제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

기본음소 설정을 위한 음소인식률 이용 방안 연구

김 호경, 구 명완
한국통신 멀티미디어연구소 음성언어연구실

A Study on the method for choosing basic phoneme units based on the phoneme recognition rate

Kim Ho-Kyoung, Koo Myoung-Wan
Spoken Language Research Team, Multimedia Technology Research Lab., Korea Telecom

{hkkim,mwkoo}@smm.kotel.co.kr

요 약

본 논문에서는 한국통신의 음성인식 시스템에서 사용하는 기본음소의 효율적인 설정을 위하여 음소인식률을 이용하는 방안의 연구에 관하여 기술한다. 본 논문에서 제안한 방식은 음성 인식 시스템의 기본 단위라 할 수 있는 기본음소를 설정할 때에 음소인식률을 구하고 유사하게 인식되는 음소들의 집합인 cohort set을 구하여, 인식률을 최대로 하는 기본음소 집합을 찾는 방법이다. 실험 방식은 기본음소 59개로부터 시작하여 음소를 1개씩 줄여가면서 최대 음소 인식률이 나오도록 하였다. 실험 결과 최고 성능을 나타내는 기본 음소 set을 구할 수 있었다.

1. 서 론

한국통신에서는 1991년부터 수행된 음성 언어 연구의 결과물로서 1994년에 음성 인식 증권정보 안내 시스템을 개발하였고, 1995년 말에는 실제로 증권 회사에 설치하여 시험운용하였다. 그리고, 올해 3월부터는 한국통신의 700번 전화정보 서비스로서 32 채널로 시험서비스되고 있다. 현재 시험서비스중인 음성인식 증권정보안내 시스템에서는 59개의 기본음소를 사용하고 조음화 현상을 고려하여 300개의 문맥종속음소로 확장하여 사용하고 있다.

음성인식 기술이란 기계가 인간의 음성을 알아듣도록 하는 것인데, 기계가 인간이 말한 단어(혹은 문장)를 분석하기 위해서는 미리 입력된 충분한 정보를 필요로 하게 된다. 이 정보중에서 가장 중요한 것의 하나는 음성인식 시스템의 인식 대상 단어를 구성하는 기본 음소들에 대한 정보이다. 이 정보는 훈련 과정에서 생성되는데 해당 음성인식 시스템의 기본 단위로서의 기본 음소는 그 정의에서부터 중요한 의미를 지닌다.

본 논문에서는 성능이 우수한 기본 음소를 효율적으로 설정하기 위한 방법에 관한 것으로, 2 장에서는 음성인식 시스템에서의 기본 음소에 대하여 논하고, 3 장에서는 음소인식률을 이용하여 기본 음소를 설정하는 방법 및 실험결과를 분석하고, 마지막으로 4 장에서 결론을 맺는다.

2. 음성인식 시스템

한국통신의 음성인식 시스템은 HMM (Hidden Markov Model) 기술에 기반을 둔 고립단어 인식시스템이며 유사음소(phoneme-like unit)를 기본단위로 사용하고 있다[1][2]. 기본음소는 HMM 파라미터를 생성하는 훈련과정에서 기본이 되는 단위이고 인식단어를 구성하는 요소가 되므로, 그 정의에 따라 음성인식 시스템의 인식 성능에 영향을 주게 된다.

2.1 특징 추출

음성신호는 8KHz, μ-law 8bit로 샘플링되고 $1-0.95z^{-1}$ 의 전달함수를 갖는 필터를 사용하여 pre-emphasize된다. 이 음성은 20msec 길이의 프레임으로 분할되며 10msec씩 중첩된다. 각 프레임들은 14차의 LPC 분석이 수행되고 이 LPC 계수를 이용하여 cepstral 계수가 구해진다. 구해진 LPC 계수는 아래의 가중치 윈도우(weighting window) $Wc(m)$에 의하여 weighting되고, cepstral 계수를 비롯하여 음성 특징을 나타내는 4가지의 계수들을 추출하게 된다. 음성인식에는 12개의 LPC cepstral 계수와 그들의 빼기(difference), 이차 빼기(second order difference), 로그 파워의 일차, 이차 빼기 값을 벡터 양자화하여 사용한다[3].

