2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

잡음 환경에서의 복수 화자 음성인식

오윤학, 허호영, 송명규, 김형순

부산대학교 전자공학과

Multi-Speaker Speech Recognition in Noisy Environments

Yoon Hark Oh, Ho Young Hur, Myung Gyu Song, Hyung Soon Kim

Dept. of Electronics Engineering, Pusan National University

E-mail:{ntoh1, hyher, mgsong, kimhs}@hyowon.cc.pusan.ac.kr

요 약

본 논문에서는 잡음 환경에서 복수 화자 음성인식 시스템의 인식 성능 향상에 관한 실험을 하였다. 복수 화자 음성인식 방식은 훈련에 참여한 복수의 사용자에 대한 등록 단어 모델을 가지므로, 인식 단계에서 등록 화자의 모든 단어 모델들을 테스트 음성과 비교하여 인식 단어를 결정한다. 그러나, 이 경우 훈련 환경과 테스트 환경의 불일치에 기인한 인식 성능 저하가 등록 화자수가 많아짐에 따라 더욱 심해지는 문제가 발생한다. 본 논문에서는 이 문제의 해결을 위해 등록 화자들의 모든 단어 모델들을 테스트 음성과 비교하는 대신 화자인식 시스템을 사용해서 발성 화자와 유사한 후보 화자들의 단어 모델들에 대해서만 테스트 음성과 비교하는 방식을 적용함으로써 기존의 방법보다 높은 단어 인식율을 얻을 수 있었다.

1. 서 론

가정의 TV 등과 같은 가전 기기에 응용되는 음성인식 시스템은 여러 명의 가족 구성원이 사용하게 된다. 여기에는 두 가지 방식의 음성인식 시스템이 적용될 수 있다. 먼저 화자독립 인식 시스템은 임의의 화자에 대해서 사용이 가능하지만, 사용자가 등록 단어를 추가하거나 변경하기가 용이하지 않다. 또한 화자독립 인식의 경우 계산량이 많이 소요되므로, 충분한 계산 처리 능력을 가지지 못한 제품에는 적용하기 곤란하다. 두 번째로 화자종속 인식 시스템은 특정 화자에 대해 우수한 인식 성능을 가지지만 복수 사용자의 이용시 자신이 누구인지를 등록시켜야 하는 불편함이 따른다. 본 논문에서는 자신의 음성으로 훈련 과정을 거친 복수의 사용자에 대해 화자종속 인식 시스템의 성능을 가지면서도 마치 화자독립 환경에서 인식을 수행하는 것과 같은 효과를 얻을 수 있는 복수 화자 음성인식 방식을 적용한다. 실험에 사용한 인식 시스템으로는 적은 훈련 데이터로 우수한 인식 성능을 가지는 동적 시간 정합(DTW) 방식을 채택하였다. 복수 화자 음성인식 시스템은 훈련에 참여한 복수의 사용자에 대한 등록 단어 모델을 가지므로, 인식 단계에서 등록 화자들의 모든 단어 모델들을 테스트 음성과 비교하여 인식 단어를 결정한다. 그러나, 이 경우 훈련 환경과 테스트 환경의 불일치에 기인한 인식 성능 저하가 등록 화자수가 많아짐에 따라 더욱 심해지는 문제가 발생한다.

본 논문에서는 이 문제의 해결을 위해 화자인식 시스템을 전단계로 사용해서 발성 화자와 특성이 유사한 후보 화자들의 단어 모델들에 대해서만 테스트 음성과 비교하는 방법을 적용한다. 이미 90년대 초 Reynolds가 소수 화자의 화자 종속 모델들을 가지고 화자 독립 인식 시스템을 구현하기 위해 화자인식 시스템을 도입하는 실험을 한 바 있다[1]. 하지만, 잡음 환경에서는 화자 인식율의 저하로 인해 현저한 음성 인식율의 저하를 초래한다. 본 논문에서는 *N*-best 기법을 적용하여 다양한 잡음 환경에서도 발성 화자가 *N*명의 후보 내에 포함될 화자 인식율이 90%이상이 되도록 하여 음성인식 시스템이 *N*명의 후보 화자의 단어 모델들만을 테스트 음성과 비교함으로써 기존의 방법보다 높은 음성 인식율을 얻을 수 있었다.

