지각적 표현에 기초한 비음 인식에 관한 연구

김 기 철, 조 정 완
한국과학기술원 전산학과 컴퓨터구조연구실

Nasal Consonants Recognition Based on the Perceptual Representation

Ki Chul Kim and Jung Wan Cho
Computer Architecture Lab., Department of Computer Science, KAIST

요 약

음성 신호에는 언어정보이외에 여러 요인에 의한 정보가 포함되어 있어서, 문자와 일대일로 대응되는 분절을 정확하게 검출하기가 어렵다. 본 연구에서는 선형예측계수(LPC) 스펙트럼의 첨두 부분을 강조한 이진(binary) 스펙트럼을 제안하고, 이를 바탕으로 음의 안정영역과 천이영역을 통합하여 음향특징을 추출하고자 한다. 각 영역의 특징은 이진 스펙트럼을 누적하여 구하며, 통합적인 특징은 각 영역의 특징을 결합한 관계적 특징으로 나타낸다. 제 2차 포르만트 주파수의 궤적을 관계적 특징으로 하여, 양순비음과 치조 비음을 구별한 결과, 모음의 문맥과 화자에 비교적 독립적인 인식결과를 얻을 수 있었다. 또한 이진 스펙트럼이 원래의 스펙트럼에 포함된 정보를 유지하는지 검토하기 위해, 같은 거리척도(distance measure)에 의해 인식실험한 결과 이진 스펙트럼의 성능이 오히려 우수하게 나타났으며, 관계적 이진 스펙트럼의 경우 화자에 따른 변화가 더욱 적었다. 음성에 백색 잡음(Gaussian white noise)을 더하여 잡음음성(noisy speech)을 만든 뒤, 같은 방법으로 실험한 결과도 유사한 인식결과를 얻을 수 있어 제안된 이진 스펙트럼의 유효성을 확인하였다.

I. 서 론

음성인식은 음성신호 가운데 포함되어 있는 언어정보를 해독하는 과정으로서, 그 궁극적인 목적은 인간과 기계 사이에 자연스런 말에 의한 통신이 가능하도록 하는 것이다. 그러나 현재 개발된 음성인식 시스템들은 지극히 제한된 범위에서만 만족할만한 성능을 보이고 있다. 그것은 음성신호의 음향특징이 문맥 또는 화자에 따라 다양하게 변화할 뿐 아니라 이웃한 음사이에 중첩 분포되어 있어서, 연속된 음성신호를 문자언어의 단위에 따라 분할(segmentation)하고, 분할된 각 분절을 분류(labeling)하는 문제가 아직 완전하게 해결되지 않고 있기 때문이다 [1, 2, 3]. 또한 음성신호에는 언어 정보 이외에 화자의 개인적인 특성, 심리상태 및 환경 등의 비언어적인 정보까지 포함되어 있어서, 음성신호에 내포된 언어정보만 따로 추출해 내기가 쉽지 않다.

현재까지의 음성인식은 주로 통계적 처리와 함께 음운론이나 문법등의 지식을 제한적으로 이용하여 한정된 어휘의 인식이나 제한된 문형의 연속 음성을 대상으로 수행되었으나, 발음이 유사한 음이나 연속 음성을 인식하는 것은 아직도 어려운 문제로 남아 있다 [3]. 이에따라 최근 음성신호에 내재된 변화와 무관하게 지각이 이루어지는 인간의 청각 계통을 모형화하려는 시도가 이루어지고 있다. 청각 인지에 관한 연구는 크게 인간의 귀를 필터(filter) 모형으로 변환해보려는 말초 청각 계통(peripheral auditory system)에 관한 연구와, 음성의 판단에 관계되는 중추 청각 계통(central auditory system)에 관한 연구로 나누어 볼 수 있다.

말초 청각 계통에 관한 연구는 음성이나 유사한 신호에 대한 기저막(basillar membrane) 또는 헤어셀(haircell) 등의 응답 특성을 조사하는 생리학적인 연구와 [4], 임계대역(critical band), 마스킹(masking) 현상과 같은 말초 신경계통에서의 지각 특성을 조사하는 심리 음향학적 연구가 있다 [5, 6]. 말초신경 계통 모델에 관한 연구는 폐 활발하게 진행되고 있어 음성인식에도 이용되고 있으나, 중추 청각 계통에 관한 이론은 주로 청취 테스트에 의존하고 있어 음성인식에 직접 이용하기가 어려운 점이 있다.

