□ 심포지움 □

음성 신호의 디지탈 신호처리 (Digital Processing of Speech Signals)

국립과학수사연구소
김 진 현

서 론

디지탈 이라는 말은 우리 일상 생활에서 흔히 듣는다. 시계, 체온계, 체중계, 자동차의 속도계, 혈압계 등 요즘에는 디지탈 표시 제품이 아주 많아 졌다. 디지탈이라는 말을 영어사전에서 찾아보면 '손가락의', '계수형의' 등의 뜻으로 쓰여져 있다. 그리고 측정분야에서 디지탈이라는 말은 '이산적', '불연속적' 이라는 뜻이 있으며, 이것은 값이 드문 드문 있다는 의미이며, 디지탈의 특징으로 미리 정해진 자리수로만 값을 표현할 수 있는 것을 뜻한다. 디지탈에 대해 반대 의미를 갖는 것이 아날로그이다.

아날로그란 '비슷한것', '연속적' 이라는 의미가 있으며, 현재 일반적으로 흔히 사용되는 아날로그로는 후자의 '연속적' 이라는 뜻으로 사용되는 경우가 대부분이다. 음성 신호는 인간의 정보교류 수단으로서 가장 편리하고 빠른 매체로, 전기를 사용하게 된 이후 인간의 가장 주요한 통신수단으로서 사용되어 왔다. 근래에 디지탈 신호처리 기술의 발전과 함께 음성 신호의 디지탈화 및 그에 따른 여러 응용분야에 대한 연구가 활발히 진행되어 왔다. 인간에 대한 음성 신호의 정보전달 측면에 있어서의 신속성 및 사용의 용이함은 음성을 중요한 정보교환의 매체로서 자리잡게 하였다. 뿐만 아니라 인간과 기계 사이의 정보교환을 위한 MMI(Man-Machine Interface)에서 음성은 아주 오랫동안 중요한 연구분야로서 매우 활발히 연구되어 왔다.

현재 음성 신호처리 연구 분야를 크게 음성분석, 음성합성, 음성인식, 화자인식, 음성부호화의 다섯가지 세부분야로 나눌수 있다. 이 Paper 에서는 디지탈 신호처리의 기본 개념들과 자동화자인식시스템에 대하여 기술한다.

디지탈 신호처리 시스템

일반적으로 디지탈 신호처리 시스템은 Fig. 1 과 같은 구성으로 되어있다. 최근에 와서는 DSP가 범용프로세서로 사용되면서 일반 마이크로프로세서와 같은 시스템 구성을 하는 것들도 있다. 그러나 대부분이 Fig. 1과 같은 시스템 구성으로 되어 있다.

저역통과 필터 LPT1은 앤티어라어싱(Anti-Aliasing) 필터라고 부르는데 이필터가 하는 역활은 아날로그 신호에서 우리의 관심있는 부분만 통과시키고 그 밖의 신호는 제거시키는 역활을 한다. 이같은 이유는 다음에 나오는 샘플링 이론에서 다루기로 한다. A/D 변환부는 아날로그 신호를 디지탈 신호로 바꾸는 것이며 DSP 프로세서는 이 디지탈 신호를 받아 여러가지 알고리즘을 구현하여서 우리의 목적에 맞는 처리를 하는 시스템의 심장부가 되는 것이다.

D/A변환부는 복호기를 통해서 부호의 위치에 대응하는 진폭을 갖는 펄스열로 복호한다. 이 복호된 신호를 LPT2에 통하면 불필요한 높은 주파수 성분이 제거되어 원신호가 재생된다.

음향신호의 디지탈 표현

Fig. 2에서 보는 바와 같이 continuous 한 신호를 아날로그 신호라 하는데 이와 같은 아날로그 신호를

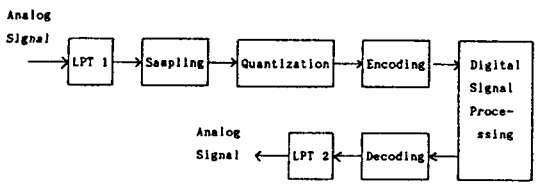


Fig. 1. Digital signal processing system.

