거절기능을 갖는 음성인식 시스템의 시험운용

구 명완
한국통신 연구개발본부 멀티미디어 연구소 음성언어 연구팀

An Experimental Field Trial of Speech Recognition System Based on Word Rejection

Myoung-Wan Koo
Spoken Language Team
Multimedia Technology Research Laboratory, Korea Telecom
mwkoo@smm.kotel.co.kr

요 약

본 논문에서는 거절기능을 갖는 음성인식 시스템의 시험운용에 대해 소개하였다. 거절기능은 소음 단어에 의한 방식과 인식 결과를 확인하는 방식을 둘 다 병행 사용하여 구현하였다. 소음단어는 필러모델을 정의하여 구현하였으며 인식결과를 확인하기 위해서는 선형변별기를 사용하였다. 연구실에서 구축한 음성 DB로 HMM파라미터를 추출한 후 시험운용 6개월 동안 구한 음성 DB로 실험한 결과 84.1%의 인식률을 구하였으며 이때 거절률은 0.8%였다.

1. 서 론

음성인식 시스템의 기술이 발전함에 따라 전화망을 통한 음성인식 시스템이 상용화되기 시작하였다[1]. 상용화 된 음성인식 시스템은 실시간으로 동작되어야 하며, 입력 음성이 인식하기에는 애매한 경우 거절을 할 수 있는 거절기능이 있어야 한다[2]. 또한 사용자가 인식 대상 단어 이외의 말을 하더라도 인식대상 단어만을 찾아낼 수 있는 keyword spotting 기능이 구현되어야 한다. 소규모의 고립단어 인식 시스템인 경우에도 사용자들은 인식대상 단어 이외의 말을 하는 경우가 많이 있으며 또한 주변 소음이 음성인식 시스템에 입력이 되는 경우가 빈번히 발생된다. 이런 경우를 위하여 음성인식 시스템은 인식결과를 내지 않고 거절하는 기능을 구비하여야 한다.

본 논문에서는 음소를 저장의 기본 단위로 구성한 HMM 화자독립 고립단어 인식시스템에서 거절기능을 구현한 방식 및 시험운용결과를 설명하고자 한다. 제 2 장에서는 거절기능에 대해서 설명하고, 제 3 장에서는 거절기능을 구현하기 위해서 필러모델의 구성방식에 대해 설명하고 선형변별기를 이용한 거절기능의 구현방식에 대해서 기술한다. 제 4 장에서는 거절기능을 구현한 음성인식 시스템의 상세구성도를 설명하고 제 5 장에서는 시험운용결과에 대해 기술한다. 마지막으로 제 6 장에서 결론을 맺는다.

2. 인식 거절기능

인식 거절이능이란 입력된 음성으로선 인식하기 어려운 상태를 나타내는 것으로 상용 시스템을 구현하기 위해서는 필수적인 기능이다. 즉 사용자가 시스템의 인식대상 단어 이외의 단어를 말하였을 경우 시스템이 사용자에게 그 단어는 인식대상단어가 아니기 때문에 인식할 수 없음을 알려주게 된다. 그러면 사용자는 자기가 말을 잘못했기 때문에 시스템이 제대로 인식할 수 있는 단어를 말하게 될 것이다. 만약에 이런 인식 거절기능이 없다면 위의 경우에 시스템은 엉뚱한 단어를 인식결과로 알려줄 것이고, 그렇게 되면 시스템의 신뢰도가 떨어지게 되므로 매우 중요한 기능이다.

현재 음성인식 시스템의 기본 알고리즘으로 가장 많이 사용되고 있는 HMM (Hidden Markov Model)방식은 패턴매칭 알고리즘의 하나이다[3]. 즉 훈련과정에서 기준 패턴을 정하고 인식과정에서는 기준 패턴과 가장 가까운 패턴을 선정하여 입력된 음성을 그 패턴으로 인식하는 것이다. 이러한 음성인식 시스템에 거절기능을 구현하기 위해서는 거절역할을 할 수 있는 기준패턴을 사용하는 방식과 인식결과를 확신하는 기능을 추가하는 방식이 있다. 거절역할을 하는 기준 패턴이 사용 될 경우에는 인식과정에서 입력음성에 가장 가까운 기준 패턴으로 거절역할을 하는 패턴이 선정이 되면 거절을 하는 것이며 인식결과를 확인하는 기능은 인식결과로 나온 비터비(Viterbi) 값을 조사하여 인식결과의 신뢰도를 구하여 낮은 경우 거절하는 것이다.

