《主 題》

음성통신 기술의 현황

김 희 동, 구 명 완
(수원대학교, 한국통신소프트웨어연구소)

■차 례■

Ⅰ. 서 론

음성은 인간의 가장 기본적이고 친밀한 정보전달의 수단으로서 공간적 제약을 극복하기 위한 음성통신의 기술의 발전이 인간과 기계사이의 통신이 가능한 단계로 이어지고 있다. 이러한 의미에서 개인휴대통신이나 멀티미디어시스템 등의 최근 각광을 받고 있는 정보산업에서 음성처리기술은 핵심을 이루게 될 것이다. 따라서, 정보과학의 발전은 음성처리기술 알고리즘의 개발 및 실용화에 좌우된다고 볼수 있다.

음성에 의한 기계와의 대화는 선진외국에서도 극히 제한적인 상황에서 이루어지고 있지만 상용화를 위한 적극적인 노력을 경주하고 있다. 한국어의 음성의 특징은 다른 언어를 대상으로한 음성처리기술을 그대로 도입하여 재현할 수 없는 일이므로 국내 연구진에 의한 연구개발의 필요성이 부각되는 부분이라는 점에서 더욱 중요하다 하겠다. 다행히 국내 연구소, 학계, 산업계에서 음성기술에 대한 연구에 박차를 가하고 있는 시점에서 국내에서의 음성통신기술분야에 대한 개괄적인 정리를 해보는 것도 의미있는 일로 여겨진다.

본 고에서는 음성통신기술을 음성부호화, 음성합성, 음성인식으로 나누어 그 기술의 현황 및 응용분야를 중심으로 살펴본 후 향후 전망에 대하여 기술하고자 한다. 각 분야마다 매우 다양한 방법들이 연구되고 있으므로, 상세한 기술내용을 소개하는 것보다는 기술을 응용측면에서 설명하고자 노력하였다.

Ⅱ. 디지탈 음성부호화 기술

음성의 디지탈 부호화 기술은 가능한 낮은 전송속도로 음성통신을 위한 것과 음성을 효과적으로 축적하기 위한 방법을 연구하는 분야로서, 음성인식이나 합성의 기초가 되기도 한다. 전송속도를 줄이는 동기는 전송비용 및 축적비용을 줄이는 것이외에도 이동통신에서와 같이 제한된 채널용량에 다수의 통화로를 제공하고자 하는 요구에 따른 것이다. 또한, 종합정보통신망에서 음악, 음성, 그래프, 화상데이타 등을 공유하기 위해서, 또는 패킷통신망에서 가변속도코딩을 지원하기 위해서 사용되기도 한다.

음성부호화 기술의 적용분야에 따라 방식도 매우 다양하지만, 대개 다음의 세가지 방법으로 분류할 수 있다. 이 중 첫째는 음성파형을 표본화하여 양자화하는 파형부호화 방식이고, 둘째는 음성의 특징을 추출하는 분석과정을 거쳐 전송된 데이타를 수신측에서 음성을 합성하여 재생하는 보코딩방식이며, 세째는 파형부호화 방식과 보코딩방식의 혼합형태인 혼합부호화 방식이다. 이들 부호화 방식들의 상세한 내용들

은 참고문헌[1][2][3] 등을 참고하길 바란다. 본 고에서는 최근의 음성부호화의 관심이 되고 있는 3.4KHz음성신호와 7KHz음성신호에 대한 음성부호화 방법과 디지탈 이동통신에서의 음성부호화방법을 중심으로 다루고자 한다.

2.1 3.4KHz 음성부호화 방식

파형부호화방식중 현재 가장 많이 사용되는 것은 CCITT의 G.711 PCM(Pulse Code Modulation)방식이다. PCM은 음성신호를 부호화 하는데 있어 64kbits/s로 변환하는 것으로, 디지탈 전화교환기의 가입자접속부분 및 국간 링크의 디지탈전송에 사용되고 있다. ISDN에서 채널속도가 64kbps를 기준으로 된 것은 바로 PCM과의 호환성을 위한 배려이다.

PCM의 전송속도는 대역폭의 사용면에서 볼 때 아날로그 통신방법보다 훨씬 비경제적이다. 따라서, 음성의 대역폭 축소에 관한 연구의 결과로 음성신호의 용장도(Redundancy)를 이용한 예측부호화방식이 제안되었는데, 대표적인 예로는 ADPCM (Adaptive Differential PCM)을 들 수 있다. 예측부호화 방식의 기본원리는 과거에 입력된 음성신 표본값으로부터 다음에 들어올 신호의 크기를 예측하여 실제 입력신호화의 차이인 오차신호만을 양자화하여 전송한다. 수신측에서도 동일한 방법으로 예측을 하고 전송된 오차신호를 더하면 정확한 입력값을 얻을 수 있다. 예측이 비교적 정확하다면, 이 오차신호의 진폭은 음성신호의 진폭보다 훨씬 작으므로 그만큼 양자화레벨수도 줄어들게 된다는 것이다. 그러나, 이와같이 차분방법을 사용할 경우 전송오류에 대해서 수신측에서는 치명적인 음질저하를 얻게 되는 단점이 있다.

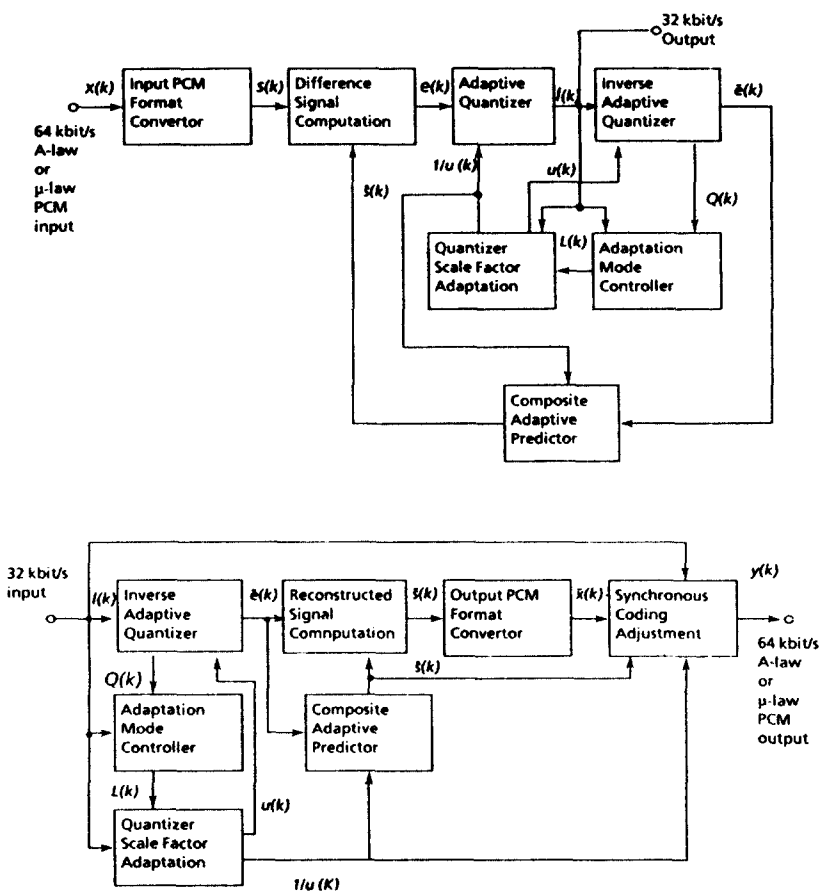

그림 1. G.721 ADPCM Coder의 Encoder와 Decoder

ADPCM방식은 73년도 최초에 발표된 이래, 예측을 정확히 하기 위한 방법과 전송오류에 대해 견인성(Robustness)를 가진 방식이 여러가지 제안되었으며, 1984년 CCITT권고안에서 표준으로 G.721로 정리되었으며 블록도를 그림 1에 나타내었다. 이러한 CCITT G.721 ADPCM의 특징을 살펴보면 다음과 같다.

