성도 면적 함수와 벡터 양자화를 이용한 음성 인식에 관한연구

*송제혁, 김동준, 박상희
연세대학교 전기공학과

A Study on Speech Recognition using Vocal Tract Area function and Vector Quantization

*Jei Hyuck Song, Dong Jun Kim, Sang Hui Park
Dept. of Electrical Engineering, Yonsei University

Abstract - We propose the vocal tract area function as the feature vector of speech recognition. Vocal tract area function is directly related to speech production. The vocal tract area function is not only showing mechanism of speech production but also can be used as an effective feature vector in speech recognition in this study .

1. 서론

음성 인식 시스템을 개발하는데 있어서 DTW, VQ, HMM, 신경망등의 인식 알고리즘의 개발도 중요하나 음성의 특징을 잘 반영할 수 있는 매개변수의 추출은 매우 중요한 문제다.

많은 음성 특징 추출방법이 제안되어 왔으나 그들 중 대부분은 음성 파형으로 부터 직접 얻어지는 음향학적 매개 변수에 의존한다. 그중 LPC 캡스트럼 계수가 가장 많이 쓰인다[1]. 그러나 효과적인 특징 추출 방법을 얻기 위해서는 음성 생성구조의 기본적인 연구가 중요하고 필수적이다. 따라서, 음성 생성과정에 대한 해석을 통하여 음성으로부터 성도 면적을 추출하여 성도에서의 음성 생성에 대한 연구가 많은 학자들에 의해 시도되어 왔다[2][3][4].

Wakita[5]는 모든 손실을 성문에 있게 하고 입술을 음향학적으로 합성된 회로로써 간주하고 성도를 추정하는 기법을 개발하였는데 추정된 결과는 상당히 실제 성도의 모양과 유사한 결과를 얻었다. 성도의 면적함수는 음성 생성에 직접적으로 연계되어 있으므로 실제 음성 생성 메카니즘의 구조를 보여 줄 수 있고 직접적인 물리적인 의미를 갖고 있기 때문에 음성 인식을 위한 효과적인 매개변수로 이용될 수 있다.

본 연구에서는 음성 인식을 위한 특징 벡터로서 성도 면적을 이용하고, 이를 벡터 양자화(Vector Quantizatio, V.Q)에 적용하여 한국어 5개 단모음 및 10개의 숫자음에 대한 인식 실험을 수행하여, 기존의 널리 이용되는 특징 벡터인 LPC 캡스트럼 계수에 의한 인식결과와 비교하여 인식을 위한 특징 벡터로써 성도 면적 함수의 이용 가능성을 검증해 보고자 한다.

2. 성도 면적 함수 추정

성도 면적 함수 추정을 위해 BURG의 알고리즘을 이용하여 우선 반사계수를 구한다. 반사 계수는 성도를 무손실 음향튜브로 모델링하였을 때의 단면적을 특징지을 수 있기 때문이다[6]. 반사 계수에서 직접 성도의 단면적을 추정하는데 있어서 경계조건으로는 Atal의 경계 조건과 Wakita의 경계 조건 2가지가 있으나, Wakita의 경계 조건이 X-선 데이타와 더 일치한다고 알려져 있으므로, 본 연구에서는 Wakita의 경계조건을 선택하였다. 경계 조건에서 입술에서는 임피이던스가 0이라고 보고, 성문에서는 저항성 임피이던스를 갖는다고 가정한다. 그림 1은 성도 면적 추정의 블럭선도를 나타낸 것이다.

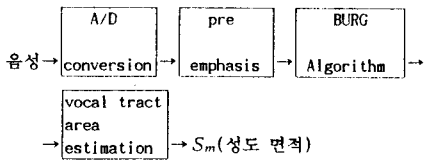


그림 1. 성도 면적 추정을 위한 블럭 선도

그림 1에서와 같이 입력 신호인 A/D 변환된 음성 데이타에 대해 프리 엠퍼시스를 거쳐 Burg의 방법을 이용하여 반사 계수를 구한다. 이는 실제의 음향학적 튜브에서 구한 반사 계수와 일치함이 알려져 있다. 추출된 반사 계수를 이용하여 음성 신호로부터, 해당 구간에서의 면적 S_m를 구할 수 있다.

$$S_m = S_{m+1}(1+k_m)/(1-k_m) \qquad (1)$$

식(1)에서, 구간 m에서의 성도 면적은 S_{m+1}과의 상대 비율과 반사 계수 k_m에 의해 상대적으로 정해지므로 성문에서 입술까지의 면적이 역으로 추정될 수 있다.