거절역할을 하는 기준 패턴으로 필러(filler) 모델을 사용한다[4]. 이 방식은 필러 모델로 구성된 소음 단어를 정의하고 후보단어와 같이 인식 작업을 수행한 후 소음 단어로 인식되면 입력을 거절하는 방식이다. 거절기능의 성능을 높이기 위해서 AT&T에서는 매 단어를 모델링 할 때 매 단어에 대해서 반단어(anti-keyword)를 동시에 모델링하여 거절기능의 성능을 향상시켰다[5]. 최근에는 음소를 기반으로 한 HMM 음성인식 시스템에서 매 음소 단위로 반음소(anti-phone)를 모델링하여 단어 독립용 거절기능을 구현하였다[6]. 인식결과를 확인하는 과정은 후 처리 과정이라고 하며 신경망을 사용하는 방법[7], likelihood ratio scoring 방법[8] 및 선형변별기를 사용하는 방식이 있다[4].

3. 거절 기능 구현

3.1 필러 모델

필러 모델이란 인식대상 단어 이외의 단어나 혹은 자동차 소리, 울음 소리 등과 같은 비 음성을 모델링하기 위해 사용된다. 인식대상 단어 이외의 단어를 표현하기 위해서는 대표적인 간투사나 많이 사용되는 단어를 독립단어로 모델링하고 자동차 소리, 울음 소리, 숨 소리 등과 같은 비 음성단어일 경우에도 각각 독립단어로 모델링 할 수 있다[9]. 그러나 본 논문에서는 필러 모델을 음소로 간주하였으며 음소 세 개(Z1, Z2, Z3)를 정의하고 이 세 개의 음소가 모여진 단어를 소음 단어로 표시하였다. 그리고 음성 소음과 비음성 소음을 동일한 소음단어로 정의하였다. 그림 1. 에는 본 논문에서 정의한 소음 단어를 표시하였다.

3.2 선형변별기

선형변별기란 매 단어의 인식 확신을 구하기 위해 음성인식 결과인 비터비 값을 미리 정해진 거절 상수와 비교하여 그 값보다 크면 인식 결과를 신뢰하고 그 값보다 작으면 인식결과를 부정

하는 방식이다. 본 논문에서는 다음과 같이 두 종류의 분류값 y1,y2를 사용하였다..

$$y1 = \log Pi(O) - \log Pg(O) \quad (1)$$

$$y2 = \log Pi(O) - \log Pk(O) \quad (2)$$

윗 식에서 Pi(O)는 첫 번째로 인식된 후보단어 i에 대한 HMM 비터비 출력 값이며 Pg(O)는 소음 단어에 대한 비터비 출력 값이다. 그리고 Pk(O)는 두 번째로 인식된 후보단어 k에 대한 비터비 출력 값이다. 이 때 단어인식 결과인 비터비 값을 그대로 사용하지 않고 다른 후보 단어에 대한 비터비 값과의 차이를 사용하는 이유는, HMM 인식 결과인 비터비 출력 값이 최적 경로만 탐색하는 데만 유효하며 입력음성과의 절대적인 유사도를 측정할 수 없기 때문이다. y1,y2로 이루어진 분류값 벡타 Y는 후보단어 i에 해당되는 투영벡타(projection vector) B와 결합되어 거절값 C를 다음과 같이 구한다.

$$C = Y^t B \quad (3)$$

식 (3)에서 C은 후보단어 i에 해당되는 거절상수와 비교되어 이 거절상수보다 크면 인식결과를 인정하고 거절상수보다 작으면 인식결과를 부정하여 거절한다. 매 후보단어에 대한 투영벡타 B와 거절상수는 훈련과정에서 구해지는데 Fisher의 분류 규칙을 사용하여 구한다.[10]

