신경망을 이용한 고립단어에서의 피치변화곡선 발생기에 관한 연구

임운천, 곽진구, 장석왕(호서대)

<Abstract>

A Study on the Pitch Contour Generator with Neural Network in the Isolated Words

Unchun Lim, Jingu Kwak, Sokwang Chang

The purpose of this paper is to generate a pitch contour which is affected by the phonetic environment and the number of syllables in each Korean isolated word using a neural network.

To do this, we analyzed a set of 513 Korean isolated words, consisting of 1-4 syllables and extracted the pitch contour and the duration of each phoneme in all the words. The total number of phonemes we analyzed is about 3800. After that we approximated the pitch contour with a 1st order polynominal by a regression analysis. We could get the slope, the initial pitch and the duration of each phoneme. We used these 3 parameters as the target pattern of the neural network and let the neural network learn the rule of the variation of the pitch and duration, which was affected by the phonetic environment of each phoneme. We used 7 consecutive phoneme strings as an input pattern for a neural network to make the network learn the effect of phonetic environment around the center phoneme.

In the learning phase, we used 3545 items(463 words) as target patterns which contained the phonetic environment of front and rear 3 phonemes and the neural network showed the correctness rate of 98.43%, 98.59%, 97.7% in the estimation of the duration, the slope, the initial pitch. In the recall phase, we tested the performance of the neural network with 251 items(50 words) which weren't need as learning data and we could get the good correctness rate of 97.34%, 95.45%, 96.3% in the generation of the duration, the slope, and the initial pitch of each phoneme.

Ⅰ. 서론

음성합성은 인간의 음향학적 정보전달수단인 음성을 기계가 소리의 합성을 통하여 가능하게 하는 기술이다.

이 기계에 의한 합성음은 올바른 정보전달능력으로서 이해도와 인간의 발성과의 유사함을 나타내는 자연성으로 평가되어 진다. 음성합성 분야의 연구들은 처음에는 이해도에 관심을 둔 명료성에 집중되었다. 그러나 음성합성의 영역이 넓어지고 보편화됨에 따라, 인간의 음성과 같이 자연스러운 합성음에 대한 요구가 증가되고 있고, 이에 따라서 특히 문-음성변환 시스템에서는 이를 실현하기 위한 수단으로서 운율제어에 대한 관심이 커지고 있다.

한국어의 운율요소에는 피치, 지속시간, 크기, 휴지기 등이 있는데, 이 운율요소들은 주로 의미론적인 요인과 구문론적인 요인에 의하여 영향을 받게 된다. 이중 의미론적인 요인의 분석은 문장의 의미론적인 이해가 선행되어야 하기 때문에 그 영역이 방대하여서 일정한 규칙을 도출하기도 어렵고, 처리 시간에도 문제가 있게 된다. 따라서 구문론적인 요인에 따른 운율요소의 변화를 살펴서 일정 법칙을 찾고, 이 법칙에 따라 운율을 제어하는 것이 통상적인 운율제어의 형태이다. 그러나 현재까지 한국어에서 구문론적인 구조와 운율요소와의 관계에 대한 체계적이고 광범위한 연구는 진행되지 않았고, 최근에야 피치와 지속시간에 대한 연구가 발표되거나 진행 중이어서 문-음성변환 시스템에 널리 사용할 수 있는 수준에는 아직 미치지 못하고 있는 실정이다.

합성음의 이해도와 자연감에 중요한 요소로 작용하는 운율요소 중 피치가 이에 끼치는 영향은 매우 큰 것으로 평가되고 있다. 이러한 이유로 피치의 변화곡선에 대한 연구는 일찍이 50년대부터 시작되어 지금까지 계속되고 있는데 이는 제시된 변화곡선모델이 자연음의 피치변화곡선을 아직도 제대로 추정하지 못하기 때문일 것이다.

피치의 변화에 영향을 주는 요인에 따라 구문론적인 측면에서 평서문을 대상으로 피치변화모델을 세분하면 문장전체에 걸친 주변화와 구.절 경계와 단어의 강세유형에 따른 부변화 그리고 분절에 의한 미세변화로 나눌 수 있다. 이중 분절에 의한 미세변화란 비록 같은 음소라 할지라도 단어 내에서 전후에 오는 유/무성과 조음방법 그리고 음절수에 따라 피치변화의 정도가 달라짐을 말한다. 때문에 이러한 모든 경우를 일일이 법칙화하여 적용한다는 것은 상당히 힘든 일이 될 것이다.

이러한 문제의 해결 방안 중의 하나로서 문-음성변환 시스템을 위한 새로운 합성단위인 CDU(Context Dependent Unit)를 사용하여 복잡한 규칙 없이도 음절내의 미세운율을 살려 줌으로서 합성음의 자연성을 확보한 예가 있다. 그러나 작성된 데이타베이스에 대한 개방을 꺼려하는 현 시점에서 CDU와 같은 음절내의 미세운율을 표현할 수 있는 새로운 합성단위를 만들어 합성음질의 개선을 꾀한다는 것은 너무나 방대한 작업이기 때문에 많은 시간과 인원을 필요로 한다는 단점을 가지고 있다.

따라서 본 연구의 목적은 우리말 고립단어내에 존재하는 분절에 의한 실제 피치와 지속시간의 변화를 자연음으로부터 직접 유출하여 이를 토대로 하여 이를 법칙화한 후, 무성자음의 경우 변이음 규칙에 의하여 처리하고 유성자음과 중성모음의 경우는

법칙에 의해 음소환경에 따른 음절내의 미세운율 조절을 통해 처리함으로써 규모가 적은 합성단위를 사용하면서도 합성음의 자연성을 높여 주기 위해 본 논문에서는 고립단어내의 분절에 의한 피치와 지속시간의 변화법칙을 신경망을 통해 학습시킴으로써 복잡한 운율법칙 알고리즘을 위한 프로그램 작성 없이 운율을 제어하는 새로운 방법을 제안하였다.

Ⅲ. 음성분석

음성합성을 위한 데이타베이스를 작성하고 합성계수에 대한 단위간 변화법칙을 구하기 위해서는 안정된 음성시료의 수집과 정확한 분석방법의 적용이 필요하다. 음성합성은 인식과 달리 여러 화자를 대상으로 하는 것이 아니고 단일 화자의 합성음이 필요하므로, 단일화자를 선택하여 고립단어 내에서의 미세운율변화에 대한 법칙생성을 위해 필요한 단어목록을 작성한 후 일정한 녹음조건하에서 수회 반복 발음하게 하여 음성시료를 채취하고, 이들 시료로부터 필요한 법칙을 추출하여야 한다.

Ⅲ-1. 음성시료채집

Ⅲ-1-1. 화자선택 및 A/D 변환

듣기에 좋고 분석하기 쉬운 화자를 선택하여야 하는데, 이러한 조건을 만족시키는 화자로는 공명특성이 좋고 피치가 길며 문장 끝에서 피치나 진폭의 급격한 감소가 없는 20-40대의 남성화자가 적합하다.

무반향실이나 조용한 장소에서 필요한 음성을 화자로 하여금 발음하게하는데, 이때 마이크와 입 사이를 지향성으로 하되 호기가 마이크에 직접 영향을 주지 않을 정도의 거리(20 cm)를 유지하도록 한후 발음하도록하여 4 KHz 저역통과 필터를 통과한 신호를 표본화주파수 10 KHz, 분해능 16 bit로 A/D 변환을 실시하여 음성시료를 저장한다.

그림 2-1에 음성시료 채집과정을 보인다.

