1999년도 대한전기학회 정기총회 및 추계학술대회 논문집 1999. 11. 20

실시간 음성타자 시스템 구현

조우연*, 최두일
공주대학교 전기전자정보통신 공학부

Development of Realtime Phonetic Typewriter

W.Y. Cho*, D.I. Choi
Dept. of Electrical Electronic & Information Eng., Kongju National Univ.

Abstract - We have developed a realtime phonetic typewriter implemented on IBM PC with sound card based on Windows 95.
In this system, analyzing of speech signal, learning of neural network, labeling of output neurons and visualizing of recognition results are performed on realtime. The developing environment for speech processing is established by adding various functions, such as editing, saving, loading of speech data and 3-D or gray level displaying of spectrogram.
Recognition experimental using Korean phone had a 71.42% for 13 basic consonant and 90.01% for 7 basic vowel accuracy.

1. 서 론

오래 전부터 음성은 인간이 가지고 있는 기본적인 능력 중에서 가장 중요한 것 중 하나로서 우리가 속박감을 거의 느끼지 않고 자유롭게 구사할 수 있는 가장 자연스럽고 효과적인 정보교류의 수단으로 여겨왔다. 이러한 인간과 인간사이의 의사소통의 수단으로만 사용되었던 음성이, 오늘날 논리적으로 사물을 생각하는 경우에 있어서도 중요한 역할을 하게되었다. 이 음성이 인간과 기계와의 통신, 즉, 정보의 교환수단으로도 사용되어 오고 있다.

최근 음성과 자연언어의 기본적인 성질의 이해에 관한 관심도 높아지고 있고 각종 미디어의 발달, 초고속 정보통신망의 구축과 더불어 멀티미디어 통신을 통한 통신판매, 물류처리, 제품홍보 등이 폭증하고 있으며 개인용 컴퓨터의 보급에 의한 신호처리기술과 정보처리기술의 급속한 발전과 더불어 음성을 통한 인간과 기계와의 직접적인 커뮤니케이션을 위한 Man-Machine Interface의 중요성도 강조되고 있다. 또 인간과 기계사이 뿐만 아니라 인간과 인간사이에 기계를 넣어 통역을 자동적으로 하고자 하는 연구도 활발히 진행되고 있다.

자기 생성 및 구조화 신경 회로망(SCONN)은 음성 신호 처리를 분석하는 적합한 환경을 제공하고 있다.

본 연구에서는 음성 인식 기술을 사용자 인터페이스로 하여 실시간 음성타자시스템을 구현하기 위한 기초가 되는 연구 중 하나인 음소인식에 대한 연구를 수행하였다. 현재 음성분석에 이용되는 상용화 된 장비는 고가일 뿐만 아니라 본 연구목적과는 적합하지 않은 환경이 대부분이므로 Visual C++ 6.0을 사용, 독자적으로 음성인식 시스템을 개발하여 연구에 이용하였다.

개발된 본 시스템의 특징은 사운드 카드가 장착된 IBM 계열의 개인용 컴퓨터와 현재 가장 대중적으로 사용되는 OS인 Windows95 환경에서 동작하도록 했고, 본 시스템의 모든 절차를 visual하게 실시간으로 구현했다.

본 연구는 음성의 분석과 학습, 라벨링, 인식결과를 그래픽과 문자로 출력하는 등의 구현이 가능하고, 또한 음성의 수집, 저장과 편집이 가능한 통합 개발환경을 구현하여 음성 타자 시스템을 구현하는데 있어서 기초적인 기반이 될 수 있는 환경을 구축하는데 그 목적이 있다.

2. 실시간 음성인식 시스템 구현

2.1 실시간 음성 인식 시스템 개요

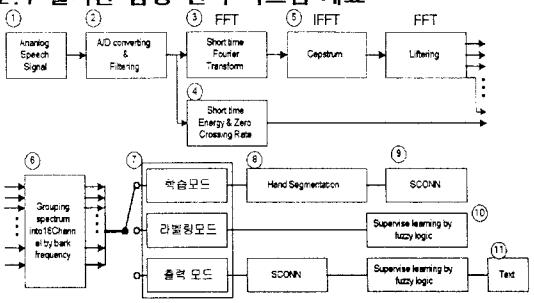

그림 2.1 음성인식시스템의 전체 블럭도

본 연구를 위하여 실시간 음성인식 시스템을 개발하였는데, 음성 인식 기술의 개요는 다음과 같다.