AD converter 를 이용해서 디지탈 형태로 바꿀수 있다. 이 신호를 같은 시간 간격으로 샘플하는데 샘플값은 샘플링 순간의 진폭값과 같은 수치이다. 샘플링 처리는 그림 (b)에 보였으며 그 샘플들의 값은 그림 (c)의 표에 실었다.

샘플링 이론

Continuous signal 안에 포함되어 있는 유효한 신호 성분중에서 가장 높은 주파수를 최소한 2배이상의 rate 로 대역제한 신호를 샘플링한다면 정보를 잃지 않고 원래의 신호를 샘플들로 부터 정확히 재구성 할 수 있는데 이것을 Sampling Theorem 이라 한다. 신호에 포함된 maximum frequency 를 Nyquist frequency 라 부르는데 최소한 두배 이상의 Nyquist frequency 로 샘플링할 필요가 있다.

Fig. 5에서 $Fs \geq 2W$ 인 경우 (a)에서는 변화되지 않으나 (b)와 같이 $Fs < 2W$ 일때는 파형의 중복현상이 발생한다.

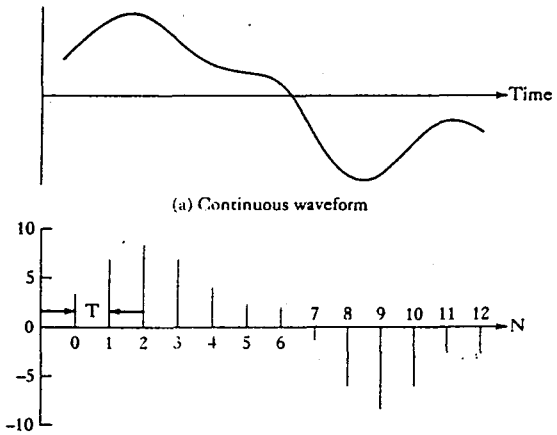
샘플링 신호를 sound 로 듣기 위해서는 DAC process 를 이용하여 아날로그 신호로 변환시켜야 하는데 원래의 대역제한 신호 스펙트럼과 겹침이 없도록 sampling frequency 를 최소한 두배의 Nyquist frequency 로 해야만 한다.

샘플링 주파수가 Nyquist Frequency 보다 낮을 경우에는 Fig. 5.(b)에서 보는바와 같이 중복현상이 발생하는데 이것을 엘리어징(aliasing) 이라 한다.

이 중복된 신호를 아날로그 신호로 변환할때 엘리어징(aliasing) 요소에 의해 찌그러짐이 발생한다.

한편 사람의 음성은 7KHz 이상의 주파수성분은 거의 갖고 있지 않으므로 양질의 음성을 얻기위한 디지탈 시스템들은 16KHz sampling rate 를 이용하고 있다. 전화 circuit는 단지 3.2KHz 의 대역폭을 가지므로 전화통화에서 당연히 통화품질이 나쁘다. 전화 시스템에서의 디지탈 기술은 그 대역폭을 제한 하게끔 되어 있기 때문에 8KHz sampling rate 을 이용하고 있다.

엘리어징(aliasing) 일그러짐을 방지하기 위해서는 A/D변환의 sampling frequency 로 부터 정해지는 Nyquist frequency 이상의 신호는 미리 A/D 변환하기 전에 없애 두던가 신호의 크기를 측정에



(b) Sample values at time interval T

N	Amplitude	N	Amplitude
0	3.38	7	-1.32
1	6.32	8	-6.08
2	7.93	9	-7.97
3	6.50	10	-5.52
4	3.73	11	-2.59
5	2.45	12	-2.41
6	1.78		

(c) Amplitude of N'th sample to 2 decimal places

Fig. 2. Uniform sampling of a continuous waveform.