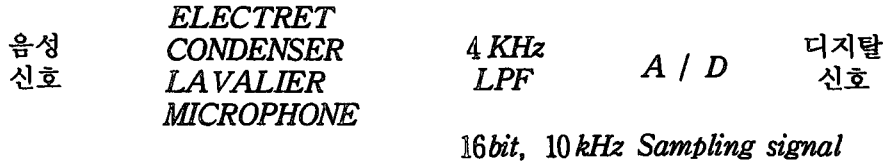


그림 2-1. 음성시료 채집과정

Fig. 2-1. The procedure of speech data collection

Ⅲ-1-2. 음성시료

본 논문에서 사용한 음성시료는 20대 남성화자로 하여금 각 단어를 3회 반복발음

케하여 이를 A/D 변환을 통해 구했다.

제시된 단어목록은 한국어에 존재하는 모든 음소와 다양한 주변음소환경을 포함하고 있으므로 이 단어목록에 의한 음성시료로부터 주변음소환경에 따른 중성모음과 유성자음(공명음)의 지속시간과 피치변화 그리고 무성자음(장애음)의 지속시간을 추출하였다.

II-2. 운율추출

그림 2-2에는 /군밤/에 대한 각음소 구간의 분리를 위한 각각의 계수들을 보인다.

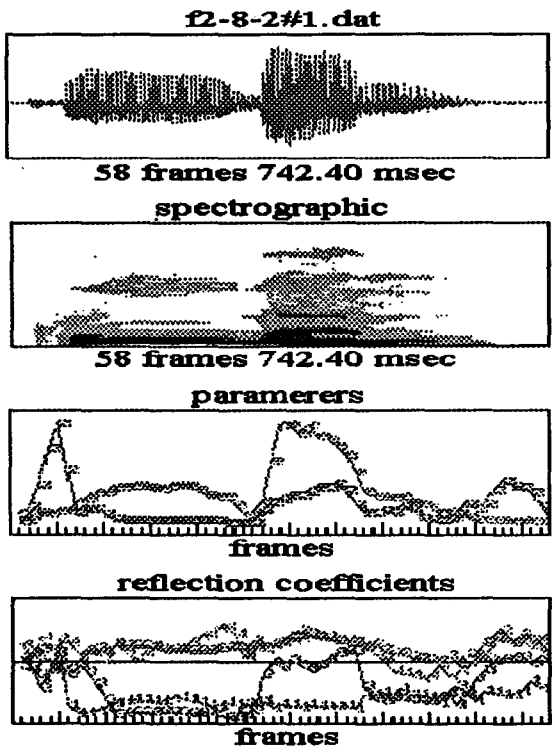


그림 2-2. 음소분할를 위한 각각의 계수들
Fig. 2-2. Parameters for phonetic segmentation

우선 각 고립단어 내의 주변음소환경에 따른 중성모음과 유성자음의 지속시간과 피치변화 그리고 무성자음의 지속시간을 추출하기 위해 파형과 스펙트로그램, 에너지, 영교차율 그리고 반사계수를 참고로 하여 파형선상에서 이들의 구간을 설정한 후, 추출된 부분이 해당 음가를 갖는지 청취를 통해 확인하여 정확하다고 인정되는 경우만을 데이타로 취하였다.

II-2-1. 모음 및 유성자음의 지속시간과 피치변화곡선 추출

모음과 유성자음 즉 공명음의 경우는 비교적 구분이 용이해 해당 구간전체를 구하

였다. 이때 음절말이 공명음으로 끝나고 다음 음절의 시작 역시 공명음인 경우 두 공명음의 구분이 힘들었는데 이는 구간을

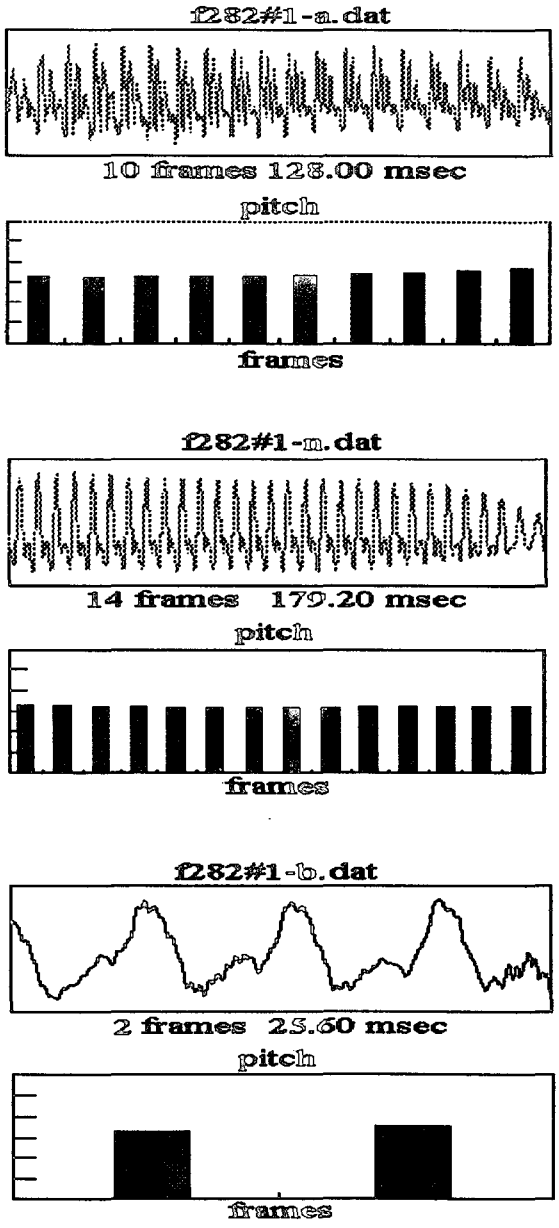


그림 2-3. /아/, /ㄴ/, /ㅂ/의 파형과 지속시간 및 피치변화곡선
Fig. 2-3. Waveform, Duration and Pitch contour of /a/, /n/, /b/

변화시키면서 청취를 통해서 구분하였다(예 달리다, 걸레, 달라, 달래다 등등). 그리고 평음의 유성음화 현상 즉, 국어의 'ㄱ, ㄷ, ㅂ, ㅈ'는 모음간에서 유성음화 된다는

현상인데 이 환경에서 기음구간이 나타나기도 하므로 실제 음향학적으로 유성화 현상이 항상 관찰되는 것이 아니지만 이것 역시 유성음으로 간주하여 피치변화와 지속시간을 측정하였다(예 군밤, 덜다, 갈비 등등).

이때 피치변화라함은 피치주기의 변화를 말한다. 일반적으로 운율을 취급할 때는 기본주파수를 구하여 그 변화를 살피게 되나 실제 합성시에 필요한 계수는 피치주기이므로 따로 기본주파수로 변환하여 사용하지 않고 피치주기를 사용하여 변화곡선을 구하였다. 참고로 피치주기(P)와 기본주파수(F0) 그리고 표본화주파수(Fs)의 관계는 다음과 같다.

$$F_0 = \frac{F_s}{P} \qquad (1)$$

그림 2-3에는 /군밤/에서 분리한 모음 /아/와 공명음 /ㄴ/ 그리고 유성음화된 장애음 /ㅂ/의 음성파형과 그 지속시간, 피치변화곡선을 나타내었다.

II-2-2. 무성자음의 지속시간 추출

무성자음 즉 장애음의 경우 단어초에서는 묵음구간을 뺀 발화개시점(VOT)을 측정하였다. 또한 음절말 불파음화 현상 즉, 국어의 음절말에 허용되는 자음은 'ㄱ, ㄷ, ㅂ, ㅁ, ㄴ, ㅇ, ㄹ'뿐이고 그중에 'ㄱ, ㄷ, ㅂ'은 허파에서 나오는 공기의 흐름이 구강에서 차단된 상태에서 음절을 마치는 현상으로서 이로 인해 무성자음의 불파된 부분은 다음 음절이 파열음으로 시작하는 경우 이 파열음의 묵음구간과 구별이 불가능하므로 음절말 'ㄱ, ㄷ, ㅂ'는 묵음구간을 포함한 구간의 길이를 지속시간으로 측정하였다(예 굿다, 걷다, 각별, 독풀 등등). 그림 2-4에는 /군밤/에서 분리한 무성자음 /ㄱ/에 대한 파형과 지속시간을 나타내었다.

