이치화 패턴을 이용한 고립단어 음성인식

°류 준 형 이 용 주 박 찬 경 김 용 호 김 경 태

한국전자통신연구소

Isolated word recognition using binary pattern

J. H. Ryoo Y. J. Lee C. K. Park Y. H. Kim K. T. Kim

E. T. R. I

ABSTRACT

This paper describes the isolated word recognition using binary patterns denoting the presence or absence of a local peak at a particular channel. In closed test, 81.3% and 76.0% of correct recognition rate were achieved in case of 10 males and 10 females with each 1500 test samples.

1. 서 론

인간이 음성을 인지할때는 각 주파수의 진폭자체 보다는 스펙트럴의 peak에 의해서 감지된다는 사실이 생리학적 실험의 결과로 나와 있다 [7] [8] [9]. 따라서 맨-머신 인터페이스에 대해서 이러한 생리학적 실험 근거를 토대로 한 음성인식의 한 방법이 사용되어 왔으며 그 유용성 또한 확인된바 있다 [1] [2] [3].

일반적으로 음운성을 가진 정보인 포먼트 주파수를 사용하는 것이 바람직하나, 포먼트 주파수 추출 자체가 용이하지 않으므로 포먼트를 포함하는 다수의 로컬 피크(local peak)들을 특징 파라미터로 사용하고자 하는 것이다. 이 로컬 피크는 S/N 비가 극히 나쁘지 않는 한 잡음의 영향을 받지 않는다는 장점도 가지고 있다.

본 연구에서는 이와같은 스펙트럴 로컬 피크의 유무에 따른 이치화 (binary) 패턴을 한국어 고립단어 인식에 적용하여 보았다. 구체적으로는 성대음원 및 조음기관의 개인차를 흡수하도록 하였으며 다이내믹 프로그래밍을 사용한 패턴매칭에 대해서도 검토하였다.

2. 실험방법

본 실험에서 사용한 인식방법은 그림 1과 같다.

(1) 주파수 분석

3.4 KHz로 저역 필터링을 한후 8 KHz로 샘플링하여 이미 데이터베이스로 구성된 음성 데이터 (숫자음 및 간단한 명령)를 Q=6, 22 채널 BPF 하여 10 msec (1 프레임) 마다 채널별로 대수에너지를 얻는다. 이때 BPF의 중심 주파수는 300 Hz~3400 Hz 에 걸쳐 1/6 옥타브 등간격으로 구성된다.

(2) 음원의 정규화

성대 음원특성의 개인차에 영향받는 음성 스펙트럴의 기울기를 정규화 하기 위해서 음성 스펙트럴 Xi와 이를 일차직선으로 근사한 Yi = A * i + B 간에 다음식의 에러

$$E = \sum_{i=1}^{22} \{ Xi - (A * i + B) \}$$

를 최소화 하는 최소 자승근사직선 Yi 를 원 스펙트럴에서 제거함으로써

$$Zi = Xi - (A * i + B)$$

하는 잔차 스펙트럴을 특징량으로 취한다.

(3) 이치화 패턴 작성

음성 스펙트럴의 로컬 피크는 주요한 음성 정보를 제공한다는 점을 감안하여 아음과 같이 이치화 패턴을 작성한다.

잔차 스펙트럴상의 모든 피크중에서, Zi가 0 dB 를 초과하는 것만을 로컬 피크로 취한다. 각 채널에 대한 로컬

피크 발생여부를 '1' 또는 '0'으로 그림2와 같이 이치화 한다.

(4) 사전 작성

기존 패턴은 일반적으로 여러 사람의 음성특징을 흡수하는 것이 바람직하다. 이를위해 기준 패턴 작성은 기본적으로 각 개인의 이치화 패턴을 중첩한다. 이때 중첩 대상 패턴들 중 가장 긴 패턴에 맞추어 나머지 패턴을 선형으로 늘임으로써 발성시간차이를 정규화 한다. 이치화 패턴을 중첩함으로써 여러 사람의 음성 특징이 흡수될 뿐 아니라 각 채널에 대해 가중치가 달라지게 된다.

(5) 입력 패턴 작성

입력 패턴과 사전을 매칭할 때 주파수 방향의 개인차를 흡수하기 위해 입력 패턴 작성시에는 잔차 스펙트럴 상에서 로컬 피크의 진폭의 0.5배 이상되는 진폭을 지닌 채널은 모두 '1'로 한다.

(6) 거리 계산

이미 작성된 사전의 각 단어와 입력의 이치화 패턴간의 거리를 계산한다. 사전과 입력의 발성길이 차이는 사전 작성시와 마찬가지로 긴쪽에 맞추어 선형으로 늘여 매칭을 한다. 2개 패턴의 거리는 다음의 절대치 거리를 취한다.

$$D = \sum_{i=1}^{FRM} \sum_{ch=1}^{22} |R(i,ch) - I(i,ch) * N|$$

여기서 FRM은 입력과 사전중 긴쪽의 프레임 길이를 나타내며, N은 사전 작성시 중첩한 회수를 나타낸다.

