한국어 단독음 인식을 위한 표준패턴 설정에 관한 연구

A Study on Creating Reference Pattern for Recognition of Korean Isolated Word

*김 계 국(Kim, K. K.)
**고 덕 영(Ko, D. Y.)
***이 종 악(Lee, J. A.)

요 약

본 연구에서는 집단화 알고리즘을 이용하여 한국어 단독음의 표준 패턴을 설정하였다.
Minimax 기법을 이용하여 각 단독음에 대하여 최고 3개까지 표준패턴을 설정하여 인식하였다.
특징 파라미터는 선형예측계수와 자기 상관 계수를 이용하였으며 패턴들 간의 유사도 비교는 Itakura가 제안한 거리측정법을 이용하였다. 표준패턴을 1개만 설정하였을때 55.9%, 2개를 설정했을 때 76.9%, 3개를 설정했을 경우는 89.5%의 인식률을 얻었다.

ABSTRACT

This paper discusses a reference pattern creation for a speaker-independent Korean isolated word by using the clustering. In this paper we permitted to top 3 clusters and created reference pattern by Minimax Criterion. The features parameter used the LPC Coefficients and Autocorrelation and simple Itakura distance measure was used to measure similarity between patterns. With word reference patterns obtained as described above the recognition rate was within one choice only 55.9%, two choice only 76.9%, three choice only 89.5%.

*전국대학교대학원 전자공학과 박사과정
**전주공업전문대학 전자과 조교수
***전국대학교 전자공학과 교수

Ⅰ. 서 론

본 연구에서는 불특정화자의 음성을 인식하기 위하여 보다 효율적인 표준패턴 설정 방안을 제시하는데 그 목적이 있다.

사람은 저마다 목소리가 다르기 때문에 각 화자의 음성의 특성 변화에 따라 인식 결과가 크게 변하게 된다. 이러한 특성 변화 때문에 생기는 오인식을 줄이고 보다 효율적인 인식을 위해 집단화 알고리즘을 도입하여 표준패턴을 설정하였다.

그림 1은 음성 인식을 실행하는데 있어서의 기본 과정을 나타내고 있다. 우선 특징 파라미터를 추출하고 설정된 표준 패턴과 시험패턴들 간의 유사도를 비교하여 인식하는 과정을 나타내고 있다.

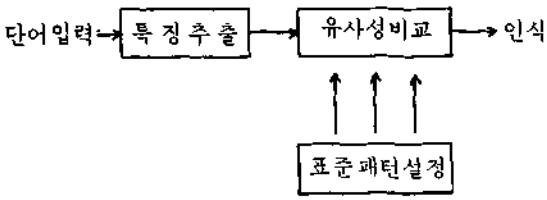

그림 1 단어인식 블럭도

특징 파라미터 선택 문제는 많은 연구진들에 의하여 연구된 바 있으며 시간 영역에서는 대수 에너지, 영교차율, 주파수 영역에서는 스펙트럼 계수, 셉스트럼계수, 선형예측계수 등 여러가지 특징 파라미터들이 이용되고 있다.

본 연구에서는 선형 예측계수와 자기 상관계수를 특징 파라미터로 선택하였다. 유사도를 비교하기 위하여 Itakura가 제안한 LPC거리 측정법을 도입하였다.

Ⅱ. 본 론

1. 표준패턴설정 알고리즘

본 연구에서는 한국어 단독음 13개를 대상으로하여 남성화자 2인중 1명은 5번 또 다른 1명은 4번 반복 발음한 117음성과 여성화자 1인에 의하여 2번씩 발음된 26음성을 집단화 대상으로 하였다.

N개의 음성 데이타(이하 토큰이라한다)로 구성된 집합을 식(1)과 같이 표현할 수 있다.

$$\Omega = \{x_1, x_2, \cdots\cdots x_N\} \quad (1)$$

x_i는 반복 음성을 표현한 음성 토큰을 의미한다.