3. 벡터 양자화

벡터 양자화란 연속혹은 떨어진 벡터들을 코드북(code book)으로 사상(mapping)시켜서 적절한 디지탈 순열(digital sequence)로 부호화 해주는 방법을 말한다. 벡터 양자화의 원래 주 목적은 데이타의 충실도를 잃지 않으면서 데이타를 압축하는 것이었다. Shannon에 의하면 스칼라대신 벡터로 조합된 신호를 부호화하는 것이 적은 비트율로 좋은 성능을 얻을 수 있다고 알려져 있다. 벡터 양자화에서는 음성파형과 같은 어떠한 형태의 패턴이라도 그 파형으로부터 샘플들의 벡터를 구성함으로써 그 패턴들을 표현 할 수 있다. 이러한 면을 고려할 때 벡터 양자화는 단순히 스칼라 양자화의 일반화뿐만 아니라 응용범위와 의미하는 바가 크다고 할 수 있다.

K차원,레벨 N의 벡터 양자화기는 각 입력 벡터 $X=\{x_0,x_1,\cdots,x_{k-1}\}$에 재합성 알파벳 $A=\{y_i; i=1\cdots,N\}$을 할당하는 사상이라고 할 수 있다.

벡터 양자화기는 source 벡터 X_n에 가장 가까운 재합성 벡터 $\hat{X}_n$을 찾아낼 수 있도록 X_n을 encoding한다. 잘 되었는지를 알기 위해서는 왜곡 측정치(distortion measure) $d(X,\hat{X})$을 사용한다. $d(X,\hat{X})$는 입력 벡터 X와 재합성벡터 $\hat{X}$의 왜곡치(distortion) 기대값 $E\{d(X,\hat{X})\}$로 표시된다. 이것은 벡터 양자화의 시스템 성능을 나타내는 인자로서 값이 작을수록 좋은 양자화기임을 의미한다. 벡터 양자화에서 평균 왜곡 측정치 $E\{d(X,\hat{X})\}$이 최소화 되는 경우 이를 최적 벡터 양자화기(optimal vector quantizer)라고 한다.

벡터 양자화기의 성능은 사용되는 코드북에 의해 결정되므로, 양자화 오차를 최소로 하는 코드북 설계를 위한 많은 연구가 행해져왔다.

1980년 Linde등은 주어진 통계적 모델이나 많은 훈련 데이타를 이용하여 코드북을 설계하는 효과적이고 직관적인 방법을 제시하였다. 이를 LBG 알고리즘이라 하며, Lloyd에 의한 1차원 스칼라 최적 양자화 알고리즘을 K차원 벡터로 확장시킨 알고리즘이다[5].

알고리즘 1. LBG 알고리즘

step 1.
Initialization : Given N=number of levels, distortion threshold $\varepsilon \geq 0$, an initial N-level reproduction alphabet A_0 and a training sequence $\{x_j; j=0, \cdots,n-1\}$. set m=0 and $MSE_{-1}=\infty$

step 2.
Given $A_m=\{y_i; i=0,1,\cdots,N\}$, find the minimum distortion partition $P(A_m)=\{S_i; i=0,\cdots,N\}$ of the training sequence : $x_j \in S_i$ if $d(x_j,y_i) \leq d(x_j,y_l)$ for *all* l. Compute MSE

$$MSE_m = MSE(A_m, P(A_m)) = n^{-1}\sum_{j=0}^{n-1} \min_{y_i \in A_m} d^2(x_j,y)$$

step 3.
If (MSE-MSE)/MSE $\leq \varepsilon$, halt with A_m final reproduction alphabet. Otherwise continue.

step 4.
Find the optimal reproduction alphabet

$$\hat{x}(P(A_m)) = \{\hat{x}(S_i); i=0,\cdots,N\} \quad \text{for } P(A_m)$$

Set $A_{m+1}=\hat{x}(P(A_m))$. Replace m by m+1 and go to step 2.

