한국음향학회지 제24권 제7호 pp. 416-422 (2005)

디지털 통신 시스템에서의 음성 인식 성능 향상을 위한 전처리 기술

Pre-Processing for Performance Enhancement of Speech Recognition in Digital Communication Systems

서 진 호*, 박 호 종*
(Jinho Seo*, Hochong Park*)

*광운대학교 전자공학과

(접수일자: 2005년 6월 30일; 수정일자: 8월 31일; 채택일자: 2005년 9월 15일)

디지털 통신 시스템에서의 음성 인식은 음성 부호화기에 의한 음성 신호의 왜곡으로 인하여 성능이 크게 저하된다. 본 논문에서는 음성 부호화기에 의한 스펙트럼 왜곡을 분석하고 왜곡된 주파수 정보를 보상하는 전처리 과정을 통하여 음성 인식 성능을 향상시키는 방법을 제안한다. 현재 널리 사용되는 표준 음성 부호화기인 IS-127 EVRC, ITU G.729 CS-ACELP, IS-96 QCELP를 사용하여 부호화에 의한 왜곡을 분석하고, 모든 음성 부호화기에 공통으로 적용하여 왜곡을 보상할 수 있는 전처리 방법을 개발하였다. 본 논문에서 제안하는 왜곡 보상 방법을 세 종류의 음성 부호화기에 각각 적용하였으며, 왜곡된 음성 신호에 대한 음성 인식률에 비하여 최대 15.6%의 인식률 향상을 얻을 수 있었다.

핵심용어: 음성 인식, 음성 부호화기, 디지털 음성 통신, 스펙트럼 보상

투고분야: 음성처리 분야 (2.5)

Speech recognition in digital communication systems has very low performance due to the spectral distortion caused by speech codecs. In this paper, the spectral distortion by speech codecs is analyzed and a pre-processing method which compensates for the spectral distortion is proposed for performance enhancement of speech recognition. Three standard speech codecs, IS-127 EVRC, ITU G.729 CS-ACELP and IS-96 QCELP, are considered for algorithm development and evaluation, and a single method which can be applied commonly to all codecs is developed. The performance of the proposed method is evaluated for three codecs, and by using the speech features extracted from the compensated spectrum, the recognition rate is improved by the maximum of 15.6% compared with that using the degraded speech features.

Keywords: speech recognition, speech codec, digital speech communication, spectral compensation

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

최근 이동 통신과 VoIP 등의 디지털 음성 통신이 급속히 보급되고 음성으로 동작하는 새로운 서비스에 대한 요구가 증가함에 따라 디지털 음성 통신 시스템에서의 음성 인식 서비스의 필요성이 강하게 나타나고 있다 [1,2]. 디지털 통신에서 소형 단말기는 크기와 계산 능력에 제약점을 가지지만 통신 서버 시스템은 메모리, 계산량 등에서 다양한 기능과 우수한 성능을 제공할 수 있으므로 음성 인식 서비스는 서버에서 제공되는 것이 효율적이며, 이를 구현하기 위한 시스템 구조 연구가 표준화 기구를 중심으로 널리 진행 중이다. 예로, IP 기반의 코어 (core) 네트워크에서 멀티미디어 기능을 정의하는 IP Multimedia Subsystem (IMS)은 음성 인식을 포함하는 모든 멀티미디어 기능을 통합된 서버인 Multimedia Resource Function (MRF)에서 담당하도록 구조가 설계되어 있다[3].

책임저자: 박 호 종 (hcpark@mail.kw.ac.kr)
서울시 노원구 월계동 447-1 광운대학교 전자공학과
(전화: 02-940-5104; 팩스: 02-913-9057)

그러나, 디지털 통신 시스템에서 통신 전송률을 감소시키기 위하여 사용하는 음성 부호화기 (speech codec)에 의하여 음성 신호의 특성이 심하게 왜곡되고, 서버 기반의 음성 인식기에 이와 같이 왜곡된 음성 신호가 그대로 입력되면 음성 인식 성능이 매우 저하된다. 이와 같은 통신 시스템에서의 음성 인식 문제점을 해결하기 위한 기존의 방법으로 단말기에서 음성 파라미터를 직접 추출하여 서버에 전달하는 distributed speech recognition (DSR) 방법, 비트 스트림 (bitstream)에서 음성 파라미터를 직접 추출하여 사용하는 방법, 음성 부호화기에 의하여 왜곡된 신호로 인식 엔진을 직접 훈련시키는 방법 등이 있다[2]. 그러나 DSR 방법은 단말기에서 파라미터를 추출하여야 하는 제약점을 가지고, 나머지 방법들은 디지털 통신에 포함되는 각각의 음성 부호화기마다 서로 다른 인식 엔진이 필요하고, 만일 새로운 음성 부호화기가 통신 시스템에 포함되면 새로운 음성 인식 모델을 다시 생성하여야 하는 문제점을 가진다. 따라서 이와 같은 기존 방법의 문제점을 해결하기 위하여 재훈련이 필요 없고 기존의 음성 인식기에 간단히 적용할 수 있는 서버에서의 새로운 음성 인식 향상 방법이 요구된다.