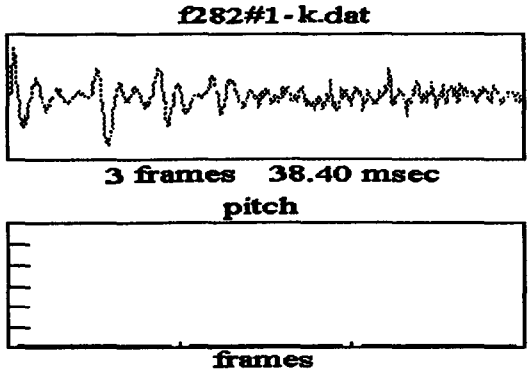


그림 2-4. 무성자음 /ㄱ/의 파형과 지속시간

Fig. 2-4. Waveform and duration of unvoiced consonant /k/

III. 신경망 구성

역전파 신경망을 이용한 초기의 응용 중에서 가장 잘 알려진 것 중의 하나가 영어로 된 문장을 음성으로 바꾸어 주는 음성합성 시스템인 네토크(Netalk) 시스템이다. 이 시스템은 1987년 세즈노우스키와 로젠버그에 의해 개발된 것으로서 이 시스템의 목표는 제시된 7개의 문자중 중앙에 해당하는 음소의 정확한 발음이며 이때 주변에 있는 6개의 문자는 중앙의 음소의 발음에 보조적인 아이디어 즉 문맥정보(context)를 제공하도록 구성되어 있다. 고립단어 내에서의 주변 음소환경에 따른 운율변화법칙을 학습시키기 위해서 본 논문에서 구성한 신경망 역시 네토크와 마찬가지로 7개의 음소중 중앙에 해당하는 음소의 피치변화곡선을 출력하도록 구성하였다.

III-1. 입력패턴

국어에서 사용되는 문자는 초성자음 18개, 중성모음 21개, 종성자음 27개로 이루어져 있다. 이를 1진 벡터화 하기 위해서는 초성과 종성의 중복을 고려한다 하더라도 하나의 문자를 코드화하기 위해서는 최소한 51 bit를 필요로 한다. 이는 신경망의 입력층 뉴런의 갯수를 증가시키게 되고 신경망의 학습속도와 직접적인 관련이 있으므로 전처리를 통해 입력층 뉴런의 갯수를 줄여줄 필요가 있다.

본 논문에서는 전처리로써 각 단어를 초, 중, 종성으로 분리하여 문-음소변환 알고리즘을 사용하여 음운변동을 적용한 후 이를 통해 얻어진 음소열 즉 초성자음 18개, 중성모음 21개, 종성자음 7개을 다음과 같이 1진벡터화하여 하나의 음소당 43 bit를 신경망의 입력패턴으로 하였다. 입력패턴 작성 방법은 다음과 같다.

음소열

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
ㄱ	ㄲ	ㄱ	ㄷ	ㄸ	ㄷ	ㅂ	ㅃ	ㄱ	ㅅ	ㅆ	ㅈ	ㅉ	ㅊ	ㅎ	ㄴ	ㄹ	ㅁ	ㅇ	이	위	에

23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40		41	42	43
외	애	으	어	아	우	오	야	여	요	유	예	애	웨	왜	워	와	의		초	중	종

입력패턴

ㄱ [1000000000000000000000000000000000000000000 100]

이 [0000000000000000000010000000000000000000000 010]

ㅁ [0000000000000000000010000000000000000000000 001]

III-2. 출력패턴

모음과 유성자음의 지속시간 동안의 피치변화에 대한 출력패턴은 엑셀5.0 상에서 회귀분석을 통한 추세선을 이용하여 1차 다항식으로 근사화한 후 그 계수인 기울기 값(a)과 y절편(늑첫프레임의 피치주기(p0))을 1진 부호화하여 작성하였다. 그리고 다항식의 변수인 지속시간(d)은 그 프레임수를 1진 부호화하여 출력패턴을 작성하였다.

그림 3-1에 단어 /김/에 대한 /이/와 /ㅁ/ 음소의 피치변화곡선과 회귀분석에 의한 추세선을 나타내었다.

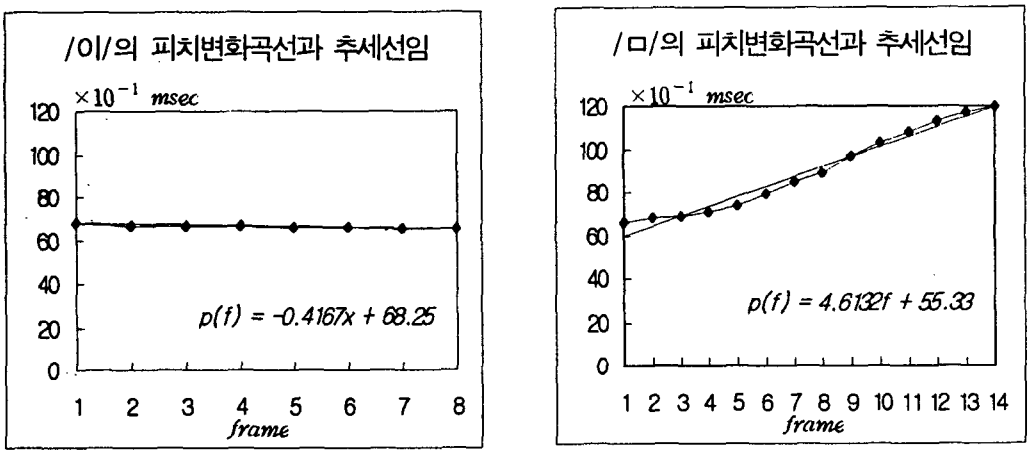


그림 3-1. /김/의 /ㅁ/, /이/에 대한 피치변화곡선과 추세선
Fig. 3-1. Pitch contour and Regression line for /m/ and /i/ of /kim/

무성자음의 경우는 피치가 존재하지 않으므로 피치변화곡선의 기울기와 y절편은 모두 0으로 부호화하였으며 지속시간은 모음과 유성자음의 경우와 마찬가지로 그 프레임수를 1진 부호화하여 출력패턴을 작성하였다.이때 각 음소의 지속시간과 y절편 즉 초기 피치값의 존재 범위는 각각 0~300 msec(0~24 frame), 5~8 msec 이므로 이를 1진 부호화하기 위해서 각각 31 bit씩를 할당하였다.

피치변화곡선의 기울기의 경우는그 최대 존재 범위가 -9~+9이므로 이를 부호화하기 위해서 지속시간, 초기 피치 값과 마찬가지로 31 bit를 할당하여 첫 1 bit는 부호를 다음 10 bit는 단자리를 다음 10 bit는 소숫점 첫째 자리를 다음 10 bit는 소숫점 둘째 자리를 나타내도록 하였다.

다음은 단어 /김/에서의 각 음소의 출력패턴 작성의 예이다.