- Codec이 32kbits/s에서 운용되도록 설계되고,
- 음성신호외에도 전화망을 이용하는 데이타통신용 모뎀신호 및 톤신호까지 부호화 할 수 있으며,
- 현재의 PCM과 직접연결이 가능하고,
- 예측기와 양자기가 입력신호에 적응하도록 설계되며,
- 채널의 전송오류에 비교적 강한 특성을 갖고 있다.

한편, G.726에는 가변속도의 ADPCM방식이 G.727에는 임베디드 ADPCM방식이 표준화되어 있다. 이경우 전송속도는 40, 32, 24, 16 kbps로서, 특히 임베디드 ADPCM의 경우 송신측에서는 40kbps로 encoding하고 네트워크의 상태에 따라 트래픽이 폭주할 경우 네트워크에서 데이타를 폐기하더라도 수신측에서는 전송된 데이타만을 가지고 음성을 재현할 수 있는 방법이다.

여기서 데이타전송속도를 낮추어 16kbps에서 ISDN 전송, 패킷음성통신, 코드리스전화, 비데오전화기 등에 사용할 수 있는 네트워크품질을 제공하는 부호화기에 대한 연구가 진행되었다. 16kbps부호화기의 설계요구조건으로는 음질은 G.721의 수준으로 유지하되 CODEC에서의 지연시간이 5ms이하고, 기존의 코더와 상호연결이 용이해야 하며, 채널에러나 패킷손실에 대해서도 음질저하가 최소화되어야 한다는 조건이다. 이 결과LD-CELP(Low Delay Codebook Excited Linear predictive coding)가 G.728 표준으로 채택되었다. 기존의 CELP가 입력신호를 버퍼링하고 처리하는 전향적응방식에 기초하고 있으나, LD-CELP는 여기서는 후향적응방식을 채용하여 부호화의 지연을 최소화하고 있다. CELP는 보코딩부호화방식 중 음성예측부호화방식에 LPC계수, 피치, 여기신호에 대하여 벡타양자화기를 두고 최적의 코드를 찾아내는 보코딩방식이다. 벡타양자화방법은 일종의 패턴매칭방법으로서, 일정한 블록의 음성샘플이나 파라메타 블럭을 하나의 벡타로 간주하여 이를 양자화하는 방법이다. 그림 2에는 전형적인 CELP부호화기, 그림 3에는 저지연 CELP부호화기의 블럭도를 비교해 두었다. 본 코더의 주관적음질평가 결과 G.721과 거의 동등한 수준의 결과를 얻었다. 현재는 이를 다시 8kbps수준으로 낮추는 연구가 진행되고 있다.

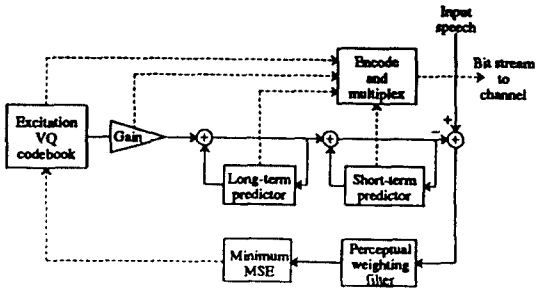

그림 2. 기존의 CELP 부호화기

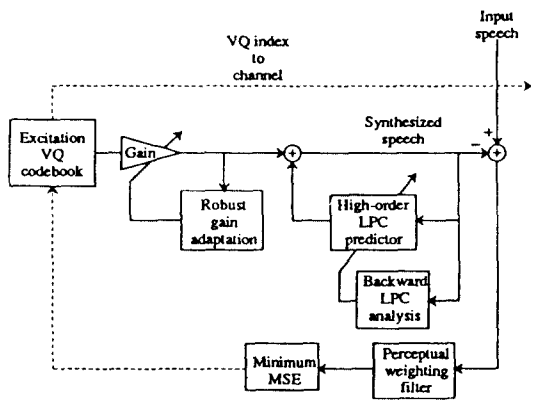

그림 3. LD-CELP 부호화기

2.2 광대역음성 부호화

음성부호화분야에서 다루고 있는 신호의 대상은 전화음성에서 광대역음성, 및 음악신호로 확장되고 있다. 3.4KHz로 대역을 제한된 전화음성은 전화기의 송수화기로 들을 때에는 별로 느낄 수 없으나, 스피커 등 통해 듣는다면 음질이 매우 나쁘다는 것을 알 수 있다. 음성의 이해도, 자연성이 높이기 위해 음성의 대역을 7KHz로 확장한 광대역음성에 대하여 1986년 CCITT에서는 64, 56, 48kbps의 G.722 표준을 제정하였다. 최근에는 이를 기초로 7KHz의 음성을 32kbps에서 16kbps 사이에서 부호화하는 검토가 되고 있다.

그림 4에는 G.722 부호화기의 블록다이어그램이 나타나 있다. 16KHz로 표본화되고, 14비트로 부호화된 PCM 데이타를 2개의 Quadrature Mirror Filter(QMF)로 분리하여 8KHz로 표본화하되 저역대역은 6비트/샘플로, 고역대역은 2비트/샘플로 G.721 ADPCM의 변형된 형태로 부호화한다. 저역대역에 대해서는 6비

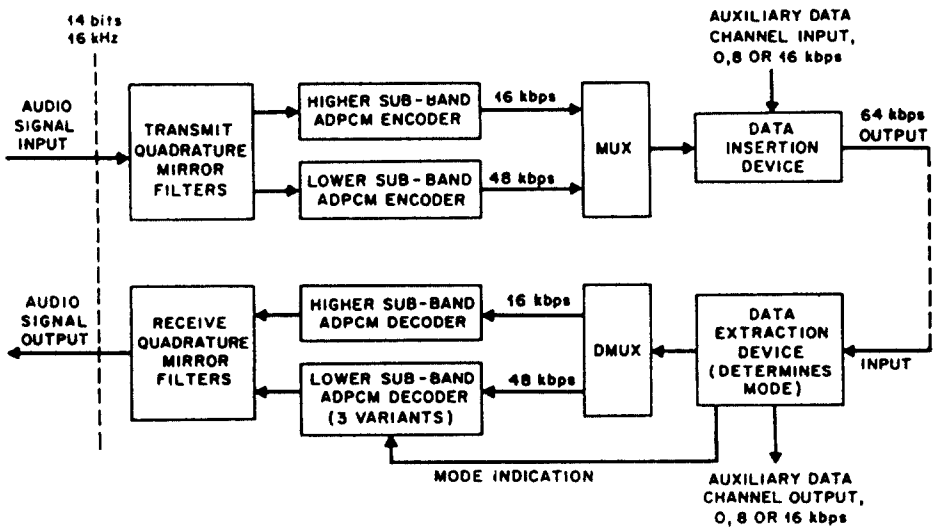

그림 4. G.722 7KHz 음성신호 부호화방식

트, 5비트, 4비트로 사용할 수 있으며, 적응알고리즘은 항상 4비트를 이용한 형태를 취하는 임베디드부호화방식을 사용한다.