4. 거절기능을 갖는 음성 인식 시스템

4.1 기본 시스템 구성

거절기능을 갖는 음성인식 시스템의 구성도가 그림 3.에 그려져 있다. 먼저 입력 음성이 들어오면 끝점 검출기에 의해 음성만 검출된다. 검출된 음성은 인식 대상 단어 이외에 에, 저 등의 간투사와 자동차 소리, 전화선에서 야기되는 소음도 포함한다. 특징추출 과정에서는 음성의 특징이 추출되고 비터비 검색과정에서는 훈련된 음소, 묵음 및 필러 모델을 사용하여 구성된 단어 발음 사전을 참조로 인식 작업이 수행된다. 비터비 검색에 의해 찾아진 인식 후보 단어는 선형변별기를 이용하여 인식결과를 출력할 것인지 거절할 것인지를 결정한다.

4.2 특징 추출

음성신호는 8khz로 샘플링되고 전달함수가 $1 - 0.95z^{-1}$인 1차 디지탈 필터로 pre-emphasis된다. 이 음성은 매 10msec단위로 LPC분석이 이루어 지고. 주변 잡음에 강하도록 weighting 함수에 의해 변환된다[11]. 사용되는 특징은 (1) 12개의 LPC 켑스트럼, (2) 12개의 LPC 켑스트럼 차, (3) 12개의 LPC 켑스트럼 이차 차이, (4) 파우워 차이 및 파우워 이차 차이로 구성된 4종류의 특징 벡타(38개)를 사용한다. 각 벡타는 훈련 과정에서 구한 4 종류의 VQ(vector quantization) 코우드 북을 사용하여 벡타 인덱스로 표현된다. 3개의 코우드 북은 256개의 코우드 워드로 구성되고 마지막 특징 벡타 (4)는 64개의 코우드 워드로 구성된다.

4.3 음소 HMM모델

기본 시스템은 이산 확률정보를 사용하는 HMM 인식 시스템이며 음소단위로 HMM 파라미터를 추출한다. 본 논문에서는 64개의 문맥독립 음소를 사용하고 이를 기준으로 문맥 종속 음소를 생성한다[12]. 그림 2. 에는 음소 HMM모델이 그려져 있다. 이 모델은 7개의 state와 12개의 transition으로 구성되며 관찰확률은 매 transition 마다 출력된다. 관찰 확률분포는 B,M,E의 3가지 분포로 tying 시켰다.

5. 인식 실험

5.1 음성 데이터 베이스

본 논문에서 사용한 음성 데이터 베이스는 전화망을 통해서 얻어진 음성으로 구성된다. 단어는 할아버지, 할머니 등과 같이 음성 다이얼링 서비스에 이용될 수 있는 150단어로 구성되며 미리 지정된 10대 ~ 50대의 화자가 발음한 음성 DB와 음성인식 시스템을 구축한 후 시험 운영기간('95.4 ~ '95.9)에 얻은 음성 DB로 나누어진다. 표 1. 에는 실험에 사용한 음성데이타 베이스의 상세내용이 나타나있다.

5.2 인식 실험

인식실험을 하기 위해서 먼저 훈련과정이 필요한데 표 1. 에서 보여진 훈련 DB를 사용하여 음소 HMM파라미터를 구하였다. 사용한 알고리즘은 forward-backward 알고리즘이었다. 일차로 61개의 문맥독립 음소의 HMM파라미터를 구하였으며 이것을 초기화하여 129개의 문맥 종속 음소의 HMM파라미터를 구하여 음성인식에 사용하였다. 인식 실험은 표 1. 에서 보여진 인식용 음성 DB를 사용하여 실험을 하였다.

