특 집

음성인식기술동향

金 炯 淳
釜山大學校 電子工學科

I. 머리말

언어의 형태를 음성언어(spoken language)와 문자언어(written language)로 분류할 때, 입으로 발음하고 귀로 듣는 음성언어는 언어의 일차적인 형태이다.[1] 문자가 만들어지기 이전에도 인간은 음성으로 의사소통을 해왔으며, 어린이들은 글자를 배우기에 앞서 음성언어를 자유롭게 사용한다. 이와 같이 음성은 인간에게 있어서 가장 익숙한 의사전달 수단이며, 이 음성을 이용하여 각종 기계나 도구들을 쉽게 조작하고자 하는 것 역시 매우 자연스러운 발상이라 하겠다. 최근 컴퓨터 및 신호처리 기술의 급속한 발전은 인간의 음성을 알아듣고(음성인식) 인간에게 음성으로 응답하는(음성합성) 기능을 가지는 시스템의 출현이 현실로 다가왔음을 보여주고 있으며, 부분적으로는 이미 상품화된 제품들이 상당수 등장하기에 이르렀다.

음성인식기술은 키보드와 같은 번거로운 수단을 사용하지 않고도 컴퓨터를 조작할 수 있게 해 주고, 멀리 떨어진 곳에서 전화를 이용하여 저장된 정보를 검색하거나 또는 각종 작업지시를 내리는 등의 일들을 가능케 해준다. 실제로 미국에서는 전화를 이용하여 음성으로 collect call을 신청하고, 원하는 종목의 증권시세를 알아보며, 상대방 이름을 말함으로써 전화연결이 되는 등의 서비스가 진행되고 있으며, 국내에서도 이와 유사한 서비스가 준비 중에 있다. 그러나, 현재까지의 기술수준으로는 공상과학영화에 나오는 것처럼 사람과 유창하게 대화를 나눌 수 있는 음성인식시스템은 개발되지 못하고 있으며, 앞으로도 가까운 장래에 이러한 시스템이 구현되지는 않을 전망이다.

본고에서는 음성인식기술의 개요를 살펴보고, 특히 실용화의 관점에서 음성인식에 사용되는 음성신호처리기술의 동향에 대해 논의하고자 한다. 최근 국내에서도 음성인식기술에 대한 관심이 높아지면서, 본지를 비롯한 몇몇 학술지에서 음성인식기술의 전반적인 동향, 구체적인 응용분야 및 국내외 연구현황에 대한 논문들이 이미 발표된 바 있으므로,[2~5] 본고에서 다루지 않은 내용에 관심있는

독자들은 이들을 참조하기 바란다.

II. 음성신호처리기술 분야

음성인식기술을 살펴보기에 앞서 음성을 대상으로 한 신호처리기술의 응용분야들을 개괄해 보기로 한다. 음성신호처리기술을 응용 분야 및 사용되는 기술의 종류에 따라 분류하면 크게 다음과 같은 분야들로 나눌 수 있으며, 이들 분야는 서로 밀접한 연관을 맺고 있다.

1. 음성부호화(Speech Coding)

음성부호화는 음성의 효율적인 전송 또는 저장을 목적으로 음성신호를 압축하는 기술로서 음성부호화 방식의 성능은 음질과 전송속도(또는 압축률), 그리고 부호화기의 복잡도로 평가된다. 음성부호화 방식은 음성파형 자체를 시간 또는 주파수 영역에서 디지탈부호화하는 파형부호화(Waveform Coding) 방식, 음성발생기관의 모델을 기반으로 추출된 모델의 파라메터들만 전송하는 보코딩(Vocoding) 방식, 그리고 이들 양자의 장점들을 사용하는 혼합부호화(Hybrid Coding) 방식으로 크게 분류할 수 있다. 최근 차세대 디지탈 이동통신과 관련하여 4kbits/s 이하의 음성부호화 방식 연구에 많은 관심이 기울여지고 있으며, 멀티미디어 통신과 관련한 광대역 오디오부호화 방식연구도 활발히 이루어지고 있다.

2. 음성합성(Speech Synthesis)

음성합성은 문자정보를 사람이 청취할 수 있는 음성신호로 변환시키는 기술이다. 음성합성의 가장 간단한 방법은 합성하고자 하는 단어, 구 또는 문장 등을 미리 저장했다가 이들을 조합하여 재생시키는 방법이지만, 이 경우 합성할 수 있는 어휘나 문장에 한계가 있다. 따라서 엄밀한 의미에서의 음성합성은 말소리의 기본단위로부터 음성학, 언어학 및 운율정보를 이용하여 음성을 만들어 냄으로써 어휘나 문장구조 등에 아무런 제한없는 합성방식을 말하며, 이를 text-to-speech 변환기술이라고도 부른다. 지금까지의 연구결과로 합성음의 명료도 면에서는 상당한 진전이 이루어졌지만, 사람이 말하는 것과 같이 자연스러운 음을 만들기 위해서는 앞으로도 해결해야 될 많은 문제점이 남아 있다.

3. 음성인식(Speech Recognition)

음성인식은 음성합성의 역과정, 즉 음성신호로부터 문자정보 또는 그 의미를 파악해내는 speech-to-ext 과정으로 볼 수 있으며, 음성의 청취과정을 다룬다는 면에서 음성의 발생과정을 다루는 음성합성보다 훨씬 난이도가 높은 과제이다. 이 분야는 본 고의 중심주제이므로 이후에 보다 상세히 다루도록 하겠다.

4. 화자인식(Speaker Recognition)

사람에 따라 발음기관의 크기와 모양에 차이가 있고 성장과정에서 습득된 발음 습관도 서로 다르므로, 동일한 문장을 말하더라도 음성신호에는 화자에 대한 정보가 포함되게 된다. 화자인식이란 음성신호를 분석하여 말하는 사람이 누구인지를 식별해내거나 특정인여부를 검증하는 기술을 의미한다. 음성은 지문이나 망막무늬 등 개인성확인의 주요방법으로 사용되고 있는 특징들에 비해 현재로서는 성능면에서 뒤떨어지지만, 사용이 편리하고 특히 전화선 등을 이용하여 원거리에서도 개인성식별이 가능하다는 점에서 계속 관심의 대상이 되고 있다. 화자인식기술은 음성인식에서 화자특성을 정규화하거나 새로운 화자의 특성에 적응화는 기술과 밀접한 연관을 맺고 있다.

