(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2008-0077874
(43) 공개일자 2008년08월26일

(51) Int. Cl.
G10L 15/00 (2006.01) *G10L 15/02* (2006.01)
G10L 21/02 (2006.01) *G10L 15/14* (2006.01)
(21) 출원번호 10-2007-0017621
(22) 출원일자 2007년02월21일
심사청구일자 2007년02월21일

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 매탄동 416
(72) 발명자
오광철
경기 성남시 분당구 구미동 까치마을롯데선경아파트 412-1102
정재훈
경기 용인시 수지구 풍덕천동 6단지 상록아파트 614-1002
정소영
서울 관악구 남현동 1072-76 야우리스위트 401호
(74) 대리인
리앤목특허법인

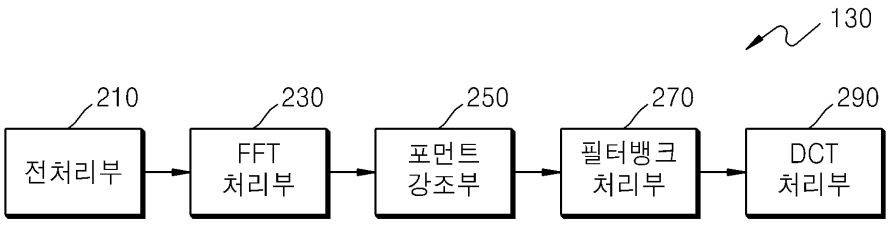
전체 청구항 수 : 총 14 항

(54) 음성 특징벡터 추출장치 및 방법과 이를 채용하는음성인식시스템 및 방법

(57) 요 약

음성 특징벡터 추출장치 및 방법과 이를 채용하는 음성인식 시스템 및 방법이 개시된다. 음성 특징벡터 추출장치는 프레임 단위로 구성된 음성신호를 주파수 영역의 신호로 변환하는 FFT 처리부; 상기 FFT 처리부로부터 제공되는 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조하는 포먼트 강조부; 및 상기 포먼트가 강조된 각 주파수 성분을 포함하는 주파수 영역의 신호를 복수개의 멜 스케일 필터뱅크를 이용하여 대역통과 필터링을 수행하는 필터뱅크 처리부를 포함하고, 상기 포먼트 강조부는 상기 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취하여 피치 하모닉 성분을 제거하는 하모닉 제거부; 및 피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시키는 스무딩부로 이루어진다.

대표도 - 도2



특허청구의 범위

청구항 1

프레임 단위로 구성된 음성신호를 주파수 영역의 신호로 변환하는 FFT 처리부;

상기 FFT 처리부로부터 제공되는 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조하는 포먼트 강조부; 및

상기 포먼트가 강조된 각 주파수 성분을 포함하는 주파수 영역의 신호를 복수개의 멜 스케일 필터뱅크를 이용하여 대역통과 필터링을 수행하는 필터뱅크 처리부를 포함하는 것을 특징으로 하는 음성 특징벡터 추출장치.

청구항 2

제1 항에 있어서, 상기 포먼트 강조부는

상기 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취하여 피치 하모닉 성분을 제거하는 하모닉 제거부; 및

피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시키는 스무딩부를 포함하는 것을 특징으로 하는 음성 특징벡터 추출장치.

청구항 3

프레임 단위로 구성된 음성신호를 주파수 영역의 신호로 변환하는 단계;

상기 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조하는 단계; 및

상기 포먼트가 강조된 각 주파수 성분을 포함하는 주파수 영역의 신호를 복수개의 멜 스케일 필터뱅크를 이용하여 대역통과 필터링을 수행하는 단계를 포함하는 것을 특징으로 하는 음성 특징벡터 추출방법.

청구항 4

제3 항에 있어서, 상기 포먼트 강조단계에서 상기 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취하여 피치 하모닉 성분을 제거하는 것을 특징으로 하는 음성 특징벡터 추출방법.

청구항 5

제4 항에 있어서, 상기 포먼트 강조단계에서 상기 피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시키는 단계를 포함하는 것을 특징으로 하는 음성 특징벡터 추출방법.

