GOVP1199511101

제 2 차년도
최종 보고서

신경망 패턴인식에 의한 Voice Commander 시스템 개발

Development of Voice Commander System Using
Neural Network Pattern Recognition

(전자도우미 시스템 개발 부록 Ⅱ)

연 구 기 관
한국과학기술연구원
시 스 템 공 학 연 구 소

과 학 기 술 처

배 포 선

사본 번호	부수	배 포 처
1/100	1	시스템공학연구소 영구 보존용
2/100	1	시스템공학연구소 도서과 보관용
3/100 - 6/100	4	시스템공학연구소 연구관리과 보관용
7/100 - 8/100	2	시스템공학연구소 인공지능연구부 보관용
9/100 - 11/100	3	과학기술처
12/100 - 100/100	89	기타 배포처

제 출 문

과학기술처 장관 귀하

본 보고서를 "신경망 패턴인식에 의한 Voice Commander 시스템 개발 (2/2)" 과제의 2차년도 최종보고서로 제출합니다.

1995년 4월 30일

주관연구기관명 : 한국과학기술연구원
시스템공학연구소

총괄연구책임자 : 한 문 성

연 구 원 : 김 도 석
박 전 규
서 상 원
이 종 현

여 백

요 약 문

I. 제 목

신경망 패턴인식에 의한 Voice Commander 시스템 개발

II. 연구개발의 목적 및 중요성

음성은 인간의 가장 자연스런 의사소통의 수단으로서 문자의 발명이전부터 통용되어온 일상 회화의 중심수단이다. 이러한 음성이라는 매체는 따라서 내재의 자연성으로 말미암아 인간과 컴퓨터간의 대화수단으로 활용하고자 하는 노력이 그간 매우 활발하게 추진되어 왔는데, 이러한 음성에 대한 기대범위는 음성합성의 경우 정보 제공 수단으로, 음성인식 및 이해는 정보 입력 수단으로써 다양한 응용대상과 부가가치 창출의 가능성을 지니고 있다.

이러한 가능성을 인정, 미국에서는 70년대 초반부터 막대한 비용의 투자를 통한 대규모 프로젝트를 수행해왔으며, 컴퓨터 계산 성능과 디지탈 신호처리 기술의 비약적인 진보에 더불어 80년대에는 HMM이라는 강력한 확률통계적 음성인식 모델의 출현으로 실시간 음성인식이 가능해졌으며 이에 따라 다양한 응용시스템 개발이 이루어지고 있는 실정이다.

그러나 현재 실정으로는 음성을 이용 컴퓨터와 자연스런 대화를 진행하는 것은 불가능하며 따라서 선진국은 단기적으로 소규모 응용영역별 음성인식 및 이해시스템의 개발을, 장기적으로 음성타자기나 자동통역시스템 등의 목표를 설정해놓고 있다. 각국별로 대표적인 연구기관 및 연구목표를 보면, 미국은 DARPA 프로젝트를 통해 CMU, MIT, BBN, SRI 등에서 ATIS(Air Travel Information System)를, 일본에서는 대표적인 음성이해시스템으로 TOSBURG(Task-Oriented dialogue System Based on speech Understanding and Response Generation), 유럽공동체에서는 ESPRIT이라는 범 EC적 통합 프로젝트하에 SUNDIAL이하 다수의 프로젝트를 운영

활발한 연구개발을 수행중에 있다. 게다가 선진국들은 이미 음성인식 성능향상의 중요한 핵심요소중의 하나인 음성 데이타베이스를 국가적 차원에서 구축하고 CD-ROM을 통해 보급하는 등 연구에 활용하고 있으며, 80년대말 HMM의 성능한계를 인지, 자연언어처리나 언어처리 등 주변지식을 폭넓게 수용하고 신경망 등 새로운 계산 패러다임을 도입하는 등의 노력을 통해 발화에 내재하는 애매성을 극복하고 인식성능을 높이는 등 제2의 도약을 추진중에 있다.

국내에서는 각 연구소, 학교, 기업체 등에서 산발적인 프로젝트를 추진하다가, 90년대초 정보통신부산하 한국통신, 전자통신연구소 등을 주축으로 자동통역 전화 프로젝트가 수행중에 있지만, 아직 이렇다할 대표적인 음성인식 응용시스템이 부재인 상황하에서 음성인식 성능향상의 관건인 한국어 음성 표준 데이타베이스에 대한 연구도 부진한 실정이다.

당 프로젝트는 이러한 배경에서 디지탈 신호처리 기술, 음성인식 알고리즘, 실시간 시스템 구현 기술, 전화망 접속 기술 등 연속음성인식 및 이해에 필요한 요소기술을 개발하고 이에 기반한 소규모 응용시스템 개발을 목적으로 하고 있다. 이에 따라 기존 HMM이나 DTW 등의 인식기술과, 신경망을 융합하는 하이브리드형의 음성 패턴인식 알고리즘을 개발하고, 이를 통해 다양한 Voice Commander 응용시스템의 개발을 목적으로 하고 있다.

III. 연구의 내용 및 범위

본 연구의 내용 및 범위는 다음과 같다.

1. 이산(discrete) 및 연속(continuous) Hidden Markov Model(HMM) 등 기존 음성인식 알고리즘의 구현을 통한 화자독립 음성인식 시스템의 개발
2. 다층 퍼셉트론(Multi-Layer Perceptron), Spatiotemporal Network, Hidden Control Neural Network 등 신경망 패턴인식 알고리즘의 개발

3. 화자독립 음성인식 응용 시스템의 개발

- 전화선을 통한 질의 응답형 Voice Commander 음성 응용 시스템의 프로토타입
- 핵심어 추출(keyword spotting)에 의한 로보트 제어용 Voice Commander 인터페이스

IV. 연구개발결과 및 활용에 관한 건의

본 연구에서 개발되는 신경망 패턴인식 알고리즘은 군사용이나 상업용에서 사용되는 다양한 패턴인식 시스템의 핵심기술로써 응용될 수 있으며, 이에 따라 개발된 질의 응답형 Voice Commander 시스템이나 로보트 제어용 음성 인터페이스는 차세대 인간-기계간 사용자 인터페이스의 하나로써 광범위하게 응용 가능하다. 이러한 시스템은 특히 연속음성인식이 가능한 분야, 컴퓨터의 인터페이스로써 다른 매체보다 입력속도의 신속을 요할 경우, 손이나 발을 다른 용도에 사용하고 있을때 동시입력수단으로 응용될 수 있으며, 키보드에 비해 훈련이 필요치 않는 등의 잇점이 있다.

또한 당 연구에서 개발된 전화망을 통한 응용시스템은 공중전화망을 통한 응용시스템이라는 관점에서 무한한 응용의 가능성이 있으며, 각종 전자기기를 무인 또는 원격제어하는 시스템등에 널리 응용될 수 있다.

여 백

Summary

I. Title

Development of Voice Commander System Using

Neural Network Pattern Recognition

II. Objectives

Voice, as the most natural means of communications has been used as a key tool for conversations before the invention of characters. Active efforts has been made to use it for communications between man and computer due to its intrinsic nature of communication media.

Since the scope of expectation can be extended to the tool for giving information with speech synthesis and to the tool for inputting information with speech recognition and understanding, voice has various application areas and the possibilities of creating high value-added applications.

U.S.A noticing these possibilities has done big projects by investing lots of money since 1970. With rapid development of high performance computer and digital signal preocesing technology along with the presence of powerful statistical HMM (Hidden Markov Model), the time has come to realize the real-time speech recognition and vaious speech related application systems.

But still it is not possible to communicate computer with natural dialog. The developed countries set the development of small

vocabulary speech recognition and understanding systems for short-term target, and the development of voice typewriter and automatic translation system for long-term target. U.S.A set the goal of ATIS (Air Travel Information System) through the DARPA by CMU, MIT, BBN, SRI, and Japan has developed a typical speech understanding system, TOSBURG (Task-Oriented dialogue System Based on speech Understanding and Response Generation), while EC has done active research and development under SUNDIAL projects in EC projects.

Furthermore they developed the nationwide speech databases, the key factors to improve the speech recognition performance, and made CD-ROM copies distributed for research. At the end of 1980s, they have pursued the succeeding jump in recognition performance by accepting natural language processing and its related knowleges and by introducing the new computational paradigms such as neural networks, fuzzy theory, etc..

Meanwhile domestic laboratories, universities, and companies have done speech projects independently, and Korea Telecom and ETRI have done jointly automatic translation project in early 1990. But the research on Korean standard speech database is still needed and the useful application system based on speech recognition.

Upon this research trend, this project has the goal of developing basic technologies needed continuous speech recognition and understanding such as digital signal procesing, speech recognition algorithm, real time implementation, and telephone line interfacing technology.

And thus the development of hybrid speech pattern matching algorithms, i.e. combination of HMM or DTW and Neural Network is

the first goal and the application to the various Voice Commander system is the another goal of this research.

III. Scope and Contents

1. Development of speaker-independent speech recognition system using discrete and continuous HMM
2. Development of neural network pattern recognition systems such as Multi-Layer Perceptron, Spatiotemporal Network, and Hidden Control Neural Network
3. Development of application systems based on speaker-independent speech recognition
   - the voice commander prototype directly applicable to various application systems through telephone line, with simple query processing
   - the speech interface for robot control by means of keyword spotting

IV. Results and Recommendations

Neural network pattern recognition algorithm as a primary result through this research can be applied both for military and commercial purposes, and the speech Q&A Voice Commander system and robot control system can be widely utilized as a new man-machine interface.

Also this system can be used as a simultaneous input device and doesn't have any keyboard trainning when faster input to computer is need and when both hands and feet are occupied.

The developed application system using telephone line has unlimited possibilities with its usage of telephone line, and can be used for controlling various electronic equipments and unmanned or remote systems.

Contents

목 차

1 장

서 론

음성은 인간의 가장 자연스런 의사소통의 수단으로서 문자의 발명이전부터 통용되어온 일상 회화의 중심수단이다. 따라서 음성매체는 자체가 가지고 있는 자연성으로 말미암아 컴퓨터 관련 분야에서는 이를 인간과 컴퓨터간의 대화수단으로 활용하고자 많은 연구를 투입해오고 있다.

음성처리에 대한 기대범위는 음성합성의 경우 컴퓨터로부터 사용자에게 정보를 제공하는 수단으로, 음성인식 및 이해는 컴퓨터에 정보를 입력하는 수단으로써 다양한 응용시스템과 부가가치를 창출할 수 있는 가능성을 내포하고 있다. 음성입력은 연속음성인식이 가능한 분야를 대상으로, 컴퓨터의 인터페이스로써 다른 매체보다 입력속도의 신속을 요할 경우, 손이나 발을 다른 용도에 사용하고 있을때 동시입력수단으로 활용이 가능하며, 특히 키보드에 비해 훈련이 필요치 않는 등의 잇점이 있기때문에 일반 범인들이 쉽게 컴퓨터를 사용하는 수단으로 그 가치가 높다. 그러나 음성인식 및 이해시스템을 실제로 구축할 경우에는 많은 분야의 학문적 통합과 교류가 필요한데, 현재 실정으로는 음성에 의해 컴퓨터와 대화를 자유롭게 수행하는 것은 불가능하며 음성부호화나 합성등의 경우도 똑같은 양상을 띠고 있다.

미국, 일본, 유럽 등에서는 일찌기 음성처리의 중요성을 인정하여 70년대 초반부터 대규모 프로젝트를 수행하기 시작했으며, 하드웨어적 측면에서는 컴퓨터 계

산 성능과 디지탈 신호처리 기술의 비약적인 진보, 소프트웨어적으로는 80년대초 Hidden Markov Model(HMM)이라는 강력한 확률통계적 음성인식 모델의 출현으로 말미암아 실시간 음성인식이 가능해지는 등 충분한 기술적 발전 교두보가 형성되어 있는 상황이다.

이러한 배경을 고려 단기적으로는 소규모 인식어휘를 대상으로 하는 영역제한형 음성응용시스템을, 장기적으로는 음성타자기나 자동통역시스템 등을 개발하는 것이 세계적인 추세이다. 음성인식 및 이해의 선두주자인 미국은 DARPA 프로젝트를 통해 CMU, MIT, BBN, SRI 등에서 ATIS(Air Travel Information System)를, 일본에서는 ATR이라는 기관을 축으로 일본내 전 연구소, 학계 등이 협력 자동통역전화 프로젝트를 수행하고 있으며, 유럽공동체에서는 ESPRIT이라는 범 EC적 통합 프로젝트하에 SUNDIAL이하 다수의 프로젝트를 운영 활발한 연구개발을 수행중에 있다. 또한 선진국들은 이미 음성인식 성능향상의 중요한 핵심요소중의 하나인 음성 데이타베이스를 국가적 차원에서 구축하고 CD-ROM을 통해 보급하는 등 연구에 활용하고 있으며, 80년대말 HMM의 성능한계를 인지, 자연언어처리나 생물학 또는 인지학 등 주변지식을 폭넓게 수용하는 노력을 통해 발화에 내재하는 애매성을 극복하고 인식성능을 높이는 등 제2의 도약을 추진중에 있다.

즉, 인간과 유사한 수준의 음성인식은 기존의 방법론만으로는 한계가 있는데 이것은 근본적으로 인간의 귀가 수학적으로 정량화하기 어려운 조음효과에 대단히 민감한 구조를 갖추고 있다는 것이다. 수년동안 언어학자들은 이러한 음성생성과정을 연구하고 보다 자연스런 음성인식 및 합성음을 위해 정형화된 규칙을 연구해왔는데, 이러한 규칙생성은 매우 긴 시간을 요하고 장황한 작업이기 때문에 새로운 전기마련을 위해 생물학적 청각 시스템의 모델링에 의한 방법과 이를 계산 패러다임화한 신경망을 도입하기에 이르렀다.

한편, 음성입력시스템 보급을 지원하는 요인으로 컴퓨터의 고속화와 멀티미디어화의 고속성장을 통해 범용 웍스테이션에 음향 입출력을 표준장비로 채택하는 추세가 일반화됨으로써, 음성신호를 받아들이고 분석하는데 필요한 음성처리비용이

저하되고 있다. 이러한 경향은 이제까지의 시뮬레이션이나 이론위주의 연구에서 탈피, 음성으로 운영되는 컴퓨터를 가능하게 하는 한 요인이 되는 것이다.

당 프로젝트는 이러한 배경에서 디지탈 신호처리 기술, 음성인식 알고리즘, 실시간 시스템 구현 기술, 전화망 접속 기술 등 연속음성인식 및 이해에 필요한 요소기술을 개발하고 이에 기반한 소규모 응용시스템 개발을 목적으로 하고 있다. 이에 따라 기존 HMM이나 DTW 등의 인식기술과, 신경망을 융합하는 하이브리드형의 음성 패턴인식 알고리즘을 개발하고, 이를 통해 다양한 Voice Commander 응용시스템의 개발을 목적으로 하고 있다.

당해년도까지의 연구결과를 요약하면 다음과 같다.

1. 기존 음성인식 알고리즘의 구현을 통한 화자독립 음성인식 시스템의 제안 (1차년도)
   - Dynamic Time Warping 알고리즘에 의한 화자독립 단어인식 알고리즘 구현
   - 신호처리에 의한 복합적 음성특징추출 및 벡터양자화에 의한 코드북작성 등 전처리를 기반으로한 이산(discrete) HMM의 구현 및 화자독립 음성인식 알고리즘 구현
2. 신경망 옴성인식알고리즘의 구현 및 신경망음성인식 프로토타입 제안 (1차년도)
   - Multi-Layer Perceptron을 통한 음성인식 알고리즘의 구현
   - Spatiotemporal Network의 구현
   - Hidden Control Neural Network의 구현
3. 실시간, 화자종속, 소용량 음성 DB (100단어이하) 단어 인식소프트웨어 개발, 화자독립 인식모델 연구 및 응용시스템 연구 (1차년도)
   - 유닉스 명령어 인식 및 편집기 vi 명령어 인식에 의한 유닉스 운영체제 구동소프트웨어의 구현

- 음성 segmentation 및 복합 음성특성추출에 의한 어학실습기의 프로토타입구현
- 음성질의응답 처리시스템을 위한 초보시스템의 설계

4. 이산 및 연속(continuous) HMM 등 기존 음성인식 알고리즘의 구현을 통한 화자독립 음성인식 시스템의 개발 (2차년도)

5. Multi-Layer Perceptron, Spatiotemporal Network, Hidden Control Neural Network 등 신경망 패턴인식 알고리즘의 개발 (2차년도)

6. 화자독립 음성인식 응용 시스템의 개발 (2차년도)

   - 전화선을 통한 질의 응답형 Voice Commander 음성 응용 시스템의 프로토타입
   - 핵심어 추출(keyword spotting)에 의한 로보트 제어용 Voice Commander 인터페이스

본 보고서의 2장에는 본 프로젝트의 기술적 배경이 되는 신경망 패턴인식과 Voice Commander 시스템의 개요를 음성처리 전반 특히 음성인식과 이해의 관점에서 기술하고 있으며, 3장과 4장은 각각 본 프로젝트에서 개발된 음성인식 응용 시스템인 공중전화망을 통한 질의 응답형 음성인터페이스 시스템과 로보트 제어용 음성 인터페이스 시스템의 실제 구현사례를 중심으로 기술하고 있다. 마지막 5장에서는 결론과 앞으로의 연구 및 개선 방향 등에 대해 기술한다.

2 장

음성언어처리와 Voice Commander 시스템

2.1 음성 언어처리 개요

음성처리와 음성인식과의 관계는 당 연구 프로젝트의 1차년도에 문헌조사와 함께 세밀한 고찰이 있었으므로 본 보고서에는 특별히 언급하지는 않을 것이다. 따라서 본 보고서에서는 1차년도에서 다루지 않은 음성의 언어처리적 관점과, 음성인식의 미래지향적인 요소인 음성언어처리와 자연언어처리와의 관계를 중점적으로 기술할 것이다.

음성처리의 분야에는 음성으로부터 자연언어로의 매체변환을 도모하는 음성인식이 있는데, 그 궁극적인 목표는 연속음성인식에 기반한 음성타자기나 음성 문서 편집기등이 있으며, HMM으로 대표되는 확률통계적인 모델에 기초하는 대어휘 음성인식 연구가 활발히 수행되고 있다. 최근에는 HMM의 한계를 인지 계산 언어학 등에서 사용한는 LR parsing이나 문맥자유문법(Context Free Grammar, CFG) 등 단순한 수준의 언어처리 모델을 사용 인식률 제고를 위한 연구가 수행되는 예도 있다[16].

표면적인 매체변환을 나타내는 연속음성인식에 반해 발화의 의미까지도 추출하는 음성이해, 음성의 언어적 측면을 중요시하는 음성언어시스템(spoken language system), 컴퓨터와의 인터페이스를 의미하는 음성대화(speech dialogue) 시스템 등 음성처리와 자연언어처리의 융합에 관한 연구도 활발히 진행되고 있다[25].

음성언어처리와 자연언어처리는 언어정보를 취급하는 점에서 공통점이 있다. 음성신호는 발화자의 개인성이나 감정등의 정보를 포함하고, 환경소음등의 영향을 받기 때문에 문자기호열에 비해서 다양한 변이(variation)를 나타낸다. 또한 연속음성인식의 경우 전후 음소환경의 영향을 통해 조음결합이라는 현상을 나타내기 때문에 음향분석의 결과로 관찰되는 특징으로는 음소환경이 불명확하게 나타나고 따라서 음소에 대응하는 특징 파라미터의 성질이 단독으로 발성되는 것이 아니기 때문에, 연속음성을 기계적인 음소등의 이산적인 기호로 변환할 수가 없다. 즉, 음성은 인터페이스로 사용시 자연성이나 신속성 등의 잇점이 있는 반면, 분석을 위해 수많은 계산시간과 비용을 요하게 되며, 음성인식과 이해의 애매성이나 에러보다 우선 해결해야 할 문제가 바로 이러한 계산 병목현상이 되는 것이다[20].

음성처리에 이용되는 언어정보의 유형에는 음소의 설정이나 단어의 음운정보의 표현(음운론), 단어의 나열과 문장의 관계를 규명하는 구문정보(통어론), 의미정보(의미론), 문맥정보나 어용론(pragmatics) 이 있는데, 현재는 자연언어처리에서 개발된 각종 언어정보나 처리기구가 이용가능하게 되었다. 이들중에는 음향특징 파라미터와 유사한 음소등에 관한 저차언어정보, 기호(symbol)화된 후의 구문, 의미, 문맥정보과 같은 고차 언어정보 등이 있는데 적절한 활용을 통해 음성인식, 이해, 대화처리 등의 성능향상이 가능할 것으로 보고 있다.

그림 2.1은 전형적인 음성이해의 처리과정을 도시하고 있다. 입력된 음성은 음향분석, 음소인식과 음성구간추출(segmentation)을 거쳐 음소 lattice로 출력된다. 다음 음소 lattice를 해석하여 단어인식을 수행하며 단어 lattice에 대해 구문해석과 의미해석을 수행해서 발화문장의 의미표현을 구한다[41].

연속음성이해는 1970년대 미국의 ARPA 음성이해 프로젝트에서 본격적으로

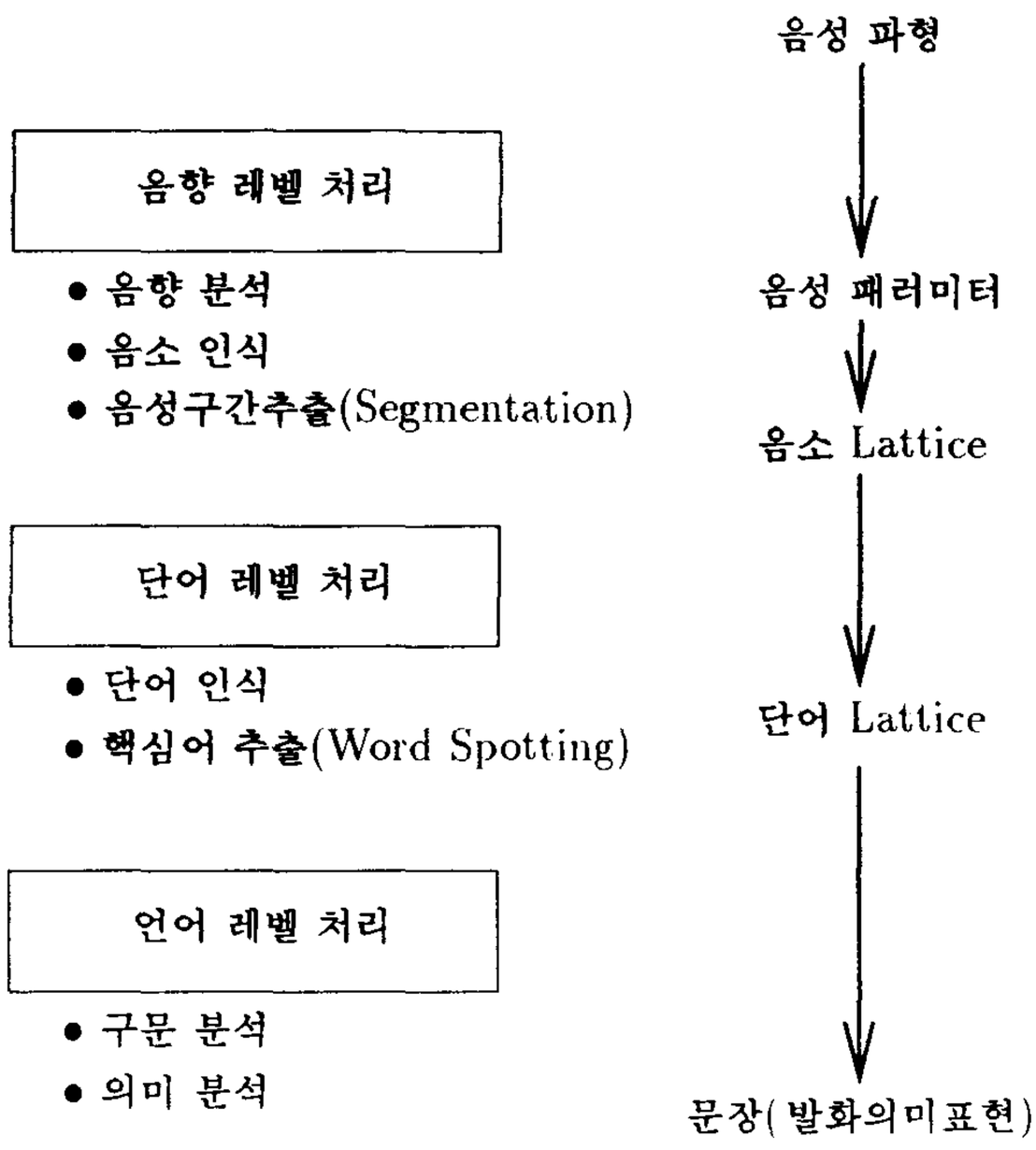


그림 2.1: 전형적인 음성이해시스템의 처리과정

시작되었으며, HARPY, HWIM, Hearsay 등 응용시스템을 명확히 설정하고 대규모 음성이해 시스템 구축을 목표로 음성처리와 자연언어처리의 통합을 시도하였다. CMU의 HARPY를 제외하면 대부분의 시스템은 미완성인 채로 종결되었지만 이러한 프로젝트의 결과 구문이나 의미레벨 지식의 유효성이 명확하게 되었다. 예를들어 Hearsay-II의 개발을 통해 각종지식을 이용하는 Blackboard 모델이 제안되고, 인공지능이나 자연언어처리의 진보에 큰 공헌을 했다. 또 음향분석, 음소, 단어레벨 패턴인식등 저차처리의 정도가 완전해야만 고차처리의 가치가 있음을 증명하였으며, 지식기반시스템의 실패를 거울삼아 그 반성과 함께 알고리즘과 데이타 중심의 확률에 기반한 음성연구가 시작되는 계기가 되었다.

