특집

음성 언어를 이용한 인간-로봇 인터페이스

유인철·조영규·이협우·육동석 (고려대학교)

I. 서 론

로봇 기술의 빠른 발전으로 가사를 돕는 지능형 서비스 로봇의 출현도 멀지 않은 미래에 가능할 것이라 여겨지고 있다. 이러한 지능형 로봇은 가사를 돕거나 거동이 불편한 환자의 일상생활을 돕는데 크게 기여할 수 있을 것이다. 기존의 로봇 기술은 주로 하드웨어 제어 측면에 초점을 맞추어 발달하여 왔으나, 최근에는 인간과 로봇 간의 자연스럽고 편리한 인터페이스 수단에 대한 연구 개발의 필요성이 높아지고 있다. 음성 언어는 인간의 가장 기본적인 의사소통 수단으로써, 일반인도 별도의 학습 없이 사용할 수 있으면서, 다양한 정보를 양방향으로 주고받을 수 있는 편리한 인터페이스 수단이다.

음성에는 "무엇을 말하고 있는지" (발화 내용) 뿐만 아니라, "누가 말하고 있는지" (화자 신원), "어디서 말하고 있는지" (화자 위치) 등의 다양한 부가 정보를 포함하고 있다. 휴머노이드 로봇이 주어진 음성 신호로부터 이와 같은 정보들을 추출하여 식별한 후, 그 상황에 맞는 상호작용을 수행할 경우 마치 실제 사람과 대화하는 듯이 친밀하고 편리하게 로봇과 상호작용할 수 있을 것이다[1]. 본 논문에서는 이러한 자연스러운 음성 기반 인간-로봇 인터페이스를 위한 기술 및 관련 연구를 소개하고, 이러한 기술이 어떻게 자연스러운 인터페이스에 적용될 수 있는지를 살펴본다.

본 논문의 구성은 다음과 같다. 제 II 장에서 다수의 마이크로폰을 이용하여 소리가 발생한 위치를 추적하는 음원 위치 추적 (sound source localization: SSL) 기술, 제 III 장에서 주어진 소리가 사람의 목소리인지를 판별하는 음성 구간 검출 (voice activity detection: VAD) 기술, 제 IV 장에서 말한 사람이 누구인지를 판별하는 화자 인식 (speaker recognition) 기술, 제 V 장에서 말한 내용이 무엇인지를 인식하는 음성 인식 (speech recognition) 기술 등을 소개하고, 마지막으로 제 VI 장에서 효과적이면서 자연스러운 음성 기반 로봇 인터페이스를 위하여 각 기술이 융합되어 적용될 수 있는 방안을 제시한다.

II. 음원 위치 추적

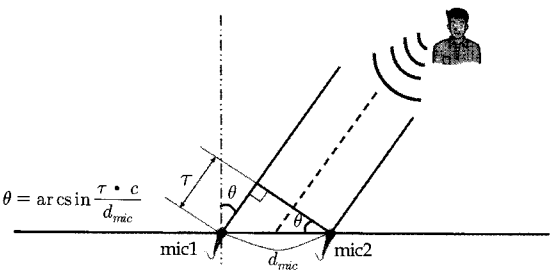
음원 위치 추적은 마이크로폰을 이용하여 소

리가 발생한 위치를 찾는 기술이다. 음원 위치 추적을 인간-로봇 상호작용에 적용할 경우, 로봇이 말한 사람의 위치를 식별하여 자연스럽게 눈을 맞출 수 있다. 사람 간의 대화에 있어 시선의 일치가 주는 영향을 감안할 때, 로봇의 음원 위치 추적은 지금 말하고 있는 내용을 로봇이 듣고 있다는 시각적 피드백 이상의 친밀감을 제공하는 데 기여할 수 있을 것이다.

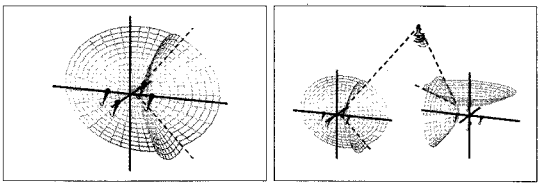
음원 위치 추적에 관한 기존 연구로는 음성 신호의 도착 시간 차이 (time difference of arrival: TDOA)에 기반한 방법[2]과 조향된 빔형성기 (steered beamformer)를 이용한 방법[3]이 있다.

