음성인식기술의 현황과 전망

이 종락

한국통신 연구개발단

(요 약)

인간의 가장 익숙한 정보교환 수단인 음성을 기계가 인식하게 함으로써 모든 기계를 말로써 작동시키고자 하는 것은 인간의 오랜 꿈이었다. 최근 컴퓨터 기술과 음성처리 기술의 급속한 발달에 힘입어 그 꿈은 현실로 다가오고 있다. 현재 고립어 인식은 충분히 실용화될 수 있는 단계에 들어 섰으며 이제 연속어 인식 내지 연속어 이해에 연구가 집중되고 있다. 인간과 기계를 인터페이스하는 언어의 전위레벨로서 날로 그 중요성이 부각되고 있는 음성인식 기술의 현황을 살펴보고 그것의 미래를 전망해 본다.

Ⅰ. 서론

음성은 인간들간의 가장 보편적이고 편리한 통신수단이다. 뿐만 아니라 인간의 사고를 담는 그릇인 언어체계 또한 음성으로부터 비롯된다. 그렇기 때문에 인간의 사고와 논리는 음성언어를 떼어 놓고는 생각하기 어렵다.

음성의 장점은 에너지면에서 극히 경제적인 입출력이 가능하다는 점과 그리 멀지않은 거리내에서는 원격 입출력이 가능하다는 점이다. 또한 음성은 인간이 일생동안 배우고 사용하는 가장 익숙한 입출력 수단이다. 이 때문에 음성인식과 합성은 인간과 기계간의 가장 자연스러운 인터페이스 수단으로 생각되고 있다. 본 고에서는 그 중에서 특히 음성인식 분야의 현황과 전망에 대해 기술하기로 한다.

제1절 서론에 이어 제2절에서 음성인식기술의 발전사와 현황을 약술하고 제3절에서는 현재까지 개발된 각종 음성인식 알고리즘을 소개하며 제4절에서는 실제 구현된 음성인식 시스팀에 관하여 살펴보고 제5절에서 음성인식기술의 전망에 대해 논의하고 결론을 맺는다.

Ⅱ. 음성인식연구 현황

음성연구의 역사는 매우 길지만 1940년경 음향학적 접근법이 도입됨으로써 활기를 띄게 되었고 Chiba, Flanagan, Fant, Stevens 등의 선구적인 연구에 의해 음성의 발생 메커니즘이 조금씩 이해되기 시작하였다. 그 무렵 Bell 연구소의 Ralph Potter 등에 의해 개발된 speech spectrograph는 이들 음향학적 접근법에 의한 음성연구를 가속시키는 획기적인 전기를 마련하였다. 한편 청각에 관한 신경 생리학 및 실험 심리학적인 연구도 활발해져서 주로 고양이를 재료로 한 해부학적 실험에 의해 귀 내부의 와우각(cochlea)의 구조가 특정 진동수에 공진하는 hair cell의 집단으로 구성된 일종의 filter bank 형태인 것과 각 filter의 공진 주파수 간격이 약 1kHz 이하에서는 균등하나 그 이상에서는 거의 선형적으로 증가한다는 mel scale 또는 Bark scale로 분포되어 있음이 noise masking 실험 등을 통하여 규명되었다. 한편 1951년에 Haskins 연구소의 Cooper에 의해 개발된 Pattern Playback은 speech spectrograph의 역기능을 가진 장치로서 Dellatre, Liberman 그리고 Cooper는 그 장치를 이용하여 실험적으로 locus principle을 발견하는 등 성과를 올리기도 했다.

그러나 이와같은 많은 노력에도 불구하고 이들 음향음성학적 연구결과를 가지고 음성인식이나 음성합성에 적용하기에는 아직 충분치 못하다. 예컨대 음성발생 메커니즘에 대한 연구도 현재까지의 결과는 낱음소에 대한 정적인 것으로서 여러 종류의 음소가 연달아 발음되는 경우 발생하는 조음결합(coarticulation)과 같은 동적인 현상에 대해서는 충분히 규명되어 있지 못한 실정이다. 또한 locus theory도 Klattalk 등 여러 formant synthesis type의 합성 시스팀에서 적용되고 있음에도 불구하고 이론적으로 규명되지 못하여 Fant 등 많은 음향음성 학자들로부터 인정을 받지 못하고 있다. 이들 주변계(peripheral system)에 대한 연구는 물론 중추신경계에서의 음성발생 및 인식에 관한 연구가 진전되어 그들의 기능이 규명되면 음성인식 또는 합성시스팀의 개발에도 큰 진전이 있을 것이다.

음성인식기술의 발전은 음성처리기술 특히 음성코딩기술의 발전과 더불어 본격화되었다고 할 수 있다. 1938년에 Bell 연구소의 Homer Dudley는 전기회로로 사람의 목소리를 흉내낼 수 있는 Voder라고 하는 일종의 음성합성장치를 개발하였다. 1939년 뉴욕에서 열린 세계 박람회에서 Voder가 선보여진 이래 그것에 음성분석기능을 부가함으로써 통신용으로 사용하려는 노력이 경주되었고 그 결과 많은 종류의 보코더(vocoder)가 고안되었는데 예를 들면 channel 보코더, phase 보코더, correlation 보코더, formant 보코더 등등 실로 보코더의 춘추전국시대를 맞이하게 되었다. 이 무렵인 1958년에 이미 몇 종류의 애널로그식 음성인식 시스팀이 고안되었으나 특정화자에 대한 인식률은 음소일 경우 60% 단어에 대해 24% 정도로서 상당히 낮은 편이었다. 이 시스팀에 제한조건을 주면 인식률이 각각 72%와 44%로 향상되었고 특정화자에 최적화시키고 임의의 다른 화자에 대해 실험했을 때의 음소 인식률은 약 45%로 떨어진 것으로 보고되었다.

이상의 애널로그식 보코더들은 곧이어 등장한 LPC(Linear Predictive Coding) 보코더나 homomorphic 보코더와 같은 디지틀 방식의 보코더들에 밀려나게 된다. 보코더들은 송신측에서 20ms 정도의 일정시간마다 음성을 분석하여 음성의 세기와 유성·무성음 여부, 성도모양을 나타내는 파라미터들, 그리고 유성음의 경우에는 추출한 피치주기를 함께 수신측으로 전송하면 수신측에서는 유성음일 경우 임펄스열, 무성음일 경우 백색잡음을 발생시키고 그것들을 성도 파라미

터가 나타내는 공진특성을 가진 필터에 입력시킴으로써 음성을 재생하는 원리를 이용한 것으로서 수신측의 구조는 모두 동일하며 송신측의 음성분석방식에 의해 특징지어진다. 이들 보코더 중에서 음질이나 계산량 면에서 유리한 LPC 보코더는 2400bps급의 저전송률 통신이 필요한 여러 분야에서 실제로 쓰이고 있다. 보코더는 전송률이 낮은 대신에 음질 특히 자연성이 나쁜 특징이 있기 때문에 지금도 꾸준히 개량되고 있는 중이며, 현재는 CELP(Code Excited Linear Prediction)나 RPE-LTP(Regular Pulse Excited with Long-Term Prediction) 보코더 등이 연구중에 있다.

보코더의 발달은 음성인식 연구에 큰 영향을 미쳤는데 특히 보코더에서 사용된 음성분석기법은 음성인식 시스팀의 첫단계인 특징추출 단계에서 그대로 이용된다. 이와같은 디지틀 음성처리기술의 발달에 힘입어 거기에다 때마침 불어온 인공지능(Artificial Intelligence) 기술의 가능성을 결합시킴으로써 고립어가 아닌 연속어를 인식하고자 하는 야심찬 계획이 미국 국방부의 ARPA(Advanced Research Project Agency)에 의해 1971년부터 5년간 총 1500만불의 연구비를 투입하여 9개의 연구소 및 대학에서 수행되었다. ARPA SUR(Speech Understanding Research) 프로젝트로 불리는 이 프로젝트의 기본 취지는 한마디로 인공지능적 기법에 의해 문법 및 의미론과 같은 상위계층의 지식을 가지고 문맥을 이해함으로써 인식률을 획기적으로 제고시킬 수 있을 것이라는 데 있었으나 5년후의 결과는 CMU(Carnegie-Mellon University)의 Harpy 시스팀을 제외하고는 모두 목표에 미달하는 것으로서 결론적으로 ARPA SUR 프로젝트는 실패라는 다소 실망스러운 것이었다.

SUR 프로젝트의 실패의 원인은 음향레벨에서의 처리가 너무나 허술한 데 있었다. 사실 Harpy가 어느정도 성공을 거둔 것도 음향레벨의 처리에 있어 다른 시스팀들보다는 효과적인 방법을 사용했기 때문이라고 볼 수 있다. ARPA SUR 프로젝트는 그러나 음성인식시스팀 개발을 위한 최초의 대형 프로젝트로서 음성인식연구를 본격화하게 하는 계기가 되어 이후 세계 각국에서 유사한 프로젝트가 속속 추진되게 되었다. 또한 ARPA SUR 프로젝트의 실패는 언어처리보다 음향적 처리능력을 강화해야 한다는 반성과 교훈을 주었다.