5. 음질개선(Speech Enhancement)

음질개선이란 잡음이 섞이거나 채널왜곡이 된 음성으로부터 발음의 명료도와 자연스러움을 개선하는 기술을 말한다. 음질개선기술은 기본적으로 단일 마이크를 이용하는 방법과 복수 개의 마이크에 의한 방법으로 나누어지며, 적응 디지탈 필터를 비롯한 다양한 종류의 신호처리 방법들이 적용되고 있다. 이 기술은 또한 잡음환경에서의 음성인식을 위한 전처리 기술로도 매우 중요한 역할을 한다.

이상에서 언급한 분야들 이외에도 음성신호로부터 어느 언어를 사용하고 있는지를 찾아내는 언어식별(Language Identification) 기술이 다국어 음성처리의 필요성과 더불어 관심의 대상이 되고 있으며,[8] 음성인식, 음성합성 및 자연어처리기술 등이 통합된 음성언어시스템(Spoken Language System)에 관한 연구들이 현재 활발히 진행되고 있다.[9]

III. 음성인식기술 개요

1. 음성인식은 왜 어려운가?

음성신호에는 언어정보, 즉 말하고자 하는 메세지에 해당하는 정보 이외에도 말하는 이가 누구이며 그의 감정상태와 태도, 그리고 주위환경이 어떠한가 하는 정보까지 포함되어 있다. 따라서, 음성신호는 언어정보와 일대일 대응관계를 가지는 것이 아니며, 동일한 언어정보라 할지라도 여러 가지 변화요인들로 인해 무수히 많은 형태의 서로 다른 음성신호로 표현될 수 있다. 음성신호에 영향을 주는 변화요인들을 정리하면 표 1과 같다.[11]

음성신호의 외적인 변화요인들 이외에도 언어정보가 음성신호로 표현되는 과정 자체도 음성인식의 어려움으로 작용한다. 음성언어는 말소리의 기본단위, 즉 음소(phoneme)들의 연결형태로 표현되며, 이들 음소들은 언어마다 차이가 있기는 하지만, 단일언어 내에서는 각각 고유의 발음방법 및 음향학적 특성을 가지는 것으로 간주된다. 그러나 음성은 발음기관의 연속적인 움직임에 의해 표현되므로 한 음소의 음향적인 특성은 인접음소에 의해 영향을 받게 되며, 이를 상호조음효과(coarticulation effect)라 부른다. 이에 따라 각각의 음소들의 음향적인 특성은 매우 다양한 형태로 나타나며, 비록 한 언어가 가지는 음소의 수는 매우 제한되어 있지만 음성인식은 소수의 명료한 음향학적 패턴들을 분류하기만 하면 되는 간단한 문제로 귀착되지 못하게 된다. 물론 음성인식의 근본적인 문제는 아직까지 인간이 어떻게 음성을 인식하는지에 관해 모르는 부분이 너무 많다는 사실에 기인한다.

〈표 1〉 음성신호에 영향을 주는 변화요인들

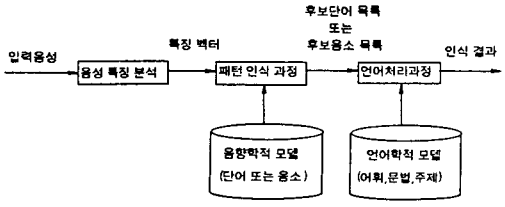
음향학적 변화요인 (Acoustic Variability)	배경잡음, 마이크의 특성 및 위치, 실내반향, 통신채널의 왜곡특성 등
단일화자에 의한 변화요인 (Intraspeaker Variability)	화자의 생리적, 심리적 상태, 발음태도, 발음속도 등
화자차이에 의한 변화요인 (Interspeaker Variability)	발음기관의 크기 및 형태, 방언, 사회적 배경, 발음습관 등

이와 같은 어려움들로 인해 음성인식의 궁극적인 목표, 즉, 잡음이 있는 실제적인 환경에서 임의의 화자가 어휘에 제한없이 자연스럽게 발음한 연속음성을 실시간에 인식 및 이해하는 수준을 만족시키는 시스템은 아직 개발되지 못하고 있으며, 현재까지의 음성인식시스템들은 여러 가지 인위적인 제약조건 하에서 운용됨을 전제로 하고 있다. 음성인식시스템의 성능에 영향을 미치는 기술적인 어려움들은 크게 다음의 다섯 가지 항목으로 나누어 볼 수 있다. 그 중 첫째가 화자 독립성(speaker independence)에 관한 것으로서, 특정인의 음성을 인식하는 것에 비해 연령과 성별, 그리고 방언이 다른 여러 사람들의 음성을 인식하는 것이 훨씬 어렵다. 두번째는 발음방법 및 속도와 관계되는 것으로서, 각 단어와 단어를 또박또박 띄어 발음하는 것보다 자연스럽게 연결시켜 발음하게 되면 단어들 사이의 상호조음현상으로 인해 인식하기 어려워지며, 이 현상은 발음속도가 빨라질수록 심화된다. 세번째로 인식대상어휘의 난이도를 들 수 있으며, 일반적으로 인식하고자 하는 어휘 규모가 커질수록 혼동되기 쉬운 단어가 많아지고 따라서 오인식의 가능성도 커진다. 물론 동일한 어휘 수라고 할지라도 단어들의 음성학적 유사성에 따라 인식난이도는 다르게 된다. 네번째는 언어의 문법구조

및 주제와 관련된 것으로서, 사람들의 일상적인 언어는 컴퓨터 언어와 같이 문법구조에 인위적인 제약을 둔 경우와는 달리 컴퓨터로 해득하기가 용이하지 않으며, 특히 대화체 음성언어는 더욱 인식하기 어렵다. 그리고, 임의의 주제를 대상으로 할 경우 특정한 주제로 발언내용을 제한할 때보다 음성인식의 난이도가 높아진다. 마지막으로 음성통신 환경에 관한 요인으로서 전화음성과 같이 주파수 왜곡이 있거나 배경잡음이 있는 경우 동일한 음성인식시스템이라도 인식 성능이 크게 떨어지게 된다.

