_{02}, \cdots, r_{0p}]^t$인 vector양이고 R_p는 r_{ij}를 갖는 $p \times p$ matrix로서

$$r_{ij} = \frac{1}{N} \sum_{n=0}^{N-1} x_{i+n} x_{j+n}$$

이다.

위의 matrix 식을 푸는 방법으로 두가지 방법이 제안되어 있는데 autocorrelation방법과 covariance방법이라 부른다. Autocorrelation방법은 입력음성신호 $\{x_n\}$에 길이 N인 window를 사용하여 그 window 밖의 sample 값을 0으로 가정하고 autocorrelation을 구하게 된다. 이 경우 autocorrelation matrix R_p는 positive definite한 symmetric Toeplitz형태를 갖게 되어 Levinson-Durbin algorithm 등을 이용할 경우 $\{\alpha_i\}$를 구하는데 계산량을 대폭적으로 줄일 수 있으면서도 안정된 계수들을 얻을 수 있다.

반면 covariance 방법의 경우는 r_{xx}를 구하는데 있어서 N개의 sample을 data sequence가 무한이 있다고 가정하고 취하여 계산을 하게 되므로 일반적으로 $r_{xx} \neq r_{xx}$가 되어 R_{xx}가 positive definite하다는 보장이 되지 않는다. 만약에 $\{x_n\}$에서 충분한 길이의 sample을 길이의 사용하고 신호가 stationary한 경우에는 위 두방법에 의한 해석 결과가 거의 같게 되지만 입력 sample의 수가 적거나 또는 nonstationary한 특성을 갖는 신호인 경우에는 두 방법이 서로 다른 $\{a_i\}$값을 갖게 된다. 이때 r_{xx}의 정의에 의하면 covariance방법이 $\{x_n\}$의 변화에 대하여 더욱 정확한 response를 내게 되지만 반면 항상 안정된 필터계수를 얻을 수 있는 것은 아니다. 시스템 이론에 의하면 입력 여기신호 $\{x_n\}$과 출력 응답신호 $\{x_n\}$ 사이의 관계식인

$$x_n + \sum_{i=1}^{p} a_i x_{n-i} = e_n$$

을 auto-regressive (AR) process라고 불린다. 양측에 z 변환을 행하면 시스템 함수 $H(z)$를 얻을 수 있는데 이 함수는 pole만을 포함하고 있다.

$$H(z) = \frac{1}{1 + a_1 z^{-1} + \cdots + a_p z^{-p}}$$

이 시스템을 안정한 all pole 시스템 또는 모델이라 한다. 그 다음의 관계식과 같이 시스템의 표현된 경우, 즉

$$x_n + \sum_{i=1}^{p} a_i x_{n-i} = e_n + \sum_{j=1}^{q} \beta_j e_{n-j}$$

입력과 시스템을 auto regressive moving average (ARMA) process라고 하며 시스템 함수 $H(z)$는

$$H(z) = \frac{1 + \beta_1 z^{-1} + \cdots + \beta_q z^{-q}}{1 + a_1 z^{-1} + \cdots + a_p z^{-p}}$$

로 표시되고 이는 pole zero 모델이라 한다.

음악음성과 같은 sample에 대한 선형예측해석은 음성과 같은 AR process로 가정하고 이에 대한 all-pole 시스템 모델을 구하는 과정이라고 할 수 있다. 따라서 앞에서 구한 최소자승오차 조건을 만족하는 matrix식은 시스템 이론에서 사용되는 Yule Walker식과 동일

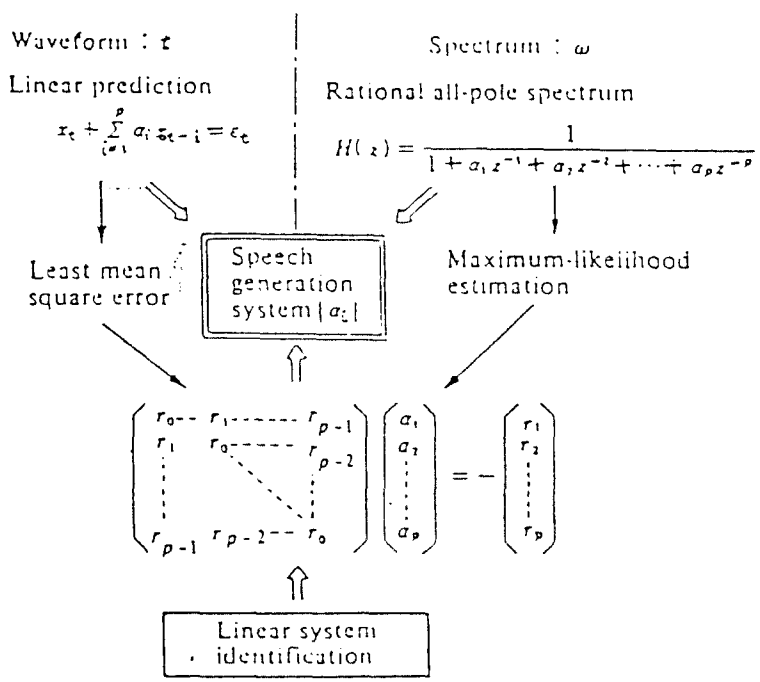


그림 6. 선형예측해석 스펙트럼 추정 방식 사이의 관계식

한 형태를 갖는다.

앞서의 선형예측 계수를 구하는 방법은 시간축 상의 sample 간의 상관계수에 의거하여 시스템의 전달함수를 구하는 것으로 곧 주파수 영역에서의 입력여기신호의 주파수 스펙트럼 envelope을 결정하여 주는 것이라 할 수 있다. 이것은 주파수 영역에서의 음성신호 스펙트럼 추정과정을 고찰하여 보면 확인이 되는 사실이다.

만약에 음성신호가 all-pole rational 스펙트럼 밀도함수를 갖는 시스템에 zero mean이고 분산이 σ^2인 random 입력 신호가 통과되어 얻어진 것이라고 가정하고 maximum-likelihood 스펙트럼 추정을 할 경우 얻어지는 스펙트럼 추정결과는 앞서의 AR-process에 대한 선형예측 해석의 결과와 동일하다. 즉 입력 신호가 all-pole 모델에 적합한 process라 가정하면 시간 영역에서의 선형예측 해석과 주파수 영역에서의 maximum-likelihood criterion에 의한 스펙트럼 추정은 동일한 matrix formula를 갖게 되어 같은 시스템함수를 나타낸다. 이 관계를 그림 6에 나타내었다.

음성에 대한 선형예측 해석의 결과로 얻어진 특징계수들은 적절히 양자화 되어 전송이 된다. 일반적으로 8 KHz로 표본화된 음성신호 해석의 경우 p = 8-10 정도가 되며 각각의 계수를 독립적으로 양자화(스칼라 양자화라함) 하기나 계수 전체를 한 vector로 보아 양자화(베터 양자화라함)하게 된다. 수신측에서는 수신된 디지탈 정보를 복호하여 음성 합성기를 구성하게 되는데 선형예측 계수를 이용한 음성합성이 그림 7에 나타나 있다.

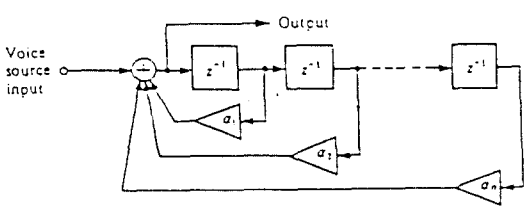


그림 7. 선형예측계수에 의한 음성합성

6.2 PARCOR 해석 방식

선형예측부호화 방식에서 얻어진 예측계수 $\{\alpha_i\}$를 양자화하여 전송할 경우에 다음과 같은 문제가 일어난다. 첫째로는 선형예측 계수들의 dynamic range가 넓어 적은 bit으로 계수들을 양자화할 경우 스펙트럼의 distortion이 크게 되는 단점이 있다. 또한 계수들을 적은 bit으로 양자화 할 때 양자화된 계수들로 음성합성 필터를 구성하면 불안정한 발진이 일어나는 수가 있다. 일반적으로 전체 전송속도를 2400bps 이하로 하기위하여는 각 계수당 양자화에 할당할 수 있는 bit의 수는 3-5bit으로 이와 같이 적은 bit으로 예측계수들을 양자화하는 것은 부적당하게 된다. 두번째로는 선형예측계수들의 값들이 선형예측 해석의 차수 p에 따라 달라지는 점이다. 그러므로 전송로의 상황에 따라서 전송하려는 예측계의 수를 적절히 증감시키기가 어렵게 된다. 위와 같은 문제를 해결하는 방법의 하나로 제안된 해석 방법이 PARCOR 해석 방법이다.

