電氣學會 論文誌 37卷 4號 1988年 4月

論 文
37~4~8

Walsh변환을 이용한 한국어 숫자음 음성분석에 관한 연구

A Study on Korean Speech Analysis using Walsh Transform

金 桂 賢*·金 俊 炫**
(Gye-Hyun Kim · Jun-Hyun Kim)

요 약

본 논문은 음성 분석에 대한 연구로서, Walsh 변환을 이용하여 여러 화자에 의해 발음된 한국어 숫자음('일' - '십')에 대한 음성 분석을 행하였다. 기존의 음성 분석 방법들(선형 예측법, Fourier 변환)에 비해, Walsh변환은 덧셈과 뺄셈만으로 수행되므로 계산속도가 빠르고, 변환 과정에서 메모리가 적게 든다. 음성분석에 있어, 동일 단어에 대한 화자들 간의 차이에 좌우되지 않는 특징을 구하는 것이 문제가 되는데, Walsh 변환 방법으로 이러한 패턴을 얻을 수 있는가를 조사해 보았다. 실험결과, 여러 화자에 의해 발음된 동일 단어(한국어 숫자음 '이') 패턴들 중 70% 정도가 유사한 패턴을 보였다.

Abstract-This work describes a speech analysis of Korean number('1'~'10') which are spoken by several speakers using Fast Walsh Transform(FWHT) method. FWHT includes only addition and subtraction operations, therefore faster and needs less memory than FFT(Fast Fourier Transfirm) or LPC(Linear Predictive Coding) analysis method. We have investigated that FWHT method can find speaker independent feature(which represents same cue about some word independent of different speakers) The results of this experiment, the 70% of same words(korean number '2') which spoken by several speakers have had slmilar patterns.

1. 서 론

본 논문은 음성 인식의 첫 단계인 음성 분석 단계에 대한 연구로서, Walsh 변환을 이용하여 한국어 단음절 숫자음 '일'부터 '십'까지를 분석하였고, 특히 한 단어를 택하여(한국어 숫자음 '이'), 여러 화자에 의해 발음된 이러 패턴들 중에서 동일한 패턴이 어느 정도인가를 조사하여, Walsh변환 방법으로, 동일 단어에 대한 화자들 간의 차이에 좌우되지 않는 특징을 구할 수 있는가를 조사해보았다.

기계가 음성을 인식하는 일련의 과정을 '자동 음성 인식'이라 하며, 음성 인식은 다음과 같은 장점이 있다.[1]

첫째, 사용자는 기계 앞에 있지 않아도 되며, 이동하면서 음성으로 기계를 제어할 수 있다.

둘째, 사용자는 손으로 더 중요한 일을 하면서, 말로 기계에 음성 명령을 줄 수 있다.

세째, 음성 명령은 key 입력보다 2배 이상 빠르게 입력된다.

음성 인식은 인식하려는 단어의 수에 따라, 고립단어 인식과 연속단어 인식으로 나누며, 특정 화자에 의해 발음된 음성만 인식하느냐 아니냐에 따라

*正 會 員 : 서울大 工大 電子計算機工學科 博士課程
**正 會 員 : 全南大 工大 電子工學科 敎授 · 工博
接受日字 : 1987年 11月 23日
1次修正 : 1988年 3月 28日

화자 종속 인식과 화자 독립 인식으로 나눈다. 음성을 인식하기 위해서는 먼저 음성을 분석해야 하는데, 음성분석은 음성의 특성을 나타내는 특징 계수를 추출함으로써 수행된다. 특징계수 추출방법은 에너지와 영교차율과 같은 간단한 것에서부터 단시간 스펙트럼, LPC, homomorphic processing 방법 등이 있다. 특징 계수 추출방법을 선택할 때, 계산시간, 필요한 메모리 양, 구현의 용이성 등이 고려되어야 한다. 음성인식에서 가장 많이 사용하는 특성 계수의 추출방법은 FFT와 LPC방법이다. 본 논문에서는 256ms 전구간의 음성에 대해 Walsh 변환을 행하여 특징계수를 추출하였고 Walsh 파우어 스펙트럼의 에너지 분포를 통해 단어들의 차이점과 공통점을 조사했다.