표 2.에는 음성 DB를 사용한 인식결과를 나타내었다. 표에서 C1, .., C6,는 인식 결과를 확신하는 과정에서 사용되는 거절상수이며 이 값이 크게 되면 거절이 되는 경우가 많아지게 된다. 표에서 CA(Correct Acceptance)는 입력 음성이 후보단어일 때 제대로 인식이 되는 단어 갯수를 말하고 CR(Correct Rejection)은 입력 음성이 소음단어일 경우 제대로 거절이 되는 단어 갯수를 말한다. FA(False Acceptance)는 입력음성이 후보단어일 경우 제대로 인식 못하는 단어 개수를 의미하는 FAI와 입력 음성이 소음단어일 때 소음단어일 때 FAO로 구성되어 있으며 FR는 입력음성이 후보단어일 경우 거절이 되는 단어 갯수를 의미한다. 인식결과는 거절되는 현상을 오 인식으로 나타내는 순 인식률과 인식률 계산에 포함시키지 않는 인식률로 나타내었다. 순 인식률은 거절기능을 적절히 사용하면 인식률이 향상되었으며 인식률은 거절기능을 강화함에 따라 인식률이 항상 증가되었다. 이러한 현상의 원인은 소음단어에 의해 기인한다. 즉 적절한 거절기능을 제공하면 소음단어가 후보단어로 오 인식될 경우를 줄여 주게된다.

6. 결론

본 논문에서는 전화망을 통한 음성인식 시스템을 개발함에 있어서 거절기능을 구현하는 방법을 기술하였으며 인식실험도 수행하였다. 거절기능은 두가지 방식을 병행 사용하여 구현하였다. 첫번째 방법은 필러 모델을 이용하는 것이다. 필러 모델을 이용하여 소음 단어를 정의하고 인식 대상 후보 단어와 동일하게 취급하여 인식 결과가 소음 단어로 인식되면 거절이 되도록 하는 것이다. 두 번째 방법은 선형변별기를 사용하는 것이다. 결과인 비터비값을 신경망의 입력으로 사용하

여 출력 노우드 결과에 따라 거절상수 값보다 크면 인식 결과를 내보내고 작으면 거절을 하도록 하는 방식이다. 6개월 시험운용중 얻은 음성 DB를 사용한 인식 실험 결과 84.1%이었으며 이때의 거절률은 0.8%이었다. 그리고 순 인식률은 83.4%였다.

참 고 문 헌

[1] J. G. Wilpon, et al., "Automatic recognition of keywords in unconstrained speech using hidden Markov models," IEEE Trans. Acost,. Speech, Signal Processing, vol. 3d, No.11, pp. 1870-1878, Nov., 1990.

[2] 구 명완, "음성인식 기술을 이용한 새로운 서비스," 제 12회 음성통신 및 신호처리 워크샵논문집, pp. 47-51, Jun., 1995.

[3] 구 명완, "음성인식 기술의 현황과 전망," 정보과학회지, 제11권 제5호, pp. 21-34, Oct. 1993

[4] R. A. Sukkar, et al., "A two pass classifier for utterance rejection in keyword spotting," Proc. of 1993 ICASSP, pp. 451-454, May, 1993.

[5] M. G. Rahim, et al., "Robust utterance verification for connected digits recognition," Proc. of 1995 ICASSP, pp. 285-288, May, 1995.

[6] R. A. Sukkar, et al., "A vocabulary independent discriminatively trained method for rejection of non-keywords in subword based speech recognition," Proc. of EUROSPEECH '95, pp. 1629-1632, Sep., 1995.

[7] D. P. Morgan, et al., "A keyword spotter which incorporates neural networks for secondary processing," Proc. of ICASSP 90, pp. 113-116, Apr., 1990.

[8] R. C. Rose and D.B Paul, "A hidden Markov model based keyword recognition system," Proc. of 1990 ICASSP, pp. 129-132, Apr., 1990.

[9] P. Jeanrenaud, et al., "Phonetic-based word spotter: various configurations and application to event spotting," Proc. of 1993 Eurospeech, pp. 1057-1060, Sep., 1993.

[10] R. O. Duda and P. E. Hart, Pattern classification and scene analysis. New York, Wiley, 1973.

[11] L. Rabiner, B-H Juang, Fundamentals of speech recognition, Prentice-Hall, 1993.

[12] C. H. Lee, et al., "Acoustic modeling of subword units for speech recognition," Proc. of 1990 ICASSP, pp. 721-724, Apr., 1990.