ARPA SUR 프로젝트의 시작과 거의 같은 시기인 1968년과 1971년에 Vintsyuk와 Sakoe가 각각 독립적으로 제안한 DTW(Dynamic Time Warping) 방식은 DP(Dynamic Programming) 기법을 이용하여 종래의 linear time alignment 방식을 dynamic time alignment가 되도록 개량한 방식으로서 segmentation과 labelling을 하지 않고 입력음성을 미리 저장해둔 단어 또는 문장단위의 speech template들과 차례로 비교하여 가장 유사한 template를 찾아내는 방법을 사용하는 것이었는데 높은 인식률을 얻어 상당한 성공을 거두었다.

DTW 방식은 그 이전의 어떤 방법보다도 높은 인식률을 보인 반면에 여러가지 단점을 가지고 있다. 첫째는 DTW 방식에서는 입력음성을 저장된 모든 template와 비교해야 하기 때문에 인식시 필요한 계산량이 매우 많은데 일반적으로 DTW에서의 계산량은 어휘규모에 비례하여 증가한다. 또한 DTW는 음소단위의 인식이 어렵기 때문에 대규모 어휘의 인식에 부적합하며 원리적으로 화자독립적인 인식에도 부적합하다. 다시 말하면 DTW는 화자종속적인 소규모 어휘의 음성인식에 적합한 방식이다. 이와같은 단점을 해결하기 위한 하나의 방법으로 등장한 것이

HMM(Hidden Markov Model)에 의한 음성인식 기법이다.

HMM에 의한 음성인식기법은 1974년과 1976년에 IBM Watson 연구소의 Baker와 Jelinek에 의해 각각 독립적으로 제안되었다. 이 방식은 이미 Harpy와 그것의 후속 시스팀인 Dragon system의 음향레벨 모델로 사용된 바 있는 Markov model을 기본으로 하고 기존의 Baum-Welch 알고리즘과 Viterbi decoding 알고리즘을 각각 re-estimation과 search 목적으로 사용한 것이었다. 그러나 HMM 방식이 급속히 보편화되고 실용화에 가까이까지 이르는 성공을 거두게 된 데에는 그 무렵부터 Gray 등에 의해 연구가 되기 시작한 VQ(Vector Quantization) 기법의 공로를 빼 놓을 수 없다. VQ 기법을 이용함으로써 segmentation과 labelling 과정없이 음성을 효과적으로 모델링할 수 있을 뿐 아니라 계산량이 크게 줄어들 수 있었다. Bell 연구소가 VQ 기법을 HMM 방식에 적용한 이래 많은 HMM 방식의 인식 시스팀이 개발되었으며 현재는 가장 각광받는 음성인식방식의 하나로 되어 있다. HMM 방식은 일종의 template matching 방식이지만 화자독립적인 인식이 가능하며 단어나 문장뿐 아니라 음소와 같은 subword unit를 기본 단위로 할 수도 있어서 대규모 어휘의 인식에 적합할 뿐 아니라 인식률도 매우 높다.

한편으로 인공지능기술에 대한 회의와 더불어 부활된 인공신경망(Artificial Neural Net) 기술을 음성인식에도 적용하고자 하는 노력이 일어났다. 그 중의 하나가 running spectrogram을 multi-layer perceptron에 입력시키고 각 hidden layer에서 일종의 time normalization을 차례로 행함으로써 output layer에서는 하나의 음소로 mapping되게 하는 TDNN(Time-Delay Neural Network) 기법이다. TDNN 기법에 의하면 비교적 작은 규모의 신경망 회로로써 상당히 높은 인식률을 얻을 수 있다. TDNN 방식이 음성인식을 위한 인공신경망 중에서 비교적 규모 또는 계산량이 적은 것은 인식대상을 음소로 하기 때문인데 그렇기 때문에 연속어 인식에 있어 유리할 것으로 생각된다. 인공신경망 방식은 아직 성능에 있어 HMM을 능가한다고할 수는 없으나 HMM과 더불어 현재 가장 활발하게 연구되고 있는 음성인식 방식이다. 최근에는 인공신경망 기법을 기존의 인식방식인 DTW나 HMM과 결합시키거나 다른 인식방식의 subsystem으로 사용하려는 연구가 많이 진행되고 있다.

앞에서 설명한 DTW, HMM, TDNN 방식 외에도 음향음성학적 지식을 이용하는 spectrogram reading 또는 SUR 프로젝트 당시에도 사용되었던 가장 오랜 음성인식 기법이라고 할 수 있는 segmentation & labelling 방식도 소수의 연구자들에 의해 아직 꾸준히 연구되고 있다. 그들은 이들 방식이 비록 현재 DTW나 HMM의 인식률에 눌려 성능면에서 열세를 면치 못하고 있지만 장차 연속어인식 시대에 진가를 발휘할 것을 기대하면서 성능을 개량하는 연구를 계속하고 있다.

이상과 같이 현재의 음향레벨에서의 음성인식기법이 ARPA SUR 프로젝트가 시작되었던 당시와 비교해 괄목할만한 진전과 성공을 거두게 되자 ARPA의 후신인 DARPA(Defence Research Projects Agency)의 Spoken Language System 연구분과에서는 SUR 프로젝트가 종료된지 10년이 지난 1985년부터 새로이 Strategic Computing Speech Recognition Program이라는 이름으로 제2차 음성이해시스팀 연구 프로젝트를 시작하였다. 이 프로젝트의 장기적인 목표는 1만 단어의 어휘를 대상으로 한 연속음성 인식이며, 단기적으로는 1000 단어의 어휘를 갖는 연속음성을 인식하되 95% 이상의 단어인식률을 얻는 것을 목표로 하고 있다. 특히 이 프로젝트에서는 패턴인식에 기초를

둔 음성인식 방식과는 달리 음향음성학, 구문론 및 의미론 등에 대한 구체적인 지식에 근거를 둔 음성인식 및 이해 연구도 병행하여 진행하고 있는데, 음성인식 시스팀의 개발뿐 아니라 관련된 세부 항목 기술, 그리고 개발환경의 조성에도 큰 비중을 두어 추진 중에 있다. 여기에는 CMU, BBN, MIT, TI 및 SRI 등의 연구기관이 참여하여 분야별로 그룹을 지어 연구를 수행하고 있다.

이 프로젝트의 첫과제는 미해군의 자료관리업무를 위한 1000단어의 연속음성을 인식하는 시스팀을 목표로 하였는데 그 목적으로 구축된 Resource Management Speech Date Base로써 시험한 결과에 의하면 대개의 시스팀들이 화자독립으로 약6%의 인식오율을 보였고 그 중에서 CMU(Carnegie-Mellon University)의 인식시스팀이 인식오율 4.3%로서 최고의 성능을 가진 것으로 나타났다고 한다.

DARPA SLS 연구분과는 1990년에 Air Travel Information System이라는 새로운 공동연구과제를 부과하여 현재 활발한 연구가 진행되고 있고 좋은 중간결과들이 보고되고 있다.

이제까지 기술한 바와 같이 외국의 음성인식기술은 1970년 초부터 본격적으로 연구되어 수십년의 역사를 가지고 있으며 그동안 많은 알고리즘과 인식방법들이 제안되고, 높은 인식률을 갖는 연구결과가 발표되었다. 특히, 음성인식 연구를 활발히 진행하고 있는 미국, 일본, 프랑스 등에서는 특정영역에서 이용할 수 있는 인식시스팀이 일부 개발되어 사용되고 있으며 전화망을 이용한 자동통역 전화시스팀의 개발에도 어느 정도 진전을 보이고 있는 것으로 보고되고 있다. 그러나, 아직까지는 완전한 인식시스팀을 구현하는 데 많은 문제점이 있기 때문에 한편으로는 제한된 문법규칙을 적용하고, 특정영역에서만 사용할 수 있는 제한적인 음성인식 시스팀이나 음성이해 시스팀을 개발하는 데 역점을 두고 연구가 진행되고 있는 실정으로서 좋은 결과가 나오고 있다.