1985년 초 DARPA의 음성인식 프로젝트는 음성신호를 자연언어로 변환한다는 전제하에 단어나 문장의 인식성능 향상을 목표로 대어휘 불특정화자 인식에 주력하게 되었는데, 그 중심이 HMM이다. 음성 DB의 정비, 컴퓨터 계산능력의 고조

를 배경으로 음성인식의 핵심기술과 고정도의 HMM 기반 불특정어휘 연속음성인식 시스템이 개발되었다. CMU의 SPHINX[16]나 BBN의 BYBLOS[5]등이 그 대표적 시스템으로서 확률통계적인 접근법에 따라, 전통적인 문법구조나 의미구조를 추출해서 음성을 문자열로 변환하는데 주로 단어나 음운등의 bigram, trigram을 많이 사용하고 있다. 이러한 표층적인 언어정보의 확률통계적 제약을 이용 저차레벨의 음성인식오류를 수정하고 애매성을 보완하는 등 문장 인식률을 향상시켰다.

그러나 HMM에 기반한 확률통계적인 음성인식 모델도 그 상세화와 부단한 개량 노력에도 불구하고 성능향상이 포화에 이르렀기 때문에 1988년경부터 음성인식과 자연언어처리의 통합에 대한 중요성을 인식, 음성언어시스템이 연구되기 시작했다. 또한 자연언어처리기술에 있어서는 종래의 표기언어(written language)로부터 발화언어(spoken language)를 연구대상으로 채택하야 한다는 새로운 자연언어처리의 중요성이 지적되어 음성과 자연언어처리의 유기적인 융합이 연구과제로 부상되었다[5][11][25].

2.1.1 음성인식을 위한 언어모델

자연언어처리에서 사용되는 언어모델(문법)은 언어의 타당한 해석적 구조를 얻는데 이용되는데 반해, 음성인식에서 사용되는 언어모델은 타당하지 않은 문자열을 배제해서 탐색공간을 줄임으로써 음성인식의 불완전한 부분을 보조하는 역할을 한다.

80년대초의 음성인식시스템에는 확률모델을 이용해서 음성의 애매성을 처리하였는데 이러한 시스템은 주로 음소레벨의 처리를 위해 HMM을, 단어나 문장 레벨의 처리는 언어모델에 의해 가능한 단어열에 대한 제약을 두는 것이었다. 이러한 제약은 전통적인 언어처리 구조와 달리 단어(품사)의 bigram이나 trigram 문법 등 단어쌍의 문법을 이용하는 경우가 많다[16]. 이러한 bigram이나 trigram 등의 핵심인 연쇄확률은 대량의 텍스트를 통계처리함에 의해 추정된다. 현재는 정규문법(regular grammar)이나 문맥자유문법을 언어 모델로 채택하는 시스템도 증가하고 있다. 이때, 연속음성인식은 상술한 각종의 언어모델에 기반하여 단어후보열(

단어 lattice)을 parsing하고 문장 전체의 완성도를 최대로 하는 단어열을 발견하는 문제로 정의할 수 있다[38].

음성이나 단어 등 저차 인식과정에서 발생하는 인식오류나 애매성에 대처하기 위해 음운이나 단어 등의 후보열로부터 lattice의 parsing을 효과적으로 수행하기 위해 여러가지의 방법이 제안되었는데, 이러한 parsing의 방식에는 인식단어검출과 단어열 평가를 구별하여 처리하도록 하는 lattice parsing 알고리즘과 두가지를 동시 진행하는 연속단어인식 알고리즘의 두종류가 있다. 이외에 자연언어처리에서 적용하는 LR parser, Earley 등의 일반적인 알고리즘등이 적용가능하며, 실시간 처리를 목적으로 하는 고속해석방법이 제안되기도 하였다. 또 유연한 제어가 가능한 Chart parser, C.Y.K 알고리즘도 음성인식에 이용되고 있다[22][23].

음성인식의 parsing시 위에 기술한 저차 음성 인식 처리의 불완전성에서 기인하는 애매성과 아울러, 단어열등의 고차 해석시 발생하는 애매성도 있다. 음성인식은 주로 전자의 처리에 주안을 두고 있지만, 음성언어시스템의 구축시는 후자의 애매성 처리와 그에 따른 수많은 계산량이 문제가 된다. 이를 위해 대어휘 연속음성인식 시스템에는 trigram등의 언어적제약에 의해 연결음성에 포함된 단어의 후보열을 탐색하고, CFG를 모델화해서 탐색후 얻어지는 N-best 단어열을 일반화된 LR parser에 따라 고속 parsing을 적용하는 방법이 제안되었다. 또한 위에 기술한 언어모델에 적용하는 탐색법에는 최적경로를 효율적으로 구하는 beam 탐색을 채용하는 것이 대부분이며, Dynamic Programming(DP)을 이용하는 A* 탐색도 연구중에 있다[38].

일종의 확률 문법에 속하는 bigram이나 trigram 문법을 보다 구체화하는 연구도 있는데, 이때 언어모델의 역할은 발화문장속의 비분법적 요소배제와 탐색공간 축소를 통해 인식률을 향상시키는데 있다. 그러나 이러한 목적으로 자주 이용되는 CFG등에도 문법적제약이 불완전하기때문에 문장요소를 제대로 갖추지 않은 문장을 생성하는 문제가 있기 때문에 CFG등의 문법을 확률화하는 확률 CFG에 의한 음성언어모델화 방법이 제안되고 있다[23]. 또한 문법 학습시 필요한 대량의 학습용의 데이타를 자동으로 작성하는 방법도 검토중에 있다.

2.1.2 음성이해와 음성대화

음성언어와 문자언어의 다른점은 음성언어가 대화적인 매체라는 점과, 사용법이 다양하고 대화에 포함되는 제스츄어나 청각 활용에 따른 영향을 받아 유동적으로 변화되는 성질이 있다는 점이다.

일상생활중에 나타나는 자유발화음성(spontaneous speech)에는 감탄사, 휴지, 호흡음 등이 포함되고, 한국어나 일본어의 경우 조사의 탈락이나 어구 도치 현상이 빈발하여 문법적 기술이 곤란한 측면이 많아서 종래의 자연언어처리 구조에서는 이를 대처하기가 어렵다. 이를 위해 음향레벨에서의 처리수단인 garbage 모델이 제안되었으며, 문장보다 작은 단위인 구구조(phrase structure)에 문법을 적용하는 유연한 parsing 방법이나 keyword를 인식기본단위로 삼는 방법이 제안되고도 있다. 자유발화음성에서 중요시되는 또 한가지는 robustness인데 미등록된 단어에 대처하는 방식에 대한 연구도 보고되고 있다[42].

인공지능이나 자연언어처리에서도 담화이해나 대화처리에 대한 연구가 수행되었지만 인간과 인간의 대화해석이나 모델화가 중심으로 주로 문자언어를 처리대상으로 삼아 음성인식 에러나 애매성에 관한 검토는 이루어지지 않고 있었다. 인간은 음성을 듣는 가운데, 발화의 상황이나 문맥, 상식이나 대화에 등장하는 화제에 관한 지식을 이용해서 다음 발화를 예측 또는 추론을 행한다. 다음 발화를 예측하기란 쉬운일이 아닌데 사실 인간도 음성을 정확히 듣지 못하는 예가 많기때문에 음성인식의 robustness 실현은 당연히 쉬운일이 아니다. 따라서 상황이나 문맥정보를 통해 화제에 관한 지식을 이용하는 것이 필요한데, 대화의 문맥정보를 이용하는 대화음성이해, 대화 상황을 응용하는 질문, 확인에 의해 애매성을 유효적절히 없애는 대화시스템 실현이 필요하다.

음성언어는 억양이나 강세등과 같은 운율과 같은 언어정보를 지니며, 구의 경계나 강조점, 감정 등의 정보를 전달하게 된다. 이러한 운율정보를 이용해서 연속음성중의 구나 절을 검출하는 방법도 보고되고 있다. 즉 음성대화시스템 구축시에는 운율정보를 이용해서 강조점이나, 의도, 감정 인식을 적극적으로 활용하는 것이

중요하다는 것이다. 일상적인 회화에서 빈번히 발생되는 "에", "저" 등의 비언어적 음성 인식도 중요한 검토대상이다. 이러한 면에서 운율정보등의 이용은 음성인식률 향상에도 유효하지만 일상회화에 빈번히 사용되는 정보라는 측면에서 신속하게 자연스런 컴퓨터와의 통신을 실현하는 핵심기술이다.

2.2 음성이해와 대화 시스템

2.2.1 연구동향

컴퓨터 하드웨어 성능 향상과 멀티미디어화의 진전에 의해 오디오 인터페이스를 표준장비로 채택하는 웍스테이션이 보급되고, 컴퓨터에서 간편하게 음성처리가 가능한 환경이 제공되고 있다.

HMM에 기반한 강력한 음성인식도 고성능 웍스테이션상에서 실시간으로 작동이 가능하게 되었으며, 음성에 의한 컴퓨터와의 대화도 실현가능성이 높아지게 되었다. 음성관련 국제회의에는 ICASSP, ICSLP, DARPA Workshop이 있는데 최근에는 음성언어시스템과 실시간시스템 개발에 대한 연구가 주류를 이루고 있는데, 즉 종래 신경망이나 HMM에 대한 연구로부터, 실제적 응용을 위한 음성언어처리 시스템에 시선이 집중되고 있는 것이다.

음성이해 및 대화 시스템 연구는 CMU, SRI, BBN, MIT 등이 중심이 되는 DARPA 프로젝트의 규모가 가장 크다. 이러한 연구기관들은 HMM을 기반으로 음성인식률 향상을 위해 음성인식 처리후에 복수의 발화후보문을 자연언어처리부에 보내서 음성인식의 애매성을 해결하고 있다. MIT에서는 음성인식과 자연언어처리등을 접속 및 통합하는 지적 음성 언어 시스템인 VOYAGER[45]가 개발되었으며, CMU는 SPHINX[16]를 기반으로 하는 multimodal 인터페이스를 개발중에 있다. 한편 유럽에는 multimodal 대화연구에 EC의 거의 모든 나라가 참여해서 음성대화연구를 수행하는 SUNDIAL 프로젝트가 수행중에 있다[25].

일본에서는 ATR의 자동번역전화 프로젝트에서 영어, 일어, 독일어의 3개국어를 대상으로 국제회의를 수행할 수 있도록 하는 실험시스템인 ASURA가 개발되었다[38]. ASURA는 음성인식, 기계번역, 번문전송, 번문음성합성을 수행한다. ATR이외의 연구기관으로는 ETL, NEC, KDD, 동지 등이 있으며, 전자협을 중심으로 음성데이타베이스를 작성하고 있다.

2.2.2 DARPA의 음성언어 시스템

1970년대 ARPA 프로젝트의 종료후 1987년에 DARPA의 음성언어이해 프로젝트가 재개되었다. 컴퓨터 계산능력의 증대와 강력한 HMM을 배경으로 대화음성에 대한 데이타베이스가 정비되고 자유발화음성을 인식대상으로 CMU, BBN, SRI, MIT 등에서 음성이해 시스템 개발을 추진하고 있다. 또한 공통의 목표시스템으로써 ATIS(Air Travel Information System)를 설정하여 사용자와 컴퓨터간의 대화를 시뮬레이션하고 있다.

BBN은 HARC(Hear And Respond to Continuous speech)라는 음성대화시스템을 개발하고 있는데, HMM에 기반한 BYBLOS라는 연속음성인식시스템을 사용하고, DELPHI라는 언어처리시스템을 사용하고 있다[5]. BYBLOS는 입력음성에 대해 bigram 모델에 기초한 forward-backward 탐색을 수행하며 N-best의 후보를 출력한다. DELPHI는 N개의 문장후보를 재평가해서 동사의 범주(category) 분류를 위해 설정된 mapping unit에 기초한 해석법을 사용하여 데이타베이스 검색언어인 SQL로 변환하는 시스템이다. BBN은 언어처리를 위해 mapping unit이란 동사분류단위를 명사나 형용사에도 확장적용하고, 어순 이용방법이나 부분 parsing등의 도입을 검토하고 있다(그림 2.2).

CMU에는 자유발화를 대상으로 하는 음성대화시스템인 PHOENIX를 개발하고 있다. ATIS를 목표시스템으로 삼고 음성인식부에는 CMU의 SPHINX[42]를 사용 입력음성을 의미 phrase slot으로 변환하고, 정보검색언어인 SQL로 변환하고 있다. 이 시스템은 잡음이나 환경잡음등에 대한 garbage HMM 모델을 채용해

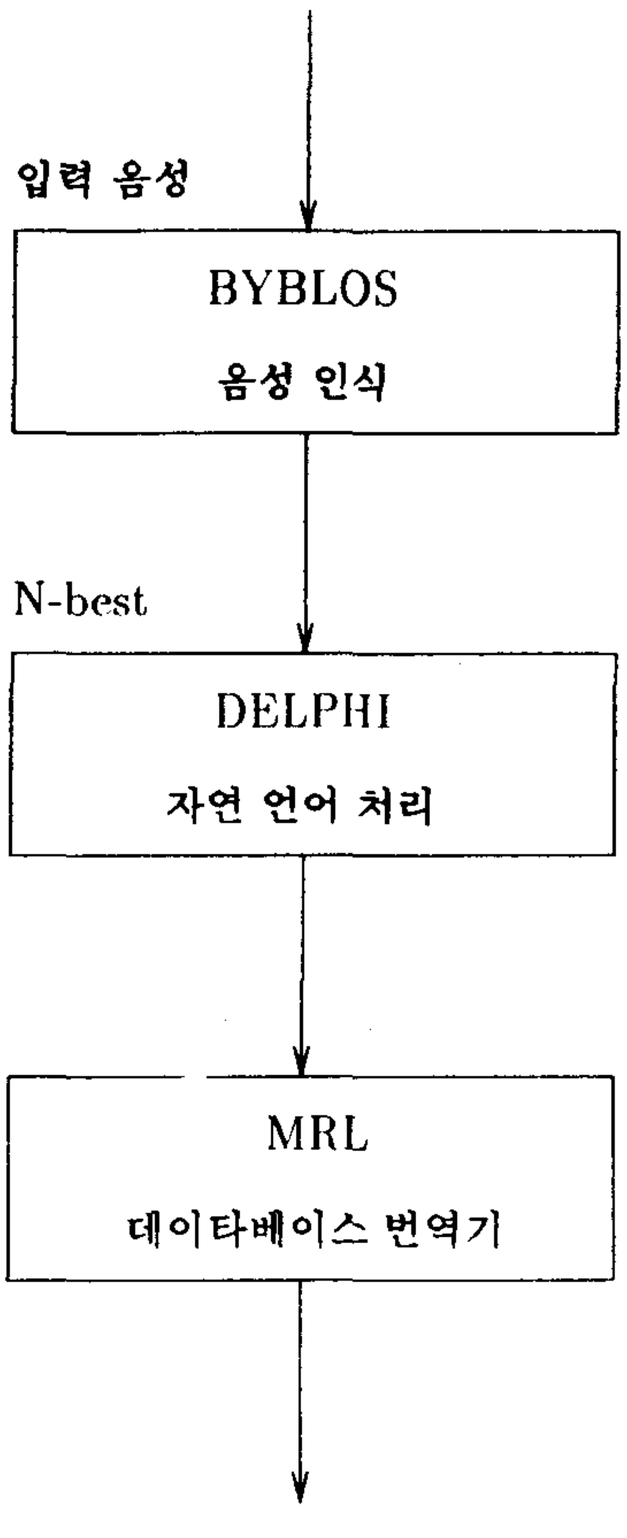


그림 2.2: BBN의 음성이해 시스템인 HARC의 처리 흐름도

서 자유발화를 처리한다. 구구조 문법을 이용하여 문법을 상태네크웍으로 표현하는 해석방식을 사용하고 있으며, 음운의 확률적 bigram 모델을 이용해서 미지어에 대처하고 있다.

2.3 Voice Commander 시스템

Voice Commander 또는 음성 명령어 시스템은 음성을 통해 간단한 명령을 입력시킬 수 있는 장치로서 일반적인 범용 컴퓨터에서부터 소형 칩을 대상으로 구현되

어 각종 전자기기나 제어시스템으로 활용가능한 시스템이라 할 수 있다.

입력매체를 음성으로 사용하기 때문에 자동적으로 근거리 조작의 잇점뿐만 아니라 소형 무전기나 워키토키 등을 이용 장거리 응용시스템에서도 다양하게 적용될 수 있다. 즉 공간적 제약이 없으며 지리적이거나 물리적인 제약사항을 극복할 수 있는 등 응용범위가 넓은 장점이 있다. 또한 기존 컴퓨터 입력장치인 키보드나 마우스에 비해 각종 기기 작동시 두손이 자유로이 다른 일에 종사할 수 있는 장점도 있다.

Voice Commander의 응용영역은 다음과 같은 것이 있다.

- 음성을 통해 가전제품을 조작할 수 있는 가전용 리모콘
- 음성을 통해 전화번호 또는 사전 기억된 명칭을 통해 전화를 거는 음성 다이얼링
- 장애자를 위한 전자기기 제어 또는 교육 시스템
- 손이나 눈이 다른 업무에 종사시의 각종 장비 및 전자기기 조작
- CAD 시스템의 명령어나 키보드의 function 키 대용
- 문서 편집기나 데이타베이스 등에 대한 명령어 입력
- 로보트 제어

미국은 이미 Voice Commander와 관련한 다양한 응용기술을 실용화해서 시장에 판매하고 있는데 이들은 대부분 소규모 어휘, 화자독립 또는 화자적응, 고립단어를 인식 대상으로 하는 패턴인식에 기반한 시스템이며, 현재까지는 자유발화음성이나 비문법적인 요소, 미지어(unknown vocabulary), 환경 잡음 등에 취약한 점이 있는데 이러한 점이 실용화의 주요 장애요인이 되고 있다.

2.3.1 핵심어구 추출

단어 인식을 통해 연속음성인식에 가장 근접한 시스템이 핵심어추출(keyword spotting) 기법을 채택한 인식방법인데, 앞절에서 소개한 TOSBURG와 같은 시스템이 전형적이라 할 수 있다[38][39][40]. 핵심어 추출이란 글자그대로 사용자의 발화문장 중 의미없는 또는 대화진행에 불필요한 단어들을 배제하고 필요한 어휘만을 추적해서 인식하는 방법으로 연속음성중의 핵심어를 인식함으로써 단어인식을 연속음성인식의 범주로 확장하고자 하는 시도로 볼 수 있다.

이러한 핵심어추출은 일반적인 음성인식이나 이해에서 필요한 문법요소 (grammar)나 단어 끝점검출(endpoint detection)등에 소요되는 부담을 줄이며, 따라서 연속음성인식시의 탐색공간을 줄이는 효과가 있다.

현재까지 보고되고 있는 핵심어추출 방법에는 Level-Building DTW에 의한 것과 비터비 탐색에 근거한 HMM 등이 일반화되어 사용되고 있으며, 신경망에 의한 방법, 신경망과 HMM 등의 하이브리드형 기법도 보고되고 있다. HMM을 이용할 시 음성신호는 곧 비핵심어 신호와 핵심어 신호의 합성된 신호로 보여지며, 대부분 garbage 모델을 설정해서 비핵심어구를 판단하고 있다. 이러한 관계를 개념적으로 다음과 같이 나타낼 수 있다[44] [19].

$$
\begin{aligned}
speech\ signal &\leftarrow \{non\_keyword\ signal\} + \{keyword\ signal\} \\
garbage\ model &\quad\ (non\_keyword\ template) \\
&\leftarrow silence + (channel\ noise) + (extra\ speech)
\end{aligned}
$$

garbage 모델에 대한 개념은 본 프로젝트에서 개발한 로보트 제어용 음성인터페이스 시스템에서도 적용한 바, 제 4장에서 연속 HMM의 구조와 시스템 구현알고리즘에서 자세히 설명할 것이다.

이러한 핵심어추출법에서 고려할 사항들은 핵심어추출과 관련하여 추출에러를 최소화하는 것, 핵심어구를 정확히 추출하기 위한 모델링 방법, 환경변화 및 잡음

에 대한 대처 방안, 연속음성인식으로의 확장성 고려 등으로 요약할 수 있다. 현재까지는 인식단위면에서 단어 또는 단어를 보다 작은 단위로 갈라놓은 subword 모델을 채택하거나, 현재의 음소를 기준으로 전후 하나씩의 음소를 묶어서 모델링하는 triphone 법, 유사한 방식으로 전후의 문맥을 동시에 고려해서 학습 및 인식을 수행시키는 문맥 종속(context dependent) subword 단위, 일반적인 음성인식시 사용되는 음향인자(acoustic unit) 등을 인식단위로 하는 시스템들이 보고되고 있다. 인식 모델은 주로 Level-Building DTW, HMM, 신경망, 및 이들의 통합형 시스템이 사용되는데 주로 핵심어와 비핵심어의 두가지 class를 분류하는데 중점을 두고 방법상 general probability descent, linear discrimination, 최우추정법(Maximum Likelihood Estimation, MLE) 등을 적용한다.

대표적인 시스템으로는 미국의 AT&T 벨 연구소에서에서 Wilpon등이 개발한 시스템이 있는데 이것은 전화선을 통한 입력음성에 대해 다섯개의 핵심어와 숫자를 인식대상으로 하고 있다[44]. 비핵심어는 무음(silence), 전송잡음(transmission noise), 인식대상외 단어(extraneous word)로 모델링되며, 핵심어와 비핵심어에 대해 서로 독립적인 HMM을 구성하고 있으며 후처리를 통해 두 그룹간의 변별력을 향상시키는 방법을 채택하고 있다. 약 7890문장을 대상으로 하는 실험에서 화자독립으로 95.1%의 단어 인식률을 나타내고 있다. 다음에 실험 및 학습시 사용된 예문의 간단한 예를 보인다.

- **<silence>** *collect* call please
- I want a *person* call
- **<silence>** Please give me the *operator*

MIT의 Rose등이 개발한 시스템은 범용으로 사용될 수 있는 task를 설정하고 20개의 핵심어를 인식하는 시스템이다. 화자적응을 이용해서 학습을 수행하여 인식률을 향상시켰으며, HMM과 음소 모델링을 통해 핵심어와 비핵심어에 대한 분별 학습(selective training)을 수행했으며, 신경망과 HMM을 융합하는 인식기법을 개발했다. 화자독립으로 74%정도의 인식률을 나타낸다[33].

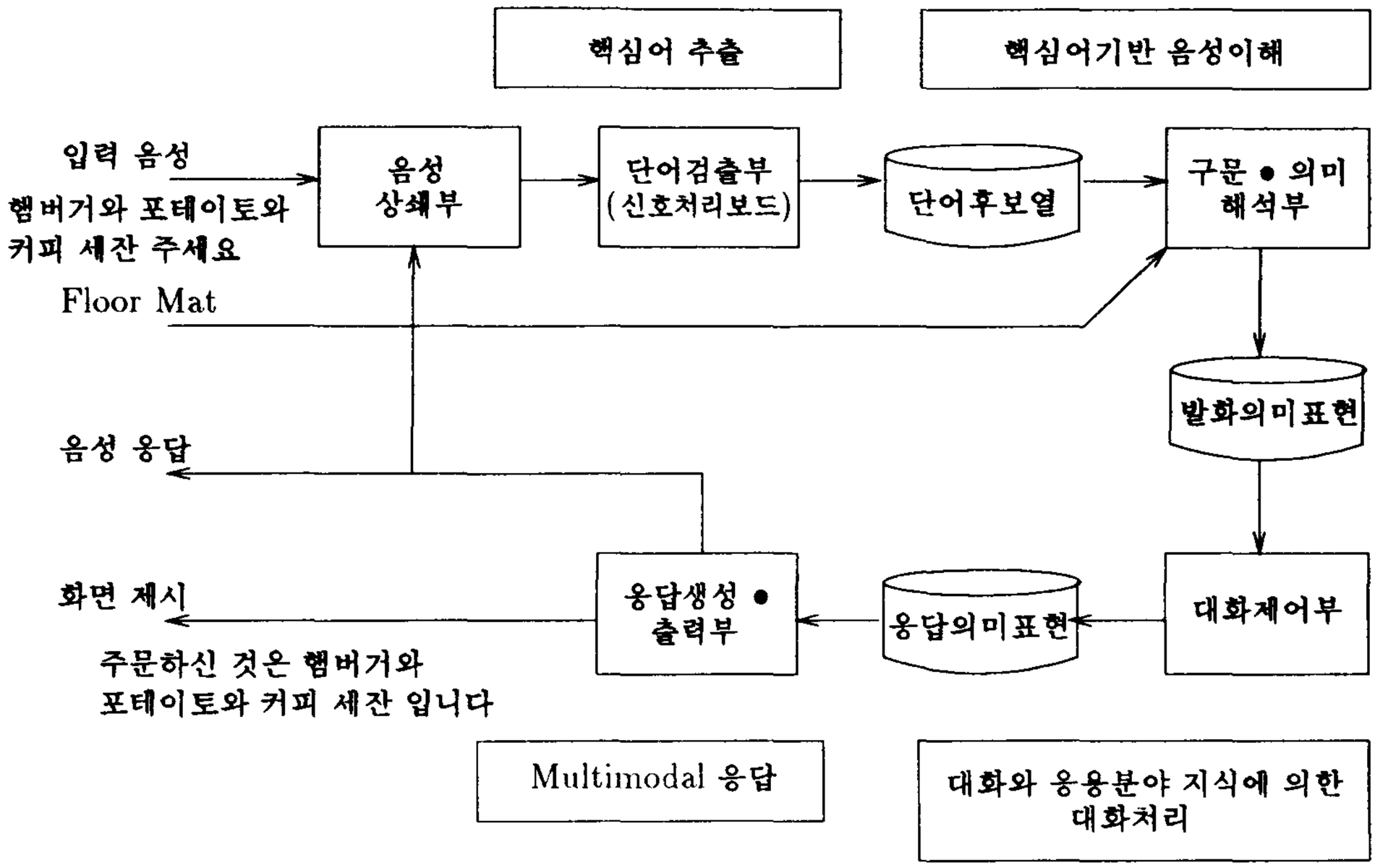


그림 2.3: 음성 자유 대화 시스템 TOSBURG II의 시스템 구성도

2.3.2 음성자유대화시스템 TOSBURG II

TOSBURG(Task-Oriented dialogue System Based on speech Understanding and Response Generation)는 핵심어 추출부, 자유발화이해부, 사용자 주도형 대화처리부, multimodal 응답생성부, 음성응답 상쇄(cancel)부 등으로 구성되어 있는데 실시간으로 동작한다[38][39][40].