본 논문에서는 디지털 통신 시스템에서의 서버 기반 음성 인식 성능 향상을 위한 새로운 방법으로서 음성 인식 전단 (front-end)에 적용되는 전처리 (pre-processing) 기술을 제안한다. 전처리 기술은 입력 신호에 직접 적용되어 음성 부호화기에 의하여 왜곡된 주파수 정보를 보상하며, 특히 입력 신호의 특성이 시간에 따라 변하는 것에 적응하여 프레임별로 서로 다른 보상 규칙을 결정한다. 인식을 위한 음성 파라미터는 왜곡이 보상된 주파수 정보로부터 추출되며, 이렇게 추출된 음성 파라미터는 기존의 음성 인식 엔진에 입력되어 음성 인식이 이루어진다. 이와 같은 방법을 통하여 새로운 인식기 설계와 재훈련 없이 매우 간단하게 기존 음성 인식기의 음성 인식 성능을 향상시킬 수 있다. 특히, 제안하는 전처리 왜곡 보상 방법은 특정한 음성 부호화기에 제한적으로 적용되는 것이 아니라, 현재 통신 시스템에서 가장 널리 사용되는 Code-Excited Linear Prediction (CELP) 구조의 모든 음성 부호화기에 공통으로 적용이 가능한 장점을 가진다. 따라서 제안하는 전처리 왜곡 보상 방법은 디지털 통신 시스템에 이미 설치되어 있는 음성 인식기에 매우 간단히 적용될 수 있다.

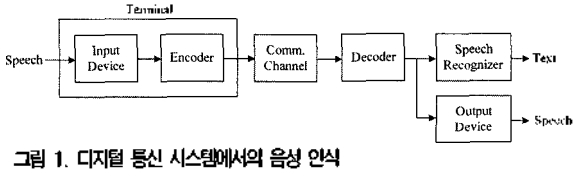

그림 1. 디지털 통신 시스템에서의 음성 인식
Fig. 1 Speech recognition in digital communication systems.

본 논문은 다음과 같이 구성된다. 2장에서 디지털 통신 시스템에서의 음성 인식기 동작을 간단히 설명하고, 3장에서 음성 인식의 관점에서 음성 부호화기에 의한 신호의 왜곡을 분석한다. 4장에서 제안하는 왜곡 보상 방법을 상세히 설명하고 5장에서 제안한 방법의 성능을 측정한다.

II. 디지털 통신 시스템에서의 음성 인식

지금까지 주로 사용되고 있는 stand-alone 형태의 음성 인식기는 최초 발성된 음성 신호가 직접 음성 인식기에 입력되는 형태로 구성되고, 이 경우 마이크 특성에 의한 변화 이외에 추가 왜곡이 발생하지 않고 원 음성 신호로부터 음성 인식이 이루어진다. 그러나 디지털 통신 시스템에 서버 기반 음성 인식기를 적용할 경우 그림 1과 같이 단말기에서 최초 발성된 음성 신호가 통신 경로를 따라 많은 과정을 거치면서 신호에 왜곡이 발생하므로 원 음성 신호가 직접 입력되는 경우에 비하여 음성 인식 성능이 크게 저하된다.

그림 1과 같이 디지털 통신 환경에 음성 인식 서비스를 적용할 경우 신호에 왜곡이 발생하는 두 가지 대표적인 이유는 통신 채널에서 발생하는 전송 오류와 음성 신호의 전송 효율을 높이기 위하여 사용하는 음성 부호화기이다. 채널에서의 전송 오류는 통신 환경에 따라 매우 불규칙적으로 발생하며 정보 왜곡이 발생하는 위치는 오류가 발생하는 위치에 의하여 결정되며, 시간적으로 왜곡의 정도가 변한다. 반면, 음성 부호화기는 정보량을

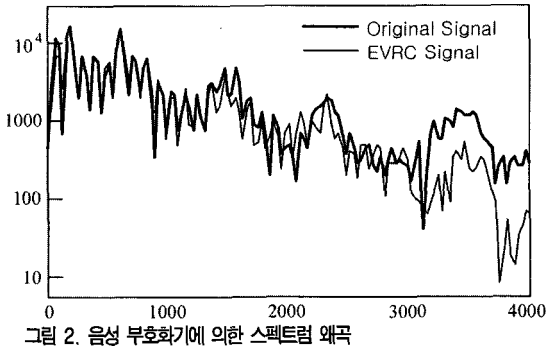

그림 2. 음성 부호화기에 의한 스펙트럼 왜곡
Fig. 2 Spectral distortion by speech codec.