국내의 음성인식에 대한 연구는 외국에 비해 10년 뒤인 1980년대 초부터 본격적으로 이루어졌으며 일부 대학이나 연구소에서만 수행되던 연구도 이제는 많은 대학에 확산되어 추진되고 있다. 연구내용은 초기에 대부분 DP 방법을 적용하여 특정화자에 의한 단독 숫자음, 지역명, 전화번호 등을 대상으로 하는 단어단위 인식이었으나 최근에는 연속 숫자음, 문장 등을 대상어휘 준패턴의 기본단위로 선정하였다면 1,000개의 기준 단위가 필요하지만 음소로 선정하였다면 40로 하여 HMM이나 신경회로망을 이용한 음소단위의 음성인식이 주를 이루고 있는 상황이다. 음성인식 알고리즘으로서는 초기의 DP 방법에서 부터 VQ 방법, HMM 방법 그리고 최근에는 NN 방법을 이용한 인식 실험결과가 발표되고 있어 10여년의 역사로 볼 때 주목할 만한 발전을 이룩하였다고 볼 수 있다. 특히 최근의 연구동향을 살펴보면 '89년도에는 VQ 와 HMM, DP 방법을 이용한 음성인식 연구가 거의 비슷하게 진행되었다가 '90년도에 들어와서는 HMM 방법을 이용한 연구가 과반수를 차지하고, 상대적으로 VQ 에 의한 연구가 감소되었다. '91년도에는 DP 와 VQ 방법에 의한 연구가 현저히 감소되면서 NN 방법을 이용한 연구가 HMM 을 이용한 연구와 동등하게 수행되고 있어 외국의 연구동향과 비슷한 경향을 보이고 있음을 알 수 있다. 또, 인식 대상어휘가 소규모에서 중규모로, 화자도 특정화자에서 불특정화자로, 인식음성도 고립어에서 연결어 또는 연속어 등으로 확장되어 연구되면서 외국의 음성인식 기술수준과 별 차이없이 연구되고 있다.

Ⅲ. 음성인식 알고리즘

1. 음성인식의 원리

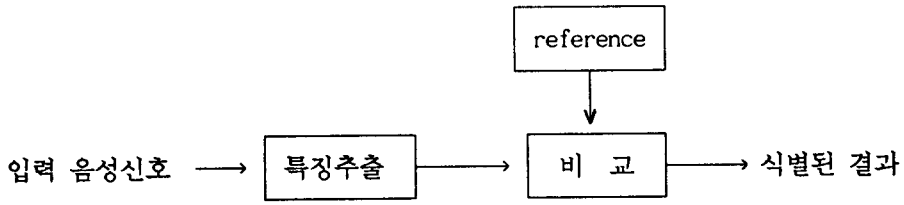

그림 1. 음성인식과정의 개념도

Reference로서 기준음성의 특징(feature) 데이타가 사용되면 template matching 방식, 음향음성학적 규칙이 사용되면 rule-based 방식이라 한다.

비교과정에서 미리 음성을 음소와 같은 작은 단위로 분할한 다음 각각을 식별하는 방식을 segmentation and labelling 방식이라 한다. 이 경우의 reference는 음소단위가 된다.

(그림 1)의 음성인식과정에 있어 가장 중요한 4가지 개념을 들면 다음과 같다.

o 특징추출(feature extraction)

선형예측(LP), homomorphic analysis, filter bank 등 각종 스펙트럼 분석법을 사용하여 음성신호에서 피치와 에너지정보가 제거된 스펙트럼을 나타내는 파라미터들을 특징으로 추출한다. 특징 파라미터로서 선형예측에 의한 반사계수, PARCOR 계수, LSP(Line Spectrum Pair), LAR(Log Area Ratio), LPC 켑스트럼 등이 있고 homomorphic analysis에 의해 얻어지는 켑스트럼(cepstrum)과 filter bank의 filter 에너지들이 있으나 그 중에서 음성인식시 인식률을 높게 해주는 특징파라미터로서 가장 각광을 받고 있는 것은 켑스트럼 계열의 파라미터들이다.

o 시간적 왜곡 제거(time alignment)

발성시의 지속시간의 차이를 극복해야 하며 그 목적으로 segmentation, dynamic time warping, hidden Markov model이 사용된다.

o 스펙트럼 왜곡 측정(distance measure)

음성특징 파라미터들을 가지고 스펙트럼의 유사도(similarity)를 측정할 수 있어야 하며 그 목적으로 Euclidean distance, LPC distortion measure 등이 사용된다.

o 사전 참조(lexical access)

인식대상을 예측함으로써 인식률을 높일 수 있다. 예측의 가장 간단한 방법은 정해진 어휘로 미리 구성된 사전(lexicon)을 인식시 참조함으로써 인식대상을 제한하는 것이다.

2. 음성인식 알고리즘의 종류

음향처리(acoustic processing and phonetic analysis) 레벨에서의 음성인식 알고리즘은 크게 knowledge-based 방식과 template matching 방식으로 나누어 볼 수 있다.

(A) Knowledge-Based 방식

Rule-based 방식이라고도 불리며 음향음성학적 지식에 의해 segmentation과 labelling을 함으로써 음소를 식별하여 음성을 인식하는 방식이다.

o Phonetic Classification

Broad phonetic class들의 특정 자질을 검출하는 property detector들의 출력을 종합함으로써 각 음소를 segmentation 및 labelling하여 음소를 인식하는 방법이다.

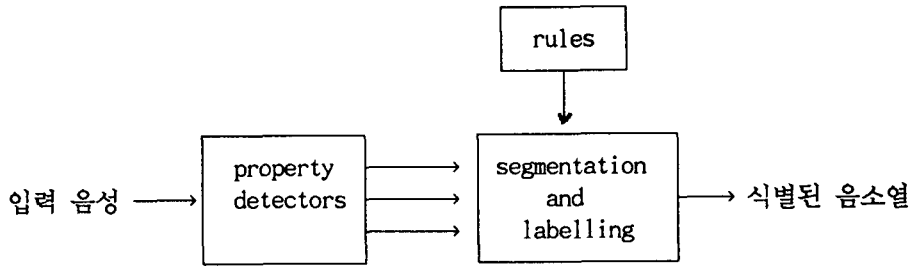

그림 2. Phonetic classification방식의 개념도

o Spectrogram Reading

음성 spectrogram을 직접 해독함으로써 각 음소를 segmentation 및 labelling하여 음소를 인식하는 방법이다.

Spectrogram을 해독하려는 시도는 speech spectrograph가 개발된 1946년부터 있었다. 예컨대 1947년에 Potter, Kopp, Green이, 1973년에 Klatt와 Stevens가, 1974년에 Svensson에 의해 시도되었으나 성공적인 결과를 얻지 못하였다. 그래서 Fant, Liberman, Lindblom, Svensson 등은 spectrogram을 해독하는 것은 불가능하다고 생각하였다. 그러나 1968년 러시아의 Lev Rubin은 부분적이나마 연속된 음소들을 음성 spectrogram으로부터 읽어낼 수 있었다.

본격적인 spectrogram reading은 1971년 MIT의 Victor Zue로부터 시작되었다. 1971년부터 현재까지 Zue는 매일 평균 0.5~1시간씩, 총 2000~2500 시간을 음성 spectrogram reading에 투입함으로써 80~90%의 음소인식률을 얻고 있다. Zue는 spectrogram reading으로 얻은 지식을 가지고 1986년에 expert spectrogram reader라는 음성인식용 expert system을 구축하였는데, 다음과 같은 간단한 형태의 rule들을 사용하였다.

```
If the VOT is short,
and the following vowel is not a schwa,
then the stop is voiced.

If there is prevoicing during closure,
then the stop is voiced.
                .
                .
                .
If the voicing of the stop is voiceless,
and the place of articulation of the stop is alveolar,
then the identity of the stop is /t/.
                .
                .
                .
```

Expert spectrogram reader의 성능은 (표1)에서 보는 바와 같이 음소인식률이 training 단계에서는 88%, testing 단계에서는 84%였으며 second candidate 음소까지 포함시킬 경우에는 각각 95%와 92%에 달하였다.

labelling 과정이 필요없는 다음과 같은 인식방식들이 각광을 받게 되었다.

o Vector Quantization 방식

음성코딩용으로 개발된 VQ 기술을 이용하여 음성을 식별하는 방식이다. Vector 양자화는 scalar 양자화를 일반화시킨 양자화 기법으로서 음성 스펙트럼을 분류하는 데 매우 적합하다.

VQ를 이용한 음성인식의 원리를 간단하게 설명한다면 utterance가 정적인 스펙트럼을 가진 것으로 보아 미리 저장해둔 특징 벡터중에서 입력음성의 특징 벡터와 가장 가까운 하나의 벡터를 선정하는 방법이다.

단순한 VQ 인식에서는 시간적 정보가 포함되어 있지 않아서 음향적 특성이 유사한 단어들 사이에 부정확한 인식이 일어난다. 따라서, 한 단어를 발성 순서에 따라 몇개의 구간(section)으로 나누고, 구간 별로 독립된 코드북을 작성함으로써 시간적 정보를 포함시키는 MSVQ(Multi-Section VQ) 방법이 Burton 등에 의해 제안되었다.

o Dynamic Time Warping 방식

음성의 지속시간의 차이로 인한 영향을 제거하기 위하여 입력음성과 기준 음성의 양끝을 서로 맞추고 선형적으로 늘이거나 줄여 pattern을 비교함으로써 최적 기준음성을 찾아내는 linear time alignment 방식의 성능을 개선하기 위해 음성의 각 부분을 늘이거나 줄여 비교하는 dynamic time alignment 방식이다. 이와같은 warping 함수를 찾는 가장 간단한 방법은 (그림 3)과 같이 일정 지속시간의 frame으로 분할한 다음 각 frame의 스펙트럼을 기준 음성의 각 frame과 모두 비교하여 total distance가 최저인 최적경로를 찾는 것이다. 이와같은 비교를 모든 기준음성에 대해 행하고 그 중에서 total distance가 최저인 기준음성을 인식된 음성으로 택하는 방법이다.