각 음성 토큰은 본래의 길이를 갖고 있으며 x_i는 n_i개의 프레임 길이를 갖는다. 이러한 음성 토큰들 사이의 거리 즉 x_i와 x_j 사이의 거리 계산식은

$$d_{ij} = \delta(x_i, x_j) = \frac{1}{n_i} \sum_{K=1}^{n_i} d(K, W(K), i, j) \quad (2)$$

와 같이 표현할 수 있으며 $d(K, W(K), i, j,)$는 x_i의 K번째 프레임과 x_j의 W(K)번째 프레임간의 거리를 나타내고 있다.

본 논문에서는 선형 예측 계수와 자기 상관 계수를 특징 파라미터로 선정하였으므로 LPC 대수확률 거리측정법을 도입하였다. 그러므로

$$d(K, W(K), i, j) = \log\left[\frac{(a'_{W(K)})' \; R_K^i \; (a'_{W(K)})}{(a_K^i)' \; R_K^i \; (a_K^i)}\right] \quad (3)$$

여기서 a_K^i는 토큰 i의 K번째 프레임의 LPC계수 벡터, R_K^i는 토큰 i의 K번째 프레임의 자동상관 계수 행렬, $a'_{W(K)}$는 표준 토큰 j의 W(K)번째 프레임의 LPC계수 벡터, $'$은 전치벡터(Vector transpose)를 뜻한다.

함수 $W_{(K)}$는 j토큰과 i토큰을 DTW (Dynamic Time Warp) 정합시켜 얻은 워핑함수로서 d_{ij}를 최소로 한다. 식(1)의 모든 음성토큰들이 M개의 집단으로 이루어 진다고 한다면 일반적으로 식(4)와 같

이 나타낼 수 있다.

$$\Omega = \bigcup_{i=1}^{M} W_i \tag{4}$$

여기서 M은 집단의 총수를 뜻하며 집단W_i의 중심토큰을 $x_p^{(i)}$로 정의하고 이 $x_p^{(i)}$는 i 번째 집단 W_i 내의 토큰이어야 한다.

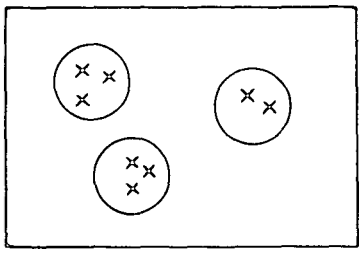

그림 2 집단화의 예.

본 연구에 제안된 알고리즘은 집단을 분류하기 위해 집단의 중심토큰을 임의로 선정하여 토큰들 간의 거리를 계산하고 수렴여부를 결정하는 과정이반복된다. 우선, 집단의 수를 결정한다. 이때 집단의 수를 M이라고 할 때 집단의 중심토큰도 임의로 M개를 선택한다. 이를 식으로 간단히 표현하면

$$x_p^{(i)} = x_i \qquad 1 \leq i \leq M \tag{5}$$

와 같다.

집단의 분류는

$$x_j \in W_i \text{ iff } \delta(x_j, x_p^{(i)}) \leq \delta(x_j, x_p^{(K)})$$

$$1 \leq K \leq M, \quad 1 \leq j \leq N \tag{6}$$

에 준한다.

집단의 중심 토큰은 minimax center를 고려하였으므로

$$\max\{\delta(x_j^{(i)}, x_K^{(i)})\} \text{ is minimum} \tag{7}$$

와 같은 기준에 따라 설정한다.