Partial autocorrelation (PARCOR)은 k-parameter라고도 불리우는데 위에서 언급한 선형예측계수의 단점을 보완할 수 있는 이점을 갖고 있다. 주어진 입력 sample sequence $\{x_n,\ n=0,1,\ldots,N-1\}$에 대한 일반적인 자기상관은 다음과 같이 주어진다.

$$r_i = \frac{1}{N}\sum_{n=0}^{N-1-i} x_n x_{n+i}$$

이 식은 i개의 sample 간격 만큼 떨어진 sample들의 곱의 합으로 i개 sample 간격만큼 떨어져 있는 신호간의 상호연관관계 정도를 나타내고 있다. 반면 partial autocorrelation은 i개의 sample 간격만큼 떨어져 있는 residual 파형의 상관관계를 표시하고 있다.

여기에서 x_i와 $x_{i+(n+1)}$ 사이의 n+2 sample들을 생각하여 보자. 이 sample 들이 AR process로 표현될 수 있다고 하면 forward 예측오차는 다음과 같이 표시된다.

$$e_{fi} = x_i - \hat{x}_i = x_i + \sum_{j=1}^{n} \alpha_j \cdot x_{i-j} = \sum_{j=0}^{n} \alpha_j \cdot x_{i-j}$$

여기에서 $\{\alpha_j\}$는 forward 선형예측 계수이고, backward 예측오차는

$$e_{bi} = x_{i-(n+1)} - \hat{x}_{i-(n+1)} = x_{i-(n+1)} + \sum_{j=0}^{n} \beta_j \cdot x_{i-j}$$
$$= \sum_{j=0}^{n+1} \beta_j \cdot x_{i-j}$$

로 나타내고 $\{\beta_j\}$는 backward 선형예측 계수이다. 이때 $\{x_i\}$가 stationary하다면

$$\beta_j = \alpha_{n+1-j} \quad ; \quad j = 1,\cdots\cdots,n+1$$

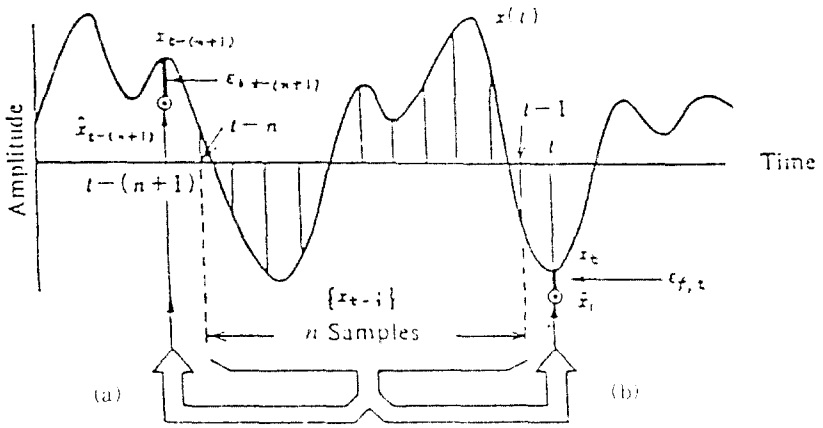


그림 8. Partial autocorrelation(PARCOR) 해석 방법

의 관계가 성립한다. 이 관계는 그림 8에 나타내었다.

Partial autocorrelation k_{n+1}은 다음과 같이 정의된다.

$$k_{n+1} = \frac{\sigma_f \sigma_b \cos\theta}{\sigma_f^2 \sigma_b^2}$$

$$= \frac{E[(\sum_{i=0}^{n} \alpha_{n+1,i} x_{t-i})(\sum_{i=0}^{n} \beta_{n+1,i} x_{t-i})]}{\{E[(\sum_{i=0}^{n} \alpha_{n+1,i} x_{t-i})^2] E[(\sum_{i=0}^{n} \beta_{n+1,i} x_{t-i})^2]\}^{1/2}}$$

여기에서 $\{\alpha_{n+1,i}\}$와 $\{\beta_{n+1,i}\}$는 각각 σ_f^2와 σ_b^2을 최소로 하는 선형예측 계수이고 $\{x_t\}$가 stationary일 경우 $\beta_{n+1,i} = \alpha_{n+1,n+1-i}$가 된다.

실제적으로 PARCOR 및 선형예측계수를 구하는 방법으로는 Levinson Durbin 방법을 들 수 있는데 일반적인 matrix inversion을 행하는 것과 비교하여 계산량을 대폭 줄일 수 있으며 k_{n+1}와 $\{\alpha_{n+1,i}\}$를 동시에 recursive하게 구할 수 있어 널리 사용되고 있다.

PARCOR 계수를 구하는 다른 방법으로 lattice 방법을 들 수 있다. 우선 forward 에러오차와 backward 에러오차를 z 변환으로 표시하면

$$e_{f,t} = (\sum_{i=0}^{n} \alpha_{n,i} z^{-i}) x_t = A_n(z) x_t$$

$$e_{b,t-(n+1)} = (\sum_{i=0}^{n} \beta_{n,i} z^{-i}) x_t = B_n(z) x_t$$

로 된다. β를 α_{n+1-i}로 치환하면

$$B_n(z) = z^{-(n+1)} A_n(1/z)$$

가 되고 α와 k에 관한 관계식을 이용하면 다음과 같은 식을 얻을 수 있다.

$$A_{n+1}(z) = A_n(z) - k_{n+1} \cdot B_n(z)$$
$$B_{n+1}(z) = z^{-1} B_n(z) - k_{n+1} \cdot A_n(z)$$

이 recursive formula와 초기조건 $A_0(z) = 1$과 $B_0(z) = z^{-1}$을 이용하면 PARCOR 계수 $\{k_i\}$를 다단 lattice 회로에서 순차적으로 구할 수 있다. 이와 같이 하여 $\{k_i\}$를 구하는 방법을 lattice 방법이라하며 그림 9에 나타내었다.

PARCOR 계수는 전송 parameter로 선택한 경우 stable한 시스템에서의 계수 값이 ± 1 내에있게 되므로 양자화하기가 쉽다. 그러나 만약 입력 음성 신호의 스펙트럼이 낮은 주파수 영역에 집중되어 있을 때에는 k_1의 값이 ± 1에 가까이 접근하게 되고 그러므로 양자화의 해석의 정확도가 떨어지게 된다. 이때에 음성의 고역 주파수 스펙트럼을 emphasis하여 주면 해석의 정확도를 높일 수 있다. 일반적으로 고역 주파수에 대한 emphasis는 상대적으로 저역 주파수 스펙트럼에 대한 에너지는 낮추는 결과가 되어 전체적인 스펙트럼의 dynamic range를 줄여주게 됨으로써 해석의 정확도를 높일 수 있다고 이해되고 있다. 고역 주파수의 emphasis 방법으로는 analog 미분 회로를 사용하거나 디지탈 신호의 차신호를 사용하는 등 여러가지를 들 수 있는데 대략 3 bit 정도의 해석의 정확도를 더 얻

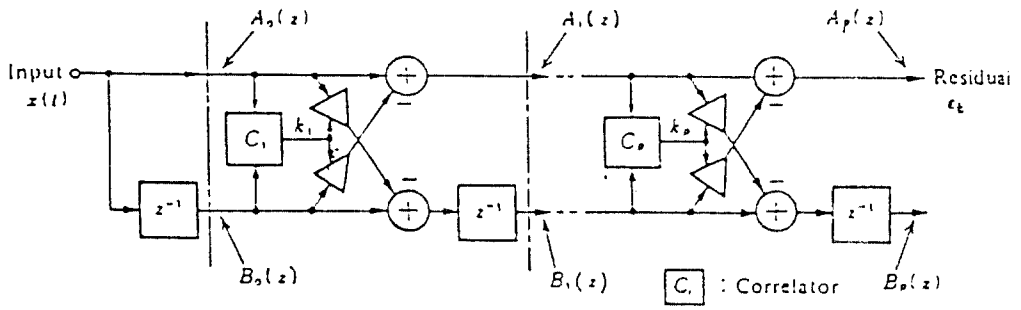


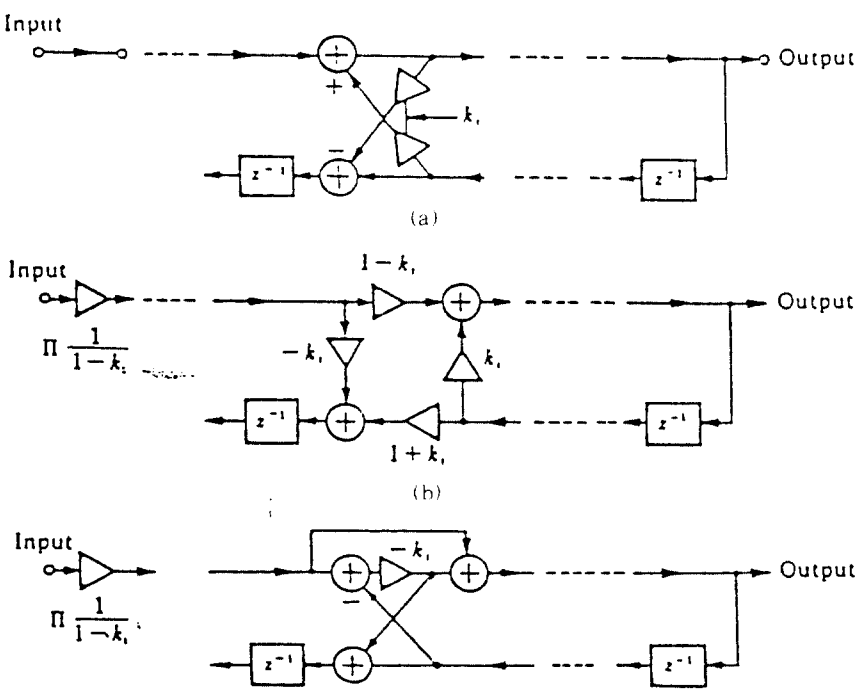
그림 9. PARCOR 계산을 위한 lattice 방법

을 수 있다.