2. 음성인식시스템의 기본 구조

그림 1에 일반적인 음성인식 시스템의 구성도가 나타나 있으며, 전체적인 동작을 개략적으로 설명하면 다음과 같다. 마이크를 통해 입력된 음성은 디지탈 신호로 변환되어 음성인식시스템으로 들어오게 되며, 음성인식의 첫 단계인 음성특징분석을 통해 단구간(short-time)별로 음성학적 특징을 잘 표현해 줄 수 있는 음성특징계수들을 추출하게 된다. 추출된 음성특징계수들은 패턴인식과정으로 넘겨져서 미리 저장된 단어 또는 음소들의 모델과 비교하게 되며, 그 결과는 일련의 후보단어 또는 후보음소들의 형태로 언어처리과정에 전달된다. 언어처리과정에서는 후보단어 또는 후보음소들의 정보를 토대로 하여, 인식대상어휘 및 문법구조, 그리고 특정 주제에의 부합 여부를 판단하여 최종 인식된 문장을 출력시키게 된다. 경우에 따라서는 언어처리과정에서 새로운 후보단어나 후보음소를 추정하여 패턴인식과정에 전달하여 이를 확인해 보도록 지시할 수도 있다. 이하에 구성도의 각 부분에 대해 보다 상세히 설명하기로 한다.



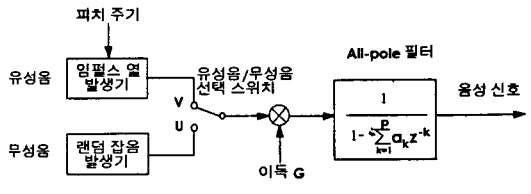
〈그림 1〉 음성인식 시스템의 기본 구성도

가. 음성특징분석

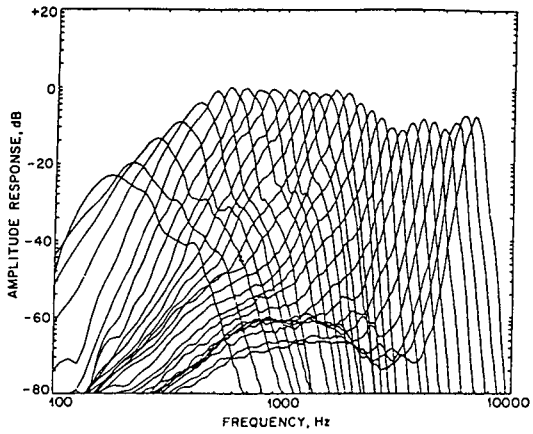
이미 언급한 바와 같이 음성신호에는 언어정보뿐만 아니라 누가 어떤 환경에서 어떤 태도로 말했는가 하는 정보까지 포함되어 있다. 따라서, 음성인식의 전처리 단계인 음성특징분석의 목적은 언어정보, 즉, 음소들간의 차이에 해당하는 음향학적 특성에는 민감하면서도 그 이외의 음향적 변화(배경잡음, 채널왜곡, 화자 차이, 발음 태도 등)에는 둔감한 음성특징 파라메터들을 추출하는 것이다. 그러나, 이러한 이상적인 음성특징추출 방법은 아직까지 알려져 있지 않으며, 일반적으로 음성인식을 위한 음성특징분석 방법으로는 음성발생기관이나 음성청취기관의 단순화된 모델에 근거를 둔 방법들이 주로 사용된다.

음성발생기관의 모델에 근거를 둔 방법의 대표적인 예로는 선형예측부호화(Linear Predictive Coding(LPC)) 방법을 들 수 있다.[12] 이 방법은 음성발생기관 중 성도(vocal tract)의 특성을 all-pole 구조를 가지는 디지탈 필터로 모델링하는 것으로서, 음성신호가 10~20ms 정도의 단구간에서는 stationary하다는 가정하에 이 구간의 음성신호로부터 디지탈필터의 계수들을 추정한다. 이 계수들이 단구간 음성신호의 주파수특성을 표현하게 되는데 특히 음성학적으로 의미있는 특징인 성도의 공명주파수에 관한 정보를 잘 나타내 주기 때문에, 음성인식을 위한 효과적인 특징추출방법으로 널리 사용되고 있다. 그림 2에 LPC 모델에 의한 음성발생기관의 모델이 나타나 있다.

음성의 청취과정에 대한 연구에 따르면 귀의 가장 안쪽 부분에 있는 달팽이관 내부의 기저막(basilar membrane)에서 소리의 주파수분석이 이루어지는 것으로 알려져 있으며, 그림 3에 사람과 유사



〈그림 2〉 LPC모델에 의한 음성 발생 과정



〈그림 3〉 청각기관 중 기저막에서의 주파수 선택특성

한 특징을 가지는 고양이의 달팽이관 내부에서의 신경세포들의 주파수 선택특성의 일부가 나타나 있다.[13] 이에 따라 청각기관의 모델에 기반을 둔 음성특징분석 방법으로 청각기관의 주파수 선택특성을 고려한 대역 필터군(bandpass filter bank)이 사용된다. 경우에 따라서는 청각기관에서의 비선형적 특성들을 포함시키기도 하지만,[13] 대부분 선형특성을 가지는 필터군의 출력을 그대로 음성특징 파라메터로 사용한다.

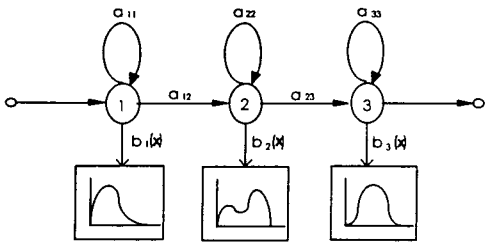
이들 음성특징분석 방법들은 실제 음성인식시스템들에 성공적으로 사용되고는 있으나, 매우 단순화된 모델에 근거를 두고 있기 때문에 여러 가지 한계점들을 드러내기도 한다. 예를 들어 LPC 방법의 경우 all-pole 모델을 사용하기 때문에 zero 특성이 중요한 비음 등의 묘사에 한계가 있는 것으로 알려져 있으며, 일부 자음의 경우 10ms 정도의 단구간에 대해서도 stationarity가 보장되지 않는다. 이에 따라 음성신호의 pole-zero 모델 및 non-stationary 모델도 검토되고 있으나, 계산량 등의 문제로 실제 음성인식시스템에 적용되는 예는 아직 별로 없다. 청각기관의 모델에 기반을 둔 대역 필터군의 경우에도 실제 청각기관에서는 수만 개의 대역 필터들이 사용되지만 구현상의 어려움으로 인해 실제로는 단지 수십 개의 필터 만으로 특징분석을 수행하고 있다. 또한, 달팽이관에서 전기적인 신호로 바뀌어 신경조직을 통해 전달된 음성신호가 두뇌에서 어떻게 처리되는지에 관해서는 현재까지 알려진 지식이 얼마 되지 않기 때문에 결국 음성청취기관의 전반부의 일부만 모델링한 셈이 된다. 앞으로 음성인식성능의 향상을 위해서는 음성발생과정 및 청취과정에 대한 보다 심도깊은 연구를 토대로 음성학적으로 의미있는 정보를 효과적으로 추출하는 음성특징분석 방법이 개발되어야 할 것이다.