이 전송속도를 더 낮추기 위하여, 광대역음성신호에도 G.728과 유사한 형식의 32kbps LD-CELP방식이 제안되어 주관적 음질평가결과 G.722와 거의 동일한 수준의 결과를 얻어내었다. 협대역 ISDN에서 64kbps 채널에 TV전화를 실현하기 위하여 48kbps의 영상신호전송과 광대역 음성신호를 16kbps로 전송하기 위한 방법이 요구됨에 따라 다시 지연을 약간 더 허용하는 범위에서 16kbps로 얻는 연구가 진행되고 있다.

2.3 이동통신에서의 음성부호화방식

언제, 누구하고, 어디서든지 통화를 하고 싶어하는 인간의 욕망을 충족시키기 위한 이동통신은 아날로그이동통신에서 디지탈이동통신으로 다시 개인휴대통신(PCS : Personal Communication Service))으로 발전해가고 있다. 아날로그이동통신에서는 하나의 Cell에서 제공 가능한 채널의 수가 제한되어 있으므로, 이를 디지탈 코딩방법을 사용하여 가능한 한 다중화의 효과를 얻고자 하는 디지탈이동통신으로 발전되고 있다. 이 때 낮은 전송속도로 효율적인 디지탈음성부호화를 필요로 한다. 일본이나 미국의 TDMA방식의 디지탈 셀룰라에서는 모토로라에서 제안한 8kbps의 고정전송속도의 VSELP(Vector Sum Excited LP)가 음성부호화의 표준으로 채택되었다. VSELP는 CELP에서 음원코드의 검색부분을 특수한 구성으로 하여 검색에 필요한 연산량과 소요메모리크기를 대폭으로 줄인 방법이다.

CDMA시스템에서는 퀄콤사가 제안한 QCELP(Qualcomm CELP)가 표준으로 제안되었다. QCELP보코더는 입력되는 음성신호의 에너지 크기에 따라 전송속도를 4가지로 변경할 수 있는 가변속도 부호화기로 무음구간에서는 800bps로, 음성구간에서는 8kbps로 전송하며 이들 구간사이에 갑작스런 변화를 방지하기 위하여 2kbps, 4kbps의 전송을 하게 된다. 이방식 역시 CELP의 변형으로 선형예측필터를 LSP(Line Spectrum Pair) 계수를 사용하여 보간이나 양자화에 이득을 얻을 수 있는 것으로 알려져 있다.[5]

한편 GSM에서는 RPE-LPT(Regular Pulse Excited LPC Long Term Prediction)으로 알려진 방법을 표준으로 채택하였다. 이 방식은 순수데이타의 속도가 13kbps로서 음성지연을 최소화하고 품질을 높이려는 의도에서 VSELP나 QCELP보다 전송속도가 높다.[4] RPE-LPT에서는 LPC의 분석부분에서 음성신호의 인접샘플간의 short term correlation을 제거하고 LPT분석부분에서 피치샘플간의 long term correlation을 다시 제거하여 백색잡음에 가까운 여기신호를 얻어 이들 펄스중 원래 음성과의 차를 최소로 하는 일정간격의 펄스열을 송신하여 수신측에서는 이의 역과정을

거쳐 원래의 음성신호를 재생한다. 그림 5에 RPE-LPT의 블록도를 나타내었다.

반면 PCS는 소형, 경량의 단말기를 개발하는 것이 관건이다. 따라서 음성부호속도를 8kbps정도로 줄이는 디지탈셀룰러의 복잡한 부호화방식에 비하여, PCS에서 사용되는 32kbps의 ADPCM이나 ADM(Adaptive Delta Modulation)은 회로가 훨씬 간단하며 이에 따른 전력소비와 신호처리 소요시간이 줄어들어 단말기의 소형경량화와 저렴화에 도움이 된다.[6]

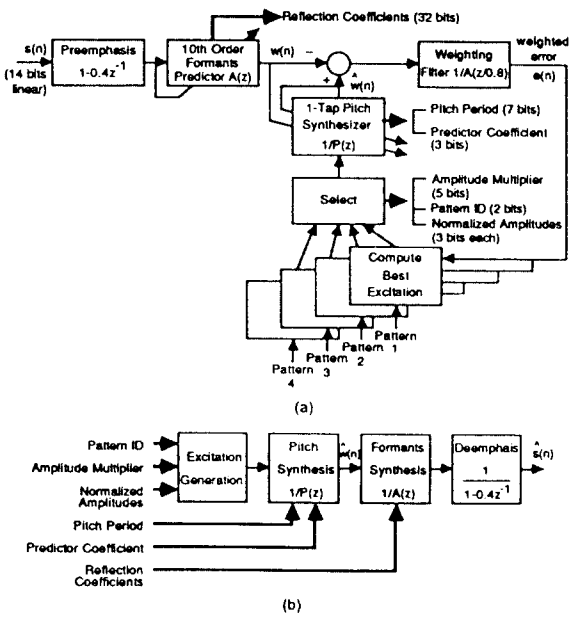

그림 5. RPE-LP 부호화기 (a) Encoder (b) Decoder

2.4 음성정보시스템의 부호화 방식에의 응용

음성정보시스템에서는 음성의 저장장치인 하드디스크의 용량 및 억세스시간을 줄이기 위한 방식이 요구되므로, 24-32kbps ADPCM을 사용한다. 이를 구현하는 방법으로는 상용의 ADPCM 칩을 채용하거나 디지탈신호처리프로세서에 의한 구현방식이 사용되고 있다. 현재 상용화된 ADPCM칩은 크게 2가지 종류로 PCM을 입력으로한 G.721 표준칩과 아날로그 음성을 입력으로 하되 독자방식에 의한 ADPCM 칩으로 분류된다. 전자의 방식을 취할 경우 음성채널에 PCM코덱과 ADPCM코덱이 2개가 필요하므로, 이보다는 경제적인 후자의 독자적인 ADPCM을 사용하는 것이 일반적이다. 반면, DSP를 사용하여 프로그램하는 방법은 경제적인 이유로 거의 사용되고 있지 않다.

음성사서함 및 음성정보시스템이 보편화되고 있는 선진외국의 경우에는 음성정보시스템상호간의 음성데이타의 송수신을 하기 위한 필요성을 느끼고, 음성정보시스템 산업체의 협의회에서 CCITT G.721을 상호 연동시의 표준부호화방법으로 채택하였다. 국내에서도, 시스템 상호간의 데이타전송을 위해서는 G.721 ADPCM을 표준으로 정해야할 당위성이 있으므로 이에 대비해야 할 것이다.

Ⅲ. 음성합성

음성합성이란 입력문장을 사람이 청취할 수 있는 음성신호로 변환시키는 과정을 말한다. 음성합성은 기술의 난이도에 따라, 단순녹음재생방식, 녹음편집방식, 규칙합성의 여러가지의 방식이 있으며 이하 각 방식과 응용사례에 대하여 설명하기로 한다.

3.1 단어조합형태의 음성합성

입력된 문장을 음성으로 변환시키는 가장 손쉬운 방법은 합성하고자 하는 문장의 음성파형을 컴퓨터에 디지탈화하여 미리 저장시켜 놓았다가 이를 저장된 데이타를 토대로 간단한 단어조합 등을 이용하여 음성 파형을 재생시키는 방법이다. 이 방식은 조합할 음성 기본단위(단어, 구 또는 문장) 및 디지탈화 시킨 음성의 억양 등에 의해 제약을 받을 뿐아니라, 사용어휘수 및 문장 형태에 제한이 가해지게 되므로 제한적 음성합성 방식이라 불리운다. 이러한 제한적 음성합성방식은 116전화시보장치, 114전화번호안내, 신용카드 조회, 은행의 잔고조회 등과 같은 조회용 자동응답음성시스템 등 단순한 음성합성 기능이 요구되는 분야에 널리 사용되고 있다.