본 연구에서는 최근의 음성지각 연구결과에 기초하여, 음성신호를 변화에 비교적 독립적이며, 지각적으로 의미있는 형태의 스펙트럼으로 변환하여 인식실험을 수행하였다. 파열음과 비음에 대한 일련의 지각연구에 따르면, 두가지 가정이 주장되고 있다 [7, 8, 9, 10]. 그 첫번째 가정은 화자나 음성학적인 문맥과 독립적이며, 음성학적인 분류기준과 일치하는 음향 패턴이 존재한다는 것이다. Kurowski와 Blumstein [7, 8]은 CV(consonant-vowel) 형태의 음절을 이용한 비음의 조음장소(place of articulation)를 구분하는 지각실험으로부터, 비음의 안정부분(nasal murmur)과 모음으로의 천이부분(nasal transition)이 각각 독립적인 특징으로 표현되는것이 아니라, 음성학적인 결정이 이루어지는 이전단계에서 하나의 특성을 가진 내부적 표현으로 청각의 말초신경 계통에 의해 통합된다고 유추함으로서 앞의 가정을 뒷받침하였다. 두번째 가정은 음성의 안정부분과 천이부분의 지각적인 통합이 청각의 말초부분에서 이루어 지는것이 아니라 중추계통에 의해 주도된다는 것이다. Repp [9, 10]은 /CV/ 또는 /VC/ 형태의 음절에 포함된 비음의 지각실험을 통해, 비음의 안정부분과

천이부분의 통합에 의해 조음장소가 지각됨을 보였으며, 그 통합과정이 중추계통의 판단에 의해 이루어진다고 주장하였다.

두가지 입장은 지각계통의 규명이라는 측면에서 보면 서로 상반되는 가정이지만, 자음의 안정부분과 모음으로의 천이부분이 통합되어 자음의 조음장소가 지각된다고 하는 점에서는 서로 일치하고 있다. 음성신호에는 안정부분(stable region)과 천이부분(transient region)이 번갈아 나타나는데, 모음이나 자음의 안정부분에서는 스펙트럼이 일정기간동안 안정된 특성을 가지며, 이중모음이나 자음과 모음사이의 천이부분에서는 스펙트럼의 특성이 점차적으로 또는 급격하게 변화하게 된다. 만일 스펙트럼의 안정부분과 천이부분을 제대로 통합한 특징을 추출할 수 있다면, 음향특징과 음성학적인 분류기준을 판련지을 수 있는 중간적인 표현형태로 볼 수 있을 것이다. 본 연구에서는 이같은 지각적 통합을 음성의 스펙트럼 통합이라는 차원에서 실현하여 비음인식에 적용하였다. 이 방법은 음향특징에 따른 음성신호의 분절이 바로 음성학적인 단위로 사상되는 것이 아니라, 음향적으로 안정된 부분과 천이부분이 통합되어 사상이 이루어지므로 분할에 따른 오류의 영향을 줄일 수 있을 뿐 아니라, 분할과정이 간단해진다는 장점이 있다.

이를 위해 일반적으로 사용되는 LPC(Linear Predictive Coefficients) 스펙트럼의 첨두부분을 강조한 이진 스펙트럼(binary spectrum)을 제안하고, 이를 바탕으로 음의 안정영역과 천이영역의 통합을 시도하였다. LPC 스펙트럼보다 단순화된 이진 스펙트럼을 제안한것은 단순히 계산량을 줄인다는 판점보다는, 일반적으로 음성학적인 정보가 많이 포함되어 있다고 알려진 스펙트럼의 첨두(peak) 부분을 강조함으로써 화자나 문맥에 따른 변화가 적은 표현을 추출하려는 목적에서이다 [11]. 안정영역과 천이영역은 세가지 형태의 동적 이진 스펙트럼으로부터 추출한 동적 특징에 의해 구분하였으며, 청각에서의 단구간 적응(*short-term adaptation*) 현상을 적용하여, 이진 스펙트럼을 누적하여 각 영역의 특징을 나타내는 스펙트럼을 구하였다. 두 영역을 통합한 특징은 각 영역에서 누적된 스펙트럼의 첨두 판계로부터 구해진다. /CV/ 음절의 비음을 조음장소에 따라 분류하기 위해서 제 2차 포르만트 주파수에 해당하는 스펙트럼의 첨두 변화를 판계적 특징으로 하여 실험한 결과, 모음의 문맥과 화자에 비교적 독립적인 인식결과를 얻을 수 있었다. 또한 이진 스펙트럼의 성능을 검토해보기 위해 현래의 LPC 스펙트럼과 같은 방법으로 각각 인식실험한 결과, 이진 스펙트럼의 성능이 오히려 우수하게 나타났으며, 안정영역과 천이영역의 통합으로 얻은 판계적 이진 스펙트럼의 경우 화자에 따른 변화가 더욱 적어, 제안된 방식의 타당성을 확인하였다.

일반적으로 비음은 음성 에너지의 전체적인 감소에 의해 그 유무가 검출되며, 제 2차 포르만트 주파수 궤적의 특성에 따라 조음장소에 따른 분류를 하게 된다 [12]. 그러나 비음은 구강(vocal tract)뿐 아니라 비강(nasal tract)을 통과하며 발음되기 때문에, 비강에 의한 반공명(antiresonance)주파수의 영향으로 포르만트 주파수의 추출이 어려워 그 분류가 쉽지 않다. 제안된 이진 스펙트럼을 이용할 경우 스펙트럼의 첨두위치의 변화가 두드러져 포르만트 주파수의 개괄적인 추이특성 검출이 용이해진다. 제 II장에서는 이진 스펙트럼의 생성및 그 통합과정을 서술하였으며 제 III장에서는 비음인식 실험 및 제안된 표현의 평가를 위한 실험과 분석 결과를 논의하였고, 제 IV장에서는 결론및 추후 연구 방향을 기술하였다.