TOSBURG II(그림 2.3)는 49개 핵심어에 대한 자유발화이해에 근거, 연속음성중의 핵심어를 추출하는 잡음면역학습에 의해 잡음이나 불필요한 단어에 대한 robustness를 높이기 위해 인식사전을 적용 특징 벡터의 시작부와 끝부분에서 연속적인 패턴의 다양한 조합을 생성한다.

구문의미해석부에는 parser에 의해 핵심어 검색을 수행하고, 핵심어 lattice를 시작부와 끝부분에서 자유롭게 해석한 후 복수의 발화의미표현 후보를 발생시켜 이

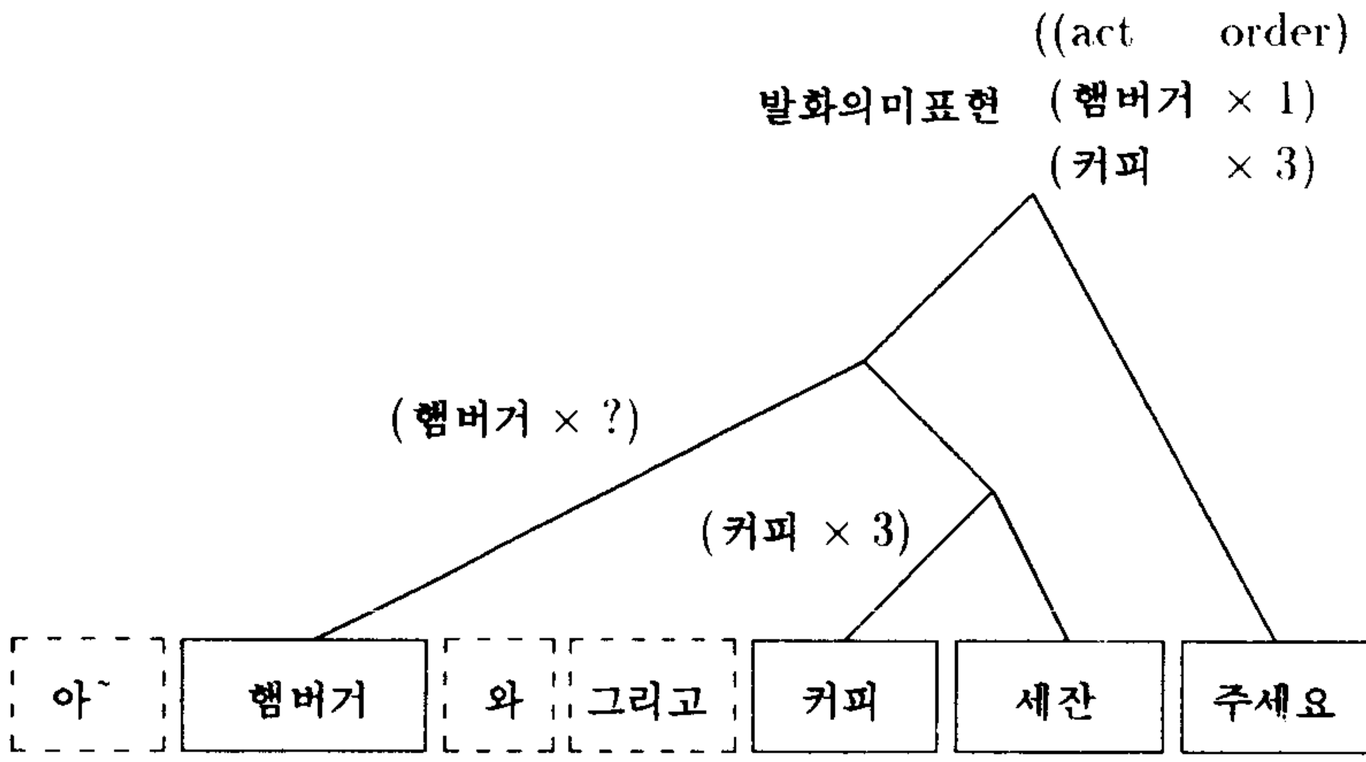


그림 2.4: 핵심어열로부터 얻어지는 해석 Tree와 의미표현

를 대화처리부에 넘긴다. 구문의미해석부는 문장의 시작과 끝의 판정, 문장 후보의 해석. 해석도중에 문장후보나 부분문장후보의 유지를 위한 확장된 LR문법을 사용한다. 의미정보는 프레임(frame)의 형식으로 표현되며 확장시 프레임이나 slot 등의 생성을 통해 의미해석 절차를 수행하며. 동시에 의미표현을 작성하고 있다. 그림 2.4는 자유발화음성의 해석예를 도시하고 있다. "햄버거", "주세요" 등의 핵심어로부터 "주문"등의 행위를 표현하는 행위(act), 주문품목, 갯수 등의 대화에 필요한 의미내용을 얻는다. 또한 대화도중에 생략된 표현이나 불필요한 단어에 대비해서 명확하지 않은 점이나 애매한 점은 대화처리에서 보충하게 된다.

사용자 주도형의 대화를 지향하여 사용자의 다양한 발화를 수용할 수 있도록 하고 있으며, 대화의 이력이나 상황을 고려하여 생략표현되는 발화의 이해를 도모하고 있다. 또한 상황에 적절한 응답을 생성하여 사용자가 부담없이 대화를 할 수 있도록 설계되었다.

그림 2.5에서는 전체적인 대화처리 흐름의 관점에서, 시스템을 두가지 상태 즉 사용자의 발화를 이해하는 사용자상태와 task를 관리하는 응답 생성시스템 상태로 나누어서 모델화하고 있는데 ATN으로 구현되어 있다. 이 모델에 따르면 사용자와 시스템의 상태변화를 대화의 정해진 수순과 독립적으로 진행 상황이 즉시 기술

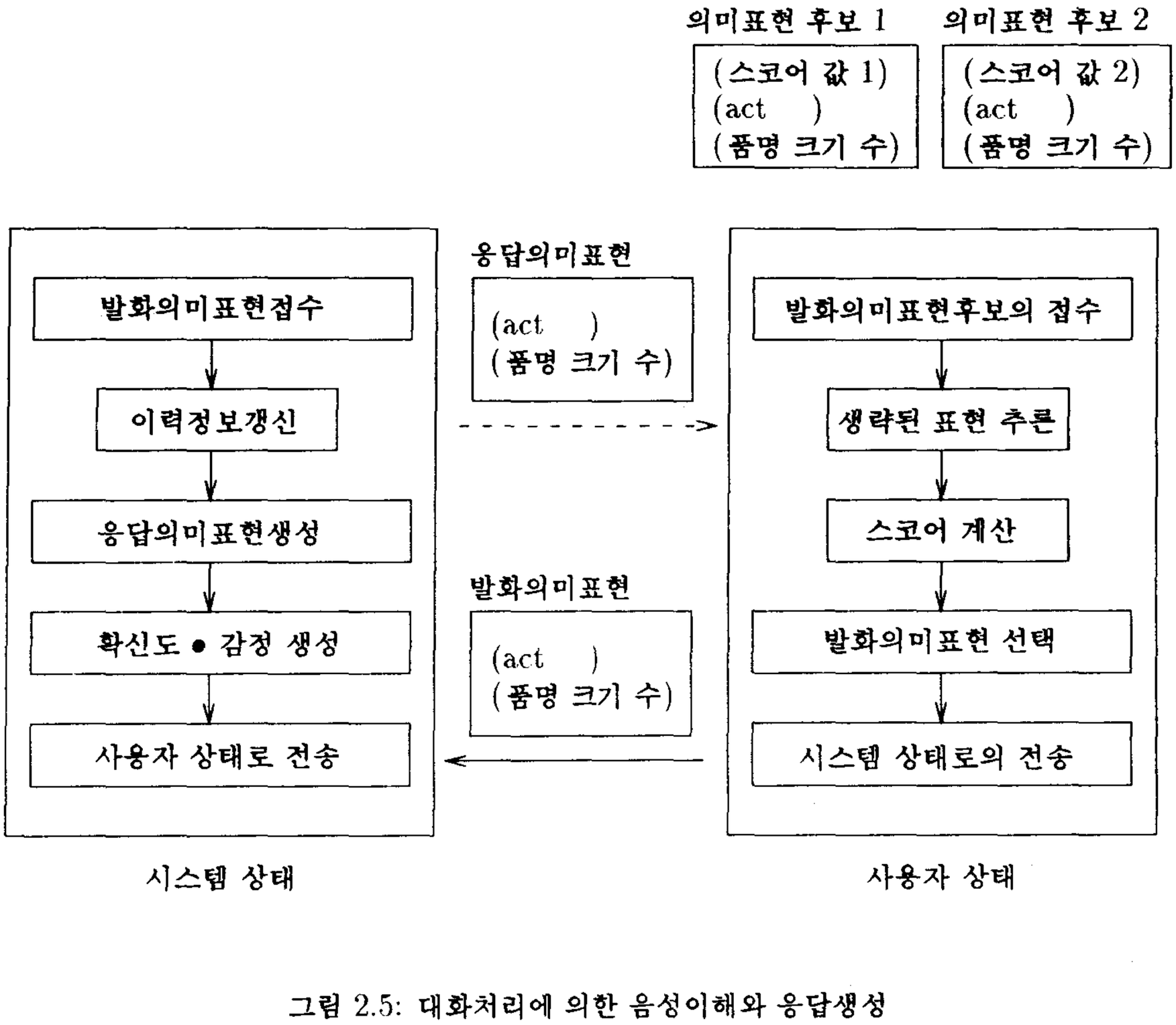


그림 2.5: 대화처리에 의한 음성이해와 응답생성

될 수 있으며, 사용자 주도의 다양한 대화 흐름에 대응할 수가 있게 된다.

사용자 상태에서는 음성이해부에서 생성된 복수의 의미표현후보에 대해 바로 전의 시스템 응답이나 대화의 이력정보와의 의미적인 조합성을 고려해서 생략된 품명, 크기, 수량 등을 유추 보충하게 된다. 직전의 시스템응답중 적절한 것이 없으면 시스템에서 정한 기준치를 설정한다. 직전 시스템응답과 입력의미표현후보의 내용이 잘 부합하지 않으면 그 후보의 스코어에 의존한 처리를 수행하며 가장 높은 스코어를 나타내는 표현이 선택된다.

한편 시스템상태에서는 음성이해의 결과를 반영하도록 대화의 이력정보를 갱신하고 응답 의미표현이 생성되는데 만일 시스템이 애매하다고 판단된 점이 있다면 이를 확인, 검증함으로써 대화를 계속 진행하게 된다. 이렇게 생성된 의미표현은 음

답생성 및 출력부에 출력되도록 하고 다음에 사용자가 발화할 내용의 이해에 이용하게 된다.

응답생성부로부터는 합성음성과 함께 화면을 통해 핵심어구, 점원의 자세, 주문품목과 갯수의 시각적 표현 등 multimodal 응답이 출력된다. 이때 점원의 표정을 대화상황에 맞도록 조절, 입술의 동작개시 시기를 합성음성에 맞추고 있다.

TOSBURG I은 음성응답도중에 음성입력을 받아들일 수 있도록 설계되어 있는데, 음성응답 상쇄 기능을 두어서 시스템이 상시 음성출력 도중 음성입력이 가능하도록 하고 있다.

3 장

전화선을 통한 음성 응용 시스템 프로토타입의 구현

3.1 시스템의 개요 및 구조

3.1.1 시스템의 개요 및 특성

본 시스템은 전화음성을 인식하는 음성인식기를 개발하고, 그 응용으로 사용자가 공중전화망을 통해 전화로써 관공서나 민간회사의 각종 서류 발급은 물론 안내업무까지 전담하도록 하는 새로운 인간-기계간 상호작용 수단의 구현에 관한 것으로, 사람의 음성을 인식하는 음성 인식기, 전화망에 대한 접속을 구현한 공중전화망접속기, 사람의 아날로그 음성신호를 컴퓨터내에서 처리가능하도록 디지탈 신호로 변환하고 이로부터 통계적 특성을 추출하는 음성특성 추출기, 개인용컴퓨터와 신호처리간을 연결하는 제어 및 통신 시스템 등으로 구성되어 있다.

특히 전화라는 매개체를 통해서 컴퓨터가 업무안내 및 서류발급 접수를 대행함으로써 관공서의 직원이 부재중인 주간이나 야간 모두 공히 24시간 대민 행정 안내 및 접수체제를 수립하여 사용자의 불필요한 대기시간을 없애고 발급업무를 자동화

함에 따른 업무수행상 능률을 기할 수 있을뿐만 아니라, 사용자가 컴퓨터와 음성을 통해 대화를 수행함으로써 사용자가 전화만 보유하고 있으면 컴퓨터 단말기를 보유한 것과 같은 기대효과를 지닐 수 있다.

본 시스템의 첫번째 구성 요소는 전화망을 통해 들어온 사용자의 음성신호를 분석하여 음성특성을 도출하는 것으로, 음성신호가 지닌 다양한 회선상 및 화자에 따라 달리 나타나는 변이요인을 고려 선형예측계수(Linear Predictive Coefficient, LPC) 분석법을 거쳐 켑스트럼, 차분 켑스트럼, 파워 및 차분 파워 등의 특성을 추출한 다음 인식기의 참조패턴으로 사용되는 세개로 구성되는 다중코드북을 구성하는 신호처리 및 음성 특성 추출부이다.

두번째 구성 요소는 음성특성추출부에서 생성된 음성특성을 받아들여 은닉 마르코프 모델을 사용해 음성인식을 수행하는 음성인식부분이다. HMM의 학습은 위에서 생성된 다중코드북을 통해 추출한 코드북의 색인을 포워드-백워드 절차법(forward-backward procedure) 및 바움-웰치(Baum-Welch) 법을 이용해서 각 단어마다의 은닉 마르코프 모델에 사용되는 확률 통계적 모수를 추출하도록 하는 과정으로 요약되며, 추출된 모수를 기반으로 비터비(Viterbi) 탐색을 거쳐 인식 단어를 결정하게 된다.

음성의 특성 추출 및 인식과정을 지원하는 환경에는 PC와 DSP를 상호 보완적으로 활용하고 있는데, 개인용 컴퓨터(PC)에는 음성인식을 위한 인식기와 화일 입출력을 전담하도록 하는 PC 관리자 시스템을 설치하고, 신호처리보드에서는 음성의 녹음, 재생 및 사용자와의 대화를 관리하도록 하는 신호처리 관리자 시스템을 설치하여 개인용 컴퓨터와 신호처리기간에 적절한 계산기능을 분담시켜 수행시간의 단축과 계산효율을 높이고 있다.

본 시스템은 PC에 네 개까지의 신호처리보드를 설비하고 다중처리를 수행하도록 하고 있는데 각각의 신호처리기에 하나씩의 전화회선을 접속시켜 개인용 컴퓨터가 이들을 관리하도록 하는 동시에 신호처리기들간의 적절한 제어와 통신을 전담하여 민원안내 및 접수를 사용자와의 대화형식으로 수행하도록 하는 것이며, 동시

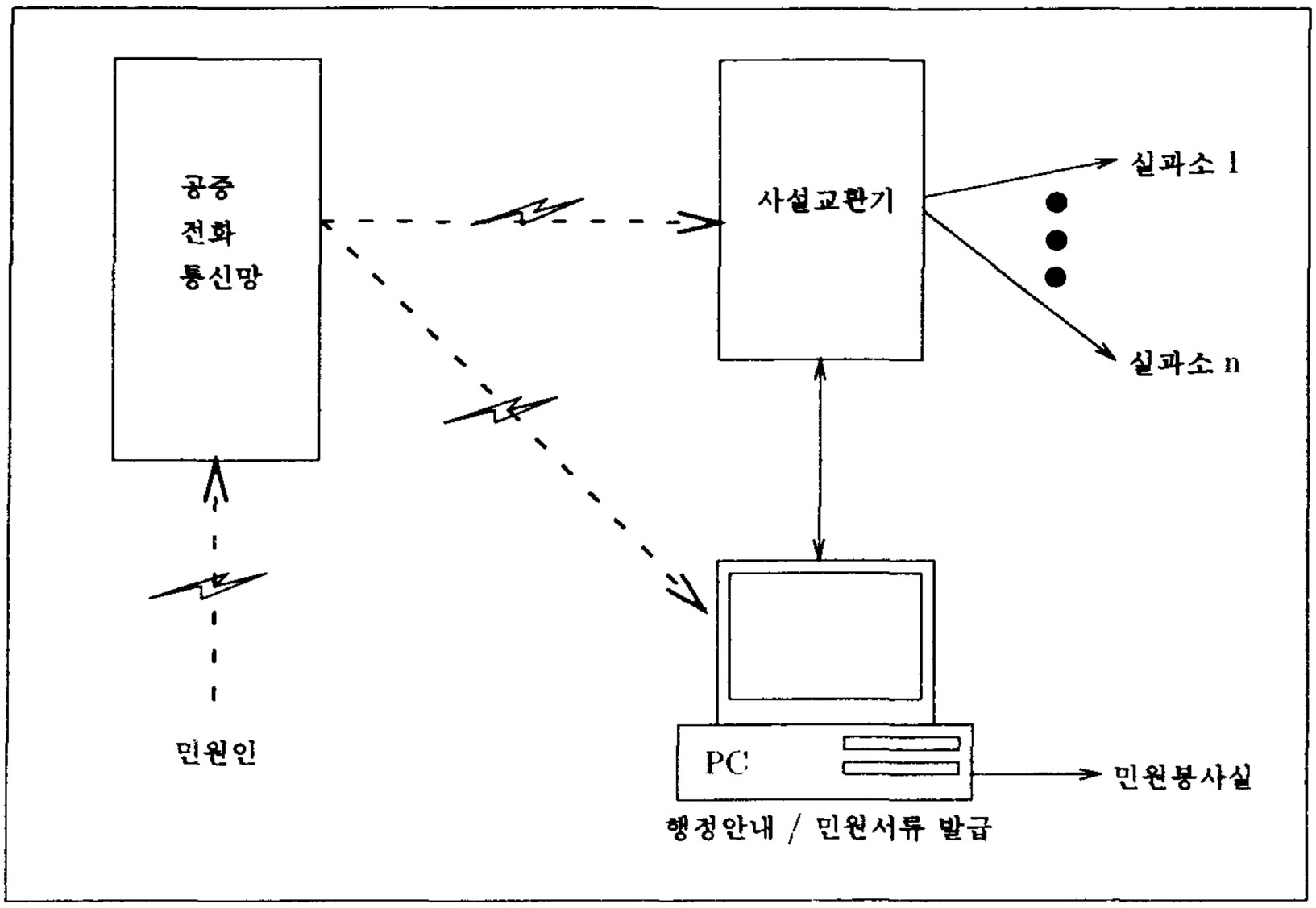


그림 3.1: 민원처리 및 안내를 위한 음성인식시스템의 개요도

에 다수의 민원 서비스를 가능하게 하고 있다.

본 시스템의 마지막 구성 요소는 일반적인 전자식 전화기에서 사용되는 DTMF (Dual-Tone Multi-Frequency) 체계를 소프트웨어적으로 구현해서 사용자가 전화기 번호판을 누를 경우 이를 감지해서 해당되는 번호를 인지하는 ARS(Audio Response System) 시스템의 기능과 음성인식기의 기능을 통합하는 것이다.

3.1.2 시스템의 구조

그림 3.1에서와 같이 사용자는 공중전화기, 가정용 전화기 등 어떠한 유형의 일반적인 전화기를 사용해서 공중전화망을 경유해서, 민간회사나 관공서의 구내 사설교환기 또는 직접회선을 통해 시민봉사실 또는 민원실에 설치된 컴퓨터와 음성을 사용 대화식의 통화를 수행함으로써 발급받고자 하는 서류를 접수할 수가 있다.

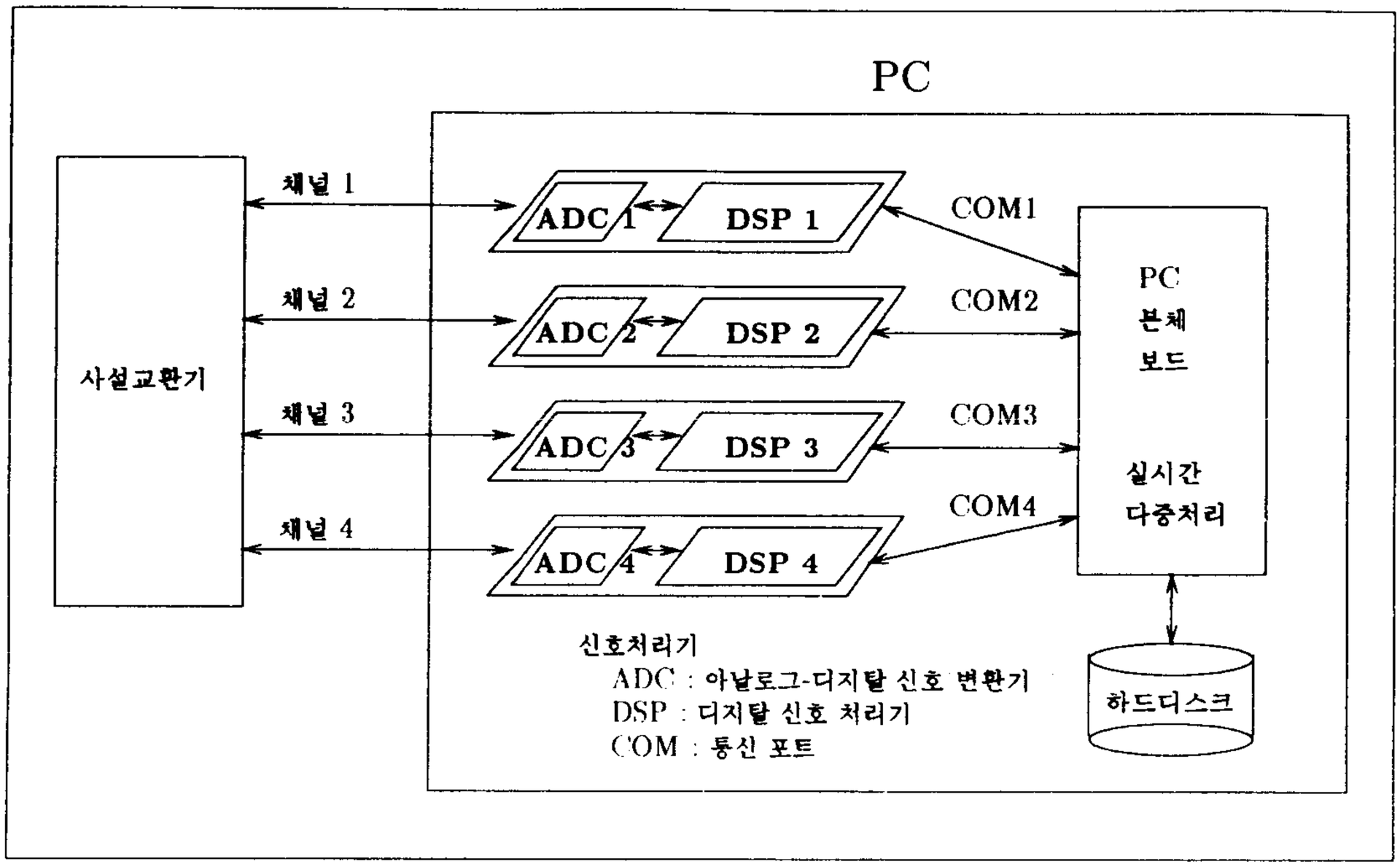


그림 3.2: 하드웨어 구성도

하드웨어

그림 3.2에서는 본 시스템의 하드웨어 구조에 대한 개념을 도해하고 있다. 음성신호처리와 음성특징 추출을 담당하는 디지탈신호처리(Digital Signal Processing, DSP) 보드는 PC의 표준 인터페이스 버스에 모두 4개까지 장착할 수 있는데 이러한 네개의 보드와 PC 시스템과의 통신수수는 PC의 표준통신 인터페이스인 RS-232 카드의 기본 주소방식(base address)을 사용 주변기기 표준인터페이스를 적용받으며 각각의 신호처리보드는 COM1, COM2, COM3, COM4의 통신포트를 할당받는다. 각 신호처리보드에는 5.5kHz-48kHz의 크리스탈 제어 표본주파수 특성을 나타내는 아날로그-디지탈 및 디지탈-아날로그 변환기가 탑재되어 있다.