이러한 방식의 가장 큰 제약점은 미리 녹음된 음성데이타베이스의 억양이나 고저 장단등의 음성내용을 변경시킬 수 없으므로 자연 합성음의 품질이 자연스럽지 못하게 된다. 이를 극복하기 위하여 동일 음성에 대해서 억양이 다른 수개의 데이타베이스를 가지도록 할 수 있으나 소요되는 메모리의 용량이 커지는 단점이 있다. 따라서 어휘수가 많고 비정형적인 단어의 합성분야에 적합하지 못하지만, 사람의 이름이나 주소 등의 합성에서와 같이 조합음의 자연스러움보다 정보의 내용자체가 중요시될 경우에 이 방식을 사용할 수 있다.

일반적으로 음성데이타베이스의 음성을 부호화하

는 방법으로서 과거에는 소요되는 메모리의 용량을 줄이기 위해 LPC, RELP 등의 저,중속도의 부호화방식을 사용하였으나, 메모리의 가격이 현저하게 줄어든 현시점에서는 ADPCM이나 PCM 등의 방법이 주로 사용되고 있다. 이방식에 대한 기술은 보편화되어 자동음성응답전화기에서 전화온 시간의 음성기록에 사용되는 등 응용분야가 넓어져 가고 있으며, 관건은 경제적으로 최적의 응용분야에 맞도록 설계하는 것이 촛점으로 되어 있다.

3.2 무제한 음성합성

이에 반하여 어휘나 문장형태에 아무런 제약없이 임의의 문장을 음성으로 합성해내는 무제한 음성합성방식이라 부른다. 무제한 음성합성방식과 제한적 음성합성방식의 구성상의 가장 현저한 차이는 언어학적 처리부의 유무에 있다.

Text-to-speech 합성 시스템의 기본 구성도는 그림 6 과 같다. 그림에서 보는 바와 같이 합성 시스템은 크게 언어학적 처리부와 음성신호 처리부의 두 부분으로 나눌 수 있다. 이들 중 언어학적 처리부는 다시 text 전처리, 문장분석 및 발음표기 변환의 세 과정으로 나누어진다. 입력 문장이 언어학적 처리부로 들어오면, 먼저 전처리 과정을 통해 약어나 특수기호 등의 표기를 구술적인 표현으로 대체시킨 다음, 문장분석 과정에서 문장의 구조를 분석하게 된다. 이 과정에서 얻어지는 정보는 음성 신호 처리부로 넘어가고, 고저, 장단 및 휴지기간 등 운율 조절에 사용된다.

언어학적 처리부의 마지막 단계인 발음표기 변환 과정에서는 발음법칙에 의해 입력 문장을 사람이 발음하는 형태의 표기로 변환시키게 된다. 이 언어학적 처리부의 구현에는 규칙으로 표현하기 어려운 많은 예외 조항이 따르므로 사전을 이용하여 이러한 문제들을 처리하도록 한다.

음성 신호 처리부는 운율조절 과정과 음성 파형 생성과정으로 나누어진다. 운율 조절 과정에서는 언어학적 처리부의 결과를 토대로 하여 음성의 고저, 장단, 강세 등 운율을 사람이 발음하는 것과 같은 자연스러움에 접근하도록 조절한다. 끝으로 음성파형 생성 과정에서는 미리 구축된 음성 데이타베이스로부터 음성 기본 단위들에 대한 정보를 제공받아 운율정보를 고려한 합성 파형을 생성시킨다.

무제한음성합성중 음성신호처리부의 성능이 전체의 합성음의 품질 및 복잡성을 좌우하는 요소로서 특히 음성합성단위의 선택, 음성합성방식의 선택에는

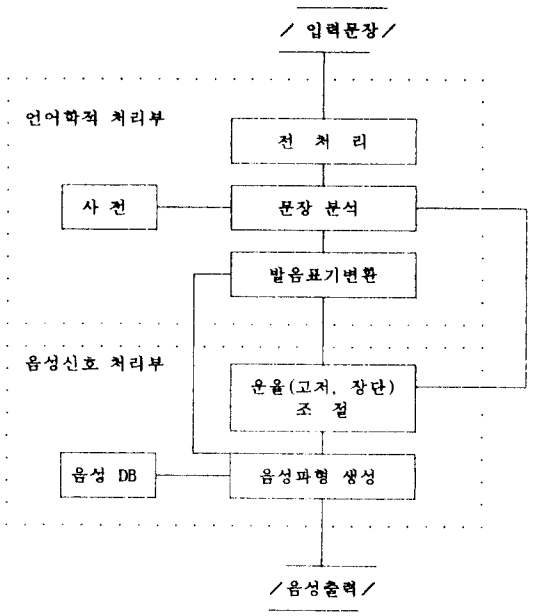

그림 6. 음성합성시스템의 블럭도

시스템의 복잡성과 데이타베이스의 크기사이에 trade-off 관계가 있다. 무제한 음성합성을 위한 가장 보편적인 접근방법은 음절, 반음절, diphone, 음소 등을 사용할 수 있는바 이들의 단위 들이 한국어의 음성합성에 사용되었을 때의 장단점에 대해서는 [1]에 상세히 기술되어 있다. 기본적으로 합성단위가 작으면 합성에 필요한 전체 데이타의 양이 작아지며 각각의 단위를 제어하기는 쉬우나, 합성단위를 연결할 때 상호 조음현상을 정확히 구현하기 어렵고 음질저하현상이 생기기 쉽다. 따라서 diphone과 같이 모음의 안정된 구간 사이의 음성데이타를 기본단위로 많이 사용하고 있다. 최근에는 합성단위를 한가지로 제한하지 않고 음소열을 음소 환경이 같은 집단으로 분할하여 음소환경에 따라 합성 단위가 자동적으로 생성되는 COC (Context Oriented Clustreing) 방법이 제안되었다.[7]

무제한 음성합성방식으로는 포만트합성과 LPC합성방법이 주로 사용된다. LPC합성방식은 근본적인 문제인 비음의 처리, 음색의 변경에 따른 LPC 계수처리문제가 난제로 되어 있으나 포만트 합성방식에 비해 단순한 구조를 가지며 분석과정을 자동적으로 수용할 수 있는 장점을 갖는다. 포만트합성은 음성신호의 특징계수들을 용이하게 조절할 수 있어, 음성의 특성 및 운율 변화가 용이하나 대신 다양한 조음현상에

따른 포만트의 변화를 정확히 규칙화하기 위하여 음성신호의 분석이 수반된다. 이들의 선택기준은 간단한 구조를 택하는가 아니면 우수한 음질 구현의 가능성이 높은 방식을 택할 것인가에 달려있다.