2. 복수 화자 음성인식 시스템

복수 화자 음성인식 시스템은 훈련에 참여한 복수의 사용자에 대한 등록 단어 모델을 가진다. 본 실험에 사용된 동적 시간 정합(DTW) 인식 시스템의 훈련 단계에서는 각 등록 화자가 등록 단어를 두 번 발성하여 얻은 두 패턴을 동적 시간 정합 방법을 이용하여 시간적으로 정합시킨 다음 평균을 취해서 reference 패턴을 얻는다. 인식 단계에서는 임의의 음성이 들어올 때 테스트 화자가 누구인지, 발성된 음성이 무엇인지에 대한 정보가

없으므로 M 명의 등록 화자 각각의 Q개의 단어 모델들을 테스트 음성과 비교하여 인식 단어를 결정한다. 일반적인 복수 화자 음성인식 시스템의 구성도는 그림 1 과 같다.

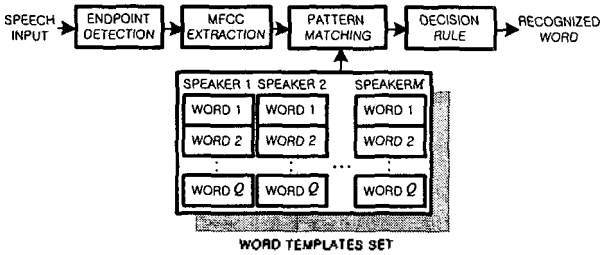

그림 1. 복수 화자 음성인식 시스템의 구성도

패턴 매칭의 결과로부터 인식 단어를 결정하는 방법은 여러 가지가 있을 수 있으나, 본 논문에서는 다음과 같은 두 가지 방법을 검토하였다.

첫 번째 방법은 모든 등록 단어 모델 $M \times Q$개중 테스트 단어와 가장 distance가 작은 단어를 선택하는 방법이다. 이 방법으로 구현한 복수 화자 음성인식 시스템에 대해 clean, 20dB, 15dB, 10dB 의 4 가지 다른 테스트 환경에서 인식 실험을 한 결과를 그림 2 에 도시하였다. Clean 환경에서는 등록 화자수가 적은 경우에 비해 등록 화자수가 많은 경우의 인식 성능이 우수하다. 이것은 테스트 시 등록 화자가 훈련 시와 조금 다르게 발성하더라도 그 단어에 대한 M명의 등록 화자들의 단어 모델과 비교하게 됨으로써 오히려 인식 성능이 향상되기 때문이다. 그러나, 훈련 환경과 테스트 환경의 불일치에 기인한 인식 성능 저하는 등록 화자수가 많아짐에 따라 더욱 심해진다. 즉, 테스트 음성에 잡음이 섞이면 음향적 공간에서의 왜곡이 일어나고, 왜곡된 테스트 음성과 등록 화자의 모든 단어 모델들을 비교한 distance 들도 영향을 받는다. 등록 화자수가 증가함에 따라 잡음에 영향을 받은 distance 들도 늘어나므로 가장 distance 가 작은 단어를 선택하는 인식 방식의 경우 오인식율이 증가하게 된다.