1. Generalized Cross Correlation

Generalized cross correlation (GCC) 기반 방법은 두 마이크로폰에 음성 신호가 도달하는 시간 차이를 이용하여 소리가 발생한 위치를 추정하는 방법이다. 예를 들어 <그림 1>과 같이 소리가 발생한 경우, 음원과 마이크로폰 간의 거리 차이로 인하여 2번 마이크로폰에 먼저 소리가 도착하고, τ만큼의 시간 후에 1번 마이크로폰에 같은 소리가 도달하게 된다.



<그림 1> 2차원 공간에서 소리의 도착 시간 차이와 음원의 방향



(a) (b)

<그림 2> 3차원 공간에서 도착 시간 차이를 이용한 음원 위치 추적

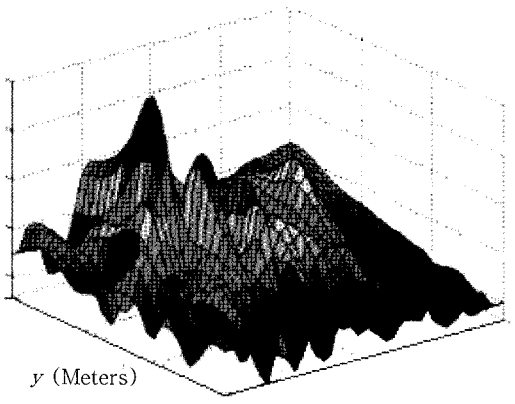
이 때 두 마이크로폰에 도착한 신호 간의 시간 차이 τ를 추정할 수 있다면, 두 마이크로폰 간의 거리 d_{mic}, 시간 차이 τ, 소리의 속도 c, 등을 이용하여 삼각 함수 공식에 의해서 두 마이크로폰의 중심과 음원 간의 각도 θ를 구할 수 있다. Generalized cross correlation은 두 마이크로폰에 입력된 소리 신호를 이용하여 신호 간의 시간 차이 τ를 구하는 방법이다. 즉, 두 마이크로폰에 도착한 신호의 cross correlation을 계산하여 그 값이 최대가 되는 τ 값을 추정한다.

두 마이크로폰을 이용하여 측정한 도착 시간 차이를 τ라고 할 때, 3차원 공간상에서 시간 차이 τ를 만족하는 모든 점들은 원뿔로 근사화될 수 있는 쌍곡면의 한쪽 면을 이루게 된다. 추가적으로 동일 평면상에 마이크로폰을 2 개 더 이용하면 <그림 2> (a)와 같이 2 개의 쌍곡면이 얻어지고, 음원은 두 쌍곡면이 만나는 곡선상에 위치하게 된다. <그림 2> (b)와 같이 8 개의 마이크로폰을 이용할 경우 곡선의 교차점을 구하여 3차원 공간상에서 소리가 발생한 위치를 추정할 수 있다.

2. Steered Response Power

이 방법은 대상이 되는 공간을 잘게 분할한 후, 모든 분할 단위에 대하여 소리의 발생 가능성

을 계산하여 그 중 가장 큰 가능성을 지니는 분할 단위에 소리가 존재한다고 추정하는 방법이다. 조향된 빔형성기를 이용하여 모든 마이크로폰에 들어온 신호들의 delay-and-sum[4]으로 steered response power (SRP)를 계산한다. 3차원 공간상에서 마이크로폰의 좌표와, 각 분할 단위의 좌표가 주어지면 각 마이크로폰에 얼마만큼의 도착 지연 시간이 발생할지를 예측할 수 있다. 이 정보를 이용하여 각 마이크로폰에 입력된 신호를 이용하여 해당 분할 단위의 좌표에 따른 도착 시간 차이로 delay-and-sum을 계산할 경우 소리가 발생한 분할 단위에서 계산한 delay-and-sum이 가장 큰 값을 가지게 된다. <그림 3>은 주어진 2차원 평면을 일정 간격으로 분할한 후, 모든 분할 단위에 대하여 delay-and-sum으로 SRP를 계산하여 에너지 크기를 표시한 예제이다. 3차원 공간에 대해서도 마찬가지로 공간을 분할하여 delay-and-sum을 계산한 후 최대가 되는 지점을 찾음으로써 음원 위치 탐색이 가능하다.