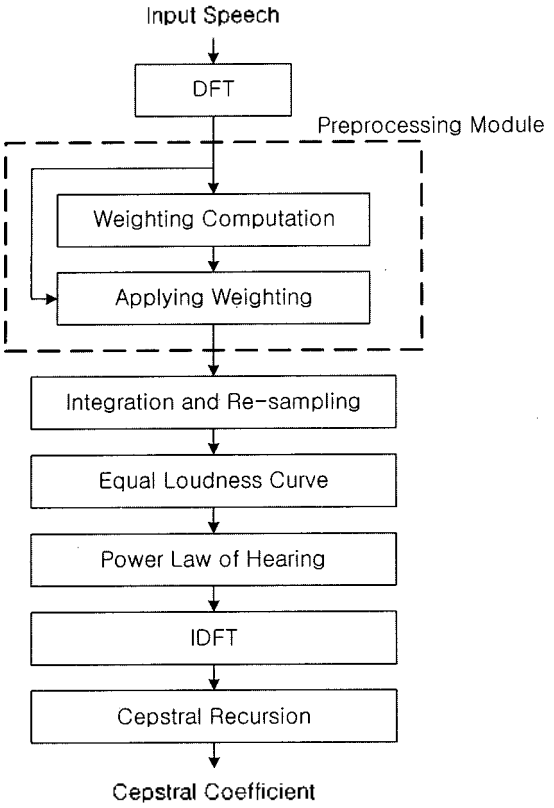

그림 3. 제안하는 전처리 모듈을 포함하는 PLP 계산 과정
Fig. 3 PLP process including the proposed pre-processing module.

줄이기 위하여 최초 입력에 대하여 정보의 손실을 포함하며 음성 특성에 왜곡을 유발시킨다. 또한 음성 부호화기는 신호에 계속적으로 적용되므로 시간에 따라 왜곡 유무가 결정되는 것이 아니라 모든 시간에 걸쳐 왜곡이 발생하고 음성 신호의 특성에 따라 왜곡의 현상이 변한다. 더욱이 통신 시스템에 사용되는 부호화기의 종류에 따라 왜곡의 형태에 차이가 발생하고 그에 따라 음성 인식기의 성능에도 차이를 나타낸다. 그러나 음성 부호화기의 동작은 미리 정해진 방식에 따라 진행되므로 어느 정도 왜곡 현상의 분석이 가능하고 이에 따라 왜곡을 보상하는 전략이 가능하게 된다. 본 논문에서는 두 가지 문제점 중 음성 부호화기에 의한 음성 신호 왜곡을 보상하여 음성 인식 성능을 향상시키는 문제만을 다룬다.

III. 음성 부호화기에 의한 왜곡 분석

디지털 통신 시스템에서의 음성 인식 향상 방법의 개발을 위하여 우선 음성 부호화기에 의한 왜곡의 특성을 분석하고 그에 따라 왜곡된 정보의 보상 방법을 개발하도록 한다. 그림 2는 원 음성 신호와 IS-127 Enhanced Variable-Rate Codec (EVRC) 음성 부호화기를 통과하여 합성된 음성 신호의 스펙트럼을 비교한 것이다[4]. 두 신호의 스펙트럼 모양이 매우 유사하며 스펙트럼 envelope이 일치하여 청취 음질에 큰 차이를 나타내지는 않지만 스펙트럼 크기가 작은 부분에서 스펙트럼 왜곡이 발생하고 특히 고주파 대역으로 갈수록 원 음성 신호와의 차이가 더 크게 나타나는 것을 볼 수 있다. 이와 같은 스펙트럼의 변형은 CELP 구조의 음성 부호화기에서의 포먼트 영역에 대한 강한 가중치 적용, 파라미터 검색 방법, 후처리 필터에서의 포먼트 및 피치 강조 동작 등에 의하여 발생하는 것으로서 CELP 음성 부호화기의 대표적인 특성이며, 음성 부호화기의 관점에서는 큰 문제가 되지 않고 청각적으로 매우 유사한 음성 신호를 합성할 수 있다. 그러나 음성 인식의 관점에서 보면 특정 주파수 정보가 왜곡되고 특히 고주파 성분의 왜곡이 크게 발생하여 음성 인식 모델을 왜곡시키고, 결국 음성 인식기에서 훈련된 스펙트럼 모델과 실제 입력되는 신호의 스펙트럼 사이에 나타나는 특성 차이에 의하여 음성 인식 성능이 크게 저하된다.

