한국지능시스템학회 논문지 2011, Vol. 21, No. 2, pp. 254-259
DOI : 10.5391/JKIIS.2011.21.2.254

GMM을 이용한 응급 단어와 비응급 단어의 검출 및 인식 기법

Detection and Recognition Method for Emergency and Non-emergency Speech by Gaussian Mixture Model

조영임* · 이대종**
Young Im Cho* and Dae Jong Lee**

* 수원대학교 컴퓨터학과
** 충북대학교 전기전자컴퓨터공학부

요 약

일반적으로 어떤 순간에 발생할지 모르는 응급 상황을 CCTV의 영상 정보만으로 상황을 항상 모니터링하기에는 인력과 비용의문제점이 발생되고 있다. 본 논문에서는 응급상황을 동적으로 보여주는 CCTV환경에서 감지하기 위해 GMM을 이용한 응급단어와 비응급단어의 검출 및 인식기법을제안하고자 한다. 제안된 방법은 Global GMM 모델에 의해 응급단어와 일반단어를 검출하고 이 모델에 의해 응급단어라 판정된 경우에는 Local GMM 모델에 응급단어 인식을 수행하게 된다. 제안된 방법은 다양한 환경하에서 취득한 응급단어와 일반단어에 대해 적용하여 타당성을 검증하였다.

키워드 : 응급상황, GMM, 음성향상, MFCC, CCTV

Abstract

For the emergency detecting in general CCTV environment of our daily life, the monitoring by only images through CCTV information occurs some problems especially in cost as well as man power. Therefore, in this paper, for detecting emergency state dynamically through CCTV as well as resolving some problems, we propose a detection and recognition method for emergency and non-emergency speech by GMM. The proposed method determine whether input speech is emergency or non-emergency speech by global GMM. If emergeny speech, local GMM is performed to classify the type of emergency speech. The proposed method is tested and verified by emergency and non-emergency speeches in various environmental conditions.

Key Words : Emergency, GMM. Speech enhancement, MFCC, CCTV

1. 서 론

인권침해의 문제에도 불구하고 효율적인 범죄예방 및 범죄수사 등에 적극적으로 활용되고 있는 CCTV 설치에 대한 요구가 지속적으로 증가하고 있다. 그러나 CCTV만을 의존하여 범죄가 발생하는 시점을 인지하여 즉각적으로 대처 하는 데는 몇 가지 문제점이 있다. 가장 큰 문제점으로 언제 발생할지 모르는 위급상황에 대처하기 위하여 관리요원 또는 담당자가 항상 화면을 관찰하여야 하나 눈의 피로감 또는 담당 인원의 부족으로 인하여 상시 화면을 감시할 수 없다는 점이다. 물론 CCTV가 사후 사건에 대하여 범죄자를 색출하는데 중요한 역할을 담당하고 있으나 발생 사건에 대한 동적인 대처의 기능으로는 한계가 있다.

이러한 문제점을 해결하기 위하여 CCTV에서 전송되는 영상정보를 이용하여 응급상황을 자동으로 검출하는 연구가 활발히 진행되고 있는데, 이러한 영상을 분석한 연구에서의 한계점은 다음과 같다[1]. 첫째, CCTV 카메라의 영상 인식이 가지고 있는 많은 기술적 문제들, 특히 기상 변화, 그림자 등 조명의 변화에 따른 오인식과 같은 문제점이 발생한다는 점이다. 둘째, 어두운 밤이나 혹은 화면상으로 구분이 불가능한 응급 상황 발생 시에 이를 확인하기 어렵다는 점이다. 셋째, CCTV가 설치된 지역이라 하더라도 CCTV 근방에서 발생한 응급상황이라 하더라도 화면에 나타나지 않는 사각지대에서 발생하였다면 이를 즉각적으로 확인할 수 없다는 점이다. 따라서 보다 효과적인 응급 상황 대처를 위해 음성인식 기술을 이용하여 보안성 강화를 고려한 연구들이 병행되어 연구되고 있다.

CCTV의 한계를 극복하기 위한 방법으로서 영상정보뿐만 아니라 음성정보까지 전송할 수 있은 디지털 CCTV에 대한 연구가 활발히 이루어지고 있다. 그러나 일반적으로 실내 환경만 아니라 실외 환경에서 발생할 수 있는 외부환경의 경우 주변에 소음이 생기는 잡음환경에 처해 있으며, 따라서 응급 상황 발생 시에 잡음으로 인하여 제한된 환경에서 음성인식시스템의 성능이 크게 저하되는 문제점이 발생된다. 이러한 문제점은 인식 시스템이 학습된 환경과 실제로 인식 시스템이 구현되는 환경에서의 음성 정보가 가지는 특성의 차이에서 오는 것이다. 마이크의 특성, 주변의 소음, 거리상의 문제 등 다양한 요소들이 인식 성능을 낮추게

접수일자 : 2011년 3월 2일
완료일자 : 2011년 4월 4일
이 논문은 경기도지역협력연구센타 지원사업으로 수행되었음 [2010수원GRRC-B3]

된다. 그 중에 주변의 소음은 자동차 소음, 주위 사람들에 의한 잡음, 거리에서 일상적으로 나오는 잡음 등 다양한 형태로 발생하여, 인식 시스템에서 인식해야 하는 음성에 합쳐져 인식 시스템의 정확성을 떨어뜨리며, 잘못된 인식 결과를 가져오게 하는 문제점을 가지고 있다.