한국어 숫자('일' - '십')의 파우어 스펙트럼과 특히 6명의 화자에 의해 발음된 한국어 숫자음 '이'의 파우어 스펙트럼의 공통점을 보여 Walsh변환을 이용하여 음성 분석을 했을 때 어느 정도 화자 독립 고립 단어 인식을 할 수 있는가를 조사했다.

본 논문에서는 인식과정까지는 구현하지 않았다. 그러나 인식 과정인 패턴의 유사도 결정 방법으로써 여러가지가 있는데 DTW(dynamic time warping) VQ(vector quantization) hidden markov model 등이 있는데 이중에 VQ에 의한 인식 시스템은 메모리양과 계산량을 DTW나 hidden markov model보다 줄일 수 있으므로 VQ방법에 의한 distortion measure(유사도 측정) 방법을 제안한다.

Walsh변환은 변환 과정중 메모리가 적게들며 덧셈과 뺄셈만으로 수행되므로 수행 시간이 빠르다. 대부분의 음성 인식 시스템에서는 Fourier변환 또는 선형 예측법 등을 사용하는데, 이 방법들은 좋은 인식율을 얻기는 하나 특징 파라미터가 복잡하기 때문에 계산 시간이 많이들고, 기억 장소도 많이 차지한다.[2),5)] 따라서 Walsh변환은 다른 음성분석들에 비하여 수행시간과 메모리가 적게 드는 장점을 갖고 있다.

예를 들면, FWHT(Fast Walsh Transform) 는 N개의 음성 샘플에 대해 N logN회의 덧셈을 수행하나 FFT(Fast Fourier Transform)의 경우 N logN회의 덧셈과 N logN회의 곱셈을 수행한다.[7)] 이것은 N의 값이 클 경우 매우 큰 계산량의 차이를 내므로 FWHT가 훨씬 적은 계산량과 적은 시간이 소요된다. 따라서 메모리의 용량과 계산시간이 FWHT의 경우가 FFT보다 훨씬 적게 든다.

2. Walsh변환

Walsh변환식은 다음과 같다.[6)]

다음은 임의의 음성 샘플 f(k)에 대한 Wlash 변환식이다.

$$F(j)=\frac{1}{N}\sum_{k=0}^{N-1} f(k)\cdot Wal(j,k) \qquad (1)$$

F(j) : 표준화된 j번째 Walsh계수
f(k) : 음성 신호의 k번째 샘플
Wal(j, k) : j번째 불연속 Walsh함수

Hadamard 변환은 Wal(j, k) (Walsh 순서 불연속 Walsh함수)대신 Had(j, k) (Hadamard 순서 불연속 Walsh함수)를 식(1)에 넣어 수행한다. 함수가 나타나는 순서에 따라 Walsh 순서 Walsh함수, Hadamard순서 Walsh함수라 한다. 본 논문에서는 계산상의 편의를 위하여 Hadamard 변환을 행한 후 계수들을 Walsh순서(영교차 순서)로 다시 배열하였다.

다음은 8개의 음성 샘플에 대한 Hadamard 변환의 예를 보이는 신호 흐름도이다.

그런데, 불연속 Walsh 함수의 차수는 2^n(n= 1, 2, 3, ……)밖에 존재하지 않으므로 본 논문에서는 단음절 한국어 숫자음을 처리하기에 적당한 2^{11} (=2048) 차수의 Walsh 함수와 2048개의 음성 샘플