[*frame*] × 10^{-1} [*msec*]

음소	지속시간(d)	기울기(a)	첫 피치주기(p_0)
ㄱ	3	0	0
이	8	-0.42	68
ㅁ	14	4.61	66

지속시간

ㄱ [0010000000000000000000000000000] [1 2 3 4 5 · · · · · · · · · 31]
이 [0000000100000000000000000000000] [1 2 3 4 5 · · · · · · · · · 31]
ㅁ [0000000000000100000000000000000] [1 2 3 4 5 · · · · · · · · · 31]

피치변화곡선의 기울기

ㄱ [0000000000000000000000000000000] [+/- 1 2 · · 0 . 1 2 · · 0 1 2 · · 0]
이 [1000000000000010000000100000000] [+/- 1 2 · · 0 . 1 2 · · 0 1 2 · · 0]
ㅁ [0000100000000000100001000000000] [+/- 1 2 · · 0 . 1 2 · · 0 1 2 · · 0]

첫프레임의 피치값

ㄱ [0000000000000000000000000000000] [51 52 53 54 55 · · · · · · 81]
이 [0000000000000000010000000000000] [51 52 53 54 55 · · · · · · 81]
ㅁ [0000000000000000100000000000000] [51 52 53 54 55 · · · · · · 81]

III-3. 신경망 구성

신경망 시스템은 그림 3-2과 같이 계층적 구조인 입력층, 은닉층, 출력층의 3개의 층으로 구성하였다.

입력층은 7개의 음소를 나타내는 유니트들로 이루어져 있는데 각 음소는 단지 하나의 유니트만을 활성화시키는 43개의 유니트들로 표현된다. 그중에서 40개는 우리말의 각 음소에 대응하고 나머지 3개의 유니트는 각각의 음소가 초성인가 중성인가 종성인가를 지정한다. 따라서 입력층에 필요한 총 유니트의 수는 301(7×43)개가 된다. 앞에서도 언급했듯이 이 신경망의 목표는 제시된 7개의 문자의 중앙에 해당하는 음소의 정확한 피치변화곡선 출력이다. 주변에 있는 6개의 음소는 중앙의 음소의 피치변화곡선 출력에 보조적인 아이디어를 제공한다. 7개 이상의 윈도우를 구성할 수도

있는데 이 경우 계산량이 매우 커지게 되며 우리말의 경우 각음소의 지속시간 등에 영향을 미치는 음소는 전후 3음소 정도로 보고된바 있으므로 본 연구에서는 7개의 윈도우를 사용하였다. 출력의 경우 각 음소의 지속시간(d)과 피치변화곡선의 기울기(a) 그리고 첫 프레임의 피치주기(p0)를 나타내는 31개의 유니트로 구성하였다 중간층에는 130개의 은닉 유니트를 구성하여 301개의 각 입력 유니트와 31개의 출력 유니트와 완전연결(full connection)하였다.

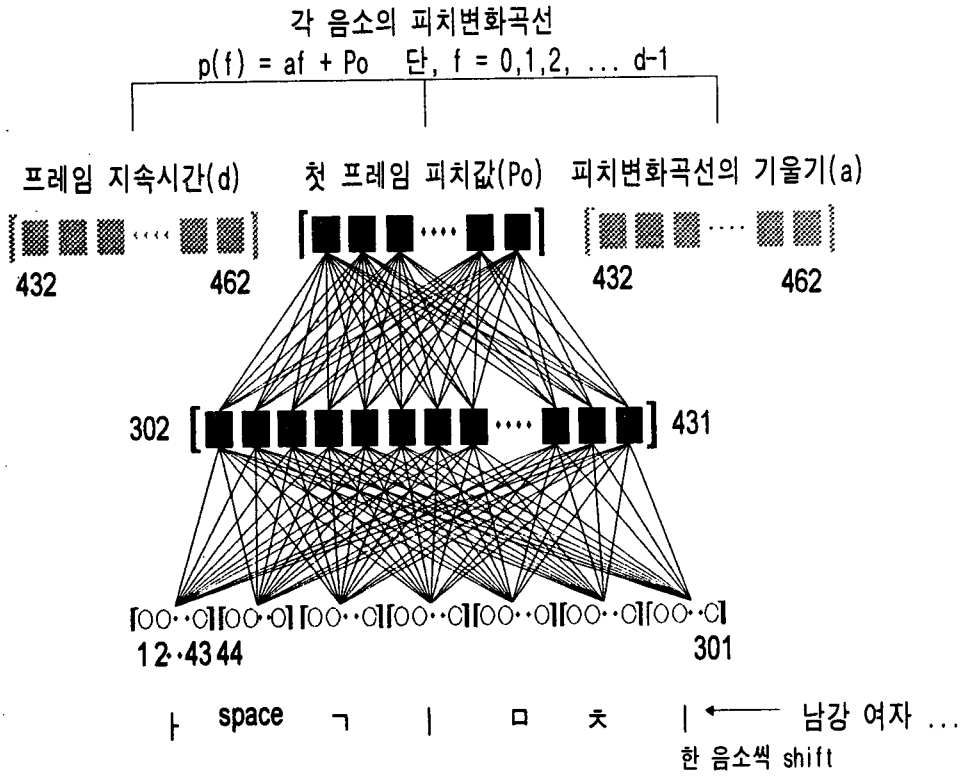


그림 3-2. 신경망 구성도

Fig. 3-2. Configuration of neural network

Ⅳ. 모의실험

학습을 위해 사용한 훈련용 단어집합을 표 4-1에 그리고 회상을 위해 사용한 회상용 단어집합을 표 4-2에 나타내었다. 제시된 단어집합은 한국어에 존재하는 모든 음소와 다양한 주변음소환경을 포함하고 있다.

이 단어집합에 의해 역전파 알고리즘과 시그모이드 함수를 사용하여 신경망의 학습을 실시하였다.

표 4-1. 훈련용 단어집합
Table 4-1. Set of words for Training

게 덱데굴 사게 개 객지 홱휙 귀뚜라미 획기적 괴물 긋다 늑대 손찌금 걷다 먹다 막다 압박 국 고리 목 유공 먹어 개 견디다 역사 강경 약속 교실 욕 궤 꽥 꽥 권유 왹더굴 과자 광대 왁새 귀족 신 넷 갠 뉘우침 쉰 뇌다 된장 근 수건 날 간나위 간 누구 눈 돈 손칼 남녀 연못 돈냥 뱐덕 감류 윤씨 뉘다 눼 웬 괜찮다 눠 원고 완수 데 대 했다 뒤 궛젓 되다 됫박 들다 듣고 덜다 걷고 달 맡고 간다 덜다 둘 굳고 돌 돋고 먿다 옛 엿 얕보다 윷 푯대 돼지다 돼지 쾟돈 둬라 구웠다 돠르르 왔다 달리다 궁도리 길 걸레 수레 고래 밸 갤 쥘손 군뢰 욀총 몰르다 다르다 횰러 굴 횰러 비렁뱅이 걸레 건류 건룡 갈랴고 멀다 달라 아람 달라 달리다 갈 날짜 불룩 군뢰 횰로 비로서 몰르다 골 결례 비례 여력 열다 가랴 걀쪽하다 건류 오륜 건룡 달려라 이뤄 월간 왈패 늴리리야 메뚜기 셈 매 남매 뛈 뫼 굄돌 금 막다 감 춤 목 먹다 염치 먈걍다 얌치 묘기 고욤 뭬 뀀지 뭘 비 입다 잡뗌 남비 우비 베 배 햅쌀 쉽다 뷥다 늪 버섯 덮다 간섭 바람 잡다 압박 사발 삼발이 부리 굽다 보리 돕다 졥시다 벼슬 옆 뱌비다 붜라 바라 실 새신 시들다 셋 새신 쇠 스님 서다 간섭 강석 사다 남산 숯 소 솜씨 셔츠 하셔요 하셔서 쉬 파리 쉽다 쉼표 쇄국 솰솰 오징어 싱싱 잉어 징 넹큼 생강 뒹굴다 굉장 등 성애 엉겅퀴 강음 강아지 강우 강요 종이 똥오줌 공예 몽외 공원 동넠 경영 경우 경위 경과 영양 양 융 용 숭어 궁둥이 웽웽 괭가리 꿩 광 닝큼 짐 반지 던지다 제기 방파제 재다 쥐 죄송 즉시 절구 고저 자리 국자 잔치 사자 감자 가자 간장 크다 칼 식칼 각하 커서 켸켸 낱켤레 캭캭 쿡쿡 코 땅콩 퀭하다 킝킝 콸콸 침장치 앞치마 체면 채롱 취 쵀고 측면 처녀 강철 차다 독창 추위 초가 취지 췌장 촬영 티 톄 태 튀기다 틔보 틀 터 탈 낙타 투구 감투 토막 손톱 밥통 튀다 틔다 피 폔 팽이 펄렁 파 감파르다 풀 독풀 날푼 폭포 폐 펴다 퍅하다 표 삽화 힘 오히려 조용히 헤치다 여행 해롭다 훠파람 훠두르다 회담 회의 호르다 허리 하얗다 훈장 간호 냉혹 혹 혜택 혀 형제 향토 영향 휴가 흉상 효도 형식 훼방 횃대 훤하다 환히 끼니 께 끼다 깨 꾀 끄다 갓끈 땅껍질 납거미 꺼라 독감 까다 껴안다 꺄웃하다 돋구다 감꽃 꼬리 뀌다 꿰다 꽹가리 꿩 꽈리 띠 각띠 뗴 때 뛰다 뙤약볕 뜰 떨다 딸 뚫다 똥 감떡 뛔라 뙤기 똬리 띄우다 뺄기 뗴 뗴다 뻗치다 빠지다 뿌리 숙뿌리 등불 뿝다 낮보기 뻐 뺨 뿌죽하다 뾰족하다 씨름 씨닭 씨 씻다 납세 쌩긋이 쐬다 쓰다 썩다 싸다 쑤시다 죽쑤다 쐐기 쏴쏴 쐬우다 찜 낙지 손찌금 째꺽 째마리 쩌다 짝 납짝코 사립짝 쭈구리다 쪽 쩍말없다 쫘르르 쪽쪽발

