지능형 반응공간을 위한 연속적 화자인식에 관한 연구

A Study of Continuous Speaker Recognition for Intelligent Responsive Space

권순일, Soonil Kwon*
*한국과학기술 연구원

요약 Human Computer Interaction 기술을 구체화 시키기 위한 Intelligent Responsive Space 의 개발에 있어서 음성정보는 여러 가지로 유용하게 활용될 수 있다. 음성신호로부터 얻을 수 있는 다양한 정보 중의 하나가 화자인식을 이용한 화자의 신원식별이다. 이 논문에서는 화자인식 인식이 어려운 환경에서도 음성 신호로부터 추출한 특성벡터들을 선택적으로 사용함으로써 화자인식 성능을 높일 수 있는 새로운 방법을 제안하려 한다. 화자를 인식하는데 있어서 인식오류를 발생시킬 가능성이 높은 특성벡터들을 인식을 위한 판단의 대상에서 배제시킴으로써 성능을 향상시킬 수 있다. 실험결과에 의하면 0.25 초에서 2 초 길이의 짧은 음성만으로도 기존의 방법에 비해 20 에서 51%의 상대적 성능 향상을 보였다. 새롭게 제안된 방법을 적용하면 기존의 방법들에 비해 세밀하면서도 정확하게 연속적으로 화자들을 인식할 수 있게 된다.

핵심어: Human Computer Interaction (HCI), Intelligent Responsive Space (IRS), Speaker Recognition, Gaussian Mixture Model (GMM)

1. 서론

Human Computer Interaction (HCI) 기술은 점차 많은 도전적이고 참신한 응용분야에 적용되어가고 있다. 그 중에서 최근 주목을 받고 있는 것이 Intelligent Responsive Space (IRS), 즉 지능형 반응공간의 개발이다. 지능형 반응공간은 지능적인 서비스 및 정보, 그리고 자연스럽고 편리한 인터페이스를 제공하는 것인데, 이러한 것들을 위해서는 상호작용을 하는 사람들에 대한 정보를 수집하는 것이 중요하다. 사람의 정보라 함은 행동 (behaviors), 말 (speech), 그리고 신원식별 (identification) 등이 있다. 최근 지능형 반응공간 개발의 일환으로 지능형 회의시스템이 연구 개발되고 있다 [1]. 지능형 회의시스템에 필요한 것들 중 하나가 회의 참가자들에 대한 화자인식 (Speaker Recognition) 이다. 연속적인 화자인식은 회의 중에 참가자들이 말하는 내용을 참가자별로 구분 지어 줄 수 있고, 그 내용을 추후 참가자 별 검색이 가능하도록 만들어 주는 기능을 가능케 해 준다 [2][3].

화자인식에 있어서 해결해야 할 문제들 중 하나가 시간적 제약이다. 회의 중 참가자들 간에 이루어지는 대화에 있어서 연속적인 화자인식을 통해 참가자들의 발언을 구분하여 저장해 줄 필요가 있는데, 짧은 한 단어 정도를 말할 경우 1초 미만의 데이터가 존재하기 때문에 화자인식을 위한 데이터 절대량 부족으로 인식성능의 저하를 초래한다. 하지만 이러한 경우가 빈번하게 발생할 수 있기 때문에, 음성 데이터가 부족한 경우에도 화자인식의 성능저하를 최소화 할 수 있는 새로운 방법을 제시하고자 한다.

음성으로부터 화자를 인식하는데 있어서 성능 저하를 초래하는 원인들은 여러 가지가 있을 수 있지만, 대표적으로 환경잡음, 침묵, 그리고 인식 대상이 되는 화자들의 공통되는 주파수 특성 등이 있다 [4]. 일반적으로 이러한 원인들은 화자인식을 위해 입력되는 음성의 길이가 짧을수록 인식을 위한 판단에 오류를 일으킬 여지가 더 많다. 그래서 화자를 인식할 때 일반적인 음성신호로부터 변환된 특성벡터들 중에서 위의 원인들에 해당되는 특성벡터들을 배제시킴으로써 성능을 향상시킬 수 있다.