두 번째 방법은 등록 단어별로 M명의 화자에 대한 모델을 가지므로, 각 단어별로 테스트 단어와의 distance 가 적은 N ($N \le M$)개를 선택하여 평균을 취하고 이 평균 distance 가 가장 작은 단어를 선택하는 방법이다. 이 방법으로 구현한 복수 화자 음성인식 시스템으로 N이 1, 2,…,10 인 경우에 대해 인식 실험을 하였다. N이 1 인 경우는 첫 번째 방법과 동일한 결정 방법이다. 실험 결과 등록 화자수가 많고 훈련 환경과 테스트 환경의 불일치가 심한 경우는 N이 2 일 때 인식율이 가장 우수하였고, 그 외의 경우에서는 N이 1 일 때, 즉 첫 번째 방법의 인식율이 가장 우수하였다. 만약 등록 화자수와 훈련 환경과 테스트 환경의 불일치에 따른 최적의 N값을 선택할 수 있다면 첫 번째 방법에 비해 약간의 인식 성능이 향상되지만, 등록 화자수와 테스트 환경에 대한 추정 알고리즘이 추가적으로 필요하다.

그러므로 본 논문에서는 첫 번째 방법을 이용한 복수 화자 음성인식 시스템을 기존 방식의 복수 화자 음성인식 시스템이라 정하고 baseline 으로 선정하였다.

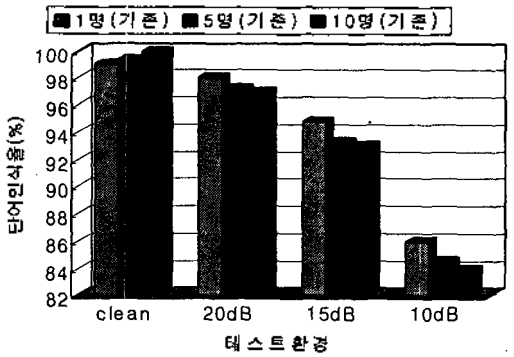

그림 2. 기존 방식의 복수 화자 음성인식 시스템에서 화자수의 증가에 따른 단어 인식율

3. 화자인식 시스템을 적용한 복수 화자 음성인식 시스템

본 논문에서는 인식 단계에서 화자인식 시스템을 전단계로 두어서, 테스트 화자와 음향적인 차이가 적은 N명의 화자들의 단어 모델들에 대해서만 테스트 음성과 비교하는 방식을 적용하였다. 실험에는 GMM (Gaussian Mixture Model)에 기반한 화자인식 시스템을 사용하였다[2].

화자인식 시스템을 전단계로 둔 복수 화자 음성인식 시스템의 전체적인 구성도는 그림 3 과 같다. 검출된 음성의 특징 벡터열이 들어오면, 화자 인식 시스템에서 등록 화자들의 모델 중 발성 화자와 유사한 후보 화자 N명의 인덱스를 등록 화자들의 단어 모델들의 집합에 전달한다. 그리고 나서, 검출된 음성의 특징 벡터열을 N명의 후보 화자 각각의 Q개의 단어 모델들 $N \times Q$개와 비교하여 테스트 단어와 가장 근접한 단어를 선택한다.

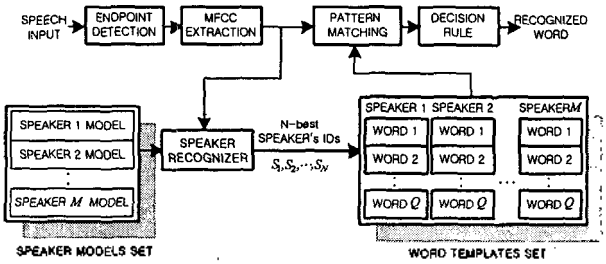

그림 3. 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 구성도

4. 실험 및 결과

복수 화자 음성인식 시스템의 훈련과 인식을 위한 음성 데이터베이스로는 Texas Instrument(TI) 46 DB[3] 중 숫자 및 명령어 20 개를 사용하였다. 이 데이터베이스 중 남성 5 명과 여성 5 명의 데이터를 추출하여, 모델 훈련을 위해서는 훈련용 DB 에서 2 회의 발성을, 인식 실험을 위해서는 인식용 DB 에서 2 회의 발성을 이용하였다. 이 데이터베이스는 12.5kHz 로 샘플링 및 16bit 로 양자화되어 있는데, 본 논문에서는 8 kHz 로 다운 샘플링하여 이용하였다. 실험에 사용된 10 개의 숫자와 10 개의 명령어는 다음과 같다.