$$Wc(m) = 1 + \frac{Q}{2}\sin(\frac{\pi m}{Q}), 1 \le m \le Q$$

2.2 훈련 과정

개개의 음성 파일들은 그 말한 내용에 따라 미리 정의한 기본 음소들로 재구성되며, 추출한 특징벡터들을 이용하여 음소 단위로 HMM 훈련 과정을 거친다. 예를 들어 "한국통신"이라고 말한 음성 파일이 있다면 이 파일은 "ㅎ","ㅏ","ㄴ","ㄱ","ㅜ","ㄱ", ... 등의 음소로써 재구성된다. 물론 같은 "ㄱ" 음소라고 할지라도 초성에 쓰이는 "ㄱ"과 종성에 쓰이는 "ㄱ"은 엄격하게 말하면 서로 다른 소리를 가지고 있다. 모든 음소는 그 음소가 사용된 위치와 주위의 음소들의 영향을 받아 그 소리가 변화하여 여러 가지의 소리값을 갖게 된다. 즉, 한 개의 음소는 다양한 소리값에 따라서 다시 몇 개의 단위로 세분화 될 수 있다. 이렇게 세분화되어 기본적인 단위로 나뉘어진 음소들을 "기본 음소"라고 부르는데, 보통 한국어 음성인식 시스템의 경우에 이 기본음소는 40~60개 정도를 사용한다. 이 기본 음소의 개수를 너무 적게 정한다면, 분명히 다르게 발음되는 소리도 한 가지로 표현하게 되므로 그 소리를 정확하게 표현하지 못하는 결과가 되어버릴 것이다. 반대로, 너무 많이 정한다면 유사하게 발음되는 소리도 너무 세분화하는 결과가 되어 인식에 혼돈을 주고 말 것이다. 이렇게 정의된 기본 음소들은 훈련 과정을 거쳐서 문맥독립음소(context-independent phone : CI)로서 생성되고 이 CI를 기반으로 하여 우리는 문맥종속음소(context-dependent phone : CD) 까지 확장을 한다. CD는 CI를 좀 더 세분화한 음소들로서 문맥에서의 전후 음소들의 영향을 받은 조음현상까지를 고려하여 만든 것이다[4]. 이렇듯 훈련 과정은 HMM 모델에서의 확률값들을 계산하여 최적화 시키는 작업이라고 할 수 있으며, CI 훈련을 거쳐서 CD 훈련으로 이어진다.

2.3 인식 과정

"인식 과정"은 현재 입력된 음성이 어떤 단어를 발화한 것인가를 판단하여 그 인식결과를 내주는 과정이다. 인식하고자 하는 단어가 입력되면 그 음성의 특징들을 추출하여, 인식대상이 되는 후보단어들 중에서 Viterbi 검색을 한다. Viterbi 검색 결과로 나오는 수치는 각 후보 단어들의 입력된 음성과의 유사도를 나타내는 것으로 그 순서대로 후보 단어를 5개 선정하고[5] 이중에서 가장 큰 유사도를 갖는 단어로서 입력된 음성의 인식결과를 내준다. 각 후보 단어들의 Viterbi 검색은 우선 이 단어들을 우리가 정의한 기본 음소에서 확장된 CD 음소들로 재구성하고 현재 입력된 음성에서 추출한 특징들을 이용하여, 훈련 과정에서 생성한 HMM 모델에 적용하여 그 유사도를 계산하는 것으로 이루어진다.

이와 같이 음성인식 시스템에서 기본 음소는 가장 기본이 되는 단위이다. 훈련 과정에서 생성되는 HMM 파라미터 값들도 이 기본 음소에서 확장된 음소들의 단위로 생성 된다. 인식 과정에서는 단어단위로 인식을 수행하지만 인식 결과로 후보 단어들을 선정할 때에는 음소단위인 HMM 파라미터를 사용하게 된다. 따라서 기본 음소의 설정 단계에서의 중요성은 음성인식 시스템의 인식성능에 큰 영향을 미치게 된다.

3. 음소인식률을 이용한 실험 과정

3.1 기본음소의 설정 방법

기존의 음성인식 시스템에서는 음성학자들이 제안한 기본 음소를 사용하였다. 이 방법은 음성학자들이 제안한 음성, 음운학적인 기본 음소들을 이용하여 훈련과정을 거쳐서 우리의 음성인식 시스템에 적합한 음소의 개수를 결정하여 사용하는 것이었다. 기본 음소를 정의하고 HMM 훈련과정을 거쳐서 훈련된 CI 음소를 생성하고, 다시 조음현상을 고려하여 CD 음소로 확장하고 훈련작업을 반복하였다. 그러나, 이렇게 학자들이 정의하였던 음소들 중에선 아주 유사하게 발음되어 음성인식 시스템이 구별하지 못하거나, 보통의 화자가 구분하여 발음하지 않는 음소들도 세분화되어 있는 경우도 있었다. 따라서 이러한 음소들은 인식 단계에서 서

로 발음이 비슷한 다른 음소로 오인식 되고 따라서 인식 시스템의 인식 성능을 저하시키는 요인이 되었다. 본 논문에서 제안하는 새로운 방법은 CI나 CD 음소를 이용한 HMM 훈련과정에 들어가기 이전의 기본 음소 단계에서, 최고의 인식 성능을 나타내는 기본 음소 집합을 찾는 방법에 관한 것이다.