이때 각각의 신호처리보드는 PC 와의 데이타 및 제어 신호 또는 코드의 송수신을 위해 PC 의 중앙처리장치에 인터럽트를 발생시키게 되는데 이를 위해 PC 에 설정되어 있는 인터럽트 요청 라인(Interrupt ReQuest line)중 IRQ10, IRQ11,

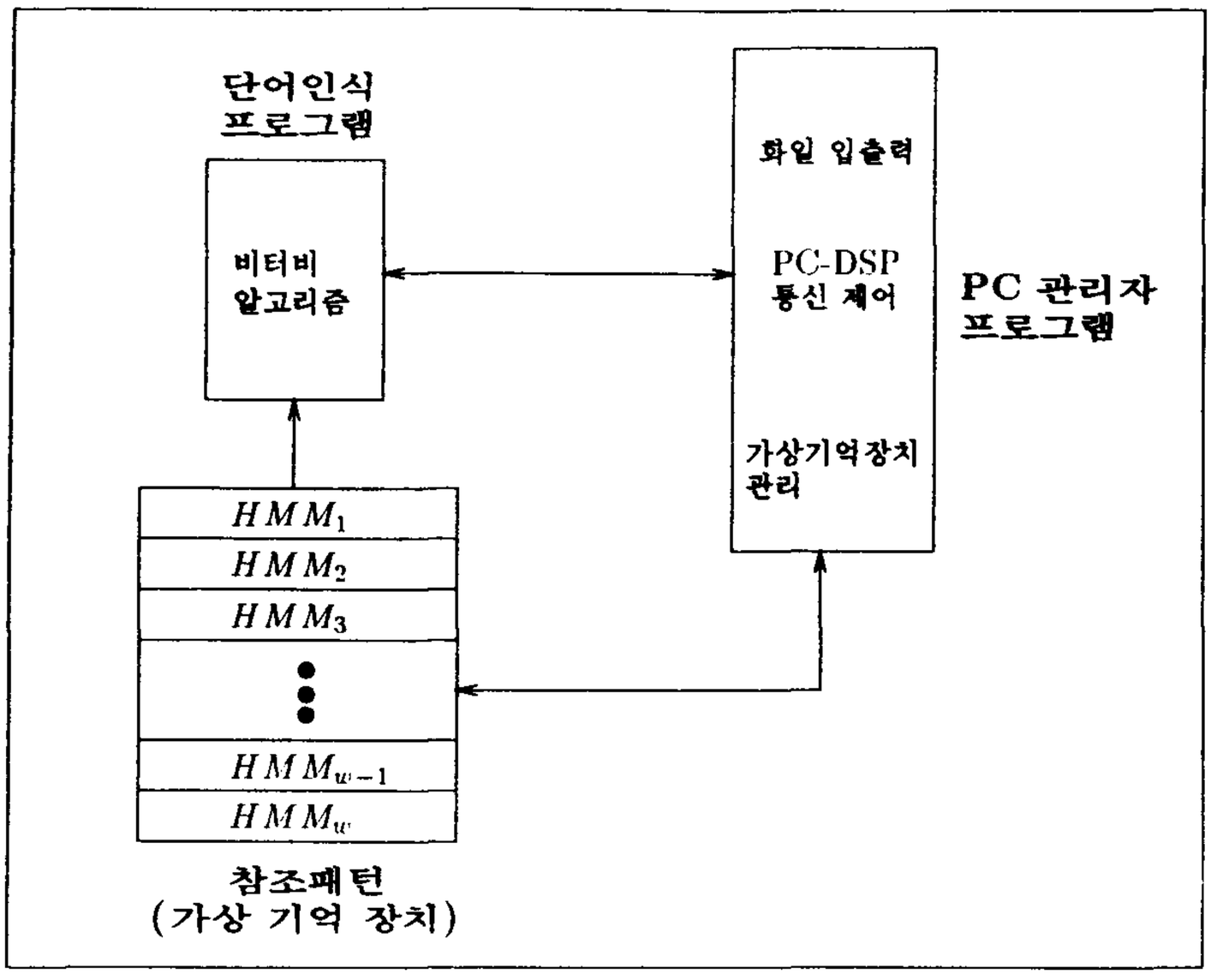


그림 3.3: PC 관리자 소프트웨어 구성도

IRQ12, IRQ15 를 각각의 신호처리보드에 할당하고 있다.

PC내에 내장되어 있는 하드디스크에는 음성인식을 위한 코드북, 인식대상이 되는 각 단어마다의 참조패턴이 이진 패턴(binary pattern)으로, 안내문이나 인사말 등과 같은 방송용 음성 데이타가 8kHz의 표본화율(sampling rate)로 디지탈 코드화되어서 저장되어 있으며, 또한 사용자와의 대화시 사용자의 음성을 역시 동일한 표본화율로써 코드화한 후 저장시키도록 되어 있다.

소프트웨어

그림 3.3과 그림 3.4에서와 같이 소프트웨어는 두 부분으로 나눌 수 있는데 PC쪽과 DSP 쪽에는 각각의 시스템 관리를 위한 관리자 프로그램이 하나씩 운영되고 있다.

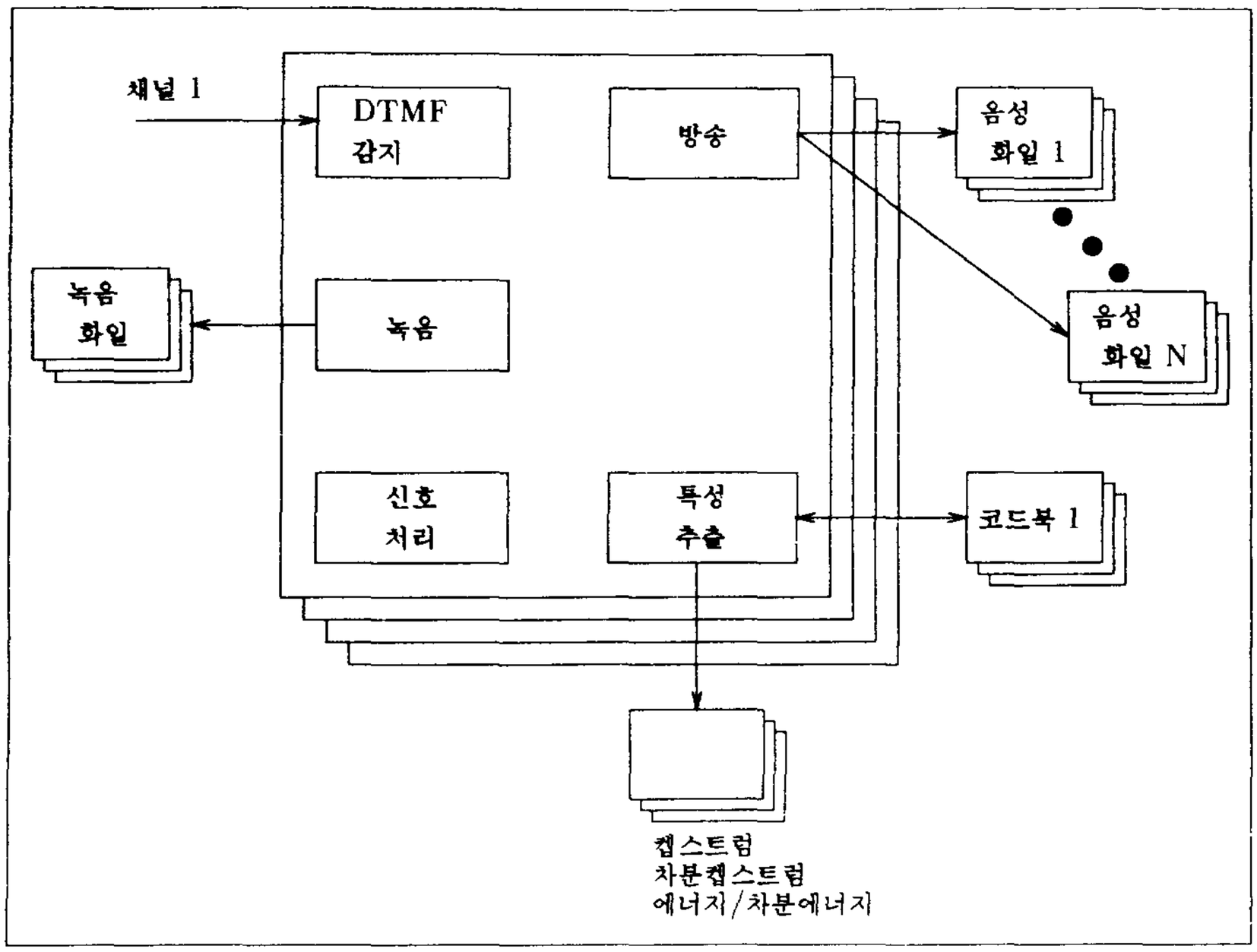


그림 3.4: DSP 소프트웨어 구성도

PC의 관리자 프로그램은 네 개의 신호처리보드 각각과의 통신을 통해 메시지와 제어코드를 주고 받음으로써 음성의 특징추출 및 인식을 수행하는 전 과정을 제어하도록 되어 있다. 관리자 프로그램은 또한 네 명의 사용자를 동시에 서비스할 수 있도록 하는 병렬 실시간 다중처리를 지원하도록 되어 있다.

PC 관리자 프로그램은 신호처리 보드의 실시간 다중처리를 지원하는 외에, DSP 관리자가 요구하는 화일의 입출력, 음성인식을 수행하는 은닉 마르코프 모델의 호출을 담당한다. 또한 최대 640K 바이트밖에 안되는 MS-DOS의 주기억장치 운용의 한계를 극복할 수 있도록 하는 가상기억장치(virtual memory)를 운영하고 있는데 그림 3.3의 가상기억장치가 그것이며, HMM의 참조패턴을 인식대상이 되는 단어의 수만큼 분할한 후에 각각의 참조패턴을 저장하게 되어 있다.

신호처리보드상에서는 전화벨을 감지하는 통화수립부터 특성을 추출하는 신호

처리의 전 과정을 관리하는 신호처리 관리자 프로그램이 수행되어진다. 신호처리 관리자 프로그램의 주요한 기능은 다섯가지로 요약할 수 있는데 DTMF 즉 전화벨을 인식하고 사용자와의 통화를 수립하며 사용자가 원하는 전화번호를 다이얼링할 수 있는 DTMF 감지 및 발생부, 사용자의 음성을 녹음하는 음성 녹음부, 사용자의 음성 또는 PC의 보조기억장치에 저장되어 있는 음성을 재생하는 음성 재생부, 녹음된 사용자의 음성으로부터 음성인식에 필요한 특성을 추출하는 음성신호처리부, 추출된 특성을 신호처리보드상의 주기억장치에 저장되어 있는 세개의 코드북과의 정합(matching)을 통해 은닉마르코프모델로 유입되는 입력 관찰 벡터(observation sequence)를 산출하는 코드북 정합부가 그것이다.

3.2 신호 처리 및 특징 추출

사용자의 음성신호는 회선상 잡음, 사용자의 불필요한 감탄사나 호흡음, 성별, 나이별, 억양별, 지역별에 따라 달리 나타나는 다양한 변이요인을 지니고 있다. 따라서 이러한 변이요소를 적절히 여과 또는 상쇄시킬 수 있는 분석법을 통해 양질의 음성신호와 신호 특성을 추출하는 것이 신호처리의 관건이다.

그림 3.5는 신호처리의 과정을 나타내고 있는데 사용자의 녹음된 음성은 대역통과필터(Band-Pass Filter, BPF) 및 증폭(preemphasis) 과정을 거쳐 잡음이나 무음성구간으로부터 음성을 단어별로 추출(word segmentation) 하게 되는데 이때 표본화된 음성신호를 30밀리초마다 묶어서 이를 하나의 프레임으로 본다. 각 프레임에 대해 음성이 존재하는지의 여부를 판단하기 위해서는 프레임별로 에너지의 크기를 계산해서 일정한 임계치(threshold)를 넘어서면 음성 프레임으로, 그보다 작으면 잡음 또는 비음성 프레임으로 본다.

- 증폭식 (y_i는 증폭된 신호, x_i는 원래의 신호, N은 표본의 수 (number of samples))

$$y_0 \leftarrow x_0$$

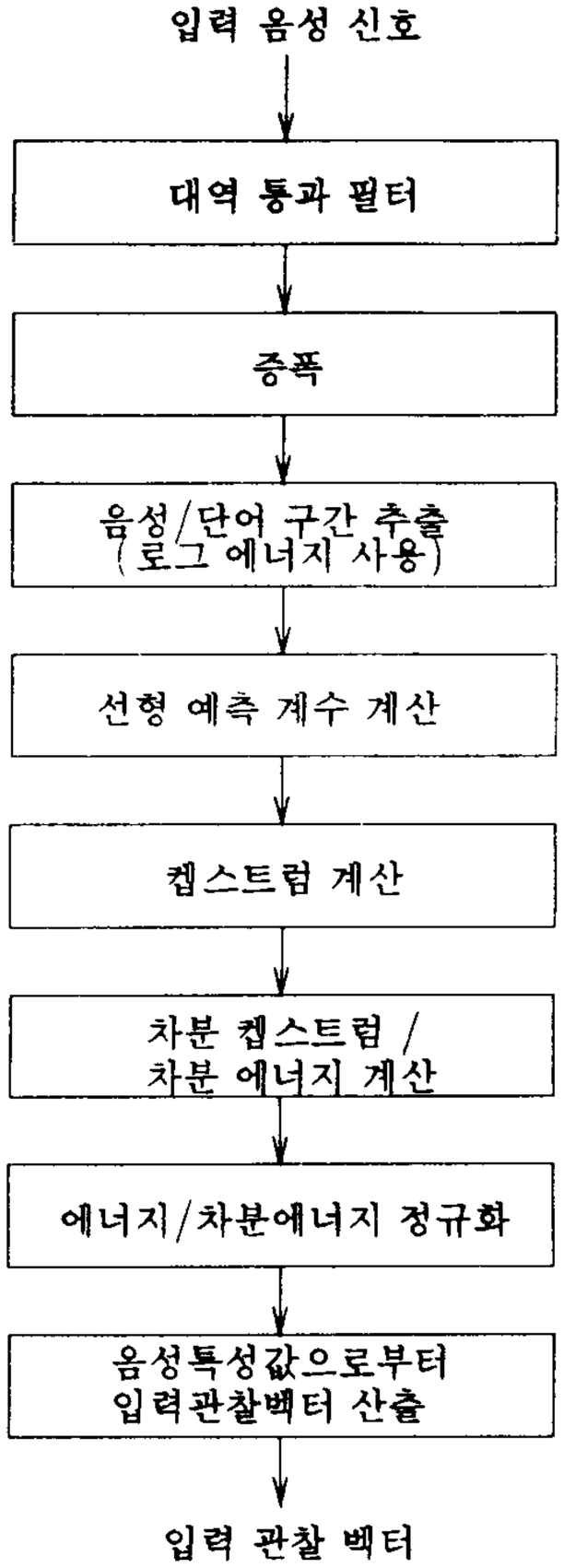


그림 3.5: 신호처리 및 특징추출 흐름도

$$y_i \leftarrow x_i - 0.95x_{i-1}, \quad 1 < i < N$$

- 대역필터식 (a_k와 b_k는 대역필터의 특성을 나타내는 모수(parameter)값, y_i는 대역필터를 통과한 값, x_i는 원래의 신호)

$$y_i \leftarrow -\sum_{k=1}^{8} b_k y_{i-k} + \sum_{1}^{9} a_k x_{n-k}$$

$$a_k \leftarrow \{-1.4363, -1.2436, 1.6427, 1.3785, -1.0043, -0.7028, 0.2155, 0.1505\}$$

$$b_k \leftarrow \{0.3878, 0.0, -1.5513, 0.0, 2.327, 0.0, -1.5513, 0.0, 0.3878\}$$

- 각 프레임은 신호 대역 특성의 표현을 위해 해밍윈도우 w 를 적용하게 되는데 다음식에 따라 얻어진 윈도우벡터값을 각 프레임에 곱하게 된다.(N : 프레임의 크기, 240)

$$w_n \leftarrow 0.54 - 0.46 \cos \frac{2\pi n}{N-1}, \quad n = 1, \ldots, 240$$

음성신호는 8kHz의 표본화율을 적용하고 있으며 이를 30밀리초마다 하나의 프레임으로 묶는다고 하면 한 프레임은 $30 \times 8 = 240$개로 구성된다. 그림 3.6에서와 같이 매 프레임은 15밀리초마다 중첩되어서 구성되는데 1번 프레임부터 시작되어서 F번째 프레임까지로 이루어지며 전체 음성 신호의 표본수에 따라 프레임의 수는 가변적이다.

이러한 프레임 구조로부터 음성 구간을 추출하기 위해 로그 에너지를 사용하는데, 프레임을 구성하는 각각의 신호를 $x(n)$이라 하면 f번째 프레임의 에너지는 다음의 수식으로 구한다.

$$E_f \leftarrow 10 \log_{10} \sum_{n=1}^{240} x_n^2 w_n$$

음성특성추출

음성의 구간이 검출되면 음성구간에 대해 10차원의 선형 예측 계수 a_k를 계산하고, 선형 예측 계수로부터 12차원 켑스트럼 벡터 c_i를 산출하게 된다.

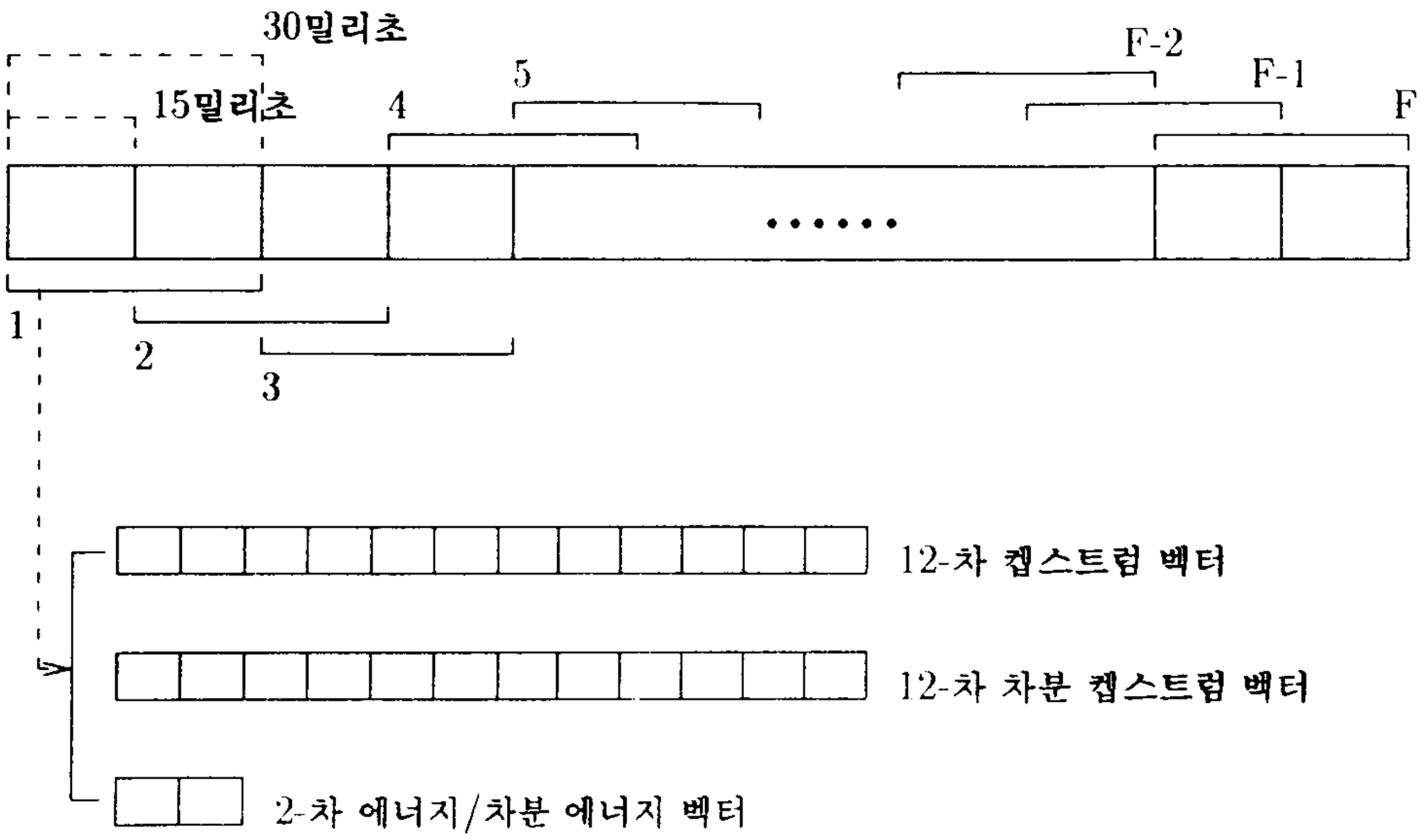


그림 3.6: 음성의 단어 구간 검출을 위한 프레임의 구성

- 레빈슨-더빈(Levinson-Durbin)법에 의한 선형 예측 계수 a_k 의 계산 (D 10)

$$R_i \leftarrow \sum_{m=i}^{240} x_m x_{m-i}. \quad i = 0, \ldots, D$$

모든 $m = 1, \ldots, D$에 대해 ($e_0 = 0. R_0 = 0$)

$$A_{mm} \leftarrow \frac{R_m - \sum_{k=1}^{m} A_{m-1,k} R_{m-k}}{e_{m-1}}$$

$$A_{mk} \leftarrow \sum_{k=1}^{m} (A_{m-1,k} - A_{mm} A_{m-1,m-k})$$

$$e_m \leftarrow (1 - A_{mm}^2) e_{m-1}$$

$$a_{k-1} \leftarrow A_{Dk}$$

- 켑스트럼 c_i 의 계산 ($D : 12, \quad i = 2, \ldots, D$)

$$c_0 \leftarrow 0$$

$$c_1 \leftarrow -a_1$$

$$c_i \leftarrow -a_i - \sum_{k=1}^{i-1} \frac{i-k}{i} c_{i-k} a_k$$

이렇게 계산된 켑스트럼에 대해 12차원의 차분(delta) 켑스트럼을 구하게 되는데 차분 켑스트럼은 현재 프레임을 기준으로 전후 프레임의 켑스트럼 벡터 값의 차를 벡터값으로 취한다. 즉 m번째 프레임의 켑스트럼 벡터를 C_m, 차분 켑스트럼을 DC_m 이라 하면

$$DC_m \leftarrow C_{m+d} - C_{m-d}$$

로 나타낼 수 있으며 d는 해당되는 프레임의 수이다(m : 2).

또한 전후 프레임의 에너지의 차이값인 차분 에너지 DE_m도 음성의 특성값으로 사용하고 있는데,

$$DE_m \leftarrow E_{m+d} - E_{m-d}$$

를 이용해서 구하며, 에너지와 차분 에너지는 하나의 벡터로 구성한다.

코드북의 구조 및 구성

LBG 법은 벡터양자화(vector quantization)의 일반화된 기법으로 현재 음성 코드북을 만드는데 있어 가장 널리 사용되고 있는데 *k-means* 법으로도 불린다.

코드북은 일단 음성의 특성 추출이 종료된 다음 참조패턴으로써 사용되는데, LBG 법에 의한 코드북은 학습용으로 채집된 음성들로부터 위에서 기술한 음성특성 추출과정을 거쳐 가장 대표적인 특성 벡터만을 추출해서 모아놓은 것이 된다. 이러한 과정이 끝나면 입력된 음성 특성과 코드북에 저장된 음성 특성과를 비교하여 그 차이값이 가장 적은 코드북의 색인값을 인식 알고리즘에서 사용하게 된다.

코드북은 그림 3.7과 같이 크기 64인 2차원 행렬로 저장된다.

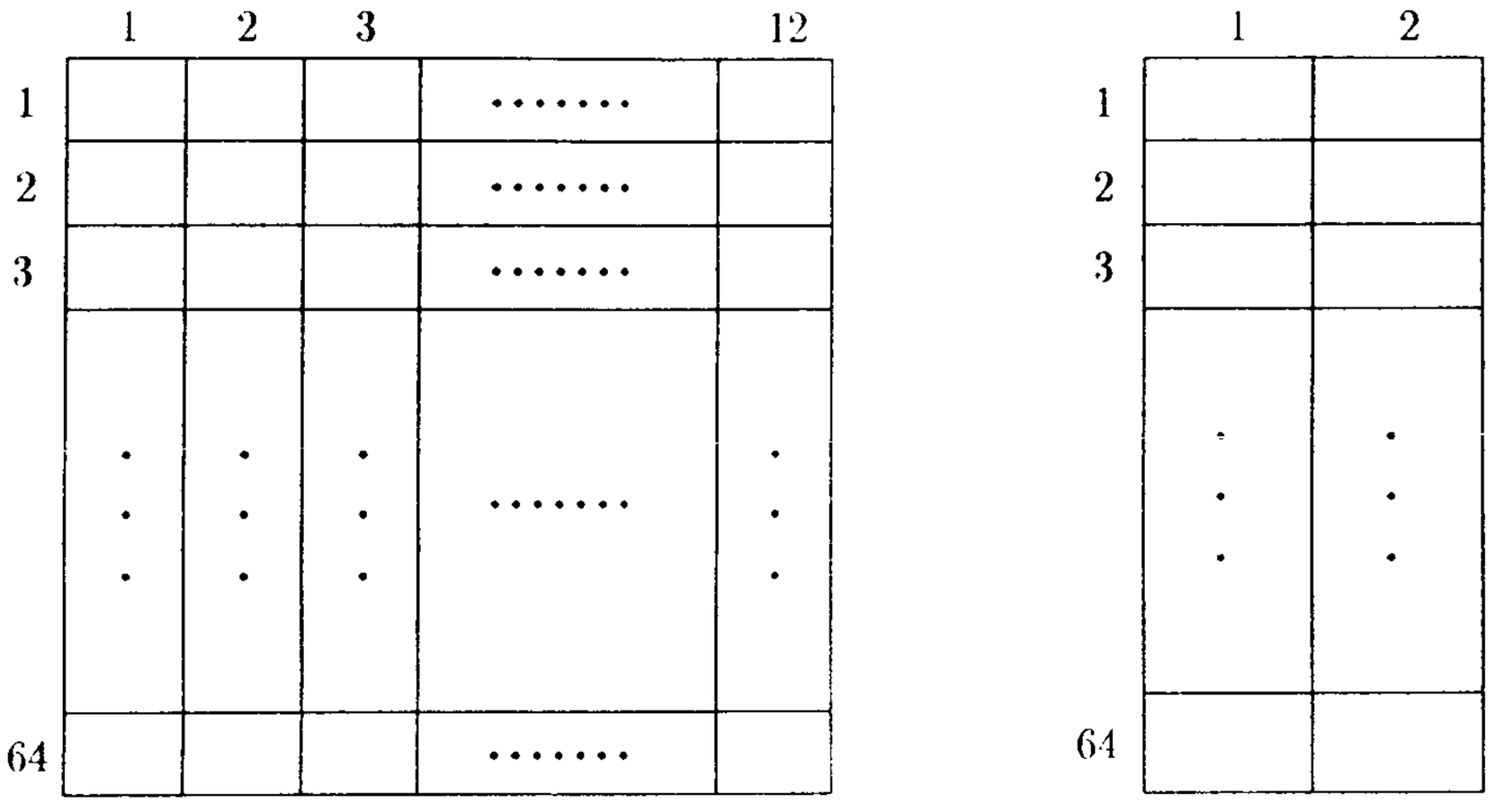


그림 3.7: 음성인식을 위한 코드북의 구성

1. 입력벡터가 저장된 데이타 화일을 읽어서 이차원 행렬 X_{mn} (m은 전체 입력 벡터의 수, n은 벡터의 차원: 켑스트럼은 12, 파워는 2)에 저장한다.

2. 적당한 방법을 사용하여 초기 코드 벡터 A_0를 정한다.