벡터 양자화기를 이용하여 음성 인식을 할때는 우선 훈련 음성 데이타를 선형예측 부호화 방법등으로 특징을 추출하고 추출된 특징을 이용하여 각 발음에 대하여 벡터 양자화를 수행한다. 코드 워드를 C_i로 표시하고 코드북을 C라 표시하면 $C=\{c_1,c_2,\cdots,c_N\}$가 된다.

S_j를 부호화한 음성의 j번째 프레임에서 추출한 음성 특징 벡터라고하면, S_j는 nearest neighbor 법칙에 의하여 코드북내의 한 코드워드 C_b로 부호화 된다.

$$d(S_j,C_b) = \min\ d(S_j,C_i) \qquad (2)$$

벡터 양자화 코드북은 연속된 여러 개의 음성 프레임들이 부호화되면서 따르게 되는 왜곡들의 평균값이 최소가 되도록 설계한다.

$$D = 1/L\sum_{j=1}^{L} \min\ d(T_j,C_i) \qquad (3)$$

여기서, L은 프레임의 수이고, T_j는 코드북 설계에 사용되는 j번째 프레임에 사용되는 음성 특징 매개변수를 나타낸다.

완성된 코드북으로 음성을 인식하는 원리는 미지의 음성과 각각의 코드북 사이의 거리를 구한 뒤 가장 최소의 거리를 가지는 코드북의 음성으로 인식한다.

음성 패턴의 특징 벡터사이의 거리 계산을 위하여 많이 사용되는 척도로는 유클리드 거리 척도, 이타쿠라-사이토 거리 척도, 대수 스펙트럼 거리 척도 등이 있으며 본 연구에서는 대수 스펙트럼 거리 척도를 이용하였다.

벡터 양자화기를 이용한 음성 인식 시스템의 개략도는 아래 그림과 같다.

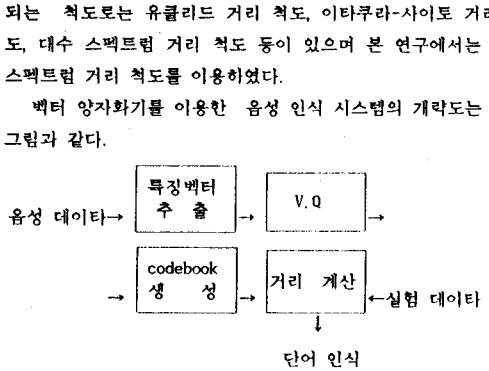


그림 2. 벡터 양자화를 이용한 음성 인식의 블럭 선도

4. 실험및 결과고찰

그림 4.1은 본 연구에서 구성한 전체 실험 과정에 대한 블럭선도이다.

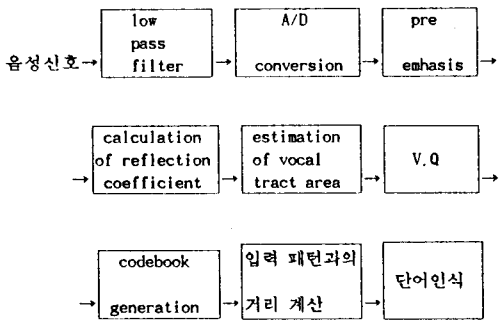
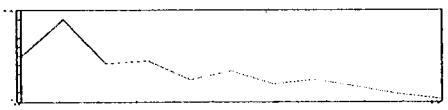


그림 3. 전체 실험 과정에 대한 블럭 선도

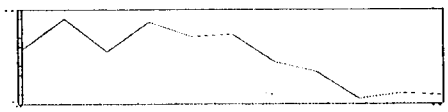
전처리단에서는 입력신호에 대해 pre emphasis $(1-\mu z^{-1})$를 거치게 되는데 음성 신호 분석시, 특히 성도에 대한 특성을 이용할 경우 분석 전에 꼭 취해져야 한다. 본 연구에서는 μ를 0.95로 고정시켰다. 다음으로 Burg알고리즘에서 반사 계수 $(-1<k_m<1)$를 구하고, 그 반사 계수를 이용하여 성도 면적을 구하고 그 성도 면적을 벡터 양자화를 하여 코드북을 생성하고 생성된 코드북과 미지의 실험 데이타와의 거리를 측정하여 가장 가까운 코드북의 발음을 미지의 데이타의 발음으로 인식하게 한다. 본 논문에서는 실험 데이타로 5개의 모음 와 10개의 숫자음에 대하여 성도 면적과 LPC 켑스트럼계수를 구하여 벡터 양자화를 하고 인식율을 구하여 인식율을 비교하였다.