ZERO	ONE	TWO	THREE	FOUR
FIVE	SIX	SEVEN	EIGHT	NINE
ENTER	ERASE	GO	HELP	NO
ROBUST	REPEAT	STOP	START	YES

특징벡터는 일반적으로 잡음환경에 강인하다고 알려진 10 차 MFCC 파라미터를 사용하였다. 입력 음성에 부가되는 배경 잡음으로는 백색잡음을 이용하였다. 본 실험에서 화자 모델과 각 화자의 단어 모델들은 clean 환경에서 훈련하였고 10dB, 15dB, 20dB, 그리고 clean 의 4 가지 신호대잡음비 환경에서 인식 실험을 하였다. 단어 인식 단계에서는 여러 가지 잡음처리 기술 중 계산량이 적으면서도 잡음에 강인한 인식 성능을 가지는 projection distance measure[4]를 적용하여 잡음 환경에서의 인식 성능 저하를 현저히 줄였다. 그리고 잡음이 부가된 음성에 대한 끝점 검출 오류로 인해 인식 성능의 저하가 야기될 수 있다. 그러나 본 논문의 실험에는 기존 방식의 복수 화자 음성인식 시스템과 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 성능 비교에 초점을 맞추고자 한다. 따라서 본 논문의 실험에는 잡음 환경에서 추가적인 끝점 검출의 오류가 없다고 가정하고, 끝점 검출을 한 clean 음성에 잡음을 부가하여 잡음 환경에서의 끝점 검출된 음성을 만들어 사용하였다. 추후 잡음 환경에서 강인한 성능을 가지는 끝점 검출 방안에 대한 연구가 더 필요하다.

화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 첫 단계 실험은 화자인식 시스템이 발성 화자와 유사한 3 명의 후보 화자들을 인식하는 실험이다. 화자 모델의 Gaussian mixture 수는 각각 8, 16, 32 를 사용하여 등록 화자가 각각 5 명, 10 명인 경우에 대해 발성 화자가 3 명의 후보 내에 들어올 인식율을 구하였다. 화자인식에 사용된 음성 벡터열은 검출된 음성의 특징 벡터열 중 전체 프레임을 다 이용하는 경우와 전체 프레임 중 유성음 구간만 이용하는 경우의 2 가지에 대해 실험을 하였다. 실험 결과 등록 화자수에 상관없이 모든 테스트 환경에서, mixture 수가 16 이고 전체 프레임 중 유성음 구간만 이용하는 경우의 화자인식 성능이 가장 우수하였다. 그림 4 는 등록 화자수가 10 명인 경우 mixture 수가 8, 16, 32 인 화자 모델에 대해 유성음 구간만 이용하여 화자인식을 하였을 때의 인식 결과를 도시한 것이다. 15dB 환경에서 테스트하는 경우에도 발성 화자가 3 명의 후보 화자 내에 들어올 인식율이 93% 이상이다. 환경 불일치가 가장 심한 10dB 에서는 78% 정도의 저조한 화자인식율을 나타내었다.