〈그림 3〉 SRP를 이용하여 측정된 공간 에너지 그래프

Phase transform을 이용한 SRP (SRP-PHAT)[3] 방법은 다른 음원 위치 추적 알고리즘에 비해 반향 등이 존재하는 실제 잡음 환경에서도 안정적인 위치 추적 성능을 나타내는 것으로 알려져 있다. 하지만 대상이 되는 공간을 잘게 분할한 뒤, 매 분할 단위에 대하여 계산을 수행해야 하므로 계산량이 매우 많은 단점이 있다. 이에 알고리즘의 정확도를 유지하면서도 계산량을 줄여서 실시간 음원 위치 추적이 가능하도록 하는 연구가 활발히 진행되고 있다.

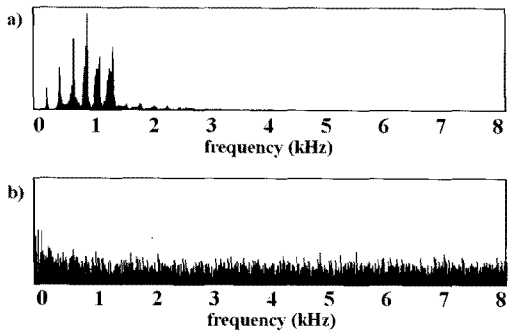
III. 음성 구간 검출

음성 구간 검출은 주어진 소리가 사람의 음성인지 기타 잡음인지를 구분하는 기술이다. 음성이 존재하는 구간을 잡음으로 잘못 거절하거나, 음성이 아닌 잡음을 음성으로 잘못 판단할 경우 이후 수행되는 음성 인식, 화자 인식 등의 성능이 크게 저하될 수 있다. 자동으로 음성 구간 검출을 적용하지 않고 버튼을 눌렀다 떼는 등 수동으로 음성이 존재하는 구간을 지정할 수도 있으나, 이러한 경우 원거리에서 별도의 기기 없이 목소리만으로 편리하게 인터페이스가 가능하다는 음성 인터페이스 고유의 장점을 잃게 되는 문제가 있다.

주어진 소리가 사람의 목소리인지 아닌지를 구분하는 기존 연구로는 크게 음성을 구분하는데 적합한 특징을 추출하여 이용하는 연구와 패턴 인식 기법을 이용하여 음성과 잡음 신호의 학습 데이터로 학습한 음성 모델과 잡음 모델을 사용하여 음성 여부를 판별하는 연구가 수행되고 있다.

1. 음성의 특징을 이용한 음성 구간 검출

음성 구간 검출을 위해 음성의 고유한 특징을 추출하는 연구 중 가장 간단한 형태는 에너지(energy)에 기반한 방법이 있다[5]. 에너지에 기반한 방법은 음성이 주변 소리에 비해 크게 발생한다는 점을 가정하여 일정 문턱값 (threshold) 이상으로 크게 발생한 소리를 무조건 음성으로 간주하는 방법이다. 이는 주변 잡음이 거의 없는 조용한 환경에서는 어느 정도 동작하지만 잡음이 조금이라도 크게 발생하면 음성으로 잘못 판단하는 문제가 생길 수 있다. 이와 같이 음성과 무관한 신호를 음성으로 잘못 판단하는 문제를 줄이기 위하여 좀 더 음성에 밀접한 관련성을 갖는 특징을 찾는 연구가 진행되어왔다. 이러한 특징의 예로는 영교차율 (zero crossing rate: ZCR), spectral entropy[6] 등이 있다. Spectral entropy는 음성의 에너지가 각 주파수에 고르게 분포하는 것이 아니라 특정 주파수에 편중된다는 점을 이용하여 스펙트럼의 에너지 편중도가 큰 소리를 음성으로 판단하는 방법이다. 예를 들어 <그림 4>를 보면 잡음에 비하여 음성의 스펙트럼이 일부 영역에 치우친 에너지 분포를 나타냄을 알 수 있다.