여러 가지 잡음에 대한 음성인식 시스템의 성능저하를 해결하기 위해 음성에 포함된 잡음을 제거하는 음질향상(speech enhancement)과 관련된 연구가 활발히 이루어지고 있다. 잡음처리를 위해 가장 대표적으로 사용되는 스펙트럼 차감법인 경우 음성이 존재하지 않는 구간에서 추정한 잡음을 잡음환경에서 차감하여 잡음을 제거하므로, 추정한 잡음의 형태가 음성인식기에 입력되는 잡음 음성에 포함된 잡음과 상이한 특성을 나타낼 경우에는 효과적인 잡음제거가 불가능하다는 문제점을 지니고 있다[2].

스펙트럼 차감법의 문제점을 해결하기 위해 위너필터링[3], 최소통계모델에 기반을 둔 MS(Minimum Statics) 방법[4-6] 등이 있다. 이러한 방법들 중에서도 음질향상을 위해 널리 사용되는 MS방법은 음성 누설량을 감소시키기 위한 최소점 추적을 위해 긴 구간의 윈도우를 요구한다. 긴 구간 윈도우는 노이즈 레벨이 급격히 변화하는 순간에 추정 능력이 저하되는 문제점이 있다. 이러한 문제점을 해결하기 위하여 데이터 기반의 재귀적 노이즈 추정법에 근거한 비정상성 노이즈 추정기법이 제안되었다[7]. 이 방법에서는 노이즈 파워 추정을 위해서 사용되는 이득함수를 데이터 기반으로 하여 얻어지며 다양한 노이즈에 대하여 효과적인 것으로 나타났다.

본 논문에서는 GMM을 이용한 응급상황에서 잡음이 섞인 음성인식 알고리즘을 제안한다. 외부환경에 의해 추가된 음성외의 잡음을 제거하기 위하여 Erkelens에 의해 제안된 노이즈 향상기법을 적용한다. 응급단어 검출 및 분류는 GMM을 이용하여 구축하였다. 응급단어는 두 단계에 걸쳐 수해된다. 첫 번째 단계에서는 Global GMM 모델에 의해 응급단어와 일반단어를 검출하고 이 모델에 의해 응급단어라 판정된 경우에는 Local GMM 모델에 의해 응급단어 중 어떤 단어에 속하는지 응급단어 인식을 수행하게 된다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안한 응급단어 검출 알고리즘을 설명하고, 3장에서는 실험방법과 실험결과에 대한 분석을 한다. 마지막 4장에서는 결론을 맺는다.

2. GMM을 이용한 응급단어 검출 알고리즘

그림 1에서는 본 논문에서 제안된 응급단어 검출 및 인식 알고리즘의 구성도를 나타냈다. 그림 1에서 보는 바와 같이 입력된 음성신호는 노이즈 제거 기법을 적용하여 음질을 향상시킨다. 그 다음 단계로 입력된 음성신호 중에서 시작점과 끝점 검출을 하는 음성구간 검출이 수행된 후 검출된 음성신호에 대한 고역강조 후 멜 켑스트럼에 기반을 둔 특징추출이 이루어지 진다. 다음 단계에서는 응급단어에 대한 모델 구축이 수행된다. 본 논문에서는 응급단어에 대한 모델을 Global GMM과 Local GMM으로 각각 구축하였다. Global GMM은 응급단어의 검출에 사용된다. 즉, Gloal GMM은 고려하고 있는 모든 응급단어의 특징벡터을 이용하여 구축하였으며, 이 모델은 응급단어의 인식이 아닌 일반단어와 응급단어의 분류에만 사용된다. Global GMM에 의해 응급단어로 검출된 경우 Local GMM에 의해 응급단어가 어떤 단어에 속하는지 응급단어 인식이 수행된다. 이와 같이 2단계 구조를 갖는 응급단어 시스템의 주된 잇점은 Global GMM에 의해 일반단어와 응급단어만을 분류함으로 빠른 인식속도가 가능하며 이는 응급상황 발생시에 효과적인 대처가 가능하도록 한다는 점이다.