이 논문은 다음과 같이 구성되어 있다. 두 번째 장에서는 지능형 반응공간에 대한 소개와 연속적 화자인식에 대해 소개 및 지능형 반응공간에서의 연속적 화자인식의 역할에 대해 기술되어 있고, 세 번째 장에서는 이 논문에서 제안하는 새로운 연속적 화자인식을 위해 새롭게 만들어지는 화자모델에 개념과 방법, 그리고 이를 이용한 화자인식 방법에 대한 설명이 되어 있다. 네 번째 장에는 실험의 방법, 그리고 결과에 대한 분석 및 토의가 서술되어 있고, 마지막 장에서는 이번 논문에 대한 결론이 기술되어 있다.

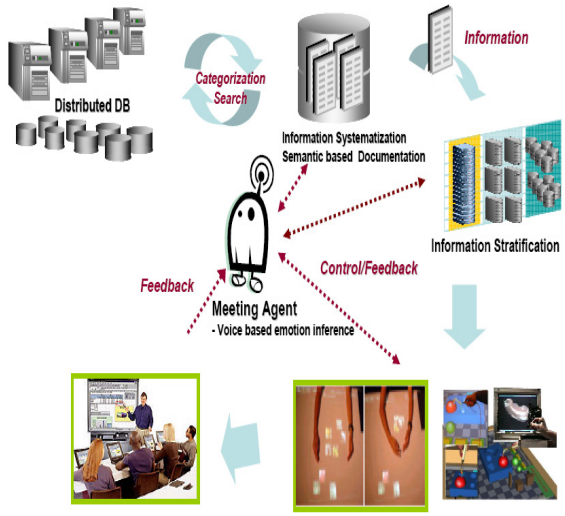

그림 1. 지능형 반응공간의 개관 ([1]로부터 인용됨)

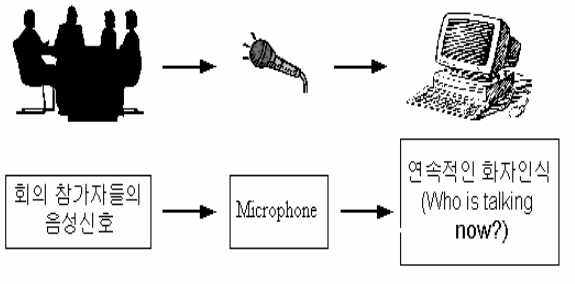

그림 2. 연속적 화자인식의 개관

2. 지능형 반응공간과 연속적인 화자인식

지능형 반응공간은 Human Computer Interaction (HCI) 기술들을 이용하여 실생활에 접목시킨 구체화된 어플리케이션이다. 지능형 반응공간이란 일정한 공간 안에서 사람이 컴퓨터로부터 지능적인 서비스 및 정보를 제공 받을 수 있고, 컴퓨터는 사람의 행동이나 말 등을 인지하고, 주변 상황도 인지하며, 또한 사람과 컴퓨터 간에 자연스럽고 편리한 인터페이스를 제공하는 것이 가능하도록 하는 것이다. 지능형 반응공간을 구축하기 위한 일환으로 개발하고 있는 지능형 회의시스템은 연구과제 진행회의를 대상으로 하여 회의참여자들에게 지능적 정보서비스와 자연스럽고 편하게 인터페이스하며 회의를 할 수 있는 공간을 제공하기 위한 것이다 [1]. 여기에는 개발회의 참여자에게 필요한 서비스 및 정보를 제공할 수 있는 지능형 회의관리 에이전트기술 개발과 회의참여자(Multi-User)의 자연스러운 인터페이스 및 회의자료 공유기술개발, 그리고 지능적인 회의록 작성, 데이터 분류, 저장 및 검색 등을 위한 다양한 연구가 진행되고 있다 [그림 1].