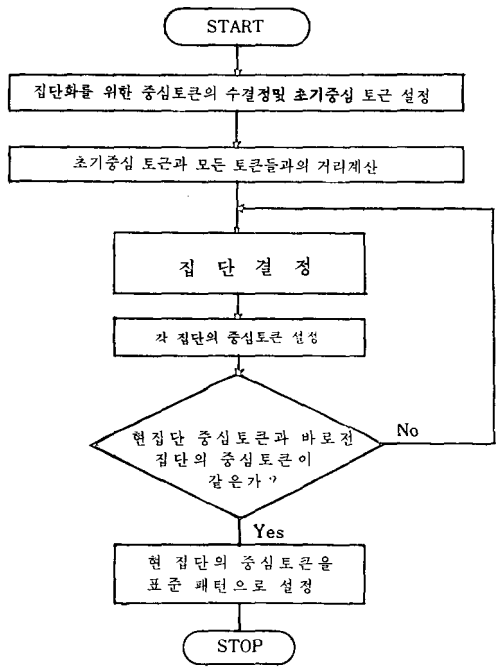

그림 3 본 연구에 제안된 집단화 블럭도

2. 실험결과

본 연구에서는 13개의 단어를 남성 화자 2인과 여성 화자 1인에 의하여 11번 반복 발음한 전철역명 143단어를 대상으로 하여 표준 패턴을 설정하였다. 3인 화자가 발음한 동일음 11단어를 집단화하여 설정한 표준패턴은 최고 3개까지 설정하여 인식 실험을 하였다.

각 단어에 대하여 표준 패턴을 1개씩 사용했을 경우 인식률은 55.9%, 2개씩 설정했을 경우 인식률은 76.9%, 그리고 3개씩 사용했을 경우 89.5%의 인식률을 얻었다. 표준 패턴을 각 단어마다 하

표 1 설정된 표준패턴.

전철역명 / 설정기준	이대	이촌	이수	옥수	수유	신촌	신사	신당	사당	대림	신림	신대방	대방
minimax cluster 3	W_1	W_2	W_2	W_2	W_1	M_{13}	M_{11}	M_{21}	W_1	W_1	W_1	W_1	W_1
	M_{24}	M_{14}	M_{13}	M_{15}	M_{11}	M_{24}	W_1	M_{23}	M_{14}	M_{14}	M_{13}	M_{14}	M_{11}
	M_{21}	M_{21}	M_{23}	M_{24}	M_{14}	M_{21}	M_{14}	M_{13}	M_{13}	M_{23}	M_{23}	M_{13}	M_{13}
minimax cluster 2	W_1	W_1	M_{13}	M_{11}	W_1	M_{13}	M_{14}	M_{14}	W_1	W_1	W_1	W_1	W_1
	M_{21}	M_{21}	M_{23}	M_{24}	M_{14}	M_{22}	W_1	W_1	M_{14}	M_{23}	M_{13}	M_{14}	M_{11}
minimax cluster 1	W_1	W_1	W_1	W_1	W_1	W_1	W_1	W_1	W_1	W_1	W_1	W_1	W_1

기호표시 { M : 남성, W : 여성
M_{ij}은 남성화자 첫번째 사람이 i번째 발음한 단어.
M_{2j}은 남성화자 두번째 사람이 j번째 발음한 단어를 말함.

표 2 1개의 표준패턴 설정시 인식결과.

입력 / 출력	이대	이촌	이수	옥수	수유	신촌	신사	신당	사당	대림	신림	신대방	대방
이 대	4	1		2			1						
이 촌		6											
이 수	2		5	1				1					
옥 수		1		8	1	2				1			2
수 유					6			1					
신 촌	2	2	2		2	8	1	1		1	2	2	
신 사							5	1	1				
신 당								4				1	
사 당					2			1	10	2	2		
대 림	2		3							6	2		
신 림							2				4		
신 대 방								1				5	
대 방	1	1	1			1	2	1		1	1	3	9

나석 설정했을 경우 표준패턴으로 설정된 단어는모두 여성화자가 첫번째 발음한 단어였는데 실제로 여성화자와 남성화자 사이에는 발음할 때 주파수의 차가 크므로 인식률이 크게 떨어짐을 알 수 있다. 여기서 표준패턴을 2개 설정했을 경우 1개를 설정했을 때보다 인식율이 크게 증진되는 것은 표준패턴 설정시 2개중 1개는 남성화자가 발음한 단어가 표준패턴으로 설정되었기 때문이다. 그러나

표 3 2개의 표준패턴설정시 인식결과

입력 / 출력	이대	이촌	이수	옥수	수유	신촌	신사	신당	사당	대림	신림	신대방	대방
이 대	5										2	2	
이 촌		8				1							
이 수	2	2	11	2									
옥 수				9	1								
수 유					8								
신 촌	1	1			2	9	1				1		
신 사							8		1				
신 당	1					1		10					
사 당									10				
대 림	2									7			2
신 림							2			4	8	1	
신 대 방								1				8	
대 방													9