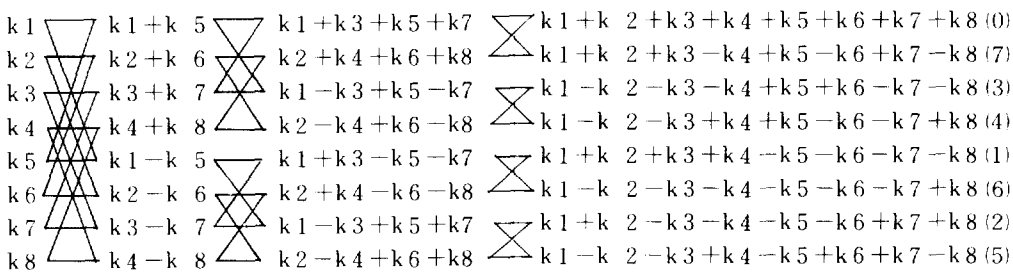


그림 1. 8개 음성 샘플에 대한 신호 흐름도

Fig. 1. Signal slow diagram of 8-point data sample

로써, 식(1)를 수행하였다. 2048개의 음성 샘플이란, 125μsec간격으로 아날로그 신호를 샘플링했을 때, 0.256초 동안의 음성 데이타이다. 그리고 단음절 한국어 숫자음('일' - '십')을 발음하는데, 약0.3초 걸리도록 음성을 입력시켰다. 사람에 따라 발음하는 속도가 다르므로, 그보다 짧거나 길게 발음하는 경우가 다르나, 음성을 입력시키기 전, 화자를 3~4회 미리 발음시켜 음성을 입력할 때에는 단음절 한국어 숫자음('일' - '십')이 0.3초 정도 걸리도록, 보통의 발음 속도로 발음시켰다.

Walsh파우어는 같은 영교차 수를 갖는 Walsh함수 계수의 제곱의 합의 제곱근이다.

$P(0) = \sqrt{\{F(0)\}^2}$

$P(i) = \sqrt{\{F(2j-1)\}^2 + \{F(2j)\}^2}$ $j = 1, \cdots, 2^{n-1}-1$

$P(2^{n-1}) = \sqrt{\{F(2^n-1)\}^2}$

P(i) : i번째 Walsh파우어 값

3. 음성 분석 시스템

본 논문에서 설계 및 구현한 시스템의 하아드 웨어 구성과 소프트웨어 구성은 다음과 같다.

3.1 하아드웨어 구성

하아드웨어 구성의 전체 블럭도는 그림2와 같다.

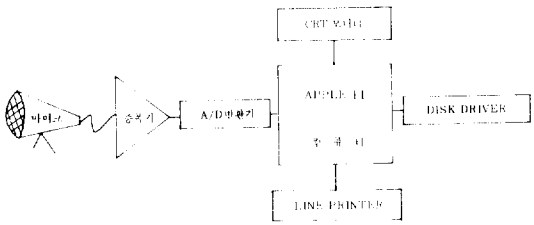


그림2. 음성 분석 시스템의 하아드 웨어 구성

Fig. 2. Hard were block dagram of speech analysis system

중심이 되는 구성 요소는 APPLE Ⅱ 컴퓨터와 A/D변환기이다. APPLE Ⅱ 키보오드의 return key를 누르면서 거의 동시에 마이크를 통해 음성을 입력시킨다. 이 아날로그 음성 신호는 증폭기에 의해 확대되어 A/D변환기로 입력된다. 본 시스템에서는 증폭기의 역할은 cassette volume이 하도록 하였다.

A/D 변환기에 입력된 아날로그 음성 신호는 A/D 변환기와 APPLE Ⅱ 메모리에 저장되어 있는 변환 프로그램에 의해 샘플링되어(125sec간격)4096개(0.5초 동안의 데이타)의 디지탈 음성 데이타(음성 데이타가 아닌 부분도 포함)가 지정된 메모리번지에 저장된다. 지정된 메모리를 CRT 화면을 통해 조사하여 음성이 입력되기 시작한 부분을 찾아 그곳에서 부터 2048개의 음성 데이타만을 diskette에 저장시킨다.

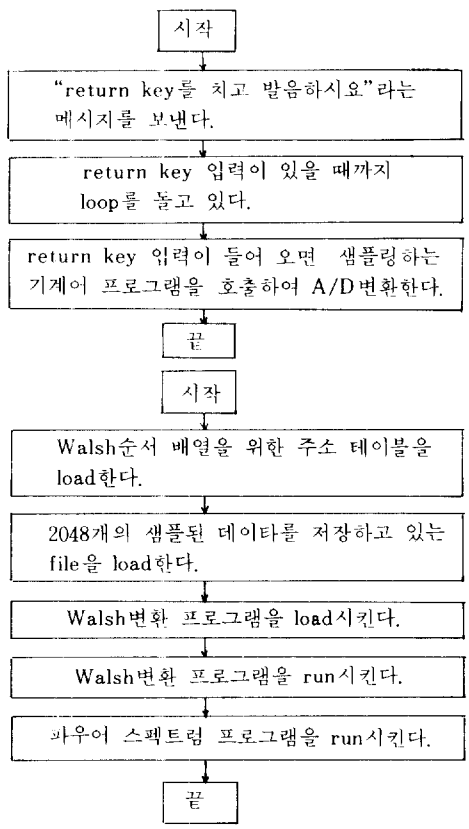
distette으로부터 2048개의 음성데이타와 변환 프로그램, 파우어 스펙트럼 프로그램을 load하여 수행시켜, 표준화된 파우어 스펙트럼(normalized power spectrum)을 CRT 모니터와 Line printer에 출력시킨다.

