特輯

研究報告

후지쯔에 있어서의 음성 자동인식의 현상과 장래

Automatic Speech Recognition Research at Fujitsu

나라 야스히로* 기무라 신타* 김 경 호*

(Yasuhiro Nara, Shinta Kimura, K. H. Loken-Kim)

요 약

본 논문에서는, 후지쯔의 음성 자동인식 관련 제품 개발의 역사, 현재의 상품, 그리고 앞으로의 연구 개발에 대해서 소개한다. 현재는 4,000단어로 부터 12,000단어를 인식하는 특정 화자형의 F2360, 17단어를 인식하는 불특정 화자형의 F2355 L/S를 판매하고 있으며, 앞으로의 연구 개발로는 음소 변형에 적극적으로 대처하고, 자연적인 발성을 인식하기 위한 기초 기술을 개발할 계획에 있다. 인식할 단어의 문자 표기에 음향 segment 변형 규칙을 적용하여 음향 segment network를 자동 생성하여서 입력 음성과의 조합을 행한다. 이 기초 기술을 대어휘 단어 음성 인식에 응용하기 위해서 필요한 단어 후보 선택 방식, 문절 발성을 문장 입력에 응용하기 위한 문절 후보 생성 방식과 문 검사 방식에 대해서도 서술한다.

ABSTRACTS

The history of automatic speech recognition research, and current and future speech products at Fujitsu are introduced here. The speech recognition research at Fujitsu started in 1970. Our research efforts have resulted in the production of a speaker dependent 12,000 word discrete / connected word recognizer(F2360), and a speaker independent 17 word discrete word recognizer(F2355L/S). Currently, we are working on a larger vocabulary speech recognizer, in which an input utterance will be matched with networks representing possible phonemic variations. Its application to text input is also discussed.

I. 서 론

당사(후지쯔)는 일본 만국박람회 (1970년)에 특정 화자 7단어 음성 인식 장치의 응용시스템을 공개[1]한 이래, 음성 자동 인식의 연구개발을 하여 왔다.

1980년에는 단음절 이산 발성의 음성 인식 장치를 시작(試作)[2]하였지만, 입력 속도가 느리고 발성이 부자연스러워 보급되지는 못하였다.

1983년에는 불특정 화자 17단어의 이산 단어 인식 기능을 개발하고, 전화기를 단말로 한 온라인시스템에 이미 개발된 F2353를 부가하여서 F2355로 상품화하였다. 본 제품은 그 후에도 기능을 강화하였고, 현재는 F2355L/S(다음 장 참조)로 판매하고 있다.

*富士通研究所

특성 화자의 단어 인식 장치는 256단어의 것을 1984년에 발표하였다. 본 장치의 특징은 음성 삽성 기능(PAROCOR 방식에 의한 음성 합성형)을 구비하고 음성 대화 시스템을 구성할 수 있도록 한 점이다. 본 제품을 개량하고, 가격을 약 1/10, 성능을 약 10배로 한 것이 다음 절에서 설명하는 F2360이다.

단음절 발성의 음성 인식 장치를 시작(試作)한 후, 문장 입력에는 문절로 부터 문정도의 발성 단위의 음성 인식 장치가 필요하다고 하는 것을 통감하였다. 문절 또는 문이라고 하는, 소위 연속 음성에는 각종 다양한 음소 변형이 포함된다. 예를 들면 그림 1은 단순한 음소 변형의 예로서 모음의 탈락(무성화)을 보여주고 있지만, 이것 이외에도 이음화, 중성화, 구개화를 비롯한 다양한 변형이 포함된다. 여기서 우리는 3장에 설명될 것과 같은 음소 변형을 예측하고, 음소 변형에 적극적으로 대처하는 방식을 개발하고 있다.

II. 현재의 제품

2.1 F2360(VCU)

F2360(VCU : Voice Communication Unit)은 음성에 의한 대화시스템을 구축하기 위한 장치로서 설계한 것이고, 음성 인식 기능에 더하여, 음성 축적