II. 이진 스펙트럼과 특징 통합

2.1. 이진 스펙트럼의 생성

일반적으로 음성신호중에서 스펙트럼의 형태및 첨두부분이 음성정보를 많이 포함하고 있으며, 스펙트럼의 천이부분도 중요하다고 알려져 있다 [11, 12, 13]. 또한 스펙트럼의 첨두부분을 강조하여 제안된 거리척도들은 그렇지 않은 척거리도에 비해 좋은 성능을 보이고 있다 [11, 14]. 즉, 연속된 스펙트럼으로 볼 수 있는 음성신호중에서 음운정보는 스펙트럼의 첨두부분 및 임계값을 넘는 변화부분가운데 주로 분포되어 있다고 볼 수 있다. 이에따라 LPC 스펙트럼을 실무율 (all-or-none principle)을 적용하여 "0"과 "1" 상태를 갖는 이진 스펙트럼으로 변환하였다 [15]. 본 연구에서는 정적, 동적, 그리고 판계적인 세가지 형태의 이진 스펙트럼을 제안하였다 [16]. 정적 이진 스펙트럼(static binary spectum)은 스펙트럼의 제 2차 미분값을 각 주파수 대역별로 임계값에 따라 "0"과 "1" 상태로 클램핑시켜 구하였다. 각 첨두부분에서 기울기가 가장 큰 위치 이상의 첨두대역만 "1" 상태로 된다. 그림 1에서 (a)와 (b)는 /나/와 /마/의 음성파형이고, (b)와 (e)는 각각의 LPC 스펙트럼, (c)와 (f)는 LPC 스펙트럼으로부터 추출한 각각의 이진 스펙트럼이다. 그림에서와 같이 이진 스펙트럼은 포르만트 주파수에 해당하는 스펙트럼의 첨두위치 및 대역폭을 포함하게 된다.

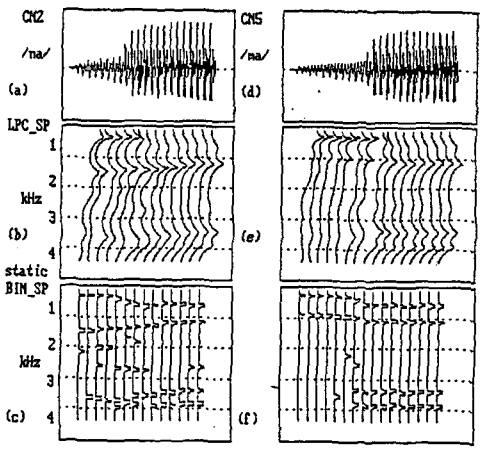

그림 1. /나/와 /마/의 음성파형과 LPC 스펙트럼 및 정적 이진 스펙트럼

2.2. 천이영역의 결정

천이영역은 스펙트럼의 변화값이 안정영역에서의 평균 변화량보다 큰 값을 가지는 영역으로 정의할 수 있다. 이를 위해 세가지 형태의 동적 이진 스펙트럼으로부터 스펙트럼의 변화 정도를 나타내는 동적 특징 (dynamic feature)을 구하였다. 동적 이진 스펙트럼을 추출한 근거는 스펙트럼의 변화가 일정한 반응 임계값을 넘을 때에 음운의 변화가 있을 것이라는 가정이다. 첫번째 형태의 동적 이진 스펙트럼은 인접한 정적 이진 스펙트럼의 변화를 나타낸다. 즉, 이진 스펙트럼간의 논리적 exclusive-OR 판계에 의해 구하며, 그 결과는 첨두 위치 및 대역의 변화를 나타낸다. 두번째 형태의 동적 이진 스펙트럼은 각 주파수 영역에서의 인접한 LPC 스펙트럼의 기울기의 차이를 나타내며, 각 LPC 스펙트럼을 미분한 스펙트럼으로부터 구한다. 즉, 인접한 LPC 스펙트럼의 미분값의 차이가 임계값 이상이 될 때에만 "1" 상태가 되도록 하였다. 세번째 형태의 동적 이진 스펙트럼은 인접한 LPC 스펙트럼간의 기울기 변화의 차이를

나타내며, 각 LPC 스펙트럼을 제 2차 미분한 스펙트럼으로부터 구한다. 즉, 인근 LPC 스펙트럼의 제 2차 미분값의 차이가 임계값을 넘을 때 "1" 상태가 되도록 하였으며, 이것은 스펙트럼 첨두 위치의 변화를 나타낸다.