〈그림 4〉 음성과 잡음의 스펙트럼; a) 모음 "a"의 스펙트럼, b) 백색 잡음 (white noise)의 스펙트럼

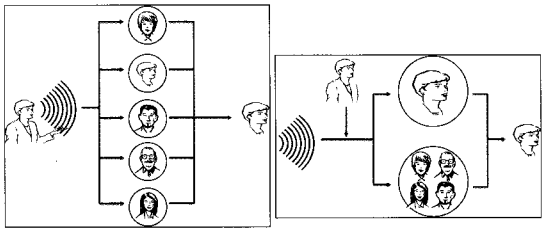
2. 패턴 인식에 기반한 음성 구간 검출

음성과 잡음의 학습 데이터를 이용하여 음성 모델과 잡음 모델을 학습시킨 후 패턴 인식 기법을 적용하여 음성과 잡음을 분류하는 방법에는 Gaussian mixture model (GMM) 또는 hidden Markov mode (HMM)을 이용한 기법이 가장 널리 사용되고 있다[7]. GMM을 이용한 음성 구간 검출 기법에는 스펙트럼뿐만 아니라 음성을 구분하는데 적합한 다른 특징, 예를 들어 low-variance spectrum과 같은 특징을 이용하여 GMM을 학습시키고 인식하는 방법 등이 가능하다[8].

Ⅳ. 화자 인식

화자 인식은 생체 인식 기술의 하나로 주어진 음성 신호가 누구의 것인지를 판별하는 기술이다. 화자 인식은 목소리로 사용자를 식별하여 사용자 인식이 필요한 다양한 환경, 예를 들어 로봇이 지금 말한 사람이 누구인지 파악하는 것에 이용될 수 있을 뿐만 아니라, 주어진 음성 신호에서 특정 화자의 발화 부분만을 추출하여 오디오 인덱스를 구성하는 분야에도 적용 가능하다.

화자 인식은 목적에 따라 구분하면 크게 두 가지로 구분할 수 있다. 화자 식별 (speaker identification)은 사전에 인식하고자 하는 화자들의 목소리를 등록한 후, 입력된 신호 중 가장 유사한 화자를 선택하는 문제이다. 화자 검증 (speaker verification)은 검증하고자 하는 화자가 고정되어있고, 입력된 목소리가 실제로 검



(a) (b)

〈그림 5〉 화자 인식의 분류; a) 화자 식별, b) 화자 검증

증하고자 하는 화자로부터 발화되었는지 여부를 판별하는 문제이다. <그림 5>는 화자 식별과 화자 검증에 대한 개념도이다.

또한, 인식에 사용하는 문맥에 따라서 사전에 정의된 단어 또는 문장으로만 인식할 경우 문맥 종속 (text dependent) 화자 인식, 임의의 단어 또는 문장으로 인식이 가능한 경우 문맥 독립 (text independent) 화자 인식으로 구분할 수 있다[9]. 미국 National Institute of Standards and Technology (NIST)에서는 주기적으로 speaker recognition evaluation (SRE)이라는 공개 대회를 주관하여 문맥 독립 적인 화자 검증 분야의 발전을 주도하고 있다[10].

화자 인식에서 중요한 연구 주제는 크게 두 가지로 나눌 수 있다. 이는 주어진 학습 데이터에 대하여 효과적으로 화자별로 내재된 고유한 패턴을 학습하여 결정을 내리는 인식 알고리즘에 관한 연구와 서로 다른 환경에서 수집된 데이터 간의 상이한 음향학적 특성을 줄이는 연구이다.

1. 화자 특징 모델링

화자의 음향학적 특징을 모델링하는 가장 대표적인 방법은 GMM을 이용한 방법이다[11]. GMM에 기반한 방법은 많은 양의 학습 데이터를 사용할 경우 안정적인 성능을 보이지만, 매우 적은 양의 학습 데이터를 사용하는 경우, 학습 데이터에 비하여 학습해야하는 모델 파라미터의 수가 많아 올바른 학습이 어렵기 때문에 성능이 저하될 수 있다. 이러한 문제를 해결하기 위하여 maximum a posterior (MAP) 또는 eigenvoice에 기반한 화자 적응 기법을 이용하여 학습해야 하는 파라미터의 수를 줄여서 적은 양의 데이터로 빠르게 화자의 특징을 모델링하는 방법이 고안되었다[11~14].

GMM에 기반하지 않은 다른 방법으로는 패턴 인식 분야에서 널리 사용되는 support vector machine (SVM)을 이용한 방법이 있다[15]. SVM의 경우 이미지와 같은 정지된 단일 데이터에 대하여 구분하는 알고리즘이므로 음성 신호와 같이 연속성을 띠는 입력에 대해서는 알고리즘을 수정할 필요가 있다. 대표적인 방법으로는 generalized linear discriminant sequence kernel (GLDS)과 같은 새로운 SVM 커널 함수를 이용하는 방법과 GMM과 결합하는 방법이 있다[16,17].