그림 2. F2360의 외관

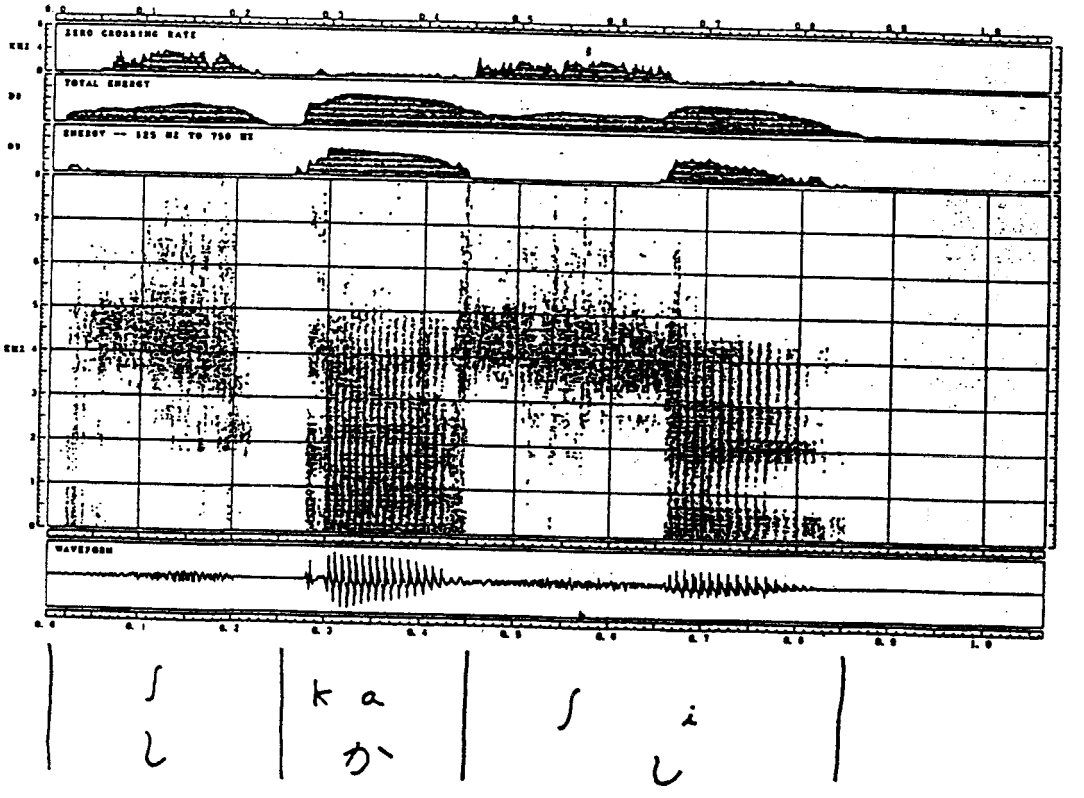


그림 1. 단순한 음소 변형의 예

세한 기능을 갖게 하였으며, Option으로 규칙 음성 합성의 기능도 제공하고 있다. 음성 인식으로는 1,000단어 (option으로 12,000단어)의 인식이 가능하고, 이산 발성 mode와 연속 발성 mode를 갖추고 있다. 장치 외관을 그림 2에 나타내었다.

본 장치는 GPIB interface를 경유하여서 personal computer등의 host computer와 접속하게 되어 있고, 음성 template와 축적 음성 data의 고속 전송이 가능하 것 외에 1개의 interface port에 15대까지의 VCU를 접속하는 것이 가능하다. 이러한 특징과 무선 기구나 회선 제어장치등의 option에 의해 다양한 application 시스템을 구성할 수 있도록 되어 있다 (그림 3).

2.2. F2355L / S[3]

F2355L / S는 공중전화망과 host computer 사이에서, host내의 음성처리 support package의 제어에 의해 dial 전화기와 pushphone을 단말로 하는 대화처리를 할 수 있고, 음성 인식 기능은 불특정 화자, 17단어의 이산 단어 인식이며, 0~9의 열개의 숫자와 「예」, 「아니오」 등 7종의 제어어를 인식할 수 있다. F2355는 host interface가 두 개이며, 최대 32개의 전화회선을 support하고, F2355는 host interface가 한 개이며, 최대 8회선을 support하는 model이다.

Ⅲ. 앞으로의 연구개발

3.1. 음소 변형에 대처한 음성 인식 방식[5]~[12]

제1장에서 서술한 것과 같은 음소 변형에 적극적으로 대처하기 위해, 음소 변형 현상을 생성 음운론으로 사용하고 있는 문맥 의존형 규칙의 형식을 사용하여서 표현하는 것을 시도하였다[6][7]. 1,000단어×10명의 음성 data에 포함되는 음소 변형을 317개의 규칙으로 기술하여, 36명(여성 6명을 포함)의