3.2 음소인식률을 이용하는 방법

음소 인식률을 최고로 하는 기본 음소 개수를 선정하는 방법은 그림 1과 같다. 우선, 기존에 우리가 사용하고 있던 기본 음소에 대한 음소 인식률을 계산하여 기준으로 삼는다. 다음으로 cohort set(유사 음소 집합)을 구하는데, 이것은 기본 음소 각각에 대하여 그와 가장 유사하게 발음되는 음소들의 집합을 말한다. 서로 유사하게 발음되어 인식에 오류를 일으키기가 쉬운 음소들을 순서대로 N개씩 구한다. 구해진 cohort set에서 가장 유사한 음소와 해당 음소가 동일한 음소라고 판명이 되는 경우, 이 음소들을 하나의 음소로 통합하여 설정한다.

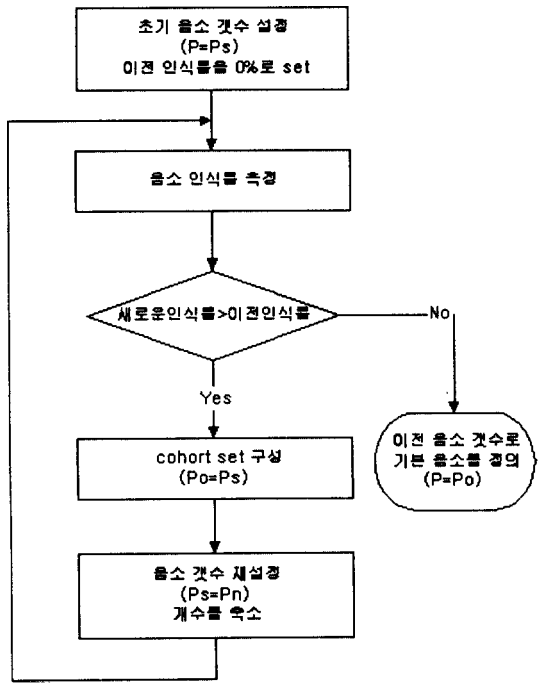

그림 1. 음소인식률을 최고로 하는 음소 개수 설정 방법

각 음소별 인식률을 살펴보면 다른 음소와의 변별력이 좋아서 인식률이 좋은 음소들도 있는 반면에, 유사한 다른 음소로 오인식되어 인식률이 저조한 음소들도 있으므로 이러한 음소들은 하나로 설정하는 것이다. 이렇게 기본 음소의 개수를 줄여서 조정하고 새로이 정의한 기본 음소들을 이용하여 HMM 초기음소 훈련을 한 후에 음소 인식률을 다시 구하여 인식률의 변화를 조사하였다. 새롭게 정의한 기본 음소를 사용한 음소 인식률이 오히려 저하되었다면 이전의 기본 음소들이 최고의 음소 인식률을 내는 것으로 결정하고 기본 음소 개수 설정 작업을 종료한다. 하지만, 음소 인식률이 향상되었다면 다시 cohort set을 구하는 작업을 반복한다. 기본 음소의 개수를 축소하는 과정을 반복하여 최고의 음소 인식률을 나타내는 기본 음소 집합을 구한다. 이렇게 향상된 음소 인식률을 갖는 기본 음소들을 이용하여 훈련 과정을 거치면 음성인식 시스템의 전체 인식 성능도 향상되는 것이다.

3.3 실험 결과

실험은 증권정보 안내 시스템용 음성 DB를 사용하였는데, 94년부터 97년 동안에 수집된 음성 파일을 자동으로 세그멘테이션한 정보를 이용하였다. 이 중에서, 기본 음소의 개수가 새롭게 정의될 때마다 HMM 기본 음소 훈련을 하기 위하여 80%를 사용하였고, 음소 인식률을 테스트하기 위하여 나머지 20%를 사용하였다.

이번의 실험에서는 기존의 음성인식 시스템에서 사용하는 이산 HMM 대신 연속 HMM을 사용하였으며, 8개의 mixture를 사용하는 가우시안 확률(Gaussian Probability) 분포를 사용하였다[6].

기존의 음성인식 증권정보 안내 시스템에서 사용하는 기본음소는 상장된 증권회사 명칭에 포함된 음소들로서 59개를 사용하였으며[7], 이 때의 음소인식률은 54.29%이었고, 이 기본 음소들 중에서 서로 유사하게 발음되는 음소들을 하나로 묶어서 44개의 기본 음소로 정의하여 다시 음소 인식률을 테스트하였다. 그 결과 오히려 인식률이 저하되어서 53개의 기본 음소로 설정하여 실험한 결과 처음의 59개 보다 향상된 음소 인식률 56.27%를 얻을 수 있었다. 다음으로는 53개의 기본 음소에서 한 개씩의 음소를 추가,삭제하는 실험을 반복하였고, 그 실험결과는 그림 2와 같았다. 실험 결과, 51개의 음소로 정의하였을 때의 인식률이 가장 우수하였다.

