Spatio-temporal방법을 이용한 지역명 인식에 관한 연구

A Study on the recognition of local name using Spatio-Temporal method

지원우°*, 김석동*, 송도선**, 이행세***

* : 호서대학교 전자계산학과

** : 중경 공업 전문대학 전자계산기과

*** : 아주대학교 전자공학과

ABSTRACT

This paper is a study on the word recognition using neural network. A limited vocabulary, speaker independent, isolated word recognition system has been built. This system recognizes isolated word without performing segmentation, phoneme identification, or dynamic time wrapping. It needs a static pattern approach to recognize a spatio-temporal pattern. The preprocessing only includes preceding and tailing silence removal, and word length determination. A LPC analysis is performed on each of 24 equally spaced frames. The PARCOR coefficients plus 3 other features from each frame is extracted. In order to simplify a structure of neural network, we composed binary code form to decrease output nodes.

I. 서 론

사회의 정보화가 급속히 진전되면서 인간과 기계와의 접촉이 빈번해 짐에 따라 인간과 기계사이의 의사 전달 방법으로서의 음성의 중요성은 더욱 증대되고 있다. 오랜 동안 사람들은 음성을 자동적으로 인식하는 기계를 만들려고 시도하고 있으나 아직 완전한 음성 인식기가 나오지 못하고 있다. 그 이유는 음성을 인식하기 위해서는 극복해야 할 몇가지 문제가 있기 때문이다. 똑같은 음성을 갖고 있는 사람은 없기 때문에 사람에 따라 음성이 변하는 문제가 가장 큰 장애물이다. 발성 당시 외부 잡음이나 간섭으로 말미암아 같은 사람의 음성이라 해도 항상 똑같지 않다. 또한 연음 현상에 따라 하나의 단어를 독립적으로 발음할 때와 그 단어가 다른 단어들 사이에 있을때 변화되는 문제는 더욱 심각하다.

일반적으로 음성 인식기는 세가지 종류로 나눌 수 있다[1]. 첫째로 단어 사이에 묵음이 있는 것을 대상으로 하는 독립 단어 인식이 있다. 둘째로는 연속된 음성을 인식하는 것이다. 이 경우는 단어 사이에 묵음이 없다. 이러한 인식기에는 발음된 음성에서 단어를 찾는 기능까지 추가되어야 한다. 물론 음성에서 단어의 시작과 끝을 찾는 것은 쉽지가 않다. 마지막으로 세번째는 음성을 이해하는 시스템이다. 위의 세가지는 모두 발성자와 무관하거나 발성자를 한정시킬 수도 있다. 보통 발성자와 무관하게 인식하는 시스템의 구성은 발성자를 한정시키는 것보다 매우 어렵다.

1950년대 후반기부터 디지탈 컴퓨터의 발전에 따라 음성 인식에 대한 연구가 시작되었다[2]. 컴퓨터에 의해 음성으로부터 음성 고유의 특징을 추출할 수 가 있었다. 1960년대에 들어와 음성을 음소, 음절 또는 단어와 같은 음성학적인 단위로 자동적으로 분할하는 기술이 개발되었다[3]. 또한 새로운 패턴 인식과 분류에 대한 알고리즘의 탄생으로 음성 인식에 영향을 끼치게 되었다. 1970년대에는 오늘날 음성 인식에 사용하는 많은 중요한 기술이 나타났다[4]. 인공 신경 회로망이 소개된 후 이것을 이용한 많은 연구가 이루어지고 있다.