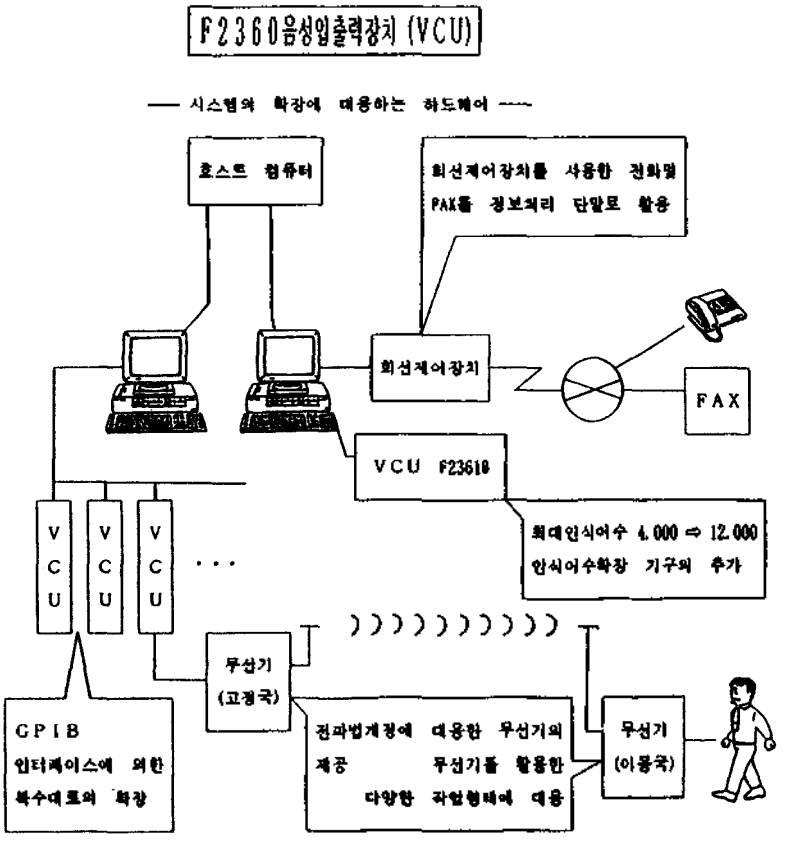


그림 3. F2360에 의한 시스템 구성 예

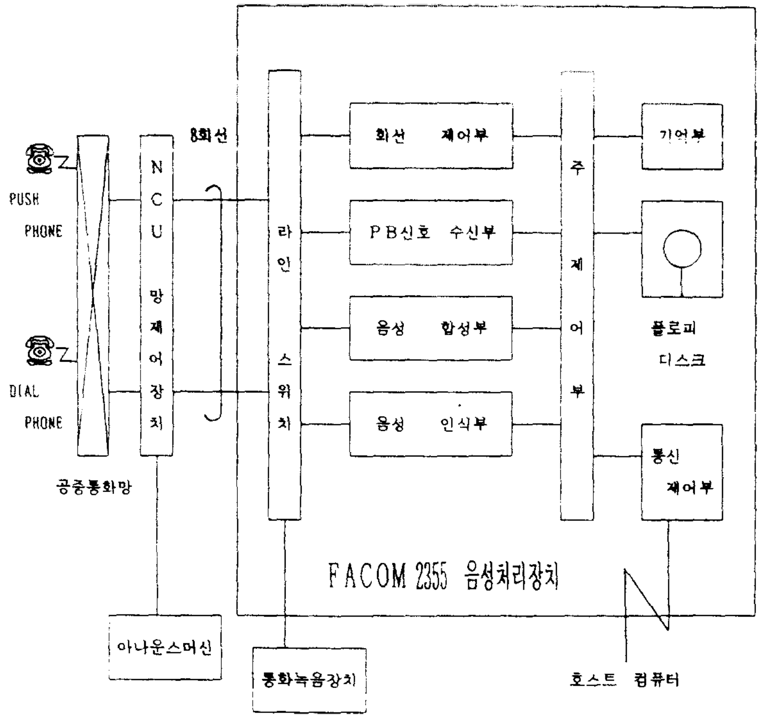


그림 4. F2355L · S의 구성

발음의 이음

N' → M' / (M, B, P)

모음의 장음화

E. I. → (E. I. . E 2)

그림 6. 음소 변형 규칙의 기술법

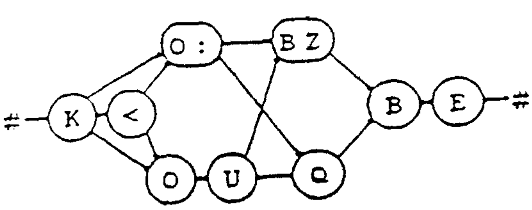


그림 7. 음향 segment · network의 예

음성 data로 이것을 평가한 결과, 평가 data에 출현하는 음소 변형중 97.6%를 cover할 수 있는 것을 확인하였다[8].