3. $NumberOfIteration \leftarrow \log_2(SizeOfCodebook)$

4. 모든 $r(r = 1, \ldots, NumberOfIteration)$에 대해

   4-1: $Distortion_{old} \leftarrow \infty$

   4-2: 모든 $i(i = 2^r - 1, \ldots, 1)$에 대해 (0.05는 perturbation factor)

   입력 벡터의 집합 A_{ij}를 가까운 것(cluster)들끼리 모으는 법칙에 의해 클러스터 C_i로 분류한다.

   $A_{2i,j} \leftarrow A_{ij} + 0.05$

   $A_{2i+1,j} \leftarrow A_{ij} - 0.05$

4-3: 모든 클러스터의 코드 벡터들을 다음식에 따라 각 클러스터에 속해 있는 입력 벡터들의 도심(centroid)으로 정한다.

$$A_i \leftarrow Centroid(C_i), \quad i = 1, \ldots, 2^r$$

4-4: 현재의 코드북에 대해 $Distortion_{new}$과 $Criterion$을 계산해서 $Criterion$이 정해진 임계치(0.005)보다 작은 값을 가질때 클러스터링을 종료하며 그렇지 않을 경우 4-2:부터 반복수행한다. ($Dist(A, B)$: 벡터 A와 B의 유클리드 거리 (Euclidean distance))

$$Distortion_{new} \leftarrow \frac{\sum_{i=1}^{m} \left[min_{j=1}^{2^r} \{Dist(X_i, A_j)\} \right]}{m}$$

$$Criterion \leftarrow \left| \frac{Distortion_{new} - Distortion_{old}}{Distortion_{new}} \right|$$

다음은 i번째 클러스터인 C_i의 도심 A_i를 구하는 절차인 $Centroid(C_i)$의 세부 알고리즘을 기술하고 있다.

- 초기 코드 벡터 A_0의 산출

$$A_{0k} \leftarrow \frac{\sum_{i=1}^{m} \sum_{j=1}^{12} X_{ij}}{m}, \quad k = 1, \ldots, 12$$

- 코드 벡터 A_i의 산출

1. $count$ 를 0으로 초기화
2. 모든 입력 벡터 $v = 1, \ldots, m$에 대해

2-1: $min_{1 \le j \le 2^r} \{Dist(X_v, A_j)\}$ 인 j를 구한다

2-2: 만약 최소거리 벡터의 색인 j가 i와 같으면

$$A_{ik} \leftarrow A_{ik} + \sum_{n=1}^{12} X_{in}$$

$$count \leftarrow count + 1$$

3. $A_{ik} \leftarrow \frac{A_{ik}}{count}$

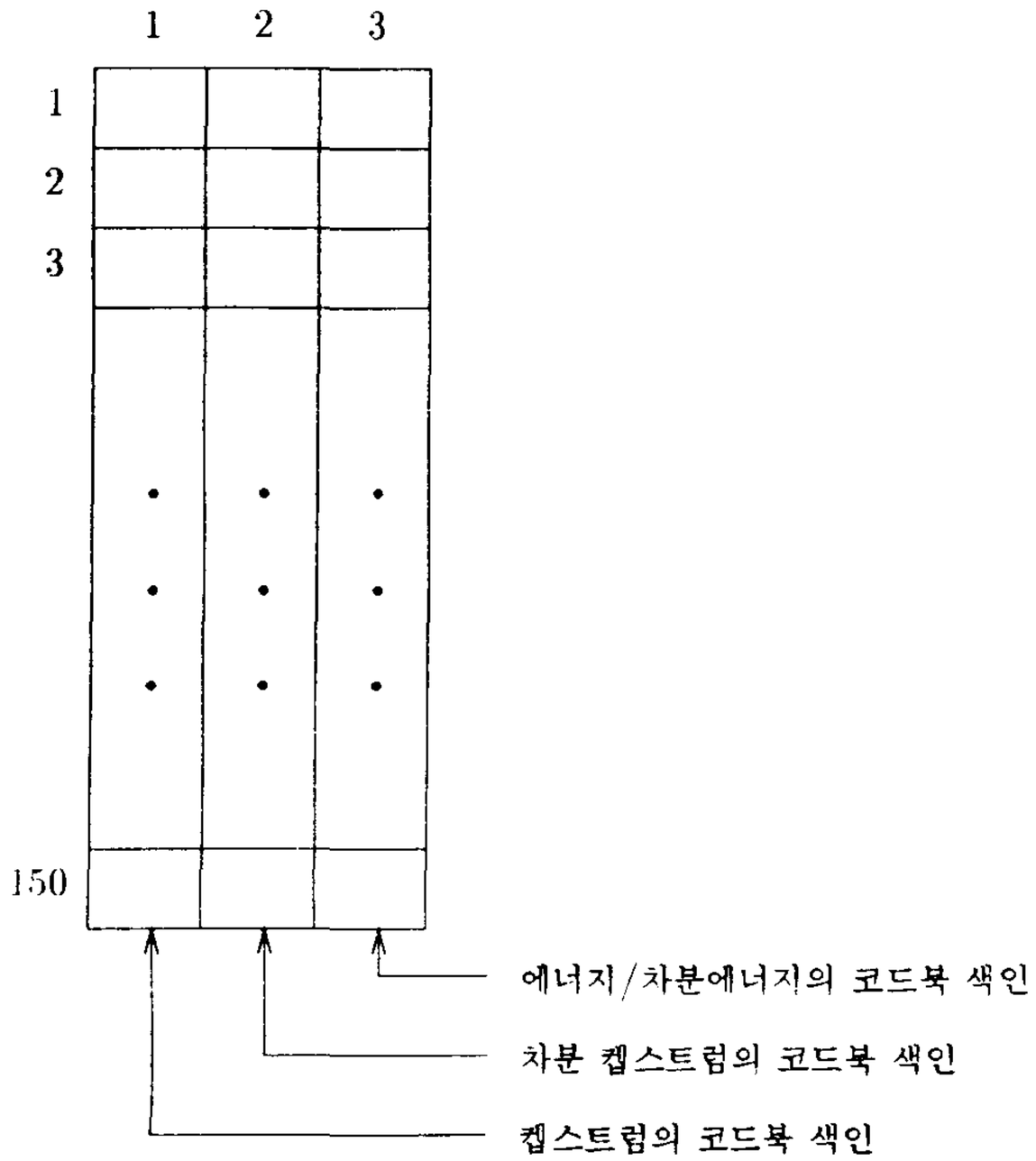


그림 3.8: HMM의 입력으로 사용되는 observation sequence 구성

그림 3.8는 HMM의 입력으로 사용되는 코드북 인덱스의 집합인 입력관찰벡터의 구성을 나타내고 있다. 가로축의 세개의 열은 각각 켑스트럼. 차분 켑스트럼 및 에너지의 코드북 색인값을 나타내며, 하나의 입력 프레임으로부터 세개의 색인값으로 구성되는 코드북 색인값 집합이 계산된다. 따라서 행의 수는 녹음된 데이타 중 음성구간이 나타나는 길이에 비례해서 나타난다. 이러한 행의 수는 사용자의 발성 속도에 따라 또 단어의 길이에 따라 다양하게 나타나기 때문에 본 시스템에서는 임의의 큰 값인 150을 채택하고 있으며, 이러한 이유로 행렬의 크기는 최대 (150행 × 3열)이 된다..

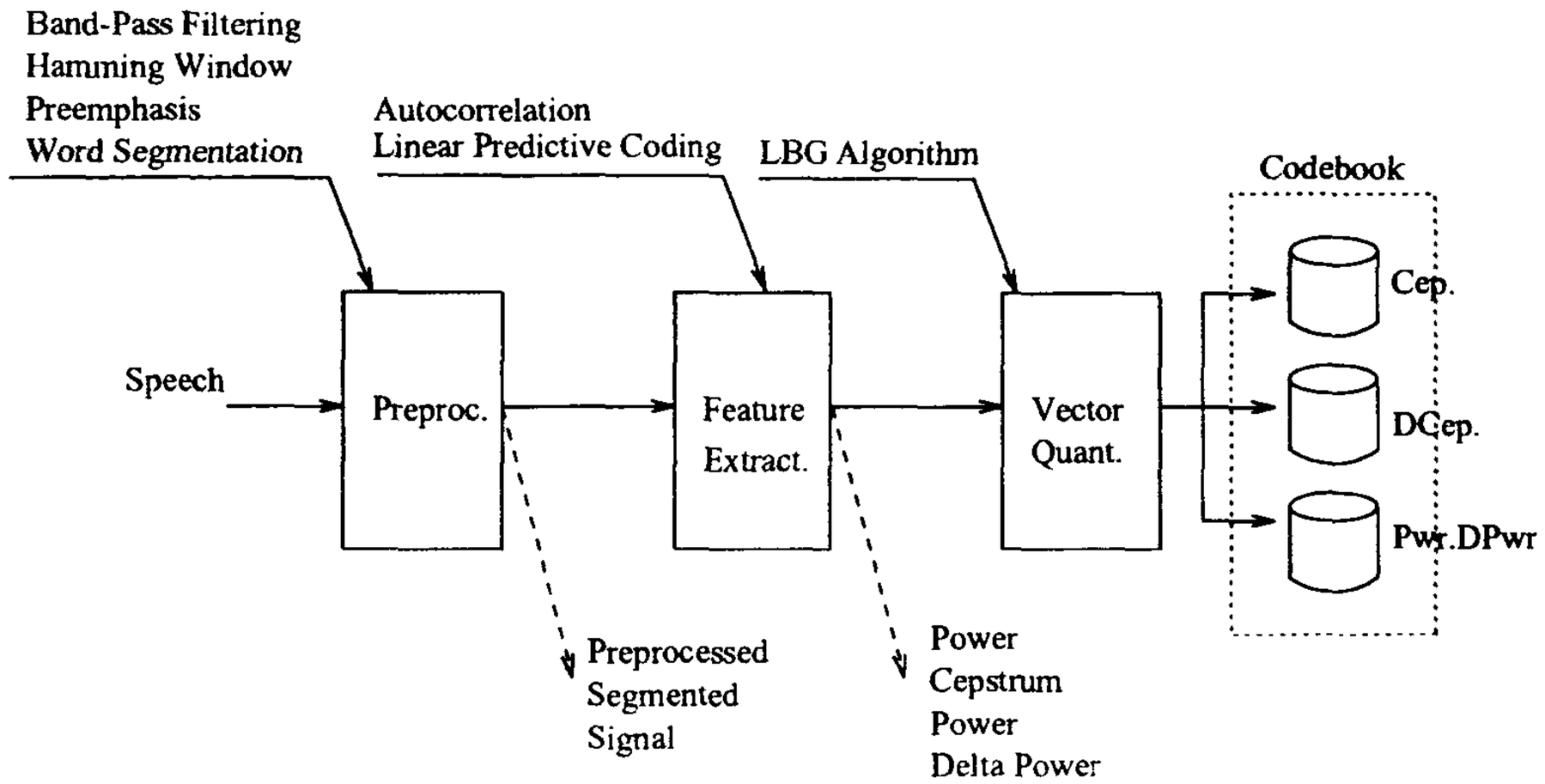


그림 3.9: 코드북 생성 과정

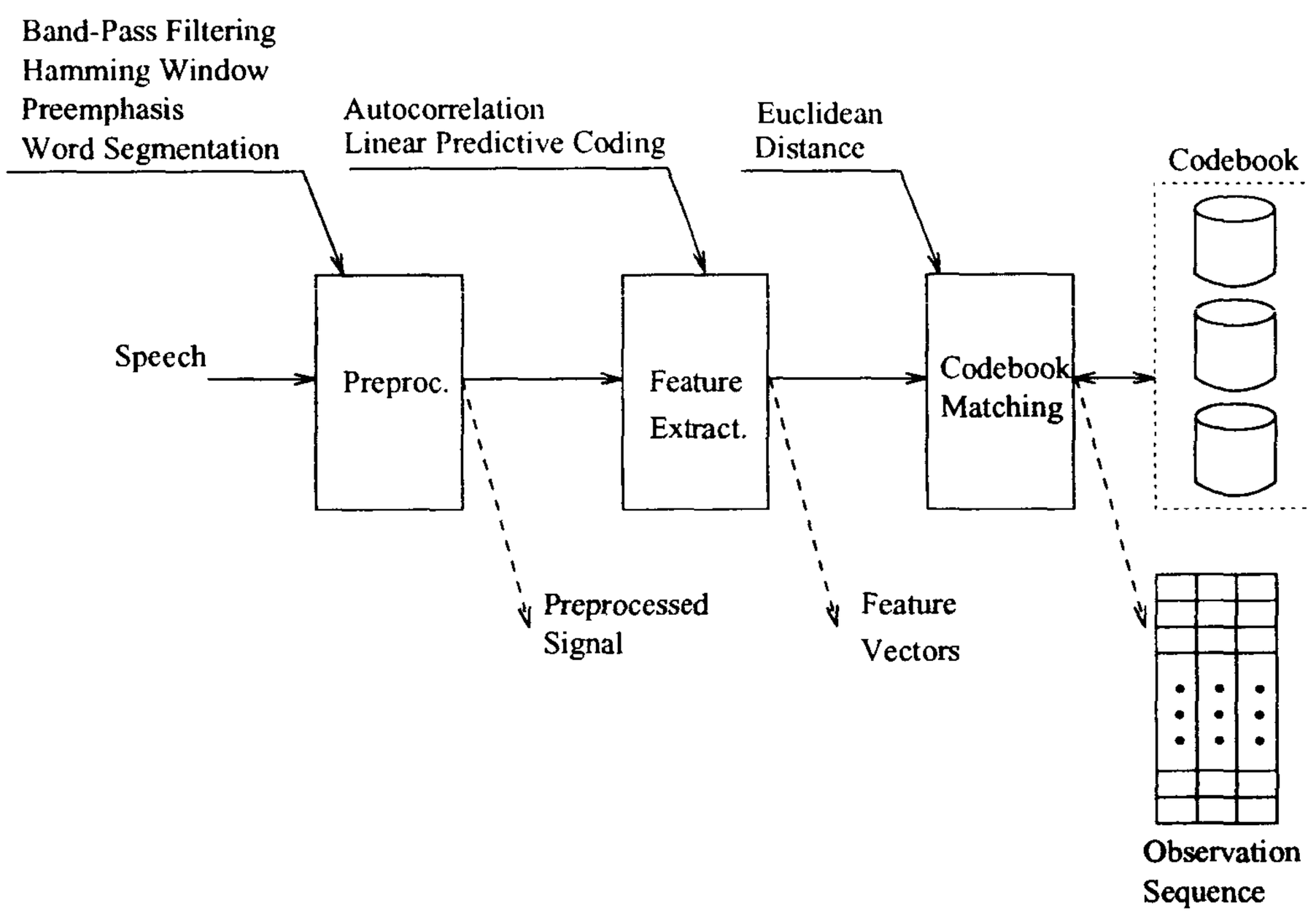


그림 3.10: observation sequence의 생성 과정

3.3 이산 HMM을 이용한 패턴인식

은닉 마르코프 모델의 중요한 특징은 마르코프 프로세스와 그 확률적 상태들에 의해 음성 신호의 통계적 특성을 명시적으로 모델링 한다는 데 있다. 이러한 통계적 처리절차는 유한상태 오토마타(Finite State Automata, FSA)에 의해 구현되어진다.

그림 3.11에서는 본 시스템에서 사용하고 있는 유한상태 오토마타의 예를 나타내고 있는데, 그림에서 원은 대략 개개의 음소를 나타내는 상태(state)로서 하나의 단어는 상태 S_1에서 출발해서 상태 S_N에서 끝나게 된다. 예를 들어 "학습"이라는 단어는 6개의 음소 또는 심볼 "/h/" + "/a/" + "/k/" + "/s/" + "/u/" + "/p/"으로 구성되는데 따라서 이 단어는 최소 6개의 상태로 표현된다고 가정한다. 화살표는 상태의 천이를 나타낸다.

HMM의 모수들은 세 가지로 나타낼 수 있는데 어떤 상태 S_i에서 S_j로의 천이를 나타내는 상태천이 확률 분포인 a_{ij}와, 어떤 상태 S_j에서 특정 음소 또는 심볼 k가 나타날 수 있는 확률인 $b_j(k)$이다. 마지막으로 초기 상태 확률 분포 π_i를 추정해야한다.

본 연구에서는 우선 벡터 양자화에 의해 작성된 세개의 음성 코드북에 대해 O개의 단어 또는 어휘 각각에 대해 그림 3.12에 보이는 것과 같이 N개의 상태로 구성된 독립적인 HMM을 구성한다. 즉 하나의 단어에 대해 독립적인 HMM이 하나씩 구성된다. 그런다음 HMM의 바움-웰치 모수추정을 통해 위에 기술한 세가지 모수들의 집합, $\lambda = (A, B, \pi)$를 최적화된 형태로 구한다. 각 상수 및 파라미터 그리고 변수들에 대한 기호들은 아래와 같다.

- N : 상태의 수
- T : 입력심볼의 길이 또는 수
- M : 각 상태에서 나타날 수 있는 심볼의 수

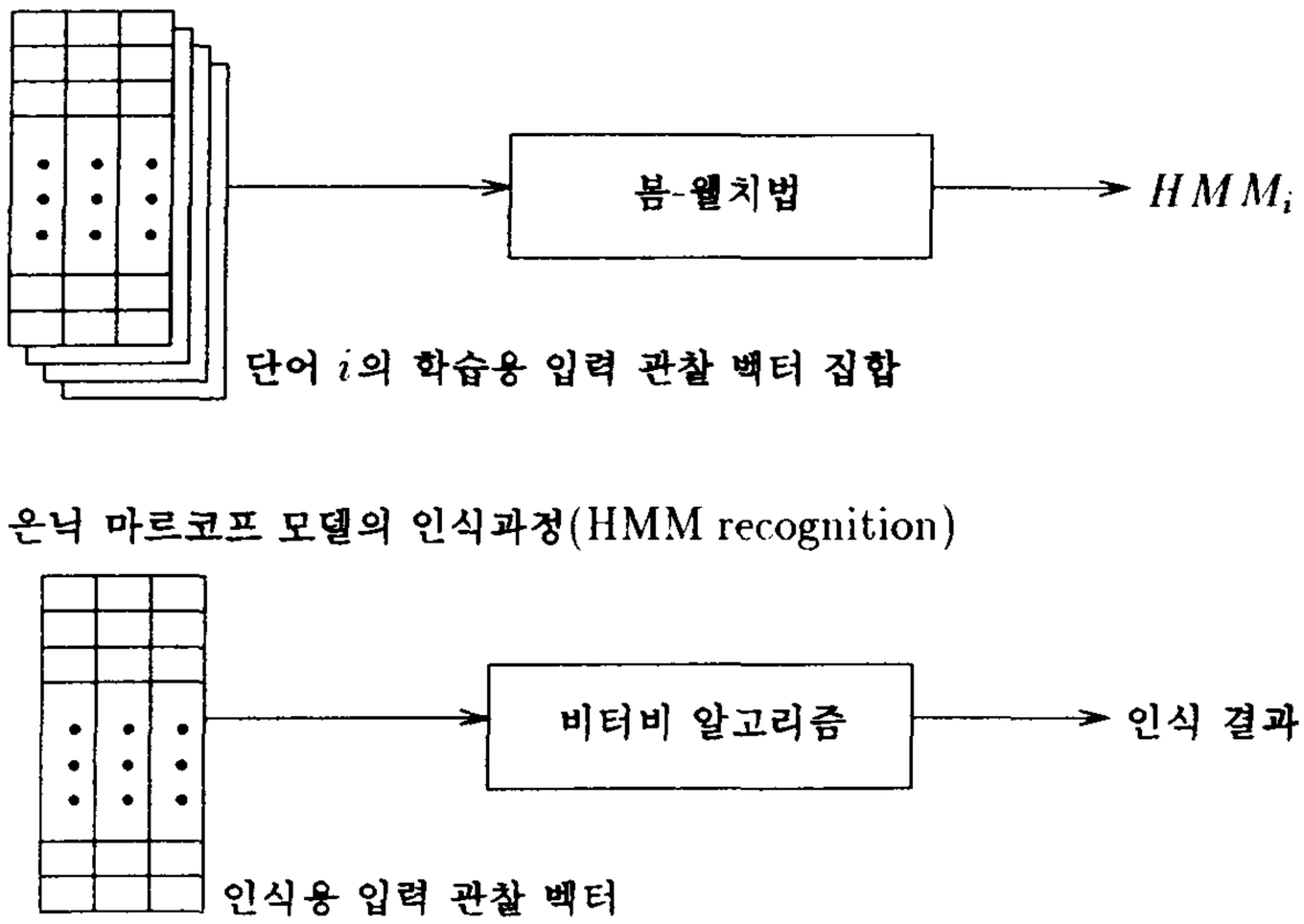


그림 3.11: HMM의 parameter estimation 및 인식도

- A : 상태 천이 확률 분포 $(1 \le i, j \le N,\ 1 \le t \le T)$

  $A = \{a_{ij}\},\ a_{ij} = P(Q_{t+1} = S_j \mid Q_t = S_i),$

- B : 입력 심볼 확률 분포 $(1 \le j \le N,\ 1 \le k \le M)$

  $B = \{b_j(k)\},\ b_j(k) = P(v_k\ at\ time\ t \mid Q_t = S_j),$

- π : 초기 상태 확률 분포

- Q : 상태의 시계열, $Q = (Q_1, Q_2, \ldots, Q_T)$

- O : 입력 심볼의 시계열, $O = (O_1, O_2, \ldots, O_T)$

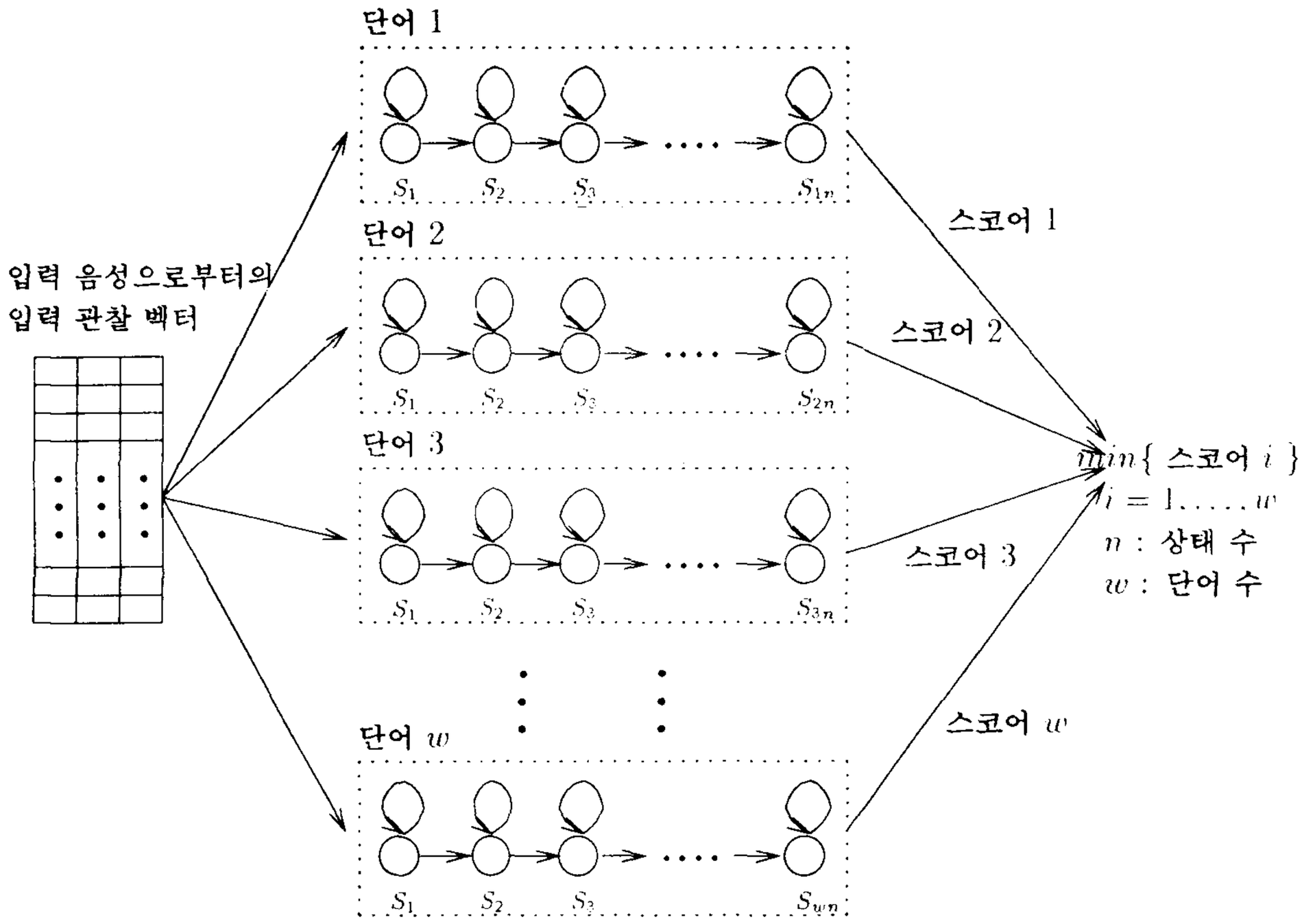


그림 3.12: HMM의 인식 과정도

포워드-백워드 절차

포워드-백워드 절차는 각 HMM에 대해 상태 천이 확률과 심볼 확률 분포를 발견하는 방법으로 포워드 변수 $\alpha_i(t)$와 백워드 변수인 $\beta_i(t)$를 정의하여 바움-웰치법에 의한 모수추정을 수행하는 알고리즘으로서 바움-웰치법의 별칭이다.

$$\alpha_i(t) \leftarrow P(O_1^{t-1} = o_1^{t-1},\ Q_t = s_i)$$

$$\beta_i(t) \leftarrow P(O_{t+1}^{T} = o_{t+1}^{T} \mid Q_t = s_i)$$

O_1^{t-1} 는 $(O_1, O_2, \cdots, O_{t-1})$ 을 표기한다. 이들 포워드 백워드 변수들은 다음의 재귀 관계식을 만족함을 보일 수 있다.

$$\alpha_i(1) \leftarrow \pi_i \cdot P(O_1 = o_1 \mid Q_1 = s_i)$$

$$\alpha_j(t) \leftarrow \sum_{i=1}^{N} \alpha_i(t-1) a_{ij} b_j(o_t),\ \ 1 \le t \le T$$

$$\beta_i(T) \leftarrow 1,\ \ 1 \le i \le N$$

$$\beta_i(t) \leftarrow \sum_{j=1}^{N} \beta_j(t+1) a_{ij} b_j(o_t),\ \ 1 \le t \le T$$

위에서 Q_t는 상태열, O_t는 출력열이다. 다음, 각 HMM 모수들의 바움-웰치법에의한 재추정은 이들 포워드 백워드 변수들을 사용하여 표현할 수 있다.