이 문제에 대한 기존의 해결 방법 중 하나는 음성 부호화기에 의하여 왜곡된 음성 신호의 정보를 인식 모델에 적용하는 방법이다. 즉, 음성 부호화기에 의하여 왜곡된 음성 신호를 이용하여 음성 인식기를 직접 훈련하는 것으로서 음성 부호화기에 의하여 저하되었던 인식률을 크게 향상시킬 수 있다. 그러나 각 음성 부호화기에 의한 데이터에 의존적이라는 약점을 가지고 있고 음성 부호화기가 변경되거나 새로운 표준 음성 부호화기가 통신 시스템에 적용되면 다시 음성 인식기를 훈련시켜야 하는 단점을 가진다. 더욱이, 이 방법이 정상적으로 동작하기 위하여 음성 인식기는 통신 시스템에 사용된 음성 부호화기의 종류를 알고 그에 따라 적절한 인식 모델을 사용하여야 하며, 이는 통신 시스템에서 음성 부호화기와 음성 인식기가 일체형 시스템에 포함되어 내부 정보를 공유할 수 있어야 가능하다. 그렇지 않을 경우 음성 인식기가 부호화기 정보를 전송받기 어려우며, 이 경우에는 통신 시스템에서 사용하는 모든 음성 부호화기를 이용하여 음성 인식기를 훈련시켜야 하며 각각의 독립적인 훈련에 비하여 음성 인식기의 성능은 저하된다. 따라서 이와 같은 문제점을 해결하기 위하여 각각의 음성 부호화기에 종속되지 않고 음성 부호화기의 종류에 관계없이 공통으로 적용 가능한 새로운 음성 인식 성능 향상

방법의 개발이 필요하며, 본 논문에서는 이와 같은 조건을 만족하는 스펙트럼 왜곡 보상 전처리 기술을 개발한다.

IV. 제안하는 전처리 기술

본 논문에서 제안하는 전처리 기술은 주파수 영역에서의 가중치 함수 적용을 통하여 스펙트럼 왜곡을 보상하는 방법이다. 그림 3과 같이 PLP 과정의 DFT 단계에서 구한 스펙트럼 계수로부터 제안한 방법에 따라 가중치 함수를 구하고, 이 가중치 함수를 원 스펙트럼 계수에 적용하여 변형된 스펙트럼 계수를 구하고, 이로부터 최종 캡스트럼 계수를 구하는 과정을 거친다. 그리고 본 논문에서는 디지털 통신 시스템에 가장 널리 사용되고 있는 세 가지 표준 음성 부호화기인 IS-127 EVRC, ITU G.729 CS-ACELP, IS-96 QCELP 부호화기를 사용하여 문제점을 분석하고 성능 향상 방법을 개발하며 최종적으로 개발된 방법의 성능을 측정 한다[4-6].

음성 부호화기에 의하여 왜곡된 음성 신호의 특징적 현상을 분석하여 다음 네 가지의 왜곡 보상 과정으로 구성된 전처리 기술을 개발 하였다.

(1) 스펙트럼 크기 정렬에 의한 보상
(2) 고주파 영역에서의 예외 처리
(3) 시간에 따른 점진적 가중치 적용
(4) 대역별 에너지 보상

각 과정의 결과물은 스펙트럼에 적용할 가중치 함수이며, 첫 과정에서 구한 가중치 함수를 시작으로 각 과정을 거치면서 가중치 함수의 변경을 통하여 최종 가중치 함수를 구하고, 마지막으로 구한 가중치 함수를 원 스펙트럼 계수에 적용하여 캡스트럼 계수를 구하게 된다. 첫 번째 과정은 왜곡된 신호의 스펙트럼을 크기 순서로 정렬하고 미리 준비된 중간 가중치 함수를 곱하는 과정을 통하여 스펙트럼 왜곡을 위한 가중치 함수를 구한다[7]. 이 방법은 음성 부호화기에 의하여 주로 크기가 작은 주파수 영역에서 많은 왜곡이 발생하는 현상을 활용하여 문제점을 해결하기 위한 것이며, 그림 4 (a)와 같이 스펙트럼을 크기 순으로 정렬하여 정렬된 스펙트럼을 구하고, 여기에 (b)의 중간 가중치 함수를 곱하여 작은 크기의 스펙트럼 영역에 큰 가중치가 적용되도록 하여 스펙트럼 왜곡을 보상하는 것이다. 이 방법에서 중간 가중치 함수를 정하는 것이 핵심이며, [7]에서 사용하였던 고정 가중치 함수의 문제점을 해결하기 위하여 가변 가중치 함수를 사용하며, 특히 가중치 함수에 의하여 실제로 크기가 변하게 되는 영역을 전체 스펙트럼의 크기를 분석하여 정하도록 한다. 즉, 스펙트럼의 전체 평균을 구하고 정렬된 스펙트럼에서 전체 스펙트럼의 평균값 이하 영역에만 완만한 가중치 함수를 적용하고 나머지 영역에는 가중치 1.0을 적용하도록 한다.