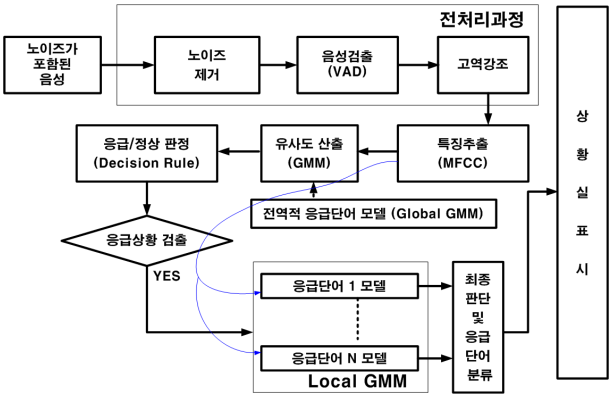


그림 1. 제안된 응급단어 검출 및 인식 알고리즘
Fig. 1 Proposed detection and recognition of emergency speech

2.1 노이즈 제거

본 논문에서는 Erkelens에 의해 데이터 기반의 재귀적 노이즈 추정법에 근거한 비정상적인 노이즈의 추정과 이를 이용한 음질향상 기법을 적용한다[7]. 적용한 방법에 대하여 간략히 서술하면 다음과 같다.

잡음이 섞인 신호모델은 식(1) 과 같다고 고려하자.

$$X(k,m) = S(k,m) + N(k,m) \tag{1}$$

여기서, $X(k,m)$는 잡음이 섞인 신호, $S(k,m)$은 깨끗한 음성, $N(k,m)$은 잡음신호를 각각 나타낸다. 이러한 신호들은 잡음 음성으로부터 신호 프레임 m에서 주파수 인덱스 k번째에서 얻어진 단구간 DFT 계산을 표현한 복소수 값을 갖는 랜덤변수이다. 신호 $S(k,m)$과 $N(k,m)$은 두 신호에 대해서 뿐만 아니라 시간과 주파수에 대해서 통계적으로 독립적이라 가정한다. 노이즈 진폭 $R = |X|$, 음성 스펙트럴 진폭 $A = |S|$ 그리고 노이즈 진폭 $D = |N|$ 이라 하자. 노이즈 DFT 계수들 N은 분산 λ_D를 갖는 복소수 가우시안 분포를 따른다고 가정한다. D^2을 (순간) 노이즈 파워라고 부르고 그의 기대값은 λ_D이다. 또한 음성 스펙트럴 분산 λ_S은 음성 파워 A^2의 기댓값이다.

사전 신호대잡음비(prior SNR) ξ와 사후 신호대잡음비(posterio SNR) ζ은 다음과 같이 각각 정의한다.

$$\xi(k,m) = \frac{\lambda_S(k,m)}{\lambda_D(k,m)}, \quad \zeta(k,m) = \frac{\lambda_S(k,m)}{\lambda_D(k,m)} \tag{2}$$

음성 진폭 A를 추정하기 위하여 식 (3)에서 보는 바와 같이 노이즈 진폭 R에 스펙트럴 이득함수를 곱하는 것이다. 일반적으로, 음성진폭인 A는 다음 식에 의해 추정된다.

$$\hat{A} = G_A(\xi,\zeta)R \tag{3}$$

식 (3)에서 보는 바와 같이 스펙트럴 이득함수 $\hat{A} = G_A(\xi,\zeta)$는 ξ, ζ값을 구한 후 데이터 기반으로 최소평균오차가 최소화되도록 데이터 기반으로 구한다. 식 (3)에서

사전 신호대잡음비 ξ의 추정치 $\hat{\xi}_{NT}(k,m)$을 식 (4)를 이용하여 구하며 α_{NT}는 평활화 변수이다.

$$\widehat{\xi_{NT}}(k,m)=\max\left[\alpha_{NT}\frac{R^2(k,m-1)}{\widehat{\lambda_D}(k,m)}+(1-\alpha_{NT})\frac{R^2(k,m)}{\widehat{\lambda_D}(k,m)},\xi_{\min}\right] \quad (4)$$

2.2 음성검출(VAD : Voice Activity Detection)

음성인식 시스템의 성능은 입력 신호 중에서 음성신호 구간을 얼마나 정확하게 검출하느냐에 크게 좌우된다. 본 논문에서는 음성의 시작점과 끝점을 검출하기 위하여 단구간 에너지와 영교차율을 이용하였다[8].

n번째 구간의 에너지 $E(n)$은 다음과 같다.

$$E(n)=\sum_{i=0}^{p-1}|x(n+i)|, n=0,1,\cdots,k \quad (5)$$

여기서, $x(n)$은 n번째 프레임의 첫 번째 음성 샘플을 의미하고, p는 프레임의 샘플수, k는 음성 프레임의 수를 각각 나타낸다. 영교차율은 프레임내의 신호파형이 영점축과 교차하는 횟수를 의미한다.

식 (5)에서 보는 바와 같이 단구간 에너지는 음성신호를 계산하기 위해서는 프레임 단위의 연산을 수행하게 된다. 음성신호는 10~30ms 정도의 짧은 시간동안에는 그 특성이 비교적 균일하다고 볼 수 있으므로 단구간 에너지 및 영교차율을 구하는 구간을 20ms로 정하였다. 따라서, 음성신호를 16㎑로 샘플링하였을 경우 한 구간의 샘플수는 320개가 된다. 그림 2에서는 단구간 에너지와 영교차율을 이용한 음성신호검출 방법 및 결과를 나타냈다.