1.	음성 입력부 개발
2.	wave 파형을 FFT를 이용한 시간-주파수 변환 (스펙트럼 파워 계산)
3.	단구간 에너지와 영교차율 계산
4.	음성여부 판별
5.	켑스트럼과 리프터링
6.	16채널 바크 스케일로 재 샘플링
7.	학습모드, 라벨링 모드, 출력모드
8.	Hand Segmentation
9.	SCONN을 이용한 학습
10.	퍼지 논리에 의한 지도학습 신경망
11.	학습된 결과를 음소 단위로의 확률적 표현과 글자 출력

표 2.1 음성 인식 기술의 개요

2.1.1 음성 입력부 개발

실용성 있는 시스템 성능을 얻기 위해서, 일반적으로 사용되는 PC용 마이크와 일반 작업 환경에서 일어날 수 있는 잡음(환경잡음)을 제거하지 않은 상태에서 음성을 취득했다. 또한 실시간으로 음성 데이터를 취득하고 음성 파형을 실시간으로 시간-주파수 영역으로 변환하기 위해서 Double Buffering 기법을 사용하였다.[1] 단계별로 간단히 요약하자면 다음 표의 4단계로 정리 될 수 있다.

Step 1. 음성데이터를 Buffer1에 저장한다. Step 2. 버퍼1이 다 채워지면 버퍼2에 저장한다. Step 3. 버퍼1의 데이터를 Data Block으로 Copy하고 버퍼1의 데이터는 다음 데이터의 입력을 대기하기 위해서 free시킨다. Step 4. 위 의 처리과정을 음성데이터 입력이 종료 될 때까지 반복한다.

표 2.2 Double Buffering

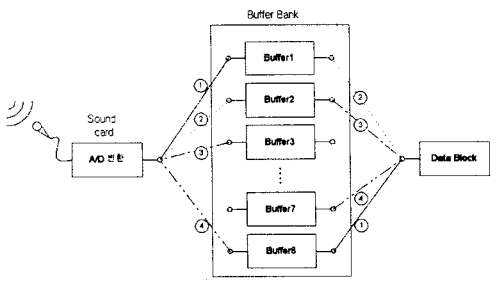

그림 2.2 Double Buffering 기법

2.1.2 음성여부 판별

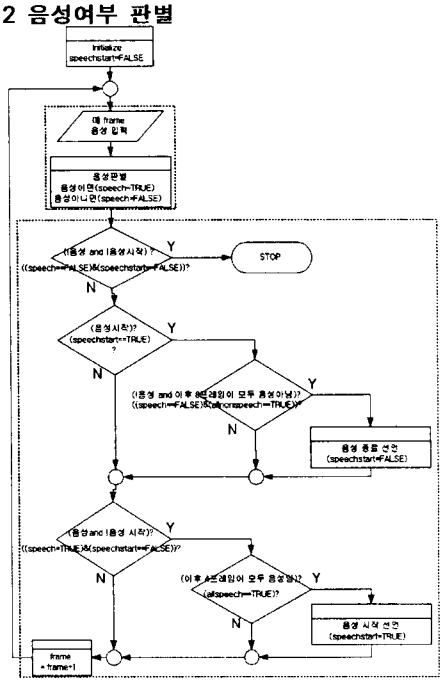

그림 2.3 음성 판별 플로차트

음성을 판별하기 위해서 첫째로 각 프레임마다 에너지와 영교차율을 가지고 에너지의 문턱값과 영교차율의 문턱값을 비교해서 음성여부를 판별을 하고, 그 다음으로 음성의 시작점과 끝점구간을 검출해내는 순서로 진행을 한다.

그림 2.3은 음성 판별을 위한 플로차트이다. 크게 두 개의 점선상자로 나누어 플로차트의 알고리즘을 프로그램화했다.