이와 같이 스펙트럼 정렬 후에 중간 가중치를 적용함으로써 매 프레임 마다 변하는 스펙트럼의 특성에 따라 가변적으로 서로 다른 가중치 함수가 적용되는 효과를 얻을 수 있다. 즉, 그림 4 (b)에 주어진 하나의 가중치 함수만을 이용하여 매 프레임마다 서로 다른 가중치 함

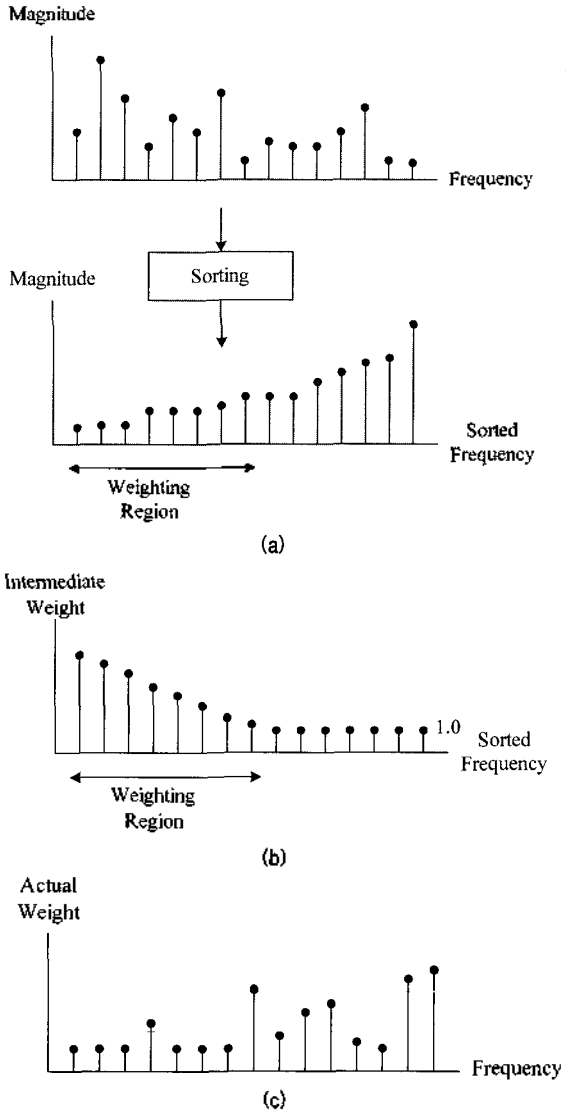

그림 4. 스펙트럼 정렬 및 가중치 함수 적용을 통한 스펙트럼 보상. (a) 스펙트럼 크기 순 정렬, (b) 중간 가중치 함수, (c) 실제 적용된 가중치 함수

Fig. 4 Spectral distortion compensation by spectrum sorting and weighting. (a) Spectral magnitude sorting, (b) Intermediate weighting function, (c) Actual weighting function.

수를 적용하는 효과를 얻을 수 있으며, 그림 4에서 입력 스펙트럼에 적용되는 실제 가중치 함수는 (c)의 가중치 함수가 되고, 이것이 첫 단계의 결과물이다.