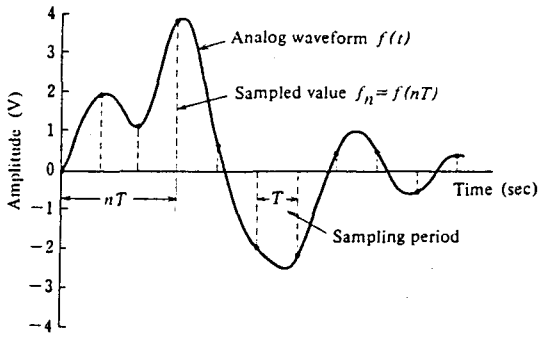


Fig. 3. Illustration of the sampling process.

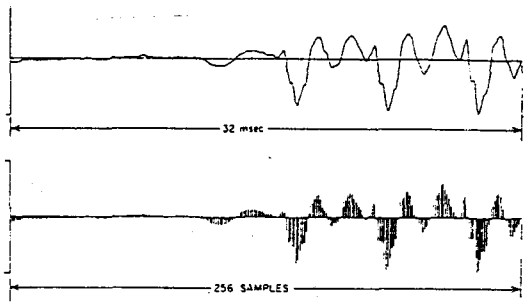


Fig. 4. Representations of a speech signal.

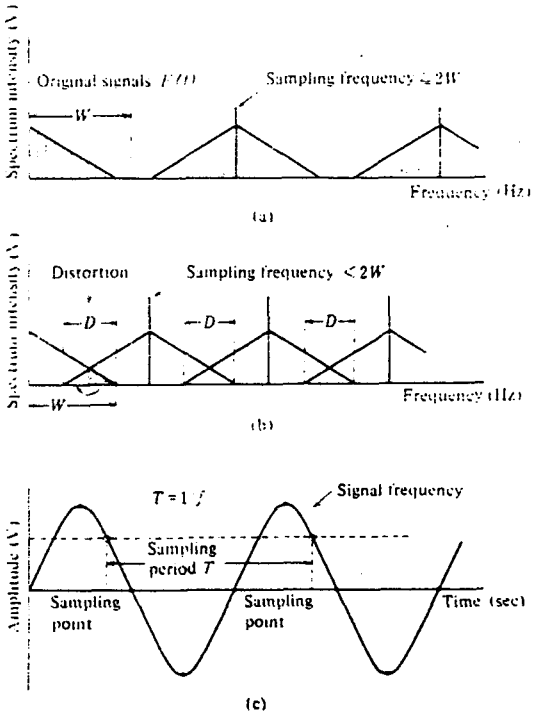


Fig. 5. Aliasing distortion by illegal sampling.(a)Properly sampled.

영향을 주지 않을 정도까지 작게 해두면 된다.

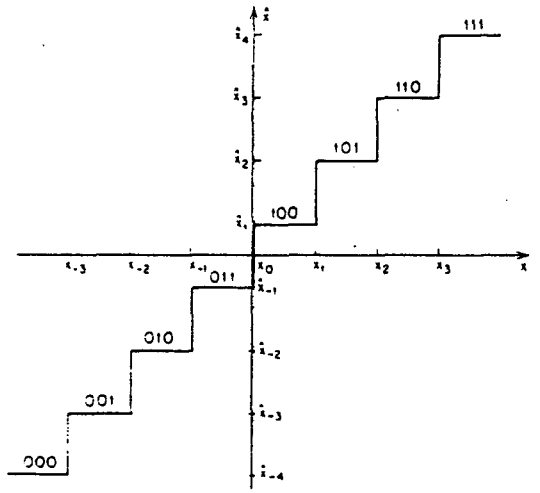
양자화(Quantization)

양자화란 표본화한 진폭을 수치로 표현하는 것이지만 진폭치를 간헐적인 스텝으로 표시한다. 따라서 표본화의 경우 시간축에 대하여 간헐적으로 진폭치를 뽑아내는 것이라 대조적으로 되어 있다. 그리고 스텝이 같은 간격의 눈금으로 되어 있는 경우를 직선 양자화라고 하며, 같은 간격이 아닌 경우를 비직선 양자화라 부르고 있다.