동적 특징은 각 동적 이진 스펙트럼의 저역(0 ~ 1 kHz), 중역(1 ~ 2 kHz), 그리고 고역(2 ~ 3 kHz) 주파수 대역에서의 크기가 된다. 각 대역에서의 동적 특징들의 값이 미리 구한 임계값을 넘는 부분이 천이영역이 되며, 그 시작점과 끝점은 천이영역 검출 알고리즘에 의해 결정된다 [17]. 먼저 저역에서의 동적 특징의 값이 임계값을 넘으면 중역 및 고역에서의 특징값도 임계값을 넘는지 조사하면서, 해당 위치의 전후 부분을 검토하여 천이영역의 시작점을 결정한다. 천이영역의 끝점도 같은 방법으로 특징 값이 임계값보다 작은지 조사하면서 구하게 된다. 이때 각 대역의 임계값은 고정된 값으로 처음부터 주어지는 것이 아니라, 신호의 변화에 적응할 수 있도록 안정 또는 천이영역에서 누적된 동적 특징 값의 평균값을 임계값으로 정했다. 따라서, 천이영역의 시작점 또는 끝점이 정해지면 해당 위치의 동적 특징값이 새로운 임계값이 된다. 그림 2에 /나/와 /마/에 대한 정적 이진 스펙트럼 및 첫번째 형태의 동적 이진 스펙트럼과 각 대역에서의 동적 특징 값의 변화와 그로부터 구한 천이영역이 도시되었다. 최종적인 천이영역은 세가지의 동적 이진 스펙트럼으로부터 구한 세가지의 천이영역을 평균한 영역으로 확정된다. 천이영역의 시작점이 결정되면, 그 이전의 안정영역에 있는 스펙트럼이 통합되면서 천이영역의 끝점 검출이 시작된다.

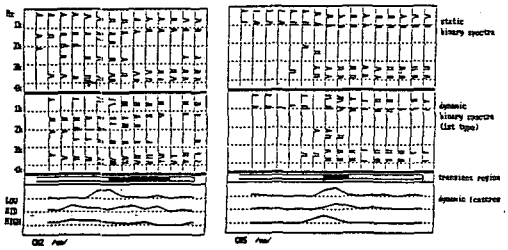

그림 2. /나/와 /마/에 대한 첫번째 형태의 동적 이진 스펙트럼 및 천이영역

2.3. 관계적 특징 (relational feature) 검출

음향학적인 천이영역은 많은 음성 정보를 포함하고 있으며, 특히 자음 인식에서는 천이영역의 포르만트 주파수 궤도가 중요한 단서가 되는 것으로 알려져 있다 [12, 13, 18]. 그러나 포르만트 주파수의 궤도를 추적하는 일은 쉽지가 않으며, 특히 비음같은 경우 반공명 주파수의 영향으로 첨두정보가 흐려지는 수가 있어 오류가 발생하기 쉽다. 본 연구는 정적 이진 스펙트럼을 각 영역에서 누적 통합한 관계적 이진 스펙트럼으로부터 안정영역과 천이영역을 판계지을 수 있는 특징을 추출할 수 있다는 가정에서 출발했으나, 실제로 두 영역의 관계를 적절하게 표현해주는 기준 특성을 결정하기란 쉽지가 않다. .여기서는 각 영역내에서 이진 스펙트럼을 누적하여, 즉, 논리적 OR 연산에 따라, 스펙트럼을 통합하였으며, 잘 알려진 특징인 제 2차 포르만트 주파수에 해당되는 첨두의 궤적 추이를 관계적 특징으로 사용하여 비음을 인식하였다. 그림 3은 /나/와 /마/에 대한 정적 이진 스펙트럼, 안정영역과 천이영역, 각 영역에서의 이진 스펙트럼 통합과정, 그리고 통합된 관계적 이진 스펙트럼을 각각 보여주고 있다.