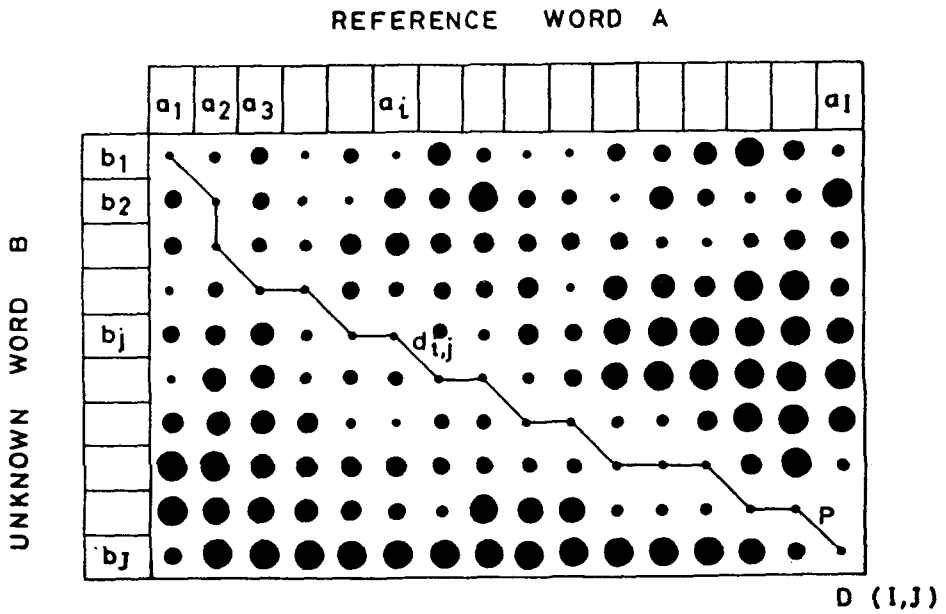

그림 3. Full search에 의한 DTW 방식의 개념도

표 1. MIT의 expert spectrogram reader와 인간과의 음소인식 성능비교

condition		first choice accuracy	top 2 choice accuracy
training	human	90	92
	system	88	95
testing	human	92	96
	system	84	92

MIT의 spectrogram reading expert 시스팀 이전에도 유사한 방식의 시스팀들이 개발된 적이 있는데, 1983년 미국 Verbex사의 Johannsen 등이 개발한 SPEX(Speech Spectrogram Expert) 시스팀이나 1984년에 Leicester Polytechnic의 Johnson 등이 개발한 knowledge-based spectrogram analysis 시스팀이 그것이다. 이에 비해, 일본 Osaka 대학의 Mizoguchi교수가 개발한 SPREX(speech recognition expert) 시스팀은 spectrogram으로부터 formant의 변화에 관한 동적인 데이터를 추출하여 음소를 인식하는 조금 색다른 방법을 사용한 것이었다.

SPREX에서는 입력음성으로부터 ZCR(Zero Crosssing Rate)을 구하고 선형예측(LPC) 분석하여 F_1과 F_2의 증가와 감소, formant의 출현과 소실, noise gap 등의 데이터를 얻은 다음 rule을 적용하여 인식하였다. Rule은 안정된 frame 검출에 10개, phone 인식과 재인식에 35개와 12개, segmentation과 음소인식에 38개와 24개가 각각 사용되었으며, 음소 90%(모음 94%, 자음 85%), 단어 87%, 문장 75%의 인식률을 얻었다.

Spectrogram reading 방식의 난점은 knowledge(즉 rule)의 획득이 어렵고 spectrogram으로부터 필요한 knowledge feature의 추출이 어렵다는 점이다. Spectrogram reading에 필요한 feature를 얻기 위해서 phonetic classification 에서와 같은 property detection 방식이 주로 사용되어왔으나, 최근(1991년) Hemdel과 Lougheed는 영상처리기법을 적용하여 spectrogram으로부터 phonetic feature를 직접 추출하는 방법을 시도하였다.

한편 time-frequency 분석방법도 spectrogram 외에 peripheral auditory model에 의한 GSD(Generalized Synchrony Detector), Wigner distribution, wavelet transform 등으로 다양화되고 있다.

(B) Template Matching 방식

일종의 pattern matching 방식으로서, 공학적인 접근법이라고 할 수 있다.

o Segmentation & Labelling 방식

ARPA의 SUR 프로젝트에서도 사용되었던 고전적인 방식이다. 흔히 segmentation 용으로는 manner cue, labelling 용으로는 place cue가 이용된다. 예컨대 SUR 프로젝트에서는 대개 segmentation용으로 ZCR, 에너지 등 간단한 소위 ZAPDASH(zero crossings and peaks in smoothed and differenced waveforms) 파라미터를 사용하였고, segmentation된 음성을 labelling하기 위해서는 LPC계수를 사용하여 미리 준비된 기준 segment와 비교하였다.

Segmentation & labelling 방식의 단점은 segmentation 및 labelling시 한번 오류가 생기면 후속처리과정에 전파되어 결정적으로 작용한다는 점이다. 그 때문에 segmentation과

(그림 3)의 방법은 너무 많은 계산을 요하므로 dynamic pattern matching이라고도 하는 DP(Dynamic Programming) 기법을 사용하여 계산량을 줄이는 방법이 보편적으로 쓰이고 있다. DP방식의 DTW에 있어서는 계산을 줄이기 위해 (그림 4) 또는 (그림 5)와 같이 경로를 제한하며 매 frame에서 (그림 6)과 같은 local constraint를 받으면서 경로를 찾아나가면 최종적으로 (그림 7)과 같이 최적경로를 찾게 된다.