2. 환경 조건에 강인한 화자 인식

화자 인식은 세션 (session)에 따라 화자 인식 성능이 크게 영향을 받는 문제가 있다. 세션은 학습 데이터 및 인식 데이터에 있어서 영향을 끼치는 음향학적 특성을 의미하는 것으로 사용하는 마이크로폰의 종류, 신호 전달 경로, 주변 잡음, 녹음 장소의 특징 등 녹음 결과에 영향을 줄 수 있는 모든 요소를 의미한다[18]. 이러한 특징이 일치하지 않을 경우, 예를 들어 학습 데이터 녹음에 사용된 마이크로폰의 종류와 인식 데이터 녹음에 사용된 마이크로폰의 종류가 다를 경

우 인식 성능이 심각하게 저하되는 문제가 있다. 이러한 문제를 해결하기 위하여 음성 신호를 분석할 때 세션의 영향을 적게 받는 특징을 추출하는 특징 추출 기반 기법과 세션의 특징을 별도로 모델링하여 상쇄하는 방법이 연구되어 왔다. 특징 추출에 기반한 방법으로는 feature warping, cepstral mean subtraction (CMS), feature mapping, relative spectral filtering (RASTA filtering) 등이 고안되었다. 이들 방법은 신호 분석 단계에서 적용되므로 이후의 학습 모델은 어떠한 것을 사용하더라도 상관없이 적용 가능하며 동시에 여러 방법을 적용하는 것도 가능하다[18]. 세션의 특징을 별도로 모델링하여 상쇄하는 방법으로는 eigenvoice 개념을 응용한 eigenchannel 방법이 있다[19]. 또한, 최종 단계에서 계산된 유사도에서 세션의 영향으로 왜곡된 값을 제거하기 위해 유사도의 분포를 보정하는 score normalization 기법이 연구되었다[20].

V. 음성 인식

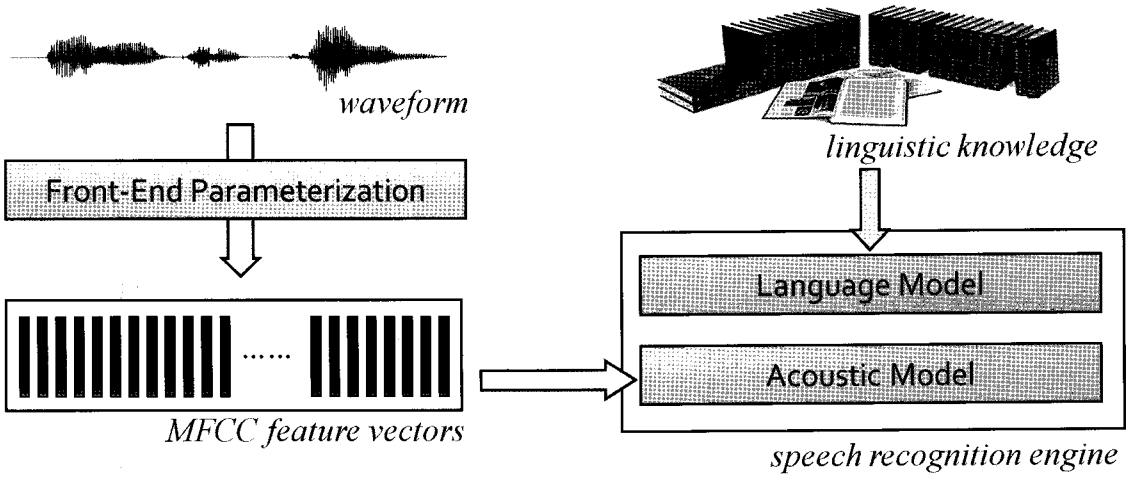
음성 인식은 주어진 음성 신호로부터 발화한 내용이 무엇이었는지를 추정하는 기술이다. 음성 인식을 통하여 목소리만으로도 로봇에게 다양한 명령을 내릴 수 있게 되면 마치 사람에게 명령하듯이 인터페이스가 가능하게 될 것이다. 음성 인식은 인식 대상 및 목적에 따라 다양하게 분류될 수 있다. 우선 인식 어휘 개수에 따라 숫자음 인식과 같이 적은 수의 어휘 (small vocabulary)에서부터 수천~수만 단어의 대용량 어휘 (large vocabulary) 인식 시스템으로 분류될 수 있다. 인식 대상이 되는 발화의 형태에 따라 단어별로 끊어지게 발음한 것을 인식하는 고립 단어 인식 (isolated word recognition) 시스템에서부터 평상시 대화체처럼 끊어짐 없이 자연스럽게 연결하여 발화한 음성을 인식하는 연속 음성 인식 (continuous speech recognition) 시스템으로 분류 가능할 수 있다[21].