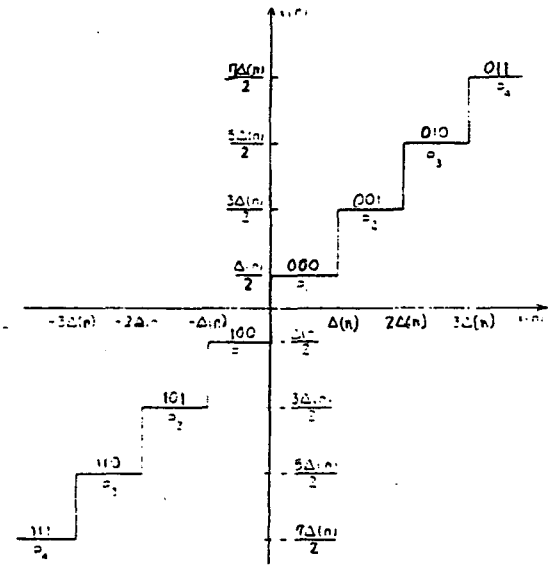
양자화의 스텝은 일반적으로는 같은 간격으로 비트수에 의하여 스텝의 간격이 결정된다. 비트수를 n이라하면 양자화 레벨수는 2 으로 표시된다.

실제로 전화처럼 narrow-band 를 이용하는 데에는 샘플당 8bit 즉 256 레벨이 적당하나 컴팩트 디스크나 디지탈 오디오 테이프와 같이 고품질을 얻기위해서는 샘플당 16bit(65,536 레벨)가 이용된다.

음성코딩 분야나 음성압축 분야는 보다 낮은 샘플링 레이트와 보다 낮은 비트로 고품질의 sound 와 music을 실현하기 위한 알고리즘 개발에 힘쓰고 있다.



(a) Input-output characteristic of a 3-bit quantizer.



(b) Input-output characteristic of a 3-bit adaptive quantizer.

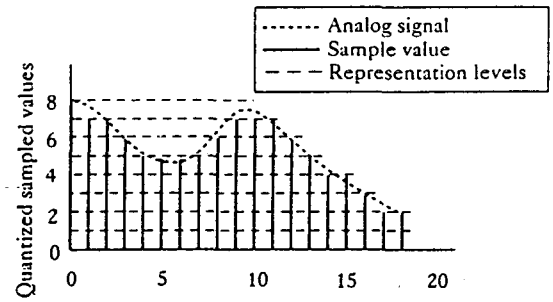
Fig. 6. Uniform and adaptive quantization.

Fig. 6에서 직선양자화와 비직선 양자화를 나타낸다. Fig. 6.(a)에서는 입력과 출력의 관계가 직선으로 되어 있으므로 직선 양자화라고 부른다. 이것에 대하여 그림 6.(b)의 경우는 입력과 출력의 관계가 직선이 아니다. 다시 말해서 입력진폭이 크게 되면 점점 스텝이 거칠게 된다. 이것은 입력신호 진폭이 작을때는 자세한 눈금이 되지만 점점 입력이 크게 되면 눈금이

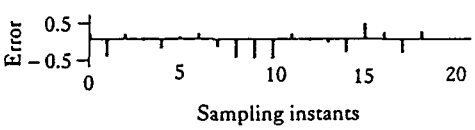
거칠게 된다. 이 경우 신호 진폭이 클때는 그것만큼 변형이나 S/N 이 나쁘게 되지만 실제로는 청감상의 문제는 나타나지 않는다. 결국 비트수를 절약할 수 있다는 뜻으로 음질상은 절약하지 못하는 경우와 거의 같다는 뜻이다.

Fig. 7.(a) 는 샘플링 순간에 아날로그 신호(점선)가 8개로 주어진 양자화 레벨(가로축 점선)중의 하나로 전환되는 것을 보이고 있으며, 샘플값들은 각 샘플 시간에 대해 수직직선으로 그리고 양자화기는 샘플링 시간의 신호값에 가장 가까운 표시레벨을 이용하고 있다. 정확하게 샘플값을 얻기 위해서는 레벨이 많으면 많을수록 그 만큼 더 아날로그 신호에 근사하게 된다.