위의 과정을 수행하는 소프트웨어 구성은 다음과 같다.

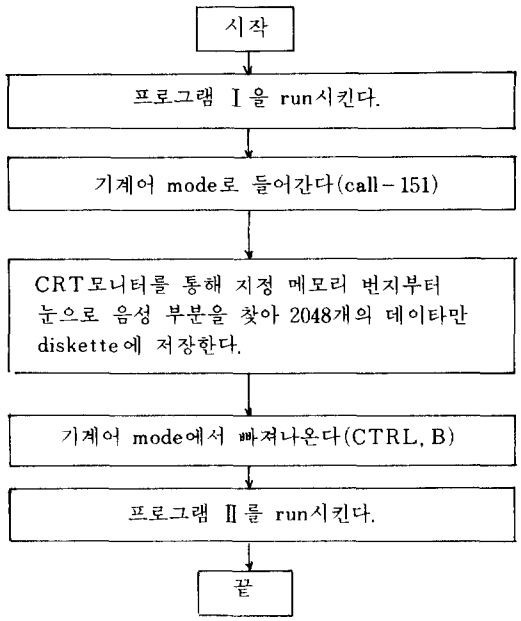
3.2 소프트웨어 구성

프로그램은 크게 두 부분으로 구성이 되는데, 샘플링하는 기계어 서브루우틴을 포함하고 있는 BASIC 프로그램Ⅰ과 샘플된 디지탈 음성 데이타를 Walsh변환하고 Walsh 파우어 스펙트럼을 구하는 BASIC 프로그램Ⅱ로 구성되어 있다.

다음은 프로그램 Ⅰ과 Ⅱ의 흐름도이다.



프로그램 Ⅰ과 프로그램 Ⅱ를 수행시키는 과정은 다음과 같다.



4. 실험결과

음성 분석은 스펙트럼의 에너지 분포로써 한다.

'일'부터 '십'까지 단어들의 대표적인 패턴은

'일'은 0, 3, 5와 27, 29, 32부분에,

'이'는 0, 2, 6과 26, 30, 32부분에,

'삼'은 0, 2와 30, 32부분에,

'사'는 전체적으로,

'오'는 0, 2, 4와 28, 30, 32부분 특히 양옆 1/8 선에,

'육'은 0, 2와 30, 32부분에,

'칠'은 전체적으로 특히 양옆 1/16선 바깥 부분에,

'팔'은 전체적으로 특히 양옆 1/16선 바깥 부분에,

'구'는 0, 2, 4과 26, 28, 30, 32부분 특히 2와 30에,

'십'은 0과 32부분에, 가운데 부분은 특히 낮은 분포를 보인다.

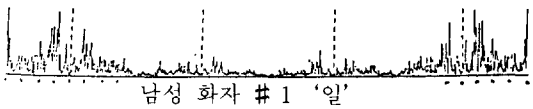
분석을 위해 스펙트럼의 x축을 32구간으로 나누었다.

각 단어는 위와 같은 위치에서 에너지 집중(energy concentration)을 보인다.

한국어 숫자음 '이'에 대한 14개 데이타 중 10개(70% 정도)가 0, 2와 30, 32에 에너지가 집중되어 있음을 보여주었다. 이 14개의 데이타는 여성 화자 3사람과 남성 화자 3사람이 발음한 것이다.