2.1.3 16채널 바크 스케일로 재 샘플링

얻어진 스펙트럼 파워의 데이터 수는 프레임 당 256개로 인식 실험을 효과적으로 수행하기 위해서는 데이터 수를 줄여야 한다. 이 때 256개의 선형 주파수 영역을 16개의 bark 주파수 영역으로 구루핑 하였다. Bark scale은 인간의 귀 구조와 유사한 형태로 나누는 방법이다. Bark Scale은 선형 주파수 영역을 인간의 달팽이관 구조와 유사하게 비선형 주파수 영역으로 환산하는 것으로 Zwicker 와 Terhardt[2] 가 제안한 식 2-1을 이용하였다.

$$f(Hz) = \frac{\tan(\frac{Z_c(Bark)}{13})}{0.76}, \quad f \le 1KHz$$

$$f(Hz) = 10^{(\frac{Z_c(Bark)}{14.2} - \frac{8.7}{14.2})}, \quad f > 1KHz \qquad (2\text{-}1)$$

여기서, bark scale은 코클리어의 기저막의 위치와 공진 주파수와의 관계를 나타내는데, 1 bark는 기저막 상의 1mm에 해당된다.

2.1.4 SCONN을 이용한 학습

1994년 최두일이 제안한 SCONN의 알고리즘은 표 2.3과 같다.[3]

SCONN의 알고리즘을 그림 2.4의 플로차트와 같이 구현하였다.

Step 1. Initialize Weights Step 2. Present New Input Step 3. Calculate Distance to All Node(s) Step 4. Find Active Node(s) and a Winner Node Step 5. If Active Node Does not Exist, then go to step 8 Step 6. Decrease Response Ranges of Active Node(s) Increase Response Ranges of Inactive Node(s) Step 7. Adapt Weights of Winner Node(or Winner node and its family nodes) go to Step 2 Step 8. Creat a Son Node from an Inactive Winner (Mother) Node go to Step 2

표 2.3 SCONN 의 알고리즘

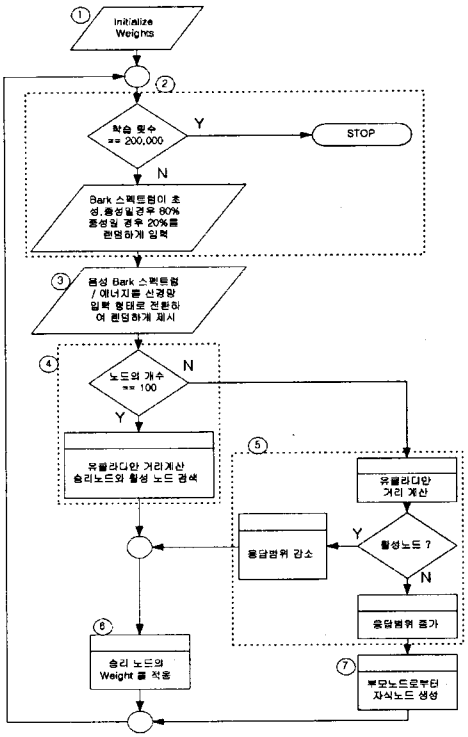

그림 2.4 SCONN의 플로차트

2.1.5 퍼지 논리에 의한 지도학습 신경망을 이용한 라벨링

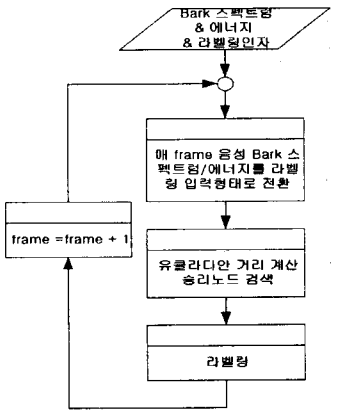

그림 2.5 라벨링의 플로차트

지금까지 학습된 각 노드(100개)는 입력되는 Bark 스펙트럼에만 적응이 되었을 뿐 text(자음, 모음)와는 적응이 되지 못한 상태이다. 따라서 그림 2.5에 보이는 플로차트 순으로 라벨링 과정을 통해, 학습된 노드와 text와 적응시키는 과정이 요구되는데 이를 라벨링이라 한다.

2.1.6 학습된 결과를 음소 단위로의 확률적 표현과 글자 출력

학습과 라벨링을 통해 생성된 데이터를 가지고, 음성인식을 테스트를 하면 학습에 의하여 라벨링 된 데이터와 가장 근접한 결과를 각 프레임마다 2차원 Gray level 그래픽을 사용하여 확률로 표현되어 출력하고, 매 프레임별로 가장 높은 확률을 기록한 글자를 텍스트로 출력하여 결과를 볼 수 있게 하였다.