2.3 음성합성시스템의 구현사례

국내외에서 상용화된 음성합성시스템의 제품을 표 1에 나타내었다. 국내에서 음성합성시스템이 구현된 사례로서 한국통신에서 제공하고 있는 KT-MaiL서비스의 음성전달서비스를 위한 FAX/VOICE-AU시스템을 들 수 있다. 전자메일은 문자정보를 주고받는 서비스로서 데이타단말을 가진 사람만이 이용할 수 있으나, 전자메일에 대한 권고안이 CCITT의 X.400 시리즈로 발표되면서 기존의 텔레매틱단말기 즉 팩시밀리나 전화기를 가진 수신자에게 텍스트메세지를 전달하는 기능을 Access Unit로 정의하였다. 한국통신에서 운영하는 KT-MAIL 메세지핸들링시스템에는 FAX/ VOICE-AU시스템을 설치하여 호스트컴퓨터에 수록된 메세지정보를 팩스이미지변환 또는 무제한 음성합성기능을 이용하여 전달하는 시스템을 구축하였으며, 이의 구성도는 그림 7에 나타내었다.[8]

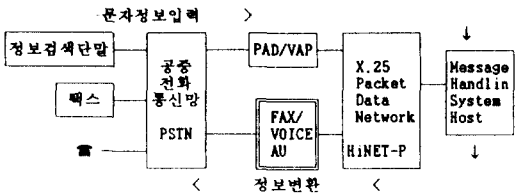

그림 7. FAX/VOICE-AU 시스템의 구성도

이 시스템에서는 전화, 팩스등의 단말은 발신기능이 없고 수신기능은 없다.

반면, 한국전자통신연구소에서 PC통신서비스의 개념을 전화, 팩스단말의 이용자에게 확장하여, 이용자의 성우에는 음성합성으로, 팩스이용자에게는 팩스이미지변환하여 데이타베이스 호스트에 수록된 정보를 검색할 수 있도록 하는 통신처리시스템을 개발하였다. 통신처리장치의 기능블럭도는 그림 8에 나타내었다.

표 1. 대표적인 음성합성 시스템들의 예

회사명	제품명	형태	합성방식	기 능
DEC	DECtalk	독립형	Formant	- 7명 음성 - 발음속도 조절 - 음색변경 기능 - 전화망 접속 기능
Speech Plus	PROSE 4000	PC 내장형	Formant	- 남성 3명 음성 - 발음속도 조절 - 전화망 접속기능
Street Electronics	Echo PC+	PC 내장형	LPC	- 남성음성 - 발음속도 조절
Artic Technology	SynPhonix	PC 내장형	Allophone 연결방식	- 남성음성 - 발음속도조절
NTT Data	VCS-II-1	NEC PC 내장형	LSP	- 남/녀음성 - 발음속도 조절
삼양전기 (일본)	VSS-300	독립형	LSP	- 남/녀음성 - 발음속도 조절
디지콤 (한국어)	가라사대	IBM-PC 내장형	LPC	- 남자음성 - 발음속도, 톤조절
디지콤	팩스/보이스 억세스유니트	독립형	LPC	- PC형을 MHS와 연동 - 전화망접속기능

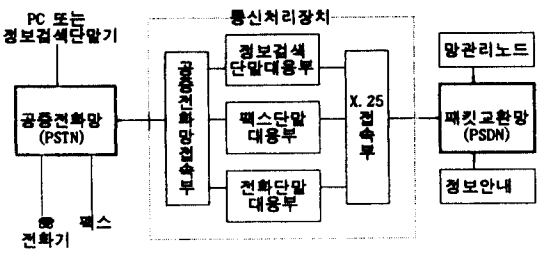

그림 8. PC통신서비스시템의 구성도

Ⅳ. 음성인식

4 1 음성인식 시스템의 기본 개념도

음성인식기술의 궁극적인 목표는 임의의 화자가 문법에 제한없이 일상적으로 발음한 문장을 인식하여 그 의미를 파악하는 시스템의 개발이다. 그러나, 현재의 기술수준으로는 이러한 목표달성에 크게 못 미치고 있으며, 인식기술의 연구도 특정한 제약조건 아래에서 제한된 목표를 가지고 이루어지고 있는 것이 현실이다.

현재의 음성인식 기본 개념도는 그림 9와 같이 기본적으로 음성으로 부터 음성 패턴(단어, 음소등)의 특징을 추출하여 기준패턴을 만드는 훈련과정과 미지의 음성이 입력되면 저장된 기준패턴과 비교하여 가장 유사한 기준패턴을 찾아내는 인식과정으로 나눌 수 있다. 이러한 알고리즘을 일반적으로 패턴 매칭(pattern matching) 알고리즘이라고 부른다.

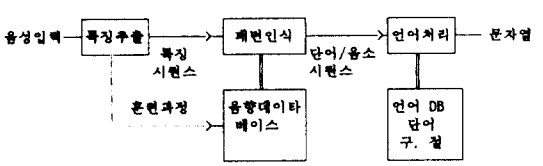

그림 9. 음성인식시스템의 블럭도

4.2 음성인식 시스템의 분류

음성인식 시스템은 표 2와 같이 단어를 인식하는 고립단어 인식시스템과 연속적으로 발음된 문장을 인식하는 연속음성 인식시스템으로 나눌 수 있다. 연속음성 인식시스템은 연결단어(connected word) 인식시스템과 대화체음성(conversational speech) 인식시스템으로 세분화 될 수 있다. 연결단어 인식시스템은 상대적으로 작은 단어를 매 단어마다 또박 또박 발음하는 단어를 인식하는 시스템인 반면 대화체음성 인식시스템은 상대적으로 대용량 단어를 매 단어마다 인식하는 것이 아니라 문장의 의미를 파악하는 것이다. 이와 같은 대화체 음성인식 시스템은 음성이해 시스템이라고도 불리우며 언어지식을 사용하는 것이 중요하다. 대화체 음성인식 시스템은 상당히 어렵기 때문에 문장내에서 필요한 단어만 선별하여 인식할 수 있는 단어선별(word spotting) 인식시스템의 연구가 최근 활발히 진행되고 있다.

표 2. 음성인식 시스템의 분류

	화자종속	화자독립
고립단어인식시스템	상용화	상용화
연속음성인식시스템		
◦ 연결단어인식	상용화	상용화
◦ 대화체음성인식	연구중	연구중

또한 음성인식시스템은 훈련이 된 특정화자 만을 인식할 수 있는 화자종속 인식시스템과 어떠한 사람의 음소도 인식할 수 있는 화자독립 인식시스템으로도 나눌 수 있다. 일반적으로 화자독립 인식시스템이 화자종속 인식시스템보다 훨씬 어려우나 이용 범위가 넓기 때문에 많이 연구되어지고 있다.

한편 기준패턴의 단위를 무엇으로 사용하느냐에 따라서 음성인식시스템의 특징이 구분될 수 있다. 단어를 기준패턴으로 사용하면 인식율이 높은 반면 인식대상 단어수가 많아질수록 메모리와 계산량이 증가한다는 단점과 연속음성에 내재되어 있는 단어사이의 연음(coarticulation)현상을 표현할 수 없다는 단점도 있다. 반면 훈련과정에 전혀 사용하지 않는 어휘도 인식할 수 있는 단어독립(vocabulary-indepent) 인식시스템에서는 기준패턴으로서 주로 음소를 사용한다. 이경우 대상 단어수가 늘어나더라도 계산량 및 메모리 사용량이 많이 증가되지 않고, 훈련과정도 간단하며, 단어사이의 연음현상도 쉽게 표현될 수 있다. 그러나 음소의 발음규칙이 명확히 해결되지 않았기 때문에 인식율이 떨어지는 단점이 있다. 최근에는 이러한 현상을 극복할 수 있는 문맥종속음소(context-dependent phoneme)를 기준패턴으로 사용하기도 한다.