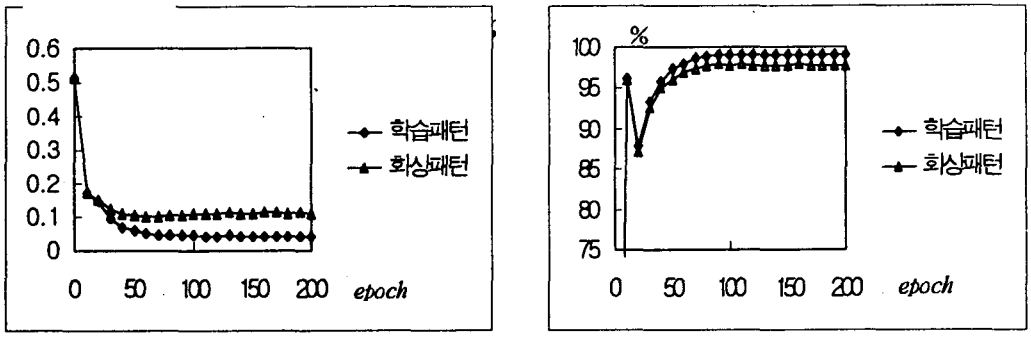
- 이상 463 단어(3545 패턴) -

표 4-2. 회상용 단어집합
Table. 4-2. Set of words for recalling

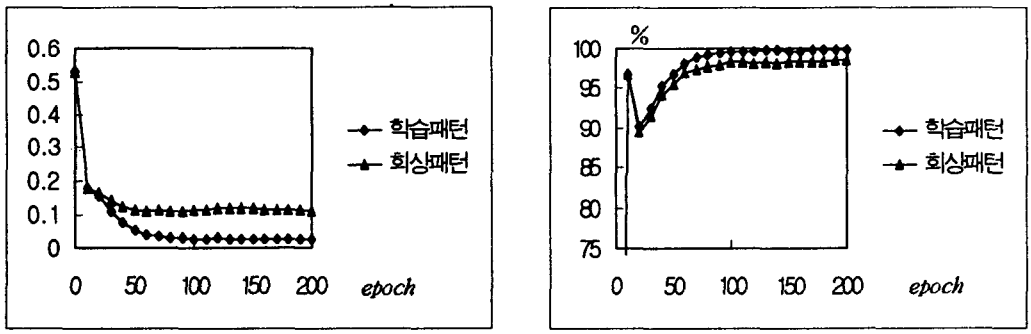
김 익다 가다 각별 구리 계시다 갸웃 육성 괘종 꿱꿱 이야기 감기 간마기 날개 남강 난간 관광 내일 느끼다 논 공룡 엔담 맨주먹 갑판 손칼 디디다 믿고 빗 나도 감독 달래다 불룩 다리 아뢰다 벼루 멜땡 꿀 미술 뱀 주검 묵다 곰 며느리 멥새 갈비 남봉 군밤 달력 귤 치료

- 이상 50 단어(251 패턴) -

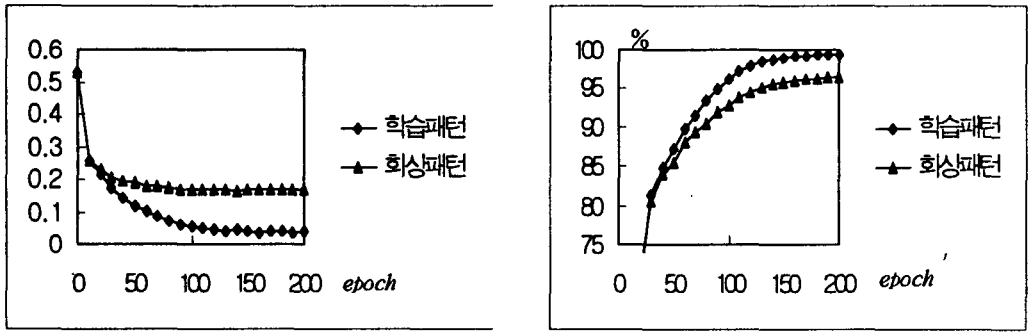
그림 4-1에는 학습횟수에 따른 RMSError와 정확도를 보인다. 또한 회상결과의 일부를 실제 목표값과 비교하여 표 4-3에 나타내었다.



(a) 지속시간



(b) 첫프레임피치



(c) 기울기

그림 4-1. 학습횟수에 따른 출력의 RMS에러와 정확도
Fig. 4-1. RMSError and correctness of output vs epochs