두 번째 과정은 고주파 영역의 예외 보상 처리이다. 그림 2에서 볼 수 있듯이 왜곡된 신호는 고주파 영역으로 갈수록 그 정도가 심해진다. 이러한 고주파 영역에서의 왜곡 현상은 거의 대부분의 음성 신호에서 발견되므로 고주파 영역의 스펙트럼은 항상 가중치 함수의 적용을 받아야 한다. 그러나 첫 번째 가중치 함수 결정 과정에서 고주파 영역의 스펙트럼 보다 더 작은 스펙트럼이 많이 존재하는 경우 고주파 영역의 스펙트럼이 가중치를 받는 순위에 들지 못 하거나 또는 매우 작은 가중치를 가지게 되는 경우가 생긴다. 이 상황에 대한 예외 처리를 두어 첫 단계에서 매우 작은 가중치를 적용 받은 고주파 영역에 대하여 첫 과정에서 얻은 가중치 함수를 증가시켜 변형된 가중치 함수를 새로 결정하며, 다음의 단계로 진행된다. (i) 첫 과정에서 결정된 가중치 함수의 고주파 영역에서 가중치 값과 기준값 비교, (ii) 기준값 보다 작은 가중치 영역 선택, (iii) 선택된 가중치 영역에 대하여 증가된 가중치 적용, (iv) 새로운 가중치 함수 확정. 그림 5(a)와 (b)에 두 번째 과정의 적용에 대한 예가 주어졌다. (a)는 첫 과정에서 구한 가중치 함수이며 여기서 고주파 대역의 작은 크기의 가중치 영역을 예외 처리 영역을 선택하고 (b)에서는 (a)에서 선택된 영역의 가중

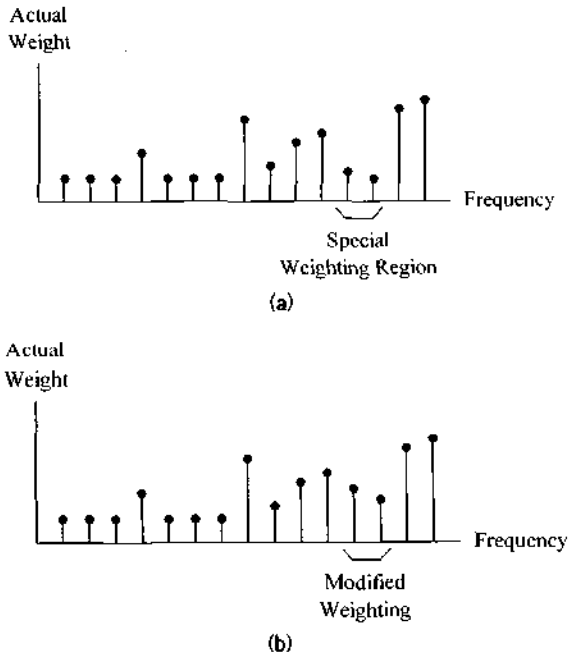

그림 5. 고주파 영역의 예외 처리. (a) 예외 처리 주파수 영역. (b) 변경된 가중치 함수

Fig. 5 Special weighting for high frequency region. (a) Special weighting region. (b) Modified weighting function.

표 1. 프레임 에너지 사이의 상관 계수

Table 1. Correlation coefficients between adjacent frame energy.

	Frame N and N-2	Frame N and N-1
'청와대'	0.9307	0.9739
'그야말로'	0.9253	0.9684
'컴퓨터'	0.9118	0.9685
'러시아'	0.9297	0.9736

치 크기를 증가시킨 새로운 가중치 함수를 얻는다.

세 번째 과정은 시간에 따른 점진적 가중치 적용(temporal smoothing)이다. 본 논문에서 사용하는 음성 인식기에서 각 프레임의 길이는 25msec이고 현재의 프레임 N과 프레임 N-1은 15msec 중첩되며 프레임 N-2과는 5msec 중첩된다. 중첩 된 영역에서 음성의 특성은 높은 상관도를 가지며 중첩 영역이 클 수로 상관도도 증가하게 될 것이다. 따라서 현재 프레임의 가중치 함수를 적용 할 때 상관도가 높은 이전 프레임들의 가중치 정보를 동시에 이용하면 성능을 더 높일 수 있다. 이 방법을 적용하기 위하여 우선 시간 영역에서 중첩되는 프레임 사이의 특성에 높은 상관도가 존재하는지 확인하였다. 표 1은 각기 다른 화자의 단어 발성 중에서 단어를 무작위로 뽑아 프레임 에너지의 상관 계수를 계산한 결과이며, 예상대로 시간 영역에서 중첩 되는 프레임들 사이에 상당히 높은 상관도가 존재하는 것을 확인할 수 있다. 따라서 이를 토대로 시간에 따른 점진적 가중치 적용 방법을 적용 할 수 있다.

시간에 따른 점진적 가중치 적용 방법에 의하여 현재 프레임의 가중치 함수는 이전 프레임에서 사용하였던 가중치 함수에 따라 변형되어 최종 결정된다. 즉, 현재 프레임에 대하여 앞의 두 단계에 의하여 가중치 함수를 우선 구하고, 이전 프레임의 가중치 함수를 고려하여 추가 변형을 적용한다. 가중치 함수에 대한 추가 변형은 크게 두 가지 상황으로 나누어진다. 첫 번째는 중첩 되는 이전 프레임에서 가중치를 받은 주파수 영역이 현재 프레임에서 가중치를 받지 못 한 경우이며, 이 때에는 현재 프레임의 해당 주파수 영역에 대하여 이전 프레임 가중치의 35%를 추가로 적용한다. 두 번째는, 중첩되는 이전 프레임에서는 가중치를 받지 못하였으나 현재 프레임에서 가중치를 받은 경우이며, 이 때에는 해당 주파수 영역에 대하여 현재 프레임의 가중치를 35% 감소시켜, 새로운 가중치 함수를 최종 생성한다.