편의상, O 가 주어졌을 때 Q 가 시간 t에서 상태 s_i에 있고, 시간 $t+1$에서 상태 s_j로 전이할 확률을 $\xi_t(i,j)$ 라고 하면 (즉, $\xi_t(i,j) = P(Q_t = s_i, Q_{t+1} = s_j \mid O, \lambda)$) 다음과 같다.

$$\xi_t(i,j) \leftarrow \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$

또한, $P(Q_t = s_i \mid O, \lambda)$ 를 $\gamma_t(i)$ 라고 하면

$$\gamma_t(i) \leftarrow \sum_{j=1}^{N} \xi_t(i,j)$$

이다.

초기 상태 분포 확률 및 상태 천이 확률 그리고 입력 심볼 확률 분포의 모수들은 다음과 같이 구해지며 각기 직관적인 해석을 가지고 있다.

$$\begin{aligned}
\tilde{\pi}_i &\leftarrow P(Q_1 = s_i \mid O, \lambda) \\
&= \sum_{j=1}^{N} \xi_1(i,j) \\
\tilde{a_{ij}} &\leftarrow \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\
b_i(\tilde{k}) &\leftarrow \frac{\sum_{t=1 .s.t. O_t = v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}
\end{aligned}$$

즉, $\tilde{\pi}_i$ 는 주어진 특징 벡터 O 에 대해 첫번째 상태가 S_i 이었을 확률과 같다. $\tilde{a_{ij}}$ 는 상태 S_i 으로부터 다른상태로 천이가 일어날 경우의 수의 기대값 분의 상태 S_i 에서 상태 S_j 로 천이하는 갯수의 기대값으로 표현되며 $b_i(k)$ 는 상태 S_i 에 있는 갯수의 기대값 분의 상태 S_i 에서 심볼 v_k 를 관찰할 경우의 갯수의 기대값으로 표현된다.

비터비 알고리즘

비터비 알고리즘은 HMM 에 의한 분류에서 원래 사용되는 각 범주(단어)에 대한 확률 분포함수

$$P(O \mid \lambda) \leftarrow \sum_{\text{all } Q} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T)$$

의 은닉상태(hidden state)에 대한 가산(summation)의 부담을 줄이기 위해 주

어진 특징 벡터로부터 가장 확률이 높은 상태열($\overline{Q}$)을 결정하여 $P(O, \overline{Q} \mid \lambda)$를 근사로서 계산함으로써 높은 계산상의 효율을 얻는 방법이다. 이를 위해 Dynamic Programming(DP)의 방법론을 적용하였는데 DP 를 최단경로탐색문제(여기서는 최장경로)에 적용하는 맥락이다. 즉, 시간 t의 상태 i 에서 시간 $t+1$ 에서 상태 j 로 가는 경로의 비용(log 확률)은 $\log a_{ij} + \log b_j(O_{t+1})$ 으로서 변수 길이 T를 갖는 주어진 특징 벡터마다 정해지며 이에 따른 전체 탐색비용의 계산에 DP법이 이용된다. 본 연구에서 사용한 알고리즘은 다음과 같다.

계산목적 $\max_{q_1 q_2 \cdots q_T} P(q_1, q_2, \cdots, q_T, O_1, O_2, \cdots, O_T \mid \lambda)$ 의 계산

변수정의 $\delta_1(i) \equiv \max_{q_1, q_2, \cdots, q_t} P(q_1, q_2, \cdots, q_t, O_1, O_2, \cdots, O_t \mid \lambda)$

$\psi_t(i) \leftarrow$ 시간 t에서 상태 i에 있을때 시간 $t-1$ 의 최적 상태 색인(optimal state index)

초기화 $\delta_1(i) \leftarrow \pi_i \cdot b_i(O_1)$

$\psi_1(i) \leftarrow 0, \ 1 \le i \le N$

반복 $\delta_t(j) \leftarrow \max_i(\delta_{t-1}(i)a_{ij})b_j(O_t)$

$\psi_t(j) \leftarrow \arg\max_i \delta_{t-1}(i)a_{ij} \ \ 1 \le i, j \le N, \ 1 < t \le T$

종결 $q_T^{opt} \leftarrow \arg\max_i \delta_T(i), \ \ 1 \le i \le N$

백트래킹 $q_{t-1}^{opt} \leftarrow \psi_t(q_t^{opt}), \ \ 1 < t \le T$

$P(O, \overline{Q} \mid \lambda)$ 는 $\max_i \delta_T(i), \quad 1 \le i \le N$ 로 얻어진다. 이 최우 상태열(most likely state sequence) 의 확률이 다른 상태열의 확률에 대해 dominant 함으로서 $P(O, \overline{Q} \mid \lambda)$ 가 $P(O \mid \lambda)$ 의 좋은 근사가 될 수 있음이 알려져 있다.

3.4 PC, DSP 간 통신 및 제어

그림 3.13에서는 PC와 신호처리보드의 전체 처리 흐름과 통신 방식을 나타내고 있다. PC의 관리자는 그래픽 사용자 인터페이스를 초기화해서 인식기의 운

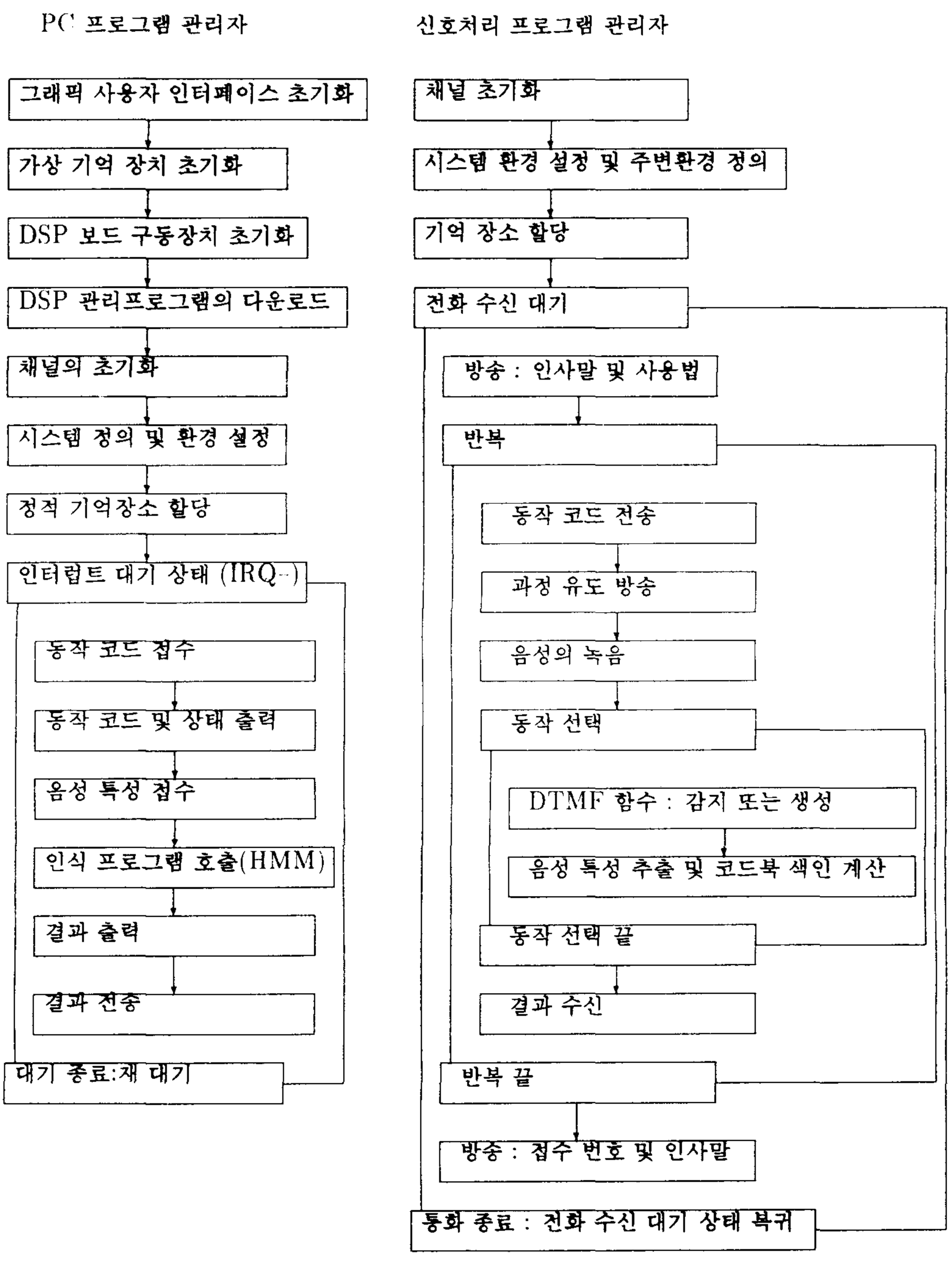


그림 3.13: PC와 DSP간의 상호 통신 방식 및 처리 흐름도

영 상태를 화면에 출력하도록 한 다음, 가상기억장치를 초기화하고 여기에 은닉 마르코프 모델의 인식용 참조 패턴들을 가상기억장치에 저장시킨다.

신호처리 보드와 PC는 서로간에 통신을 주고 받는 통신전용 메시지를 주고받음으로써 채널을 수립하는 등 환경을 수립한다. 이러한 PC와 신호처리보드간의 통신은 그림 3.14에 도시한 채널 관리 자료 블럭에 의해서 수행되는데 채널 관리 블럭은 그림과 같이 여섯개의 세부 레코드로 구성된다.

채널수립후 PC는 신호처리보드의 명령을 기다리는 대기 상태로 들어가며, 신호처리 보드는 필요할 경우 항상 인터럽트체계를 사용해서 명령을 보내고 결과를 접수한다. 신호처리 관리자도 일단의 환경설정이 끝나면 대기 상태로 들어가는데 이러한 대기 상태는 전화벨을 감지할 경우 종료된다.

전화벨이 감지되면 사용자와의 사이에 메시지 송수신을 위한 대화 서비스 상태로 들어간다. 우선 인사말과 간단한 안내말을 사용자에게 방송하고 PC 관리자에게 현재의 상태를 그대로 채널 운영 블럭을 통해 전송한다.

신호처리관리자는 사용자로부터 두가지 유형의 신호를 접수하게 되는데 하나는 사용자가 발성하는 음성이며, 하나는 사용자의 전화기에서 발생되는 숫자음이다. 일단 대화상태로 들어가면 신호처리 관리자는 필요한 단어를 발성 또는 전화기의 번호판을 누르도록 사용자에게 요청하게 되는데 사용자는 요구하는 단어를 발성하거나 전화기 번호판을 누르며 신호처리 관리자는 이를 녹음한 다음 음성 특성 추출을 수행한다. 그런 다음 신호처리 관리자는 PC 관리자에게 코드북 색인 리스트를 전송하게 되는데, PC 관리자는 이를 받아서 은닉 마르코프 수행 모듈을 호출한다. 은닉 마르코프 모듈은 이를 이용해서 인식 과정을 수행한 다음 인식 결과를 PC 관리자에게 제시하고 이를 DSP 관리자에게 전송해서 신호처리 관리자가 인식 결과를 사용자에게 알려주고 정확한 인식여부를 재 검증하는 절차를 거치게 된다. 검증시 사용자는 "예"나 "아니오"의 간단한 대답만을 하면 된다.

이러한 과정은 매번 PC 관리자에게 전송되어서 진행 과정을 화면에 출력하도록 되어 있다. 사용자와의 대화가 종료되면 인식기는 종료 메시지와 인사말을 방송

채널 관리 데이타 블럭(Channel Management Data Block)의 구조

화일 지정자	채널 지정자	명령어블럭	신호처리기 지정자	데이타 버퍼	인식지정자

화일 지정자 : 현재 디스크로부터 읽거나 쓰여질 화일 지정자
채널 지정자 : 지정된 신호처리 보드에 해당하는 채널 번호 또는 지정자
명령어 블럭 : PC 또는 신호처리 관리자에게 보내지는 명령어 블럭
신호처리기 지정자 : 명령어, 제어신호, 데이타가 송수신되는 신호처리기의 번호
데이타 버퍼 : PC 또는 신호처리 관리자에게 보내지는 데이타를 위한 버퍼
인식 지정자 : 인식된 단어들을 저장하는 벡터

명령어 블럭의 구조

길이	명령어	반환값	기타

길이 : PC 또는 신호처리 관리자에게 전송되는 데이타의 전체 길이
명령어 : 제어신호, 특정한 행위를 유발시키는 등의 명령어
반환값 : 명령어의 수행 결과를 나타내는 지정된 값
기타 : 명령어의 확장으로 필요시만 사용

인식 지정자

인식단어1	인식단어2	인식단어3	인식단어4	.....	인식단어N

그림 3.14: PC와 DSP간의 채널운영 및 메시지, 데이타 전송을 위한 채널운영블럭의 구성 / 인식된 단어의 리스트를 저장하는 인식 단어 벡터의 구성

		고주파 군			(Hz)
		1209	1336	1477	1633
저주파 군 (Hz)	697	1	2	3	A
	770	4	5	6	B
	852	7	8	9	C
	941	*	0	#	D

하고 회선을 끊게 되며, 다시 전화벨 감지 대기 상태로 들어 간다.

3.5 DTMF의 감지 및 발생

일반적인 전자식 전화기의 번호판 배열방식과 이와 관련한 주파수 분할표는 위의 표와 같다. DTMF란 위의 표와 같이 고주파군(high frequency group)의 주파수와 저주파군(low frequency group)의 주파수를 합성한 개념으로 이의 합성에 따라 일반적인 전화기의 신호음을 생성하게 되는 것이다.

DTMF 신호의 검출은 음성 특성 추출과 마찬가지로 일단 DTMF 신호를 녹음한 뒤 이로부터 DTMF의 신호영역을 추출하는 신호추출(segmentation) 과정을 거쳐 앞서 언급한 프레임을 구성한다. 검출된 신호에 대해 영교차율(Zero Crossing Rate, ZCR)을 구하고 고속 푸리에 변환(Fast Fourier Transform, FFT)을 거치게 된다. 본 시스템에서는 각 추출된 각 DTMF 신호영역을 대상으로 1024 포인트 FFT를 거친후에 20-300, 102-300번째 주파수 대역중에서 각각 피크를 추출하여 큰 쪽을 고주파군, 작은쪽을 저주파군에 정합시켜 DTMF 신호를 검출하고 있다.

FFT 함수는 다음 식에 근거한다.($N = 1024, j = \sqrt{-1}$)

$$X(k) \leftarrow \sum_{n=1}^{N} x(n) e^{-j\frac{2\pi}{N}kn}, \quad k = 1, 2, \ldots, N$$

3.6 실험 및 결과

본 시스템은 대전시 모 구청의 자동 민원처리 시스템의 음성 인터페이스로서 실용화 과정을 거쳤으며 총 86개의 어휘로 구성되어 있다.

이산 HMM의 학습 및 인식에 사용된 음성 데이타베이스는 다음과 같은 어휘로 구성되어 있다.

- 번지수를 나타내기 위한 27개 숫자음

  천, 이천, ···, 구천, 백, 이백, ···, 구백 십, 이십, ···, 구십

- 숫자음 11개 : 주민등록번호나 번지수 인식에 사용

  영, 공, 일, 이, 삼, 사, 오, 육(륙), 칠, 팔, 구

- 발급 통수 12개

  한통, 두통, 세통, ···, 열통, 석통, 넉통

- 인식 검증용 2단어와 사이음

  예, 아니오, 다시

- 서류명 3종

  호적초본, 호적등본, 건축물관리대장

- 동명 30개

  가수원동, 가장동, ···, 내동, 도마동, ···, 흑석동 등

위의 어휘는 실제 다음과 같은 문자열 포인터 어레이에 저장되어 있다.

```
char    *table[86] = {

    "1000", "2000", "3000", "4000", "5000", "6000", "7000", "8000", "9000",

    "100", "200", "300", "400", "500", "600", "700", "800", "900",
```

"10", "20", "30", "40", "50", "60", "70", "80", "90",

"0", "0", "1", "2", "3", "4", "5", "6", "7", "8", "9",

"1", "2", "3", "4", "5", "6", "7",

"8", "9", "10", "3", "4", "예 ", "아니오 ", "-",

"호적등본", "호적초본", "건축물관리대장",

"가수원동", "가장동", "갈마동", "관저동", "괴곡동", "괴정동",

"내동", "도마동", "도마1동", "도마2동", "도안동", "둔산동",

"매노동", "변동", "복수동", "봉곡동", "산직동", "삼천동",

"오동", "용문동", "용촌동", "우명동", "원정동", "월평동",

"장안동", "정림동", "탄방동", "평촌동", "흑석동", "기성동" };

음성 데이타베이스의 구성은 시내 전화를 대상으로 200명분의 음성 데이타 수집를 수집하였으며, 이중 100명분은 학습용으로 나머지 100명분은 실험용으로 사용하였다.

이중 서류명에 대해 약 98%의 인식률, 숫자음에 대해 약 92%의 인식률, 동명에 대해 약 90%의 인식률을 나타내었다.

여 백

4 장

로보트 제어용 음성 인터페이스 시스템

4.1 시스템의 개요 및 구조

본 시스템은 전형적인 Voice Commander 시스템의 하나인 로보트 제어를 위한 음성 인터페이스를 공중전화망을 대상으로 구현한 시스템이다. 이러한 시스템은 일반적으로 격리된 지역간에 수립되어 있는 전화망을 대상으로 격리된 지역에 존재하는 전자적 기기나 설비를 원격조종할 수 있는 기능의 실현을 음성이라는 매체를 통해 구현한 것이 된다.

현재 가정용 또는 산업용 시스템중에는 전화기의 DTMF 버튼을 통해 특정한 가전기기나 산업기기를 제어할 수 있는 시스템이 이미 상용화되어 사용중에 있다. 이러한 시스템의 주요 단점은 필요할 때마다 사용자가 제어 코드를 기억해서 해당되는 제어코드와 관련된 번호를 전화기의 버튼을 통해 입력시키는 방식으로서, 자연성이 없으며 입력속도가 느린 단점이 있다. 또한 전화를 받는 ARS 서버시스템

에서 제공하는 획일적인 시나리오에 의해 운영되기때문에 사용자는 이러한 대화과정에 적극적으로 참여할 수가 없게 되며 따라서 거부감이 생기게 된다.

본 시스템은 전화망을 통해 사용자 또는 기기관리자가 직접 전화를 걸어서 로보트 시스템과 대화를 하듯이 로보트를 제어할 수 있는 잇점이 있으며, 2장에서 기술한 핵심어 추출(keyword spotting)법을 적용하여 사용자의 발화문장 가운데에 존재하는 로보트 제어에 필요한 어휘를 추출하게 되어 있다.

핵심어 추출법에 의해 추출된 제어에 관계된 명령어는 PC에 설치된 관리자 프로그램에 의해 제어신호가 발생되도록 되어 있어서 실제 로보트 시스템에 제어신호를 전달해서 로보트가 구동되는 것을 실연되도록 하고 있다.

본 시스템의 구조는 그림 4.1에서와 같이 전장에서 기술한 민원처리 시스템과 유사한 하드웨어 사양을 가지고 있으며 상이한 점은 실제 로보트의 팔이 설치되어 있고 PC로부터 serial port를 통해 제어신호가 전송되도록 되어 있다는 점이다.

소프트웨어는 역시 크게 두부분으로 구분되는데 PC 관리자 프로그램과, DSP 관리자 프로그램으로 구성된다. PC 관리자 프로그램은 다시 DSP 보드와의 통신 수립과 네개의 DSP 보드의 다중처리를 지원하는 주 프로그램, 로보트의 동작을 그대로 화면상에 제시하는 그래픽 사용자 인터페이스 부분, 사용자가 원하는 로보트에 원하는 제어신호를 발생시킬 수 있도록 하는 사용자 선택형 로보트 기기 제어부분, 로보트를 실제 구동하도록 제어 신호를 발생시키고 serial port를 통해 전달하도록 하는 네 개의 하부 구조로 구성되어 있다.

DSP 보드상에서 수행되는 프로그램도 역시 네 개의 부분으로 구성되어 있는데 PC 관리자와의 통신을 전담하고 전화벨 감지후 사용자와의 회선을 수립하는 주 프로그램, 사용자의 음성을 녹음하고 재생하는 부분, 녹음된 음성으로부터 신호처리를 통해 음성 특성을 추출하는 부분, 마지막으로 본 시스템의 핵심 부분인 연속 HMM에 의한 음성인식 부분이 그것이다.

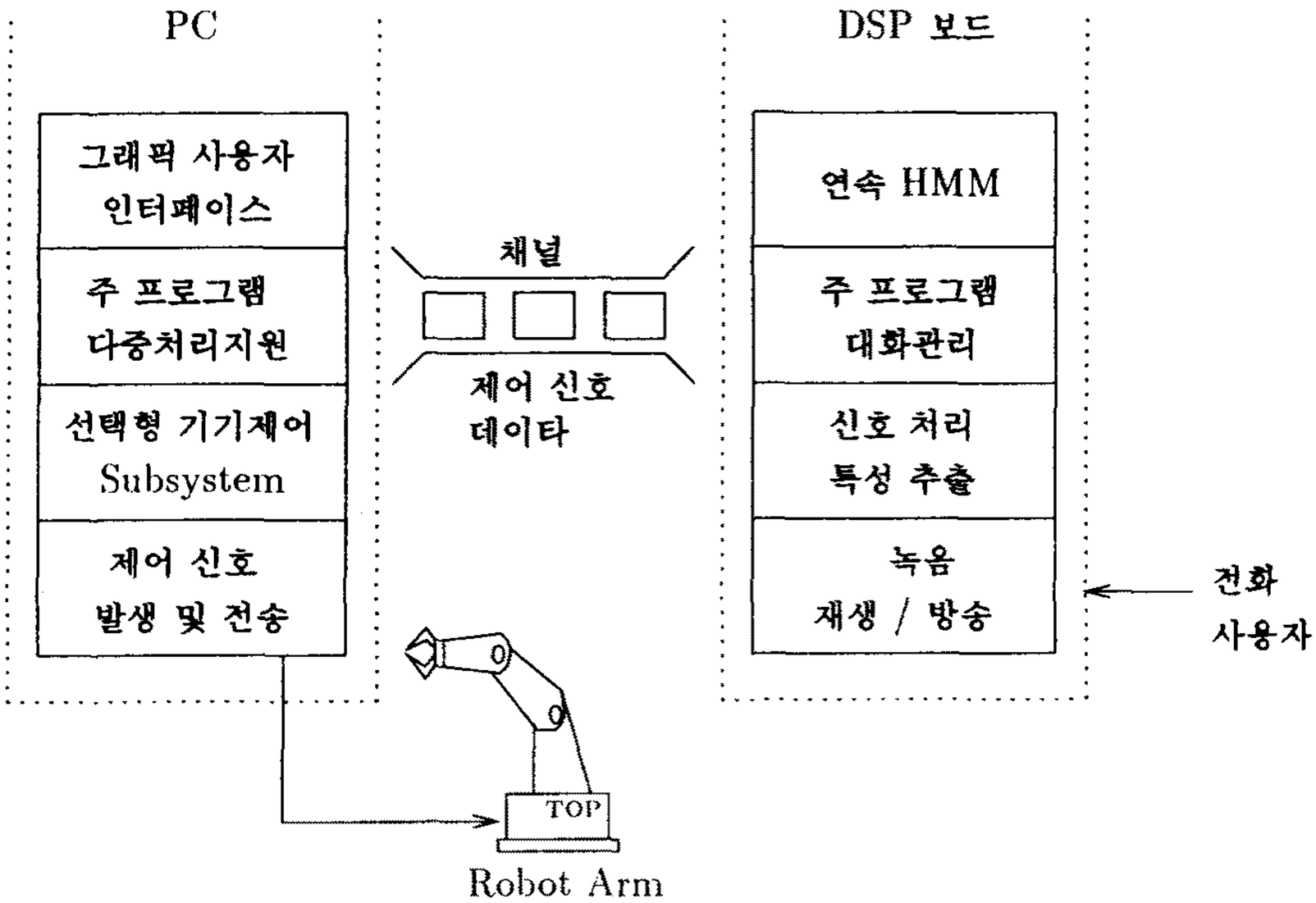


그림 4.1: 로보트 제어 음성 인식 인터페이스의 구조

4.2 신호 처리 및 특징 추출

전화 회선을 통해 입력된 사용자의 음성에는 전화회선상에 존재하는 잡음, 백색잡음(white noise), 환경잡음 등 다양한 출처의 잡음이 존재한다. 신호처리부에서는 이러한 잡음 요소를 고려해서, 특징추출을 수행하는데 입력 음성 신호를 전장에서 기술한 방법에 따라 대역 통과 필터, 증폭, 음성 구간 검출을 거쳐 LPC 분석을 수행한다. LPC 계수에 대해 켑스트럼과 차분 켑스트럼을 구하면 특징추출이 종료된다. (그림 4.2 참조)

이산 HMM과는 달리 연속 HMM에서는 코드북이 불필요하며, 신호처리 및 특징추출부에서 만들어진 켑스트럼(12차), 차분 켑스트럼(12차), 파워 및 차분파워(2차)를 하나의 벡터내에 연속으로 붙여서 만들어진 26차의 특징벡터를 각 음성 구간 프레임마다 도출해서 그대로 이산 HMM의 모수추정과 인식에 사용하고 있

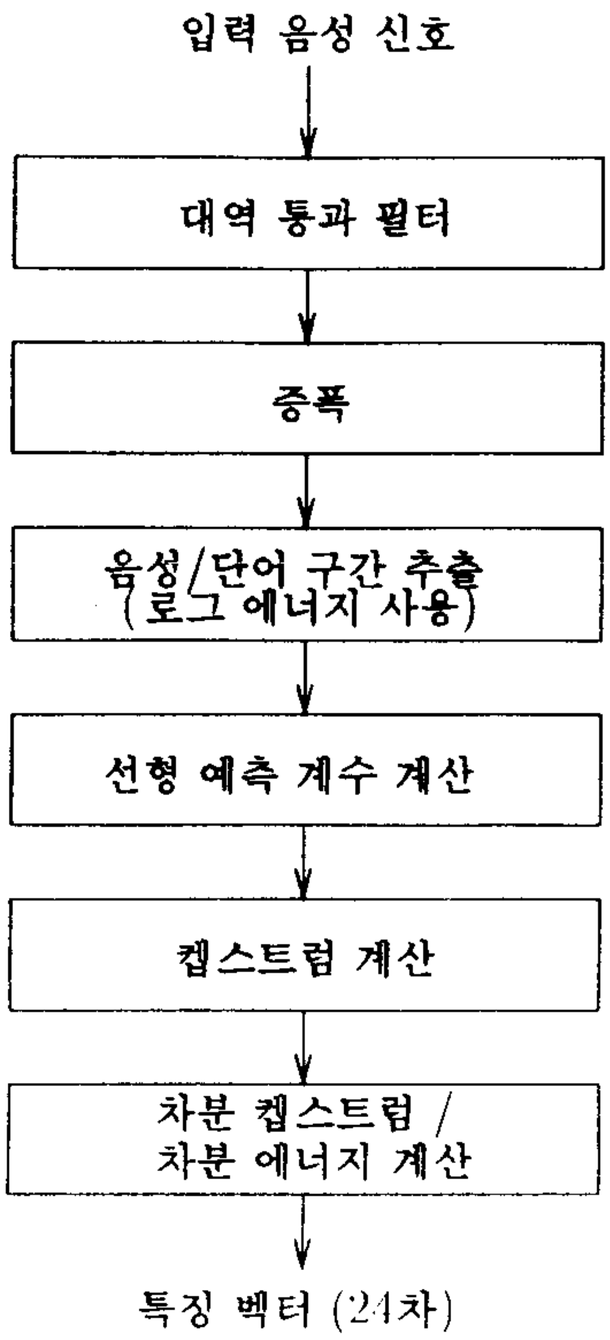


그림 4.2: 음성신호로부터의 특징 추출

다.