대 어휘 연속 음성 인식 (large-vocabulary continuous speech recognition: LVCSR)은 대용량의 어휘를 대상으로 하고 단어 사이에 끊어짐이 없이 자연스럽게 발화하는 음성을 인식하는 것을 목표로 한다[22]. 이 장에서는 이러한 LVCSR의 구성 요소를 살펴보고, LVCSR의 연구 주제에 대하여 알아본다.

1. 대 어휘 연속 음성 인식 시스템의 구조

기존의 전형적인 LVCSR 시스템은 HMM에 기반한 확률 모델을 이용한다. HMM에 기반한 인식 시스템의 목표는 주어진 음성 신호 X에 대하여 가장 가능성이 높은 단어열 W를 찾는 것이라고 할 수 있다. 즉 주어진 X에 대하여 확률 $P(W|X)$를 최대화할 수 있는 단어열 W를 찾는 문제로 정의할 수 있다. 하지만, 가능한 음성 신호 X의 종류가 무한하여 확률 값을 추정하기 어려우므로 Bayes' rule을 이용하여 $P(W|X) = P(W) \cdot P(X|W)/P(X)$로 나누어 분석한다. 주어진 음성 신호 X에 대하여 최대 값을 찾는 것이므로 $P(X)$의 값은 일정하여 계산에서 제외하여도 무방하다. $P(W)$는 해당 단어열이 발생할 확률로 언어 모델 (language model)을 이용하여 구할 수 있고, $P(X|W)$는 주어진 단어열이 입력된 음성 신호처럼 발화될 확률로서 음향 모델 (acoustic model)을 이용하여 구할 수 있다.

<그림 6>은 대 어휘 연속 음성 인식 시스템의 일반적인 구조를 나타낸 것이다. 입력된 음성 신



〈그림 6〉 대 어휘 연속 음성 인식 시스템의 구조

호에 대하여 먼저 front-end parameterization 단계에서 음성 인식에 유용한 특징 벡터를 추출하게 된다. 음성 인식에 유용한 특징은 말하는 내용이 바뀔 경우 특징 값이 민감하게 바뀌면서 말하는 내용과 무관한 특성 (잡음, 화자의 발화 성향 등)에 대해서는 값이 잘 바뀌지 않는 것이 유용한 특징이라 할 수 있다. 현재 가장 널리 사용되고 있는 특징 벡터는 mel-frequency cepstral coefficient (MFCC) 벡터로, 인간의 발성 특징과 청각 구조의 특징을 모델링하는 특징 벡터이다[23].

LVCSR 시스템에서는 인식해야 하는 대상 어휘가 매우 많기 때문에 어휘별로 모델을 구축하고 학습하는 것은 현실적으로 어렵다. 따라서 주로 발음의 최소 단위인 음소 (phoneme) 별로 모델을 구축하고 학습하며 단어 모델은 해당 단어를 구성하는 음소들의 모델을 연결하여 생성하는 방식을 취하게 된다. 음소를 모델링하는데 있어서는 음소당 모델을 하나씩 생성하는 모노폰 (monophone) 모델링부터 인접 음소에 따른 발화의 변화를 모델링하기 위하여 음소당 다수의 문맥 종속 모델을 생성하는 트라이폰 (triphone) 모델링 방법 등이 존재한다[24].

언어 모델의 경우 유사한 발음을 지니는 여러 단어열 후보 중 문법적으로 올바른 단어열을 선택하고, 불필요한 후보들을 줄이는데 이용된다. LVCSR에서는 통계학적 언어 모델인 n-gram 모델이 가장 널리 사용되고 있다. 이것은 각 단어의 발생 확률을 추정할 때, 선행하는 $n-1$ 개의 단어열이 무엇인가에 따라 확률을 다르게 예측하는 모델이다.