청구항 6

제5 항에 있어서, 상기 스무딩 단계는 다음 수학식

$$\hat{X}(k) = \frac{1}{U} \frac{\sum_{m=k+1}^{m=k+U} (m-k) \cdot \tilde{X}(m)}{\overline{X} + \sum_{m=k+1}^{m=k+U} \tilde{X}(m)}$$

$$\overline{X} = \frac{\sum_{m=1}^{m=N} \tilde{X}(m)}{N/P}$$

(여기서, $\tilde{X}(k)$ 는 피치 하모닉 성분이 억제된 k 번째 주파수 성분을 나타내고, $\hat{X}(k)$ 는 스무딩된 k 번째 주파수 성분을 나타내고, U는 국소적인 무게 중심을 구하는데 사용되는 주파수 성분의 수를 나타내고, $\overline{X}$ 는 전체 스펙트럼의 평균과 관련있는 파라미터이며, N은 FFT 포인트의 수, P는 $\overline{X}$ 가 전체 스펙트럼의 평균보다 큰 값이 되도록 하는 파라미터이다)

에 의해 수행되는 것을 특징으로 하는 음성 특징벡터 추출방법.

청구항 7

프레임 단위로 구성된 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조한 스펙트럼을 얻고, 상기 포먼트가 강조된 스펙트럼을 이용하여 음성인식을 위한 특징벡터를 추출하는 특징추출부; 및

데이터베이스를 참조하여 상기 추출된 특징벡터에 대한 인식과정을 수행하는 인식부를 포함하는 것을 특징으로 하는 음성인식시스템.

청구항 8

제7 항에 있어서, 상기 특징추출부는 상기 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취하여 피치 하모닉 성분을 제거하는 것을 특징으로 하는 음성인식시스템.

청구항 9

제8 항에 있어서, 상기 특징추출부는 상기 피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시키는 것을 특징으로 하는 음성인식시스템.

청구항 10

프레임 단위로 구성된 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조한 스펙트럼을 얻고, 상기 포먼트가 강조된 스펙트럼을 이용하여 음성인식을 위한 특징벡터를 추출하는 단계; 및

데이터베이스를 참조하여 상기 추출된 특징벡터에 대한 인식과정을 수행하는 단계를 포함하는 것을 특징으로 하는 음성인식방법.

청구항 11

제10 항에 있어서, 상기 특징추출단계는

상기 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취하여 피치 하모닉 성분을 제거하는 단계; 및

피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시키는 단계를 포함하는 것을 특징으로 하는 음성인식방법.

청구항 12

제11 항에 있어서, 상기 스무딩 단계는 다음 수학식

$$\hat{X}(k) = \frac{1}{U} \frac{\sum_{m=k+1}^{m=k+U} (m-k) \cdot \tilde{X}(m)}{\overline{X} + \sum_{m=k+1}^{m=k+U} \tilde{X}(m)}$$

$$\overline{X} = \frac{\sum_{m=1}^{m=N} \tilde{X}(m)}{N/P}$$

(여기서, $\tilde{X}(k)$ 는 피치 하모닉 성분이 억제된 k 번째 주파수 성분을 나타내고, $\hat{X}(k)$ 는 스무딩된 k 번째 주파수 성분을 나타내고, U는 국소적인 무게 중심을 구하는데 사용되는 주파수 성분의 수를 나타내고, $\overline{X}$ 는 전체 스펙트럼의 평균과 관련있는 파라미터이며, N은 FFT 포인트의 수, P는 $\overline{X}$ 가 전체 스펙트럼의 평균보다 큰 값이 되도록 하는 파라미터이다)

에 의해 수행되는 것을 특징으로 하는 음성인식방법.

청구항 13

제3 항 내지 제6 항 중 어느 한 항에 기재된 음성 특징벡터 추출방법을 실행할 수 있는 프로그램을 기재한 컴퓨터로 읽을 수 있는 기록매체.

청구항 14

제10 항 내지 제12 항 중 어느 한 항에 기재된 음성인식방법을 실행할 수 있는 프로그램을 기재한 컴퓨터로 읽을 수 있는 기록매체.