4.3 연속 HMM을 이용한 패턴인식

자연스러운 로보트 구동 명령 문형으로부터 핵심단어를 추출(keyword spotting), 인식하는 기능을 연결단어 인식의 framework 으로 구현하였다.

즉, 문법 network 에서 핵심 단어들(로보트 몸체의 부분, 방향, 각도, 동작명령) 이외의 단어들 (조사,감탄사 등) 에 대해서 garbage HMM, 그리고 silence 에 대해 silence HMM 을 별도로 형성, 학습 및 인식에 포함시켜 연결단어 인식

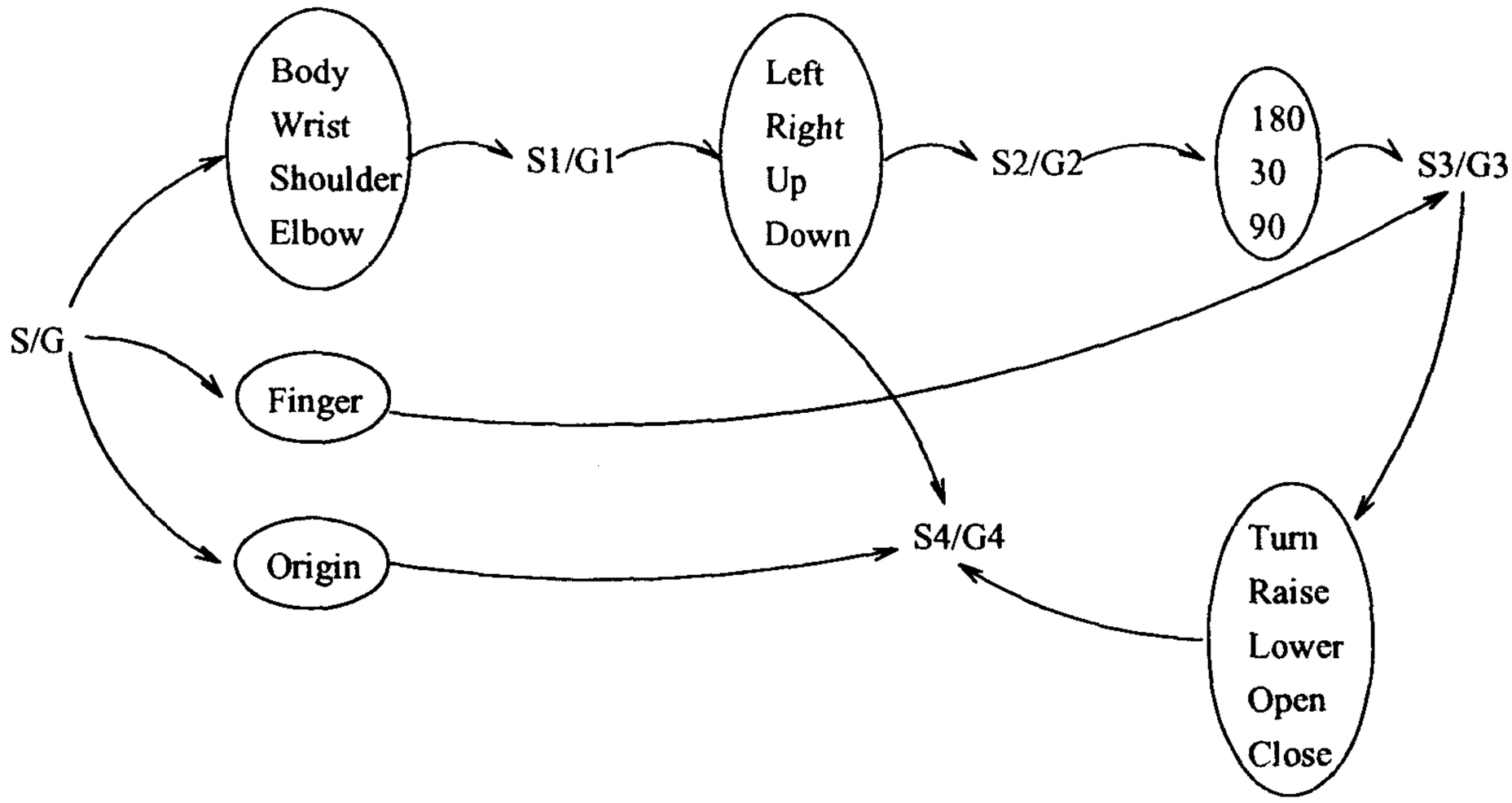


그림 4.3: 로보트 암 제어를 위한 문법 network의 구성

을 수행함으로써 핵심단어 인식의 효과를 얻었다. 인식된 단어 열에서 garbage 단어를 제외한 단어들이 인식된 핵심단어들이 된다. 그림 4.3은 본 시스템에서 사용하는 문법을 Finite State Network (FSN) 으로 도해하고 있는데, 원안의 문자는 인식 단어 즉 핵심 단어를 의미하며 원밖의 문자는 각각 S#는 silence, G#은 garbage 모델을 의미한다. 또한 연속음성을 위한 FSN의 구성은 그림 4.4에 제시하고 있다.

HMM 의 상태 확률 모델은 continous mixture density 를 사용하였다.

$$b_j(O) = \sum_{m=1}^{M} c_{jm} N(O \mid \mu_{jm}, \Sigma_{jm})$$

$j = 1, \cdots, N \;\; m = 1, \cdots, M$. $N(O \mid \mu_{jm}, \Sigma_{jm})$ 은 정규분포의 확률밀도 함수를 나타낸다

$$\mid 2\pi\Sigma \mid^{-p/2} \exp\{-\frac{1}{2}(O - \mu_{jm})' \Sigma_{jm}^{-1} (O - \mu_{jm})\}$$

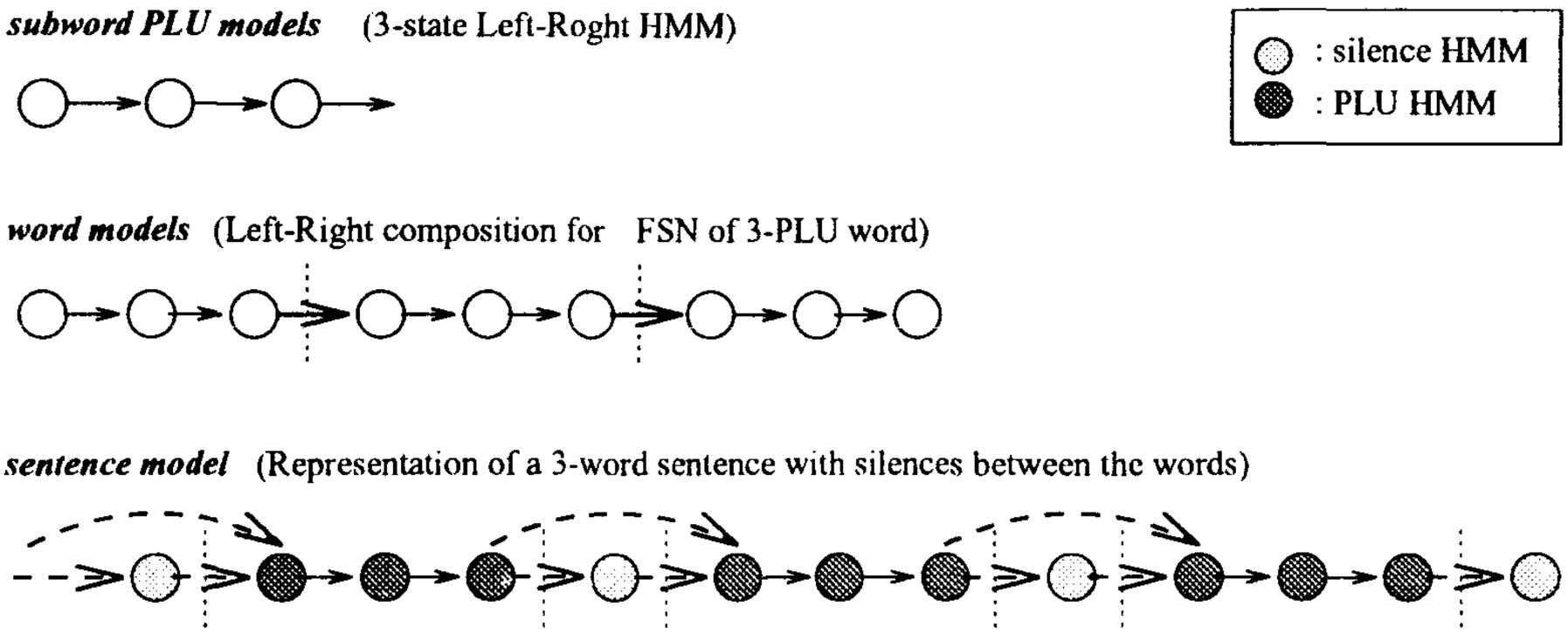
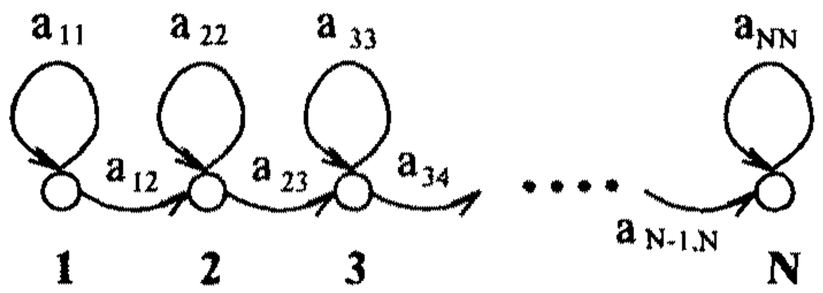


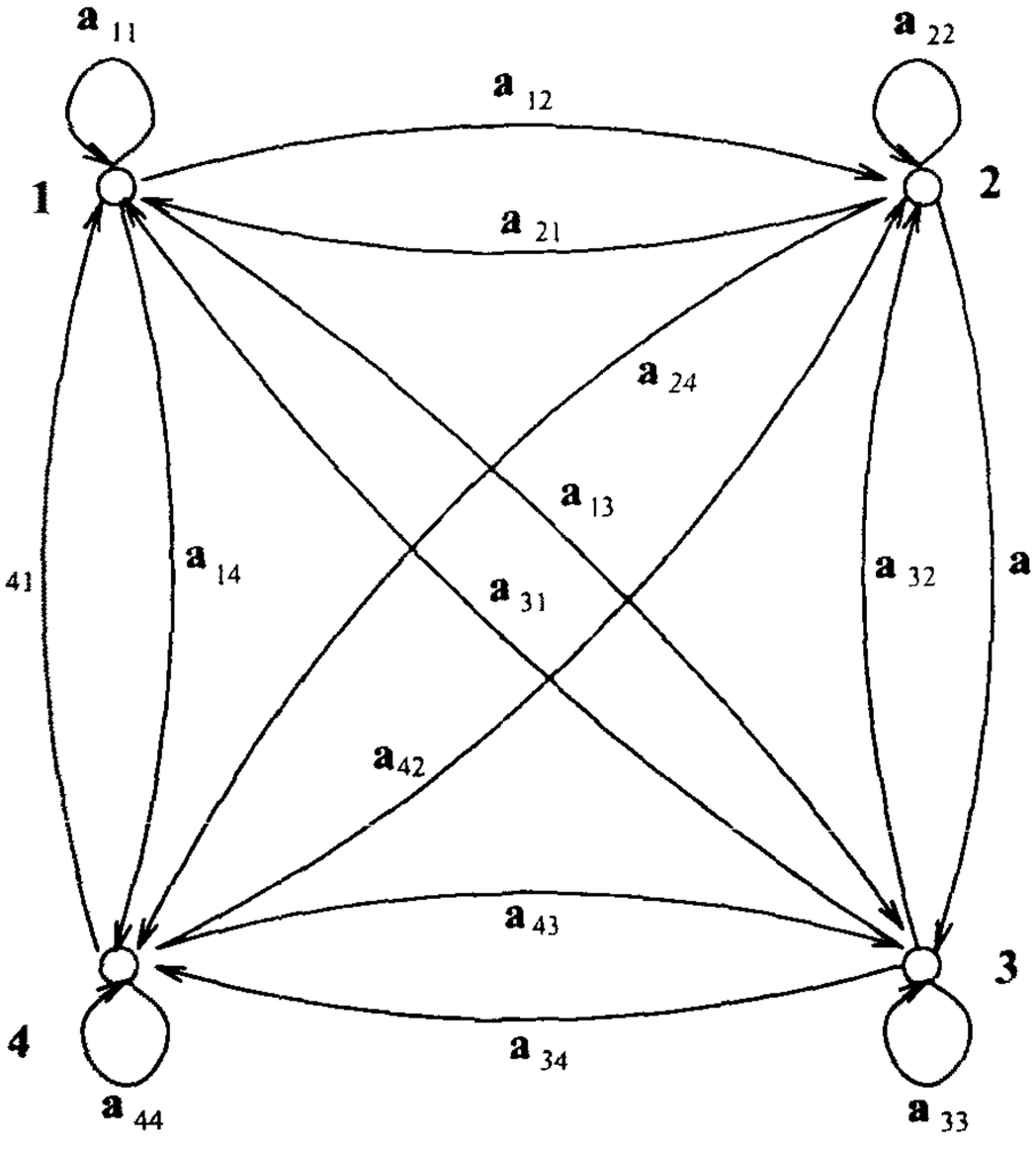
그림 4.4: 연속 음성인식을 위한 FSN의 구성

이때, M 은 mixture 의 갯수, N 은 단어 HMM 의 상태의 갯수이며 p 는 특징 벡터의 차원을 나타낸다. 상태 천이 확률에 있어 silence 및 garbage 모델의 경우 상태간의 full communication을 가정하였으며 핵심단어 모델의 경우 left-to-right 구조를 채택하였다. 그림 4.5 의 (가) 와 (나) 는 각각 left-to-right 형과 ergodic 형의 상태 천이 구조를 나타낸다. N 은 핵심단어의 경우에 음소의 갯수 더하기 3 ~ 5을 주었으며, silence 및 background 잡음 에는 5 개, garbage HMM 에는 10 개 정도를 주었다.

학습은 각 핵심단어 별 데이타, garbage 단어들의 집합의 데이타, 그리고 silence 및 background 잡음의 데이타에 대하여 단어별 HMM 학습을 수행하였다. 학습으로부터 얻어진 단어별 파라미터 집합(parameter set)을 위에서 설명한 FSN에 따라 loading 하여 연결 단어 인식을 수행한다. 인식 탐색 알고리즘은 비터비 탐색의 단어 level 을 포함한 연장으로서 frame synchronous one-pass 알고리즘[31] 을 구현하였다.



(가)



(나)

그림 4.5: (가) left-to-right 형 상태 천이 (나) ergodic 형 상태천이

4.4 단어별 학습

학습 알고리즘은 standard EM 알고리즘 (바움-웰치 알고리즘)을 사용하였다. 연속형 mixture density HMM 의 forward-backward 절차에의한 학습과정은 잘 알려져있다[28].

Baum 의 auxiliary 함수의 반복적인 maximization 을 통해 HMM 우도함수(likelihood function)의 local maxima 로 수렴하도록 하는 파라미터 추정치 열을 얻을 수 있다.

$$Q(\lambda, \lambda') = \sum_{\mathbf{q}} \log P(\mathbf{O}, \mathbf{q} \mid \lambda) P(\mathbf{O}, \mathbf{q} \mid \lambda')$$

$$Q(\lambda, \lambda') \geq Q(\lambda', \lambda') \Rightarrow P(\mathbf{O} \mid \lambda) \geq P(\mathbf{O} \mid \lambda')$$

log 함수의 성질로부터 $Q(\lambda, \lambda')$ 은 HMM 의 각 파라미터 성분에 대한 Q 함수들로 분해되며 그 합으로 표현된다.

$$Q(\lambda, \lambda') = Q(\pi, \lambda') + Q(A, \lambda') + Q(\mathbf{\Theta}, \lambda')$$

여기서 π 는 HMM 의 초기 상태 분포 확률, A 는 상태 천이 확률 행렬이며 $\mathbf{\Theta}$ 는 각 상태의 observation 확률을 결정하는 연속형 mixture 분포들의 파라미터 집합을 나타낸다. ($\mathbf{\Theta} = \{(c_{jm}, \mu_{jm}, \Sigma_{jm})\}, \quad j = 1, \cdots, N \quad m = 1, \cdots, M$) HMM 파라미터 성분에 대한 바움-웰치 알고리즘의 재추정 공식은 이 Q 함수의 성

분들을 해당 파라미터에 대해 maximize 함으로써 구할 수 있으며 다음과 같다.

$$\tilde{\pi}_i = \frac{P(\mathbf{O}, q_1 = i \mid \lambda)}{P(\mathbf{O} \mid \lambda)}$$

$$\tilde{a}_{ij} = \frac{\sum_{t=2}^{T} P(\mathbf{O}, q_{t-1} = i, q_t = j \mid \lambda)}{\sum_{t=2}^{T} P(\mathbf{O}, q_{t-1} = i \mid \lambda)}$$

이때, 주어진 $\mathbf{O}$ 에 대하여 시간 t 에 상태 j 의 k 번째 mixture component 에 있을 확률을 $\gamma_t(j,k)$ 라 하면

$$\gamma_t(j,k) = \frac{P(\mathbf{O}, q_t = i \mid \lambda)}{P(\mathbf{O} \mid \lambda)} \cdot \frac{c_{jk} N(\mathbf{o}_t \mid \mu_{jk}, \Sigma_{jk})}{\sum_{m=1}^{M} c_{jm} N(\mathbf{o}_t \mid \mu_{jm}, \Sigma_{jm})}$$

그리고,

$$\tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)}$$

$$\tilde{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \mathbf{O}_t}{\sum_{t=1}^{T} \gamma_t(j,k)}$$

$$\tilde{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (\mathbf{O}_t - \mu_{jk})(\mathbf{O}_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)}$$

이들 재추정 공식의 계산은 다음과 같이 정의되는 forward 변수 $\alpha_t(i)$ 와 backward 변수 $\beta_t(i)$ 를 이용한다.

$$\alpha_t(i) = P(\mathbf{o}_1, \cdots, \mathbf{o}_t,\ q_t = i \mid \lambda)$$

$$\beta_t(i) = P(\mathbf{o}_{t+1}, \cdots, \mathbf{o}_T \mid q_t = i, \lambda)$$

이 변수들은 다음의 관계를 만족한다.

$$\alpha_1(i) \leftarrow \pi_i b_i(\mathbf{o}_1) \quad 1 \le i \le N$$

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(\mathbf{o}_{t+1}), \quad 1 \le t < T$$

$$\beta_T(i) \leftarrow 1, \quad 1 \le i \le N$$

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(\mathbf{o}_{t+1}), \quad 1 \le t < T$$

또한 위의 재추정 공식 계산에서,

$$P(\mathbf{O}, q_t = i \mid \lambda) = \alpha_t(i)\beta_t(i)$$

$$P(\mathbf{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

$$P(\mathbf{O}, q_{t-1} = i, q_t = j \mid \lambda) = \alpha_{t-1}(i) a_{ij} b_j(\mathbf{o}_t) \beta_t(j)$$

가 성립함을 보일 수 있다.

4.5 one-pass 비터비 탐색 알고리즘

연결된 단어들로서의 연속 단어 인식의 방법론은 보통, 인식을 위해 사용되는 모든 지식들(예, 단어 표현, 언어 모델)을 확률적인 또는 결정적인 네트워크로 나타내고, 음성을 나타내는 subword 또는 단어 모델의 기본 네트워크와 결합하고, 전 네트워크를 Dynamic Programming 을 이용하여 효과적이고 정확하게 탐색할 수 있다는 생각에 기반을 둔다.

이 문제를 풀기위한 여러 알고리즘들이 개발 발전되어 왔는데, Stack 알고리즘, Jelinek[14], Level-Building 알고리즘, Myers와 Rabiner의 알고리즘[21], one-stage DP 알고리즘, Ney[24]. frame-synchronous DP 탐색 알고리즘, Lee와 Rabiner[31] 등이 알려져 있다. 이들은 모두 문법적 제약 하에서 최적의 단어열을 구하는 기능을 가지고있다. 이들 탐색 알고리즘들의 주된 차이점은 알고리즘 구현상의 특색들이다. (예, 프레임 synchronous 또는 단어 synchronous)

그림 4.6 는 HMM 단어 모델을 기반으로한 Finite State Network 의 frame-synchronous 비터비 탐색 알고리즘을 설명하기 위한 도표이다. 단어인식에서의 상태와 상태사이의 천이를 고려하는 비터비 탐색으로부터 한 레벨 더 확장하여 매 시간 $t = 1, \cdots, T$ 에서 단어와 단어 사이의 천이를 또한 함께 고려하는 탐색을 하는 것이 기본적인 생각이다. 보통 FSN 을 이야기 할 때 grammar node

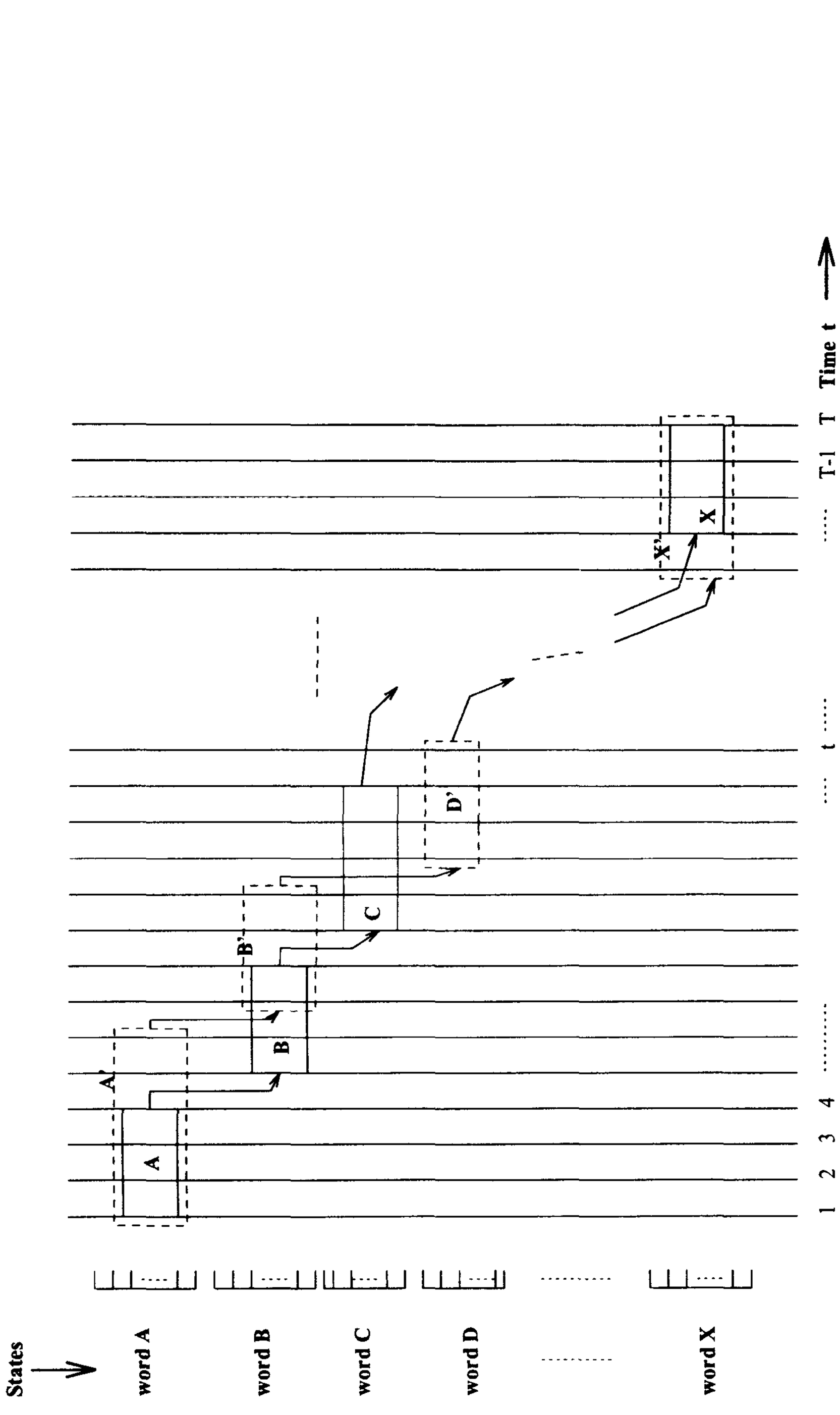


그림 4.6: one-pass 비터비 알고리즘을 이용한 네트워크 탐색

라는 자료 구조가 추가되어 다루어지나 grammar node 를 따로 표현하지않고 단어인식의 경우의 직접적 연장으로서 설명할 수 있겠다.