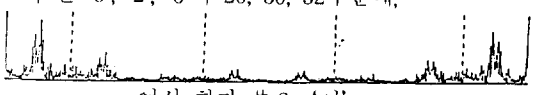
실험결과 데이타

'일'은 0, 3, 5와 27, 29, 32부분에,



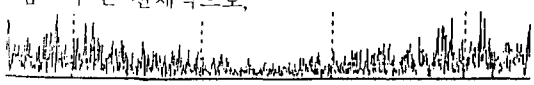
남성 화자 # 1 '일'

'이'는 0, 2, 6과 26, 30, 32부분에,



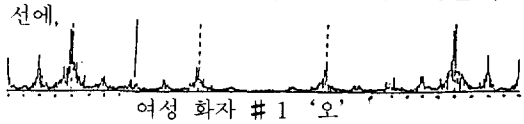
여성 화자 # 2 '이'

'삼' '사'는 전체적으로,



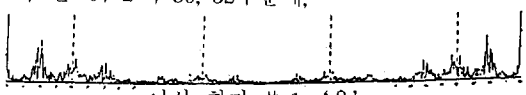
남성 화자 # 1 '사'

'오'는 0, 2, 4와 28, 30, 32부분 특히 양옆 1/8 선에,



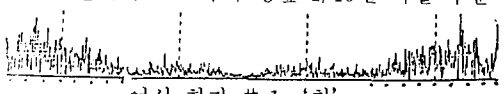
여성 화자 # 1 '오'

'육'은 0, 2와 30, 32부분에,



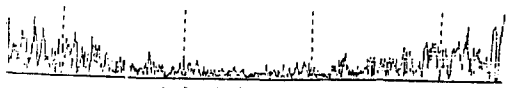
여성 화자 # 1 '육'

'칠'은 전체적으로 특히 양옆 1/16선 바깥 부분에,



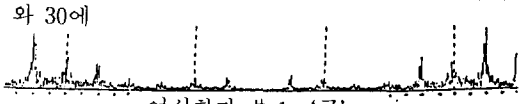
여성 화자 # 1 '칠'

'팔'은 전체적으로 특히 양옆 1/16선 바깥 부분에,



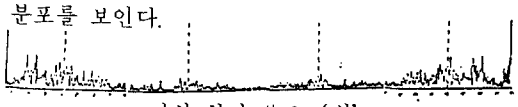
여성 화자 # 1 '팔'

'구'는 0, 2, 4, 6과 26, 28, 30, 32부분 특히 2와 30에



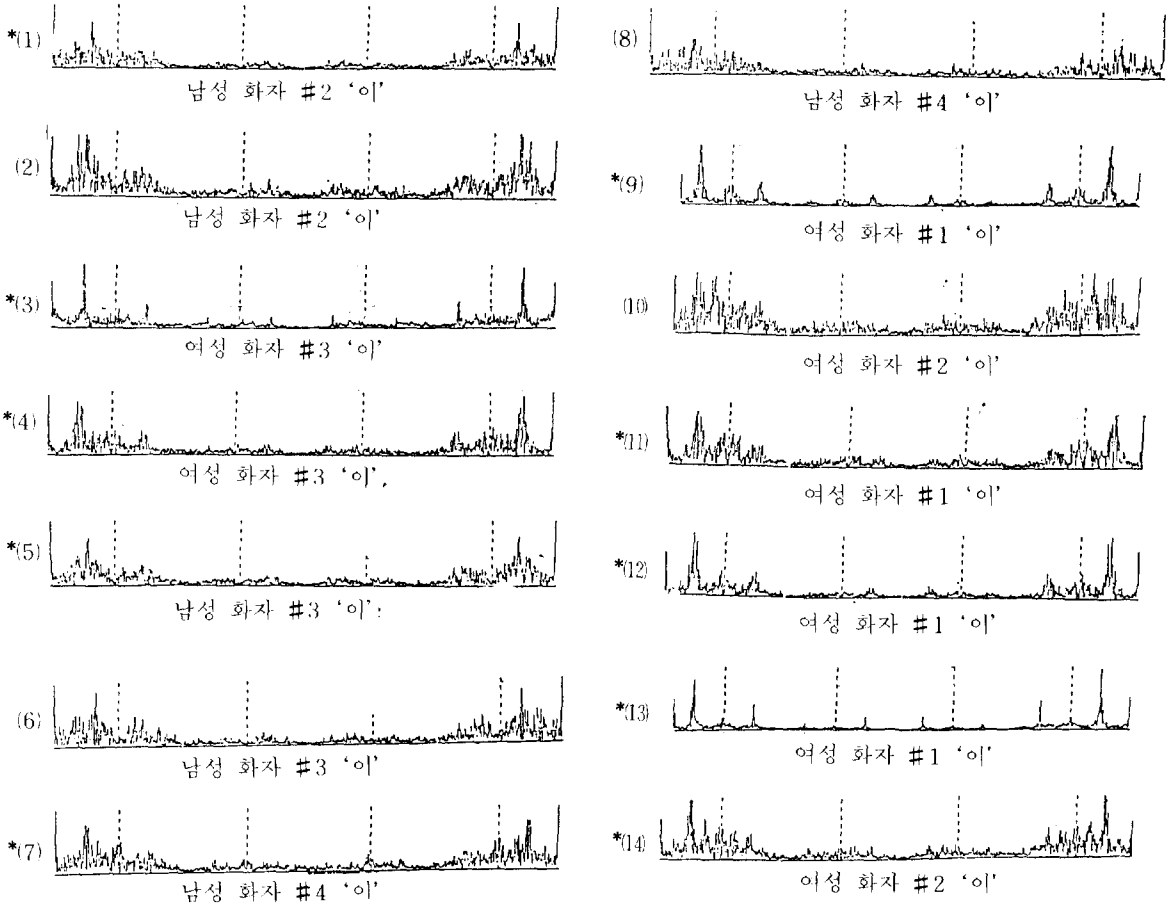
여성화자 # 1 '구'