네 번째 과정은 급격한 에너지 변화에 대한 보상 처리이다. 음성 부호화기에 의하여 왜곡된 음성 신호는 위에서 언급하였듯이 원 음성 신호보다 스펙트럼의 에너지가

감소된다. 신호의 부호화 성능이 높아 왜곡이 작은 경우 음성 신호의 스펙트럼 에너지의 감소량은 비교적 작아지지만 그렇지 않을 경우에는 스펙트럼 에너지의 감소 현상이 두드러지게 나타나며, 실험에 의하면 특정 프레임에서 갑자기 많은 에너지 감소가 발생하는 경우가 종종 발생하는 것을 확인할 수 있다. 프레임의 진행 시간이 10msec에 불과하므로 프레임 에너지에 급격한 감소가 발생하는 것은 입력 신호의 특성보다는 음성 부호화기의 순간적인 성능 저하로 인하여 발생할 확률이 높으며, 이에 따라 프레임 에너지에 큰 감소가 발생하면 예외 처리를 하도록 한다. 즉, 필터 뱅크에서 각 대역의 에너지가 이전 프레임의 해당 대역에 비하여 크게 저하되면 해당 대역에서 부호화에 의한 스펙트럼 왜곡이 매우 크다고 판단하여 대역에 일정한 추가 가중치를 적용하여 대역의 에너지를 증가시키도록 한다.

이상의 네 가지 과정을 통하여 스펙트럼에 적용될 가중치 함수가 최종 결정되며, 이와 같은 가중치 함수 적용을 통하여 음성 부호화기에 의하여 왜곡된 스펙트럼이 보상되는 효과를 얻게 된다. 특히 각 프레임의 입력 음성 특성에 따라 서로 다른 가중치 함수가 결정되므로 음성 특성에 따른 가변 가중치 함수의 적용이 가능하다.

V. 성능 평가

본 논문에서 제안한 스펙트럼 보상 방법의 성능 분석을 위하여 HTK 3.1을 이용하는 고립 단어 인식기를 설계하여 사용하였다. 특징 파라미터는 Perceptual Linear Prediction (PLP) 기반의 13차 켑스트럼 계수를 사용하였고 실험에 사용된 음성 신호는 현재 이동 통신에 서비스되는 규격과 동일한 조건으로 하기 위하여 8kHz의 샘플링을 갖는 음성 데이터를 사용하였다[8]. 음성 데이터들은 SITEC에서 제작한 '클린 스피치 PBW452단어 DB'이며 훈련 데이터는 클린 스피치 30명분, 개인당 452단어씩 총 13,560단어이며 테스트 데이터는 8명분의 3,616단어이다.

제안한 스펙트럼 보상 방법의 성능을 상대적으로 비교하기 위하여 두 개의 음성 인식 모델을 생성하였다. 하나는 기존의 음성 인식기에 해당하는 것으로서 원 음성 신호를 가지고 만든 모델이며, 다른 한 가지는 원 음성 신호를 3개의 음성 부호화기 (IS-127 EVRC, ITU G.729 CS-ACELP, IS-96 QCELP)에 입력하여 얻은 왜곡된 음성 신호를 가지고 훈련시킨 모델이다. 따라서 음성 인식기 동작 환경은 크게 세 가지(원 음성 모델, 전처리 모듈이 포함된 원 음성 모델, 왜곡된 음성 모델)가 있으며 각각의 경우에 대하여 동일한 입력에 대한 음성 인식률을 측정하였다.

표 2. 제안한 전처리 기술을 사용한 단어 정확도
Table 2. Word accuracy using the proposed pre-processing method.