다음에, 이들 규칙을 음성 인식에 이용하기 위한 방법을 고안하였다 인식할 단어의 가나(かな) 표기를 음소 변형을 포함하지 않는 발성의 음향 segment 계열로 변환한 뒤, 음소 변형 규칙을 적용하고, 가능한 변형을 추가하여서 network로 하였다[11][12]. 인식할 전 단어의 network를 준비하여, 입력 음성과 network 와의 matching 및 거리계산을 하고, 최소 거리의 category를 인식 결과로 한다. matching으로는 음향 segment길이 규칙을 병용한다. 이것은 network의각 node인 음향 segment의 지속 시간의 평균치와 표준

[illegible] 어두(語頭), 어중(語中), 어말(語末)로 구분하여서 기술할 것이다.

3.2. 음소 상세 식별[13][20]

3.2.1. 음성 현상 동기(同期) 방식

음성 인식을 위한 음성 분석의 종래 방식으로서는 음성 현상과는 무관계한 등간격의 분석주기(예를 들면 5ms)로 분석하였다. 따라서, 예를 들면 파열자음의 파열 시점과 분석창과의 위치 관계가 동기하지 않고, 분석창의 중심으로 되기도 하고 가장자리가 되기도 하는 것이 불안정하게 일어났다(그림 8). 구체적인 예로서는, 녹음 tape에 녹음된 음성을 되풀이 인식하면 옳게 인식하기도 하고 잘못 인식하기도 하는 것이 불규칙하게 발생하였다.

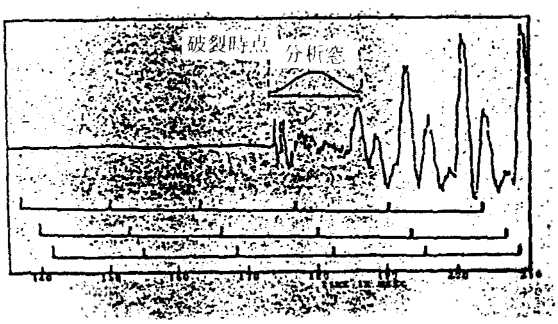


그림 8. 종래의 비동기 분석실

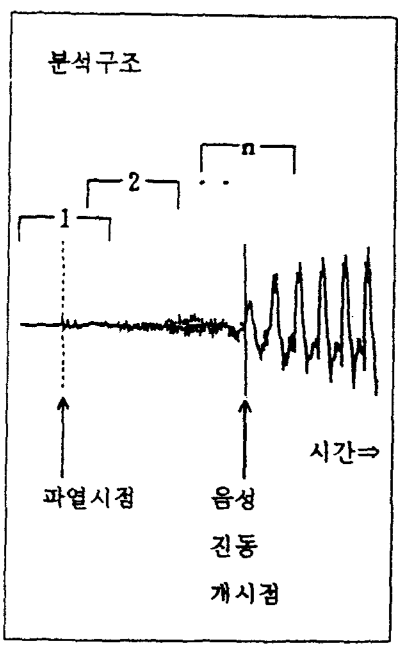


그림 9. 자음부 시(時) 계열의 특징 추출법

우리는 음성 인식 중에서도 특별히 정밀도가 요구되는 유사 자음간의 식별에는 음성 현상에 동기한 분석을 하여야 한다고 생각하여 파열자음[13]-[19], 비(鼻) 자음[20], 무성 마찰 자음[21]의 각 자음군 내의 식별에 대해서 음성 현상 동기분석법을 연구하여 왔다. 그림 9에 파열 자음의 경우의 분석창 설정 예를 나타내었다. 파열시점과 성대 진동 개시점의 사이를 (n−1) 등분하고, 각각 등분점이 중심이 되도록 하여 n개(예를 들면 4개)의 분석창을 설정하였다.

3.2.2. Neural network의 이용

자음의 식별로 neural network(이하, 뉴럴넷)의 이용을 검토하였다[21]. 앞절에서 서술한 동기 분석의 결과를 주성분 분석에 의해 차원 압축한 것을 parameter로 하고, 각종 식별 수법과의 비교를 하였다. 그 일부를 그림 10에 나타내었다. Multi-template법 I(이하 MT I)는 미지 pattern근방의 K개의 template중에서 가장 많이 속한 class를 선택하는 방법이고, multi template법 II(이하 MT II)는 class별 K개의 근방 template와의 평균 거리에 의한 것이다. 뉴럴넷 I는 hidden layer 한층과 12소자, 초기 weight로서 난수(亂數)를 사용한 것이다. 선형 판별함수법은 뉴럴넷보다도 높은 식별율을 나타내고,