음성을 받기 위해서는 보통 두가지 도구가 사용된다. 마이크로폰과 음성파를 디지탈 신호로 바꾸는 A/D 변환기가 사용된다. 디지탈 신호 처리(Digital Signal Processing) 모듈의 목적은 음성 인식에 보편적으로 사용하는 스펙트럼적인 표현을 가공하고 향상시키는데 있다. 음성 데이타는 양이 매우 커서 인식알고리즘을 적용하기 위하여 음성 데이타를 일시적으로 저장하는 버퍼가 필요하다. 이를 위하여 전처리된 신호의 저장부분이 사용된다. 음성 데이타를 전처리한 신호와 미리 저장되어 있는 기준 음성과의 유사성을 측정하기 위하여 패턴 정합 알고리즘을 적용한다. 가장 잘 정합되는 음성을 결정한다. 가장 좋은 결정을 하기 위해서는 가능한 여러번의 정합 알고리즘을 적용한다.

지금까지 가장 일반적인 인식 알고리즘으로서 저장되어 있는 표준 패턴과 미지의 음성을 동적인 프로그램 기법을 이용하여 정합시키는 방법으로 시간측상에서 표준패턴의 각부분과 정합시켜 가장 비슷한 것을 찾는 방법인 DTW(Dynamic Time Wrapping)[5]와 음성을 인식하는데 통계적으로 접근하는 방법으로 음성데이타를 관찰열로 사용

하여 여러형태의 관찰열 중에서 가장 유사한 것을 선택하는 알고리즘인 HMM(Hidden Markov Model)[6]이 있다.

음소와 음절을 기본으로하는 인식이 신경망에 의해 행해지고 있다. Kohonen의 자기 구조 특징도(self-organizing feature map)는 음소 인식 모델중의 하나이다. 이것은 음소를 인식하기 위한 학습 방법이 자율 학습으로 비슷한 음소끼리 묶어 분류하였다[7]. 음소 인식에 대한 또다른 성공적인 인공 신경 모델은 시간 지연 신경 회로망(Time-Delay Neural Network)이다[8][9]. TDNN은 여러층을 갖는 역전파(Back Propagation) 모델을 이용하여 각각의 연속적인 층으로 오랜시간에 대하여 음성을 적용하는 방법으로 가장 위층에서 시간에 불변하는 특성을 가진 음소를 인식하는 방법이다. 본 논문의 신경회로망에 사용한 알고리즘으로는 지도학습의 대표적인 방법인 역전파 알고리즘을 이용하였다[10].

II. 음성구간 검출

발음된 음성을 단어별로 분할하는것 즉 음성의 시작점과 마지막점을 검출하는 것을 음성 구간 검출이라 하는데 이것은 음성인식, 합성및 분석등 음성공학의 모든 분야에 걸쳐 매우 중요한 일이다. 음성 구간 검출은 음성에 해당하는 입력 부분만 처리하도록 함으로써 계산량을 줄이고 인식률을 향상시키는데 중요한 역할을 한다. 정확한 유음구간 검출은 S/N비가 높은 환경에서는 가장 낮은 레벨의 음성(에너지가 낮은 마찰음) 에너지 조차도 배경 잡음 에너지를 초과한다. 대부분의 실제적인 음성 시스템은 15~20 dB 정도의 낮은 S/N 비에서 동작된다. 일반적으로 음성의 시작이 에너지가 낮은 마찰음(/ㅅ/, /ㅈ/, /ㅊ/)이나 파열음(/ㅋ/, /ㅌ/, /ㅍ/)이 있을때나 음성의 끝부분에 비음(/ㄴ/, /ㅇ/) 등이 있을때는 음성의 시작점과 끝점을 분리해 내기가 어렵다. 본 논문에서는 프레임내의 평균에너지와 피크 신호 크기를 이용해 음성구간을 검출하였다[11]. 음성 신호의 통계적 특성이 시간에 따라 거의 변하지 않는 준 정상적(quasi-stationary)인 사실을 이용해 분리하였다.

1) 묵음이 포함된 전 음성 구간을 일정한 시간 간격으로 분할(segmentation)하여 프레임(frame)으로 나눈다

2) 각 프레임에서 평균 에너지와 최대 신호 크기의 비를 구한다.

3) 평균에너지와 최대값의 비(rate)가 최대가 되는 프레임을 찾는다.