4.3 음성인식 시스템의 난제

현재의 음성인식 기술의 발전이 더딘 이유는 다음

과 같은 어려움이 따르기 때문이다. 첫번째로 연음과 축약(reduction)현상이다. 단어 혹은 문장내에서의 음소는 이웃하는 음소에 따라 발음현상이 매우 달라진다. 특히 연속음성인 경우 음소의 발음이 사라지는 현상도 있다. 두번째로 음소를 정확히 분할하기 힘들다. 음성은 글자와 대응되는 음소구간을 시간축에 대해서 정확히 분할하기는 어렵다. 무성자음(unvoiced consonant)은 어느 정도 분할이 가능하나 모음은 상당히 어렵다. 세번째로 개인성이다. 다른 사람이 동일한 단어를 발음하여도 발성속도 및 발성기관에 따라 음성 특징은 달라진다. 네번째로 언어학 정보의 부재이다. 일반적으로 사람이 음성을 이해하는데 언어정보를 사용한다. 예를 들면 문법적으로 이해하기 어려운 음성을 발음하더라도 사람은 그 음성을 구문해석하여 문법에 맞는 적당한 단어로 해석하거나 장의 의미가 파악되도록 의미적해석을 하게 된다. 그러나 아직까지 언어학에서 이러한 의미인식과정에 대한 연구결과를 얻지 못하고 있다.

4.4 음성인식 시스템의 시장성

음성인식 시스템에 대하여 현재 미국에서 활성화되고 있는 분야는 데이타 입력, 전화망 및 음성문자변환 분야이다. 데이타 입력 분야는 음성인식 시스템 시장의 주류를 이루고 있으며 점차적으로 증가 추세에 있다. 자동차 조립공장에서 페인트 칠의 상태를 감시시키기 위하여 사용되는 Vocollect와 같은 데이타 수집용 인식시스템 등이 대표적인 응용사례이다. 또한 우체국에서 소포, 편지 등을 자동으로 분류하기 위하여 음성명령을 이용할 수 있는 시스템이 개발되어 있다. 이런 분야의 인식시스템은 적용분야에 따라 비용이 달라지나 기본적인 시스템인 경우 대 인식 터미날당 2~4천 달라 정도이라고 한다. 현재 판매자들은 좀더 낮은 가격, 짧은 훈련시간 및 설치 경험 등의 향상을 위해서 노력하고 있다.

전화망을 통한 음성인식 시스템은 3.4KHz 전화음성을 인식하는 것으로서 자동응답시스템의 메뉴선택이나 오퍼레이터의 작업을 줄이는데 사용된다. 이 시스템은 전화기의 푸쉬버튼을 사용하는 기존의 자동음성응답시스템을 대처할 수 있어 전화망 분야는 음성인식 시스템의 시장중 가장 빠르게 성장하고 있는 분야중의 하나이다. 현재 미국에서는 많은 지역 수신자 부담 전화(collect call)가 안내양의 도움없이 자동음성인식시스템에 의하여 이루어진다. 이러한 시스템은 "네", "아니오" 등의 간단한 단어를 인식할 수 있다. 또한 오디오텍스 시스템 공급사인 Advance Telecom Service 및 West Interative 회사에서는 음성인식 기술을 그들의 시스템에 도입하고 있다. 이와 같이 전화망을 통한 음성인식 시스템의 보급이 활발해짐에 따라 음성인식 기능을 첨가하는데 필요한 부가적인 비용이 전화회선당 수백달러 이하로 떨어지고 있다.

음성문자변환 분야도 최근 몇년전 부터 빠른 성장을 보여왔다. 미국내의 중소기업인 Dragon Systems 사는 음성으로서 자유롭게 텍스트를 만들 수 있는 30,000단어 인식기를 판매하고 있다. 이 시스템은 워드 프로세싱 소프트웨어의 입력용으로 사용되고 있는데 현재까지 판매된 약 반 정도가 기존의 키이보드를 이용할 수 없는 장애자들이 구매하였다고 한다. Kurzweil AI 회사도 의사들의 진료기록작성용의 인식시스템을 제공하고 있다. 이러한 시스템들은 8 Mbyte 메모리를 갖는 IBM PC AT 호환 컴퓨터를 포함하여 $15,000~$30,000이라고 한다. 현재 1년에 수천개 이하로 판매되고 있지만 인식대상 어휘가 늘고 사용하기가 편리해짐에 따라 판매댓수는 증가할 것이다.

앞으로는 단말기, 화자 확인 및 명령과 통제 분야에서도 비약적인 성장이 있을 것이다. 단말기가 더욱 복잡해짐에 따라 사용자는 지능적인 제어능력이 단말기에 부가되기를 요구할 것이다. 지능적인 응용이 가능하려면 일상적인 환경에서 적절한 단어를 화자 독립적으로 인식할 수 있는 기술이 개발되어야 하며 싼 가격으로 IC chip화 되어야 할 것이다. 이런 능력을 갖는 소프트웨어 및 하드웨어는 앞으로 3~5년 이내에 개발될 것으로 예상된다.

화자확인 분야는 전화망을 통한 컴퓨터 액세스가 넓게 사용됨에 따라 중요성이 부각될 것이다. 음성은 개인성을 나타내는 상세한 정보를 가지고 있으므로 컴퓨터나 전화망의 사용권을 제한하는 보안시스템에 이용될 것이다. 현재 미국 Sprint 전화회사에서는 개인식별번호(personal identification number)에 화자 확인 과정을 부여하여 전화를 사용할 수 있는 지의 여부를 알아내는 시스템의 성능실험을 하고 있다고 한다.

명령과 제어분야는 개인 컴퓨터에 펜 및 마우스 입력 이외에 음성인식 기능이 부여됨에 따라 진보할 것이다. 현재에서도 음성으로서 컴퓨터에 간단한 명령을 내릴 수 있는 시스템이 개발이 되어 있지만 앞으로 몇년 이내에 음성 입 출력이 가능한 멀티미디어 인터페이스 기능이 부가된 컴퓨터가 출연할 것이다. Voice Information Associates에서는 미국내에서 1990년부터 1995년까지 음성인식 시스템의 시장의 성장

률은 년간 44% 정도가 될 것으로 추정하고 있다. 음성인식 시스템의 시장규모는 1992년도에 약 2억3천3백만 달러 정도이며 1995년도 까지 5억7천6백만 달러로, 세기말에서는 십억 달러까지 도달할 것이라고 추정한다. 만약 유럽과 일본에서의 시장을 포함한다면 이 수치는 2배정도가 될 것이다.

4.5 음성인식기술을 응용한 응용시스템

가. 음성인식 기술을 이용한 부서안내 시스템

한국통신에서는 1992년부터 음성인식 기술을 이용한 정보검색 시스템인 KARS(Korea telecom Automatic Recognition System)를 개발하여 왔다. 현재 개발된 시스템은 전화를 자동으로 수신하여 사용자의 명령음성을 인식하여 한국통신 연구센타내의 부서명에 따른 전화번호 및 위치를 안내하여 준다. 현재 KARS는 한국통신 연구센타내의 부서명 188단어를 인식할 수 있다. 구내 교환시스템을 통할 경우 화자독립 인식율이 현재 92.0%이며 PSTN을 통한 음성에 대해서 성능평가를 수행하고 있다.

그림 10에는 KARS의 개략적인 구성도를 나타내었다. 사용자가 KARS로 전화를 걸면 시스템이 자동응답하여 사용자의 음성명령을 요구하는 안내문을 출력한다. 사용자의 음성명령이 입력되면 음성구간이 검출되고 스팩트럼상의 특징추출, 벡타 양자화 과정이 수행된 후 대화관리기로 전달된다. 대화 관리기는 시스템의 사용을 쉽게하기 위해 KARS의 모든 흐름을 제어한다. 대화관리기에 의해 제어되는 인식단어 검색기는 음소모델을 근거로 하여 인식대상 단어중 입력음성의 특징과 가장 유사한 단어를 선택한다. 그림 11은 KARS와의 실제적인 대화과정을 나타내었다.