아날로그 신호와 샘플값 사이의 차를 양자화 에러라고 하는데 Fig. 7.(b)에서 처럼 그것을 노이즈로 볼수 있기 때문에 종종 양자화 노이즈라고 부른다. 양자화 간격을 좁게 결국 스텝 간격을 좁게 해가면 아날로그 신호와 양자화 신호의 차가 작게되어 하부에 나타낸 오차신호의 진폭이 작게되는 것을 알 수 있다. 이와같이 오차신호의 진폭을 작게 하려면 비트수를 늘려가는 것이 좋다. 한편 표본화 간격의 폭은 어떻게 될까? 실은 이 경우도 표본화 간격을 좁게 해가면 원래 아날로그 신호와 비슷해 지므로 양자화 오차는



(a) Quantized sampling with 8 representation levels (3 bits per sample).



(b) Quantization error introduced by sampling in (a).

Fig. 7. Quantization error arising from finite spacing between representation levels.

작게 된다. 단 이론적으로는 나이키스트 주파수의 1.2배 이상이면 충분하므로 주로 효과가 기대되는 것은 양자화 간격의 방법이라고 생각된다. 양자화 오차는 변형이나 잡음으로 되지만 이 잡음 성분에는 꽤 넓은 주파수 대역까지 포함하고 있으므로 음성은 일반 잡음같은 것으로 또 영상 신호는 자세한 랜덤 잡음으로 되어 화면에 나타난다.

에러 검출 코드

에러 검출방법으로서 패리티 체크(parity check) 방식이 있는데, 가령 데이타를 원거리에 전송하거나 대량의 정보를 주고 받을때 주위의 조건에 따른 노이즈(noise)라든지 회로상의 잘못으로 에러가 생길 가능성이 많다. 이러한 에러를 완전히 없애기는 어렵지만 일반적으로 채용되고 있는 방법은 정보를 나타내는 부호의 여분으로 에러 검출용 비트를 하나 더 추가시켜 언제나 전체 부호속에 포함되어 있는 1의 수가 홀수 또는 짝수개로 되도록하는 것이다. 만약 이때 에러가 생기면 이 규칙이 무너지기 때문에 에러가 생겼다는 것을 곧 알게 된다. 이와 같은 방법을 일반적으로 패리티 체크 방식(parity check method)이라고 하며 이때 사용된 검출용 비트를 패리티 비트(parity bit)라고 한다.

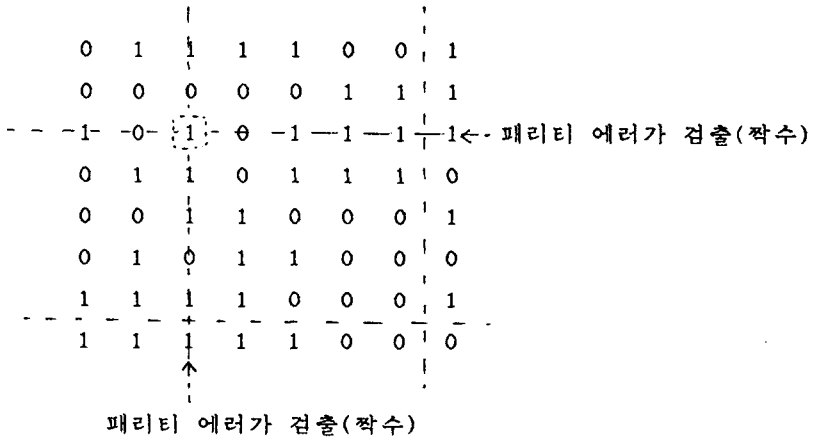
만약 3행째의 코드에서 10001111 에 에러가 발생해서 10101111 로 되었다고 하자. 이런 경우는 아래에서 보는 바와 같이 동시에 홀수 패리티 체크 비트와 홀수 패리티 체크 워드로 검출되어 나타난다.

자동 화자인식 시스템

화자인식(Speaker Recognition)에는 크게 화자 확인(Speaker Verification)과 화자 식별(Speaker Identification)의 두 분야가 있다. 이 두분야는 모두 N명에 대한 참조패턴(reference pattern) 데이타베이스를 갖고 있다는 점에서는 서로 같지만, 어떠한 일을 하느냐에 따라 다음과 같이 구별된다. 화자 확인은 입력 신호로서 음성 신호와 그 음성 신호의 화자에 대한 Identity 가 같이 주어지며, 시스템은 그 Identity 에 해당되는 화자에 대한 참조패턴(reference pattern) 과 입력된 음성 신호가 일치하는지를 검사하여 일치 또는 불일치라는 판정을 내리게 된다.