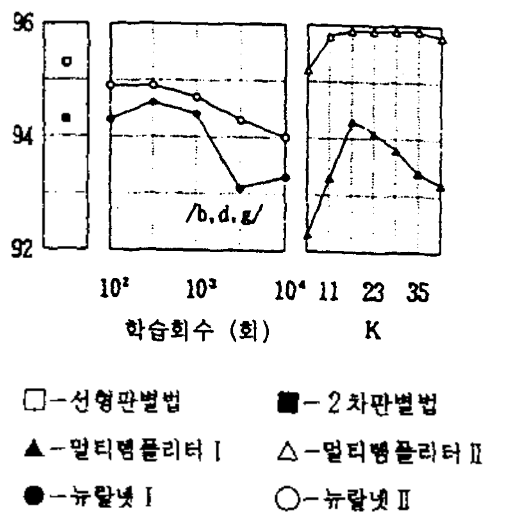
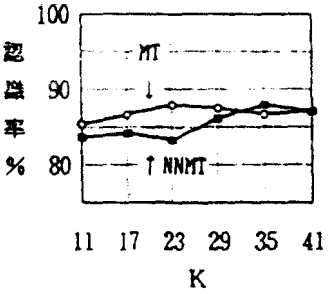


그림 10. 각 수법의 식별율



동시에 뉴랄넷과 동일하게 내적(內積) 연산으로 구성되는 것에 주목하고, 선형 판별 함수의 계수를 뉴랄넷과 초기치로 하는 방법을 고안하였다(그림 중의 뉴랄넷Ⅱ). 뉴랄넷의 식별율이 낮은 것은 학습에 사용한 data중에 고립한 algorithm으로도 식별 곤란한 data가 있어서 그것을 억지로 학습하려고 하기 때문에 판별면이 복잡하게 되어 버렸다고 생각하였다. 여기서, 계산량이 많지만 정밀도가 높은 MTⅡ의 인식 결과를 뉴랄넷에 학습시키는 시도(NNMT)를 하였다[22]. 그 결과, 적절한 K를 선택하면 NNMT는 MTⅡ 이상의 식별율을 나타내는 것이 판명되었다.

3.3. 대어휘 단어 음성 인식으로의 전개[23] [24].

소수 단어의 음성 인식에서는 복수 등록등으로 어느 정도의 음소 변형에 대처하는 것이 가능하다. 3.1 절에서 서술한 음소 변형에 대처한 음성 인식 방식(그림 2)은 전 단어를 등록할 수 없는 대어휘 단어의 음성 인식에서 위력을 발휘하는 것이지만, 이 방식은 매우 복잡한 처리를 필요로 하기 때문에 처리 시간에 문제가 있다.

따라서, 분소합(本照合)을 하기에 앞서 비교적 처리량이 적은 방법으로 후보를 선택하는 것을 검토하였다. 음향 처리부에서 출력되는 frame주기 마다의 분석 결과를 8frame씩 정리하여서 압축함과 동시에, 음향 segment 계속 시간길이법칙을 사용하지 않는 것에 의해서 처리량을 1/80로 절감하였다. 10만 단어를 이 방식으로, 1,000단어로 선택하는 경우에 대해서 평가한 결과, 탈락율은 0.3%였다.

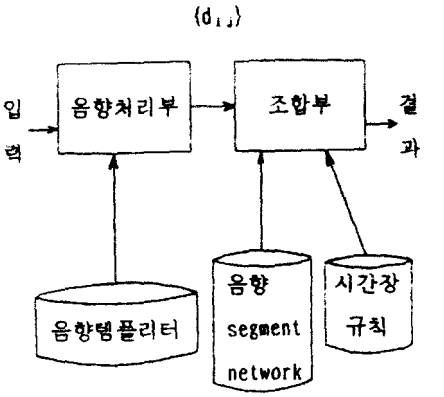


그림 12. 음향 segment · network를 사용한 음성 인식

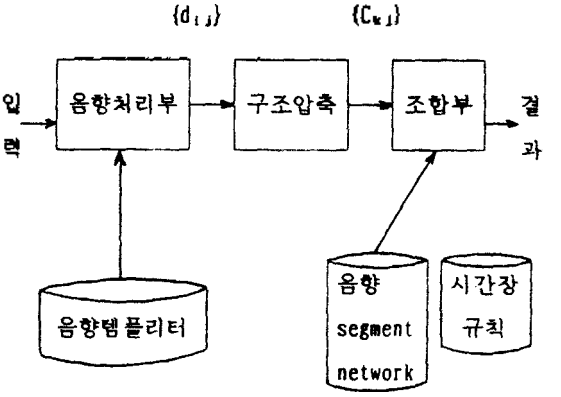


그림 13. 단어 후보 선택 방식

3.4. 문장 입력으로의 전개[25]~[27].