명 세 서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

<7> 본 발명은 음성인식에 관한 것으로서, 보다 구체적으로는 포먼트(formant)를 강조하기 위하여 피치 하모닉 성분을 억제함으로써 음성인식에 필요한 특징벡터를 보다 정확하게 추출하는 장치 및 방법과 이를 채용하는 음성인식시스템 및 방법에 관한 것이다.

<8> 현재, 음성인식 기술은 개인용 휴대 단말에서 정보 가전, 컴퓨터, 대용량 텔레포니 서버 등에 이르기까지 응용 범위를 점차 넓혀가고 있지만, 주변 환경에 따라 달라지는 인식성능의 불안정성을 개선하기 위하여 음성인식 성능 자체를 높이려는 시도와 잡음환경에서 인식율 저하를 방지하려는 시도와 관련하여 다양한 연구가 진행되어 왔다.

<9> 이중, 잡음환경에서 인식율이 저하하는 것을 방지하기 위하여, 음성인식 기술의 첫 단계인 음성 특징벡터 추출 과정에서 기존의 멜-주파수 켑스트럼 계수(mel-frequency cepstral coefficient, 이하 'MFCC' 이라 칭함) 특징벡터를 시간적인 특성을 고려하여 선형적으로 또는 비선형적으로 변환하는 기술들이 다양하게 연구되고 있다.

<10> 먼저, 특징벡터의 시간적인 특성을 고려한 기존의 변환 알고리즘에는 켑스트럼 평균 차감법(cepstral mean subtraction), 평균-분산 정규화(mean-variance normalization, On real-time mean-variance normalization of speech recognition features, P. Pujol, D. Macho and C. Nadeu, ICASSP, 2006, pp.773-776), RASTA 알고리즘(RelAtive SpecTrAl algorithm, Data-driven RASTA filters in reverberation, M. L. Shire et al, ICASSP, 2000, pp. 1627-1630), 히스토그램 정규화(histogram normalization, Quantile based histogram equalization for noise robust large vocabulary speech recognition, F. Hilger and H. Ney, IEEE Trans. Audio, Speech, Language Processing, vol.14, no.3, pp. 845-854), 델타 특징 증강 알고리즘(augmenting delta feature, On the use of high order derivatives for high performance alphabet recognition, J. di

Martino, ICASSP, 2002, pp. 953-956)등이 있다.

<11> 그리고, 특징벡터들을 선형적으로 변환하는 기술들에는 LDA(linear discriminant analysis) 및 PCA(principal component analysis, Optimization of temporal filters for constructing robust features in speech recognition, Jeih-Weih Hung et. al, IEEE Trans. Audio, Speech, and Language Processing, vol.14, No.3, 2006, pp. 808-832)를 이용하여 시간-프레임 상의 특징 데이터를 변환하는 방법들이 있다.

<12> 또한, 비선형 신경망을 사용하는 방법으로는 시간적인 패턴 알고리즘(TempoRAl Patterns, 이하 'TRAP' 이라 칭함, Temporal patterns in ASR of noisy speech, H. Hermansky and S. Sharma, ICASSP, 1999, pp. 289-292), 자동 음성 속성 전사 알고리즘(automatic speech attribute transcription, 이하 ASAT, A study on knowledge source integration for candidate rescoring in automatic speech recognition, Jinyu Li, Yu Tsao and Chin-Hui Lee, ICASSP, 2005, pp. 837-840) 등이 공지되어 있다.

<13> 한편, 음성인식 성능 자체를 높이는 시도와 관련해서는 음성인식과는 관련성이 적은 피치 하모닉을 포함하는 스펙트럼으로부터 MFCC 특징벡터를 추출하므로 그 성능 개선에는 한계가 있었다.

발명이 이루고자 하는 기술적 과제

<14> 본 발명이 이루고자 하는 기술적 과제는 포먼트를 강조하기 위하여 피치 하모닉 성분을 억제함으로써 음성인식에 필요한 특징벡터를 보다 정확하게 추출하는 장치 및 방법을 제공하는데 있다.

<15> 본 발명이 이루고자 하는 다른 기술적 과제는 상기한 음성 특징벡터 추출장치 및 방법을 채용하는 음성인식시스템 및 방법을 제공하는데 있다.