4) 음성의 앞과 뒤에서 문턱값(threshold) 이상이 되는 프레임을 음성의 시작점과 끝점으로 한다.

III. 인식 알고리즘

본 논문에 사용한 전체적인 방법은 그림 1에 나타내었다.

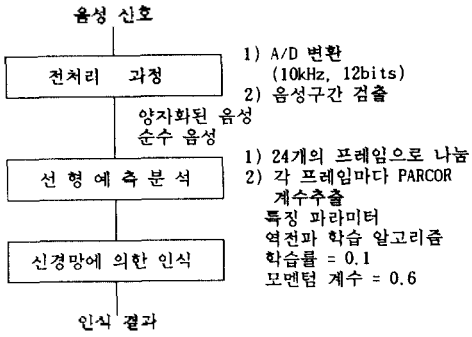
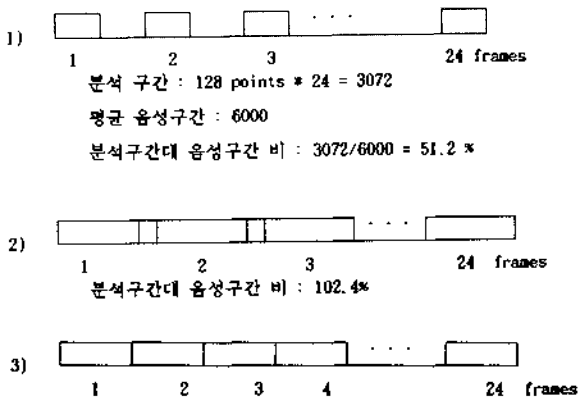


그림 1. 전체 개략도

음성은 시간에 따라 수시로 변화한다. 시간에 따라 변화하는 패턴을 신경망으로 인식하기는 쉽지 않다. 본 논문에서는 음성의 시간에 따른 변화를 흡수하기 위해 음성의 전 구간을 일정한 갯수의 프레임(frame)으로 나누고 각프레임마다 특징을 구하고나서, 모든 프레임에서 구한 특징을 한꺼번에 신경망의 입력으로 하였다. 각 프레임에서 구한 음성 특징으로는 선형 예측 분석에 의한 PARCOR계수와 영교차율, 프레임내의 평균에너지, 잔차에너지를 이용하였다. 본 논문에서 대상으로 하는 단어는 DDD지역명으로 음성구간의 길이가 500 msec - 700 msec로 평균 길이가 약 600msec(6000 samples)이다. 음성 분석 구간은 3가지로 나누어 실험하였다. 하나는 프레임 길이를 128samples로 하고 두번째는 프레임 길이를 256 samples로하고 세번째는 프레임 길이를 가변적으로 하여 음성 구간을 24등분하여 이를 프레임 길이로 하였다.



첫번째의 경우 평균 음성 구간이 분석 구간인 3072 (24 frame * 12.8 msec)개 보다 길므로 분석구간 사이의 간격이 있으며 두번째의 경우는 평균 6개 ((24 frames * 25.6 msec - 6000)/23)가 중복이 되고, 세번째의 경우는 음성구간을 24개로 나누어 분석 구간으로 정하였으므로 음성의 길이에 따라 프레임의 길이가 가변된다. 각 프레임마다는 해밍윈도우(Hamming window)를 사용하였다.

또한 각 프레임마다 특징을

1) 5차의 PARCOR계수

2) 8차의 PARCOR계수

3) 5차의 PARCOR계수 + 영교차율 + 프레임 에너지 + 잔차 에너지등 3가지로 나누어 조사하였다. 첫번째인 경우 신경망의 입력 갯수는 총 120개 (24 frames * 5차 PARCOR 계수)로서, j번째 입력 노드의 의미는

$$i = ((j - 1) * 5) + k$$

이며, 여기서 j = 1, 2, 3,..., 24 (프레임 번호)

k = 1, 2, 3, 4, 5 (PARCOR차수)

로 표현된다면 j번째 프레임의 k차 PARCOR 계수가 된다. 두번째와 세번째의 신경망의 입력 갯수는 192개(24 frames * 8 개 특징)이다.