표 4 3개의 표준패턴 설정시 인식결과

입력 / 출력	이대	이촌	이수	옥수	수유	신촌	신사	신당	사당	대림	신림	신대방	대방
이 대	8												
이 촌		11				1							
이 수			11										
옥 수				11									
수 유					9								
신 촌					2	9	2						
신 사							9	1				1	
신 당								10				1	
사 당									10				
대 림	1									11			2
신 림	2					1					11		
신 대 방									1			9	
대 방													9

이 경우에 있어서도 남성 화자간의 발음상의 차이가 있었다. 첫번째 남성화자는 각 음절 사이의 길이(duration)가 짧은 데 반하여 두번째 남성화자는 각 음절 사이의 길이가 길었다.

따라서 화자가 바뀌면 같은 단어라 할지라도 거리 계산값이 커져서 인식이 안되는 것으로 나타났으며 3개를 표준 패턴으로 설정했을 경우는 2개를 설정했을 때보다 인식률이 더욱 증가되었으며표준패턴 2개을 설정했을 때의 단점을 상당히 보완해 줌을 알 수 있었다. 그러나 표준패턴을 3개 설정했을 경우는 컴퓨터의 거리계산 시간이 많이 걸린다는 것도 염두에 두어야 할 것같다.

본 연구에서 설정한 표준패턴은 표1에 나타내었으며 인식 결과를 표2, 표3, 표4에 표시하였다.

Ⅲ. 결 론

불특정 화자의 음성을 인식하기 위해서는 각화자의 성대변화를 모두 수렴할 수 있는 표준패턴을 설정해야 한다.

최고 3개의 표준패턴을 설정하는 과정에서 3명의 화자가 발음한 음성이 하나씩 3개의 토큰이 표준패턴으로 설정된 것도 있었지만 3개 모두가 남성화자의 음성이 표준음으로 설정된 경우도 있었다. 또, 2개의 표준패턴을 설정하는 과정에서는 남녀 각각 1인이 발음한 토큰이 대부분 표준음으로 설정되었는데 실제로 인식하는 과정에서 문제가 되었다. 어느 화자든 자기자신이 발음한 단어가 표준음으로 설정이 된 경우는 자신이 발음한 대부분의 단어가 인식되었으나 똑같은 단어라 할지라도 화자가 달라지면 이들 단어들간의 거리값이 커지기 때문에 오인식되는 경우가 많았다.

이는 각 화자의 성대특성변화와 여성화자의 불규칙한 음폭 때문에 발생하는 현상으로 생각된다.

앞으로의 연구과제는 보다 효율적인 또 다른표준패턴설정 알고리즘개발도 중요하지만 인식결정알고리즘 개발에도 역점을 두어야 할 것으로 생각된다.

參 考 文 獻

1. L.R. Rabiner and J.G. Wilpon "Application of Clustering Techniques to Speaker-Trained isolated Word Recognition" Bell Lab. Vol. 58. No. 10. 12. 1979.
2. Peter U. de Souza. "Statistical Tests and Distance Measures for LPC coefficients" IEEE Vol. ASSP 25, No. 6, 12. 1977.
3. Ifroaki Sakoe, Seihi Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition" IEEE Trans. Vol. ASSP-26, No. 1, 2. 1978.
4. L.R. Rabiner. "On creating Reference templates for speaker independent recognition of isolated word. "IEEE. Vol. ASSP-26. No. 1, 2. 1978.
5. Helmuth Spath. Ellis Horwood Limited. "Cluster Analysis Algorithms for DATA reduction and classification of objects."
6. 김계국 "집단화를 이용한 한국어 숫자음성의 표준패턴 설정에 관한 연구" 한국음향학회 논문집 Vol 5. No 2. 1986. 6.
7. FUMITADA ITAKURA. "Minimum Prediction Residual Principle Applied to Speech Recognition" IEEE Vol. ASSP-23, No. 1, 2. 1976.