6.2.1 PARCOR

PARCOR 음성 합성 방식은 입력 파형 sample에서 순차적으로 correlation을 제거하여 나가는 방식으로 all-pole 스펙트럼의 formant 구조를 차례차례로 inverse filtering 하여 스펙트럼을 평탄하게 만들어가는 과정이다. 따라서 그 해석 결과인 residual 신호는 유성음의 경우 pulse train과 같은 모양을 갖고 무성음의 경우에는 random noise와 같게 된다.

PARCOR 음성 합성은 해석방법과 반대로 residual 파형에 순차적으로 correlation을 더해주는 과정으로 평탄한 주파수 스펙트럼을 갖는 residual 입력에 formant 구조를 만들어가는 것이라 할 수 있다. 만약 이 때 음성해석에서 얻은 residual 파형을 입력으로 사용하면 PARCOR 음성합성 후 원래의 입력 음성과 동일한 파형을 얻을 수 있다.



(a)2-승산기 lattice 형 (b)ladder 형 (c)1 승산기 형

그림 10. PARCOR 음성 합성 회로

PARCOR 음성합성 방식을 식으로 표시하면 다음과 같다.

$$A_m(z) = A_{m+1}(z) + k_{m+1}B_m(z)$$
$$B_{m+1}(z) = z^{-1}[B_m(z) + k_{m+1}A_m(z)]$$

위식을 세가지의 다른방식으로 구현한 합성회로가 그림 10에 표시되어 있다.

6.2.2 음성 여기신호에 대한 모델링 및 특성추출

PARCOR 해석의 결과로 얻어진 residual 파형을 그대로 PARCOR 합성회로의 입력으로 사용하면 원래의 음성파형이 그대로 재생이 되지만 이 경우 요구되는 data 감축 효과는 없게 되어 저전송속도의 음성 통신은 불가능하게 된다. 그러므로 음성신호 발생 모델에 입각하여 residual 파형을 다음과 같이 모델링하여 전송하게 된다. 즉 유성음의 경우는 pulse train으로 또 무성음의 경우는 random noise로 근사시켜 전송한다.

유성음의 여기 신호는 진폭 A와 pitch 주기 T로 양자화하여 전송하며 T = 0는 noise source를 의미하고 T > 0인 경우 유성음의 pulse 신호를 나타낸다. T와 A가 모두 0인 경우는 묵음 구간을 나타내며 A와 T에 대한 정보는 모두 residual 파형에서 얻는다.

Residual 파형의 에너지는 A로 나타내며 residual 파형의 자기상관이 주기적인가의 여부에 따라 유·무성음이 구별된다. 이때 주기로 pitch 주기 T를 결정하게 된다. 그러나 실제적으로 정확한 pitch 주기를 구하는 것은 어려운 문제로 다음의 여러가지 방법들이 사용된다.

(1)Residual 파형의 correlation
(2)입력 음성파형의 자승과 양의 신호의 correlation
(3) 지연 이래된 residual 파형의 correlation
(4)입력음성 또는 입력음성의 자승과 성분의 average magnitude difference function(AMDF)
(5)입력 음성의 cepstrum

또한 pitch pulse 파형으로는 순수한 impulse가 아닌 다른 파형들의 사용이 연구되고 있으며 음질개선의 효과가 있음이 발표되고 있다.

이와같은 음성특성계수들을 추출하여 전송하고 수신된 계수들로 음성을 합성하는 시스템에 대한 블럭도가 그림 11에 나타나 있다.

6.2.3 PAECOR 음성해석및 합성시의 유의점

PARCOR 계수를 보호화하여 전송할 경우 그러하여야 할 점으로 다음 사항을 들 수 있다.

(1)PARCOR 계수의 분포특성

PARCOR 계수는 그 차수에 따라 분포 특성이 변화

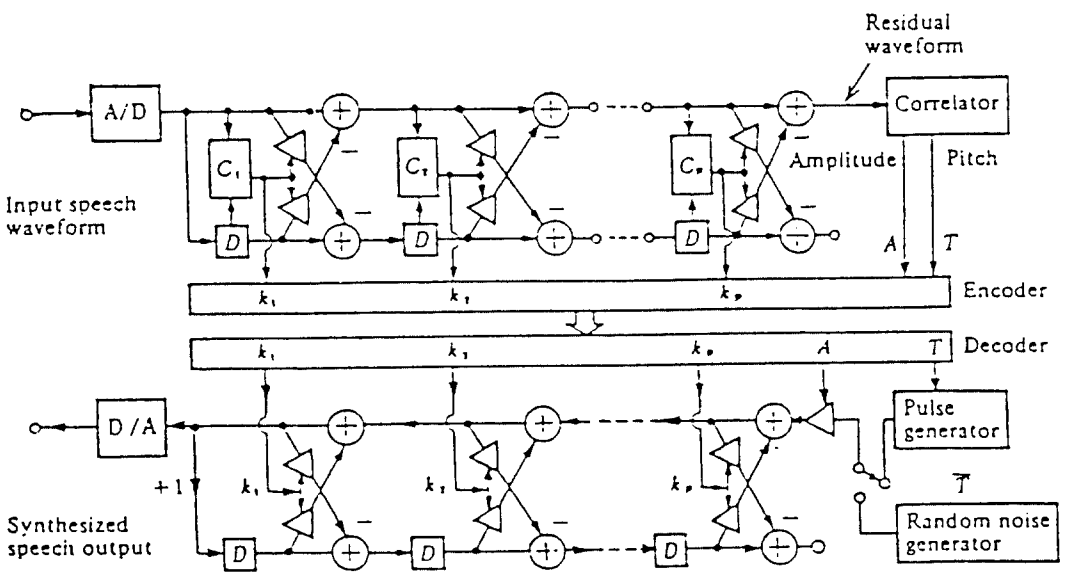


그림 11. PARCOR 해석 및 합성 시스템

하는데 k_1의 경우는 +1 근처에 집중적으로 분포되며 k_2는 -1 근처에 높은 분포를 갖는다. 차수 i가 증가하면 k_i의 분포는 0을 중심으로 넓게 분포된다. 또한 k_i의 분포는 화자의 성별에는 별 영향을 받지 않는 것으로 알려져 있다.

이와 같은 k_i의 분포외에도 양자화로 인한 변화량 $\triangle k_i$에 대한 스펙트럼 sensitivity도 k_i의 양자화에 중요한 요소로 작용을 한다. 같은 양의 $\triangle k_i$에 대하여는 계수의 order i가 작을수록 스펙트럼 distortion이 증가하며 특히 낮은 차수의 계수 변화는 적은 주파수 영역에서 큰 distortion을 유발하고 높은 차수의 계수 변화는 큰 주파수 영역에 걸쳐 적은 스펙트럼 distortion을 발생시킨다.

(2)PARCOR 계수의 양자화 및 비선형 bit 할당

단위시간당 전송속도가 한정되어 있을 때 각 계수당 양자화에 필요한 bit 수를 비선형으로 할당하는 것이 인위적으로 동일하게 할당하는 것보다 스펙트럼 distortion을 줄일 수 있어서 합성음질이 향상된다. 이 때에 전체 bit 양이 증가하면 양자화 distortion은 지수적으로 감소한다.

(3)Frame Rate

청취 실험 결과에 의하면 좋은 합성음질을 얻기 위해서는 frame rate이 15 msec 이하이어야 한다. 그러나 5 msec 이하의 frame rate 될 경우는 더 이상 음질이 향상되지 않는다.

(4)PARCOR 계수의 수

PARCOR 계수의 수 p가 8을 넘을 경우 음질향상의 정도가 상당히 적으나 p가 6이하일 때에는 음질저하가 심하게 생긴다. p값에 대한 최적치는 전체 전송속도에 따라 변화되는데 여성음성의 경우 9600 bps 이상의 전송속도에서의 p의 최적치는 10이고 그보다 낮은 전송속도에서는 8 정도이다.