III. 실험 및 분석

3.1. 비음인식 실험

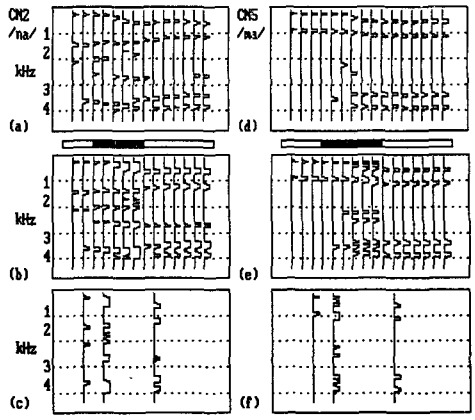

그림 3. /나/와 /마/의 이진 스펙트럼 통합

실험용 음성자료로 세명의 20대 한국인 남자에 의해 각각 세번씩 발음된 /나, 마, 너, 머, 노, 모, 누, 무/ 등의 /CV/ 음절을 사용하였으며, /나, 마/ 등은 두번씩 발음되어 모두 90개의 음절이 사용되었다. 음성신호는 방음장치가 된 녹음실에서 녹음되어 차단주파수가 4.7 kHz인 analog 저역 필터로 여과된 뒤, 10 kHz 주파수로 표본화되어 IBM PC/AT에서 처리되었다. 분석 방법 이외에 의한 오류를 최소화하기 위해서, 자음의 시작점부터 모음의 안정부분까지 파형 편집기에 의해 수작업으로 편집하였다. 먼저 특징 벡터를 평탄화시키기 위하여, 음성 데이타에 대해 0.95 비율로 프리엠퍼시스(preemphasis)를 취한 후 20 밀리초의 해밍 창함수(Hamming window)를 10 밀리초씩 이동시키며 얻은 각 프레임마다 14차 LPC 분석을 수행하였다. LPC 스펙트럼은 자기상관법에 따라 구한 14개의 LPC 계수들을 256 포인트 FFT(fast Fourier transform)를 적용하여 구하였으며, 크기에 따른 변화를 줄이기 위해 0과 1 사이로 정규화시켰다 [19].

정규화된 LPC 스펙트럼에 대해 앞 장에서 기술한 방법에 따라 정적 이진 스펙트럼을 구한 뒤, 동적 특징에 따라 안정영역과 천이영역을 구분하고 각 영역에서 이진 스펙트럼을 통합한다. 누적 결과 구해진 관계적 이진 스펙트럼을 17개의 임계 대역(critical band)으로 변환한 뒤 [6], 제 2차 포르만트 주파수에 해당되는 스펙트럼 첨두의 궤적 기울기를 추적하였다. /나/와 /마/에 대한 관계적 이진 스펙트럼 및 임계 대역으로 변환된 스펙트럼이 그림 4에 나타나 있다. 양순음 /m/과 치조음 /n/을 구분하기 위해, 궤적 기울기에 따른 휴리스틱 룰(heuristic rule)을 적용하였는데, 이 룰은 모음의 상황과는 무관하게 적용된다. 즉, 비음과 모음 사이의 천이부분에서, 치조음은 600 ~ 1000 Hz 사이에 위치한 제 2차 포르만트 주파수가 단조 감소 추세를 보이는 반면, 양순음 /m/은 900 ~ 1500 Hz 사이에 위치한 제 2차 포르만트 주파수가 증가 추세를 보이므로, 궤적의 가능한 기울기 형태에 따른 치조음 또는 양순음의 가능성을 계산한 뒤, 값이 더 큰 음을 인식 결과로 결정한다.

세명의 화자에 대해 실험한 결과, 92.1 %의 인식률을 얻을 수 있었으며, 치조음의 인식률이 양순음의 그것보다 6.6 % 더 높았다. 이 결과는 /VCV/ 상황의 비음을 인식하기 위해, 문맥에 따른 포르만트 주파수의 천이를 fuzzy 관계로 나타낸 De Mori [18]의 결과와 비교할 만하다. 그의 결과는 4명이 발음한 200개의 데이타에 대해 6 %의 오류율을 보여준다. Gubrynowicz 등[20]은 /CVCV/ 환경하의 비음을 인식하기 위해, 자음이 끝나는 부분에서 모음의 상황과

무관하게 스펙트럼의 전체적인 모양에 따른 틀을 만들어 적용하였다. 그의 예비실험 결과는 /m/, /n/, /gn/에 대해 평균 53.1 %인데, /m/과 /n/에 대해서는 대략 80 %에 이른다.

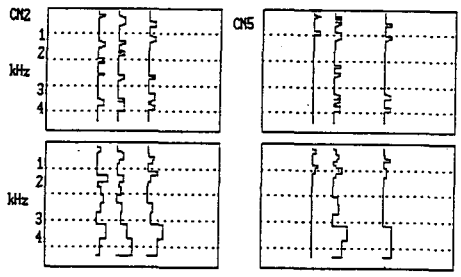

그림 4. /나/와 /마/에 대한 판계적 이진 스펙트럼의 임계 대역 변환

3.2. 이진 스펙트럼의 성능 평가

제안된 이진 스펙트럼의 성능 평가를 위해, LPC cepstral 계수에 의한 비음인식 결과와 비교하였다. 화자종속(speaker-dependent) 인식을 위해서, 각 사람의 3 회 발음중 2 회의 발음으로 번갈아가며 표준패턴을 만들고 나머지 발음으로 인식실험을 수행하였다. 화자독립(speaker-independent) 인식을 위해서도 3 사람중 번갈아 가며 2 사람의 각 2번씩의 발음을 이용하여 표준패턴을 만들고, 나머지 한사람의 발음으로 인식실험을 수행하였다. 각 발음마다 길이가 다르므로 DTW(dynamic time warping)에 의해 최소의 거리값을 갖는 경로에 따라 평균을 취한 표준패턴을 만들었다. 인식 대상 패턴과 비교할 때에도, DTW 경로에 따라 비교한 뒤 가장 유사도가 큰 표준 패턴을 인식 결과로 하였으며, 유사도 비교 방법은 각 계수에 대한 유클리디안 거리척도를 사용하였다 [21]. 표 1에 판계적 특징에 의해 비음을 인식한 결과와 LPC cepstral 계수를 이용한 인식결과가 비교되었다.