<16> 상기 기술적 과제를 해결하기 위하여 본 발명에 따른 음성 특징벡터 추출장치는 프레임 단위로 구성된 음성신호를 주파수 영역의 신호로 변환하는 FFT 처리부; 상기 FFT 처리부로부터 제공되는 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조하는 포먼트 강조부; 및 상기 포먼트가 강조된 각 주파수 성분을 포함하는 주파수 영역의 신호를 복수개의 멜 스케일 필터뱅크를 이용하여 대역통과 필터링을 수행하는 필터뱅크 처리부를 포함하여 이루어진다.

<17> 여기서, 상기 포먼트 강조부는 상기 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취하여 피치 하모닉 성분을 제거하는 하모닉 제거부; 및 피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시키는 스무딩부를 포함하는 것이 바람직하다.

<18> 상기 기술적 과제를 해결하기 위하여 본 발명에 따른 음성 특징벡터 추출방법은 프레임 단위로 구성된 음성신호를 주파수 영역의 신호로 변환하는 단계; 상기 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조하는 단계; 및 상기 포먼트가 강조된 각 주파수 성분을 포함하는 주파수 영역의 신호를 복수개의 멜 스케일 필터뱅크를 이용하여 대역통과 필터링을 수행하는 단계를 포함하여 이루어진다.

<19> 상기 다른 기술적 과제를 해결하기 위하여 본 발명에 따른 음성인식시스템은 프레임 단위로 구성된 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조한 스펙트럼을 얻고, 상기 포먼트가 강조된 스펙트럼을 이용하여 음성인식을 위한 특징벡터를 추출하는 특징추출부; 및 데이터베이스를 참조하여 상기 추출된 특징벡터에 대한 인식과정을 수행하는 인식부를 포함하여 이루어진다.

<20> 상기 다른 기술적 과제를 해결하기 위하여 본 발명에 따른 음성인식방법은 프레임 단위로 구성된 주파수 영역의 신호에 대하여, 각 주파수 성분에 포함된 피치 하모닉 성분을 억제하여 포먼트를 강조한 스펙트럼을 얻고, 상기 포먼트가 강조된 스펙트럼을 이용하여 음성인식을 위한 특징벡터를 추출하는 단계; 및 데이터베이스를 참조하여 상기 추출된 특징벡터에 대한 인식과정을 수행하는 단계를 포함하여 이루어진다.

<21> 상기 음성 특징벡터 추출방법 및 음성인식방법은 바람직하게는 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체로 구현될 수 있다.

발명의 구성 및 작용

<22> 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 대하여 상세하게 설명하기로 한다.

<23> 도 1은 본 발명이 채용되는 음성인식시스템의 구성을 나타낸 블록도로서, 잡음제거부(110), 특징벡터 추출부

(130), 인식부(150) 및 데이터베이스(170)를 포함하여 이루어진다.

<24> 도 1을 참조하면, 잡음제거부(110)는 입력되는 음성신호에 대하여 잡음을 제거한다. 음성신호에서 잡음을 제거하기 위해서는 공지되어 있는 다양한 방법 예를 들면, 스펙트럼 차감법(spectral subtraction) 등을 적용할 수 있다.

<25> 특징벡터 추출부(130)는 스펙트럼을 비선형적으로 변환함으로써, 피치 하모닉 성분이 억제되어 포먼트가 강조된 스펙트럼을 얻고, 얻어진 스펙트럼으로부터 MFCC 특징벡터를 추출한다.

<26> 인식부(150)는 특징벡터 추출부(130)에서 추출된 특징벡터를 대하여 학습된 데이터베이스(170)에 저장된 파라미터를 이용하여 유사도를 계산한다. 인식부(150)는 HMM(Hidden Markov Model), DTW(Dynamic Time Warping), 및 신경회로망(neural network) 등과 같은 다양한 음성인식 모델을 사용할 수 있다.

<27> 데이터베이스(170)는 인식부(150)에서 사용하는 모델의 파라미터를 미리 학습되어 저장한다. 인식부(150)가 신경회로망 모델을 사용할 경우 데이터베이스(170)에 저장되는 파라미터는 BP(Back Propagation) 알고리즘에 의해 학습된 각 노드들의 가중치값이고, 인식부(150)가 HMM 모델을 사용할 경우 데이터베이스(170)에 저장되는 파라미터는 Baum-Welch 재추정 알고리즘에 의해 학습된 상태천이 확률과 각 상태의 확률분포이다.