2.2 실험 및 결과 고찰

2.2.1 실시간 음성 분석 시스템 블럭도

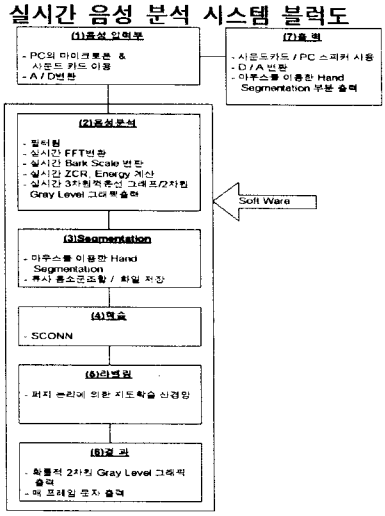

그림 2.6 실시간 음성 분석시스템 Block 도

음성 타자기를 구현하기 위한 실시간 음성 분석 시스템은 위와 같은 블록도로 구성되어 있다.

2.2.2 음성 인식

라벨링 과정까지 마치면 음성인식을 할 수 있는 상태가 된다. 본 시스템의 라벨링 옵션 창에서 라벨 보기 모드를 선택 한 후 마이크로 음성을 입력하면 그와 동시에 음성을 분석하여 그림 2.7와 같이 음성인식의 결과가 나타난다.

그림 2.7의 가로축은 시간 축을 나타내는 것이고 세로축은 아래에서부터 /ㄱ/ㄴ/ㄷ/ㄹ/ㅁ/ㅂ/ㅅ/ㅇ/ㅈ/ㅊ/ㅋ/ㅌ/ㅍ/ㅎ/ㅏ/ㅐ/ㅣ/ㅗ/ㅜ/ㅡ/ㅓ/_ㄱ/_ㄴ/_ㄷ/_ㄹ/_ㅁ/_ㅂ/_ㅇ/순으로 되어있어서 명암이 가장 진한 부분이 학습된 결과와 가장 유사한 부분이 된다. 바로 윗 부분은 매 프레임별 최대값에 해당하는 글자를 출력하도록 하여 음성인식을 확인하도록 하였다.

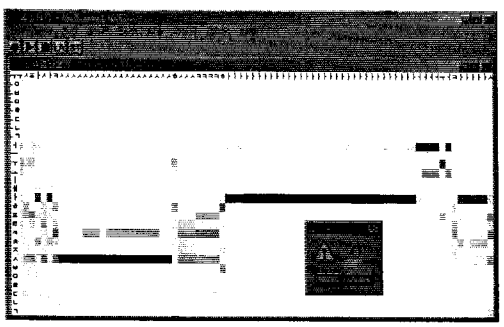

그림 2.7 /사/ 음성인식의 결과보기

2.2.3 인식 결과

표 2.4의 조건으로 학습한 후 그 인식 결과는 표 2.5에 보여진다.

옵션 구분	종류	입력값
사운드 카드 설정	channel 수	모노
	sampling rate	22kHz
	sample 당 bit 수	16 bit
	음성 문턱에너지	120,000
	음성 문턱 영교차율	0.5
스펙트럼 분석	window size	512 byte
	중첩율	0.75
	window 형태	Hamming
신경망 입력	응답범위 초기값	100
	감소인자	1.01
	증가인자	0.999
	학습 횟수	200,000번
	초성, 중성 학습율	80%

표 2.4 실험 조건 옵션

모음(중성) / 자음(초성)	아	애	어	에	오	와	외	우	워	위	으	의	어	이	○	×	인식률
가	가	개	거	게	고	과	괴	구	궈	귀	그	긔	거	기	4	3	
나	나	내	너	네	노	놔	뇌	누	눠	뉘	느	늬	너	니	6	1	
다	다	대	더	데	도	돠	되	두	둬	뒤	드	듸	더	디	1	6	
라	라	래	러	레	로	롸	뢰	루	뤄	뤼	르	릐	러	리	7	0	
마	마	매	머	메	모	뫄	뫼	무	뭐	뮈	므	믜	머	미	7	0	
바	바	배	버	베	보	봐	뵈	부	붜	뷔	브	븨	버	비	5	2	
사	사	새	서	세	소	솨	쇠	수	숴	쉬	스	싀	서	시	6	1	71.42%
자	사	재	저	제	조	좌	죄	주	줘	쥐	즈	즤	저	지	3	4	
차	차	채	처	체	초	촤	최	추	춰	취	츠	츼	처	치	6	1	
카	카	캐	커	케	코	콰	쾨	쿠	쿼	퀴	크	킈	커	키	4	3	
타	타	태	터	테	토	톼	퇴	투	퉈	튀	트	틔	터	티	4	3	
파	파	패	퍼	페	포	퐈	푀	푸	풔	퓌	프	픠	퍼	피	6	1	
하	하	해	허	헤	호	화	회	후	훠	휘	흐	희	허	히	6	1	
○		13		12		13		10		12		9		13			
×		0		1		0		3		1		4		0			
인식률	90.01 %																