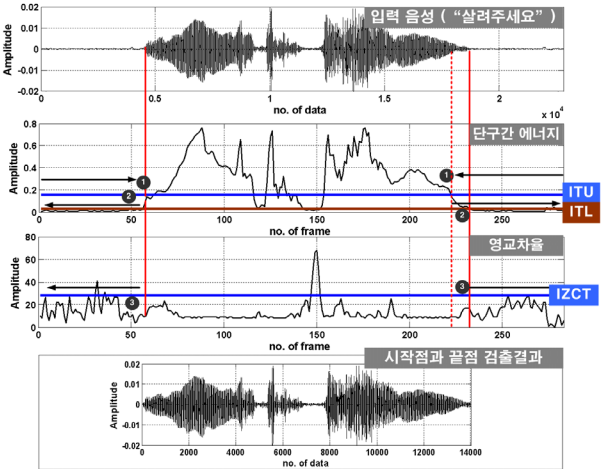


그림 2. 음성구간 검출 방법 및 결과
Fig. 2 A method and result of voice activity detection

음성신호 검출과정을 간략히 설명하면 다음과 같다.

[1단계] 입력된 음성신호에 대해 정방향으로 단구간 에너지 값을 계산한 후, 계산된 단구간 에너지값이 미리 설정된 ITU값을 처음으로 넘는 점을 잠정적인 시작점이라 간주한다.

[2단계] [단계 1]에서 선택한 프레임을 기준으로 역방향으로 미리 설정된 ITU보다 큰 값을 갖는 프레임을 시작점이라 간주한다. 그러나 ITU보다 큰 값을 갖는 프레임이 존재하지 않을 경우 [단계 1]에서 구한 프레임을 시작점이라 간주한다.

[3단계] [단계 1] 또는 [단계 2]에 의해 선택된 프레임을 기준으로 역방향으로 영교차율을 구한다. 계산된 영교차율이 미리 설정된 IZCT 값을 초과하는 프레임이 연속적으로 5회 이상 존재할 경우 이 점을 시작점이라 간주하고 존재하지 않을 경우 [단계 1] 또는 [단계 2]에서 결정된 프레임을 시작점이라 결정한다.

시작점과 끝점을 결정하기 위해서는 ITU, ITL과 IZCT 값을 미리 설정해야 한다. ITL 값은 음성신호의 처음 5 프레임의 평균값을 설정하고 ITU값은 ITL값의 4배로 설정한다. IZCT값은 맨처음 5개의 묵음구간동안에 영교차율의 평균 IZC, 표준편차 σ_{IZC}을 이용하여 $IZCT=IZC+2\sigma_{IZC}$에 의해 결정한다. 그리고 음성의 끝점을 검출하기 위해서는 음성의 끝점을 기준으로 한다는 점을 제외하면 시작점 검출 방법과 동일하다.

2.3 특징추출

사람의 귀가 주파수 변화에 반응하게 되는 양상이 선형적이지 않고 로그스케일과 비슷한 멜(Mel) 스케일을 따르는 청각적 특성을 반영한 켑스트럼 계수 추출 방법이다. 멜 스케일에 따르면 낮은 주파수에서는 작은 변화에도 민감하게 반응하지만, 높은 주파수로 갈수록 민감도가 작아지므로 특징 추출시에 주파수 분석 빈도를 이와 같은 특성에 맞추는 방식이다. 처리 과정은 그림 3과 같다.

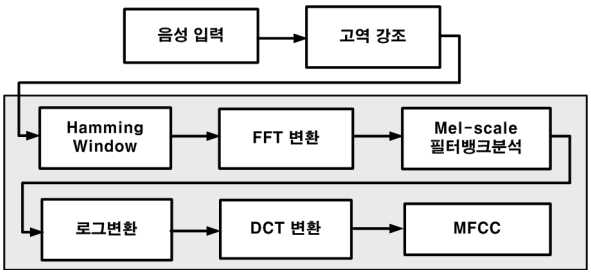


그림 3. MFCC 과정
Fig. 3. The process of MFCC

① 분석구간의 음성 신호에 푸리에(Fourier) 변환을 취하여 스펙트럼을 구한다.
② Mel 스케일에 맞춘 삼각 필터뱅크를 대응시켜 각 밴드에서의 크기의 합을 취한다.
③ 필터뱅크 출력값에 로그를 취한다.
④ 로그를 취한 필터 뱅크 값에 이산 코사인 변환(DCT, Discrete Cosine Transform)을 하여 최종 MFCC를 구한다.

2.4 GMM을 이용한 응급단어 모델 구축

가우시안 혼합모델(Gaussian mixture model)을 이용한 모델 구축과정을 그림 4에 나타냈다. 그림 4에서 보는 바와 같이 음성신호에 대한 특징벡터를 추출한 후 추출된 특징벡터들을 이용하여 GMM의 모델을 구축하게 된다.