현재 본 시스템은 SUN/UNIX 워크스테이션에 장치된 2개의 DSP 보드(TMS320C30)를 사용한다. 하나의 DSP 보드는 전화선을 통한 A/D 및 D/A 변화를 위한 아날로그 모듈을 포함하고 있으며 특징추출 및 벡타양자의 과정이 프레임 동기적으로 수행되고 있다. 다른 하나의 DSP 보드는 단어검색 알고리즘을 위

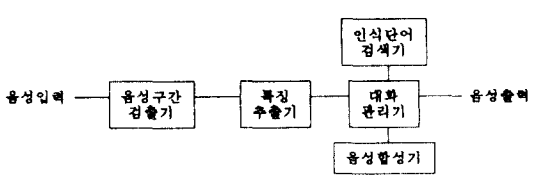

그림 10. KARS의 개략도

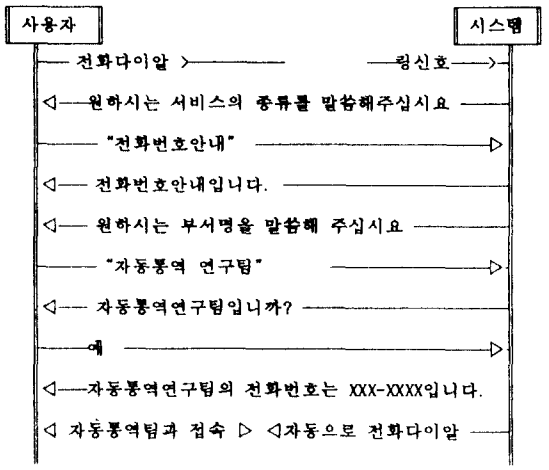

그림 11. KARS와의 대화과정 예

해 사용되며 프레임 비동기적으로 수행된다.

나. AT&T의 음성인식기술 응용 시스템

음성인식기술이 실용화되기 위해서는 일반적인 음성에서 의미적으로 중요한 단어를 인식하는 키워드 인식기술, 시스템이 안내문을 송출하고 있는 경우에 사용자의 음성을 자동으로 인식하는 Barge-In이라는 안내문 인터럽트 기술과 기준패턴을 만들 때 단어를 사용하지 않고 음소등과 같이 서브워드를 사용하는 기술이 요구된다. 이러한 3가지 기술적인 문제를 해결한 시스템이 현재 미국 AT&T에서는 자동 호 과금 시스템(Automated alternate billing system)에 적용하고 있다. 이 시스템은 "0"번을 dialing한 후 콜렉터 콜(callect call), 전화카드 콜(calling card call) 등 중 한 단어를 음성으로 말하면 자동으로 인식하여 관련 서비스를 수행하여 준다. 이 시스템은 10단어를 인식할 수 있는 화자 독립 고립단어 인식 시스템이다. 사용되는 기술은 HMM(hidden Markov modeling) 기술이며 이 시스템의 사용빈도수는 1년에 10억호 정도가 된다고 한다.

다. NYNEX의 음성인식기술 응용 시스템

NYNEX에서는 현재 2종류의 음성인식 기술을 이용한 응용시스템을 운용하고 있다. 하나는 음성 다이얼링 서비스로서, 사용자가 음성으로 상호나 사람이름을 전화기에 직접 말하면 음성인식하여 자동으로

전화를 걸어주는 서비스이다. 이 서비스를 위하여 NYNEX는 1990년에 모델 시스템을 개발하였으며 2년에 걸쳐 시험서비스 기간을 거친후 1992년 3월 중순부터 400명 가입자를 대상으로 시험 중에 있다. 이 서비스에 가입하면 한달에 $4~6를 지불해야 한다. 현재 가입자는 50명의 상호나 사람이름을 등록할 수 있다. 시스템에 사용된 기술은 DTW(dynamic time warping)를 이용한 화자종속 고립단어 인식기술이다.

또 다른 하나는 전화번호 안내보조 시스템(directory assitance system)이 있다. 이 시스템은 사용자가 보스톤 내의 Smith 사람의 전화번호를 알고 싶을 때 "보스톤 내의 Smith씨 전화번호를 알고싶습니다"라고 발음하면, 인식시스템이 보스톤이라는 단어를 인식하여 보스톤 내의 전화번호를 안내해 줄 수 있는 안내양으로 자동으로 접속되는 기능을 수행한다. 이 시스템은 HMM 기술을 이용하는 화자독립 고립단어 인식시스템이며 도시이름이 없을 경우 서비스가 거부되는 기능도 있다. 현재 94.3%의 서비스 성공율을 기록하고 있다.

라. 캐나다의 음성인식 기술 응용 시스템

캐나다의 Northern Telecom에서는 전화음성을 인식하여 회사명을 인식하고 해당회사의 주가등의 정보를 안내해주는 Stock Talk 시스템을 운용하고 있다. 그림 12는 Stock Talk시스템의 망 구성도를 나타낸 것이다. Stock Talk 시스템은 X.25 네트워크를 통하여 증권정보 데이타베이스 컴퓨터와 연결되어 있으며, 공중전화망을 통하여 사용자와 접속된다. 사용자가 특정번호(1-800-661-STOC)로 전화를 걸어 뉴욕 증권시장에 상장된 1092개의 회사와 토론토 증권시장에 상장된 883개의 회사중 하나를 음성입력하면 이를 인식하여 증권정보를 알려준다. 이 시스템에 사용된 기술은 HMM이며 1561개의 회사명에 대한 인식실험 결과 96%의 인식율을 얻고 있다.

마. NTT의 음성인식 기술 응용 시스템

일본 NTT의 대표적인 음성인식 기술 응용 시스템은 ANSER이다. 이 시스템은 전화 한 통화로 은행 고객에게 수표 조회, 신용 정보등을 제공하는 금융정보 서비스 시스템이다. 시스템에 사용되는 기술은 DTW를 이용하는 화자독립 고립단어 인식기술이며 15단어를 인식할 수 있다. 현재 한달에 3천호 정도 사용되고 있으며 년간 3천만 달러의 수입이 있다고 한다.

바. PC와 workstation용 음성인식 소프트웨어

앞에서 설명한 바와 같이 특정목적을 위한 음성인식 기술을 이용한 전용 응용 시스템이 있는 반면 PC와 워크스테이션용 음성인식 소프트웨어도 현재 많이 판매되고 있다. 대표적인 소프트웨어로서는 워크스테이션용 Hark 소프트웨어와 PC용 DragonDictate가 있다. Hark는 BBN에서 개발한 연속음성인식용 소프트웨어로서 오디오가 장착된 Sun과 Silicon Graphics/Unix 워크스테이션에서 동작한다. 이 소프트웨어는 DSP 혹은 array processor와 같이 부가적인 하드웨어를 필요로 하지 않으면서 2,000단어를 인식할 수 있는 기능이 있다. 또한 사용된 기술은 HMM이며 새로운 단어 혹은 사람의 음성을 쉽게 훈련시킬 수 있으며 1000단어의 연속음성 데이타베이스 사용시 94.5%의 인식율을, TI 숫자 데이타베이스를 사용했을 때는 99.5%의 인식율을 얻었다. 그림 13에는 Hark 인식기의 구

