유·무성음 및 묵음 식별에 관한 연구

김 명 환 유 형 근 김 순 협
경운대학 대학원 전자공학과 OPC중앙연구소 경운대학 전자계산기 공학과

A Study on the Voiced,Unvoiced and Silence Classification

KIM MYUNG WHAN RYU HYEUNG KEUN KIM SOON HYOB
Kwang Woon University OPC R&I Kwang Woon University

(Abstract)
This paper reports on a Voiced-Unvoiced-Silence Classification of speech for Korean Speech Recog nition.
In this paper, it is describe a method which uses a Pattern Recognition Technique for classifying a given speech segment into the three classes.
Best result is obtained with the combination using ZCR,P1,Ep and classification error rate is less than 1%.

1. 서론

주어진 음성파형을 유성음,무성음,묵음으로 식별하기 위하여 일반적으로 피치분석(Pitch analysis)으로 행하여져 왔으나 몇가지 큰 단점이 있다.
첫째,유성음은 주기적인신호라는 단순한 특징에 기본을 두었는데 성대진동의 갑작스러운 변화는 유성음일지라도 비주기적인 음성파형을 만든다.
둘째,음성합성분야에서는 피치분석이 가능하나, 음성분할(Speech Segmentation),음성인식(Speech Recognition)과 같은 음응분야 에서는 복잡성 분만아니라 유성음과 무성음 경계부분 에서는 대우 나쁜수행을 초래한다.
따라서, 본 논문은 3가지음성신호를 식별하기 위하여 패턴인식방법을 사용하였다.
한편 본연구에 사용된 분석파라메타는 음성신호의 영고 차율,대수 에너지,자동상관계수,선형예측 분석에서 얻은 첫번째 예측계수,예측오차의 에너지 이다.
그리고 유사도를 측정하기 위한 거리측정법은 Mahalanobis 거리측정법을 사용하였고 3가지 음성신호의 판정은 계산된 거리가 최소일때를 선택하여 결정하였으며 분석을 위한 음성데이타는 경운대학 디지탈 신호 처리연구실의 시스템에 의해 전처리 되었으며 전체적인 식별시스템은 그림 1 과 같다.

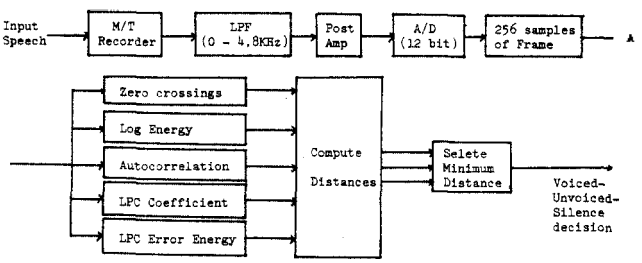


그림 1. 한국어 음성식별에 관한 시스템 블록도.
Fig 1. A Block diagram of the Spoken Korean Speech Classification System.

2. 음성신호의 분석이론

가) 프레임 길이와 프레임 간격
프레임: 음성신호를 한번 처리하는 기본단위
프레임길이: 프레임에 포함되어있는 샘플 수.
본 연구에는 프레임길이는 256 샘플

나) 영고 차율

$$Zn=\sum_{m=-\infty}^{\infty}\left|Sgn\left[x(m)\right]-Sgn\left[x(m-1)\right]\right|\cdot W(n-m) \quad ...(1)$$

유성음은 스펙트럼의 하향때문에 에너지는 낮은 주파수에 집중되어 있다.
반면,무성음은 에너지의 대부분이 높은주파수에 집중되어 있으며 높은주파수는 높은영고 차율을 낮은주파수는 영고 차율이 낮다.
그림 (2)는 각음성에 대한 영고 차율을 나타낸그림이다.