그림에서 직선이나 점선의 직사각형들은 어느 한 단어의 상태들에 입력 음성이 머무르고 있는 경우를 나타낸다. 즉, 처음의 **A** 로 표시되어 있는 상자는 시간 $1 \leq t \leq 4$ 동안 단어 **A**에 머무르고 있음을 나타낸다. 그림에서는 문법 네트워크상 가능한 두개의 다른 단어열을 나타내고 있는데 ($\mathbf{A}-\mathbf{B}-\mathbf{C}-\mathbf{X}$ 와 $\mathbf{A}'-\mathbf{B}'-\mathbf{D}'-\mathbf{X}'$), 시작부분의 $\mathbf{A}-\mathbf{B}-$ 와 $\mathbf{A}'-\mathbf{B}'-$ 는 같은 단어열 **A** 와 **B** 이나 시간상의 다른 duration 을 나타내는 경우이며, 첫번째 열에서는 단어 **B** 다음 단어 **C** 로 가지만 두번째 열에서는 단어 **B** 다음 단어 **D** 로 감을 나타낸다.

비터비 탐색은 이러한 가능한 duration 과 천이를 가진 모든 단어열 중에 가장 높은 우도 점수를 갖는 (또한, 그 열의 각 단어 안에서는 각기의 HMM 단어 모델에 따라 optimal 상태열을 갖는) 열을 탐색해 내는 문제로서, 2-레벨 Dynamic Programming 의 방법론이되며, 매 시간 t 에 단어 k 의 HMM 상태 i 에 도달하는 optimal path 의 누적 점수들과 앞의 단어 j 로부터 천이후 단어 k 의 초기 상태에 도달되는 optimal path 의 누적 점수들을 계산하여 path updating을 수행한다.

이 path updating 에서 고려해야 할 변수들은 아래와 같으며 HMM 의 일반적인 천이구조에 대해 성립한다. (변수명은 구현된 로보트 구동 시스템에 사용한 것들이다.)

sum : 시간 t 에서 특징 벡터의 local 우도 점수

tempmax : 바로 전 단어의 끝으로부터 다음단어 jj 의 상태 k 로 천이가 일어날 경우 best path 의 누적 점수

likmax : 단어 jj 의 바로 전 상태에서 같은 단어 jj 의 다음 상태 k 로 천이가 일어날 경우 best path 의 누적 점수

lik[jj][k] : 시간 t 에 단어 jj 의 상태 k 에 도달하는 best path 의 누적 점수

likprev[jj][k] : 시간 $t-1$ 에 단어 jj 의 상태 k 에 도달하는 best path 의 누적 점수

LIKprev[ii] : 시간 $t-1$ 에 단어 ii 의 마지막 상태에 도달하는 best path 의 누적 점수

LIK[ii] : 시간 t 에 단어 ii 의 마지막 상태에 도달하는 best path 의 누적 점수

wtrack[t][jj] : 시간 t 에 가장 좋은 우도 점수의 전 단어

elapse[jj][k] : 단어 jj 에 들어온 후 시간 t 까지의 duration 길이

wlength[t][jj] : 시간 t 에 끝나는 단어 jj 의 총 duration

위의 변수들에는 다음의 관계가 있다.

$$\textbf{tempmax} = \max_{ii}\{\textbf{LIKprev}[ii] + A[ii][jj] + init[jj][k]\}$$

여기서 $A[ii][jj]$ 는 단어 ii 에서 단어 jj 로의 천이 파라미터이며, $init[jj][k]$ 는 단어 jj 의 초기 상태 k 의 확률의 로그값 이다.

$$\textbf{wtrack}[t][jj] = \arg\max_{ii}\{\textbf{LIKprev}[ii] + A[ii][jj] + init[jj][k]\}$$

$$\textbf{likmax} = \max_{j\in\{\text{states of word } jj\}}\{\textbf{likprev}[jj][j] + a[jj][j][k]\}$$

여기서 $a[jj][j][k]$ 는 단어 jj HMM 의 상태 j 에서 상태 k 로의 천이 확률의 로그 값이다.

$$\mathbf{lik}[jj][k] = \max(\mathbf{tempmax}, \mathbf{likmax}) + \mathbf{sum}$$

Decoding 을 위한 변수들은,

MLWS[k] : optimal 단어 열의 끝에서 k 번째 단어

WLENGTH[k] : **MLWS**[k] 의 duration 길이

이다. 그러면 path update 에서 얻어진 정보를 사용하여 역추적 (backtracking) 은 다음과 같이 된다.

시간 $t = T$, (프레임 길이) 에 대해서,

$\mathbf{MLWS}[1] = \arg\max_i \mathbf{LIK}[i]$

$\mathbf{WLENGTH}[1] = \mathbf{wlength}[T][MLWS[1]]$

$k \leftarrow 1, \quad t \leftarrow T$

$t \leftarrow t - \mathbf{WLENGTH}[k]$

if $(t < 0)$ stop

$\mathbf{MLWS}[k+1] = \mathbf{wtrack}[t+1][\mathbf{MLWS}[k]]$

$\mathbf{WLENGTH}[k+1] = \mathbf{wlength}[t][\mathbf{MLWS}[k+1]]$

$k \leftarrow k + 1$

Repeat

최적의 단어열을 탐색하는 decoding 알고리즘의 전체적인 플로우는 다음과 같다.

탐색 알고리즘 :

1. DP 탐색 알고리즘의 네트워크상의 초기조건을 부여한다.
   관련 변수들의 초기화
2. 매 시간 $t(t = 1. \cdots .$ 프레임 길이 $T)$ 에서. 누적된 우도 점수들과 optimal path 를 역 추적 하는데 필요한 관련 변수들을 update 한다.
   각 단어 k 에 대하여 optimal path 의 update 는 다음 중 한가지로 된다.
   - 단어 k 모델 안에서 :
     해당 프레임에 대한 local distance (observation 과 상태 천이의 로그 확률 값) 의 계산
     시간 t 에 단어 k 의 각 상태에 도달하는 best path 의 누적 점수 update
     시간 t 에 단어 k 가 끝나는 경우에 대한 누적 점수 update
     duration 변수의 증가
   - 단어들의 문법 네트워크 (FSN) 레벨에서 :
     전 단어로 부터의 탈출과 다음 가능한 단어의 초기 상태로의 천이 maximum 누적 점수를 주는 가장 좋은 전 단어를 찾는다. (optimal 단어 열 역 추적을 위한 정보 저장)
     시간 t 에 단어 k 의 초기 상태를 시작하는 best path 의 누적 점수

update

duration 변수 1 로 setting

3. 시간 T 에서의 optimal 최종 단어로 부터 추적 변수를 사용 역추적하여 가장 좋은 우도 점수의 단어열을 얻는다.

위의 알고리즘에서 duration 변수를 1 로 설정하는것은 $\mathbf{elapse}[jj][k] \leftarrow 1$ 을 말한다.

4.6 실험 및 결과

로보트 제어용 학습 데이타는 그림 4.7 같은 개념적 문 단위를 통해 수집하였다.

A : 본체, 팔목, 어깨, 팔꿈치, 팔목, 손가락

B : 왼쪽, 오른쪽, 좌, 우, 위, 아래, 아래로

C : 삼십도, 구십도, 백팔십도

D : 올려, 내려, 돌려, 벌려, 닫아

G : garbage 단어들 (를,을,어,주세요,봐요,음,등등)

연속 HMM의 학습 및 실험에 사용된 음성데이타 예문은 다음과 같이 수집하였으며, 15초 이내에 주어진 문장을 자연스럽게 세 번 발성하도록 유도함과 동시에 숨소리, 감탄사 등의 비핵심어도 문장내에 포함시켜서 녹음하였다.

- [본체|팔목] {를|을} [왼쪽|오른쪽] {으로} {<각도>} {돌려} {봐} {주세요}

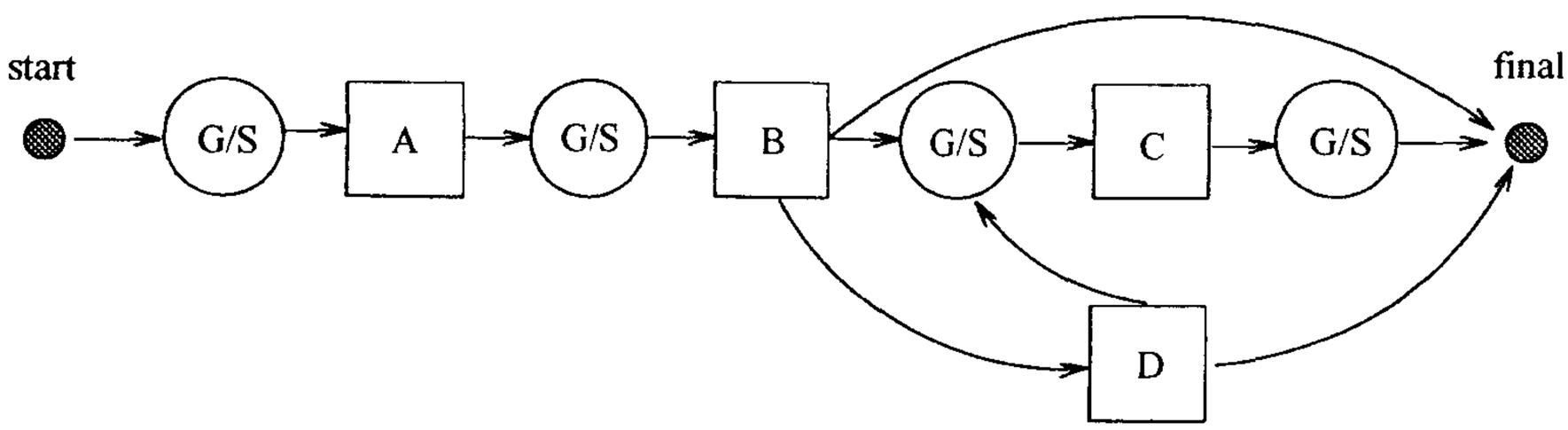


그림 4.7: 학습 데이타 채집용 개념적 문장 구성

- [어깨|팔꿈치|팔목] {를|을} [위로|아래로] {<각도>} {올려|내려} {봐} {주세요}
- 손가락 {을} [벌려|열어|닫아] {봐} {주세요}
- 취소
- [전|원래] 위치 {로} {가}
- 집어
- 놓아
- 조금더
- 커피 {한잔} [따라|따르시오]
- <각도> ::= [30도|90도|180도]

학습과 인식실험을 위해서 시내전화를 대상으로 40명분의 데이타를 수집하였으며, 이중 20명의 데이타를 학습에, 20명분의 데이타는 실험용으로 활용하였다. 수집된 음성 데이타는 본 연구에서 개발된 음성 신호 처리 도구의 그래픽 기능을 이용, 수동으로 음성구간을 추출하고 각 단어별로 학습을 수행 연속 HMM의 모수를 산정하였고 그림 4.3 의 문법 제약 조건에 따라 연결 단어 인식을 위한 FSN 을 구성하였다. 위에서 설명된 단어와 단어 사이의 천이 파라미터는 deterministic 하게 결정이 된다. 즉 천이 가능의 경우 1 을 불가능의 경우 0 을 주는 방식을 사용하

였다. 초기 단어에 대한 확률은 초기 silence 와 garbage 나 A 군에 속한 핵심단어들에게 양수로 주었다.

단어별 학습에 있어, 초기 HMM 파라미터의 추정치는 uniform 분할과 k-means clustering 알고리즘을 이용하여 구하였다. 즉, 매 학습 utterance 데이타를 해당 단어의 상태 수로 uniform 하게 분할하여 각 상태에 속하는 특징벡터들의 집합을 형성하고 k-means clustering 알고리즘으로 clustering 하여 observation 확률 밀도 함수(probability density function) 의 파라미터들(cluster 의 sample mean vector 와 covariance matrix)에 대한 초기 값을 정하였다. 초기 상태 확률은, ergodic 형 모델의 경우 단어의 모든 상태에 uniform 하게 주었으며 left-to-right 형의 경우 단어의 처음 상태에 1 을 주었다. 마찬가지로 천이 확률도 ergodic 형 모델의 경우 모든 상태에서 모든 상태로의 천이 확률을 uniform 하게 주었으며 left-to-right 형의 경우 self-transition 과 바로 다음 상태로의 천이 확률에 대해 적당한 양수를 주었다.

mixture 의 수 M 은 $2 \sim 3$ 개 주었다. 학습 데이타의 양의 부족으로 M 이 클 경우 covariance matrix 의 determinant 가 0 에 가까워지는 경우가 많았다. 이는, covariance matrix 는 diagonal한 형을 사용하였으므로, 형성 가능 mixture 의 수가 많아지면서 데이타중에 특징 벡터를 구성하는 component 가운데 거의 변하지 않는 부분이 하나의 mixture component 의 재추정에 영향을 미침으로서 발생한다고 이해된다. EM 알고리즘 학습의 반복 수는 maximum 10 번으로 했다. 대부분의 경우에 10 번 이전에 주요한 우도 점수의 수렴이 이루어졌다.

완성된 인식 시스템에 대한 인식 실험 결과 핵심어 추출의 성공도는 99%이며, 인식실험시 95%이상의 높은 인식률을 나타내고 있다. silence 및 배경 잡음과 비핵심단어들에 대해서 ergodic 형의 HMM 에 의한 학습은 다수의 단어(비핵심 단어)들을 한데 묶어서 하나의 범주를 만들어 그에 대한 전체적인 통계적 특성 – 즉 관심없는 단어들에 대한 대체적 성질– 을 형성하는 데 성공적이었다고 평가된다. 이

는 구현된 시스템에 대한 실 시간 인식 실험에서 비핵심 단어들의 사용에 대해 매우 flexible한 인식 성능을 관찰함으로써 확인된다. 중간 중간의 불규칙한 duration의 감탄사의 삽입에 대하여, 또 심지어 본 실험 문장의 시작이전이나 끝난 이후에 비 핵심 단어들로 구성된 다른 어떤 문장을 삽입하는 경우에도, 시스템의 올바른 핵심단어 인식 결과를 종종 볼 수 있었다.

인식의 출력은 역 추적에 의해 얻어진 optimal 단어 열, 즉 단어들의 index 열과 그 각 단어들의 duration 의 열이다. 예를들면 길이 119 (frame 의 갯수) 의 실험 발음에 대하여,

G	본체	G1	S1	오른쪽	G2	삼십도	S3	돌려	G4	S4
11	15	6	4	20	7	26	3	13	12	2

와 같은 결과를 얻을 수 있다.

인식 속도는 실험 문장의 발음 길이에 따라 변하지만 보통 속도의 발음에 대해 위에 설명한 하드웨어상에서 각도 부분을 생략하여 구성할 경우 2 ~ 3 초 각도 부분을 포함시킬 경우 7 ~ 13 초 걸렸다.

5 장

결 론

본 연구에서는 현재 음성인식 및 이해의 분야에서 성공적인 응용시스템이 기대되는 Voice Commander 시스템의 전반적인 기술 동향과 적용되는 신호처리 및 음성 패턴인식 알고리즘의 전반적인 고찰과 연구를 통해 소규모 어휘를 대상으로 하는 응용시스템을 개발했다.

고성능 음성인식의 가장 근간이 되는 음성패턴인식 알고리즘은 80년대 중반부터 가장 우수한 성능을 나타내고 있는 HMM을 이산 HMM과 연속 HMM의 관점에서 분석 및 구현하는 한편 이를 이용한 응용시스템을 개발하였다. 특히 전화선을 통해 응용되는 질의응답형 인터페이스는 전화선을 통한 응용이라는 점에서 많은 응용 가능성을 지니고 있다. 음성에 의해 제어되는 로보트는 원격지의 전자적 설비나 기기를 구동하는 새로운 인간-기계간 인터페이스뿐만 아니라 가정용, 산업용에 널리 쓰일 수 있는 가능성을 지닌다.

신호처리 및 특성추출을 위해서는 일반적으로 사용되는 LPC 분석과 이를 이용한 켑스트럼 추출을 통한 특징벡터를 구성하여 사용했는데, 이산 HMM을 위해서는 벡터 양자화에 의한 다중코드북을 적용해서 이산 HMM에 유입되는 입력 관

찰 벡터의 집합을 계산하였으며, 이산 HMM을 위해서는 특징벡터를 그대로 적용하였다.

또한 본 연구에서는 핵심어 추출기법을 적용하여 사용자의 입력음성중에서 인식에 필요한 핵심어를 추출하고 비 핵심어를 배제하는 방법을 이용해서 로보트 제어용 음성인터페이스를 구축하였다.

일반적으로 음성인식시스템의 구현에는 위에 기술한 음성인식이나 신호처리기법외에 이를 실시간으로 구현할 수 있는 기술이 필요한데, 본 연구에서는 고성능 신호처리 칩인 Texas Instrument사의 TMS-320C31을 채용한 Elf-31 보드와, AT&T사의 DSP-32C 칩을 채용한 S-32C 보드릉 통해 실시간 응용 시스템을 구현하고 있다. 또한 PC나 웍스테이션에 공히 응용될 수 있는 일반적인 시스템의 구현 및 다수의 신호처리 보드와 신호처리보드의 디바이스 드라이버 시스템을 용통성있게 재구성하여 실시간 다중처리를 지원하는 기술도 개발하였다.

향후 본 연구는 과제종료와 동시에 "전자도우미 시스템 개발" 과제로 발전적으로 흡수 통합되었으며, 장기적인 관점의 음성인식률 제고와 대화형 음성처리 시스템이나 자동통역 등에 응용되는 음성이해 및 음성대화 시스템을 위한 핵심기술개발 과제로서 연속성을 지니는 동시에 음성인식기술의 발전에 기여할 것으로 기대된다.

참고 문헌

[1] J. B. Allen, "Cochlear Modeling", *IEEE ASSP Magazine*, pp. 3–29, Jan. 1985.

[2] J. B. Allen, "Cochlear micromechanics - A physical model of transduction", *J. Acoust. Soc. Amer.*, vol. 68, no. 6, pp. 1660–1670, 1980.

[3] L. R. Bahl, F. Jelinek and R. Mercer (1983), "A maximum likelihood approach to continuous speech recognition", *IEEE Trans. Pattern Anal. Machine Intell. PAMI* **5**(2): 179–190.

[4] L. R. Bahl, P. F. Brown, P. V. De Souza and R. L. Mercer (1998), "Acoustic Markov models used in the Tangora speech recognition system", *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, April, pp. 497–500.

[5] W. Bates, "The BBN/Harc Spoken Language Understanding System", *ICASSP 93*, pp. 111–114, 1993.

[6] J. Bernstein, M. Cohen, H. Murveit and M. Weintraub (1989), "Linguistic contraints in hidden Markov model based speech recognition", *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May, pp. 699–702.

[7] W. Das, "Influence of Background Noise and Microphone on the Performance of the IBM Tangora Speech Recognition System", *ICASSP 93*, pp. 71–74, 1993.

[8] A. M. Derouault (1987) Context-dependent phonetic Markov models for large vocabulary speech recognition. *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, April, pp. 360–363.

[9] M. Franzini, M. Witbrock and K. F. Lee (1989), " A Connectionist Approach to Continuous Speech Recognition", *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, May, pp. 425–428.

[10] O. Ghitza, "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition", *IEEE Trans. on Speech and Audio Processing*. vol. 2, no. 1, pp. 115 132, Jan. 1994.

[11] E. Giachin, C. H. Lee, L. R. Rabiner, A. E. Rosenberg and R. Pieraccini, "On the Use of Interword Context-Dependent Units for Word Juncture Modeling", *Computer Speech and Language*, **6**: 197–213, 1992.

[12] D. Greenwood, "A Cochlear Frequency-position Function for Several Species-29 Years Later", *IEEE J. Acoust. Soc. Amer.*, vol. 87, no. 6, pp. 2592–2605, 1990.

[13] M. Y. Hwang, H. W. Hon and K. F. Lee (1989), "Modeling between-word coarticulation in continous speech recognition", *Proc. Eurospeech*, September.

[14] F. Jelinek, "A fast sequential decoding algorithm using a stack", *IBM J Res. Develop.*, vol. 13, pp. 675–685, Nov. 1969.

[15] J. M. Kates, "A Time-Domain Digital Cochlear Filter", *IEEE Trans. on Signal Processing*, vol. 39, no. 12, pp. 2573–2592, Dec. 1991.

[16] K. F. Lee (1989) "Automatic Speech Recognition: The Development of the SPHINX System", *Kluwer Academic Publishers*, Boston.

[17] C. H. Lee, L. R. Rabiner and R. Peraccini, "Speaker Independent Continous Speech Recognition Using Continuous Density Hidden Markov Models", *Proc. NATO-ASI, Speech Recognition and Understanding: Recent Advances, Trends and Applications*, P. Laface and R. DeMori (eds.), Srpinger-Verlag, Cetraro, Italy, pp.135-163, 1992.

[18] C. H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved Acoustic Modelling for Large Vocabulary Continuous Speech Recognition", *Computer Speech and Language*, **6**: 103–127, 1992.

[19] R. P. Lippman and E. Singer, "Hybrid Neural-network/HMM Approaches to Wordspotting", *ICASSP 93*, Vol. 1, pp. 565–568, 1993.

[20] F. Mclaughlin, "ICSE ii Prompts Engineers to Reflect on Yestarday, Look to Tomorrow", *IEEE Computer*, pp. 110-112, Jul. 1989.

[21] C. S. Myers and L.R. Rabiner, "A level building dynamic time warping algorithm for connected work recognition", *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-29, pp. 284-297, Apr. 1981.

[22] H. Ney, D. Mergel, A. Noll and A. Paeseler, "A data-driven organization of the dynamic programming beam search for continous speech recognition", *ICASSP 87*, pp. 833–836, 1987.

[23] H. Ney, "Dynamic Programming Speech Recognition Using A Context-Free Grammar", *ICASSP 87*, pp. 69–72, 1987.

[24] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition", *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-32, pp. 263-271, April 1984.

[25] J. Peckham,"Speech Understanding and Dialogue Over the Telephone: An Overview of Progress in the SUNDIAL Project", *Proc. EUROSPEECH 91*, pp. 1469-1472, 1991

[26] R. Pieraccini and A. E. Rosenberg,"Automatic Generation of Phonetic Units for Continuous Speech Recognition", *ICASSP 89*, Glasgow, UK, pp. 623-626, May 1989.

[27] P. J. Price, W. Fischer, J. Bernstein and D. Pallett, "A Database for Continuous Speech Recognition in a 1000-Word Domain", *Proc. ICASSP 88, New York*, pp. 651–654, April 1988.

[28] Lawrence Rabiner and Biing-Hwang Juang (1993), *Fundamentals of Speech Recognition*, Ch 7-8, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[29] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, **77**(2): 257–286, February, 1989.

[30] L. R. Rabiner, J. G. Wilpon and B. H. Juang (1986) "A segmental k- means training procedure for connected word recognition based on whole word reference patterns", *AT & T Tech. J.*, **65**(3): 21–31.

[31] L. R. Rabiner and C. H. Lee, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", *IEEE Transactions on Acoustics, Speech and Signal processing*, **37**:1649–1658, November 1989

[32] L. R. Rabiner, C. H. Lee, R. Pieraccini and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition", *Computer Speech and Language*, **4**, January 1990.

[33] R. C. Rose, "Definition of subword acoustic units for wordspotting", *Eurospeech 92*, pp. 1049–1052, 1992.

[34] Sadaoki Furui and M. Sondhi (eds.) (1992), *Advances in Speech Signal Processing*, Ch 3, MARCEL DEKKER, INC.

[35] T. Svendsen and F. K. Soong, "On the Automatic Segmentation of Speech Signals", *Proc. ICASSP 87*, Dallas, TX, pp. 77 80, April 1987.

[36] R. Stern it et al., "Sentence Parsing with Weak Grammatical Constraints", *ICASSP 87*, pp. 380–383, 1987.

[37] R. Schwartz it et al., "The N-best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses", *ICASSP 90*, pp. 81–84, 1990.

[38] Y. Takebayashi, "Natural Language Processing for Speech Understanding and Dialogue", *Information Processing(Japanese version)*, Vol. 34, pp. 1287–1296, 1993.

[39] Y. Takebayashi *et al.*, "A Robust Speech Recognition System Using Word-spotting with Noise Immunity Learning", *ICASSP 91*, pp. 905–908, 1991.

[40] Y. Takebayashi *et al.*, "Keyword-spotting in Noisy Continuous Speech Using Word Pattern Vector Subtraction and Noise Immunity Learning", *ICASSP 92*, pp. II-85–88, 1992.

[41] M. Tomita, "An Efficient Word Lattice Parsing Algorithm for Continuous Speech Algorithm", *ICASSP 86*, pp. 1569–1572, 1986.

[42] W. Ward, "Understanding Spontaneous Speech: The Phoenix System", *ICASSP 91*, pp. 365–367, 1991.

[43] M. Weintraub *et al.*, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition", *Proc. ICASSP 89*, Glasgow, Scotland

[44] J. G. Wilpon and L. Rabiner, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. ASSP*, Vol. 38, No. 11, pp. 1870-1878, 1990.

[45] V. Zue *et al.*, "The VOYAGER Speech Understanding System: Preliminary Development and Evaluation", *ICASSP 90*, pp. 73–76, 1990.