일본어 문장의 음성 입력을 고려할 경우, 제1장에서 서술한 것 같이 단음절 발성으로는 사용하기 어려울 뿐만이니라 단이 발성 또한 비현실적이다. 이것은 단어가 복잡하게 변화되어서 집속하기도 하고, 극단으로 짧은 조사등의 단어가 존재하기 때문이다. 예를 들면, 「감사 합습니다」라고 하는 문은 「감+지+않+습+니+다」로 분할하여서 발성하지 않으면 안된다. 문 단위의 발성이 이상적이지만, 현재의 기술로 다룰 수 있는 어휘수는 1,000단어 정도이다. 음성으로 일반 문장을 입력하는데 10만단어 이상의 어휘를 대상으로 하지 않으면 안되기 때문에, 우리는 문절 단위의 발성을 당면의 target으로 생각하고 있다.

각 문절을 각각의 단어, 예를 들면 「나는」과 「내

入力	あなたは	本を	読んでいますか
テ	あなたと(2058)	本を(2237)	住んでいますか(2407)
テ	あなたは(2497)	本も(2636)	読んでいますか(2706)
ィ	あなたがたは(2582)	本の(2644)	読んでいるのですか(2779)
ス	あなたが(2706)	本と(3100)	泳いでいますか(2847)
	花子は(2731)	公園の(3176)	読んでください(2938)
	山田は (2753)	本が(3180)	読んでもらえますか(2975)
	あなたがた(2804)	ボールを(3267)	乗ってもらえますか(3012)
	中は(2821)	物も(3314)	答えていますか(3042)