(5)PARCOR 계수의 비선형 변환

PARCOR 계수의 스펙트럼 sensitivity를 균일하게 하여 양자화를 쉽게 할 수 있도록 하기위하여 k_1, k_2에 대한 arctanh 또는 arcsin 변환을 하는 수가 있다. 이 경우 k_1, k_2에 대한 비교적 균일한 sensitivity를 얻을 수 있어 양자화 과정에 적절히 사용될 수 있다. 실제적으로 k_i의 부호화는 다음과 같이 행할 수 있다.

우선 k_1과 k_2의 경우 $x = \text{arctanh } k_i$ 변환을 한 후 $|x|<3$의 범위에 대하여 선형 양자화를 행한다. k_3 이상의 계수에 대하여는 비선형 변환이 필요가 없으며 각 계수의 범위에 맞도록 선형 양자화를 행한다.

(6)시간영역에서의 전송계수들의 Interpoation

PARCOR 계수의 frame 단위의 표본화 과정에 있어서 frame 시간이 길어질 경우 적당한 시간 간격으로 계수의 값들에 대한 interpolation이 필요하게 된다. Interpolation 방식으로는 일반적으로 선형 interpolation이 사용된다. 선형 interpolation을 사용할 경우 모음구간에서는 상당히 효과가 좋으나 무성자음 또는 음성의 변이 부분에서는 스펙트럼 distortion이 심하게 나타난다. 따라서 이와 같은 interpolation 시의 문제점들을 해결하기 위한 한 방법으로 LSP(line spectrum pair)방식이 제안되었다.

6.3 선 스펙트럼 해석 방식

선형예측해석방식에서 추출된 선형예측계수 $\{\alpha_i\}$를 양자화하여 전송할 경우 각 α_i를 10 bit 미만으로 양자화하게 되면 스펙트럼 diatortion이 크게 될 뿐만 아니라 합성필터자체가 불안정하게되어 발진하는 수가 생긴다. 이와 같은 결점을 보완하기 위하여 PARCOR 해석 방식이 개발되었다. 그러나 이 PARCOR 시스템에서도 하나의 결점은 해석구간이 길어져 계수의 interpolation을 수행할 때 스펙트럼 distortion이 크게 되는 점이다. 이것은 $\{k_i\}$의 특성이 선형 interpolation에 맞지 않기 때문으로 선 스펙트럼 pair (LSP, line spectrum pair)계수가 제안되었으며 이 점을 보완할 수 있게 되었다. 이 LSP 해석 방식의 개요를 PARCOR 합성과정부터 시작하여 살펴보기로 한다.

PARCOR 합성과정이 그림 12에 도시되어 있다. 그림에서 보면 입력단자 X에서 출력단자 Y까지의 전달함수 $H_p(z) = \dfrac{1}{A_p(z)}$ 이고 Y에서 Z까지의 전달함수는 $R_p(z)$이다. 따라서 X에서 Z까지의 전달함수 $R_p(z) = \dfrac{R_p(z)}{A_p(z)}$ 로 주어지며 이 시스템이 안정한 조건은 $|k_i|<1$, $i=1,2,\ldots,p$이다.

PARCOR 음성 합성은 유파가 손실이 없는 음향관을 전파되어 나가는 과정으로 설명할 수 있다. 이 시스템에서의 손실은 Z 단자에서의 backward 에너지 형태로 나타나게 되는데 이 음향관 모델에서 Z 단자를

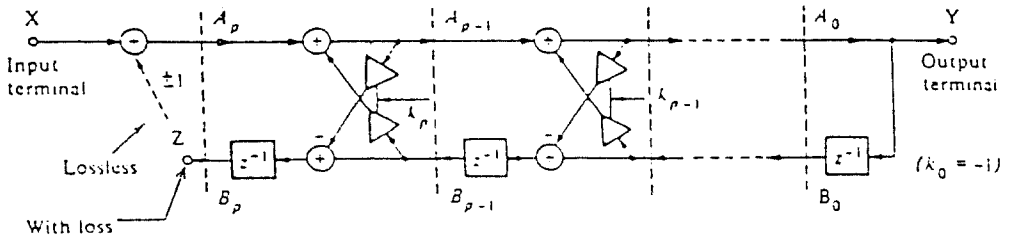


그림 12. PARCOR음성 합성 및 선 스펙트럼 변환

$k_{p+1} = \pm 1$인 선로를 통하여 입력단자에 연결하여 주게되면 완벽한 무 손실 시스템이 된다. 그러므로 각 공진점에서의 Q가 무한대가 되기 때문에 에너지의 스펙트럼이 여러 개의 선 스펙트럼에 집중된다. Feedback 경로 $k_{p+1} = -1$은 입력 터미널이 완전히 닫혔을 때에 해당하고 $k_{p+1} = \pm 1$은 무한 자유공간에 입력 터미널이 열려 있을 때에 해당한다.

만약에 음향관이 무 손실 시스템이 되고 공진 Q가 무한대 일때 p차 방정식의 근은 $z = e^{j\omega}$로 되어 z평면상의 단위원상에 존재하게 된다. 방정식의 차수는 단위원을 x축상에 투사할 경우 $(x = \cos\lambda)$ $\frac{p}{2}$로 줄어들게 된다. 이 식을 풀면 $\lambda = \cos^{-1}x$로 되고 λ에서의 $H_p(z)$의 값이 intensity m을 결정한다. 이 (λ, m_i), $i = 1,2,\cdots,\frac{p}{2}$를 $H(z)$의 선 스펙트럼 표시라 하고 (λ, m_i)의 값은 음성 입력 파형의 correlation으로 부터 구한 $\{\alpha_i\}$나 $\{k_i\}$, $i = 1,2,\cdots,p$로 부터 구할 수 있다.

그림 12에서 feedback 경로 $k_{p+1} = \pm 1$일 때의 전달함수를 각각 $P_p(z)$와 $Q_p(z)$라 하면 다음과 같은 식을 얻을 수 있다. $k_{p+1} = 1$일 때

$$P_p(z) = A_p(z) - B_p(z) = A_p(z) - z^{-(p+1)}A_p\left(\frac{1}{z}\right)$$

이고 $k_{p+1} = -1$ 일때

$$Q_p(z) = A_p(z) + B_p(z) = A_p(z) + z^{-(p+1)}A_p\left(\frac{1}{z}\right)$$

이다. 그런데 $A_p(z) = 0$의 근이 $|z| = 1$인 단위원 내에 존재하면 $P_p(z) = 0$와 $Q_p(z) = 0$는 서로 다른 근을 갖고 그 근들은 단위원상에 존재하게 된다. 따라서 $P_p(z)$와 $Q_p(z)$는 다음과 같이 나타낼 수 있다. p가 우수일 때

$$P_p(z) = (1 - z^{-1})\prod(1 - 2\cos\omega_i z^{-1} + z^{-2})$$

$$Q_p(z) = (1 + z^{-1})\prod(1 - 2\cos\omega_i z^{-1} + z^{-2})$$

이고, p가 기수일 때는

$$P_p(z) = (1 - z^{-2})\prod(1 - 2\cos\omega_i z^{-1} + z^{-2})$$

$$Q_p(z) = \prod(1 - 2\cos\omega_i z^{-1} + z^{-2})$$

가 된다.

여기에서 $A_p(z)$는 $\frac{[P_p(z)+Q_p(z)]}{2}$로 주어지고 $H_p(z) = \frac{1}{A_p(z)}$ 이므로 LSP 음성 합성방식은 그림 13과 같이 표시할 수 있다. 위 LSP 합성 방식의 실제적인 구현 방식이 그림 14에 나타나있다.