'십'은 0과 32부분에, 가운데 부분은 특히 낮은 분포를 보인다.



여성 화자 # 1 '십'

각 단어는 이와 같은 위치에서 에너지 집중을 보인다.



한국어 숫자음 '이'에 대한 14개 데이타 중 10개(70% 정도)가 0.2와 30, 32에 에너지가 집중되어 있음을 보여주었다.

5. 결 론

Walsh함수를 이용하여 음성 분석을 행했을 때 여러 화자가 발음한 동일 단어(한국어 숫자음 '이'에 대하여 70%(14개의 패턴 중 10개)정도가 Walsh파우어 스펙트럼 상에서 유사한 에너지 집중(energy concentration)을 보였다.

음성 분석에 있어 같은 단어에 대한 화자들 간의 차이에 좌우되지 않는 특징(speaker independent feature)를 구하는 것이 문제가 되는데 Walsh 변환 방법을 통해 이러한 특징을 얻을 수 있는가 조사해 보았다. 실험 결과를 통해 여러 화자에 의해 발음된 동일 단어 패턴들 중 70% 정도가 유사한 패턴을 보임을 알 수 있었는데 이 결과는 화자독립 음성 인식 단계에서 높은 인식율을 얻을 수 있는 결과라고 생각한다.

본 논문에서는 또 Walsh 변환 방법으로 여러 화자에 의해 발음된 한국어 숫자음('일' - '십')에 대한 대표적인 패턴들을 구해보았다.

음성 인식이나 합성의 문제는 그 연구 결과가 사용되는 언어에 따라 다르기 때문에 한국어의 경우 국내에서 해결할 문제이지 외국의 연구 결과를 사용하거나, 비교할 수 없다. 따라서 국내에서의 독자적인 연구가 필요하다.[8]

참 고 문 헌

1) Wayne A. Lea, "Trends in Speech Recognition," Prentice-Hall, INC., Englewood Cliffs, NewJersey 07632, p24-38, 1980

2) 김 인경, "LPC 방식을 이용한 한국말 단어 인식에 관한 연구," 서울 대학원 전자과, 석사 학위

논문, 1985
3) 김낙현, "한국어 단음절어의 분류 인식에 관한 연구," 서울대학원 전자과, 석사 학위 논문, 1984
4) 이 찬길, "Dynamic Programming에 의한 한글 고립 단어 인식에 관한 연구", 서울대학원 전자과, 석사 학위 논문, 1983
5) J. D. Markel and A. H. Gray, Jr, 'Linear Prediction of Speech,' SpringerVerlag, Berlim Heidelberg New york, 1976
6) N. Ahmed, K. R. Rao, 'Orthogonal Transfrmos for Digital Signal Processing', Springer-Verlag. Berlim. Heidelberg, New York 1985. 1975
7) Alan V. Oppenheim/Ronald W. Shafer, 'Digital Signal Processing', Prentice-Hall, INC, Englewood Cliffs, New Jersey, 1975, p284~287
8) 한국과학기술원 전기 및 전자 공학과 통신 연구실, '디지탈 음성 처리 기술 연구 개발 최종 보고서,' 1987.