2진화 10진 코드와 패리티 비트

10진수	2진화 10진 의 홀수 패리티					1의 합계
	2^3	2^2	2^1	2^0	패리티 비트	
0	0	0	0	0	1	1
1	0	0	0	1	0	1
2	0	0	1	0	0	1
3	0	0	1	1	1	3
4	0	1	0	0	0	1
5	0	1	0	1	1	3
6	0	1	1	0	1	3
7	0	1	1	1	0	3
8	1	0	0	0	0	1
9	1	0	0	1	1	3
10	1	0	1	0	1	3
11	1	0	1	1	0	3
12	1	1	0	0	1	3
13	1	1	0	1	0	3
14	1	1	1	0	0	3
15	1	1	1	1	1	5



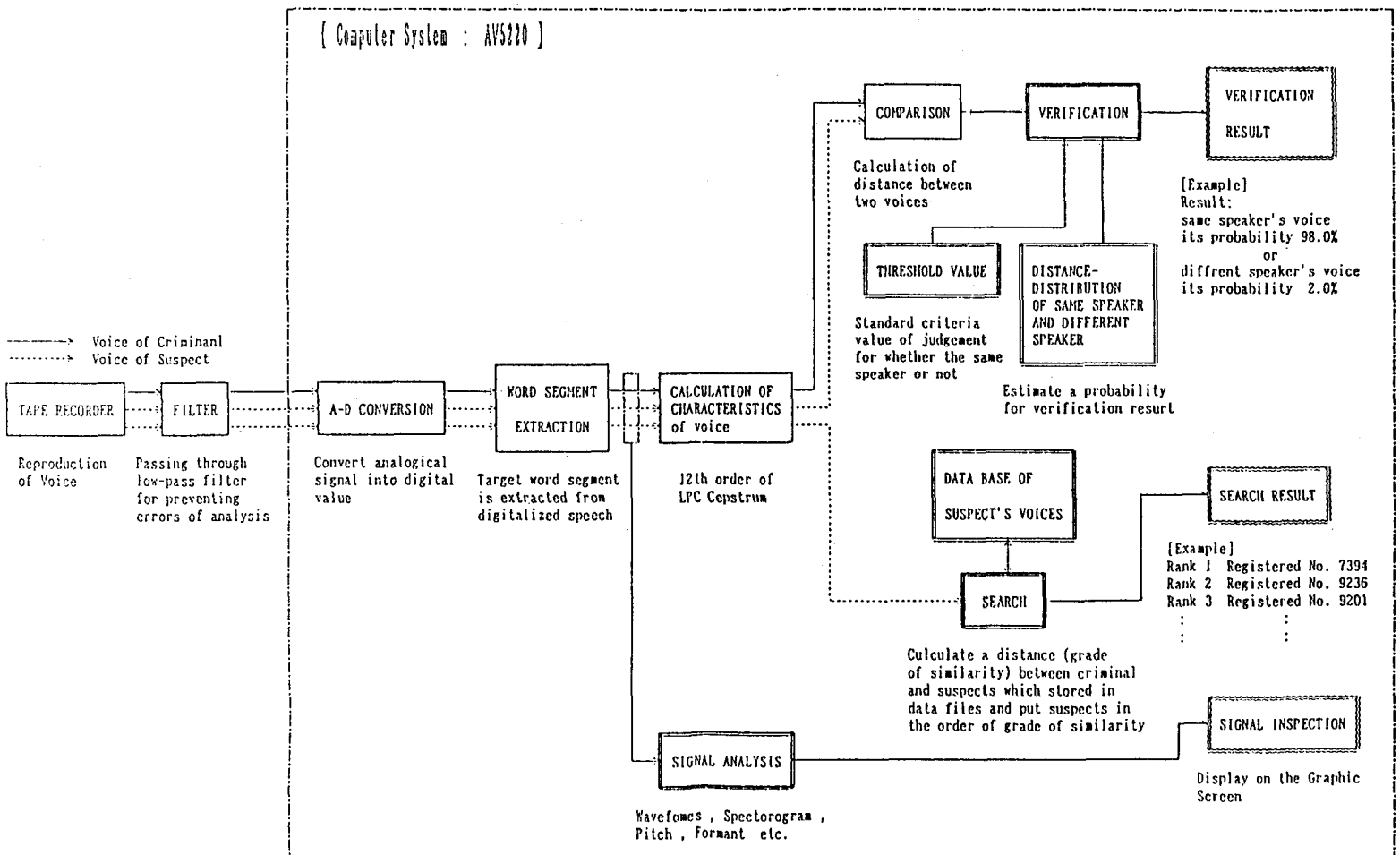
반면, 화자 식별은 입력신호로서 들어온 음성 신호가 검토 대상인 N명 중 누구에게 해당되는 지를 찾아 해당 화자가 누구인지를 알려주게 된다.