그림 14에 보듯이 LSP 합성 회로는 (c_i, c_{i+1}), $c_i = \cos\omega_i$, $i = 1, 3, 5, \ldots, \frac{(p+1)}{2}$ 또는 (ω_i, ω_{i+1})이 주어

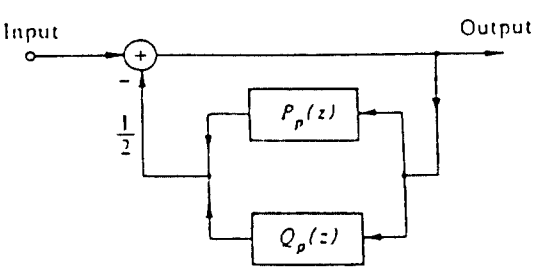
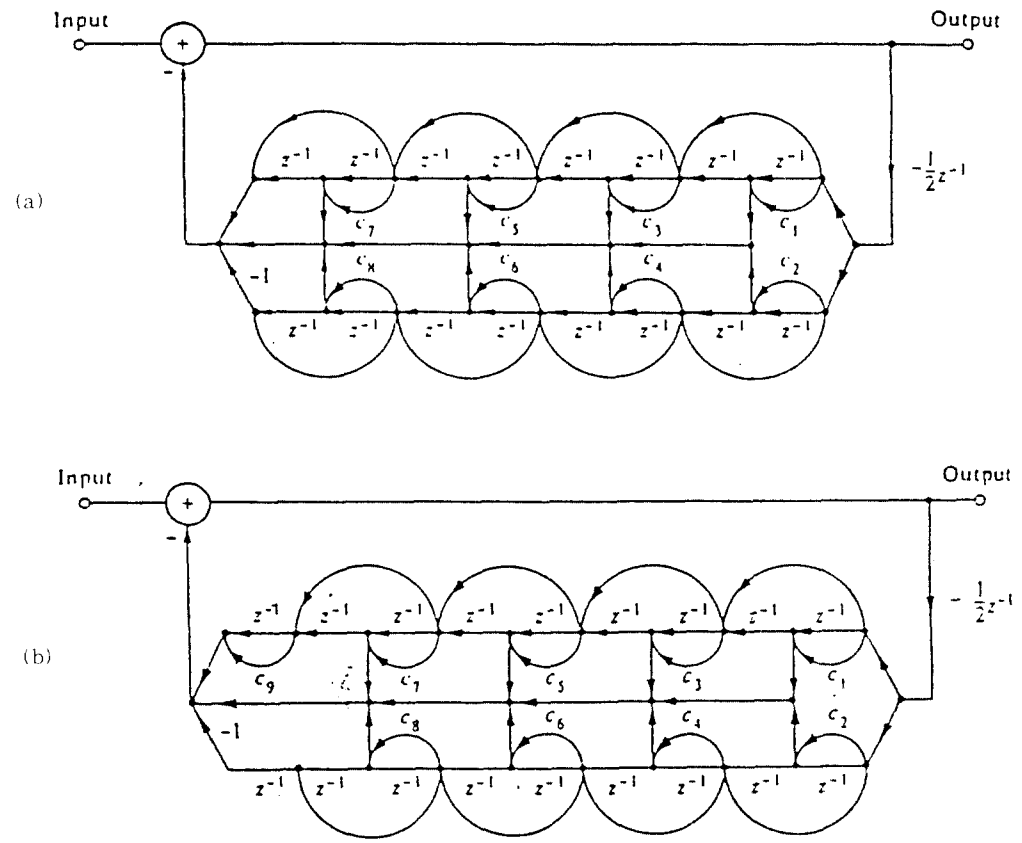


그림 13. LSP 합성도



(a)p = 8(even) (b)p = 9(odd)

그림 14. LSP 합성의 실제 $c_i = 2\cos\omega_i$

지면 구성할 수가 있다. 여기에서 (ω_i, ω_{i+1})이 한 수 파수 쌍을 나타내기 때문에 음성 스펙트럼의 LSP 표시라 한다. 선형 예측 계수 $\{\alpha_i\}$, $i = 1,2,...,p$와 PARCOR 계수 $\{k_i\}$, $i = 1,2,...,p$ 선 스펙트럼 표시 (λ_i, m_i), $i = 1,3,...,p-1$과 LSP 표시 (ω_i, ω_{i+1}), $i = 1,3,...,p-1$은 모두 all-pole 스펙트럼의 equivalent한 표시 방법으로 그림 15에 LSP 해석방식에 대한 예가 표시되어 있다.

LSP 특성 계수는 스펙트럼 sensitivity가 균일하고 저 전송 속도에서 스펙트럼 distortion이 적으며 선형 interpolation이 양호하다. 따라서 2400 bps 이하의 저 전송 속도에서 같은 bit rate의 경우 PARCOR 계수의 전송시보다 양질의 합성 음성 출력을 얻을 수 있다.

Ⅶ. Pitch 검출과 pitch 주기 추출

Pitch 검출의 목적은 pitch 주기에 연관된 두가지의 model 계수를 얻는데 있다. 이것은 excitation mode 즉 유성음, 무성음의 여부에 대한 것과 pitch 주기의 값이다. Pitch 주기의 여부 검출과 pitch 주기 추출로 이 계수들을 구하게 된다.

Pitch 정보는 음성의 naturalness에 큰 영향을 미치기 때문에 인간의 청각은 pitch 변화에 매우 민감하게 반응한다. 그러므로 정확한 pitch 해석은 합성음성의 음질을 좌우하는 중요한 요소이다.

Pitch 해석을 위하여 많은 해석방법들이 제안되었고 실제적으로는 그 중 한두가지를 조합하여 사용하는 수가 많다. 본 절에서는 기본적인 pitch 해석 방법만을 설명하기로 한다. 일반적으로 pitch 해석의 정확

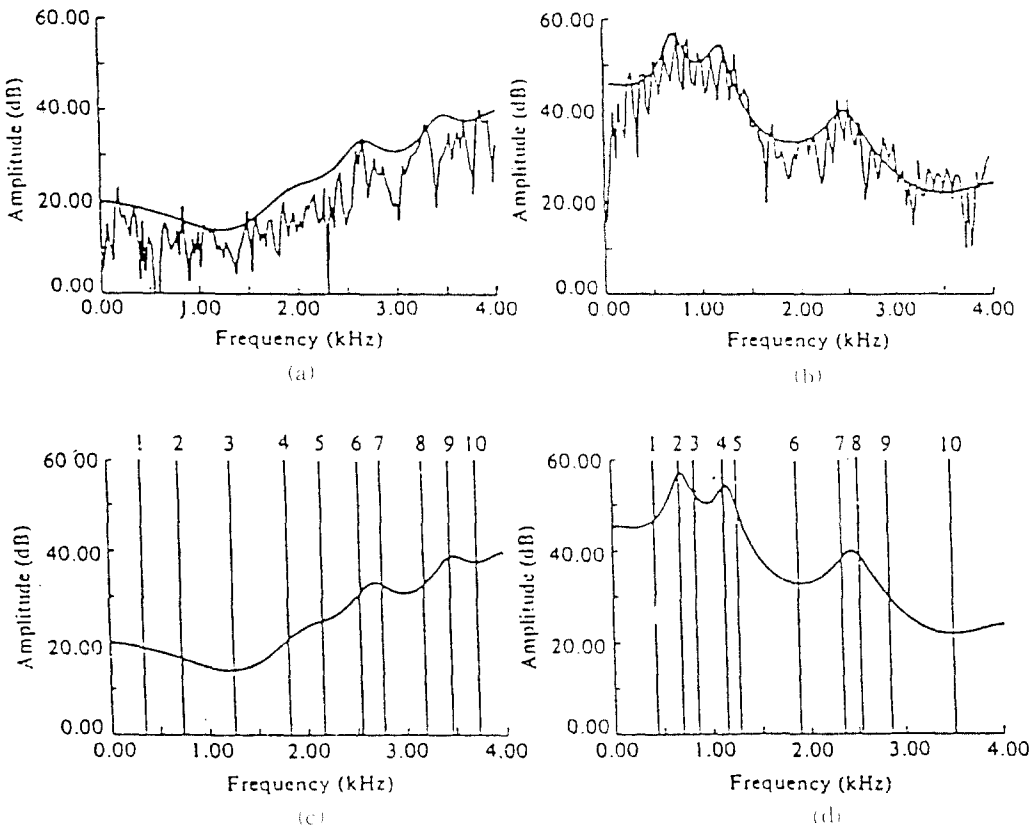


그림 15. LSP 해석의 예

도를 높이기 위하여는 음성신호의 preprocessing과 postprocessing 과정이 필요하다. Preprocessing으로 수행되는 신호처리 기법으로는 normalization, 적응 filtering, clipping 등이 있고 pitch 해석 후 postprocessing으로는 smoothing, correction 등을 들 수 있다.

Pitch 검출을 시간영역, 주파수영역에서 수행할 수 있으며 두가지가 혼합된 hybrid 방법도 사용된다.

7.1 시간영역에서의 pitch 검출

Pitch 검출의 기본원리는 formant들로 인하여 나타나는 quasiperiodicity를 preprocessing을 통하여 제거한 후 간단한 시간영역의 pitch 검출 algorithm을 사용하는 것이다.

1)영교차

영교차율은 입력 음성신호를 low-pass filtering한 후 추정하는 데 low pass filter의 차단주파수와 검출할 수 있는 pitch 주기 주파수가 연관되어 있다. 그러므로 formant 주파수에 따라 이 차단주파수를 변화시키는 adaptive filtering의 사용이 바람직하다.

2)Peak와 Valley

Low pass filter된 음성신호에서 peak 점과 valley 점들을 찾아내어 이들의 주기성에서 pitch를 검출한다. 실제에 있어서 6개의 검출기가 병렬로 연결되어 사용되는 데 그 각각의 검출기가 추출하려는 정보는 peak, valley, peak to valley, valley to-peak, peak to previous-peak와 valley to-previous-valley에 대한 것이다. 이 검출기들을 사용하면 입력신호에 pitch 주기에 대한 기본파와 2차 harmonic이 포함되어 있을 경우에도 정확한 pitch 주기를 검출할 수 있다.