지능형 회의시스템을 위한 기술들을 활용하는데 있어서 화자들에 대한 정보는 유용하게 쓰일 수 있다. 연속적인 화자인식 기술을 이용하여 회의 시 발언 중인 화자가 누구인지를 실시간으로 인지하여 회의관리 에이전트는 발언자 중심의

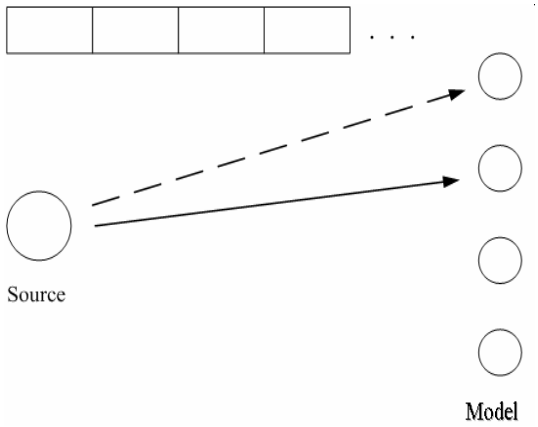

그림 3. 연속적 화자인식 방법의 예

자료를 생성, 정리, 검색 등의 관리할 수 있다. 또한 회의 시 어떤 특정시간에 누가 발언 하였는지를 인지하여 지능적인 회의록 작성을 위해 화자 별로 발언 내용을 분류하여 관리를 할 수 있다.

일반적인 화자인식에서는 사람이나 회사, 도시의 이름과 같은 단어들이나 신용카드 번호, 전화번호 등의 일련의 숫자들을 단편적으로 발성한 음성정보를 이용하여 누가 말한 것인지를 찾아내는 것이다. 반면에 연속적 화자인식은 전화통화 내용이나 방송 뉴스 등의 오디오 데이터에 있는 화자들을 인식하는 것으로, 연속적으로 들어오는 오디오 입력 데이터를 이용하여 연속적으로 화자인식을 행한다 할 수 있다. 그림 2는 회의를 예로 들어 설명하고 있는데, 회의 참가자들의 대화가 마이크로폰을 통해 입력이 되면, 화자인식 시스템에서 각 참가자들의 발언에 대한 시간적 정보를 얻는다. 즉 화자 별로 언제 누가 발언을 했는지가 분류되고, 저장, 검색이 가능해진다.

연속적으로 화자를 인식하기 위해서는 일반적인 화자인식 방법과 같이 미리 인식하려는 화자들의 음성 데이터를 이용하여 화자 별로 통계적 모델을 만든다. 화자의 통계적 모델로는 일반적으로 Gaussian Mixture Model (GMM)을 사용한다 [5]. 화자모델을 만들 때에는 음성 신호를 변환하여 그것으로부터 특성벡터를 추출하고 그것들을 이용해 모델을 만든다. 이렇게 미리 만들어진 화자 모델들을 이용하여 그림 2에 표현되어 있는 예와 같이 화자인식을 하게 되는데, 인식 대상이 되는 화자의 음성 데이터가 들어오면 추출된 특성벡터들과 화자 모델들을 이용하여 Likelihood를 계산하여 비교한 후 확률적으로 가장 유사한 화자모델이 선택되어 그 모델에 대응되는 화자로 인식결과를 얻게 된다 [5][6][7]. 이에 해당하는 수식은 다음과 같다.

$$\hat{M} = \arg\max \Pr(V|M_i), \quad i = 1,\ldots,S. \qquad (1)$$

위의 수식 (1)에서 M은 화자모델이고, V는 일정한 개수의 특성벡터들로 이루어진 특성벡터 세트이며, S는 인식대상이 되는 화자의 총 수이다. 즉, 각 화자모델 별로 특성벡터 세트의 확률 값들 중 최고를 나타내는 화자모델을 찾는 것이다.