나. 패턴인식과정

패턴인식과정은 음성특징분석과정에서 추출된 음성특징 계수들과 가장 잘 부합되는 언어적 표현을 찾아내는 과정이라고 말할 수 있다. 이를 위해서는 먼저 패턴인식을 하기 위한 음성의 기본 단위(단어, 반음절, 음소, 변이음 등)를 정한 다음, 훈련용 음성 데이타로부터 미리 이들 음성단위에 해당하는 각각의 대표패턴 또는 모델을 구해서 저장한다. 그 다음으로 인식하고자 하는 입력음성의 특징패턴이 분석되면 이를 저장된 대표패턴 또는 모델들과 비교하여 가장 가까운 패턴들에 해당하는 음성단위들을 인식된 단어 또는 음소의 후보로 결정하게 된다.

패턴인식과정의 가장 기본적인 방법은 소위 template matching이라는 방법으로서, 이 방법은 미지의 입력특징패턴의 시간열(time sequence)과 저장된 음성단위들의 대표패턴(들)의 시간열을 직접 비교하는 것이다. 이 경우 발음속도의 차이에 따른 영향을 보상해 주기 위하여 보통 Dynamic Time Warping(DTW)이라 불리우는 시간축 정규화 기술이 사용된다. 이 방법은 인식대상 어휘 수가 적고 훈련용 음성데이타가 얼마 되지 않는 상황에서는 매우 효과적인 방법으로 알려져 있다.

두번째는 Hidden Markov Model(HMM)이라는 통계적 방법으로서, 이 방법은 음성단위에 해당하는 패턴들의 통계적인 정보를 확률모델 형태로 저장하고 미지의 입력패턴이 들어오면 각각의 모델에서 이 패턴이 나올 수 있는 확률을 계산함으로써 이 패턴에 가장 적합한 음성단위를 찾아내는 방법이다.[14] HMM 방법에서는 그림 4에서 보는 바와 같이 음성신호를 상태천이확률 $\{a_{ij}\}$와 각 상태에서의 관찰확률 $\{b_i(x)\}$라는 두 단계에 걸친 확률



〈그림 4〉 Hidden Markov Mode(HMM)을 이용한 음소 모델의 예

과정으로 표현하며, 이 때 Markov 특성을 갖는 상태열은 직접 관찰될 수 없기 때문에 hidden Markov 모델이라 불리운다. 이 방법에서는 음성신호로부터 모델의 파라메터들을 추정하고 추정된 모델과 입력된 음성패턴과의 유사도를 측정하는 과정들이 명확하게 정의되어 있으며, 모델훈련에 필요한 양의 음성 데이타가 준비될 경우 성능 면에서도 가장 우수한 것으로 평가되고 있어서 현재 음성인식을 위한 패턴인식 방법으로 가장 널리 사용되고 있다. 특히 통계적인 언어모델이 사용될 경우 HMM 방법은 음성처리 및 언어처리를 단일구조로 처리할 수 있다는 큰 이점을 가진다. 물론 HMM 방법에도 각 상태에서의 출력패턴들이 서로 독립적이라는 가정이나 상태지속시간이 지수함수분포로 주어지는 점 등 실제의 음성신호특성과 부합되지 않는 부분들이 지적되고 있으며, 이러한 문제점들을 극복하기 위한 시도들이 계속되고 있다.

세번째로 최근 패턴인식의 새로운 접근방법으로 관심을 모으고 있는 인공신경망에 의한 방법을 들 수 있으며, 이 방법은 부분적으로나마 인간의 두뇌모델에 기초를 두고 학습이 진행됨에 따라 점차적으로 정보분류능력이 향상되는 신경망 구조를 이용하고 있다.[15] 그러나, 이 방법은 template matching이나 HMM 방법에 비해 시간영역에서의 동적인 처리능력 면에서는 아직 뒤떨어지는 것으로 알려지고 있으며, 이에 따라 HMM 방법과 인공신경망 방법을 통합하려는 시도들도 많이 이루어지고 있다.

마지막으로 지금까지의 접근방법과는 전혀 다른 방법으로서 지식기반 또는 규칙에 의한 방법을 들 수 있다. 이 방법은 인공지능 기술에서의 전문가 시스템의 일종으로 생각할 수 있으며, 음성특징분석 결과로부터 음성의 기본단위들로 분류하는 과정에서 일련의 음성학적인 규칙들을 선별적으로 적용하여 최종 인식결과를 도출해내는 방법이다. 현재로서는 음성분류에 적용될 음성학적 지식이 불충분하며 또한 이들 지식을 효과적으로 운용하는 방법론에 한계가 있기 때문에 성능 면에서 통계적인 패턴인식 방법에 비해 저조한 결과를 나타내고 있으나, 여건이 성숙됨에 따라 음성인식의 핵심적인 방법으로 부각될 것으로 전망된다.

다. 언어학적 처리

수십 단어 정도의 고립단어인식이 목표인 응용분야에서는 단어를 음성인식의 기본 단위로 삼으면 되고 그 이상의 언어학적 지식은 불필요할 수도 있다. 그러나, 대용량 어휘에 의한 문장형태의 음성인식을 위해서는 음소와 같이 단어보다 작은 음성단위를 인식의 기본단위로 택하는 것이 필수적이다. 그러나, 언어학적 지식을 전혀 동원하지 않은 상태에서 현재의 기술로 도달할 수 있는 화자독립 음소인식율은 70% 정도에 불과하다. 그리고, 실제로 사람의 경우에도 무의미 단어로 구성된 문장의 청취도는 이보다 그리 높은 수준이 못되는 것으로 알려져 있으며, 이는 음성인식에서 언어학적 정보가 차지하는 중요성이 얼마나 큰가를 보여주는 좋은 예이다. 일반적으로 언어학적 지식은 단어구성에 관한 어휘론적(lexical) 지식, 문법구조에 관한 구문론적(syntactic) 지식, 문장 의미를 다루는 의미론적(semantic) 지식, 그리고 주제에의 부합여부를 판단하는 실용론적(pragmatic) 지식으로 이루어진다. 그 중에서 의미론 및 실용론적 지식을 음성인식에 적용하는 데에는 구현상의 어려움이 있기 때문에, 현재 음성인식에 사용되는 언어학적 모델은 주로 어휘론 및 구문론적 지식에 기반을 두고 있다.