매 입력에 대해 사전의 단어와 매칭을 하여 거리가 가장 짧은 단어를 인식결과로 취한다.

(7) DP를 이용한 거리계산

로컬 피크의 정보만을 사용하는 이치화 패턴에서 DP를 사용한 매칭의 유효성의 정도를 타진하도록 하였다. 2개 패턴의 각 프레임 사이의 거리를 선형 매칭의 경우와 동일하게

$$d_{ij} = \sum_{ch=1}^{22} |R(j,ch) - I(i,ch)|$$

로 둔다. DP를 이용하는 경우에는 기준 패턴으로서 중첩하지 않은것을 사용하므로 입력 패턴에 가중치는 부과하지 않는다

3. 실험 및 결과

(1) 음성 데이터

남성 10명, 여성 10명이 25단어에 대해 6회씩 발성한 데이터를 사용하였다. 단어는 /이/, /육/, /공/, /일/, /사/, /오/, /육/, /삼/, /영/, /칠/, /팔/, /구/, /예/, /넷/, /둘/, /셋/, /여덟/, /다섯/, /아홉/, /여섯/, /다시/, /하나/, /뒤로/, /일곱/, /아니오/ 이며 숫자음과 약간의 명령어로 구성된 음성 데이터 베이스[6]를 이용한다.

(2) 특정화자의 경우

남여 각 10 명에 대해 각각 자신의 1~3회 발성으로 기준 패턴을 만들고, 4~6회 발성을 입력으로 하였다.

	1	2	3	4	5	6	7	8	9	10	평균
남	97.3	94.7	92.0	90.7	94.7	93.3	94.7	96.0	85.3	85.3	92.4
여	93.3	90.1	89.3	96.0	54.7	89.3	82.7	58.7	93.3	85.3	83.3

(3) closed test 경우

남여 각각 10명의 1~6회 발성으로 기준 패턴을 만들고 기준 패턴을 만든 데이터를 다시 입력한 결과 남자의 경우 81.3%, 여자의 경우 76.0%의 인식율이 얻어졌다.

(4) open test 경우

남여 각각 5명의 1~6회 발성으로 기준 패턴을 만들고 기준 패턴 작성에 참여하지 않은 남녀 각각 5명의 1~6회 발성을 입력으로 한 결과 남자의 경우 67.2%, 여자의 경우 59.7%로 closed 실험과는 많이 떨어지는 인식률이 얻어졌다. 이는 불특정화자 인식을 위한 적극적인 개인성 제거 방법을 이용하지 않은 탓이라 생각된다.

(5) DP를 이용한 경우

DP의 구체적인 방법은 참고문헌[5]를 이용하였다. 음성의 동적 변화에 대한 제한으로서 경로는 P=1 즉 발성길이가 2배

혹은 1/2배 되는 경우는 제거 되도록 하였다.

두 패턴간의 전체길이의 차에 대한 제한 조건으로 윈도우(Window)를 주는데, 본 실험에 사용되는 음성 시료는 같은 단어 일지라도 화자 또 발성 회수별로 3배 가까이 차이나는 경우도 존재한다는 점을 감안하여 L=20 으로 취했다.

가. 특정화자의 경우

남여 각각 1회 발성을 기준 패턴으로 취하고 2~6회 발성을 입력으로 하였다.

	1	2	3	4	5	6	7	8	9	10	평균
남	92.8	88.0	88.8	87.2	88.0	95.2	88.8	90.4	87.2	69.6	87.6
여	86.4	87.2	94.4	96.0	52.0	89.6	78.4	47.2	91.2	79.2	80.2

나. 불특정 화자의 경우

남여 각 1인의 1회 발성을 기준 패턴으로 취하고 기준 패턴 작성에 참여하지 않은 남녀 각 9인의 1~3회 발성을 입력으로 하였다. 이 경우 남성은 55.6%, 여성은 45.0%의 인식률을 나타냈다.

4. 결 론

본 논문에서는 BPF 군을 통해 얻어지는 에너지 스펙트럴의 정보를, 이치화 패턴의 유무를 나타내는 '1' 또는 '0'의 이치화 패턴으로 압축하고 이를 이용하여 한국어 고립단어 음성 인식을 시도했다.

DP 이용한 인식률이 선형 매칭을 사용한 경우 보다 저조하다는 사실은, 중첩된 기준 패턴을 사용하지 못하였다는 점을 감안하더라도 이치화 패턴간의 거리로서 절대치 거리를 사용하는 것은 부적절하다는 점을 나타낸다고 볼 수 있다.