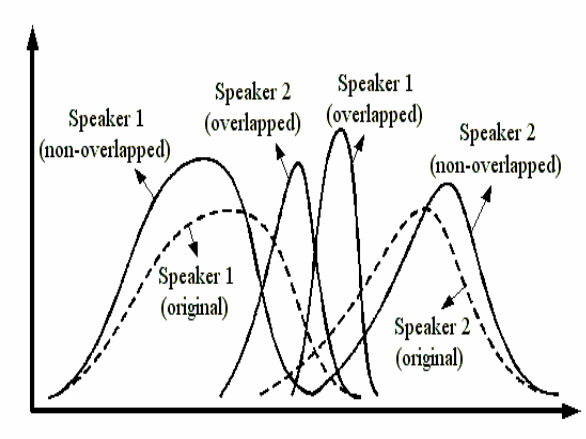

그림 4. 화자모델 분할의 예

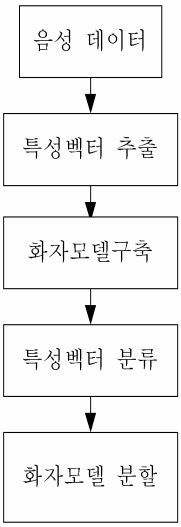

그림 5. 화자모델 분할 방법

화자인식을 하기 위해서는 연속적으로 입력되는 음성 데이트를 일정한 길이로 나누든지 아니면 화자가 바뀌는 시점을 찾아내어 나누게 되는데, 그림 3에서는 전자의 방법을 사용하여 설명하고 있다. 일정한 길이로 나누어진 음성데이터는 각각이 어느 화자로부터 나온 음성인지를 찾아서 정보를 표시하게 된다. 이때 나누어진 음성데이터는 그 길이에 따라 전체적인 화자인식 성능에 많은 영향을 끼치게 된다. 지금까지 연구되어온 결과에 의하면, 음성 데이터의 길이가 짧아질수록 화자인식이 성능이 저하되는데, 그 길이가 대략 2초 이하가 되면 급격한 성능 저하를 가져오게 된다 [5][8]. 하지만, 화자가 빠르게 바뀌는 경우, 그 길이가 길면 나누어진 음성데이터가 복수의 화자를 포함하게 되어 인식성능을 저하시킬 우려가 있다. 그래서 짧은 음성만을 가지고 더 좋은 화자인식 성능을 보여줄 수 있는 방법이 필요한데, 이에 대한 연구가 계속되고 있다.

3. 새로운 연속적인 화자인식 방법

기존의 화자인식에 있어서 성능저하의 주된 원인들 중에

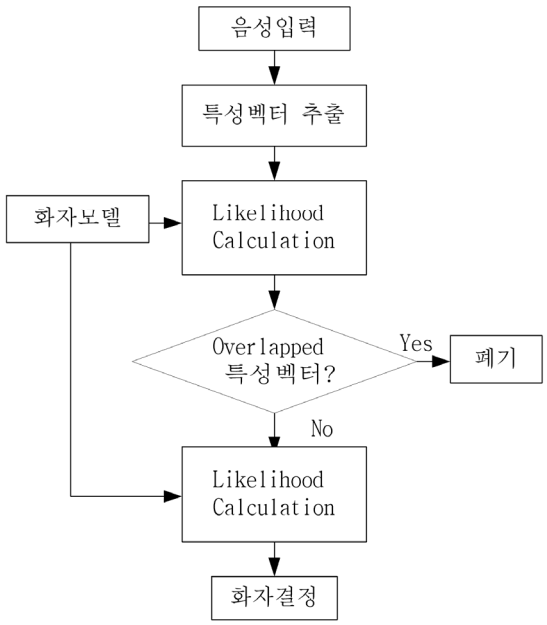

그림 6. 분할된 화자모델을 기반으로 한 연속적 화자인식 방법

하나는 통계적 모델을 사용함으로써 발생될 수 있는 문제인데, 인식 대상인 화자들의 모델들이 서로 겹쳐있기 때문이고, 그 겹친 부분에 해당되는 특성벡터들은 인식오류에 상당한 영향을 끼친다. 화자모델들이 겹치는 이유로는 환경잡음, 침묵, 그리고 인식 대상이 되는 화자들의 유사한 주파수 특성 등 화자들의 음성데이터들이 지니고 있는 공통된 특성들 때문이다.