그림 15. 문절 라티스의 예

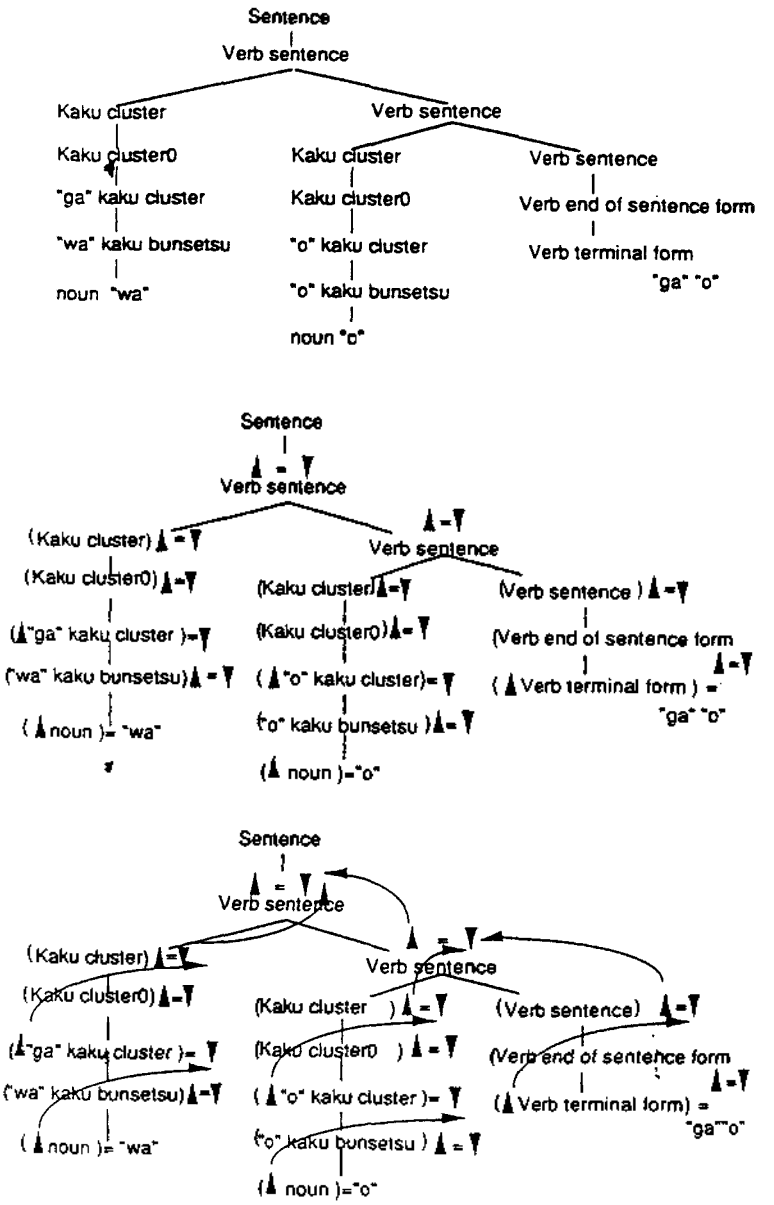


그림 16. C구조

가」를 서로 전부 관련이 없는 별개의 단어라고 고려한다면, 대어휘 단어 음성 입력 시스템에 의한 실현이 가능하다. 대략 10만어 어휘로부터 20억의 문절의 파생이 가능하게 된다. 여기서, 우리는 무절 음성인식의 제일보는 발성으로부터 개괄적인 특징을 추출하고, 문절 문법에 비춰서 수천 정도의 문절후보를 생성하는 것이라고 생각하고 있다. 약 46,000어휘로 비교적 단순한 문절 문법 model을 사용하고, 문절 발성으로부터 VCV(모음－자음－모음 연쇄)를 추출한 경우에 문절 후보가 어느 정도 삭감될 수 있는가를 검토하였다[25]. 그 결과, 하나의 VCV가 정확하게 추출될 수 있으면 문절 후보의 수는 약 1/10이 되는 것이 판명되었다. 실제로는 복수의 VCV를 추출할 수 있고, 그것들의 순서 관계도 이용할 수 있는 긍정적인 면과, VCV 추출에는 애매함이 있으므로 복수 후보를 고려하지 않으면 안된다고 하는 부정적인 면을 생각할 필요가 있다.

다음에, 삭감한 문절 후보를 단어라고 간주하고, 앞절에 서술한 대어휘 음성 인식 방식으로 인식하게 된다. 대어휘 음성 인식 방식은 높은 정밀도를 갖고 있지만, 인식 결과를 언제나 하나의 의미로 결정할 수 있다고는 생각하기 어렵다. 거기서 그림 15에 나타내는 것과 같은 문절 라티스로부터 문으로서, 문법적 또는 의미적으로 옳은 것을 선택하는 것과 같은 언어 처리가 필요하게 된다.

라티스 중에서 차례로 후보 문을 생성하고, 각 후보 문에 대해서 언어 처리에 의한 check를 한다. 후보문 생성은 depth first search를 사용하는 것이 가장 단순하지만, 정해(正解)의 가능성이 높은 것을 가능한한 먼저 생성하는데는 인식 score가 근접한 후보들을 그룹화하여[26] best first search를 하는 것이 효과적이다.

현재는 문법 처리를 검토하고, 또한 의미 처리에의 제일보로서 표층격의 검사를 검토하고 있다. 표층격이라는 것은, 예를 들면 「讀む」라고 하는 동사가 「～が, ～を, ～に, ～で」라고 하는 격을 지배할 수 있다고 한 것이다. 우리는 언어 처리의 당면의 과제로 LFG(Lexical Functional Grammar)를 고려하고 있다. 단문을 대상으로 한 단순한 문법을 BNF 기법으로 정의하고, 구문 해석한 결과를 그림 16에 나타내었다. Parser가 문법에 의해 해석한 결과를 먼저 phrase structure(LFG로는 C구조)로 나타낸다.

다음에 이 결과에 따라서 정보를 전파시키면 그림 17에 나타내는 것과 같은 F구조를 얻을 수 있다. 이 예로는 「が格, を格, に格, で格」을 지배할 수 있는 동사에 대해서, 실제로 「が格」과 「を格」가 나타난 것을 표시하고 있고, 옳다고 판단하지만 지배할 수 있는 격 이외의 격이 나타나면 옳지 않다고 판단한다. 금후는 이것을 술어가 지배할 수 있는 어(語)의 의미 속성을 검사할 수 있도록 확장할 필요가 있다.

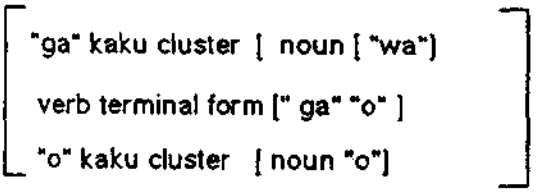


그림 17. F구조

인사말

본 논문은 1990년 5월, 동북대학에서 개최된 「음성 인식의 현상과 장래」에 관한 심포지움에 발표한 것을 ATLAS-JK 일한기계 번역시스템을 사용하여 번역한 것이다. 번역에 힘이 되어 주신 한국후지쯔의 김병원과장님, 김동욱씨께 감사드린다.

參 考 文 獻

1. 에다가와(枝川), 이또오(伊勝) 등 : 음성정보처리의 approach, FUJITSU Vol. 21, No. 3, pp. 401～419, 1970.
2. 시라도리(白鳥) : 음성 처리의 실적, 사무와 경영, 1980년 10월호, pp. 54～56, 1980.
3. 도미따(富田), 나카무라(中村) 등 : 음성 처리장치, FUJITSU Vol. 32, No. 3, pp. 419～425, 1983
4. FACOM 2370 음성입력시스템, FUJITSU Vol. 35, No. 5, pp. 581～588, 1984.
5. M. Yoshida, et al : Technology for Pattern Information Processing, FUJITSU Scientific and Technical Journal Vol. 25, No. 2, pp. 81～112, 1989.
6. 기무라(木村), 나라(奈良) : topdown 음소 segmentation