표 4-3. 희상결과

Table. 4-3. The result of recall

단어	희상데이타							d^{rc}	p_0^{rc}	a^{rc}	d^{org}	p_0^{org}	a^{org}
김	ㄱ	ㅏ		ㄱ	ㅣ	ㅁ		3	0	0	-	-	-
			ㄱ	ㅣ	ㅁ		ㅣ	8	65	-0.27	-	67	-
		ㄱ	ㅣ	ㅁ		ㅣ	ㄱ	14	68	3.27	-	-	-
익다	ㅣ	ㅁ		ㅣ	ㄱ	ㅍ	ㅏ	4	76	-2.3	-	-	-
	ㅁ		ㅣ	ㄱ	ㅍ	ㅏ		16	0	0	-	-	-
		ㅣ	ㄱ	ㅍ	ㅏ		ㄱ	2	0	0	-	-	-
	ㅣ	ㄱ	ㅍ	ㅏ		ㄱ	ㅏ	17	59	4.72	-	-	-
가다	ㅍ	ㅏ		ㄱ	ㅏ	ㄷ	ㅏ	3	0	0	-	-	-
	ㅏ		ㄱ	ㅏ	ㄷ	ㅏ		5	76	-0.7	-	-	-
		ㄱ	ㅏ	ㄷ	ㅏ		ㄱ	3	75	0	-	-	-
	ㄱ	ㅏ	ㄷ	ㅏ		ㄱ	ㅏ	15	66	3.11	-	-	3.09
각별	ㄷ	ㅏ		ㄱ	ㅏ	ㄱ	ㅛ	3	0	0	-	-	-
	ㅏ		ㄱ	ㅏ	ㄱ	ㅛ	ㅕ	3	77	-0.5	-	-	-
		ㄱ	ㅏ	ㄱ	ㅛ	ㅕ	ㄹ	17	0	0	-	-	-
	ㄱ	ㅏ	ㄱ	ㅛ	ㅕ	ㄹ		2	0	0	-	-	-
	ㅏ	ㄱ	ㅛ	ㅕ	ㄹ		ㄱ	5	60	5.32	-	-	3.19
	ㄱ	ㅛ	ㅕ	ㄹ		ㄱ	ㅜ	9	70	4.13	-	-	-
구리	ㅕ	ㄹ		ㄱ	ㅜ	ㄹ	ㅣ	3	0	0	-	-	-
	ㄹ		ㄱ	ㅜ	ㄹ	ㅣ		4	70	-1.4	-	-	-
		ㄱ	ㅜ	ㄹ	ㅣ		ㄱ	2	66	-1	-	-	-
	ㄱ	ㅜ	ㄹ	ㅣ		ㄱ	ㅖ	15	62	2.88	-	-	-
계시다	ㄹ	ㅣ		ㄱ	ㅖ	ㅅ	ㅣ	3	0	0	-	-	-
	ㅣ		ㄱ	ㅖ	ㅅ	ㅣ	ㄷ	8	66	-0.2	-	-	-
		ㄱ	ㅖ	ㅅ	ㅣ	ㄷ	ㅏ	2	0	0	1	-	-
	ㄱ	ㅖ	ㅅ	ㅣ	ㄷ	ㅏ		5	64	0	-	-	-
	ㅖ	ㅅ	ㅣ	ㄷ	ㅏ		ㄱ	2	75	0	-	0	-
	ㅅ	ㅣ	ㄷ	ㅏ		ㄱ	ㅑ	16	58	3.43	-	-	2.65
갸웃	ㄷ	ㅏ		ㄱ	ㅑ	ㅜ	ㄷ	3	0	0	-	-	-
	ㅏ		ㄱ	ㅑ	ㅜ	ㄷ		13	66	-0.03	-	-	-
		ㄱ	ㅑ	ㅜ	ㄷ		ㅠ	8	68	5.06	-	-	-
	ㄱ	ㅑ	ㅜ	ㄷ		ㅠ	ㄱ	0	0	0	-	-	-

단어	희상데이타							d^{rc}	p_0^{rc}	d^{rc}	d^{org}	p_0^{org}	d^{org}
욱성	ㅜ	ㄷ		ㅠ	ㄱ	ㅆ	ㅓ	4	75	-1.65	5	68	-0.6
	ㄷ		ㅠ	ㄱ	ㅆ	ㅓ	ㅇ	16	0	0	-	-	-
		ㅠ	ㄱ	ㅆ	ㅓ	ㅇ		2	0	0	-	-	-
	ㅠ	ㄱ	ㅆ	ㅓ	ㅇ		ㄱ	10	57	1.71	-	-	1.53
	ㄱ	ㅆ	ㅓ	ㅇ		ㄱ	ㅐ	11	77	3.25	-	-	-
패종	ㅓ	ㅇ		ㄱ	ㅐ	ㅈ	ㅗ	3	0	0	-	-	-
	ㅇ		ㄱ	ㅐ	ㅈ	ㅗ	ㅇ	11	64	-0.65	9	63	-0.24
		ㄱ	ㅐ	ㅈ	ㅗ	ㅇ		4	68	2.12	-	-	-
	ㄱ	ㅐ	ㅈ	ㅗ	ㅇ		ㄲ	13	57	1.54	7	61	0.77
	ㅐ	ㅈ	ㅗ	ㅇ		ㄲ	ㅞ	10	64	3.35	15	-	-
꽥꽥	ㅗ	ㅇ		ㄲ	ㅞ	ㄱ	ㄲ	3	0	0	-	-	-
	ㅇ		ㄲ	ㅞ	ㄱ	ㄲ	ㅞ	15	56	2.76	9	-	0.07
		ㄲ	ㅞ	ㄱ	ㄲ	ㅞ	ㄱ	14	0	0	-	-	-
	ㄲ	ㅞ	ㄱ	ㄲ	ㅞ	ㄱ		2	0	0	-	-	-
	ㅞ	ㄱ	ㄲ	ㅞ	ㄱ		ㅣ	10	56	2.26	17	55	2.16
	ㄱ	ㄲ	ㅞ	ㄱ		ㅣ	ㅑ	0	0	0	-	-	-
이야기	ㅞ	ㄱ		ㅣ	ㅑ	ㄱ	ㅣ	9	80	-2.18	-	-	0.32
	ㄱ		ㅣ	ㅑ	ㄱ	ㅣ		10	62	3.2	-	-	-
		ㅣ	ㅑ	ㄱ	ㅣ		ㄱ	4	66	1.5	-	-	-
	ㅣ	ㅑ	ㄱ	ㅣ		ㄱ	ㅏ	18	70	2.67	-	-	-
감기	ㄱ	ㅣ		ㄱ	ㅏ	ㅁ	ㄱ	3	0	0	-	-	-
	ㅣ		ㄱ	ㅏ	ㅁ	ㄱ	ㅣ	7	78	-1.46	-	-	-
		ㄱ	ㅏ	ㅁ	ㄱ	ㅣ		13	66	1.05	-	-	-
	ㄱ	ㅏ	ㅁ	ㄱ	ㅣ		ㄱ	2	65	3	-	-	-
	ㅏ	ㅁ	ㄱ	ㅣ		ㄱ	ㅏ	17	67	3.01	-	-	-
간마기	ㄱ	ㅣ		ㄱ	ㅏ	ㄴ	ㅁ	3	0	0	-	-	-
	ㅣ		ㄱ	ㅏ	ㄴ	ㅁ	ㅏ	3	80	1	-	-	-
		ㄱ	ㅏ	ㄴ	ㅁ	ㅁ	ㄱ	3	79	1.12	-	-	-
	ㄱ	ㅏ	ㄴ	ㅁ	ㅏ	ㄱ	ㅣ	9	69	1.23	-	-	-
	ㅏ	ㄴ	ㅁ	ㅏ	ㄱ	ㅣ		10	61	3.8	-	-	0.38
	ㄴ	ㅁ	ㅏ	ㄱ	ㅣ		ㄴ	2	65	2	-	-	-
	ㅁ	ㅏ	ㄱ	ㅣ		ㄴ	ㅏ	19	68	2.54	-	-	-
날개	ㄱ	ㅣ		ㄴ	ㅏ	ㄹ	ㄱ	2	80	-2	-	-	-
	ㅣ		ㄴ	ㅏ	ㄹ	ㄱ	ㅒ	9	76	-1.43	-	-	-
		ㄴ	ㅏ	ㄹ	ㄱ	ㅒ		6	62	0	-	-	-
	ㄴ	ㅏ	ㄹ	ㄱ	ㅒ		ㄴ	3	64	1.09	-	-	-
	ㅏ	ㄹ	ㄱ	ㅒ		ㄴ	ㅏ	16	59	4.30	-	-	-
남강	ㄱ	ㅒ		ㄴ	ㅏ	ㅁ	ㄱ	2	84	-3.5	-	-	-
	ㅒ		ㄴ	ㅏ	ㅁ	ㄱ	ㅏ	6	80	-2	-	-	-
		ㄴ	ㅏ	ㅁ	ㄱ	ㅏ	ㅇ	13	66	0.98	-	-	-
	ㄴ	ㅏ	ㅁ	ㄱ	ㅏ	ㅇ		2	62	-2	-	-	-
	ㅏ	ㅁ	ㄱ	ㅏ	ㅇ		ㄴ	10	57	1.78	-	-	-
	ㅁ	ㄱ	ㅏ	ㅇ		ㄴ	ㅏ	9	85	3.5	10	-	-
난간	ㅏ	ㅇ		ㄴ	ㅏ	ㄴ	ㄱ	2	78	0	-	-	-
	ㅇ		ㄴ	ㅏ	ㄴ	ㄱ	ㅏ	6	69	-1.74	-	79	-
		ㄴ	ㅏ	ㄴ	ㄱ	ㅏ	ㄴ	14	71	0.76	-	-	-
	ㄴ	ㅏ	ㄴ	ㄱ	ㅏ	ㄴ		2	64	-3	-	-	-
	ㅏ	ㄴ	ㄱ	ㅏ	ㄴ		ㄱ	9	58	1.03	-	-	-
	ㄴ	ㄱ	ㅏ	ㄴ		ㄱ	ㅘ	12	58	3.25	-	-	-