① 모음 인식 실험

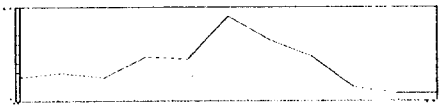
모음 인식 실험에서는 4명의 화자가 3번씩 발음한 /아/, /에/, /이/, /오/, /우/ 것을 대상으로 하여 실험을 하였다. 3번씩 발음한 데이타중에서 첫번째 발음한 것을 훈련 데이타로 하여서 코드북을 작성하였고 나머지 2개의 발음으로 인식 실험을 하였다. 아래 그림 4와 그림 5에 /아/, /에/, /이/, /오/, /우/에 대한 성도 면적과 LPC 켑스트럼을 이용한 특징 패턴 추출 결과를 차례대로 나타내었다.



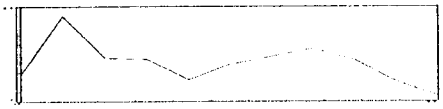
(가) /아/의 성도 면적



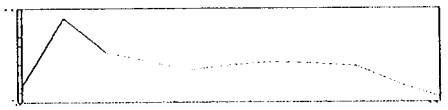
(나) /에/의 성도 면적



(다) /이/의 성도 면적

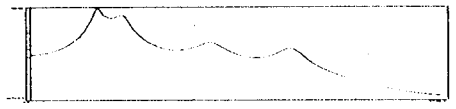


(라) /오/의 성도 면적

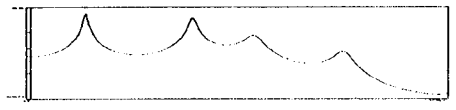


(마) /우/의 성도 면적

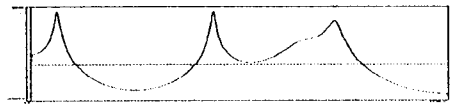
그림 4. 성도면적을 이용한 5개의 모음에 대한 특징 패턴



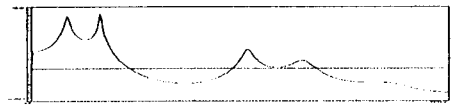
(가) /아/의 스펙트럼



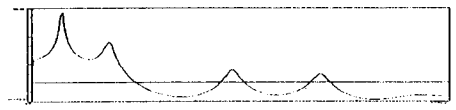
(나) /에/의 스펙트럼



(다) /이/의 스펙트럼



(라) /오/의 스펙트럼



(마) /우/의 스펙트럼

그림 5. LPC 켑스트럼을 이용한 5개의 모음에 대한 특징 패턴

그림 4와 그림 5를 볼 때 LPC 켑스트럼을 이용한 특징 패턴보다는 성도 면적을 이용한 특징 패턴이 /오/, /우/의 구별에 있어서 확실함을 알 수 있다. 벡터 양자화를 수행할 때 코드워드 갯수로는 4개를 선택 하였다. 16개, 32개등의 코드워드를 가진 코드북도 만들 수 있으나 메모리도 많이 차지하고 인식 속도도 느려지므로 코드워드 갯수 4개로 벡터 양자화를 하였다. 성도 면적을 이용하여 음성 인식 실험을 할 때는 각 프레임에서 성도 면적을 가장 큰 값에 대하여 정규화(normalization)를 하

여야만 제대로 인식을 할 수 있기 때문에 정규화를 꼭 하여야 한다. 성도 면적을 특징 벡터로한 코드북과 LPC켑스트럼을 특징 벡터로 한 코드북을 가지고 화자 종속 실험을 한 결과는 다음과 같다.