Condition / Test Data	Trained by Original Speech		Trained by Degraded Speech
	without Pre-Processing	with Pre-Processing	
Original	84.3%	-	-
EVRC	64.9%	80.5%	82.7%
G.729	74.1%	79.3%	85.6%
QCELP	66.2%	77.7%	83.1%

표 2에 정리되어 있듯이 EVRC에 의하여 왜곡된 음성 신호를 그대로 기존의 음성 인식기 (원 음성 모델)에 입력시키면 단어 인식률이 64.9%이지만, 동일한 신호를 제안한 전처리 모듈을 통과한 후 인식기에 입력시키면 인식률이 80.5%가 되어 15.6%의 인식률 향상을 얻을 수 있다. 만일 EVRC에 의하여 왜곡된 신호로 직접 훈련한 모델의 인식기에 입력시키면 인식률이 82.7%가 되어 제안한 방법에 비하여 성능이 향상되지만, 제안한 방법은 새로운 훈련이 필요 없이 기존의 인식 모델을 그대로 사용하는 장점을 가진다. G.729와 QCELP에 의하여 왜곡된 신호에 대하여서도 동일한 전처리 방법을 사용하여 많은 인식률 향상을 얻을 수 있으며, 이를 통하여 동일한 전처리 방법으로 다른 종류의 음성 부호화기에 의한 스펙트럼 왜곡을 보상하고, 그 결과 제안한 방법이 모든 통신 환경에 공통적으로 적용될 수 있는 것을 확인할 수 있다.

VI. 결 론

디지털 통신 시스템에서 음성 인식기를 동작시키면 음성 부호화기에 의한 음성 신호의 왜곡에 의하여 음성 인식 성능이 크게 저하된다. 이를 해결하기 위한 기존의 방법은 각 음성 부호화기의 특성을 포함하여 인식기를 새로 훈련하는 것이며, 이는 음성 인식기가 각 음성 부호화기에 종속되고 음성 부호화기마다 서로 다른 모델이 적용되는 문제점을 가진다. 본 논문에서는 이과 같은 문

제점을 해결하기 위하여 모든 CELP 음성 부호화기에 공통적으로 적용 할 수 있는 음성 인식 향상 방법을 제안하였다. 음성 부호화기에 의하여 음성 신호의 스펙트럼에 왜곡이 발생하고 이로 인하여 인식기의 성능이 저하된다. 이를 해결하기 위하여 왜곡된 스펙트럼을 보상하는 전처리 방법을 제안하며, 스펙트럼의 크기 정렬에 따른 보상, 고주파 영역에서의 예외처리, 시간에 따른 점진적 보상, 대역별 에너지 보상 등의 네 단계에 따라 스펙트럼 가중치 함수를 구하여 스펙트럼 왜곡을 보상하며, 보상된 스펙트럼으로부터 음성 특성을 추출하여 기존의 음성 인식기에 입력시킨다. 제안한 전처리 보상 방법을 적용하여 인식 성능이 크게 향상되는 것을 확인하였고, 특히 서로 다른 음성 부호화기에 대하여 동일한 보상 방법을 적용하여 성능이 향상되는 것을 확인하였다.

감사의 글

본 연구는 2004년 광운대학교 연구비의 지원으로 이루어졌습니다.

참 고 문 헌

1. 3GPP TS 22.243, "Speech recognition framework for automated voice services," Sept. 2003.
2. H. K. Kim and R. V. Cox, "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," IEEE Trans. Speech and Audio Processing, 9 (5), July 2001.
3. 3GPP TS 23.228, "IP Multimedia Subsystem(IMS)", March 2004.
4. TIA/EIA/IS-127 "Enhanced variable rate codec, speech service option 3 for wideband spectrum digital systems," 1996.
5. TIA/EIA/IS-96, "Speech service option standard for wideband spread spectrum digital cellular system," 1994.
6. ITU G.729, "Coding of speech at 8kb/s using conjugate-structure algebraic-code-excited linear prediction," 1996.
7. 한상욱, 박호종, "Modified PLP Feature를 적용하여 이동 통신 시스템에서의 음성 인식률 향상", 한국음향학회 추계 학술대회, 고려대학교, 2003.
8. H. Hermansky, "Perceptual linear predictive(PLP) analysis of speech," J. Acoust. Society of America, 87, 1738-1752, 1990.

저자 약력

• 서 진 호 (Jinho Seo)

2003년 2월: 대전대학교 정보통신공학과(공학사)
2005년 2월: 광운대학교 전자공학과(공학석사)
2005년 1월~현재: (주)벨웨이브 연구원
※ 주관심분야: 음성 신호처리, 음성 인식, 이동통신 시스템

• 박 호 종 (Hochong Park)

1986년 2월: 서울대학교 전자공학과(공학사)
1987년 12월: Univ. of Wisconsin-Madison 전자공학과(M.S.)
1993년 5월: Univ. of Wisconsin-Madison 전자공학과(Ph.D.)
1993년 9월~1997년 8월: 삼성전자 선임연구원
1997년 9월~현재: 광운대학교 전자공학과 부교수
※ 주관심분야: 음성/오디오 신호처리, 영상신호처리, 이동통신 시스템