[illegible] 음학(音學)강론, 1985년 [illegible] 1985.
[illegible] segmentation에 [illegible] 음학(音學)강론, 1985년 9월~10월, 3-1-11, 1985.
8. 기무라(木村) : [illegible] 시스템, 음학(音學)강론, 1986년 3월, 2-1-11, 1986.
9. 기무라(木村) : [illegible] 화자의 음소 패턴에 관한 검토, 음학(音學)강론, 1986년 10월, 3-3-1, 1986.
10. 기무라(木村) : 음소 segment network에 의한 단어 음성 인식, 음학(音學)강론, 1987년 10월, 1-5-21, 1987.
11. 기무라(木村) : 음성 중의 음향 segment의 [illegible] 변동 [illegible], 음학(音學)강론, 1988년 3월, 1-2-11, 1988.
12. 기무라(木村), 사나다(眞田) 등 : 음향 segment network를 사용한 단어 음성 인식에 있어서의 규칙의 평가, 음학(音學)강론, 1988년 10월, 1-3-20, 1988.
13. 고바야시(小林), 이와미다(巖見田) 등 : 유성 파열 자음 식별에 관한 검토, 음학(音學)강론, 1985년 9월~10월, 2-1-11, 1985.
14. 고바야시(小林), 이와미다(巖見田) 등 : 유성 파열 자음 식별에 있어서의 과도부 특징 추출법의 검토, 음학(音學)강론, 1986년 3월, 2-1-1, 1986.
15. 이와미다(巖見田), 고바야시(小林)등 : 무성 파열 자음 식별에 관한 검토, 음학(音學)강론, 1986년 3월, 2-1-2, 1986.
16. 이와미다(巖見田), 나라(奈良) : 무성 파열 자음 식별에 있어서의 시계열 특징 추출법의 검토, 음학(音學)강론, 1986년 10월, 1-3-8, 1986.
17. 이와미다(巖見田), 나라(奈良) : 파열 자음 식별에 있어서의 파열 시점의 자동검출, 음학(音學)강론, 1987년 3월, 2-5-18, 1987.
18. 이와미다(巖見田), 사나다(眞田) 등 : 연속 음성 중의 파열 자음 식별의 검토, 음학(音學)강론, 1988년 10월, 1-3-2, 1988.
19. 고지마(小島), 기무라(小村) : 불특정 화자 무성 마찰 자음의 식별, 신학전(信學全) 1990년 봄, A-227, 1990.
20. 이와미다(巖見田) : 개방점 동기 분석에 의한 비(鼻) 자음 식별의 검토, 음학(音學)강론, 1988년 3월, 1-2-9, 1988.
21. 사나다(眞田), 이와미다(巖見田) 등 : 불특정 화자 자음 식별에 있어서의 neural network와 각종 수법의 비교, 음학(音學)강론, 1988년 10월, 2-P-13, 1988.
22. 사나다(眞田), 기무라(木村) : multi template법의 식별기구를 학습한 neural network, 음학(音學)강론, 1989년 10월, 1-1-24, 1989.
23. 기무라(木村) : 음향 segment network에 의한 단어 후보 선택, 음학(音學)강론, 1989년 10월, 2-P-7, 1989.
24. 야마사끼(山崎), 기무라(木村) : 음향 segment network 를 사용한 10만 단어 인식, 음학(音學)강론, 1990년 3월, 1-3-24, 1990.
25. 다나까(田中), 나라(奈良) : VCV에 의한 문절 후보 삭감 효과의 검토, 음학(音學)강론, 1989년 3월, 3-6-1, 1989.
26. Loken-Kim, Nara, et al, : A POST-PROCESSOR FOR A LARGE VOCABULARY JAPANESE SPEECH RECOGNITION SYSTEM, Eurospeech 1989, Vol. 2, pp. 1~4, 1989.
27. Loken-Kim, Nara, et al, : LANGUAGE PROCESSING FOR A LARGE VOCABULARY ISOLATED SPEECH RECOGNITION SYSTEM, 정처전(情處全), 1989년 후기, 7-E-8, 1989.

筆者紹介

▲나라 야스히로

1975년 : 게이오대학 전기공학과 졸업(공학사)
1977년 : 게이오대학 전기공학과 (공학석사)
1977년~현재 : 후지쯔 연구소

▲ 기무라 산다

1978년 : 고오베대학 전기공학과 졸업(공학사)
1980년 : 고오베대학 전자공학과 (공학석사)
1980년~현재 : 후지쯔 연구소

▲김 경 호

1977년 : 산양대학 금속공학과 졸업(공학사)
1982년 : 노스케롤라이나대학 산업공학과(공학석사)
1988년 : 노스케롤라이나대학 산업공학과(공학박사)
1988년～현재 : 후지쯔 연구소