이때 입력신호가 검토대상인 N명 중 어느 누구에게도 해당되지 않을 경우도 고려해야 한다. 이처럼 대상외의 화자가 발성할 가능성을 고려한 경우를 open-set 화자 구성이라 하고, 발성할 화자가 반드시 N명 내에 있다는 가정하에 구현된 시스템은 close-set 화자 구성이라 한다.

특징벡터는 화자의 개인성 정보를 충분히 표현할 수 있어야 하며, 화자간 유사성을 평가하는 척도는 화자내의 변이를 수용하면서도 화자사이의 변이는 최대화할 수 있는 것이어야 한다.

좋은 특징은 사용자내(Intra speaker)에서는 그 변

Block Diagram of Automatic Voice Identification System



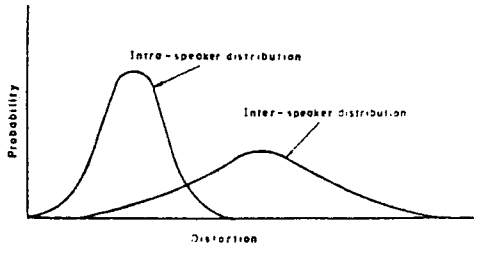


Fig. 8. 에러율과 문턱치 결정의 관계.

이가 작고 사용자간(Inter speaker) 의 벡터 사이에서는 그 변이가 크다. 즉, Fig. 10와 같은 분포를 갖는 특징벡터가 좋은 벡터라 할 수 있다.

좋은 특징벡터가 갖추어야 할 조건은,

① 화자간 정보를 효과적으로 나타낼수 있어야 하고

② 측정이 쉬워야 하고

③ 시간에 대해 안정해야 하고

④ 음성신호에서 자연적이고 반복적으로 일어나야 하며

⑤ 발성환경이 바뀌어도 영향이 적어야 하고

⑥ 흉내를 허용하지 않아야 한다

그러면 실제로 구성하는 화자인식 시스템에 어떤 특징 벡터를 선택할 것인가 하는 문제가 생기게 된다. 이를 해결하기 위해서는 특징의 효용성을 나타내는 척도를 정의해야 한다. 널리 사용되는 효용성 측정 척도로 F-ratio 가 있다. F-ratio 는 다음과 같이 정의된다.

F = inter-speaker variance / intra-speaker variance

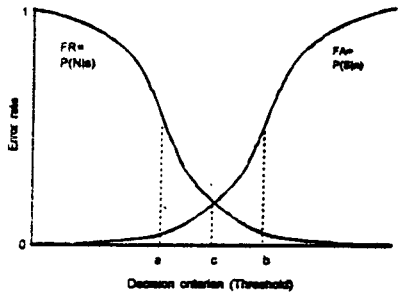
$$= \left[\frac{1}{m-1}\sum_{i=1}^{m}(\mu_i - \mu)^2\right] \Big/ \left[\frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{k=1}^{n}(X_{ik} - \mu_i)^2\right]$$