선형 매칭의 경우는 DP에 비해 정확한 음성검출이 요구되므로 음성검출방법을 개선해야 한다는 점과 음성인식시 유성음과 무성음 구간을 구분처리하는 방법, 매칭하는 단어간 거리계산에 음소의 지속시간등의 정보를 첨가하는 방법 등등을 검토하여 인식률의 개선을 이룩하고자 한다.

참고문헌

[1] T. Matsuoka, K. Kido, "Spectral의local Peak의 검토", 동북대 전기통신 담화회의 기록, 42, No.3, pp. 61-71, 1973 (일본어)

[2] K.T. Kim et al, "Recognition of vowels in Japanese words using spectral local peaks", J. Acoust. Soc. Jpn (E). 5, 4, (1984)

[3] K.T. Kim et al, "Recognition of stop consonants in Japanese words using local spectral peaks", J. Acoust. Soc. Jpn (E), 7,6 (1986)

[4] H. Matsumoto, M. Nakagawa, M. Yoneyama, "Local Peak 하중평균 사서를 이용한 불특정화자 단어음성 인식", 일본전자통신학회 논문지 Vol.J68-A No.1, pp.78-85, Jan. 1985 (일본어)

[5] H. Sakoe and S. Chiba, "Dynamic Programming Alogrithm Optimization for Spoken Word Recognition". IEEE Trans. ASSP., Vol. ASSP-26, pp. 43-49, Feb. 1978

[6] 이용주외, "단어 음성데이타의 수집 및 DB 구성 시스템", 대한전자공학회 추계 종합학술대회 논문집 Vol.9 No.2 (86.12)

[7] R. Galabos, "Neurophysiology of the anditory system", J. Acoust. Soc. Am. 22, pp. 785-791 (1950)

[8] R. Galabos and H. Davis, "The respoonse of single auditory nerve fibers to acoustic stimulation", J. Neuro-physiol. 6. pp. 39-59 (1943)

[9] Y. Katsuki, et al, "Activity of auditory neurons in upper level of brain of cat," J. Neuro-physiol. 22, pp. 343-359 (1959)

본 연구는 과기처 특정연구과제의 일환으로 이루어진 것이다.

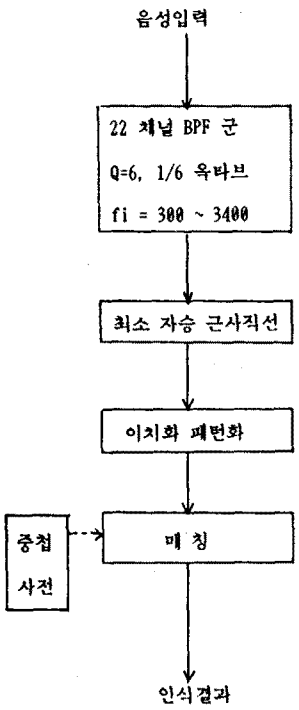
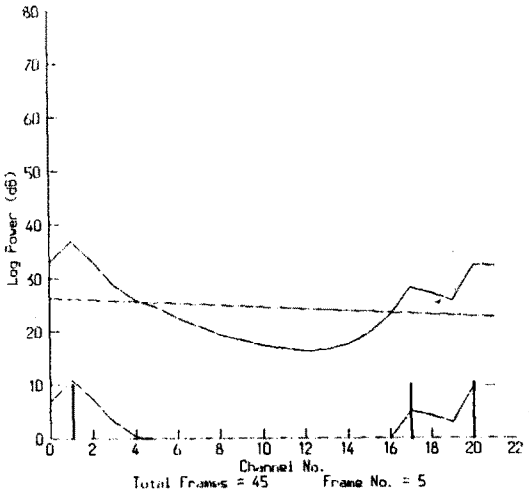
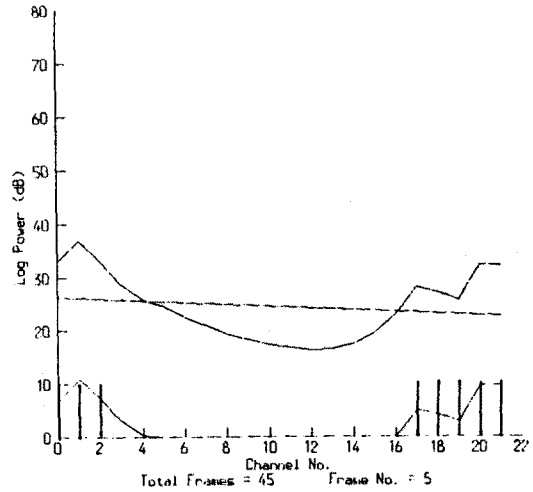


그림1. 음성인식 순서도

Fig. 1 Flowchart of speech recognition



(1) 사전 작성인 경우



(2) 입력인 경우

그림2. 이치화 패턴의 추출

Fig. 2 Extraction of binary patterns