음성인식의 응용분야에 따라서는 컴퓨터 언어와 같이 매우 제한된 언어 모델을 사용할 수도 있겠으나, 일반적으로 음성인식은 자연언어를 대상으로 하고 있기 때문에 형식언어의 문법을 적용하기는

곤란하다. 따라서, 음성인식에는 N-gram 형태의 통계적 언어 모델이 주로 사용되며, 이 모델에서는 일련의 단어들로 구성된 문장이 생성될 확률을 N−1개의 단어로 구성된 단어열 다음에 특정 단어가 나타날 확률들로써 표현한다. 이를 위해서는 방대한 양의 문장 corpus로부터 이들 단어열에 대한 확률을 추정해야 하며, 실제로 두 개 또는 세 개의 연속된 단어 사이의 연관관계를 확률적으로 모델링한 bigram 및 trigram 언어 모델 등이 제한된 주제에 의한 음성인식에 성공적으로 사용되고 있다.[16] 그러나, 주제에 제한을 두지 않은 음성이나 자연스러운 대화체 연속음성을 인식하기 위해서 어떠한 언어 모델을 사용하고 이를 어떻게 구현할 것인가에 관해서는 아직도 해결해야 할 많은 문제점이 남아 있다. 그리고, 의미론적 지식과 운율정보 등을 음성인식에 효과적으로 활용하는 방법도 앞으로 연구되어야 할 과제이다.

3. 음성인식시스템의 성능

현재의 기술수준으로도 실험실 환경에서의 고립단어 및 연결단어 인식성능은 매우 높다. 예를 들어 방음실에서 녹음한 10개의 숫자음에 대한 고립단어인식의 경우 훈련을 거치지 않은 불특정화자에 대해서도 99.7%의 인식률을 나타내며, 어휘에 따라서는 수백 단어 이상의 어휘에 대해서도 97% 이상의 인식 성능이 얻어지고 있다.[16] 물론 영어의 alphabet과 같이 혼동을 일으키기 쉬운 단어(예를 들면, B, C, D, E, G, P, T, Z 등)들이 포함된 어휘에 대해서는 수십 단어라도 90%를 조금 넘는 수준 밖에는 얻을 수 없으며, 이는 앞서 언급한 바와 같이 현재의 기술로 도달할 수 있는 불특정화자 음소인식율이 70% 정도에 불과함을 고려하면 당연한 결과이다.

보다 심각한 문제는 잡음이나 채널왜곡이 있는 실제적인 환경에서 인식실험을 할 경우 인식성능이 방음실에서의 실험결과에 비해 현저하게 떨어진다는 점이며, 10개의 숫자음의 경우 실제 환경에서의 단어인식률은 약 98% 정도로 방음실에서의 인식률 99.7%에 비해서는 매우 저조한 결과이다. 잡음 및 채널왜곡에 대한 문제는 다음 절에서 다시 언급하기로 하겠다.

연속음성인식의 경우 미국 국방성에서 1980년대 중반이후 음성언어(spoken language) 프로그램이라는 이름으로 추진되는 과제에서 매년 참여연구기관들의 연구결과를 동일한 음성 데이타베이스를 대상으로 비교평가하는 자리를 마련하고 있으며, 해군 자원관리를 주제로 한 1000단어 규모의 음성 데이타베이스에 대해 Carnegie Mellon 대학을 비롯한 몇몇 연구기관이 96% 이상의 인식성능을 얻었다. 지난 1994년 말에는 이보다 훨씬 난이도가 높은 NAB(North American Business) News를 대상으로 한 평가가 진행되었는데 그 결과는 표 2와 같다.[17] 표에서 보는 바와 같이 가장 우수한 시스템의 경우 단어오인식률이 7.2%라는 뛰어난 성능을 기록했지만, 실용화의 관점에서는 아직 요원한 결과이다.

〈표 2〉 1994년 미국 국방성 연속음성인식 성능평가결과

인식 test 참여기관	단어오인식률
Cambridge University(HTK Group)	7.2 %
IBM T.J. Watson Research Labs	8.6%
CNRS-LIMSI, Paris	9.2%
AT&T Bell Laboratories	10.0%
BBN Systems & Technologies	10.2%
SRI International	10.3%
Dragon Systems Inc.	10.3%
Philips Research Aachen	10.6%
Boston University	10.9%
Carnegie Mellon University	10.9%
New York University	11.0%
Cambridge University(NN Group)	12.4%
MIT Lincoln Laboratory	17.4%
CRIM, Montreal	20.2%
Karlsruhe University	22.8%

IV. 음성인식의 실용화를 위한 음성처리기술

앞으로의 음성인식기술 연구는 크게 두 가지 방

향으로 나누어 질 수 있을 것이다. 그 중 첫번째는 음성인식의 궁극적인 목표인 대화체 연속음성의 인식을 추구하는 방향으로서 이를 위해서는 대화체 언어처리기술이 관건이 될 것으로 보인다. 두번째는 현재의 기술수준을 토대로 음성인식기술의 실용화를 추구하는 방향으로서 잡음 및 채널왜곡 환경에서의 음성인식기술, 새로운 화자에 대한 적응기술, 그리고 핵심어검출기술 등이 이 범주에 포함된다. 본 절에서는 이들 음성인식의 실용화를 위한 음성처리기술에 대해 살펴보기로 한다.