인식 단계에서는, 입력된 음성의 내용을 추정하기 위하여 음향 모델과 언어 모델을 모두 불러들인 후, 무수히 많은 가능한 단어열 후보들 중 가장 확률이 큰 단어열을 추정할 필요가 있다. 이러한 추정 과정을 decoding이라 부른다. 그러나 음향 모델과 언어 모델의 규모가 크고 가능한 후보열이 거의 무한하기 때문에 검색의 효율성을 높이기 위하여 beam pruning, two-pass search 등 다양한 기법이 적용된다. Beam pruning은 매 단계마다 가능한 후보열 중 가능성이 낮은 후보들을 제거하여 탐색 수를 줄이는 방법이며,

two-pass search는 상대적으로 간단한 모델 (monophone-unigram 모델 등)을 이용하여 일차적으로 유력한 후보들을 추려낸 후, 이 결과에 대해서만 복잡한 모델 (triphone-trigram 모델 등)을 적용하여 후보를 얻는 방법이다.

2. 음성 인식 연구 방향

화자 인식의 경우와 마찬가지로 음성 인식 역시 학습 데이터와 인식 데이터의 환경이 일치할수록 높은 성능을 나타낸다. 하지만 환경이나 화자가 바뀔 때마다 전체 시스템을 처음부터 다시 학습시키는 것은 필요 학습 데이터가 너무 많아 현실적이지 않다. 따라서 많은 수의 기존 발화 데이터로 음성 인식 시스템을 학습시킨 후 새로운 환경이나 화자의 데이터를 소량 받아들여 적응 (adaptation) 기법으로 변화된 환경 및 화자의 특징을 반영하여 빠른 학습을 수행한다 [25,12,13].

LVCSR 시스템과 같은 대형 시스템은 모델을 저장하기 위한 대용량의 저장소와, 많은 연산량을 요구하는데, 기존의 인식 성능을 유지하면서도 소형 장치에서 동작할 수 있도록 효율적인 알고리즘 및 저장 방식이 연구되고 있다. 예를 들어 유사한 특성을 지니는 음향 모델들을 클러스터링 (clustering) 알고리즘을 적용하여 모델을 공유하도록 함으로써 적은 데이터로 학습이 가능할 뿐 아니라 음향 모델 자체의 크기도 줄일 수 있다[26].

VI. 결 론

음성에는 해당 음성의 발화 내용 뿐 아니라, 말하는 사람의 위치 정보와 신원 등 다양한 정보가 담겨있다. 음성으로부터 이러한 정보를 추출하여 활용함으로써 기존의 인터페이스보다도 훨씬 자연스러우면서 편리한 로봇 제어가 가능해질 것이다. 예를 들어 기존의 인터페이스로는 사용자가 바뀔 때 마다 사용자가 직접 그 사실을 로봇에 알려주어야 할 것이나, 화자 인식 기술을 이용하여 말하는 사람이 바뀌었음을 자동으로 추정하는 것이 가능할 것이다.

음원 위치 탐색, 음성 구간 검출, 화자 인식 기술, 음성 인식 기술은 각각으로도 자연스러운 인간 로봇 인터페이스에 적용될 수 있지만, 이들이 결합되어 더욱 더 효과적인 인터페이스로 기능할 수 있다. 음원 위치 탐색 기술과 음성 구간 검출 기술을 결합할 경우 잡음이 존재하는 실제 환경에서도 음성이 존재하는 위치를 추정하여 해당 위치에서 발생하는 음성 신호만을 빔형성 (beamforming) 기법으로 강화하여 원하는 음성 신호를 강화하고 다른 방향에서 발생하는 원치 않는 잡음 신호를 약화시킬 수 있다. 또한 음성 구간 검출 알고리즘에서 음성이 없는 묵음 구간이라 판별한 구간의 특징 분석을 통하여 잡음 보상 (noise compensation) 알고리즘을 적용하여 화자 인식이나 음성 인식 등의 성능을 증대시킬 수 있다. 또한, 일반 가정집 같이 복수의 사용자가 존재할 수 있는 환경에서는 화자 인식 기법을 이용하여 각 사용자별 발화 특징을 별도로 모델링한 후, 현재 발화하는 사용자가 누구인지 식별하여 해당 화자 전용의 모델을 적용하여 인식 성능을 높일 수 있다. 이와 같이 화자 인식과 음성 인식을 결합할 경우 화자별로 전용의 학습 데이터를 이용한 화자 적응 및 환경 적응이 가능하여 실제 환경에서도 높은 인식 성능을 기대할 수 있다. 또한, 입력되는 소리에서 목소리뿐만 아니라 그 이외에 어떤 종류의 소리가 (예를 들면 전

화벨, 초인종 소리 등) 언제 발생하였는지 자동으로 인식하는 auditory scene analysis 또는 audio diarization[27]은 로봇의 청각 시스템에 필요한 기술이다.