연속적 화자인식을 할 경우에 화자인식 대상의 음성데이터가 충분하지 못하면, 데이터의 시간적 길이가 짧을수록 화자인식의 성능이 급속히 저하된다. 특히 이러한 현상은 음성데이터의 시간적 길이가 2초 미만일 경우 더 확연히 나타난다. 왜냐하면, 일정한 길이의 음성데이터로부터 추출한 일정한 개수의 특성벡터들이 각각 화자모델 별로 확률 값을 가지게 되고, 이 값들을 모두 포함하는 총 확률 값을 구하여, 최고의 확률을 보이는 모델에 대응되는 화자가 인식의 결과가 된다. 이때 화자모델들이 서로 겹쳐있는 부분과 대응되는 특성벡터들은 원래 대응 되어야 하는 화자모델이 아닌 다른 화자모델에서 더 높은 확률을 보일 수 있는데, 이러한 현상들이 화자인식 오류의 주 원인이 된다. 그런데 이런 특성벡터들은 영향은 특성벡터 세트내의 벡터 수가 작을수록, 즉 음성데이터의 길이가 짧을수록 커지게 된다.

위와 같은 문제점을 해결하기 위해 새로운 화자인식 방법을 제안한다. 이 방법은 인식오류에 영향을 끼치는 특성벡터들을 화자인식을 위한 데이터에서 제외시킴으로써 인식성능 저하를 최소화 할 수 있다는 취지를 가지고 있다. 이를 위해 인식오류 문제의 소지가 있는 특성벡터를 구분해 내는 방법이 필요하다. 이를 위해 다음과 같은 방법으로 화자모델을 분할시킨다. 먼저 기존의 방법과 같이 각 화자마다 GMM 모델을 만든다. 그 다음 만들어 놓은 화자모델들을 각각 두 개의 모델로 분할시킨다. 모델분할을 위해서는 최초에 화자모델을 만들 때 사용되었던 특성벡터들을 다시 사용하는데, 먼저 특성벡터들을 두 가지 카테고리로 분류한다. 즉, 특성벡터

가 인식오류를 일으킬 여지가 있는지 확인한 후에 오류를 일으키지 않는 벡터들과 오류를 일으키는 벡터들을 구분하여 보관한다.

예를 들어, 도시한 그림 4를 보면, 점선으로 도시된 것이 기존의 화자모델들이고, 점선으로 도시된 것이 분할된 화자모델들이다. 두 개의 화자모델이 겹치는 부분에 해당하는 특성벡터들 중에는 Likelihood를 계산하여 비교해 볼 때 Speaker 1에 속하면서도 Speaker 2에 속하는 것으로 인식되는 것들이 있으며, Speaker 2에 속하면서도 Speaker 1에 속하는 것으로 인식되는 것들이 있다. 이렇게 실제와 다르게 인식되는 특성벡터들이 결국 화자인식 오류를 일으키므로 실제 연속적 화자인식을 할 때 이들을 구분해 내기 위해 각 화자모델을 두 개로 분할시킨다. 이를 위해 각 화자마다 모델을 다시 만드는데, 이번에는 처음과 달리 오류를 일으키는 않는 벡터들로 만든 모델 (non-overlapped)과 오류를 일으키는 벡터들로 만든 모델 (overlapped)을 각각 따로 만든다 [그림 5].