가우시안혼합모델 식 (6)과 같이 음성신호를 M개의 각 성분분포들의 선형결합으로 표현된다[9][10].

$$p(\vec{x}\mid\lambda)=\sum_{i=1}^{M}p_i b_i(\vec{x}) \quad (6)$$

식 (6)에서 $\vec{x}$는 음성의 특징벡터, p_i는 혼합가중치 또는 사전확률 ($\sum_i w_i = 1$)이며, $b_i(\vec{x_i})$는 식 (7)과 같이 평균백터들과 공분산행렬인 (μ, Σ_i)에 의해 계산된다.

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} exp\left[-\frac{1}{2}(\vec{x}-\mu_j)^T \Sigma_i^{-1}(\vec{x}-\mu_j)\right] \quad (7)$$

따라서, 가우시안 분포를 표현하기 위해서는 평균벡터들과 공분산행렬, 그리고 사전행렬이 필요하다. 이들 세가지 파라미터의 집합이 응급단어의 가우시안 혼합분포를 표현할 수 있는 모델이 되며 이 집합을 GMM이라고 하고 식 (8)와 같이 표현된다.

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \quad (8)$$

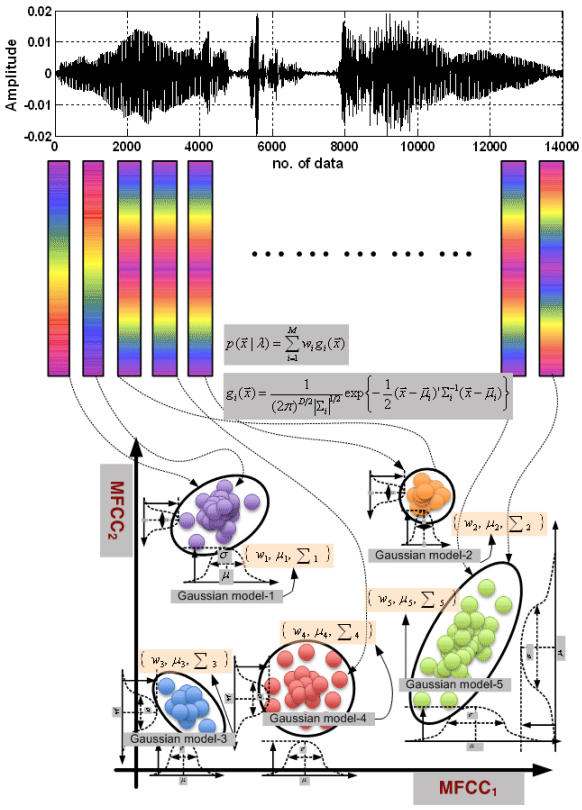


그림 4. 응급단어의 GMM 학습과정
Fig. 4 Training process of GMM for emergency speech

GMM의 세가지 파라미터들은 임의로 초기값을 선택한 후 Expectation 단계와 Maximization 단계로 구성된 EM 알고리즘에 의해 파라미터의 값들이 수렴할 때까지 반복 수행하면서 파라미터의 값을 ML(Maximum Likelihood) 함수가 최대화 될 때까지 추정한다. 이때 대부분의 반복 알고리즘과 마찬가지로 파라미터의 수렴성은 보장되지만 수렴이 전역적 최대값에 대한 보장은 할 수 없으며 단지 초기 시작점에 의존하는 지역적 최대값으로 수렴함으로 알려져 있다.

본 논문에서는 응급단어와 비응급단어 검출을 위한 Global GMM과 응급단어 인식을 위한 Local GMM을 각각 나누어서 구축하였다. Global GMM은 모델 구축에 사용될 모든 훈련용 응급단어에 대한 특징을 추출한 후 추출된 모든 특징벡터를 이용하여 GMM 모델의 파라미터를 추정하였다. 응급단어와 비응급단어의 검출은 입력음성에 대한 특징벡터를 추출한 후, 추출된 특징벡터에 대한 GMM 확률값을 각각 구한 후 구해진 확률값의 로그 평균값을 산출하고, 산출된 로그 평균값을 이용한 결정법칙에 의해 응급단어와 비응급단어의 검출을 수행한다. Local GMM은 모델 구축에 사용될 훈련용 음성데이터를 응급단어별로 분류하여 특징을 추출한다. 추출된 특징값을 이용하여 응급단어별로 독립적으로 GMM 모델 파라미터를 추정한다. 응급단어 인식은 입력음성에 대한 특징을 추출한 후 추출된 특징벡터를 응급단어별로 구축된 GMM 모델에 적용하여 로그 평균값을 산출한 후 가장 높은 확률값을 갖는 모델을 선정하여 응급단어 인식이 수행된다.