표 2.5 기본 음성 실험 결과

위 실험은 자음 13개와 모음 7개의 조합으로 이루어진 한국어의 음소에 대하여 인식 실험을 한 결과이고, 특히 자음의 뒤에 오는 조음효과(이음)에 주목하여 인식결과를 관찰했다.

자음은 71.42%, 모음은 90.01%의 인식 결과를 나타냈다. 자음 중 특히 /ㄷ/과 /ㅈ/이 인식률이 저조한데, 통계적으로 예사소리(ㄱ,ㄷ,ㅂ,ㅅ,ㅈ) 계열이 전반적으로 인식률이 낮은 편이다. 이는 발음 구간이 매우 짧고 비정상적(non-stationary)이어서 인식의 어려움이 보이고 있다. 반면 울림소리(ㄴ,ㄹ,ㅁ,ㅇ) 계열은 발음 구간이 길기 때문에 인식률이 높다. 거센소리(ㅊ,ㅋ,ㅍ,ㅌ) 계열은 평균 이상의 인식률을 보이고 있고, 목청소리(ㅎ) 또한 높은 인식률을 보이고 있다.

3. 결 론

본 연구는 실시간 음성 타자기를 구현하기 위한 기초 기반이 될 수 있는 음소 단위 분석 시스템을 개발하였다. 본 시스템은 음성을 시각적인 방법을 통해 효과적으로 분석하고, 음성 인식에 적합한 모델인 자기 생성 및 구조화 신경 회로망(SCONN)의 최적성과 빠른 적응성을 이용하여 음소 인식 실험을 수행하였다. 이를 위하여 음성의 취득 및 저장, 편집 등을 할 수 있는 프로그램, 음성 신호를 시간 영역으로 시각화하고, segmentation하고 출력하는 프로그램, 시간-주파수 영역을 3차원과 2차원 Gray-level로 시각화 한 프로그램, Bark scale로 그루핑하여 시각화하고 segmentation 할 수 있는 프로그램, SCONN을 이용한 학습 프로그램, 퍼지 논리에 의한 지도 학습 신경망을 이용한 라벨링 프로그램, 인식결과를 그래픽과 문자로 출력하는 프로그램을 MS Windows 95 기반의 통합 환경으로 visual하게 개발하였다.

본 연구에서 구현한 음소 단위의 인식은 순수하게 음소 인식만을 구현했다. 하지만 지금 우리가 구현 하고자하는 음성 타자 시스템은 음소 인식과정을 거친 후 그 후처리 과정으로 단어사전과 비교하여 단어 인식과정을 거쳐야만 완벽한 음성 타자기가 구현됨을 미루어 볼 때 현재까지 각 음소에 대한 인식률만 가지고도 만족 할만 하겠다.

따라서 향후 연구 과제로는 음절과 음절사이 단어와 단어사이의 높은 segmentation율을 얻는 것과 단어사전과 비교하여 인식하는 시스템을 개발한다면 보다 완벽한 음성 타자기를 구현 할 수 있으리라 기대한다.

[참 고 문 헌]

[1] Richard j. Simon, "멀티미디어&ODBC", 대림, 617-618, 1997

[2] E. Zwicker and E. Terhardt, "Analytical expressions for critical - band rate and critical bandwidth as a function of frequency," JASA, vol. 68(5), pp. 1523-1525, 1980.

[3] D. Choi and S. Park, "Self-Creating and Organizing Neural Networks," IEEE Trans. on Neural Networks, Vol. 5, No. 4., pp. 561-575, July 1994.