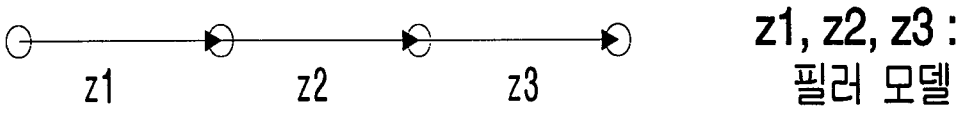

그림 1. 소음 단어 모델

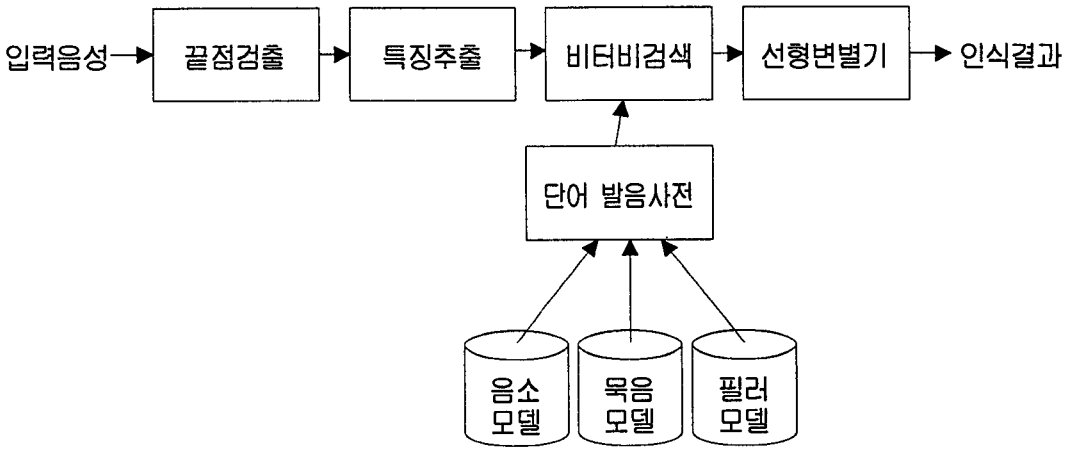

그림 2. 거절기능을 갖는 음성인식 시스템

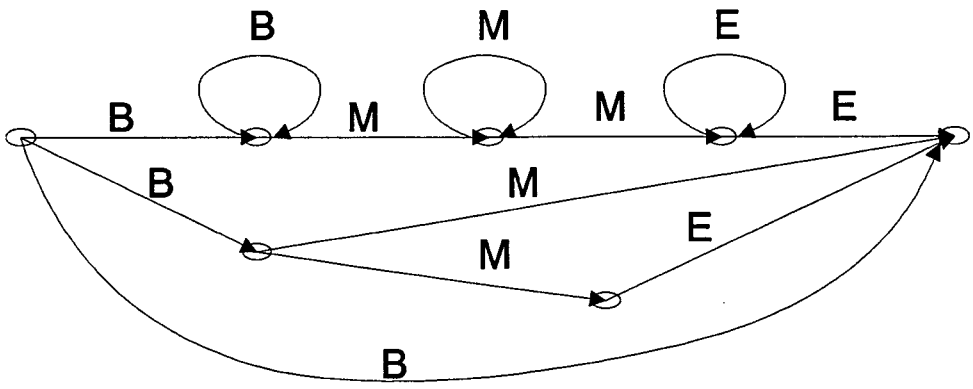

그림 3. 소음 HMM 모델

표 1. 음성 데이터 베이스 구성

훈련용			인식용		
화자	단어		화자	단어	
	후보 단어	소음 단어		후보 단어	소음 단어
500명	72000	375	시험운용	939	55

표 2. 음성인식 실험
(C5 > C4 > C3 > C2 > C1)

방법	CA	CR	FA		FR	인식률(%)	거절률(%)	순인식률(%)
			FAI	FAO				
거절기능 사용안함	791	32	148	23	0	82.8	0.0	82.8
C1	790	36	144	19	5	83.5	0.5	83.1
C2	790	39	141	16	8	84.1	0.8	83.4
C3	785	43	136	12	18	84.8	1.8	83.3
C4	771	38	116	10	59	86.5	5.9	81.4
C5	753	34	98	10	99	87.9	10.0	79.2