3)자기 상관 함수(Autocorrelation Function)

이 방법은 주기적인 신호의 자기상관함수도 같은 주기를 갖는 주기함수가 된다는 성질을 이용하는 것으로 신호의 주기에 해당하는 점에 peak점이 나타나게 된다. 주기적인 신호의 자기상관함수의 peak점의 값은 원점에서의 값과 상응하는 크기가 되며 음성과 같은 quasiperiodic한 신호에 있어서는 peak점의 값이 조금 작게 된다.

실제 음성신호에서 자기상관함수를 구하기 위하여는 window를 사용한다. 그러므로 자기상관함수는

$$r_n(m) = \sum_{n=1}^{N-1} x(n)w(i-n)x(n+m)w(i-n-m)$$

으로 주어지고 직접 이 함수를 구하거나 또는 FFT를 이용하는 수도 있다. 이 함수의 첫번째 최대치에 해당하는 m값이 구하려는 pitch값이 된다.

자기상관함수를 이용한 pitch 주기 검출에 있어서 입력 신호의 작은 값들을 0로 center clipping한 후 구한 자기상관함수를 사용하면 pitch 검출기의 성능을 향상시킬 수 있다. 또한 입력신호의 level을 +1, 0, -1로 양자화하여 자기상관을 구하면 계산량을 대폭 줄일 수 있다.

4)Average Magnitude Difference Function(AMDF)

이 방법은 자기상관함수계산에 소요되는 계산량을 줄이려는 목적으로 제안되었으며 실제적인 구현도 간단히 할 수 있는 이점이 있다. 자기상관함수의 계산과 비교하여 보면 두 신호의 곱을 구하는 대신 차를 구하는 것으로 주기적인 신호에 있어서 주기 p만큼 떨어져 있는 신호의 차가 0가 되는 즉 $x(n)-x(n+p)=0$ 인 성질을 이용하는 것이다. 그러므로 AMDF는 다음과 같이 나타낼 수 있다.

$$AMDF(m) = \sum_{n=1}^{N-1} |x(n)w(i-n) - x(n+m)w(i-n-m)|$$

여기에서 검출하려는 pitch 주기는 AMDF(m)가 최소치를 갖는 점의 m값이다.

7.2 Spectral Methods

주파수 영역에서의 pitch 검출기는 주기적인 신호의 스펙트럼이 기본 주파수와 그 harmonics에서 impulse형태로 나타난다는 사실을 이용하려는 것이다. 이와 같은 harmonics 구조는 유성음에서만 나타나며 각 harmonics의 주파수 간격을 측정함으로서 pitch 주기를 추출할 수 있다.

1)Cepstrum Method

음성신호의 스펙트럼 X(k)가 여기신호의 스펙트럼 E(k)과 성도의 스펙트럼 R(k)의 곱으로 표시된다면 X(k)의 logarithm은

$$\ln|E(k)R(k)| = \ln|E(k)| + \ln|R(k)|$$

가 된다. 위 식의 IDFT를 취하면 cepstrum $\hat{x}(n)$을 얻는 데 이는

$$\hat{x}(n) = \hat{e}(n) + \hat{r}(n)$$

이다. 여기에서 여기신호 e(n)은 주기적인 신호이기 때문에 cepstrum $\hat{e}(n)$은 pitch 주기의 위치에 큰 peak를 갖는다. 반면 r(n)은 slow, aperiodic oscillation을 하게 되므로 평탄한 특성을 갖는다.

Cepstrum을 이용하여 실제로 pitch를 구하기 위하여는 주파수영역에서의 resolution이 좋아야한다. 그러나 이때에도 F_0가 너무 작아 다음 spectral line들과 인접하게 되면 pitch 주기의 추출이 어려워 진다.

2)주파수 영역에서의 Comb Filtering

이 방법은 음성의 short-time spectrum $X_i(k)$를 주파수영역에서 variable comb filtering하는 것으로 comb filter 함수를 $C(k, k_0)$라 하면 이 함수와 $X_i(k)$와의 cross-correlation이 최대가 되는 값을 찾는 것이다. 여기에서 n_0는 comb filter에서 2개의 teeth간의 간격이다. n_0를 변화시켜가는 과정에서 최대값이 얻어졌을 때는 comb filter가 $X_i(k)$의 harmonic line들과 정확히 일치할 때이다. 이때 $F_0 = k_0$가 된다. 이것을 식으로 표시하면

$$F_0 = argmax \{ \sum_{k=1}^{N-1} X_i(k) C(k, k_0) \}$$

이다.

7.3 Hybrid Method

Hybrid 방법은 시간영역과 주파수영역의 특성들을 동시에 사용하는 방법이다.

1)Simplified Inverse Filtering 기법

이 방법은 우선 성도에 의한 입력신호 x(k)의 spectral envelope의 영향을 줄인 후 자기상관법에 의하여

pitch를 검출하는 것이다. Spectral envelope의 영향을 줄여 평탄한 주파수 특성을 얻는 과정을 signal whitening이라 하는데 이것은 시간영역에서의 inverse filtering을 통하여 얻는다.

Ⅷ. Vocoding 방식의 개선및 연구 방향

Vocoder의 음질개선을 위한 연구는 앞에서도 언급한 바와 같이 음성신호발생모델에서 spectral envelope의 정확한 추출과 여기신호에 대한 모델링에 초점이 집중되어 진행되고 있다. spectral envelope의 정확한 추출에 대한 연구로는 앞에서 설명한 대역 여과기로는 FFT를 이용한 short time Fourier 해석, homomorphic 신호해석, 선형예측해석방식등이 제안되어 사용되고 있으며, 이 중에서도 선형예측해석방식은 그 성능의 우수함이 입증되어 미 국방성 표준 방식인 LPC-10에 채용되어 사용되고 있으며 그 변형된 형태의 PARCOR 해석과 신 스펙트럼 해석방식에 대한 연구가 활발히 진행되고 있다. 그 외에도 음성신호중 무음의 정확한 표현을 위한 pole-zero 모델링 방법에 대한 연구도 계속되고 있다. 그러나 vocoder의 음질개선을 위한 음성발생모델연구는 선형예측방식의 제안이후 최근에는 주로 여기신호의 모델링 분야에서 중점적으로 진행되고 있다.

합성음질 개선을 위한 여기신호의 모델링에 대한 연구로는 초기의 간단한 방식 즉 유성음의 경우에는 주기적인 impulse신호로, 무성음의 경우에는 random noise 신호로 사용하던 방식에서 최근에는 다중 pulse 여기신호방식, random code 벡터여기신호방식등에 대한 연구가 진행되고 있으며 계산량을 줄일 수 있는 고속 알고리즘의 개발이 주된 과제가 되고 있다.

본 절에서는 음성발생모델에 기초한 합성음질 개선에 대한 연구중 최근 연구가 활발히 진행중인 분야로 spectral envelope에 관련하여 noise spectral shaping에 대한 연구와 여기신호 모델링에 관련하여 다중 pulse 여기신호방식과 random code 벡터여기신호방식에 대하여 기술하고 이외 다른 음성발생모델 parameter들을 효율적으로 부호화하는 방법의 벡터양자화 방법에 대하여 다룬다.

8.1 Noise Spectral Shaping

일반적인 파형부호화기나 선형예측기를 사용하는 음성부호화기의 경우 평균자승오차 또는 예측오차를 최소화하는 방향으로 음성을 부호화한다. 이와 같이 하여 추출 전송된 음성부호화 parameter들을 이용하여 음성을 합성하게 되면 합성음성에는 백색잡음이 부가된다.

높은 전송속도를 갖는 파형부호화기의 경우는 신호대잡음비가 비교적 높기 때문에 수신측에서의 합성음성에 포함된 백색잡음이 전체 부호화기의 음질에 그 영향을 크게 느끼지만 저 전송속도의 부호화기에서는 이 백색잡음이 합성음질에 큰 영향을 미친다.

Noise spectral shaping은 저 전송속도 음성부호화기에서 이 백색잡음의 영향을 감소시키려는 목적으로 연구되고 있는 방법으로 사람의 청각특성의 일부를 이용하여 잡음의 효과를 줄이는 방향으로 연구가 진행되고 있다. 즉 청각특성의 잡음에 대한 masking효과를 이용하여 잡음의 spectrum을 음성신호의 spectrum과 유사하게 되도록 변형시킴으로서 잡음의 subjective한 효과를 줄이도록 하고 있다.

Noise spectral shaping의 예로 이를 APC 부호화기에 적용한 것을 그림 16에 나타내었다. 이 그림에서 B(z)가 noise shaping을 위한 feedback filter로 추가된 것을 제외하면 음성합성부분의 구조는 전과 변함이 없다. 즉 전송속도는 일대대수 유지하면서 noise shaping 효과를 얻을 수 있다.