3. 실험결과 및 분석

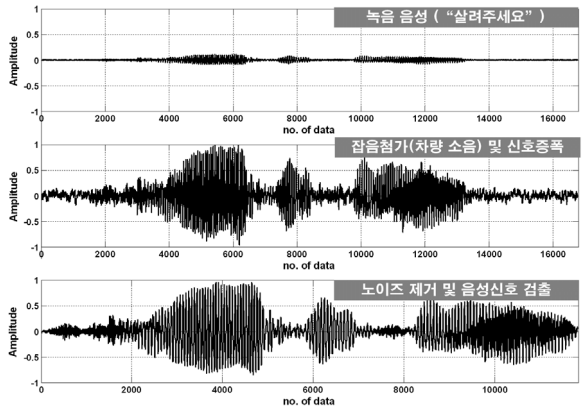
제안된 알고리즘의 성능을 평가하기 위해 세 종류의 응급단어을 녹음하였다. 응급단어 중에서 “살려주세요”는 마이크 앞에서 작은 목소리로 애절한 감정상태에서 녹음하였으며, “도와주세요”와 “불이야”는 마이크에서 5[m] 떨어진 지점에서 큰 목소리로 긴급한 상황을 고려하여 녹음하였다. 녹음에 사용된 마이크는 Infranonic 사의 UFO를 이용하여 16kHz/16 bit로 녹음하였다.

실험에 사용된 응급단어 음성 데이터는 한 가정을 고려하여 40대 남자와 40대 여자, 20대 대학생과 중등 여학생 1명, 초등여학생 1명으로 총 5명으로부터 취득하였다. 녹음 횟수는 각각의 응급단어당 크기와 감정을 달리하여 8회 녹음하였으며 따라서 총 120(응급단어 3 ×5명×8회)개의 응급단어를 구축하였다. 이 중에서 60개의 응급단어는 모델 구축을 위한 학습용으로 사용하였고 나머지 60개의 응급단어는 제안모델의 평가를 위한 검증용으로 사용하였다. 제안모델의 성능평가를 위한 일반단어는 SiTEC DB 중에서 500명의 화자로 구성된 4,178 음성파일을 이용하였다[11].

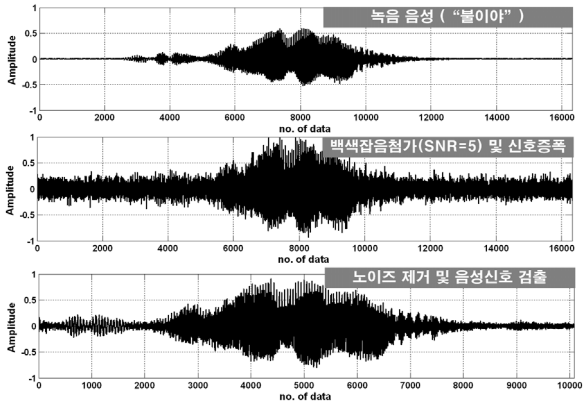
또한, 주변 잡음에 의한 제안 알고리즘의 성능을 평가하기 위해서 차량소음, 오토바이 소음과 백색잡음을 고려하였다. 여기서 차량 소음과 오토바이 소음은 도로에서 5m 떨어진 지점에서 취득하였다. 백색잡음은 신호대잡음비를 25, 15, 5[dB]로 변경하면서 각각 성능을 분석하였다. 그림 5에서는 응급단어 “살려주세요”와 “불이야”에 대한 음성파형과 잡음이 추가된 파형, 그리고 음질향상 후의 파형을 각각 나타냈다. 그림 5(a)에서 보는 바와 같이 “살레주세요”의 응급파형은 진폭이 매우 적기 때문에 차량소음을 첨가한 후 -1과 1사이로 신호를 증폭하였다. 그림 5(b)에서는 응급단어 “불이야”에 대한 파형을 나타냈다. 입력된 음성파형은 -1과 1사이로 신호증폭을 한 후 백색잡음(SNR=5)을 추가하였다. 그림 5에서 보는 바와 같이 노이즈 향상기법을 적용한 결과 차량소음뿐만 아니라 백색잡음에 대해서도 음질이 크게 개선되었음을 확인할 수 있다.

GMM 모델 구축을 위한 실험과정은 다음과 같다. 잡음이 없는 음급단어에 대하여 -1과 1 사이로 정규화한 후 preemphasis 계수 0.96으로 전처리한 후 20ms의 해밍 윈도우를 10ms 간격으로 오버랩하여 구간단위 분석하였으며, 각 구간에서 1차의 에너지와 12차의 멜 켑스트럼을 구하여 총 13차의 특징벡터를 이용하여 Global GMM과 Local GMM을 구축하였다. 검증데이터에 대한 제안된 모델의 평가는 모델구축과정과 동일한 과정을 거친다. 다만 노이즈에

대한 평가를 위해서 입력음성에 노이즈를 첨가한 음성에 대하여 특징벡터의 추출이 이루어진다.