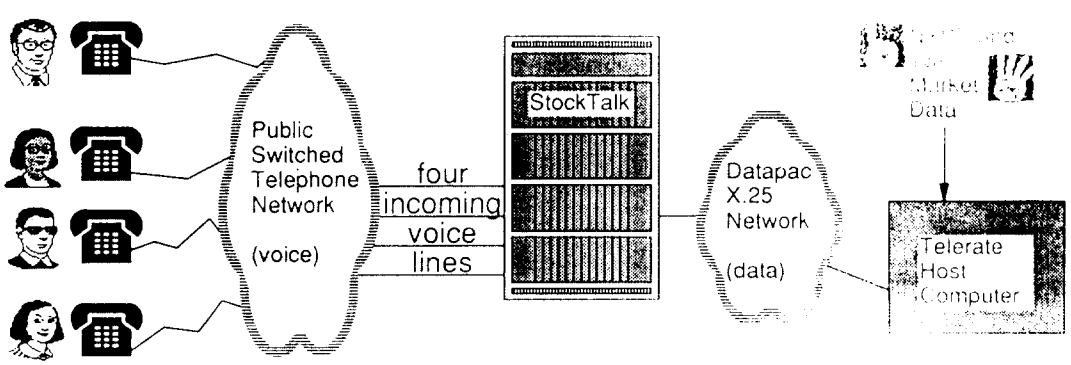

그림 12. Stock Talk의 망 구성도

조를 나타내었는데 응용 프로그램 인터페이스를 통하여 사용자가 필요로 하는 응용시스템을 개발할 수 있다. 대표적인 응용으로서 기업체 내에서 사람이름을 전화로 통해서 발음하면 그 사람의 전화번호로 자동적으로 다이얼링하는 시스템이 있다.

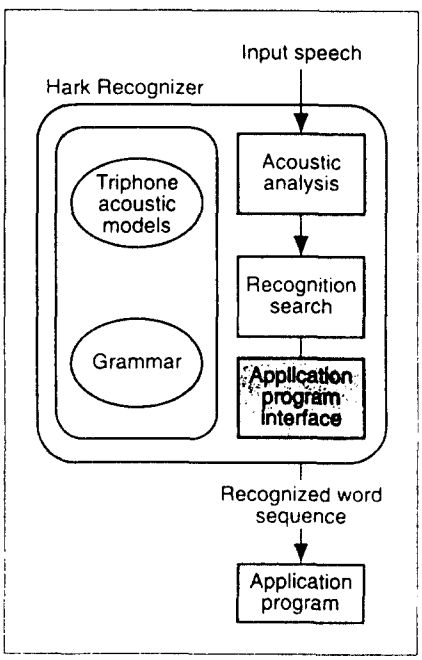

그림 13. Hark 인식기의 구조

대표적인 PC용 음성인식 소프트웨어인 Dragon-Dictate는 1990년도 부터 Dragon Systems 회사가 판매해 왔는데 최근에 DragonDictate-30K Version 2.0으로 성능 향상되었다. 이 소프트웨어는 PC 386 혹은 486에서 MS-DOS, OS/2 및 windows 3.1 등의 OS상에서 동작되며 최소 12Mb의 메모리와 22Mb의 하드디스크 크기를 필요로 하고 있으며 IBM ACPA 보드를 포함해서 $4,995에 팔리고 있다. DragonDictate 2.0의 대표적인 특징은 17과의 tutorial이 있어서 사용자가 2시간 반동안 tutorial 과정을 습득하면 습득과정에서 얻은 음성을 이용해서 자체내에 있는 화자독립용 파라미터와 결합되어 사용자의 음성인식율을 향상시키는 기능을 갖고 있다. 이 소프트웨어는 음성으로서 워드프로세서를 작업할 수 있게 하는 hand-free 소프트웨어로서 음성으로만 계속 작업을 시킬 수 있는 eye-free 기능도 있다. 현재 3만 단어를 인식할 수 있으며 사용자의 단어사용 빈도수에 따라 단어내용도 변경시킬 수 있다.

V. 결 론

지금까지 디지탈 음성통신기술을 음성부호화기술, 음성합성, 음성인식기술로 나누어서 그 기술의 주요 개념과 현황을 살펴보았다. 이러한 음성처리 기술들은 음성에 의한 입출력이 요구되는 분야가 계속 증가됨에 따라 다양한 형태로 실용화되어 가고 있다. 다행히 국내의 기술도 상당히 진척되어 있어, 음성정보시스템은 완전 국내 개발되었으며, 음성합성보드도 상용화되었으며, 전자메일 내용을 음성으로 출력하는 응용까지 실용화되었다. 한국어의 음성처리는 국내 기술진에 의해서만이 해결되어야 하는 현실을 볼 때, 국내 연구진의 노력은 고무적인 것으로 평가될 수 있다. 앞으로도 요소기술들의 기술개발은 물론 현재 기술개발수준을 활용하여 시스템화하고 서비스를 개발하는 방향으로 전개되어야 할 것이다.

(주) CCITT가 ITU로 명칭을 변경하였으나 본 고에서는 CCITT로 명기하였다.

참 고 문 헌

1. 김회동, "음성처리 기술의 현황과 전망", 한국통신학회지 8권 6호, pp. 11-31, 1991년 6월.
2. 이황수, "저 전송속도 음성부호화를 위한 디지탈 음성처리 신호처리기술", 한국통신학회지, 9권 10호, pp. 4-24, 1992년 10월.
3. 한국음향학회, "자동통역 전화시스템의 기술개발 전략(III)", 음성통신 및 신호처리 워크샵 논문집, 1993년 8월
4. 유지만, 황인태, "CDMA와 GSM의 기술비교", 전자공학회지 제21권 1호, pp. 34-41, 1994년 1월.
5. 하순회, "CDMA시스템 디지틀 신호처리 기술", 전자공학회지 제21권 1호, pp. 24-33, 1994년 1월.
6. 최두환, "PCN의 무선접속방식과 CDMA", 텔레커뮤니케이션 리뷰 Vol. 3, No. 9, pp. 42-53, 1993.
7. S. Nakajima and H. Hamada, "Automatic generation of sysnthesis units based on context oriented clustering", ICASSP 88,pp. 133-136, 1988.
8. 구준모, "매체변환기술과 응용", 전자공학회지 제20권 5호, pp. 9-15. 1993년 5월.

9. 구명완, "음성인식기술의 현황과 전망", 전자공학회지 제20권 5호, pp. 43-49, 1993년 5월.
10. 구명완, "음성인식기술의 현황과 전망", 정보과학회지 제11권 55호, pp. 21-34, 1993년 10월

김 희 동

• 1957년 11월 3일 생
• 1981년 : 서울대학교 전기공학과 (학사)
• 1983년 : 한국과학기술원 전기및전자공학과 (석사)
• 1987년 : 한국과학기술원 전기및전자공학과 (박사)
• 1987년~1992년 : 디지콤 정보통신연구소 연구소장
• 1992년~현재 : 수원대학교 정보통신공학과 조교수
※ 관심분야 : 음성신호처리, 정보통신망, 정보통신시스템

구 명 완

1960년 4월 26일생
• 1982년 : 연세대학교 전자공학과 (학사)
• 1985년 : 한국과학기술원 전기및전자공학과 (석사)
• 1991 : 한국과학기술원 전기및전자공학과 (박사)
• 1985~현재 : 한국통신 소프트웨어연구소 선임연구원, 자동통역연구팀 팀장
※ 관심분야 : 자동통역전화시스템 개발, 음성인식, Neurl Network 음성합성.