8.2 다중 펄스 여기 신호방식

초기의 음성발생 모델을 사용한 경우 합성음질 저하의 원인은 하나로 유·무성을 분석에 의한 여기신호 모델을 얻을 수 있다. 음성신호는 실제로는 유·무성음의 완벽을 강화해 얻을 수 있는 신호 부분이있는 반면 유·무성음의 전이 부분 또는 유성음화된 무성음 신호 등 유·무성음 구분을 명확히 할 수 없는 신호가 많이 존재한다. 따라서 이와 같은 문제를 해결하기 위한 방법으로 혼합 여기 신호를 사용하거나(그림 17) 다중 펄스 여기신호를 사용하는 방법이 제안되어 연구되고 있다.

혼합여기신호 사용은 음성신호중 유·무성음의 성질이 어느 정도 포함되어 있는 가중 추출하여 여기신호를 만드는 방법인 반면 다중펄스 여기신호 방식은 유·무성음의 구분이 없이 적절한 개수의 pulse로 하frame의 음성신호에 대한 여기신호를 발생시킨다. 즉 유·무성음의 판별을 필요로 하지 않으므로 유·무성음 판별 오류에 대한 음질저하는 일어나지 않는 반면 여기신호의 pulse 위치 설정에 많은 주의가 필요하다.

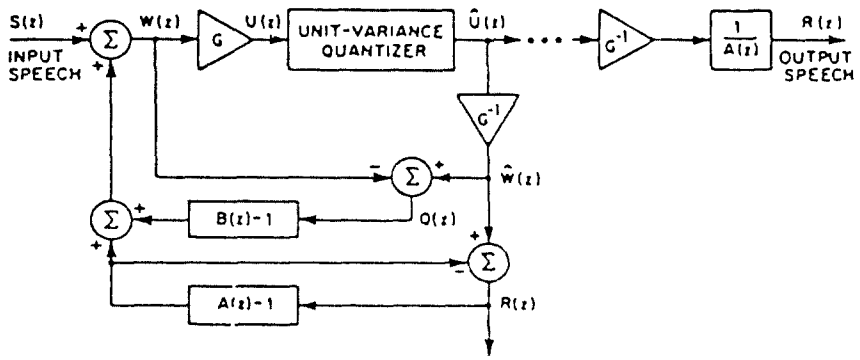


그림 16. APC-NS 시스템의 구현 예

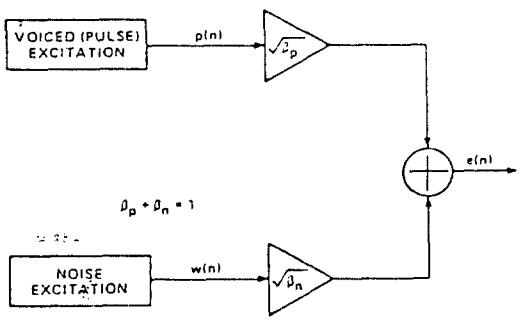


그림 17. 혼합여기 신호 발생기

일반적으로 한 pitch 주기정도의 구간에 8개의 pulse를 사용하며 pulse의 최적위치 및 크기는 abalysis-by-synthesis 방식으로 결정한다. 이와 같은 다중펄스 여기신호방식의 vocoder 구조를 그림 18에 나타내었다.

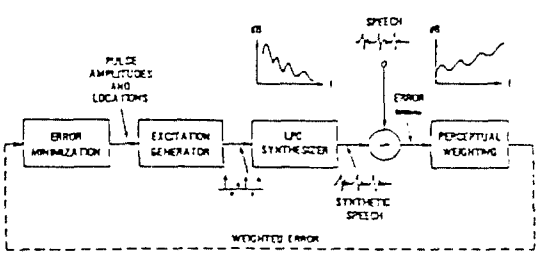


그림 18. 다중펄스 여기신호방식 vocoder

이 그림에 나타난 error weighting 필터는 앞절에서 설명한 noise shaping 필터로 vocoder의 subjective한 음질 향상을 위하여 사용한다.

8.3 Random code 벡터 여기 신호방식

이 방식은 앞서의 다중펄스 여기신호방식과 달리 여기신호로서 Gaussian 잡음과 같은 형태의 파형 벡터들을 사용하며 어떠한 벡터를 선택할 것인가는 앞에서와 마찬가지로 analysis by-synthesis 방법을 사용한다. 이 방식에서도 합성음의 subjective한 음질을 높이기 위하여 noise weighting을 채용한다. Random code 벡터 여기신호방식에 대한 블럭도가 그림 19에 나타나 있다.

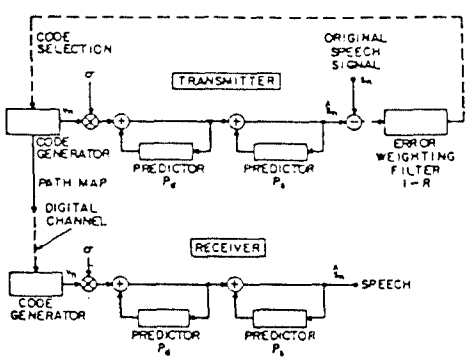


그림 19. Random code 벡터 여기신호를 이용한 vocoder의 블럭도

이 그림과 같은 구조의 vocoder를 사용할 경우 합성음질이 매우 좋은 부호화기를 구현할 수 있으나 가장 큰 문제점은 최적의 code 벡터를 찾는 과정으로 full search를 할 경우 계산량이 매우 많게 된다(1초의 음성신호처리에 Cray-1 연산속도로 125초 정도의 시간이 소요된다). 따라서 현재로서는 실제적인 방법이

되지는 못하고 있으나 최근에 이 인간속도를 높이는 방법에 대한 연구가 활발히 진행되고 있어 좋은 결과를 가져올 수 있으리라 예상된다.

8.4 벡터 양자화를 이용한 전송속도의 감축

Vocoder에서의 전송속도 감축은 단순한 음성신호 발생 모델의 가정으로 얻을 수 있으며 전송을 위한 모델 parameter들의 양자화 과정을 통하여 더욱 그 전송 속도를 줄일 수 있다. 이것은 전송 parameter들의 scalar 양자화가 아닌 벡터 양자화를 통하여 가능하며 특히 80년대 들어 효율적이고 실제적인 각종 벡터 양자화방식이 제안되면서부터 이의 응용연구가 활발히 진행되고 있다.

그림 20에 이 벡터 양자화기의 기본 개념을 표시하였다. 이 그림에서

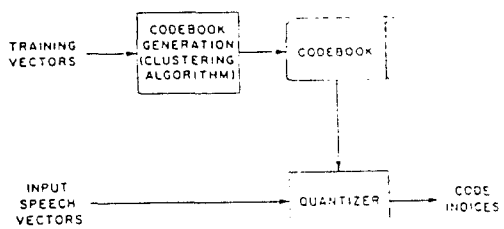


그림 20. 벡터 양자화의 기본 개념

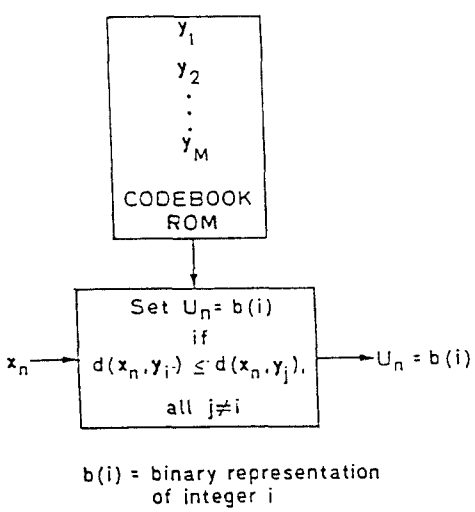


그림 21. 벡터 양자화기의 encoder 부분

보면 우선 많은 양의 training 벡터를 이용, clustering을 통하여 표준 codebook을 작성한다. 여기에서 codebook의 크기가 바로 전송속도와 직결된다. 다음 부호화하려는 입력 음성 벡터를 이 표준 codebook과 비교하여 가장 유사한 벡터를 추출한 다음 이 벡터의 index만을 전송하면 된다. 이 부호화 과정을 그림 21에 표시하였다.