<28> 도 2는 본 발명에 따른 음성 특징벡터 추출장치의 일실시예의 구성을 나타낸 블록도로서, 전처리부(210), FFT(Fast Fourier Transform) 처리부(230), 포먼트 강조부(250), 필터뱅크 처리부(270) 및 DCT(Discrete Cosine Transform) 처리부(290)를 포함하여 이루어진다.

<29> 도 2를 참조하면, 전처리부(210)는 음성신호에 대하여 예를 들면 10 msec 마다 20~30 ms 길이로 한 프레임을 구성하고, 프레임 단위로 프리앰퍼시스(pre-emphasis) 처리를 수행하여 고주파 성분을 강조함으로써 자음성분을 강화한다. 프리앰퍼시스가 수행된 신호 x(n)은 다음 수학식 1과 같이 나타낼 수 있다.

수학식 1

<30> $x(n) = s(n) - \alpha(n-1)$

<31> 여기서, s(n)은 음성신호이고, α는 프리앰퍼시스에 사용되는 상수값으로서 통상 0.97을 사용한다.

<32> 한편, 프레임 간의 경계값의 갑작스러운 변화에 의해 주파수 정보가 왜곡되는 것을 방지하기 위하여, 전처리부(210)는 고주파 성분이 강조된 프레임 단위의 신호에 윈도우 함수 예를 들면 다음 수학식 2와 같이 나타낼 수 있는 해밍 윈도우 함수 h(n)를 적용한다.

수학식 2

<33> $h(n) = 0.6 - 0.4 \sin(2\pi n/M)$

<34> 여기서, M은 해밍 윈도우의 길이이다.

<35> FFT 처리부(230)는 윈도우가 적용된 신호를 N-포인트 FFT(Fast Fourier Transform) 처리하여 주파수 영역의 신호로 변환한다. N-포인트 FFT 처리는 다음 수학식 3과 같이 나타낼 수 있다.

수학식 3

<36>
$$X(k) = X(e^{j2\pi kf_s/N}) = \sum_{n=0}^{N-1} x(n)(e^{j2\pi kf_s/N})$$

<37> 여기서, f_s 는 샘플링 주파수이고, k= 0, 1, ...,N-1이다.

<38> 포먼트 강조부(250)는 FFT 처리부(230)로부터 제공되는 FFT 처리된 신호로부터 각 주파수 성분의 크기를 구하고, 각 주파수 성분의 크기에 대하여 인접한 주파수 성분의 크기를 차감하여 그 절대치를 취함으로써 피치 하모닉 성분을 제거하고, 피치 하모닉 성분이 제거된 각 주파수 성분의 크기를 국소적으로 스무딩하여 포먼트를 강조한다.

<39> 필터뱅크 처리부(270)는 포먼트 강조부(250)를 통해 제공되는 포먼트가 강조된 주파수 영역의 신호에 대하여, 인간의 청각특성에 따라 저주파수 영역은 좁게, 고주파수 영역은 넓게 그 대역폭을 멜 스케일로 분할한 복수개

의 필터뱅크를 이용하여 대역통과 필터링을 수행한다. 즉, 하나의 프레임내에서 특정 주파수성분에 대한 스펙트럼을 멜-스케일 필터링을 통하여 특징을 보다 잘 나타낼 수 있는 차원공간으로 변환한다. 이러한 멜-스케일 필터링은 다음 수학식 4와 같이 나타낼 수 있다.

수학식 4

$$E[j] = \sum_{m=1}^{m=N} \hat{X}(m) \cdot H_j(m), \quad 1 \le j \le J$$

<40>

<41> 여기서, E[j]는 필터뱅크 j의 출력을 나타내며, J는 필터뱅크의 수이고, $H_j(m)$은 필터뱅크 j의 전달함수를 나타낸다.