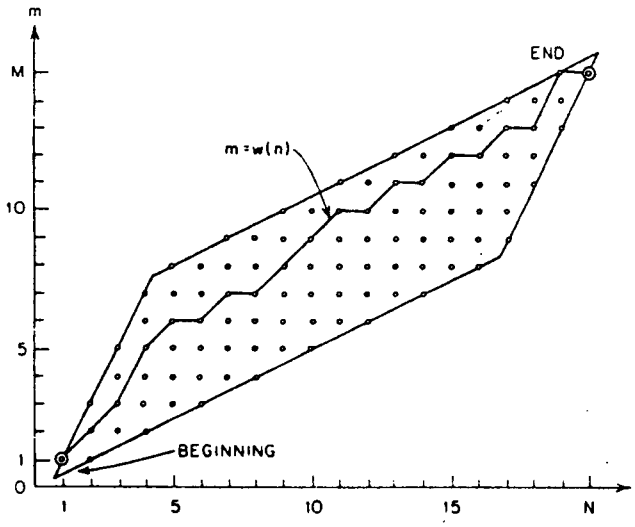

그림 4. DTW에서의 계산량 감소를 위한 영역제한 방법

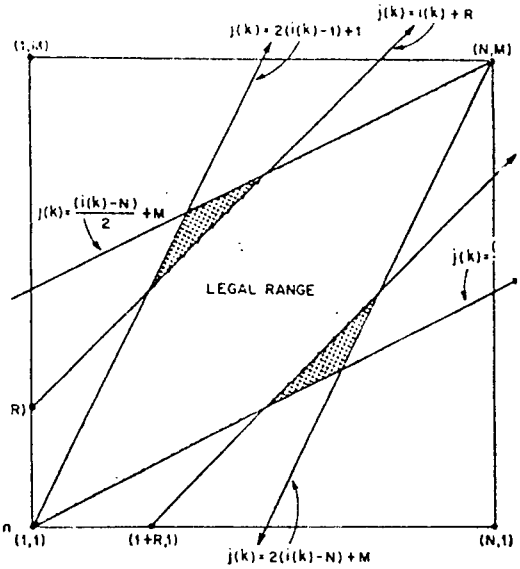

그림 5. DTW에서의 계산량 감소를 위한 영역제한 방법

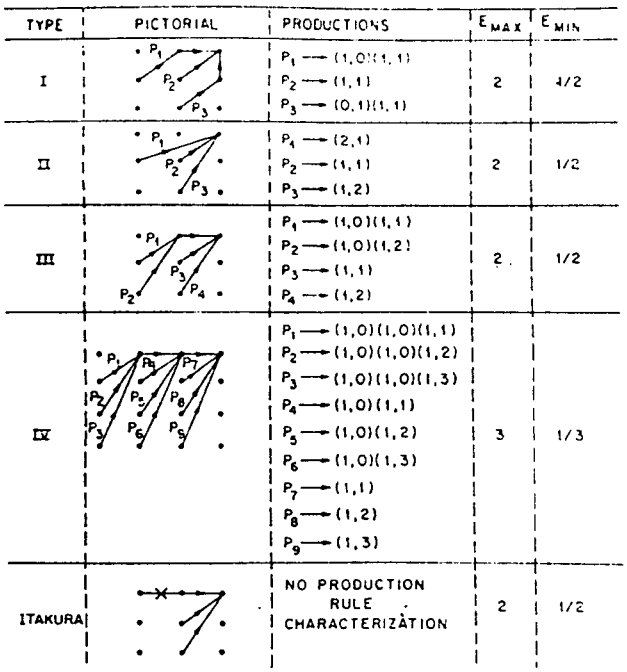

그림 6. DTW에서의 계산량 감소를 위한 local path constraint

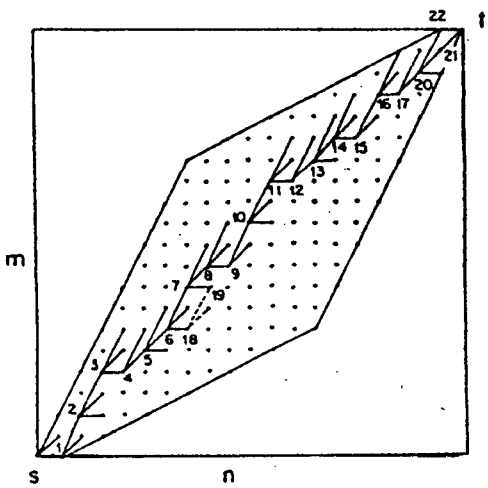

그림 7. DP에 의한 DTW의 개념도

DP를 이용하여 연결음성을 인식할 때 발생하는 문제점을 극복하기 위한 방법으로 DP를 변형한 LB(Level-Building) DTW와 OS(One-Stage) DP 방법이 있다.

LB DTW 알고리즘은 입력단어수를 지정할 수 있으며 계산량을 줄이기 위해 여러가지의 DP range 감축을 시도할 수 있는 장점이 있는 방법이다.

OSDP 알고리즘은 표준패턴을 몇개로 구분하고 입력패턴이 가장 잘 매칭되는 표준패턴을 결정하는 기법이다. 즉 입력패턴과 미지의 template 열을 가장 잘 매칭시키는 warping 경로를 찾는다.

DTW알고리즘은 고립단어 인식에 주로 이용되며 어휘가 작고 인식 시간이 많이 소요된다는 단점이 있지만 인식률이 높기 때문에 VLSI chip으로 구현되어 현재 널리 상용화되어 있다. 또한 기준패턴을 쉽게 만들 수 있기 때문에 음성인식시스팀의 업무내용을 사용자의 요구에 따라 용이하게 변경할 수 있다.

o Dynamic Interpolation 방식

(그림 8)은 DTW 방식에 의해 입력음성과 template 음성간의 distance를 구하는 방법을 보여주며 그 결과를 (그림 9)와 같이 두 스펙트럼 파라미터 m_1, m_2로 구성되는 phonotopic map 상에 표시하였다. DTW 방식에서는 모든 template마다 warping 함수를 계산해 내야 하기 때문에 계산량이 많다. 1978년 Ruske와 Schotola에 의해 개발된 Dynamic Interpolation 방식에서는 각 패턴마다 일정한 수의 interpolation point 만을 만든다. 예컨대 (그림 10)에서는 각 패턴을 9개의 등거리 부분들로 나누고 9개의 interpolation point를 구하여 거리를 계산함으로써 스펙트럼이 적게 변하는 부분보다 급히 변하는 부분이 강조된다. 이 방식의 장점은 모든 기준패턴에 있어 interpolation을 한번만 수행하여 정규화로 된 형태로 저장하면 되므로 계산량이 적다는 점이다. Dynamic interpolation 방식의 음성인식 성능은 dynamic time warping 방식과 거의 같은 것으로 알려져 있다.

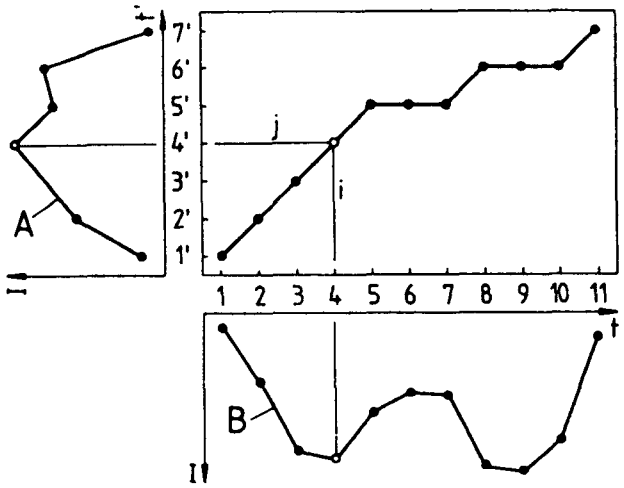

그림 8. DTW에 의해 구해진 time warping function

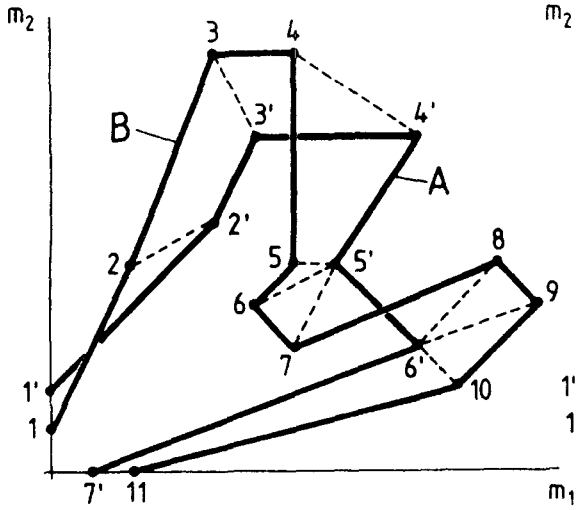

그림 9. DTW의 phonotopic map 상의 도시

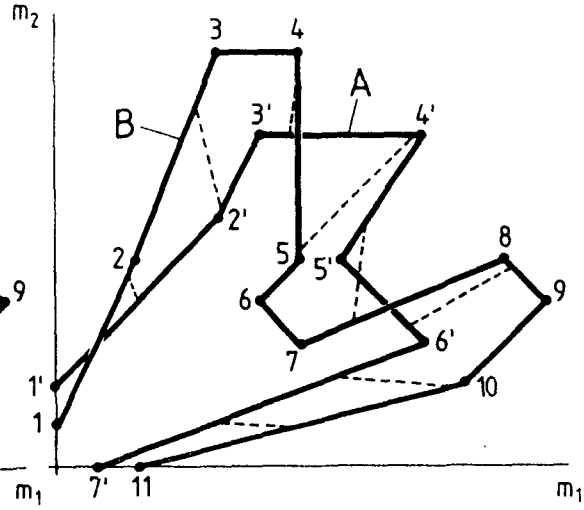

그림 10. Dynamic interpolation 방식의 phonotopic map 상의 도시

o Hidden Markov Model 방식

HMM 방식의 기본적인 개념은 음성이 Markov 모델로 모델링될 수 있다는 가정하에 훈련과정에서 Markov 모델의 확률적인 파라미터를 구하여 기준 Markov 모델을 만들고 인식과정에서는 입력 음성과 가장 유사한 기준 Markov 모델을 추정해냄으로써 인식한다는 것이다. 모델로서 hidden Markov 모델을 사용하는데 그것은 음성패턴의 다양한 변화를 수용하기 위해서이다. 여기서 hidden 이란 용어는 state가 음성패턴에 관계없이 모델속에 숨어있다는 것을 뜻한다.

HMM은 관측이 불가능한 한 process를 관측이 가능한 심볼로 발생시키는 process를 통하여 추정하는 이중의 확률 process이다. 그러므로, 음성과 같이 다변성이 많고, 발생 과정을 알 수 없는 process들을 표현하는 데 적당한 모델링 방법이다.