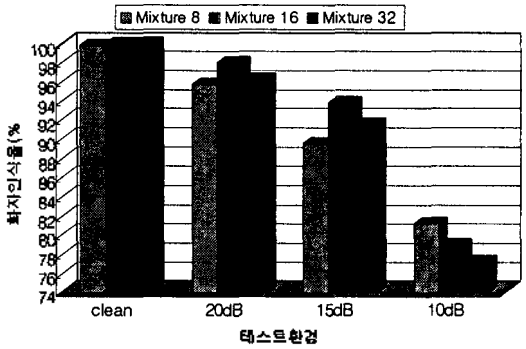

그림 4. Mixture 수가 8, 16, 32 인 경우 화자인식 시스템의 best-3 화자인식율

그 다음 단계로 화자인식 시스템의 결과인 발성 화자와 유사한 3 명의 후보 화자들의 단어 모델들과 테스트 음성을 비교하여 가장 distance 가 작은 단어를 인식하는 실험을 수행하였다. Mixture 수가 각각 8, 16, 32 인 화자인식 시스템의 인식 결과인 3 명의 후보 화자들의 등록 단어 모델들을 테스트 음성과 비교하여 단어 인식율을 구하였으며, 이 실험 결과를 그림 5 에 도시하였다.

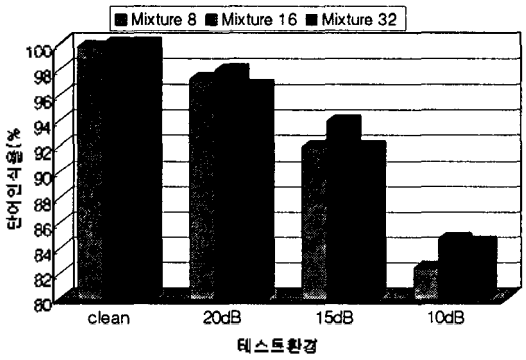

그림 5. Mixture 수가 8, 16, 32 인 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 단어 인식율

화자인식 시스템을 적용한 복수 화자 음성인식 시스템은 mixture 수가 16 인 화자인식 시스템을 적용하는 경우 가장 우수한 단어 인식율을 가지는데, 기존 방식의 시스템과 비교하기 위해 등록 화자가 각각 5 명, 10 명인 경우의 단어 인식율을 그림 6 에 함께 도시하였다. 모든 테스트 환경에서 등록 화자수가 같은 경우 기존 방식의 시스템보다 인식 성능이 개선되었다.

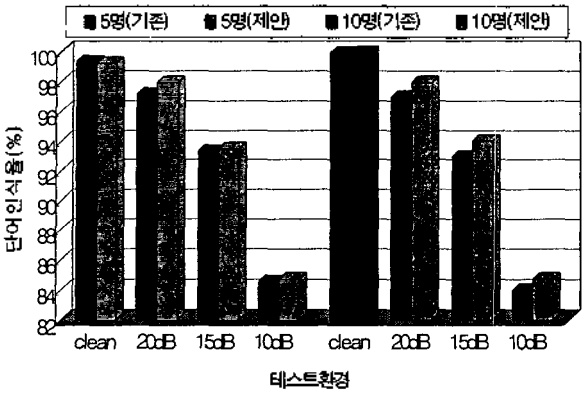

그림 6. 기존 방식의 시스템과 화자인식 시스템을 적용한 복수화자 음성인식 시스템의 단어 인식율

만약 어떤 잡음 환경에서도 발성 화자가 3 명의 후보 화자 내에 포함될 화자인식율이 100% 에 가깝다고 한다면 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 단어 인식율은 더욱 향상될 것이다. 본 논문에서는 clean 환경에서의 화자인식 시스템의 인식 결과인 3 명의 후보 화자들을 잡음 환경에서의 화자인식 시스템의 인식 결과로 사용하고 그 3 명의 후보 화자의 단어 모델들과 테스트 음성과 비교하는 경우의 실험을 참고용으로 수행하였다. 이 실험 결과를 그림 7 에 도시하였다. 훈련 환경과 테스트 환경의 불일치가 심해짐에도 불구하고 등록 화자수가 적은 경우보다 많은 경우 오히려 인식 성능이 향상되었다. 즉, 테스트 음성에 잡음이 섞이면 음향적 공간에서의 왜곡이 생기고, 왜곡된 테스트 음성과 등록 화자의 모든 단어 모델들을 비교한 distance 들도 영향을 받는다. 그러나 등록 화자수가 증가하더라도 발성 화자와 유사한 3 명의 후보 화자의 단어 모델들만 비교하므로 잡음에 영향을 받은 distance 들이 늘어나지 않는다. 따라서 등록 화자수의 증가에 따라 오인식율이 증가하지 않는다. 또한 사용자가 훈련시와 조금 다르게 테스트 음성을 발성하더라도 등록 화자가 1 명인 경우보다 3 명 이상인 경우 각 단어에 대한 3 명의 후보 화자의 단어 모델들과 비교하게 됨으로써 오히려 단어 인식율이 향상된다. 이 실험의 결과로부터 잡음 환경에서 화자인식 시스템의 성능을 개선시킨다면, 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 단어 인식율은 향상될 것이라는 것을 알 수 있다.