<42> DCT 처리부(290)는 필터뱅크 처리부(270)로부터 제공되는 각 필터뱅크 신호에 대하여 DCT 처리를 수행하여 최종적인 MFCC 특징벡터를 추출한다. 현재 음성인식 기술에서 널리 사용되고 있는 MFCC 특징벡터는 각 프레임당 12차의 벡터로 표현된다. DCT 처리를 통하여 출력되는 m차 MFCC 특징벡터 C(m)은 다음 수학식 5와 같이 나타낼 수 있다.

수학식 5

$$C[m] = \sum_{j=1}^{j=J} E(j) \cdot \cos\left[\frac{\pi \cdot m}{J}(j-0.5)\right]$$

<43>

<44> 여기서, J는 필터뱅크의 수이고, j는 각 필터뱅크를 나타낸다.

<45> 도 3은 도 2에 도시된 포먼트 강조부(250)의 세부적인 구성을 나타낸 블록도로서, 크기 계산부(310), 하모닉 제거부(330) 및 스무딩(smoothing)부(350)를 포함하여 이루어진다.

<46> 도 3을 참조하면, 크기 계산부(310)는 FFT 처리부(230)로부터 제공되는 신호로부터 각 주파수 성분의 크기를 구한다. 즉, FFT 처리부(230)로부터 제공되는 신호는 복소수이므로 그 크기를 취하여 실수값으로 변환함으로써 각 주파수 성분의 크기를 구할 수 있다.

<47> 하모닉 제거부(330)는 크기 계산부(310)로부터 제공되는 각 주파수 성분의 크기와 이웃하는 하위 주파수 성분의 크기를 차감하고, 차감된 결과의 절대치를 취함으로써 피치 하모닉 성분을 억제한다. 이는 다음 수학식 6과 같이 나타낼 수 있다.

수학식 6

$$\tilde{X}(k) = |X(k) - X(k-1)|$$

<48>

<49> 여기서, $\tilde{X}(k)$ 는 피치 하모닉 성분이 억제된 k 번째 주파수 성분을 나타낸다.

<50> 스무딩부(350)는 피치 하모닉 성분이 억제된 각 주파수 성분을 국소적인 무게 중심을 이용하여 스무딩시킨다. 스무딩 처리는 다음 수학식 7 및 8과 같이 나타낼 수 있다.

수학식 7

$$\hat{X}(k) = \frac{1}{U} \frac{\sum_{m=k+1}^{m=k+U} (m-k) \cdot \tilde{X}(m)}{\overline{X} + \sum_{m=k+1}^{m=k+U} \tilde{X}(m)}$$

<51>

수학식 8

$$\overline{X} = \frac{\sum_{m=1}^{m=N} \tilde{X}(m)}{N/P}$$

<52>

<53> 여기서, $\hat{X}(k)$ 는 스무딩된 k 번째 주파수 성분을 나타내고, U는 국소적인 무게 중심을 구하는데 사용되는 주파수 성분의 수 즉, 윈도우의 길이를 나타내고, $\overline{X}$ 는 전체 스펙트럼의 평균과 관련있는 파라미터이며, N은 FFT 포인트의 수, P는 $\overline{X}$ 가 전체 스펙트럼의 평균보다 큰 값이 되도록 조정하는 파라미터이다.

<54> 도 4a 및 도 4b는 본 발명과 종래기술간의 성능을 비교하기 위하여, 모음의 스펙트럼을 보여주는 도면이다. 도 4a는 종래기술에 의한 모음의 스펙트럼, 도 4b는 본 발명에 따른 모음의 스펙트럼을 각각 나타낸다. 종래기술에 따르면 피치 하모닉 성분에 의하여 두번째 포먼트와 세번째 포먼트를 구분하는 것이 어려우나, 본 발명의 경우에는 명확하게 구분됨을 알 수 있다.

<55> 도 5a 및 도 5b는 본 발명과 종래기술간의 성능을 비교하기 위하여, 한 문장의 스펙트로그램을 보여주는 도면이다. 도 5a는 종래기술에 의한 한 문장의 스펙트로그램, 도 5b는 본 발명에 따른 한 문장의 스펙트로그램을 각각 나타낸다. 이에 따르면, 마찬가지로 본 발명의 경우 포먼트의 궤적을 정확하게 추적할 수 있음을 알 수 있다. 한편, 도 6a 및 도 6b는 본 발명과 종래기술간의 성능을 비교하기 위하여, 필터뱅크들의 스펙트로그램을 보여주는 도면으로서, 마찬가지로 본 발명의 경우 스펙트로그램이 안정되어 있어서 포먼트의 궤적을 좀 더 명확하게 추적할 수 있음을 알 수 있다.