다) 대수 에너지
대수 에너지는 샘플값 자승의 합에 대수를 취한것으로

$$LEn = 10 * \log \sum_{m=-\infty}^{\infty} x^2(m)\cdot W(n-m) \quad .....(2)$$

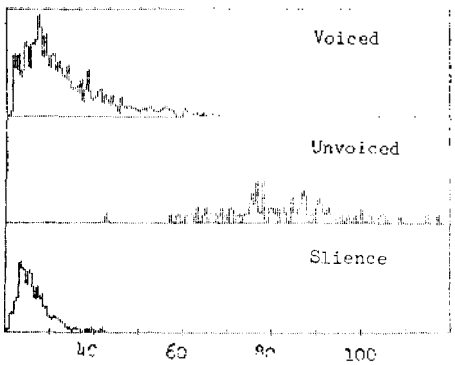


그림 2(a). 25.6ms 당 각음의 영고 차율의 통계분포

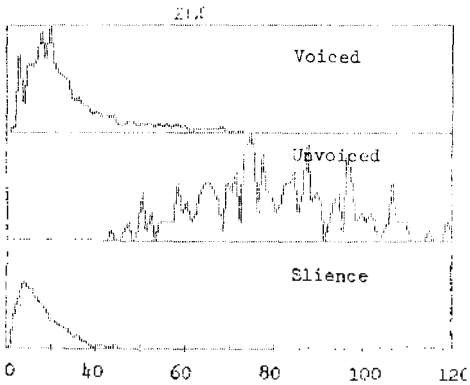


그림 2(b). 12.8ms 당 각음의 영고 차율의 통계분포

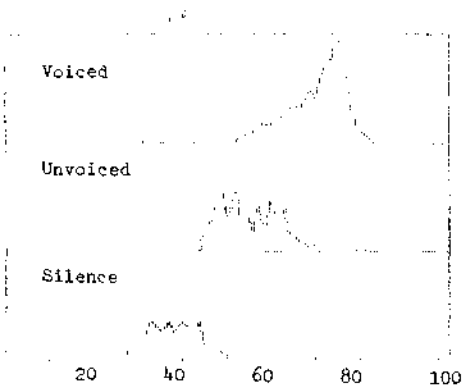


그림 3. 25.6 ms당 각음의 대수 에너지의 통계분포

다) 자동 상관계수

$$C1 = \frac{\sum_{m=1}^{N} x(m)\cdot x(m-1)}{\sqrt{\sum_{m=1}^{N}\left(x^2(m)\right)\cdot\sum_{m=0}^{N-1}\left(x^2(m)\right)}} \quad \cdots\cdots(3)$$

자동 상관계수는 인접한 샘플의 상관관계를 나타내며 정의에 의하여 1 - -1 사이에 변화한다.

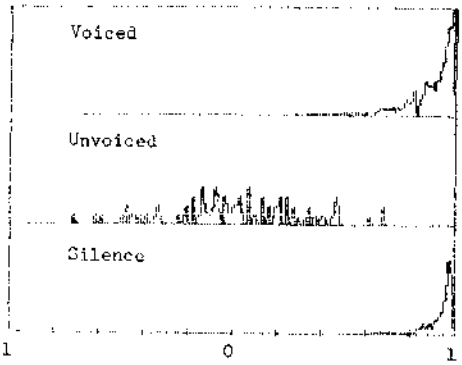


그림 4. 25.6 ms 당 각음의 첫번째자동상관계수의 분포

라) 선형예측법

한음성샘플을 각각음성샘플들의 선형결합으로 근사화 할수있다는 데 있으며 선형시변 시스템을 특성화하는 매개변수를 추정하는데 확실하고 정밀한 방법 이다.

$$S(n)=\sum_{m=1}^{p} a_k\cdot s(n-k) + G\sum_{l=0}^{q} b_l\cdot u(n-l), b_0=1 \quad \cdots(4)$$

(1) 예측계수

평균제곱오차 E_p를 최소화 함으로써 얻을수 있다.

$$E_p = \sum_{m=-\infty}^{\infty} e_n(m)^2 \quad \cdots\cdots(5)$$

$$= \sum_{m=-\infty}^{\infty}\left(s_n(m) - \sum_{k=1}^{p} a_k\cdot \hat{s}_n(m-k)\right)^2 \quad \cdots\cdots(6)$$

여기서 E_p 가 최소가 되도록 하려면

$$\frac{\partial E_p}{\partial a_i} = 0 \quad i=1,2,3\ldots\ldots,p \quad \cdots\cdots(7)$$

식(6),(7) 을 풀면

$$\sum_{m=-\infty}^{\infty} s_n(m-i)\cdot s_n(m) = \sum_{k=1}^{p} a_k \sum_{m=-\infty}^{\infty} s_n(m-i)\cdot s_n(m-k) \quad \cdots(8)$$

(2) 예측오차의 에너지

$$E_p = \sum_{m=-\infty}^{\infty} s_n^2(m) - \sum_{k=1}^{p} a_k \sum_{m=-\infty}^{\infty} s_n(m)\cdot s_n(m-k) \quad \cdots\cdots(9)$$