1. 잡음 및 채널왜곡 환경에서의 음성인식

대부분의 음성인식시스템이 방음실에서 녹음한 음성에 대해서는 우수한 성능을 나타내다가도 잡음과 채널왜곡이 있는 실제환경에서는 그 성능이 급격히 저하되는 특성을 보이는데, 이는 이들 시스템이 실제환경에서의 음향학적 변화요인들에 대한 대처능력을 가지고 있지 못하기 때문이다. 인식성능에 영향을 미치는 음향학적 변화요인들로는 배경잡음과 입력 마이크의 특성, 전화회선 등에서 비롯되는 채널왜곡 등을 들 수 있다. 그 외에도 배경잡음이 화자의 발성에 영향을 미쳐 음성신호 자체에 왜곡을 가져다 주는 문제도 고려해야 할 대상이며, 이를 Lombard 효과라고 부른다.

잡음 및 채널왜곡 환경에서 음성을 인식하기 위해 시도되고 있는 방법들은 크게 다음과 같은 부류들로 나누어 볼 수 있다. 그 중 첫째가 음성인식을 위한 훈련과정을 잡음 및 채널왜곡 환경에서 수행하는 방법으로 잡음 및 채널왜곡 특성이 일정한 경우에는 효과가 있지만, 시간에 따라 환경특성이 변화되는 경우에는 사용하기 곤란하다. 이 방법의 변형된 형태로서 다양한 환경 및 다양한 발성방식 하에서 음성데이타를 수집하여 인식 모델을 구성하는 multi-style 훈련방법도 시도되고 있다. 두번째로 잡음만이 문제가 되는 상황에서는 음성인식의 전처리 과정으로 잡음음성의 음질개선을 수행하는 방법을 사용할 수 있다. 이러한 전처리 과정으로는 가장 고전적인 spectral subtraction 방법으로부터 다수의 마이크 배열을 이용한 적응잡음제거방법에 이르기까지 많은 방법들이 개발되어 왔다. 세번째는 잡음이나 채널왜곡으로 인한 영향에 강인한 음성특징 표현을 사용하는 것으로서, 청각기관의 모델에 기반을 둔 몇몇 방식들이 잡음 및 채널왜곡 환경에서 상대적으로 우수한 성능을 나타내는 것으로 보고되고 있다.[13,18] 네번째는 음성특징 패턴들을 비교하는 과정에서 잡음 영향에 강인한 스펙트럼 비교척도를 사용하는 방법으로 projection-based distortion measure 등이 대표적인 예이다.[24] 그 외에도 다양한 종류의 마이크를 사용할 때에 마이크 특성 차이에 의한 영향을 보상해주기 위한 방법들도 다수 제안되고 있다.[19]

현재의 대부분의 음성인식시스템이 마이크와 입사이의 거리를 일정하게 유지시키기 위해 headset 마이크를 머리에 쓰거나 손에 마이크를 들고 음성을 입력하는 방법을 사용하고 있으나, 이는 기계와 대화하기 위한 편리한 수단을 제공하고자 하는 음성인식의 취지상 바람직하지 못하다. 이에 따라, 시스템에 장착된 마이크가 말하는 사람을 추적하면서 양호한 음질의 음성입력을 유지토록 하는 방식에 대한 연구가 진행되고 있으며, 이를 위해 다수의 마이크 배열을 이용하여 화자를 추적하는 방식이 검토되고 있다. 그러나, 현재로서는 실내반향(reverberation)이나 다른 화자들의 음성에 의한 영향 등 해결해야 할 많은 과제를 안고 있는 실정이다.[9]

2. 화자적응 및 화자정규화 기술

이미 언급한 바와 같이 사람마다 발음기관의 크기와 모양에 차이가 있고 성장과정에서 습득된 발음 습관도 서로 다르므로, 음성신호에는 화자에 따른 영향이 나타나게 된다. 따라서, 인식을 위한 음성모델(또는 패턴)들을 훈련시키기에 충분한 만큼의 특정화자의 음성 데이타가 준비될 수 있다면, 훈련에 참여한 특정화자의 음성만을 인식하는 화자종속(speaker-dependent) 음성인식의 성능이 훈련에 참여하지 않은 임의의 화자의 음성을 인식하는 화자독립(speaker-independent) 음성인식에 비해 훨씬 우수하다. 그러나, 이러한 훈련과정은 사용자의 불편함과 더불어 많은 시간과 노력이 투자되는 일이며, 특히 인식대상 어휘가 많을 경우에

는 비현실적이다.

이 문제에 대한 현실적인 대안으로서 미리 많은 사람의 음성 데이타로부터 화자독립 음성모델을 구성한 다음 특정화자가 발음한 소규모의 훈련용 음성 데이타에 포함된 정보를 효과적으로 활용함으로써 화자독립 음성인식시스템의 성능을 향상시키는 방법들이 검토되고 있다.[20] 이 때, 특정화자의 음성에 따라 화자독립 음성모델을 변화시키는 방식을 화자적응(speaker adaptation) 기술이라 하고, 화자독립 음성모델에 잘 부합되도록 특정화자의 음성특징을 변화시키는 방식을 화자정규화(speaker normalization) 기술이라 부른다. 이들 화자적응 및 화자정규화 기술은 특정화자의 훈련용 음성데이타에 대한 text 정보가 제공되느냐 여부에 따라 supervised 방법 및 unsupervised 방법으로 분류되며, 훈련과정과 인식과정의 구분여부에 따라 다시 off-line 방법과 on-line 방법으로 나누어진다. 특정화자에 의한 훈련을 최소화시키면서 화자종속 음성인식에 가까운 인식성능을 얻기 위한 다양한 방법들이 제안되고 있다.[20]

3. 핵심어검출(Keyword Spotting) 기술

핵심어검출기술은 어휘에 제한없이 자연스럽게 발음한 연속음성으로부터 미리 정해진 특정 단어(keyword)들을 검출해 내는 것이다. 많은 경우 이러한 keyword들은 강세가 주어진 상태에서 발음되거나 충분히 명료하게 발음되며, 이러한 핵심어를 검출하는 일은 연속음성 전체를 인식하는 것에 비해서는 훨씬 용이한 작업이다. 따라서, 핵심어검출은 고립단어인식에서 사용자가 단어를 또박또박 띄어 발음해야 하는 불편함과 연속음성인식이 지니는 성능저조의 문제점을 모두 해결할 수 있다. 그러므로, 이 기술은 자연스럽게 발음된 문장 내에서 핵심어들만 검출해 내면 의미가 통할 수 있는 많은 응용분야에 매우 효과적으로 적용될 수 있다. 현재 미국 AT&T사 등에서는 핵심어검출기술을 이용한 전화교환업무의 자동화 서비스를 1992년부터 시행하고 있으며,[21] 음성메세지의 자동분류에 핵심어검출기술을 이용하는 방안도 검토되고 있다.[22]

핵심어검출기술은 기본적으로 검출하고자 하는 keyword들과 그 밖의 음성부분(non-keyword) 그리고 비음성구간(silence)을 각각 별도의 HMM들로 모델링하는 것을 근간으로 하고 있다. 여기서 non-keyword 모델이 keyword 음성부분을 잠식하지 않으면서 그 밖의 음성부분을 얼마만큼 효과적으로 표현해 줄 수 있는가에 따라 핵심어검출 시스템의 성능이 크게 좌우된다.