기본 음소의 개수가 같다고 하더라도 그 안에 정의되어 있는 음소들에 따라 음소 인식률이 서로 다른 것은 당연하다. 그림 2의 실험 결과는 해당하는 기본 음소의 개수 집합에서 실험한 중에서 가장 우수한 인식성능을 나타낸 집합의 인식 성능을 보여준다.

어떤 한 음소를 추가하거나 제외시키거나(제외한다

는 것은 가장 유사한 다른 음소와 하나의 음소로 정의하는 것을 말한다)하여 새로이 실험 과정을 되풀이하여 인식률을 계산할 때 많은 시간이 소요되므로, 음소 인식률에 변화가 있을 만한 음소들만으로 음소 인식률을 미리 구하는 방법을 사용하였다. 하지만 이 시간은 HMM CI훈련이나 CD 훈련 작업에 비하면 상대적으로 아주 짧은 시간이므로 여러 가지의 반복 실험이 가능하다.

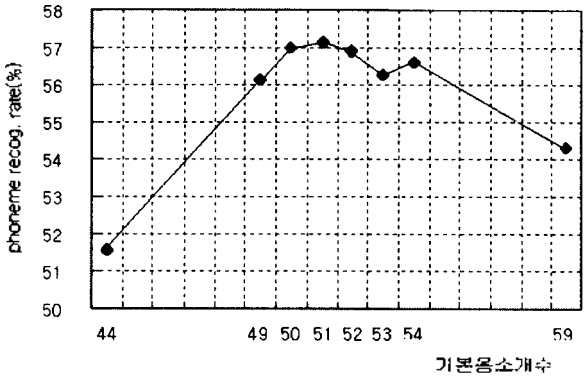

그림 2. 기본 음소 개수에 따른 음소 인식률

4. 결 론

본 논문에서는 음성인식 시스템에서 사용하는 기본 음소의 설정을 위하여 음소 인식률을 이용하는 방법에 대하여 기술하였다. 지금까지는 음성,음운학자들이 한국어의 음소로 정의한 기본 음소를 그대로 사용하였다. 그러나, 기본 음소는 음성 인식 시스템의 기본 단위로서 그 정의에 따라서 HMM 훈련 과정을 거쳐 생성되는 파라미터나, 인식 성능에 큰 영향을 줄 수 있다.

최고의 음소 인식률을 나타내는 기본 음소를 구하기 위하여, 음소 인식률과 유사음소집합(cohort set)을 구하는 실험을 반복하였다. 실험 결과 51개의 기본 음소를 정의하였을 때의 음소 인식률이 기존의 59개로 정의하였을 때보다 2.7%가 향상되었다.

지금도 최적의 기본 음소 집합을 구하기 위한 계속적인 연구가 진행되고 있으며 실험결과로 구해진 기본 음소는 현재 시험 서비스 중인 "음성인식 증권 정보 안내 시스템"에 적용되어 인식 성능을 향상시키게 될 것이다.

참 고 문 헌

[1] K. -F. Lee, Automatic speech recognition: the development of the SPHINX system. Kluwer Academic Publisher, Norwell, Mass., 1989.

[2] C. H. Lee et al., "Acoustic modeling of subword units for speech recognition," in Proc. 1990 IEEE Int. Conf. Acoust., Speech, Sognal Prcessing, pp. 721-724, April 1990.

[3] M. W. Koo et al., "KT-STOCK : A speaker-independent, large-vocabulary speech recognition system over the telephone," in Proc. 1994 Int. Conf. on Spoken Lang. Processing, pp. 1387-1390, Sep., 1994.

[4] 구 명완, "N개의 최적문장을 찾을 수 있는 한국어 연속음성인식 시스템," 제 11회 음성통신 및 신호처리 워크샵 논문집, pp. 48-51, 1994년 10월.

[5] Y. Chow, et al., "The N-best algorithm: an efficient procedure for finding top N sentence hypotheses speech recognition performance," Proc. Speech and Natural language, pp. 147-149, Oct. 1989.

[6] 구 명완, "HMM 훈련 알고리즘에 따른 음소인식률 비교 연구,", 음성통신 및 신호처리 워크샵 논문집 발표 예정, 1998

[7] 김 재인, 구 명완, "음성인식 증권정보시스템의 개발 및 시험운용결과 분석,", 제 13회 음성통신 및 신호처리 워크샵 논문집, pp. 185-191, 1996.