화자인식의 오류를 일으키는 특성벡터들을 구분해 내기 위하여 분할된 화자모델들을 이용하여 화자인식을 하게 된다. 입력된 음성데이터로부터 추출한 특성벡터들 중 오류를 일으키는 특성벡터들로 만든 모델(overlapped)에 최고의 유사성(Likelihood)를 나타내는 것들은 판단에서 제외시키고, 오류를 일으키지 않는 특성벡터들로 만든 모델(non-overlapped)을 선택하는 벡터들만을 판단의 대상으로 삼아 화자인식의 판정을 낸다 [그림 6]. 이러한 방법을 이용하면 상대적으로 적은 음성데이터만을 가지고도 향상된 화자인식 성능을 얻을 수 있다.

4. 실험 결과

이 논문에서 제안된 방법의 우수성을 평가해 보기 위해 400명의 화자데이터를 이용한 실험을 실시하였다. 화자데이터는 Speaker Recognition Benchmark NIST Speech (1999) Corpus에서 얻어진 것이고, 400명 중 240명은 여성이고, 나머지 160명은 남성이다. 이 실험에서는 화자들이 회의를 하고 있고, 회의 참석인원이 8명이라는 가정하에 사람들을 8명씩 (여성 4명, 남성 4명) 50개 팀을 만들어 실험을 하였다. 인식대상이 되는 음성데이터의 길이는 025, 0.5, 1, 2초로 실시하였는데, 데이터 길이를 맞추기 위해 연속적인 음성데이터를 인위적으로 분할하여 실험을 하였다. 하지만 각 음성데이터가 단순히 한두 단어의 인명 또는 지명이거나, 일련의 숫자가 아니고, 일반적인 대화에서 얻은 것이기 때문에 전자의 데이터들에 비해 자연스러우며, 화자인식이 어려운 데이터라 할 수 있다.

각 팀마다 일정한 음성데이터 길이(025, 0.5, 1, 2초)에 있어서 화자변화 감지방법을 배제시킨 상태로 8명의 음성데이터를 뒤섞어 8명의 화자를 포함한 하나의 연속적인 음성데이터를 만들었다. 예를 들어, 1초의 음성데이터에 대한 실험이라면, 팀 내의 화자들의 1초 길이의 음성데이터들을 순서 없이 뒤섞은 뒤 정확하게 화자들을 찾아내는지 알아보았다. 오류확률은 인식 오류 화자의 수를 총 화자의 수로 나누어 계산하였다. 이러한 실험 방법을 이용하면 실질적인 연속적 화자인식이라 할 수 없지만, 연속적 대화의 음성데이터에 있어서 길이의 변화에 따라 어떠한 화자인식 오류확률을 보이는

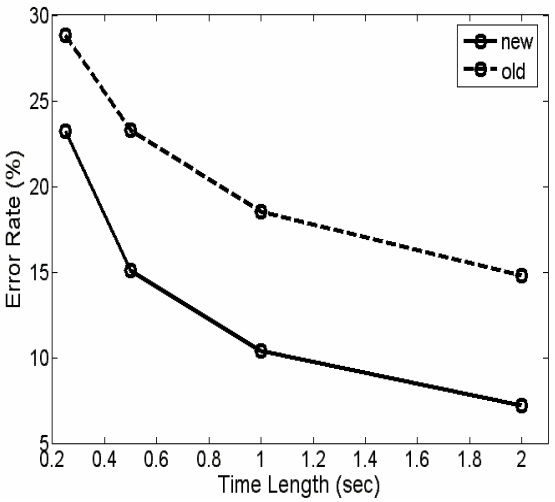

그림 7. 화자인식 오류확률에 대한 실험결과

지 알아보기 위한 차선의 방법이라 할 수 있겠다.

그림 7은 연속적 화자인식의 결과로 오류확률을 보여주고 있다. 새로 제안한 방법이 기존의 방법에 비해 입력음성 길이에 따라 평균적으로 5.6% 에서 8.2% 정도의 절대적 성능향상 (상대적으로 20% 에서 51%)을 보였다. 이 실험 결과에서 특히 주의해서 볼 점은 0.5초의 음성에 대해 새로운 방법을 이용할 경우 얻어지는 연속적 화자인식 성능은 기존방법을 이용할 경우 2초의 음성이 요구된다는 것이다. 즉 일정한 성능을 유지하면서 기존의 방법에 비해 더 세밀하게 연속적 화자인식을 할 수 있다.