대량의 어휘를 인식하는데 따른 신경망의 구조를 간단히 하기 위하여 본 논문에서는 출력층의 노드를 이진 형태로 구성하였다. 즉 출력노드가 N개일 때 binary값은 2^N 개가 되므로 이를 이용하여 50개의 단어를 인식하기 위해 출력 노드의 수를 6개로 하였다.

IV. 실험및 결과

실험에 사용한 음성은 표1에 나타난것 같이 장거리 전화에 이용되는 지역명 50개를 대상으로 하였다.

1. 서 울	11. 문 산	21. 여 주	31. 설 악	41. 태 백
2. 부 산	12. 미 금	22. 연 천	32. 양 구	42. 평 창
3. 대 구	13. 발 안	23. 오 산	33. 원 당	43. 홍 천
4. 광 주	14. 수 원	24. 용 인	34. 양 양	44. 남양주
5. 대 전	15. 시 흥	25. 의 왕	35. 영 월	45. 의정부
6. 강 화	16. 성 남	26. 일 산	36. 원 주	46. 장호원
7. 고 양	17. 안 산	27. 파 주	37. 인 제	47. 주문진
8. 과 천	18. 안 성	28. 하 남	38. 정 선	48. 연무대
9. 광 명	19. 안 양	29. 화 성	39. 철 원	49. 장승포
10. 구 리	20. 양 주	30. 도 계	40. 춘 천	50. 동장양

표 1. 인식 대상 단어

남성 2사람이 50개의 지역명을 각각 5번씩 발음한 총 500개의 음성으로 학습을 시켰으며, 인식에 사용한 음성은 학습에 참여하지 않은 2사람의 음성으로 5번, 2번씩 발음한 총 350개의 음성을 대상으로 하였다.

음성은 data acquisition board인 DT2801로 음성 크기를 최대치가 ±10V 이내, 샘플링 주파수 10kHz, 양자화 준위 12 비트로 아날로그-디지탈 변환하여 저장하였다. 학습 횟수는 2000회 이내에서 각 음성에 대해 실제의 출력값과 원하는 출력값의 차, 즉 허용오차가 0.1 이내에 들도록 하였다. 실험에서 사용한 음성 데이타 내용과 특징 파라미터를 표2에 나타내었다.

대 상 음 성	장거리 전화에 사용하는 50개의 지역명	
음 성 채 집	학습	2명의 남성 각각 5번 발음한 500개 음성
	인식	2명의 남성이 5번,2번 발음한 총 350개음성
샘플링 특성	샘플링 주파수 : 10 kHz 양자화 비트수 : 12 bits	
분 석 구 간	프레임 길이	1) 12.8 msec 2) 25.6 msec 3) 20.8 - 29.1 msec
	프레임 갯수	24개
	각프레임 마다 해밍윈도우 사용	
특징 파라미터	1) 5차 PARCOR 계수 2) 8차 PARCOR 계수 3) 5차 PARCOR 계수, 영교차율, 에너지, 잔차에너지	

표 2. 음성 파라미터

본 논문의 제안한 방법에 의한 실험결과를 표3에 나타내었다. 이때 하나의 음성에 대한 프레임 갯수를 모두 24개로 제한하여 사용하였다.