VII. 감사의 글

이 논문은 2006년도 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-311-D00822). 또한 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구 결과로 수행되었음 (IITA-2009-C1090-0902-0007).

참고문헌

[1] I. Yoo and D. Yook, "Speech processing techniques for efficient human-robot interface," *한국언어학회 2008년도 겨울 학술대회 자료집*, pp.1-4, 2008.

[2] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-32, No.2, pp.280-285, 1984.

[3] J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Array*, Ph.D. Thesis, Brown University, 2000.

[4] D. Johnson and D. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1993.

[5] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-29, No.4, pp.777-785, 1981.

[6] B. Wu and K. Wang, "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments," *IEEE Transactions on Speech and Audio Processing*, Vol.13, No.5, pp.762-775, 2005.

[7] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol.6, No.1, pp.1-3, 1999.

[8] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.14, No.2, pp.412-424, 2006.

[9] J. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, Vol.85, No.9, pp.1437-1462, 1997.

[10] http://www.itl.nist.gov/iad/mig/tests/sre

[11] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, Vol.10, pp.19-41, 2000.

[12] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, Vol.2, No.2, pp.291-298,

1994.

[13] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, Vol.8, No.6, pp.695-707, 2000.

[14] B. Mak, J. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, Vol.13, No.5, pp.984-992, 2005.

[15] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1995.

[16] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, Vol.20, No.2-3, pp.210-229, 2006.

[17] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, Vol.13, No.5, pp.308-311, 2006.

[18] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Transactions on Audio, Speech and Language Processing*, Vol.15, No.7, pp.1979-1986, 2007.

[19] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, Vol.15, No.4, pp.1435-1447, 2007.

[20] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, Vol.10, No.1-3, pp.42-54, 2000.

[21] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[22] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, Vol.13, No.5, pp.45-57, 1996.

[23] S. Davis and P. Mermelstein, "Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, No.4, pp.357-366, 1980.

[24] J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. Thesis, Cambridge University, 1995.

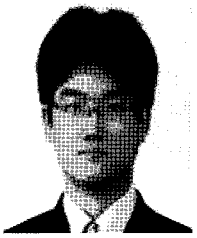
[25] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol.9, No.2, pp.171-185, 1995.

[26] E. Bocchieri and B. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Transactions on Speech and Audio Processing*, Vol.9, No.3, pp.264-275, 2001.

[27] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems,"

IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.5, pp.1557-1565, 2006.

저자소개



유 인 철

2006년 3월 고려대학교 컴퓨터학과 학사
2008년 3월 고려대학교 컴퓨터학과 석사
현재 고려대학교 컴퓨터학과 박사과정

주관심 분야 : Robust speech recognition / speaker recognition

조 영 규

2001년 3월 한성대학교 컴퓨터학과
2003년 9월 고려대학교 전산학 석사
2005년 9월 고려대학교 전산학 박사

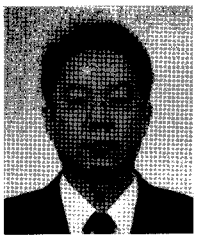
주관심 분야 : 음원 위치 추적

저자소개

이 협 우

2007년 2월 고려대학교 컴퓨터학과 학사
2009년 2월 고려대학교 컴퓨터전파통신공학부 석사
2009년 고려대학교 모바일솔루션학과 박사과정

주관심 분야 : speaker recognition, speech recognition, signal processing



육 동 석

1990년 8월 고려대학교 컴퓨터학과 학사
1993년 2월 고려대학교 컴퓨터학과 석사
1999년 10월 Rutgers University, Ph.D.
1999년 9월~2001년 2월 IBM T.J. Watson Research Center, Senior Software Engineer
2001년 3월~현재 고려대학교 컴퓨터학과 교수

주관심 분야 : Machine Learning, Speech Processing