(a) 응급단어("살려주세요")



(b)응급단어("불이야")

그림 5. 응급단어에 대한 음질향상 전과 후의 파형
Fig. 5 Waveforms before and after speech enhancement for two emergency speeches

그림 6에서는 노이즈가 없는 상태에서 응급단어와 비응급단어의 Global GMM 출력값을 나타냈다. 그림 6에서 보는 바와 같이 응급단어의 98.3%는 -22보다 큰 출력값을 나타냈고, 비응급단어는 100% 모두 -30보다 작은 출력값을 나타냈다. 이러한 값들을 기준으로 하여 본 논문에서는 그림 6에서 보는 바와 같이 응급상태, 준응급상태, 비응급상태 등으로 세 구간으로 구분하였다. 여기서 비응급상태는 응급단어와 비응급단어가 존재할 확률이 높은 구간으로서, 입력음성값이 준응급상황에 존재한다면 관리자가 입력음성값을 직접 들음으로서 응급과 비응급을 판단하는 구간이라 가정한다. 노이즈가 없는 상태에서 이러한 준응급상태에 존재하는 응급단어는 1.3%로 나타났다.

표 1에서는 음질향상기법 적용 전과 후의 응급단어와 비응급단어의 검출결과를 나타냈다. 성능지표로서 오거율과 오인식률 사용하였다. 여기서 오거부율은 응급단이지만 응급단어로 판단하지 않은 경우를 의미하여, 오인식률은 비응급단어임에도 불구하고 응급단어로 판단한 것을 의미한다.

음질향상기법을 적용전에는 오인식률은 0.0%이지만 오거부율이 노이즈가 존재할 경우 증가한 것으로 나타났다. 특히, 오토바이 소음인 경우 오거부율이 6.7%로 나타났고 특히, 판정보류영역인 준응급상태에 속한 경우가 51.7%로 나탔다. 또한 SNR이 5인 백색잡음을 첨가한 경우 오거부율이 11.7%, 판정보류 영역이 65%로 나타나 노이즈에 의해서 성능이 현저히 저하됨을 확인할 수 있다. 그러나 음질향상기법을 적용한 결과 본 논문에서 고려하는 모든 소음에 대해서 오거부율과 오인식률이 0.0%로 나타났으며 판정보류에 속하는 준응급상태에 속한 경우도 음질향상기법을 적용하기 전과 비교하면 성능이 현저히 향상됨을 확인 할 수 있다.

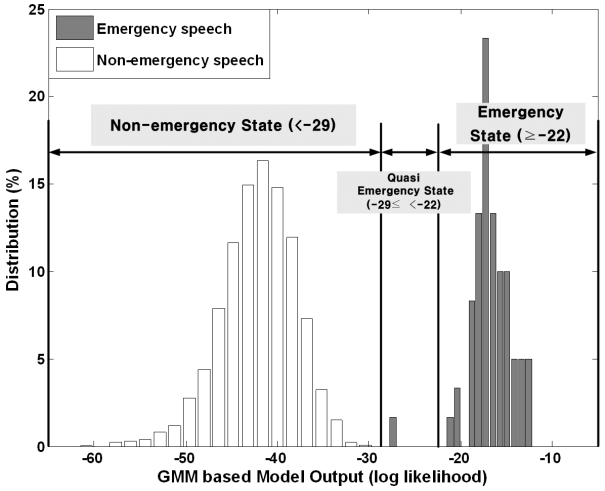


그림 6. Gloabl GMM의 출력값 (노이즈 무)
Fig. 6 GMM based model output

표 1. 응급단어와 비응급단어의 검출 결과
Table 1. Detection results between emergence and non-emergency speeches

(a) Results before speech enhancement method

실험 조건		오거부율	오인식률	판정 보류
노이즈 무		0.0%	0.0%	응급단어:1.7% 일반단어:0.0%
차량 소음		5.0%	0.0%	응급단어:13.3% 일반단어:0.01%
오토바이 소음		6.7%	0.0%	응급단어:51.7% 일반단어:0.0%
백색잡음	SNR:25	0.0%	0.0%	응급단어:8.3% 일반단어:1.1%
	SNR:15	3.3%	0.0%	응급단어:41.7% 일반단어:0.2%
	SNR:5	11.7%	0.0%	응급단어:65.0% 일반단어:0.0%

(b) Results after speech enhancement method

실험 조건		오거부율	오인식률	판정 보류
노이즈 무		0.0%	0.0%	응급단어:1.7% 일반단어:0.0%
차량 소음		0.0%	0.0%	응급단어:1.7% 일반단어:0.04%
오토바이 소음		0.0%	0.0%	응급단어:1.7% 일반단어:0.9%
백색잡음	SNR:25	0.0%	0.0%	응급단어:1.7% 일반단어:0.04%
	SNR:15	0.0%	0.0%	응급단어:1.7% 일반단어:0.9%
	SNR:5	0.0%	0.0%	응급단어:6.7% 일반단어:1.1%

표 2에서는 Global GMM에 의해 응급단어로 판단된 음성파일에 대하여 음성향상기법 적용 후 그 다음 단계인 Local GMM에 의해 응급단어를 인식한 결과를 나타냈다. 표 2에서 A단어는 “살려주세요”, B단어는 “도와주세요”, C단어는 “불이야”를 의미한다. 표 2에서 Local GMM에 응급단어 인식에 사용된 단어는 총 음성입력의 수인 4238단어(비응급단어 4178, 응급단어 60) 중에 응급단어 60개만 해당됨으로 Global GMM에 의해 1.42%만이 선택됨으로 모든 입력음성에 대해 응급단어 인식을 수행하는 것과 비교하여 처리속도가 우수함을 알 수 있다.