	128 samples/frame	256 samples/frame	가변길이프레임
5차 PARCOR	95 %	85 %	81 %
8차 PARCOR	96 %	87 %	90 %
5차 PARCOR + 영교차율 + 프레임에너지 + 잔차에너지	93 %	90 %	79 %
평 균	94.7 %	87.3 %	83.3 %

표 3. 파라미터별 인식률

일반적으로 한 사람이 같은 단어를 발음하더라도 발성한 단어의 길이는 발음할 때마다 각기 다르게 나타난다. 이 경우 시간적인 신축 방법(DTW)을 이용하면 단어의 길이를 비선형적으로 정규화시킬수 있으나 본논문에서는 계산량을 줄이기 위해서 단어의 길이를 균등한 크기로 정규화 하였다. 프레임 길이를 짧게 한것이 비록 프레임 사이의 데이타는 손실은 있지만 인식률이 가장 높게 나타났다. 즉 음성은 시간적으로 변화가 많으므로 분석구간이 짧아야 할 필요가 있음을 보인다. 데이타의 손실을 없애기 위해서 각기 다른 크기의 음성을 24개로 균등히 나눈 프레임의 인식률이 제일 낮았다. 즉 음성 분석 구간이 길아야 인식률이 높아짐을 알 수가 있다. 특징 파라미터로 순수한 PARCOR계수를 이용한 것이 PARCOR계수와 다른 특징을 이용한 것보다 높게 나타났다. 에너지는 상당히 큰값이므로 신경망에 그대로 입력시킬 수가 없으므로 $\log_{10}(\cdot)$을 취하여 크기를 1자리 숫자로 정규화 하였고, 영교차율도 100으로 나누어 사용하였다. 학습 시간은 5차의 PARCOR계수를 사용할 때가 입력 노드의 수가 적으므로 평균 6시간 38분정도 걸렸으며, 프레임마다 8개의 특징 파라미터를 사용한 8차의 PARCOR계수와 ZCR등을 사용한 경우가 평균 8시간 7분정도 소요되었다.

V. 결 론

본 연구에서는 우리나라 대표적인 50개의 지역이름을 대상으로 인식을 하였다. 시간에 따른 음성의 변화를 입력노드수에 확장하여 포함시켰다. 즉 음성을 24개의 부분으로 균등히 나누고 각 프레임에서 구한 파라미터를 모두 신경망의 입력으로 사용하였다. 신경망의 구조를 줄이기 위해 출력노드를 이진 코드형식으로 하여 학습 계산량을 감소시켰다. 인식률은 분석구간은 비교적 짧을수록 좋았으며, 파라미터는 5차보다 8차의 PARCOR계수가 다소 좋음을 보인다.

앞으로 연속적으로 발음한 단어를 대상으로 하기을 위해서는 단어의 검출과정이 보다 일반적으로 되어야 하며, 분석구간에 대한 연구가 필요하다.

참고 문헌

[1] T.W. Parsons, *Voice And Speech Processing*, McGraw-Hill Book Company, 1986

[2] K.H.Davis, R.Biddulph, and S.Balashek, "Automatic Recognition of Spoken Digits", *J.Acoust. Soc. Am.*, Vol.24, 1952.

[3] R.D. Peacoke and D.H.Graf,"An Introduction to Speech and Speaker Recognition ", *Computer*, vol. 23, no. 8, August 1990.

[4] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Engelwood Cliffs, N.J., Prentice-Hall, 1978.

[5] H.Sakoe and S.Chiba, "Dynamic Programing Optimazition for Spoken Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, pp. 43-49, Feb. 1978.

[6] Lee, Kai-Fu and H.W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", *Proceedings IEEE ICASSP*, pp.123-126, 1988.

[7] T. Kohonen, "The 'neural' Phonetic Typewriter", *IEEE Computer Special Issue on Neural Networks and Neural Computing*, pp. 11-22, March 1988.

[8] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang , "Phoneme Recognition using Time-Delay Neural Network", *IEEE Trans. Vol. ASSP-37*, No. 8, Aug. 1989.

[9] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang : "Phoneme Recognition: Neural Networks vs. Hidden Markov Models", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, April 1988

[10] R.P. Lippman, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, Vol.4, No. 2, pp. 4-22, April 1987.

[11] 김석동,이행세, "신경망을 이용한 우리말 인식에 관한 연구", 한국음향학회지 제11권 3호, 1992.6.