3. 식별 알고리즘

확률밀도함수 $P(X/W_i)$

$$P(X/W_i) = \frac{1}{(2\pi)^{n/2}\left|C_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(X-\hat{m}_i)^t C_i^{-1}(X-\hat{m}_i)\right\} \quad \cdots(10)$$

식(10)에서 정규밀도함수의 지수형을 단조증가함수인 자연대수로 취하고 최적결정함수 $\hat{d}_i(X)=P(X/W_i)P(W_i)$에 대입하면 다음식과 같다.

$$\hat{d}_i(X)=\ln P(W_i)-\frac{1}{2}\ln|C_i|-\frac{n}{2}\ln 2\pi-\frac{1}{2}\left[(X-\hat{m}_i)^t C_i^{-1}(X-\hat{m}_i)\right] \quad \cdots\cdots\cdots\cdots\cdots\cdots (11)$$

또 윗식의 처음3항들은 측정치벡타 와 무관함으로 다음식과 같이 간단하게 쓸수있다.

$$\hat{d}_i(X)=(X-m_i)^t\, c_i^{-1}\,(X-m_i) \quad \cdots\cdots\cdots(12)$$

윗식은 측정치벡타 X와 평균벡타 $\hat{m}_i$ 사이의 거리인데 통계적인 특징이 분명히 고려되기 때문에 유사도를 측정하는데 유용한 식이며 이를 Mahalanobis거리측정법이라고 한다.

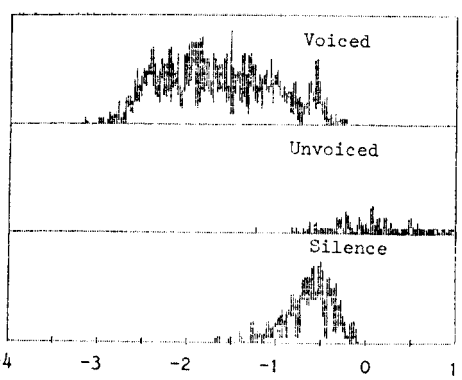


그림5. 첫번째 예측계수에 관한 통계분포

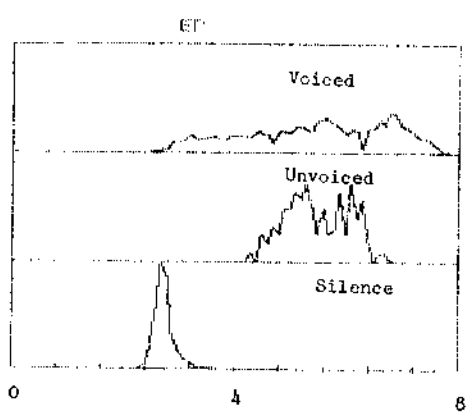


그림 6. 예측오차의 에너지에 관한 통계분포

4. 실험 및 결과

성인남성 3명이 발성한 한국어 숫자음(/영/-/구/)를 7번 반복한음성중 임의로 선정한 154개숫자음을 대상으로 실험하였으며 전체적인 식별시스템은 그림7과 같다.

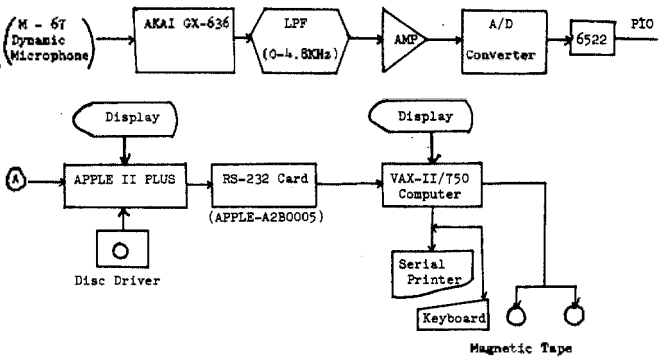


그림 7 한국어 음성분석 및 식별에 관한 하드웨어 시스템.

Fig. 7 A Hardware System for Korean Speech analysis and Classification.

(1) 표준집합 설정

본연구는 확자독립방법으로 음성을 식별하였기 때문에 표준집합설정은 2명의 남성화자가 발성한 숫자음 파형을 손수식별하여 3가지음성신호부류로 구분하고 각부류에 대한 평균치,표준편차,공분산 행렬을 구하였다. 다음 그림은 표준집합에 사용한 음성파형 이다.

그림 8. 표준집합에 사용한 숫자음 /칠/의 음성파형

표 1. 표준집합에서의 평균치,표준편차,공분산행렬

Table 1. Means, Standard Deviations, and Covariance Matrices in Reference set.