계산량의 관점에서 볼 때, 기존 방식의 복수 화자 음성인식 시스템은 인식 단계에서 등록 화자들의 모든 단어 모델들을 테스트 단어와 비교해야 하므로 등록 화자수와 등록 단어수가 증가할수록 인식 시간의 증가를 초래한다. 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 경우 단어 모델들을 비교하는 계산량은 감소하고 화자를 인식하는 계산량은 추가되므로 전체적으로 볼 때 등록 화자수가 증가하고 인식 단어가 많아질수록 화자인식 시스템을 적용한 복수 화자 음성인식 시스템이 계산량 측면에서 유리하다. 인식 어휘가 20 개일 때 등록 화자수가 5 명인 경우는 기존의 방식의 시스템에 비해 약 1.16 배가 소요되고, 10 명인 경우는 0.87 배로 오히려 줄어든다. 계산량의 추가적인 감축을 위해서는 앞으로 더 연구가 필요하다.

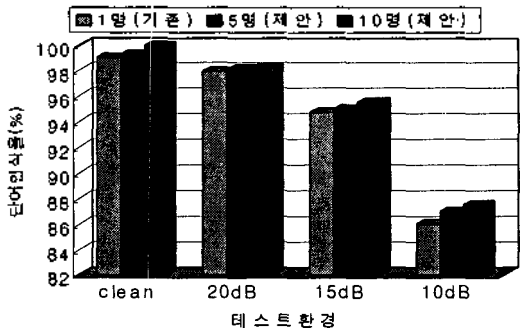

그림 7. 이상적인(clean 환경) 화자인식 시스템을 적용한 복수 화자 음성인식 시스템의 단어 인식율

5. 결 론

본 논문에서는 잡음 환경에서 복수 화자 음성인식 시스템의 인식 성능을 향상시키기 위한 수단으로 화자인식 시스템을 도입하였다. 화자인식 시스템을 전단계로 두어 발성 화자와 유사한 3 명의 후보 화자의 단어 모델들만 테스트 단어와 비교하는 방식을 적용하였다. 실험 결과 다양한 잡음 환경에서도 발성 화자가 3 명의 후보 내에 포함될 화자 인식율이 90% 이상이었고, 단어 인식율은 기존 방식의 시스템보다 우수함을 확인할 수 있었다. 앞으로 잡음 환경에서의 끝점 검출 및 화자인식 성능의 향상에 대해 계속 연구할 계획이다. 개선된 복수 화자 음성인식 시스템은 가정용 전자 제품 등에 대한 음성인식 제어장치 개발에 도움이 될 것으로 기대된다.

참 고 문 헌

[1] D. A. Reynolds and L. P. Heck, "Integration of speaker and speech recognition systems," in Proc. IEEE ICASSP, vol.2, pp.869 -872, 1991.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech and Audio Processing, vol. 31, pp.72-83, Jan. 1995.

[3] G. R. Doddington and T. B. Schalk, "Speech recognition: turning theory to practice," IEEE Spectrum, pp.26-32, Sep. 1981.

[4] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.37, No.11, pp.1659-1671, Nov. 1989.