5. 결론

지능형 회의 시스템을 개발하는데 있어서 회의 중에 얻을 수 있는 음성신호를 음성인식 시스템이나 화자인식 시스템을 이용하여 언어적인 정보와 화자의 정보로 바꿀 수 있다. 특히 화자정보를 얻기 위한 화자인식 시스템에 있어서 연속적 화자인식을 이용해 추출할 수 있는 화자정보는 회의 시스템을 지능화 시키는데 중요한 역할을 할 수 있다.

연속적 화자인식은 그 대상이 되는 음성데이터의 시간적 길이에 따라 성능이 좌우되는데, 특히 그 길이가 2초 이하일 경우 급격한 성능 저하를 초래한다. 여러 가지 원인이 있을 수 있지만, 이 논문에서는 특정 특성벡터들로 인한 영향에 주목했다. 문제를 일으키는 특성벡터들을 제거하는 방법은 연속적 화자인식의 시간적 제약을 극복함으로써 회의 시스템에 있어서 화자 별 음성 데이터 구분 및 검색 기능을 향상시킬 수 있고, 지능형 반응 공간 개발에 큰 기여를 하리라고 본다.

지능적 회의시스템을 위한 연속적 화자인식의 성능과 유용성을 증대시키기 위해서는 다음과 같은 연구가 계속되고 있다. 첫째, 화자가 바뀌는 시점을 정확하게 찾아낼 수 있는 방법이다. 화자인식의 어떤 시간적 길이에 해당되는 특성벡터들의 한 세트 안에서 화자가 바뀔 경우 화자를 찾는 것이 거의 불가능하기 때문이다. 둘째, 화자인식을 위한 화자모델

이 존재하지 않는 화자들에 대한 대책이 필요하다 [8]. 화자들에 대한 모델을 미리 만들어 놓지 못한 상황에서 그것들을 대체할 모델이 필요하게 된다. 순차적으로 입력된 음성정보를 이용하여 화자모델들을 초기화 시킬 수도 있겠지만, 이럴 경우 화자모델을 만드는데 필요한 데이터의 양이 부족하여 도리어 인식오류를 유발시키는 원인이 될 수 있다. 이러한 도전적인 문제들을 해결함으로써 보다 우수한 성능의 지능적 회의시스템을 구축할 수 있을 것이다.

참고문헌

[1] J.-H. Park, K.-W. Yeom, S. Ha, M.-W. Park, and L. Kim, "An overview of intelligent responsive space in tangible space initiative technology," Proc. Internt. Workshop on the Tangible Space Initiative (3rd), pp. 523-531, 2006.

[2] C. Busso, S. Hernanz, C.-W. Chu, S. Kwon, C. Lee, P.G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: participant and speaker localization and identification," Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1117-1120, 2005.

[3] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel, "Multimodal People ID for a Multimedia Meeting Browser," Proc. 7th ACM Internat. Conf. on Multimedia, Part 1, pp. 159-168, 1999.

[4] J. P. Campbell, "Speaker recognition: A tutorial," Proc. IEEE, 85, pp. 1436-1462, 1997.

[5] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech Audio Process., 3, (1), pp. 334-337, 1995.

[6] T. Hastie, H. R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, New York, 2001.

[7] R. V. Hogg and E. A. Tanis, "Probability and Statistical Inference," Prentice Hall, New Jersey, 2001.

[8] S. Kwon and S. Narayanan, "Unsupervised Speaker Indexing Using Generic Models," IEEE Trans. on Speech and Audio Processing, Vol. 13, Issue 5, Part 2, pp.1004-1013, 2005.