(1) Voiced(1-5)	Zero Crossing	Log Energy	First Autocorrelation	First LPC	LPC Log Error
Mean	-5.7462	12.0885	0.887	-1.013	5.46539
Standard Deviation	17.9169	8.06634	0.12053	0.6884	1.26081
Covariance Matrix	1	0.[illegible]	-0.9519	0.5210	0.01563
	0.[illegible]	1	-0.3358	0.9510	0.87403
	-0.[illegible]	-0.[illegible]	1	-0.6220	-0.6409
	0.5210	0.9510	-0.6220	1	0.38242
	0.01563	0.87403	-0.6409	0.38242	1
(2) Unvoiced (6-12)					
Mean	12.[illegible]179	10.0719	0.05178	-0.0246	5.48372
Standard Deviation	[illegible].[illegible]61	6.[illegible]730	0.10666	0.41399	0.56852
Covariance Matrix	1	0.5681	-0.8301	0.5090	0.6383
	0.5681	1	-0.3109	-0.097	0.9601
	-0.8301	-0.[illegible]349	1	-0.7553	-0.3737
	0.5090	-0.097	-0.7553	1	-0.0564
	0.6383	0.9601	-0.3737	-0.056	1
(3) Silence (13-)					
Mean	13.2095	11.3968	0.9601	-0.5879	2.7581
Standard Deviation	8.60351	3.79419	0.0115	0.25377	0.1916
Covariance Matrix	1	-0.210	-0.560	0.07120	0.6511
	-0.210	1	0.6403	-0.8016	0.3603
	-0.560	0.6403	1	-0.4056	-0.07784
	0.07120	-0.8016	-0.4056	1	-0.1090
	0.6511	0.3603	-0.0778	-0.3090	1

(2) 결과 및 고찰

(가) 1개의 파라메타로 식별하였을 때

Measured parameter	V → U	V → S	V → T	Total	Error Rate
ZCR	12	490	26	528	14.3%
LE	143	78	30	251	6%
C1	6	677	48	731	19.6%
P1	66	473	41	580	16.2%
Ep	466	88	0	543	14.7%

표 2. 각각의 파라메타로 식별했을 때의 전체식별오차
Table 2. Total Classification error using each parameter.

(나) 2개의 파라메타로 식별하였을 때

Measured parameter	V → U	V → S	U → S	Total	Error Rate
ZCR,LE	54	46	5	105	2.8%
ZCR,C1	20	565	1	586	15.9%
ZCR,P1	15	272	16	237	6.4%
ZCR,Ep	1	81	2	69	2.1%
LE,C1	31	43	2	76	2%
LE,P1	396	37	8	341	9.2%
LE,Ep	60	42	6	106	2.8%
C1,P1	32	[illegible]	16	314	8.6%
C1,Ep	9	106	3	117	3.1%
P1,Ep	560	47	3	607	16.4%

표 3 2개의 파라메타를 조합하여 식별 하였을때
Table 3 Total Classification error using two parameter.

(다) 3개의 파라메타로 식별하였을 때

Measured parameter	V → U	V → S	U → S	Total	Error rate
ZCR,LE,C1	37	59	5	96	2.6%
ZCR,LE,P1	61	46	6	113	3%
ZCR,LE,Ep	3	54	2	59	1.5%
ZCR,C1,P1	34	47	17	78	2%
ZCR,C1,Ep	5	73	3	80	2.1%
ZCR,P1,Ep	1	35	3	39	1.05%
LE,C1,P1	93	54	5	96	2.6%
LE,C1,Ep	[illegible]	47	2	56	1.5%
LE,P1,Ep	430	32	7	569	15.3%
C1,P1,Ep	7	59	3	69	1.8%

(라)4.5개의 파라메타로 식별하였을 때

Measured parameter	V → U	V → S	U → S	Total	Error Rate
ZCR,LE,C1,P1	[illegible]	14	7	98	2.6%
ZCR,LE,C1,Ep	3	31	3	86	2.3%
LE,C1,P1,Ep	0	56	0	56	1.5%
ZCR,C1,P1,Ep	1	24	1	30	0.8%
ZCR,LE,P1,Ep	0	37	3	96	2.6%
ZCR,LE,C1,P1,Ep	3	24	1	98	2.6%

표 5. 4,5개의 파라메타로 식별하였을때 전체오차
Table 5. Total Classification error using four or five parameters.