또한 이진 스펙트럼에 의한 음성정보의 손실 정도를 평가해 보기 위해, LPC cepstral 계수, LPC 스펙트럼, 정적 이진 스펙트럼, 판계적 이진 스펙트럼을 같은 유클리디안 거리척도에 의해 비음인식을 수행하였다. 그림 5에 화자종속 및 화자독립 비음 인식 인식결과가 나타나 있다. 각 경우 모두 원래의 LPC 스펙트럼보다 단순화된 이진 스펙트럼이 더 좋은 인식 결과를 보이고 있다. 따라서 이진 스펙트럼이 원래의 LPC 스펙트럼보다 음성정보를 훨씬 더 효과적으로 표현해준다고 볼 수 있다. 즉, 음성신호에는 많은 정보들이 중복되어 포함되어 있기 때문에, 음성정보만 포함된 스펙트럼 표현은 훨씬 간단하게 될 것이다. 다른 가능성은 유클리디안 거리척도가 음성정보뿐 아니라 LPC 스펙트럼에 포함된 다른 정보들도 같은 비중으로 비교해 주기 때문에, 비록 이진 스펙트럼이 LPC 스펙트럼에 포함된 음성정보를 어느정도 잃어버린다 해도 음운정보 이외의 다른 정보를 더 많이 포함한 LPC 스펙트럼의 성능이 떨어질 수 있다는 것이다 [22].

3.3. 백색잡음이 섞인 비음의 인식

잡음 환경하에서의 성능을 비교해보기 위해, 백색잡음(white noise)을 컴퓨터의 난수 발생기를 사용하여 Gaussian 분포가 되도록 한 뒤, 표본화된 음성과 더하여 잡음음성(noisy speech)을 만들었다. 이때 표본화된 음성은

표 1. 판계적 특징을 이용한 비음인식과 LPC cepstrum을 이용한 비음인식의 결과 비교

recognition method	/m/	/n/	Average
proposed method (speaker-independent)	88.9 %	95.5 %	92.1 %
speaker-dependent cepstral matching	77.8 %	88.9 %	83.3 %
speaker-independent cepstral matching	64.4 %	86.7 %	75.6 %

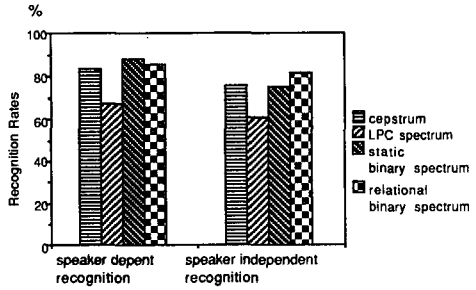

그림 5. 유클리디안 거리척도를 이용한 LPC cepstral 계수, LPC 스펙트럼, 정적 이진 스펙트럼, 그리고 판계적 이진 스펙트럼의 화자종속 및 화자독립 비음인식 결과

무잡음음성(clean speech)이라고 가정하였으며, -5 dB에서 30 dB까지 5 dB 간격의 SNR(signal to noise ratio) 조건으로 잡음음성을 만들었으며, SNR은 다음식으로 정의된다.

SNR (dB) = 20 * log (무잡음음성의 세기 / 잡음의 세기)

잡음을 더하기 이전의 음성이 4.7 kHz 이하로 여과되었기 때문에, 잡음음성도 digital 저역 필터에 의해 여과시킨 뒤 분석을 수행하였다. 그림 6에 잡음 음성 생성과정을 도시하였다.

정적 이진 스펙트럼 및 판계적 이진 스펙트럼을 LPC 스펙트럼과 비교하기 위해, 각 스펙트럼을 17 임계 대역으로 변환시켜 유클리디안 거리척도에 의해 비교하였다. 각 SNR 조건의 잡음음성을 인식하기 위한 표준 패턴은 무잡음음성을 이용하여 생성하였으며, 화자종속 /CV/ 음절인식 결과가 그림 7에 나타나 있다. 백색잡음 환경하에서도 이진 스펙트럼의 성능이 전체적으로 LPC 스펙트럼보다 더 좋은 것을 볼 수 있으며, 이것은 이진 스펙트럼이 음성의 중요 정보를 충분히 포함하고 있음을 보여준다.

그러나 사람은 거의 5 dB 정도까지의 잡음음성도 알아듣는다는 사실을 생각할 때, 더 정교한 형태의 표이 고려되어야 하며, 거리척도 또한 개선될 필요가 있다. 백색잡음은 전 주파수 대역에 분포하지만, 고주파 대역에 더 큰 영향을 미친다. 그림 7에서 판계적 이진 스펙트럼의 성능이 급격하게 저하된 것은 스펙트럼의 고주파 대역의 비중이 상대적으로 더 주어졌기 때문으로 보인다. 이를 검토하기 위해 스펙트럼중 3 kHz 미만의 성분만 비교하여 인식한 결과가 그림 8에 도시되었다. 또한 같은 방법으로 2 kHz 미만의 성분만 비교하여 인식한 결과가 그림 9에 도시되었다. 잡음음성에 대해 전체적으로 성능향상이 이루어지지만, 특히 판계적 이진 스펙트럼의 성능향상이 두드러진다. 따라서 각 소리의 특성에 따라 비교할 부분에 초점을 맞추는, 즉, 스펙트럼의 대역별로 다른 비중치를 갖는, 거리척도를 도입할 필요가 있음을 알 수 있다.