응급단어 인식결과를 나타낸 표 2에서 보는 바와 같이 노이즈가 존재하지 않는 경우 인식률을 86.7%로 나타났다. 특히 B단어(“도와주세요”)의 인식률이 다른 단어에 비하여 인식률이 낮은 것으로 나타났다. 이는 A단어(“살려주세요”)와 B단어의 뒷부분 음성이 비슷하여 B단어의 일부가 A단어로 인식되었기 때문인 것으로 분석되었다. 차량소음과 오토바이 소음, 그리고 백색잡음이 SNR이 25일 때 까지 인식률의 큰 저하를 보이지 않았으나 SNR이 15로 백색잡음의 크기가 클수록 인식률이 성능이 크게 저하됨을 알 수 있다. 특히 SNR이 5일 때 인식률은 60.0%로 나타나 노이즈가 존재하지 않는 인식률과 비교하여 26.7% 낮아짐을 알 수 있다.

표 2. 응급단어 인식결과
Table 2. The recognition result for emergency speeches

실험 조건		응급단어 인식			총 인식률
		A단어	B단어	C단어	
노이즈 무		95.0%	80.0%	85.0%	86.7%
차량 소음		95.0%	80.0%	85.0%	86.7%
오토바이 소음		100%	70.0%	80.0%	83.3%
백색 잡음	SNR:25	95.0%	80.0%	90.0%	88.3%
	SNR:15	60.0%	80.0%	65.0%	68.3%
	SNR:5	60.0%	65.0%	55.0%	60.0%

5. 결 론

GMM을 이용한 응급상황에서 응급단어와 비응급단어의 검출 및 응급단어 인식 방법을 제안하였다. 제안된 방법은 Global GMM 모델에 의해 응급단어와 일반단어를 검출하고 이 모델에 의해 응급단어라 판정된 경우에는 Local GMM 모델에 응급단어 인식을 수행하게 된다. 제안방법의 성능평가 결과 Global GMM 모델에서 응급단어와 비응급단어를 포함한 검증음성에서 1.42%만이 응급단어로 선택됨으로 모든 입력음성에 대해 응급단어 인식을 수행하는 것과 비교하여 처리속도가 우수함을 알 수 있다. 또한 응급단어에 대한 응급단어 인식결과 잡음의 매우 큰 경우를 제외하고는 80% 이상의 인식률을 나타내 제안방법의 적용 가능성을 검증하였다. 추후 연구에서는 응급단어의 인식률을 향상시킬 수 있는 방법과 다양한 응급단어에 대한 제안방법의 타당성을 하고자 한다.

참 고 문 헌

[1] 유장희, 문기영, 조현숙, “지능형 영상보안 기술현황 및 동향,” *전자통신동향분석*, 제23권, 4호, pp. 80-89, 2008.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans., ASSP*, vol. 37, no. 2, pp. 113-120, 1979.

[3] Doclo, S., Rong Dong, Klasen, T.J., Wouters, J., Haykin, S., Moonen, M., "Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids," *Applications of Signal Processing to Audio and Acoustics,* vol. 16, no. 16, pp 70-73, 2005.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," in *Proc. ICASSP*. vol. ASSP-32, no. 6, pp. 1109-1121, 1984.

[5] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EISOPCO*. pp. 1182-1185, 1994.

[6] R. Marin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans, Speech Audio Process*, vol. 9, no. 5, pp. 504-512, 2001.

[7] J. S. Erkelens and Richard Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1112-1123. 2008.

[8] Rabiner and Sambur, "An algorithm for determining the endpoints of isolated utterances," *The bell system technical journal*, vol. 54, no. 2, pp. 297-315, 1975,

[9] Ethem Alpayd, "Soft vector quantization and the EM algorithm," *Neural Networks*, vol. 11, pp. 467-477, Issue 3, 1998.

[10] P. Dhanalakshmi., S. Palanivel, V. Ramalingam, “Classification of audio signals using AANN and GMM,” *Applied soft computing*, vol. 11, No. 1, pp. 716-723,2011.

[11] http://www.sitec.or.kr

저 자 소 개

조영임(Cho Young Im)
1988 :고려대학교 컴퓨터학과 학사
1990 :고려대학교 컴퓨터학과 석사
1994 :고려대학교 컴퓨터학과 박사
1995 :삼성전자 선임연구원
2000 :Univ. of Massachusetts, post-doc.
현재 :수원대학교 컴퓨터학과 교수

관심분야 : 인공지능, 뉴로퍼지시스템, 에이전트시스템, 음성인식, 유비쿼터스 시스템

이대종(Lee Dae Jong)
제20권 3호(210년 6월호) 참조