전반적으로 많은 파라메타를 조합하여 식별하였을때 3% 보다도 적은 식별오차를 얻었다.

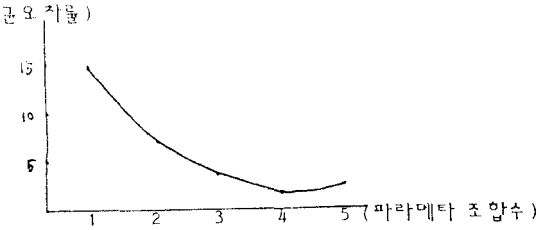
오차의 대부분은 끝점부분에서 유성음구간을 묵음으로 식별하였고, 유성음과 무성음의 경계부분에서도 오차가 발생하였다.

각 파라메타조합수와 평균오차률및 최소오차률을 살펴보면

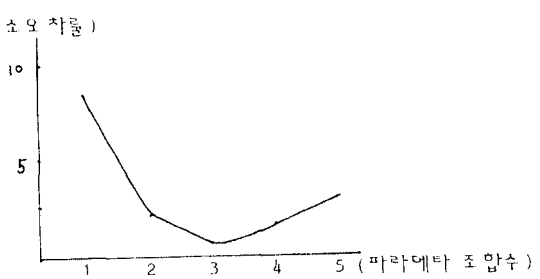
파라메타 조합수	평균오차률	최소오차률
1	14.6 %	8 %
2	7.18 %	2 %
3	3.3 %	0.95 %
4	2.2 %	1.5 %
5	2.6 %	2.6 %

표 6. 파라메타 조합수와 평균오차률및 최소오차률

(평균오차률)



(최소오차률)



5. 결론

본 논문은 한국어 음성인식을 위하여 음성구간을 유성음,무성음,묵으로 식별하기위하여 패턴인식방법(Pattern Recognition Technique)을 사용하였다.
첫째,특징집합은 음성분석에서 일반적으로 사용되는 음성신호의 영고차률,대수에너지,첫번째 예측계수, 첫번째 자동상관계수,예측오차의 에너지를 사용하였으며,
둘째, 표준집합은 3가지 음성신호부류에 대하여 손수 식별한 음성데이타들로 부터 얻은 평균벡타와 분산을 통계적으로 처리 하였으며
셋째,유사도 거리측정법은 각파라메타들의 상관관계 및 통계적인 특징이 분명히 고려되는Mahalanobis 거리측정법을 사용하였으며
마지막으로, 3가지음성신호 부류판정은 각음성에 대한 연습집합과 표준집합과의 계산된거리가 최소일때를 선택하여 결정 하였다.
그리고 각파라메타들을 여러가지로 조합하여 식별한 결과 영고차률,선형예측분석에서 얻은 첫번째예측계수 예측오차의 에너지를 파라메타로 수행하였을때 다른조합방식보다도 가장좋은 결과를 얻었으며 그때의 식별오차는 1%보다도 적었다.

참고문헌

(1) L.Rabiner, M.Sambur, and C.Schmidt,"Applications of a nonlinear smoothing algorithm to speech processing,"IEEE Trans.Acost.,Speech,Signal processing, vol. ASSP-23, PP. 552-557, Dec. 1975.

(2) B.S.Atal and L.R.Rabiner,"A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition,"IEEE Trans. Acoust.,Speech,Signal processing,vol. ASSP-24, pp.201-121,June 1976.

(3) J.J.Dubnowski,R.W.Schafer, and L.R.Rabiner,"Real-time digital pitch detector," IEEE Trans.Acoust.,Speech,Signal processing,vol.ASSP-24, pp 2-8,Feb.1976

(4) L.J.Siegel and K.Steiglitz,"A pattern classification algorithm for the voiced/unvoiced decision," in 1976 ICASSP Proc.,1976, pp. 326-329.

(5) L.F.Lamel, L.R.Rabiner,A.E.Rosenberg,and J.G.Wilpon,"An improved ending point detector for isolated word recognition,"IEEE Trans.Acoust.,Speech,Signal processing,vol ASSP-29,pp.777-785,August,1981.

(6) H.F.Silverman and N.R.Dixon,"A comparison of several speech-spectra classification methods,"IEEE Trans.Acoust.,Speech,Signal processing,vol ASSP-24, August,1976.

(7) L.J.Siegel and A.C.Bessey,"Voiced/Unvoiced/Mixed excitation classification of speech,"IEEE Trans.Acoust.,Speech,Signal processing,vol.ASSP-30,No.3 Jun.1982.