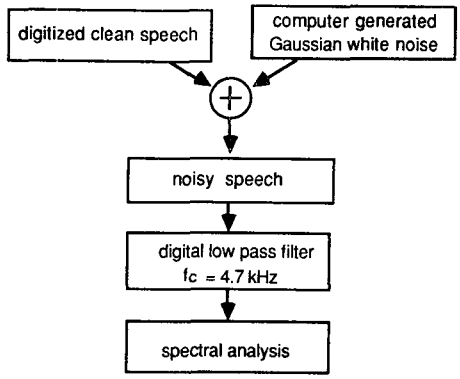

그림 6. 잡음음성(noisy speech) 생성과정

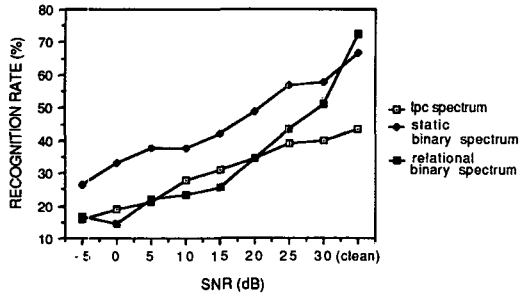

그림 7. 백색잡음(Gaussian white noise)이 섞인 /CV/ 음절인식 결과

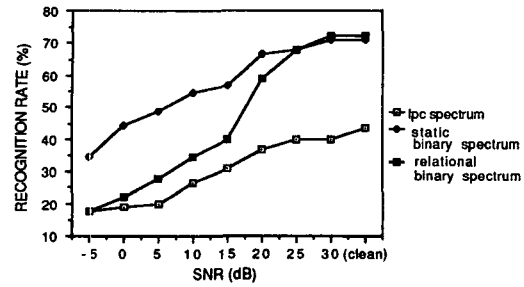

그림 8. 백색잡음(Gaussian white noise)이 섞인 /CV/ 음절인식 결과 (3 kHz 미만 성분 이용)

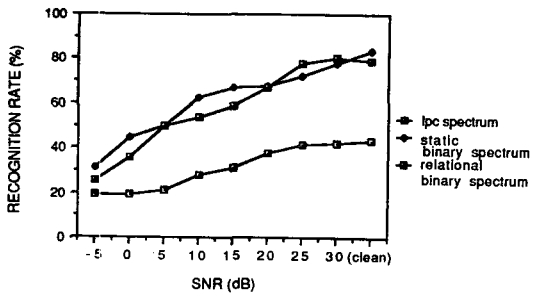

그림 9. 백색잡음(Gaussian white noise)이 섞인 /CV/ 음절인식 결과 (2 kHz 미만 성분 이용)

IV. 결 론

본 본문에서는 지각적인 표현에 기초한 이진 스펙트럼을 제안하고 비음인식에 적용하였다. 그 성능을 검토한 결과, LPC 스펙트럼이나 cepstrum보다 향상된 성능을 확인하였다. 정적 이진 스펙트럼은 아직 중복된 특징을 갖고 있으므로, 더 간결한 형태의 표현으로 변환할 수 있다. 이를 위해 정적 이진 스펙트럼을 안정영역과 천이영역에서 누적 통합하여 관계적 이진 스펙트럼을 얻었으며, 이로부터 두 영역의 통합적인 특징, 즉 비음인식의 경우엔 제 2차 포르만트 주파수의 궤적을 추출함으로써, 화자와 문맥에 비교적 독립적인 표현을 얻을 수 있었다. 또한 비음의 음성신호에 백색잡음을 추가하여 잡음음성을 만들어 인식실험을 수행한 결과에서도 역시 이진 스펙트럼의 성능이 더 안정되어 있음을 알 수 있었다. 즉, 이진 스펙트럼이 어느정도 지각적인 표현과 관련성이 있다고 볼 수 있다.

그러나 관계 이진 스펙트럼의 경우, 잡음이 증가함에 따라 그 성능이 급격히 하락하는데 이것은 이진 스펙트럼을 입계 대역으로 변환시키는 경우, 스펙트럼의 고주파 성분이 저주파 성분보다 더 가중되기 때문으로 보인다. 따라서 이를 보정해 줄 필요가 있으며, 나아가 스펙트럼내에서의 첨두의 상대적인 크기를 유지해 주는 표현으로 변환시켜줄 필요가 있다. 또한 각 소리의 특징에 따라 촛점을 맞춰 비교할 수 있는, 지각적인 처리에 바탕을 둔 거리척도에 대한 연구가 필요하며, 이를 위해서는 비음이외의 음, 예를 들면 파열음등에 대해 제안된 표현방법의 성능을 검토해야 한다. 제안된 분할 방식은 기존의 언어단위에 기초한 방식보다 오류 허용 한도가 크긴 하지만, 천이영역의 검출 자체는 중요하므로, 인접한 스펙트럼 뿐 아니라, 더 넓은 시간폭을 조사하도록 개선할 필요가 있다. 또한 관계적 특징 추출시 천이영역의 특징을 단일한 스펙트럼으로 통합시키는 것은, 천이영역의 특성을 충분히 나타내지 못한다고 볼 수 있으므로, 천이영역의 특징이 차지하는 비중이 더 조사되어야 한다.