where, $\mu_i = \frac{1}{n}\sum_{k=1}^{n} X_{ik}$; mean of each speaker

$\mu = \frac{1}{m}\sum_{i=1}^{m}\mu_i$; mean of all speaker

X_{ik} : feature parameter

즉, 전체 화자들의 평균치의 변이 값을 화자 각각의 변이의 평균으로 나눈것으로 정의된다. 이는 Fig. 10의 화자내 편차로 화자간의 편차를 나눈 효과를 갖는데 화자내 편차는 작을수록 좋고 화자간의 편차는 클수록 좋다. 그러므로 F-ratio 값이 크면 좋은 특징이라 말할 수 있고 값이 작으면 성능이 좋지 못한 특징으로 생각할 수 있다. F-ratio 는 좋은 특징을 선택하는 데 보다 나쁜 특징을 가려내는 목적으로 많이 사용된다.



FR(False Rejection), FA(False Acceptance), N:거부, S:확인, n:사칭자, s:올은 화자

Fig. 9. 에러율과 문턱치 결정의 관계.

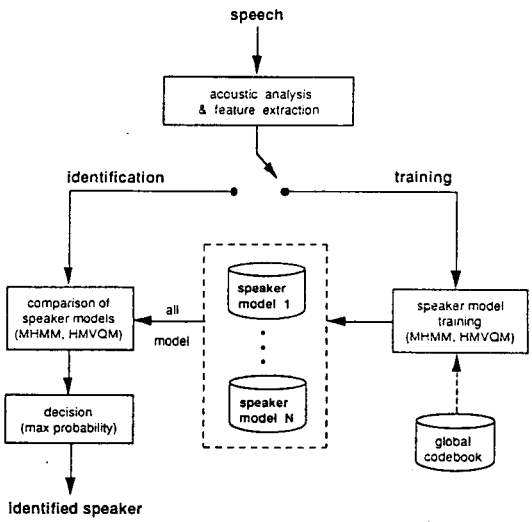


Fig. 10. 화자식별 시스템 구성도.

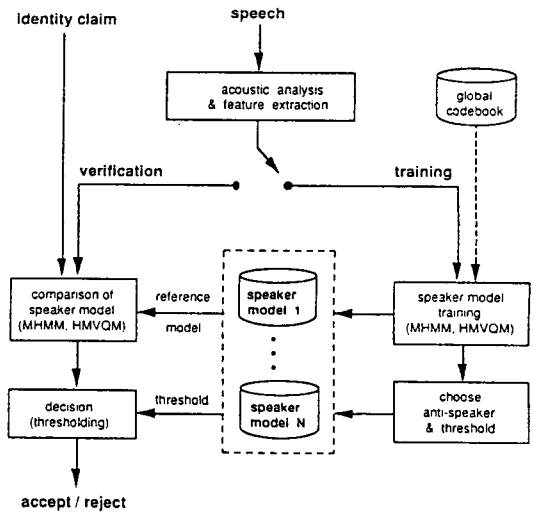


Fig. 11. 화자확인 시스템 구성도.

화자확인 시스템의 에러는 false rejection 과 false acceptance 의 두 가지 에러로 나눌 수 있다. false rejection 은 옳은 사용자에 대해 잘못 거부하는 것이고 false acceptance 는 반대로 사용권한이 없는 화자를 올바른 사용자로 오인식 하는 것이다. 이 두가지 에러는 서로 trade-off 관계에 있다. 즉, false acceptance 에러율을 줄이기 위해 유사성 비교의 문턱치를 낮추어 주면 false rejection 에러율이 증가하고 반대로 false rejection 에러율을 줄이기 위해 문턱치를 너무 높이면 false acceptance 에러가 증가하게 된다. 그림 9에 두 에러 사이의 관계를 보이고 있다.

그러므로 시스템 구성시 이러한 관계를 잘 고려하여 적절한 문턱치를 가질수 있도록 신중한 실험이 필요하다. 아울러 이는 시스템의 사용 용도에 따라 특정 에러를 줄이는 방향으로 결정할 수 있다. 예를 들어 철저한 보안이 요구되는 경우엔 false rejection 에러율이 좀 높아 지더라도 false acceptance 에러율을 특정기준 이하로 떨어뜨려 놓아야 한다.