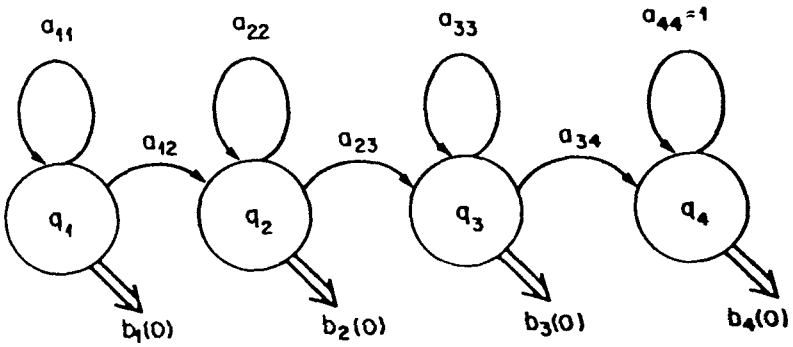

그림 11. Hidden Markov model의 개념도

(그림 11)에서 보는 바와 같이 HMM은 천이들에 의해 서로 연결된 상태들의 모임으로서 각 천이에는 2가지 종류의 확률이 관련되어 있다. 하나는 상태의 선정에 관한 천이 확률이고, 또 하나는 천이가 이루어졌을 때 유한개의 관측 대상으로부터 각 출력 심벌이 방출되는 조건부 확률을 규정하는 출력 확률 밀도 함수(output pdf)이다. 즉 음성패턴의 각 특징을 state의 천이확률과 출력확률로 표현하여 준다.

HMM은 두가지의 가정을 기본으로 하고 있는데 첫번째 가정은 현재의 상태는 그 바로 이전의 상태에만 의존한다는 것이고 두번째 가정은 출력이 독립적이라는 것이다.

HMM을 이용하여 음성인식을 하고자 할 때 평가(evaluation), decoding, 학습(learning)의 세 가지 문제점을 해결하여야 하는데, 첫번째 평가 문제는 모델과 관측열이 주어졌을 때 관측열의 확률을 계산하는 법으로서 forward 알고리즘을 이용한다.

두번째 문제는 인식과정에 필요한 decoding에 관한 것으로 관측열이 주어졌을 때 최적인 상태열을 선택하는 문제인데 이는 Viterbi 알고리즘을 이용한다. 세번째 문제는 훈련(training) 과정에 필요한 학습에 관한 문제로 최적의 모델링을 하기 위해 각 파라미터를 조정하는 문제인데 이는 Baum-Welch의 재평가(reestimation) 알고리즘으로 해결한다.

HMM에는 정적인 특징 파라미터만 이용하는 일반적인 HMM 인식방법 외에도 다중 특징을 이용한 D(Dynamic)HMM 인식방법이 있다. DHMM은 정적 스펙트럼의 특징 파라미터와 동적 스펙트럼의 특징 파라미터를 함께 모델링한 것이다. DHMM에서는 정적 특징과 동적 특징 파라미터를 모두 이용하는데 정적 특징과 동적 특징사이의 상관 관계가 매우 적다는 사실에 의해 파라미터의 크기에 따른 계산량의 증가가 그다지 없이 모델링될 수 있으며 인식 알고리즘은 정적 특징 및 동적 특징이 조합된 점만 제외하면 HMM에 의한 인식과 동일하다.

HMM 알고리즘은 1970년 말부터 음성인식 알고리즘으로 많이 사용되었고 최근에는 높은 인식률과 빠른 인식시간때문에 대용량 음성인식시스팀에 널리 적용되고 있다. HMM 방식에서는 기준패턴을 단어나 문장과 같은 단위로 설정할 수 있음은 물론이지만 음소, 음절과 같이 단어 이하의 발음 길이를 갖는 패턴으로 설정하여도 입력음성으로 단어, 문장들을 인식할 수 있으므로 대용량 음성인식시스팀에 주로 이용된다. 예컨대 만약 1,000단어 음성인식시스팀에서 단어를 기준패턴의 기본단위로 선정하였다면 1,000개의 기준 단위가 필요하지만 음소로 선정하였다면 40~50개의 음소의 파라미터만 저장하면 된다.

HMM 모델의 훈련에 필요한 노력은 DTW 방식에서의 template 구축에 비하여 대단히 크지만 반대로 인식시의 계산량은 훨씬 적으며, 시간과 스펙트럼의 양쪽에 확률 및 통계적인 방법을 사용함으로써 averaging 또는 clustering 효과를 얻을 수 있어 inter-speaker 및 intra-speaker variability를 수용하기가 용이한 장점이 있다.

HMM을 실제 음성인식시스팀에 적용하기 위해서는 smoothing, scaling, corrective training 등을 고려하여야 한다.

HMM은 segmentation과 labelling을 피할 수 있는 방법으로 도입되었으나 현재 이 방법은 자동 segmentation 및 labelling 용으로 이용되기도 한다.

o Neural Network 방식

인공신경망의 개념은 뇌신경세포의 구조와 동작을 모방함으로써 인식, 연상, 논리, 판단, 분류와 같은 뇌의 기능을 흉내낼 수 있다는 것이다. (그림 12)에 가장 간단한 형태의 인공신경망인 single-layer perceptron의 구조를 보였다. Single-layer perceptron으로는 선형분리만 가능하기 때문에 (그림 13)과 같이 그것들을 여러층으로 포갠 형태의 multi-layer perceptron이 많이 사용되고 있다.

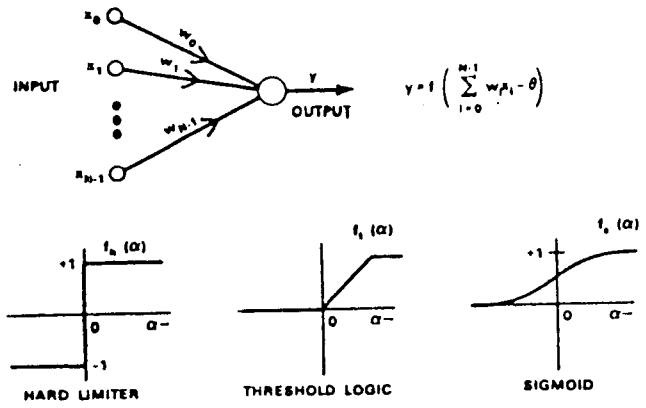

그림 12. Single-layer perceptron의 개념도

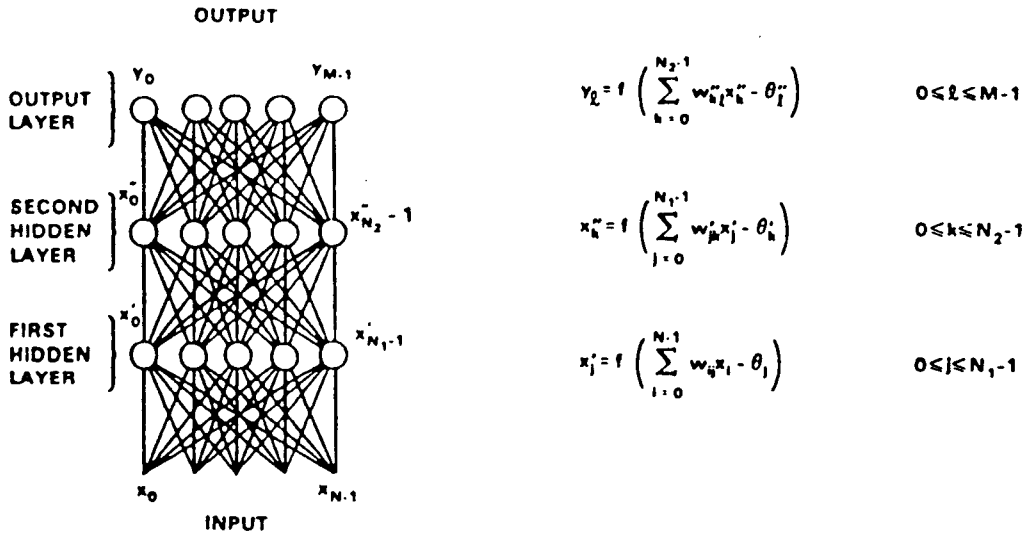

그림 13. Multi-layer perceptron의 구조

Perceptron에서의 weight는 훈련과정을 통하여 구해진다. 즉 훈련할 음성의 특징이 입력데이터로, 그 때의 음성의 의미가 출력데이터로 주어지면 학습 알고리즘에 의해 모든 weight들의 값이 구해지며, 훈련을 거듭할 수록 weight 값들은 더욱 정확성을 가지게 된다. Multi-layer perceptron의 훈련에는 error back propagation이라는 알고리즘이 널리 사용되고 있다.

신경망을 음성인식에 이용한 초기에는 프레임 단위로 분할된 음성을 음소로 분류하는 등 시간적 변화를 고려하지 못하였으나 TDNN(Time Delay Neural Network)의 출현으로 동적인 패턴을 검출할 수 있게 되었다.

TDNN의 입력은 (그림 14)에 보인 바와 같이 filter bank로부터 출력된 10ms 단위로 구성된 running spectrogram이다. (그림 15)에 나타낸 바와 같이 그것들은 2개의 hidden layer를 거치면서 일종의 time normalization이 행해져서 output layer에서는 하나의 음소로 mapping된다.