이상에서와 같이 463개의 고립단어를 사용하여 신경망을 학습시킨 후 실험결과를 검토하기 위해서, 학습에 참여하지 않은 50개의 데이터를 선정하여 회상시켰을 때 입력과 출력의 오차에 대하여 살펴보았다. 신경망의 학습정도는 훈련이 끝난 신경망의 오차로 설명할 수 있는데 각각의 오차는 31개의 출력 유니트 중 가장 큰 값으로 반응한 것 하나만을 '1'로 하고 나머지는 '0'으로 하여 이것을 원래의 값으로 변환 시킨 후 목적패턴과 비교하여 구하였다. 이러한 방법으로 구한 각음소의 지속시간, 첫피치주기, 기울기에 대한 파라미터 추정율은 그림 4-1에서 알 수 있듯이 학습데이타의 경우 각각 98.43%, 98.59%, 97.7%이었으며 회상데이타의 경우는 각각 97.34%, 95.45%, 96.3%이었다.

표 4-3에 나타난 회상 결과를 보면 무성음의 지속시간과 같은 경우는 거의 100% 올바른 결과를 보였으며 모음과 유성자음의 지속시간 역시 에러가 목표값과 불과 ±2 프레임 정도의 오차를 나타내어 이것을 감안한다면 모든 음소에 대한 지속시간의 추정율은 상당히 높은 것으로 평가할 수 있다. 또한 첫프레임의 피치주기도 ±0.5 msec 정도의 오차를 감안한다면 거의 정확한 출력 결과를 얻었다. 이는 신경망이 유사한 패턴에 대해서 바로 이웃해 있는 출력 유니트에 강하게 반응하기 때문인 것으로 생각된다. 그러나 피치변화곡선의 기울기의 경우는 상당히 많은 에러를 나타냈는데 이는 유사한 패턴에 대해 이웃한 유니트에 반응을 하면 이것이 기울기값 자체에 매우 큰 영향을 미치기 때문이다. 특히 ±부호에 대한 오차는 전혀 다른 변화곡선의 추정을 초래할 수 있는 요인이 될 수도 있다. 이러한 에러는 출력패턴의 부호화 방법이 적절하지 못한데서 기인하는 것으로 생각된다.

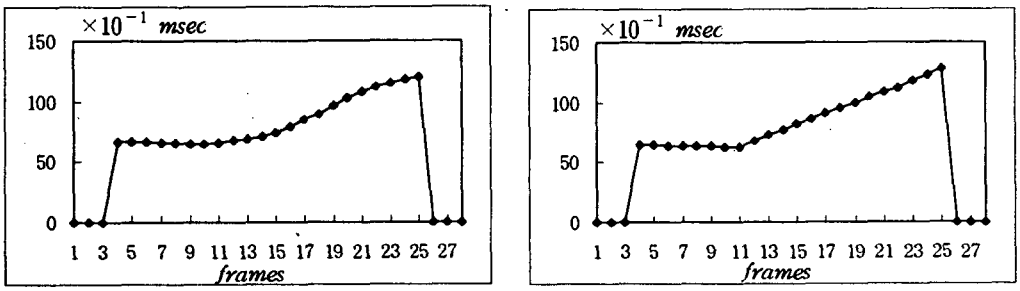
Ⅴ. 결과 및 검토

신경망이 출력한 각 음소에 대한 운율 파라미터를 사용하여 고립단어 전체에 대한 피치변화곡선을 생성시킬 수 있었다. 이것은 신경망의 출력 파라미터가 각각 지속시간, 기울기, 첫피치주기이므로 이 세개의 파라미터를 사용하여 피치변화곡선에 대한 1차 다항식을 구성함으로서 이루어 진다. 우선 고립단어내의 각 음소에 대한 피치변화곡선을 1차 다항식을 사용하여 추정한 다음 모든 음소의 추정된 변화곡선을 프레임축상에 음소순으로 나열시키면 고립단어 전체의 피치변화곡선을 구할 수 있다.

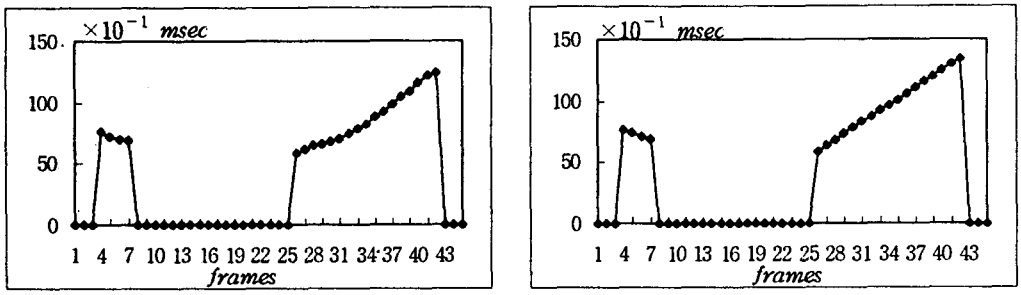
각 음소의 변화곡선 추정을 위한 1차 다항식은 지속시간, 기울기, 첫피치주기를 각각 d, a, p0라 한다면 각 프레임에서의 피치주기 P(f)는

$$P(f) = af + p_0 \qquad 단,\ f = 0, 1, 2, \cdots, d-1 \qquad (2)$$

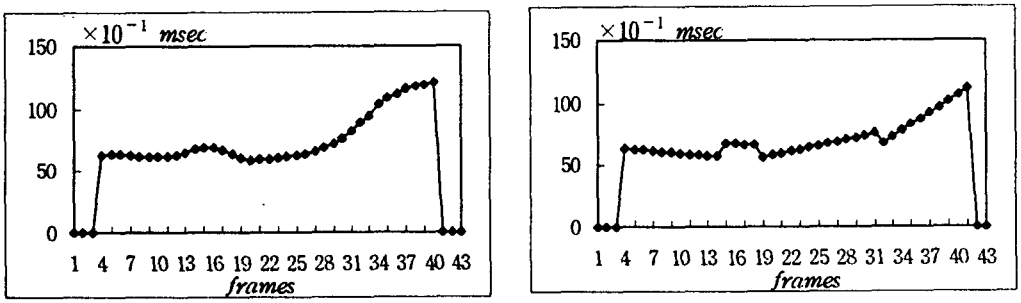
로 구할 수 있다.