참고 문헌

[1] Elman, J. L., and McClelland, J. L., "Exploiting Lawful Variability in the Speech Wave," in *Invariance and Variability in Speech Processes*, Perkell, J. S., and Klatt, D. H. (eds.), Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 360-385, 1986.

[2] Pisoni, D. B., and Luce, P. A., "Acoustic-Phonetic Representation in Word Recognition," in *Spoken Word Recognition*, Frauenfelder, U. H., and Tyler, L. K. (eds.), MIT Press, Cambridge, pp. 21-52, 1987.

[3] Vaissiere, J., "Speech Recognition: A Tutorial," in *Computer Speech Processing*, Fallside, F., and Woods, W. A. (eds.), Prentice-Hall, London, pp. 191-242, 1985.

[4] Delgutte, B., and Kiang, N. Y.S., "Speech Coding in The Auditory Nerve: I. Vowel-like Sounds," J. Acoust. Soc. Am., Vol. 75, pp. 866-878, 1984.

[5] Goldhor, R. S., "Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features," Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Mass. Inst. Tech., Cambridge, MA, 1985.

[6] Zwicker, E. and Terhards, E., "Analytically Expressions for Critical Bandwidth as a Function of Frequency," J. Acoust. Soc. Am., Vol. 68, pp. 1523-1525, 1980.

[7] Kurowski, K., and Blumstein, S. E., "Perceptual Integration of the Murmur and Formant Transitions for Place of Articulation in Nasal Consonants," J. Acoust. Soc. Am., Vol. 76, pp. 383-390, 1984.

[8] Kurowski, K., and Blumstein, S. E., "Acoustic Properties for Place of Articulations in Nasal Consonants," J. Acoust. Soc. Am., Vol. 81, pp. 1917-1927, 1987.

[9] Repp, B. H., "Perception of the [m]-[n] Distinction in CV Syllables," J. Acoust. Soc. Am., Vol. 79, pp. 1987-1999, 1986.

[10] Repp, B. H., "On the Possible Role of Auditory Short-term Adaptation in Perception of the Prevocalic [m]-[n] Contrast," J. Acoust. Soc. Am., Vol. 82, pp. 1525-1538, 1987.

[11] Klatt, D. H., "Prediction of Perceived Phonetic Distance From Critical-Band Spectra: A First Step," Proc. ICASSP-82, pp. 1278-1281, 1982.

[12] Borden, G. J., and Harris, K. S., *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*, Williams & Wilkins, Baltimore, 1980.

[13] Furui, S., "On the Role of Spectral Transitions for Speech Perception," J. Acoust. Soc. Am., Vol. 80, pp. 1016-1025, 1986.

[14] Hanson, B. A., and Wakida, H., "Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise," IEEE, Trans. *ASSP-35*, pp. 968-973, 1987.

[15] Goldsby, R. A., *Biology*, Harper and Row Publisher Inc., 1976.

[16] Kim, K. C., Lee, H. S., and Cho, J. W., "Phonetic Recognition Using Peak Weighted Binary Spectrum," Proc. ICASSP-89, pp. 330-333, 1989.

[17] Kim, K. C., Maeng, S. R., and Cho, J. W., "Application of Perceptual Integration to Speech Recognition: Some Preliminary Experiments in the Recognition of Nasal Consonants," CAL-TR-89-033, Computer Architecture Lab., Dept. of Computer Science, KAIST, Seoul, Mar. 1989.

[18] De Mori, R., *Computer Models of Speech Using Fuzzy Algorithms*, Plenum Press, New York, 1983.

[19] Markel, J. D., and Gray, A. H., *Linear Prediction of Speech*, Springer-Verlag, 1976.

[20] Gubrynowicz, R., Le Guennec, L., and Mercier, G., "Detection and Recognition of Nasal Consonants in Continuous Speech - Preliminary Results," in *New Systems and Architectures for Automatic Speech Recognition and Synthesis: NATO ASI Series F. Vol. 16*, De Mori, R., and Suen, C. Y. (eds.), Springer-Verlag, pp. 613-628, 1985.

[21] Sakoe, H., and Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE, Trans. *ASSP-26*, No. 1, 1978, pp. 43-49, 1978.

[22] Gevins, A. S., and Morgan, N. H., "Ignorance-Based Systems," Proc. ICASSP-84, Vol. 3, Paper 39A.5/1-4, 1984.