표 1. 5개의 모음에 대한 인식 실험

모음	모음에 대한 오인식 갯수							
	성도 면적				LPC 켑스트럼 계수			
	KDJ	SJS	RKK	JSW	KDJ	SJS	RKK	JSW
/아/	0	0	0	0	1	0	0	0
/에/	0	0	0	0	0	0	0	0
/이/	0	0	0	0	0	0	0	0
/오/	0	0	0	0	0	0	0	0
/우/	0	0	0	0	0	0	0	0
인식율 (%)	100				97.5			

② 숫자음 인식 실험

숫자음 인식 실험은 4명의 화자가 /영/, /일/, /이/, /삼/, /사/, /오/, /육/, /칠/, /팔/, /구/의 10개의 발음을 5번씩 한것을 이용하였다. 첫번째 발음을 훈련 데이타로 이용하였고 나머지 4번씩의 발음을 실험 데이타로 이용하였다.

표 2. 10개의 숫자음에 대한 인식 실험

음소	음소에 대한 오인식 갯수							
	성도 면적				LPC 켑스트럼 계수			
	KDJ	SJS	RKK	JSW	KDJ	SJS	RKK	JSW
/영/	0	0	0	0	0	0	0	0
/일/	1	0	0	0	0	0	0	0
/이/	0	0	0	0	0	0	0	0
/삼/	0	0	0	0	0	0	2	0
/사/	0	0	0	0	1	0	0	0
/오/	0	0	0	0	0	0	0	0
/육/	0	0	0	0	0	0	0	0
/칠/	0	0	0	0	0	0	0	0
/팔/	0	0	0	0	0	0	0	0
/구/	0	0	0	0	0	0	0	0
인식율 (%)	99.3				98.1			

성도 면적을 이용한 인식 실험도 LPC켑스트럼 계수를 이용한 인식 실험과 같이 입력패턴과의 거리를 계산 할 때 유클리드 거리를 사용하는 것보다는 대수 거리를 이용하는 것이 인식율을 높일 수 있었다.

5. 결론

본 논문에서는 음성 인식을 할 때 특징 추출 벡터로 조음적 특징을 반영한 성도 면적을 이용하여 음성 인식을 하였다. 그 결과를 전통적으로 음성 인식에 쓰이는 음성 특징으로 가장 많이 쓰이는 LPC 켑스트럼 계수와 비교하였다. 그 결과로는 첫째, 5개의 모음에 대하여 성도 면적을 이용 할 때는 100%로서 LPC 켑스트럼 계수를 이용한 인식율 97.5%와 비교할 때 더 나은 인식율을 얻었다. 둘째, 10개의 숫자음에 대한 인식 실험에서도 성도 면적과 LPC 켑스트럼 계수을 이용한 인식율은 각각 99.3%와 98.1%로서 성도 면적이 더 높은 인식율을 얻었다. 셋째, /오/, /우/등의 특정 발음에서는 LPC 켑스트럼 계수보다 더 확실하게 패턴을 구분할 수 있었다. 이러한 결과로 볼 때 성도 면적을 특징 벡터로 사용한다면 음성 인식 실험은 더 좋은 성능을 가질 수 있을 것으로 기대된다.

참고문헌

[1] Panos. E. Papamichalis, "Practical Approaches To Speech Coding.", Prentice Hall, 1987.

[2] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustical measurement." JASA, vol. 41, pp 1002 - 1010. 1967.

[3] P. Mermelstein, "Determination of the vocal tract shape from measured Formant frequencies," JASA vol. 41, pp 1283 - 1294, 1967.

[4] A. Paige, V. W. Zue, "Computation of vocal tract area function," IEEE Trans. Audio Electroacoustics, Vol. AU-18, pp 7 - 18,1970.

[5] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Processing, Vol. AU-21, NO.5, oct. 1973

[6] S. Furui, "Digital Speech Processing, Synthesis, and Recognition," Dekker, 1992

[7] L. R. Rabiner and R. W. Schafer "Digital Processing of Speech signals," Prentice Hall,1978.

[8] J. D. Markel and A. H. Gray, Jr., "Linear Prediction of Speech," Springer Verlag, 1980.

[9] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. commun., vol. com-28, pp.84-95, Jan.1980