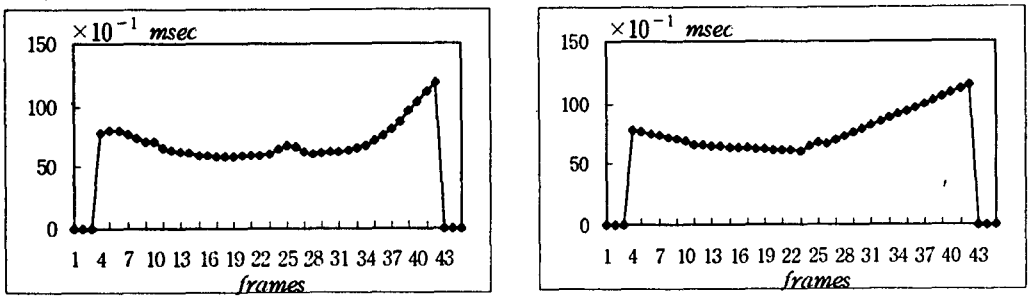
(a) 고립단어 /김/의 피치변화곡선



(b) 고립단어 /익따/의 피치변화곡선



(c) 고립단어 /괘종/의 피치변화곡선



(d) 고립단어 /남강/의 피치변화곡선

그림 5-1 원 피치변화곡선과 신경망에 의해 추정된 피치변화곡선

Fig. 5-1. Original pitch contour and extracted it by neural network

Ⅵ. 결론

각 고립단어의 원래의 피치변화곡선과 신경망을 사용하여 추정한 피치변화곡선을 그림 5-1에 나타내었다. 두 개의 변화곡선을 살펴 보면 지속시간의 경우는 거의 일치하나 추정된 변화곡선은 자연음으로부터 추출된 피치의 비선형적 변화를 제대로 추정하지 못함을 알 수 있다. 각 음소의 연결부분에서 피치의 점프현상이 두드러지는데 이것은 기울기와 첫피치주기에 대한 신경망의 출력에러에 기인하는 것으로 볼 수 있다. 이는 피치변화곡선에 대한 회귀분석을 3차 이상의 다항식으로 한다면 해결할 수 있을 것이다.

본 논문의 목적은 신경망을 사용하여 한국어 고립단어 내에서의 각 음소환경과 음절수에 따른 피치변화곡선을 추정하는 것이다. 이를 위해 본 논문에서는 1~4음절로 이루어진 513개의 한국어 고립단어를 음소별로 분리한 후, 이로부터 3800여개에 달하는 음소의 피치와 지속시간을 조사하였다. 그리고 여기서 얻어진 각각의 지속시간 동안의 피치변화를 회귀분석을 통한 1차 다항식으로 근사화하여 이 다항식의 계수인 기울기와 y절편(≒첫피치주기), 그리고 변수인 지속시간을 신경망의 목적패턴으로 하고 주변음소환경을 고려한 음소열을 입력패턴으로 함으로서 신경망이 각 음소의 주변환경에 따른 피치와 지속시간의 변화법칙을 스스로 학습하도록 하였다. 실험결과 전후 3음소의 음소환경을 포함하고 있는 3545개의 학습패턴(463단어)의 경우 신경망은 지속시간, 기울기, 첫피치주기에 대해 각각 98.43%, 98.59%, 97.7%의 추정율을 보였으며 학습에 참여하지 않은 251개의 희상패턴(50단어)의 경우 각각 97.34%, 95.45%, 96.3%라는 좋은 추정율을 나타내었다. 그리고 신경망의 출력에 의해 얻어진 세 개의 파라미터를 다시 1차 다항식의 계수와 변수로 사용하여 각 음소의 피치변화곡선을 구하고 이를 프레임축상에 순서대로 나열함으로서 고립단어 전체에 대한 피치변화곡선을 추정할 수 있었다.

추정된 변화곡선은 실제 자연음으로부터 추출된 변화곡선과 비교하여 볼 때 거의 유사함을 나타내었다. 그러나 원래의 변화곡선 자체를 1차 다항식으로 근사를 하였기 때문에 자연음에서 추출된 변화곡선의 비선형적인 요소를 제대로 추정할 수 없었다. 그리고 각 출력 파라미터에 대한 신경망의 출력에 큰 오차가 존재하는 경우에는 변화곡선을 제대로 추정할 수 없었다. 이러한 오차는 출력패턴 작성에도 문제점이 있지만 대부분의 오차가 희상데이타에 존재하는 것으로 보아 이는 부족한 학습데이타에 기인하는 것으로 생각된다. 그러므로 이러한 문제점을 해결하기 위해서는 우선 우리말 고립단어 내에 존재하는 거의 모든 음소환경에 대한 더 많은 학습데이타를 신경망에 학습시킴으로써 신경망이 더욱 정확히 변화법칙을 일반화 시킬 수 있게 함으로써 해결할 수 있을 것이다. 또한 자연음성에 존재하는 비선형적 피치의 변화를 살려주기 위해서는 1차 다항식이 아닌 3차 이상의 다항식을 사용하여 변화곡선을 근사시킴으로써 해결할 수 있을 것이다.

음성합성기술은 인간의 음향학적 정보 전달 수단인 음성을 기계가 소리의 합성을 통해서 가능하게 하는 기술로서, 이 기계에 의한 합성음은 올바른 정보 전달 능력으로서의 이해도와 인간의 발성과의 유사함을 나타내는 자연성으로 평가되어 진다. 이

러한 합성음의 이해도와 자연감에 피치가 미치는 영향은 매우 큰 것으로 평가되고 있다. 그래서 본 논문는 자연음성에 존재하는 무성자음, 모음, 유성자음의 단어 내에서 위치나 전후 음소배열 그리고 음절수에 따른 피치와 지속시간의 변화를 신경망으로 구현함으로써 합성음질을 개선시킬 수 있는 방법을 제시하였다.

<참고문헌>

[1] H. Witten, Principles of Computer Speech, Academic Press,1982.

[2] 허 준, "무제한 단어 한국어 음성합성 시스템에서의 운율정보 구현에 관한 연구", 서울대학교 석사학위논문, 1990.

[3] 임 운천, "한국어 법칙합성을 위한 운율법칙 구현에 관한 연구", 서울대학교 박사학위논문, 1991

[4] R. M. Meli and F. Fallside, "The modelling of F0 contours", IEEE Proc. ICASSP'82, pp.947-949, 1982.

[5] 최 운천외, "고품질의 한국어 문장음성변환 시스템", 제9회 음성통신 및 신호처리 워크샵 논문집. pp.193-196, 1992.

[6] 이 양희, "음성합성기술 개발의 현황과 과제", 제1회 음성학 학술대회 자료집, pp.145-154, 1994.

[7] 이 상억, "국어의 음운규칙", 제1회 음성학 학술대회 자료집, pp.45-53, 1994.

[8] H. Fujisaki and K. Hirose, "Comparison of acoustic features of word accent in English and Japanese", J.Acoust.Soc.Jpn(E)7.1, pp.57-63, 1986.

[9] Do-Heung Ko, "The nature of temporal relationship between adjacet segments in spoken Korean", Phonetica, vol.31, pp.259-273, 1975.

[10] 안 수길, "한국어 규칙합성에 관한 연구", 서울대학교 생산기술연구소 연구보고서, 1987.

[11] 김 종미, "자음의 단어내 음운환경별로 본 음가 변화", 한국음향학회 지, vol.13 no.5, pp.69-76, 1994.

[12] P. D. Wasserman, Neural computing theory and practice, Van Nostrand Reinhold, New York, 1989.

[13] R. P. Lippman, "An introduction to computing with neural nets", IEEE vol.ASSP-59 no.2, pp.4-22, 1987

[14] E. R. Kandel and J. H. Schwartz, Principles of neural science, Elsevier, New York, 1985.

[15] B. Gold, R. P. Lippmann and M. L. Manpass, "Some neural net recognition results on isolated words", Proc. IEEE Int. Conf. on Neural Networks, vol.4, pp.472-434, 1987.

[16] J. M. Zurada, Introduction to Artificial Neural System, West publishing company, 1992.

[17] D. E. Rumelhart and H. L. Mccleland, Parallel Distributed Processing, The MIT Press, Cambridge, Massachusetts London, England, 1986.

[18] T. J. Sejnowski and C. R. Rosenberg, "A parallel network that learns to read aloud", JHU/EECS Technical Report, 1986.