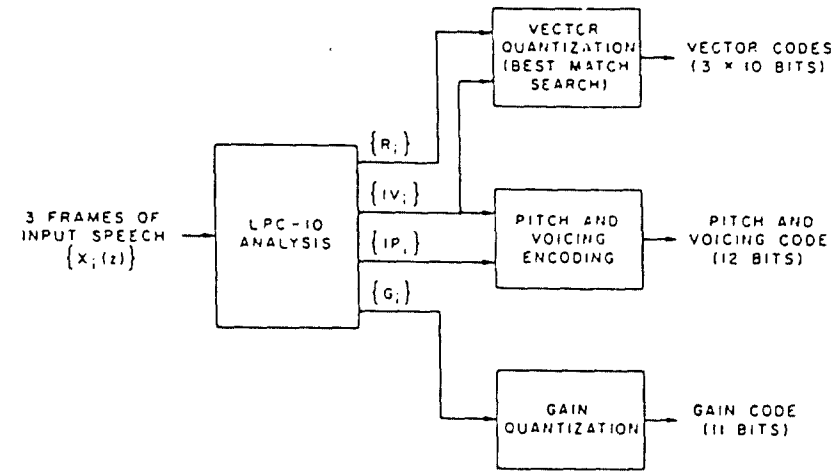
이 벡터 양자화를 처리하려는 벡터가 파형이건 또는 LPC 스펙트럼 계수이건간에 벡터간의 distortion과 centroid만 정의되면 적용할 수 있음으로 벡터 양자화를 적용하여 기존 표준 LPC 10 vocoder의 전송속도 2400 bps를 800 bps로 낮출 수 있음이 발표되었다. 이 경우 LPC vocoder의 음질은 전송 속도의 감소에도 불구하고 그대로 유지된다. 그림 22에 벡터 양자화를 이용한 800 bps LPC vocoder의 블럭도가 나타나있다.

Ⅳ. 결 론

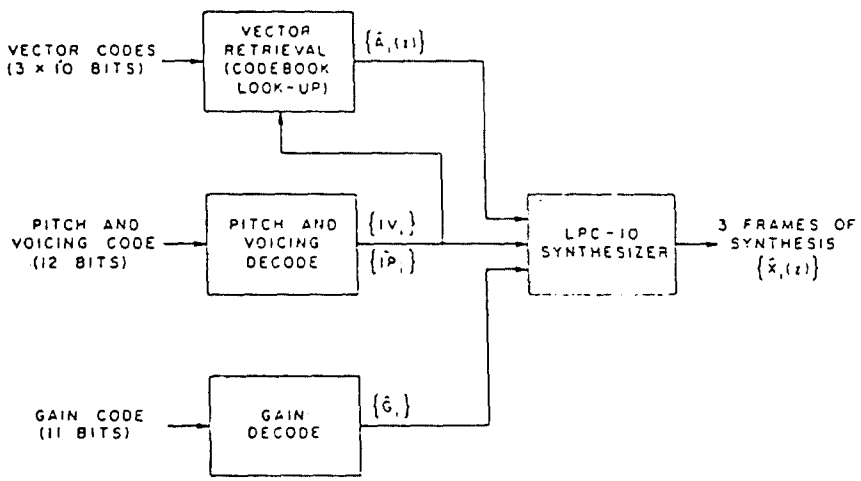
이상으로 vocoder 기술의 현황과 최근의 기술 동향에 대하여 음성발생 모델로 부터 short-time 음성해석에 기초한 channel vocoder, formant vocoder 및 현재 가장 연구가 활발한 선형 예측 부호화방식에 대하여 그 기초 및 해석 방법을 살펴보았다. 현재의 vocoding 방식은 성도의 특성을 나타내는 성도특성계수 또는 음성의 spectral envelope와 여기신호의 모델에 기초를 두고 있으므로 유성음에 대한 여기 신호로 pitch 주기 검출에 관한 문제를 그 다음에 다루었다.

Vocoder에 대한 최근의 활발한 연구에도 불구하고 음성발생 모델의 형태는 지금의 간단한 모델을 거의 그대로 사용하고 있다. 특히 선형예측방식에 의한 성도특성계수의 추출방식은 여타 방식보다 우수한 것으로 평가되어 이의 변형된 방식들에 대한 연구가 활발히 행하여지고 있다. 최근에 특히 연구가 집중되고 있는 분야는 여기신호의 모델링에 대한 것으로 종래의 유·무성음 각각에 대한 pitch 주기의 impulse 또는 random noise에서 유·무성음 분류를 필요로 하지 않는 다중펄스 여기신호 및 random code 벡터 양자화에 대한 연구로 매우 활발하여 기존 vocoder의 전송속도를 더욱 줄일 수 있게 되었다.

그러나 새로운 여기신호모델을 사용하거나 벡터 양자화를 도입한 경우의 문제점으로는 현재의 hardware 기술로는 실시간 처리가 불가능한 많은 계산량을 들 수 있다. 그러나 현재 이러한 계산은 고속화하



(a)



(b)

(a)LPC 해석 / 부호화기 (b)부호기 / 합성기

그림 22. 800bps LPC vocoder의 블럭도

려는 연구가 많이 진행되고 있고 약간의 성능저하를 감수할 경우 suboptimum 부호화 방식에 대한 연구가 수행되고 있어 멀지 않아 음질이 대폭 향상된 vocoder가 출현하리라 예상된다.

본 고에서는 저 전송속도도 음성부호화기의 기본이 되는 신호처리 알고리즘에 대하여만 언급하였다. 여기에서 언급된 방식들은 최근에 들어 디지탈 이동통신, 전화선을 이용한 디지탈 음성전송등에 관하여는 지면관계로 언급하지 못하였으나 추후 실제적인 각 부호화 방식에 대하여 따로 서술할 기회를 갖게 되기를 기대한다.

참 고 문 헌

1. J. Makhoul and M. Berouti, "Adaptive noise spectral shaping and entropy coding in predictive coding of

speech," IEEE Trans. Acoust., Speech, Signal processing, vol. ASSP-27, pp.63-73, Feb. 1979.
2. J. L. Flanagan, et al., "Speech Coding," IEEE Trans. Commu., vol. COM-27, pp.710-736, Apr. 1979.
3. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol. COM-28, pp.84-85, Jan. 1980.
4. A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp.562-574, Oct. 1980.
5. B. Gold, P. E. Blankenship, and R. J. McAulay, "New application of channel vocoders," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, 13-23, Feb. 1981.
6. B. S. Atal, "Predictive coding of speech at low bit rates," IEEE Trans. Commun., vol. COM-30, pp. 600-614, Apr. 1982.
7. D. Y. Wong, B. H. Juang, and A. H. Gray, Jr., "An 800 bit/s vector quantization LPC vocoder," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-30, pp.770-780, Oct. 1982.
8. R. V. Cox, "Recent trends in digital speech coding," in Proc. 1983 IEEE GLOBECOM, pp.23.1.1-23.1.5, Nov. 1983.
9. S. Singhal and B. S. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in Proc. ICASSP., vol.1, pp.1.3.1-1.3.4, Mar. 1984.
10. 은종관, "디지털 음성통신기술의 현황," 전자공학회 학술지, 제11권, pp.30-40, Dec., 1984.
11. R. M. Gray, "Vector quantization," IEEE Assp magazine, pp.4-29, Apr. 1984.
12. V. K. Jain and R. Hangartner, "Efficient algorithm for multi-pulse LPC analysis of speech," in Proc. ICASSP, pp.1.4.1-1.4.4, Mar. 1984.
13. S. Y. Kwon and A. J. Goldberg, "An enhanced LPC vocoder with no voiced/unvoiced switch," IEEE Trans. Acoustic., Speech, Signal Processing, vol. ASSP-32, pp.851-858, Aug. 1984.
14. M. R. Schroeder, "Linear predictive coding of speech: review and current directions," IEEE Commun. Magazine, vol. 23, pp.54-61, Aug. 1985.
15. M. R. Schroeder and B. S. Atal, "Code-excited linear prediction(CELP): high quality speech at very low bit rates," in Proc. ICASSP, pp.25.1.1-25.1.4, Mar. 1985.
16. J. M. Trancoso and B. S. Atal, "Efficient procedures for finding the optimum innovation in stochastic codes," in Proc. ICASSP., pp.44.5.1-44.5.4, Mar. 1986.
17. N. S. Jayant, "Coding speech at low bit rates," IEEE spectrum, pp.58-63, Aug. 1986.
18. R. M. Gray, "Fundamentals of vector quantization," in Proc. IEEE TENCON, pp.37.1.1-37.1.10, Aug. 1987.

이 황 수

• 1952년 9월 19일 생
• 1971년 3월 - 1975년 2월 : 서울대학교 공과대학 전기공학과(공학사)
• 1976년 3월 - 1978년 8월 : 한국과학기술원 전기 및 전자공학과(공학석사)
• 1978년 9월 - 1983년 2월 : 한국과학기술원 전기 및 전자공학과(공학박사)
• 1975년 1월 - 1975년 10월 : 한대조선공업(주) 설계부 사원
• 1983년 3월 - 1989년 2월 : 한국과학기술원 전기 및 전자공학과 조교수
• 1983년 3월 - 1992년 1월 : 한국과학기술원 전기 및 전자공학과 부교수
• 1992년 3월 - 현재 : 한국과학기술원 정보통신연구센터 정보 및 통신공학과 부교수
• 1984년 4월 - 1985년 5월 : 미국 Stanford대학교 Information Systems Lab. Post Doc. 연구원
• 연구분야 : 디지털 통신, 이동통신, 신호처리(통신, 음성, 레이다)