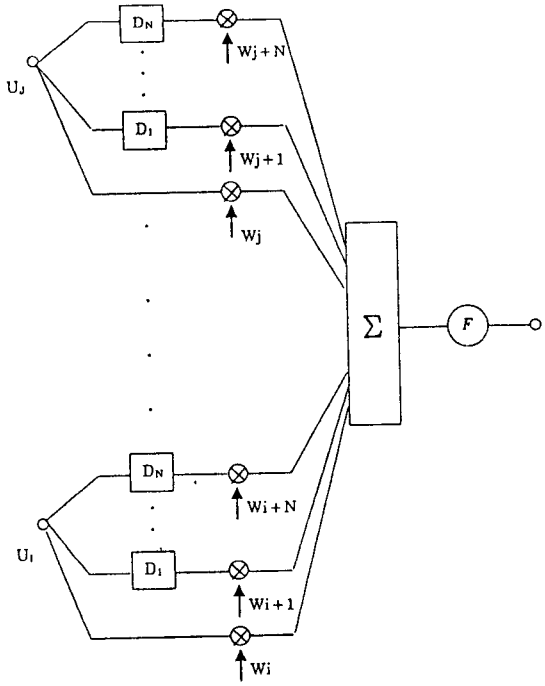

그림 14. TDNN의 입력부

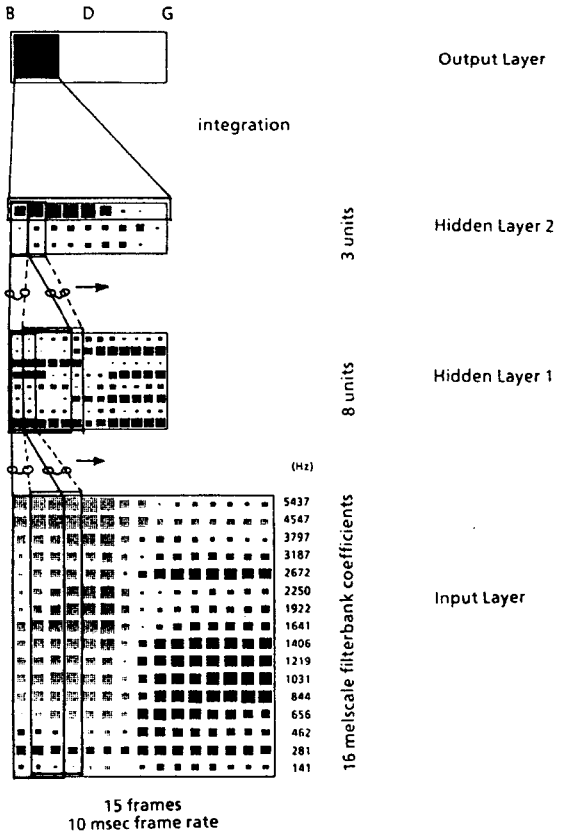

그림 15. TDNN의 개념도

TSNN(Time-State Neural Network)은 TDNN과 비슷하나 running spectrogram을 시간축 방향으로 4개의 state로 분할하고 음소의 세부구조를 전담하여 인식시킴으로써 TDNN보다 높은 인식률을 얻기 위해 고안된 방법이다.

DPNN(Dynamic Programming Neural Network)은 시간적 distortion과 스펙트럼 distortion을 각각 DP와 신경망을 이용하여 처리하는 방법으로서 화자독립 인식시스팀에 적합하다.

이상의 방법들은 훈련데이타와 그것의 의미를 사용하여 훈련시키는 supervised learning neural net이었다. 또다른 supervised learning 알고리즘으로 Kohonen이 제안한 LVQ(Learning Vector Quantization)가 있는데 이것은 뇌의 특성을 VQ에 적용한 것으로서 음성인식시스팀에 사용되어 높은 인식률을 보였다.

Supervised learning과 반대로 훈련시 훈련데이타의 의미를 제공하지 않는 unsupervised learning 알고리즘의 대표적인 것으로 Kohonen의 self-organizing feature map이 있다. Feature map 알고리즘은 음성데이터를 의미에 관계없이 훈련시키면 요소를 대표할 수 있는 특징이 저절로 형성된다는 것이다. 이 알고리즘은 실제 인간의 청각작용과 유사하나 음성인식시스팀에 적용하였을 경우 supervised learning 알고리즘 보다는 낮은 인식률을 보이는 것으로 알려져 있다.

Ⅳ. 음성인식시스팀 개발현황

여기에서는 전절에서 설명한 음성인식 알고리즘을 이용하여 구현된 음성인식시스팀들을 소개한다. 현재의 음성인식기술은 과거에 비하여 획기적인 발전을 이룩하였으나 현재 부분적으로 상용화되어 있는 시스팀들은 대부분 특정영역의 제한된 조건에서 사용되고 있는 실정이다.

음성인식시스팀은 입력음성이 고립어, 연결어, 연속어인가에 따라, 화자종속인가 화자독립인가에 따라, 또한 어휘의 규모에 따라 구분되며 고립어보다 연속어 쪽이, 화자종속보다 화자독립 쪽이, 또한 어휘가 커질수록 인식이 어려워진다.

1. 외국의 상용 음성인식시스팀 현황

외국의 상용 음성인식시스팀들은 DTW 알고리즘을 기반으로 하였으며 화자종속 시스팀들이 대부분인 것이 특징이다. 극히 최근에 와서 HMM을 기반으로 한 상용시스팀이 등장하고 있다.

표 2. 외국의 상용 음성인식 시스팀

제 조 업 체	국명	시스팀명	특 징	단어수	인식율 (%)
AUDEC	미국	SSB-1000 vodialer	화자종속고립단어 화자종속고립단어	144 48	95 95
Dragon System	미국	Voice Scribe 1000 Dragon Dictate	화자종속고립단어 화자적응고립단어	1000 30,000	
IBM	미국	Voice command	화자종속고립단어	64	95-98
INFOVOX AB	스웨덴	RA 201/PC	화자종속고립단어	179	
INTEL	미국	IWS	화자종속고립단어	200	
Interstate Voice Products	미국	Vocaline SRL/LC Vocaline CSRB	화자종속연결단어 화자종속연속음성	400 100	98 99
Kurzweill Applied Intelligence	미국		화자종속고립단어	1000	
NEC America		SAR-10 SR-10 DP 200	화자종속고립단어 화자종속고립단어 화자종속연결단어	250 128 150	98 98
Speech Systems Inc.	미국	Phonetic Engine -200 Phonetic engine -400	화자종속연속음성 화자독립연속음성	20,000 40,000	90 97
Texas Instruments	미국	Speech Command	화자종속연속음성	1000	
Voice Connection		Introvoice V	화자종속연속음성	400	98
Voice Control System	미국	VC 1000	화자독립연속음성	40	98.5
VOTAN	미국	Voice-Card Voice-Card Voice Key	화자독립연속음성 화자종속연속음성 화자종속고립단어	13 640 64	93 98

2. 외국의 연구용 음성인식시스템 현황

외국의 연구용 음성인식시스템들은 HMM 방식을 기반으로 한 대용량 연속음성인식시스템들이 주축을 이루고 있다.

표 3. 외국의 연구용 음성인식시스템

연구기관	국 명	시스템명	특징	단어수	인식률 (%)
CMU	미국	SPHINX PHOENIX	화자 독립 연속 음성 자연 음성 이해	1000 100	96.2 80
BBN	미국	BYBLOS	화자 종속 연속 음성 화자 적용	1000	87.5 94.8
Lincoln Lab	미국		화자 독립 연속 음성	1000	87.4
ATR	일본		화자 종속 연속 음성 화자 적응	1035	88.4 81.6
IBM	미국	Tangora	화자 종속 고립 단어	20,000	95
NEC	일본		화자 종속 고립 단어	1,800	97.5
NTT	일본		화자 종속 연속 음성	500	83
SRI	미국	DECIPHER	화자 독립 연속 음성	1,000	97

3. 국내의 음성인식시스템 현황

국내에는 아직 상용화된 음성인식시스템은 나오지 않고 있고 주로 대학에서 시뮬레이션을 통해 알고리즘을 연구하고 있으며 연구용 음성인식시스템도 아직은 보고된 바 없다.

여기에서는 연구용 음성인식시스템으로 구축되어 있는 KARS(Korea Telecom Automatic Recognition System)에 대해 소개하기로 한다. KARS는 음성인식에 의한 정보검색 모델시스템으로서 KT Research Center 내의 116개 부서명을 전화기를 통해 입력시키면 부서소개, 위치안내 및 전화번호 안내를 해 주는 서비스를 목표로 하고 있다. 입력어휘의 규모는 제어용 단어 7개를 더한 123개이고 인식단위는 44개의 유사음소를 기본으로 하고 인식률을 개선하기 위해 좌우측 음소를 고려한 context-dependent unit에 대해서도 실험하였다. Discrete HMM을 기본 모델로 하였으며 계산량을 줄이기 위해 decoding 알고리즘으로 Viterbi beam search 방식을 사용하였다. 또한 인식률을 제고하기 위해 3가지 codebook을 사용하였고 음소의 duration 정보를 반영하였으며 훈련시 corrective training 과정을 추가 하였다.