4. Barge-in 기술

일반적으로 사람들은 기계를 대상으로 말하는 것이 익숙하지 않으며, 따라서 음성인식의 실용화를 위해서는 적절한 음성안내어(voice prompt)를 통해 사람으로 하여금 기계가 알아듣기 쉽게(음성인식이 용이하게) 말하도록 유도하는 것이 중요하다. 그러나, 음성인식시스템을 자주 사용하게 되는 사용자들에게는 매번 나오는 음성안내어를 다 듣고나서 음성을 입력해야 한다는 것은 매우 불편한 일이 아닐 수 없다. Barge-in 기술은 음성안내어가 나오고 있는 도중에 음성을 입력하더라도 자동적으로 음성안내어를 중단시키고 음성안내어와 중복되어 입력된 음성부분에 대해서도 정확하게 인식하도록 하는 기술로서, 이를 위해서 adaptive echo cancellation 기술이 사용된다.

V. 음성인식기술의 응용

사람과 자연스럽게 대화할 수 있는 음성인식기술이 개발된다면 그 응용분야는 본고에서 논의할 필요조차 없을만큼 무궁무진할 것이다. 그러나, 현재의 음성인식기술은 사용 어휘나 발음방식 면에서 제한을 두는 것이 대부분이며, 이러한 제약조건 하에서도 어느 정도의 인식오류를 피할 수 없다. 따라서, 현상태에서 음성인식기술이 적용될 수 있는 분야들은 다음과 같은 조건을 만족시켜야 한다.[6] 첫째로 인식오류가 치명적인 영향을 주지 않아야 한다. 둘째로 기계를 대상으로 말하는 불편함을 감수하고자 하는 사람들에게 실질적인 유익이 있어

야 한다. 세째로 실시간에 인식결과를 얻을 수 있어야 한다. 물론 지속적인 이용을 위해서는 인식성능이 어느 수준이상 되어야 하며, 성능의 요구수준은 응용분야에 따라 차이가 있을 것이다. 이러한 제한조건들에도 불구하고 여러 분야에서 음성인식기술의 실용화가 추진되고 있는데, 본고에서는 몇가지 분야만 살펴보기로 한다.

그 중 첫번째로 정보통신분야를 들 수 있다. 음성인식은 사람이 전화를 이용하여 컴퓨터와 정보를 주고받는 것을 가능케 해준다. 예를 들어, 미국의 전화회사들은 증권시세를 비롯한 각종 생활정보 조회, 은행업무 등을 전화음성인식에 의해 수행하는 서비스를 제공하기 시작하였다. 이러한 서비스는 기존에는 버튼식 전화기로만 가능했던 기능들을 다이얼식 전화기를 가지고도 이용할 수 있도록 해줄 뿐만 아니라, 전화기의 버튼입력으로는 곤란한 회사이름 등을 음성을 통해 컴퓨터에 전달하여 관련 정보를 조회할 수 있게 한다. 그 외에도 전화교환원이 수행하던 서비스를 자동화하는 데에도 음성인식기술이 사용되고 있는데, 전화번호 안내서비스나 음성 dialing 서비스 등이 그 예이다.[6,23]

두번째로 음성인식에 의한 원고작성(voice dictation)을 들 수 있으며, 영어의 경우 2만 단어에서 5만 단어 정도의 어휘를 인식할 수 있는 제품들이 상품화되었다. IBM, Dragon Systems, 및 Kurzweil AI사 등에서 개발된 이들 제품들은 사용자가 자신의 목소리로 수 분에서 수십 분 정도 말을 하여 인식시스템이 자신의 목소리에 익숙해지도록 하고, 각 단어와 단어 사이를 또박또박 띄어 읽어야 한다는 등의 제약조건이 있기는 하지만, 분당 50단어 이상의 입력이 가능하다. 단어 오인식률은 3%에서 5% 정도이며, 오류가 발견될 경우 음성명령에 의해 즉시 수정할 수 있는 기능을 가지고 있어서, 방사선과 의사들의 X선 사진 판독보고서 작성 등의 응용분야에 활용되고 있다.

음성인식의 세번째 응용분야로 일반 가정용 또는 산업용 기기들에 음성인식기능이 부가되는 것을 들 수 있다. 사람이 전화번호 또는 상대방의 이름을 말하면 자동적으로 전화를 걸어주는 음성인식 전화기, 음성으로 방송 채널을 선택하는 카 스테레오, 음성으로 예약녹화를 지시하는 비디오기기(VCR), 그리고 음성명령에 의해 작동하는 퍼스널 컴퓨터(PC) 등이 이미 등장하고 있으며, 장난감이나 오락기기에도 음성인식기능이 적용되고 있다.

이 외에도 음성인식기술은 신체장애자들에게 많은 도움을 줄 수 있으며, 음성합성 및 기계번역기술과 결합되어 만들어지는 자동통역전화 서비스도 현재 여러 나라에서 많은 연구가 진행되고 있는데, 제한된 주제에 대해서나마 성공될 경우 지대한 파급효과가 기대되는 분야이다.

VI. 맺 음 말

음성인식기술의 향후 전망과 관련해서는 섣부른 낙관이나 비관, 그 어느 쪽도 할 수 없는 상황이라고 보여진다. 1970년대 음성인식연구가 본격적으로 추진되기 시작한 시점만 해도 1980년대 후반이면 음성인식분야의 시장이 크게 형성될 것이라는 예측이 나왔었지만, 1990년대 중반에 접어드는 현재 시점에서 보더라도 음성인식기술의 실생활에의 응용은 아직까지 시작단계라고 밖에 볼 수 없으며, 사람과 자유롭게 대화하는 컴퓨터가 과연 언제 출현할 수 있을지에 대해서는 예측조차 하기 어려운 실정이다. 또한 일부의 비관론자들은 음성인식이 너무나 어려운 과제이므로 이 분야의 연구에 투자하는 것은 낭비에 불과하다는 주장을 했지만, 오늘날 전화음성을 통해 컴퓨터에 수록된 정보를 조회하고 음성인식에 의해 원고를 작성하는 일 등이 현실 세계에서 이미 진행되고 있다.

실제로 미국 등의 선진국에서는 특정 응용분야를 대상으로 하는 불특정화자 대용량 어휘의 음성인식시스템이 금세기 안에 실용화될 것이라고 전망하고 있다. 그리고 전화음성을 이용한 정보조회 및 음성명령에 의한 컴퓨터 조작 등은 더욱 가까운 시장래에 일반화될 것으로 예상된다. 그러나, 현재의 기술수준, 특히 대화체 언어처리에 관한 기술수준으로 볼 때, 사람과 자유롭게 대화할 수 있는 컴퓨터의 개발은 다음 세기 중에나 실현될 수 있을

것으로 보인다. 따라서, 앞으로의 음성인식연구는 대화체 연속음성인식에 큰 비중이 두어질 것으로 판단되며, 이와는 별도로 현재 실험실 환경에서 진행되어 온 음성인식기술의 실용화를 위한 작업들(잡음환경에서의 음성인식 및 화자적응 방식연구 등)에 관심이 집중되고 있다.

음성인식기술이 이미 목도한 정보통신시대의 주요 핵심기술 중 하나가 될 것임에는 이론의 여지가 없다. 다국어 음성언어처리를 겨냥한 외국의 유수 기관들이 한국어 음성처리에도 상당한 관심을 보이고 있어서 머지않아 우리말을 포함한 다국어 음성언어시스템들이 출현할 수 있음을 고려할 때, 우리말을 사용하는 분야에서마저 외국기술에 종속되지 않기 위해서는 한국어의 음성학 및 언어학을 비롯하여 반도체, 신호처리, 컴퓨터공학에 이르기까지 제 분야의 국내 연구인력들의 협력연구가 필수적으로 요청된다.

참 고 문 헌

〔1〕 이현복, "음성학의 중요성," 제1회 음성학 학술대회 자료집, 대한음성학회, pp.1～4, 1994년 2월

〔2〕 구명완, "음성인식기술의 현황과 전망," 대한전자공학회지, 제20권 제5호, pp.41～50, 1993년 5월

〔3〕 김순협, "음성인식 기술 현황," 한국통신학회지, 제11권 제5호, pp.40～56, 1994년 9월

〔4〕 구명완, "전화망을 통한 음성인식시스템 개발현황," 한국통신학회지, 제11권 제5호, pp.8～16, 1994년 9월

〔5〕 김형순, 김희동, 임병근, 은종관, "음성처리 기술의 현황과 전망," 한국통신학회지, 제8권 제6호, 1991년 6월

〔6〕 L. R. Rabiner, "Applications of Voice Processing to Telecommunications," Proc. IEEE, vol.82, no.2, pp.199～228, Feb. 1994.

〔7〕 J. R. Deller, Jr., J. G. Proakis and J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, 1993.

〔8〕 Y. K. Muthusamy, E. Barnard and R. A. Cole, "Reviewing automatic language identification," IEEE Signal Processing Magazine, vol.11, no.4, pp.33～41, Oct. 1994.

〔9〕 R. Cole and L. Hirschman et al., "The challenge of spoken language systems : research directions for the nineties," IEEE Trans. Speech and Audio Processing, vol. 3, no.1, pp.1～21, Jan. 1995.

〔10〕 L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

〔11〕 V. W. Zue, "The use of speech knowledge in automatic speech recognition," Proc. IEEE, vol.73, no.11, pp.1602～1615, Nov. 1985.

〔12〕 J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, 1976.

〔13〕 O. Ghitza, "Auditory nerve representation as a basis for speech processing," in Advances in Speech Signal Processing, S. Furui and M. Sondhi, Eds., Marcel Dekker, pp.453～485, 1992.

〔14〕 L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257～286, Feb. 1989.

〔15〕 A. Waibel, "Neural network approaches for speech recognition," in Advances in Speech Signal Processing, S. Furui and M. Sondhi, Eds., Marcel Dekker, pp.555～595, 1992.

〔16〕 F. Jelinek, R. L. Mercer and S. Roukos, "Principles of lexical language modeling

for speech," in Advances in Speech Signal Processing, S. Furui and M. Sondhi, Eds., Marcel Dekker, pp.651~699, 1992.

[17] D. S. Pallett and J. G. Fiscus et al., "1994 Benchmark Tests for the ARPA Spoken Language Program," Proceedings of the Spoken Language Technology Workshop, Jan. 1995.

[18] H. Hermansky et al., "RASTA-PLP speech analysis technique," in Proc. 1992 IEEE ICASSP, pp.I-121-124, Mar. 1992.

[19] A. Acero, Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, 1993.

[20] S. Furui, "Speaker-independent and speaker-adaptive recognition techniques," in Advances in Speech Signal Processing, S. Furui and M. Sondhi, Eds., Marcel Dekker, pp.597~621, 1992.

[21] J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidded Markov models," IEEE Trans. Acoust., Speech, Signal Processing, vol.38, no.11, pp.1870~1878, Nov. 1990.

[22] 김형순, "Keyword Spotting 기술," 한국통신학회지, 제11권 제5호, pp.57~66, 1994년 9월

[23] D. B. Roe and J. G. Wilpon, "Whither speech recognition : the next 25 years," IEEE Communication Magazine, Vol.31, No.11, Nov. 1993.

[24] D. Mansour and B. H. Juang, "A family of distortion measures based upon projectin peration for robust speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.37, no.11, pp.1659~1671, Nov. 1989.

저 자 소 개

金 炯 淳

1960年 8月 21日生

1983年 2月	서울대학교 공과대학 전자공학과(공학사)
1984年 2月	한국과학기술원 전기 및 전자공학과(박사과정 조기진학)
1989年 2月	한국과학기술원 전기 및 전자공학과(공학박사)
1987年 1月~1992年 6月	디지콤 정보통신연구소(연구부장)
1992年 7月~현재	부산대학교 공과대학 전자공학과(조교수)

주관심분야 : 음성신호처리, 디지탈 통신