KARS는 화자독립형 고립어 인식시스템으로서 8명의 음성으로 훈련시키고 그에 속하지 않는 2명의 음성을 전화기를 통해 입력하였을 때 (표 4)와 같은 인식률을 얻었다. 현재까지는 전화망을 이용하지는 않고 있으나 추후 전화망을 통한 실제의 전화음성을 인식하도록 할 계획이다.

표 4. KARS의 음성인식 성능 [%]

대 역	5kHz	전화대역 (3.3kHz)
top 1	98.6	98.17
top 2	99.6	99.8

V. 전망 및 결론

기계에 의한 음성인식이 실현된다면 그것의 응용가능성은 말할 것도 없이 실로 무한하다고 할 수 있을 것이다. 음성에 의한 기계의 제어는 물론이고 음성에 의한 데이터의 입력이라든가 자동통역 등의 분야에 응용됨으로써 사회생활 및 문화생활에 있어 혁신적 변화가 일어날 수 있을 것이다. 그 중에서도 최근 특히 관심을 끄는 것이 자동통역전화이다. 자동통역전화 시스템은 한 나라의 대화체 언어를 상대국의 언어로 통역하여 상대방의 언어를 모르더라도 자유롭게 전화를 통화할 수 있게 해주는 전화시스템이다. 앞으로 국제간의 교류가 빈번해짐에 따라 이러한 시스템이 상대방의 언어에 능숙하지 못한 많은 사람들의 관심을 끌 것은 당연하다.

자동통역전화시스템은 음성인식부, 기계번역부 그리고 음성합성부로 구성된다. 자동통역전화 시스템은 음성신호생성, 음성인식 및 합성, 음성재생 및 잡음제거, 음성신호처리, 기계번역, 그리고 컴퓨터와 통신 접속기술등 다양한 기술이 종합적으로 접목되어야 하는 시스템이지만, 그 중에서도 음성인식기술이 bottleneck으로 되어있다. 자동통역전화에 필요한 음성인식기술은 화자독립 대어휘 연속음성인식기술이어야 하므로 다음과 같은 문제점들이 해결되지 않으면 안된다.

첫째, 화자독립적인 음성인식을 위하여는 인간의 음성발생 특성을 감지하여 신속하게 적응할 수 있는 화자적응기술이 개발되어야 한다. 예컨대 남녀별 차이를 감지하는 데 기본주파수를 이용하는 것이 하나의 방법이 될 수 있을 것이다.

둘째, 대용량의 어휘를 인식하기 위해서는 음소인식방식이 적합하나 인식률이 낮고 template matching 방식은 인식률이 높으나 어휘가 커지면 훈련과정에 필요한 노력이 매우 커진다. 이러한 현상을 극복하기 위해 인식의 기준단위로서 단어 대신 음소와 같은 sub-word unit가 사용되고 있으나 조음현상을 반영하기가 단어보다 나쁘기 때문에 인식률이 저하된다. 따라서 조음결합현상을 반영하는 새로운 인식단위의 연구가 개발되어야 하며, 그 경우에는 단 한번의 훈련으로 족하게 될 것이다.

셋째, 연속음성을 인식하기 위해 상위계층인 언어처리 기능을 강화하는 연구가 필요하다. 문장을 인식하기 위해 영어와 같은 굴절어에서는 단어를 인식대상으로 삼을 수 있으나 우리말과 같은 교착어에서는 구를 기본 인식대상어로 삼아야 하는데 이 때문에 어휘의 규모가 엄청나게 증가된다. 이 경우 언어처리 알고리즘은 문장내에서의 단어(혹은 구)의 위치에 따라 인식대상 단어(혹은 구)를 한정함으로써 인식시 계산량을 줄이고 인식률을 높여 준다. HMM에서는 기준단위를 음소로 하더라도 단어, 구 혹은 문장단위로 조합하여 인식하게 되므로 어휘가 더욱 증가될 것이다. 그 경우에는 필요한 단어만을 선별하여 인식시킬 수 있는 word spotting 알고리즘을 이용하는 것이 하나의 해결방법이 될 수 있을 것이다.

넷째, 고속발성음성(fluent speech)에서 발생하는 제반 현상 - 예컨대 formant의 undershoot라든가 /ㄱ/ 또는 /ㅇ/ 음소의 smearing, 종성 /ㄷ/ 음가의 변질 등 - 에 대한 음성학적 연구결과가 인식시스템에 반영되어야 할 것이다.

다섯째, 주변 환경 및 배경 잡음에 의한 인식률 저하 문제를 해결하여야 한다. 이를 위해서는 주변잡음을 감쇠시킬 수 있는 전처리기술이나 주변잡음이 있는 음성을 효과적으로 인식할 수 있는 인식 알고리즘이 개발되어야 할 것이다.

완벽한 음성인식 시스템을 설계하기 위해서는 이 외에도 처리해야만 하는 요인들이 많이 남아 있다. 그러나, 이러한 문제점들을 해결하기 위한 노력들이 관련 연구자들에 의해 계속해서 이루어지고 있어 많은 발전이 있었으며 지금과 같은 속도로 발전한다면 앞으로 10년 이내에는 획기적인 해결책이 등장할 것으로 기대된다. 그러므로 2000년대 초에는 자동통역전화나 자동통역 컴퓨터가 등장하는 것도 불가능한 일만은 아닐 것이다.

그러나 한편 그 때를 기다리기만 하는 것보다는 현재까지 개발된 화자독립형 대어휘 고립어 음성인식기술을 적절한 응용분야에 적용하여 실용화하려는 노력도 중요하다. 그러한 일례로서 독일의 BMFT(Federal Minister of Reseach and Technology)가 제안한 Verbmobil을 들 수 있다.

Verbmobil은 한마디로 "상면(face-to-face) 대화를 위한 휴대용 번역 시스팀" 이다. Verbmobil의 최종목표는 물론 다른 언어를 사용하는 회의등 모임에서 사람들간의 대화를 상대방의 언어로 번역해 주는 휴대용 동시 번역기이지만 우선은 일종의 번역보조수단으로 개발하는 것이 검토되고 있다. 즉 2000년까지 두가지 제품생산을 목표로 하는 것인데 첫번째 제품은 서로 다른 언어를 사용하는 대화자 간에 극히 제한된 주제와 대화의 목적이 잘 정의된 상태에서 상면 대화 수단을 거의 실시간으로 제공하는 것이다. 현재 가장 유력한 응용분야는 매매협상(contract negotiation)으로 꼽히고 있다. 두번째 제품은 자기 모국어가 아닌 언어를 사용하는 두 대화자간에 모르는 단어나 구를 해석해 주는 것이다. 즉 외국어로 대화시 말문이 막힐 경우 버튼을 누르고 자기 모국어로 해당 단어나 구를 입력시키면 그것을 번역해 준다. 또한 대화중에 상대방이 한 말의 의미를 모를 때 해당 의미를 제공하는 것도 가능하다. 현재의 국내외 음성인식기술은 화자독립형 대어휘 고립어 인식시스팀의 실용화가 가능한 수준에 있으며, 연구실에서의 노력이 연속어인식에 집중되고 있어 조만간 제한적이지만 연속어 인식기술을 사용한 상용시스팀이 개발될 전망이다.

본 고에서는 최근 고조되고 있는 man-machine interface 기술의 핵심이 되는 음성인식기술을 방식별로 설명하고 국내외 음성인식의 연구동향을 살펴보았으며 음성인식기술의 문제점 및 향후 발전방향에 대해서 기술하였다. 특히, 자동통역 전화시스팀의 개발을 위한 음성인식기술에 대해 기술하였다.

최근의 음성인식은 HMM과 신경회로망을 이용한 인식방법을 사용하는 추세이며 인식대상 음성도 고립어 또는 연결어보다 연속음성을 처리하고 있고 어휘수도 대어휘(수만단어 이상)를 목표로 잡고 있다.

한편, 국내에서도 자동통역 전화시스팀을 개발하기 위한 연구가 국가적 차원에서 진행되고 있으므로 향후 국내의 음성인식기술은 가일층 발전할 것이며, 한국어의 음성인식은 한국인이 해결 해야 된다는 언어상의 특수성을 감안한다면 이러한 프로젝트의 시작은 매우 바람직하다. 그러나 인간의 인식과정이 아직 밝혀지지않은 시점에서 인간의 청취능력 수준의 인식기술을 얻기 위해서는 응용시스팀 개발 일변도에서 벗어나 좀 더 많은 기초적인 연구가 이루어져야 할 것이다. 이를 위해서는 음성신호처리 기술전문가 뿐 아니라 음성학, 음운학, 언어학 등 유관 학문분야의 전문가들 간에 긴밀한 협조체제 구축 및 기술교류가 어느 때보다도 절실하다 하겠다.