<56> 다음, 본 발명에 의한 효과를 검증하기 위하여 영어 발성에 대한 음성인식 실험을 수행하였다. 음성데이터는 미국 LDC(Linguistic Data Consortium)에서 제공하는 TIMIT 코퍼스로서, 이 데이터베이스는 음소에 대한 레벨이 부가되어 있어서 음소인식 성능을 측정하는데 기준이 되고 있다. 한편, 이 데이터베이스는 미국 전역을 8개 지역으로 나누어 각각 그 지방의 언어를 사용하는 사람의 음성을 수집하였으며, 총 6천여개의 음성문장으로 구성된다. 또한, 이 음성문장들을 음성인식에 사용하기 위하여 학습문장과 테스트문장으로 나누었으며 본 실험은 이를 기준으로 수행하였다.

<57> 음소인식에 사용된 알고리즘은 HMM이며, 영국 캠프리지 대학에서 제공하는 KTK 툴을 사용하였다. 성능 비교를 위하여 종래기술의 MFCC 특징벡터는 13차를 기준으로 델타(delta) 계수와 델타-델타(acceleration) 계수를 포함한 39차 특징벡터를 사용하였고, 20차의 필터뱅크를 사용하였다. 1 프레임은 20 ms의 길이를 가지며, 512-포인트 FFT 처리가 수행되었고, 프리앰퍼시스 상수는 0.97을 사용하였다. 한편, 본 발명에서는 프리앰퍼시스 상수로 0.97을, 20 msec의 해밍 윈도우, 512-포인트 FFT, 20차 필터뱅크를 사용하였고, 스무딩부(350)에서 사용되는 U는 5를 사용함으로써 현재 주파수 성분의 값을 주변 4개의 주파수 성분에 해당하는 값에 대하여 비교하여 스무딩을 수행하였다.

<58> 또한, P는 4를 사용함으로써 전체 스펙트럼의 평균의 4배에 해당하는 값을 기준으로 함으로써, 묵음구간에서와 같이 스펙트럼의 크기가 너무 작은 경우에도 작은 값의 변동이 국소적인 무게중심 스무딩으로 크게 변화하여 마치 포먼트 성분처럼 커지는 것을 방지하였다.

<59> 다음 표 1은 상기와 같은 실험환경에서 본 발명에 의해 얻어지는 특징벡터와 종래기술에 의해 얻어지는 특징벡터에 대한 음소인식 실험결과를 나타낸 것이다.

<60> [표 1]

<61>

	인식율	정확도	# 히트	# 삭제	# 대체	# 삽입	# 총 단어
종래기술	73.74 %	70.48 %	47,303	5,562	11,280	2,094	64,145
본 발명	74.23 %	71.17 %	47,618	5,500	11,027	1,963	64,145

<62> 표 1을 살펴보면 본 발명에 의한 특징벡터를 사용한 결과, 종래의 음소인식율인 73.34 %보다 높은 74.23 %을 나

타내며, 삽입 에러를 포함한 정확도에서도 향상된 결과를 나타냄을 확인할 수 있다.

<63> 본 발명은 또한 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플라피디스크, 광데이터 저장장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다. 그리고 본 발명을 구현하기 위한 기능적인(functional) 프로그램, 코드 및 코드 세그먼트들은 본 발명이 속하는 기술분야의 프로그래머들에 의해 용이하게 추론될 수 있다.

발명의 효과

<64> 상술한 바와 같이 본 발명에 따르면, FFT 처리되어 얻어지는 음성 스펙트럼에 대하여 피치 하모닉 성분을 억제하여 포먼트를 강조한 음성 스펙트럼을 얻고, 이로부터 보다 정확한 특징벡터를 추출하여 음성인식에 사용함으로써 음성인식 성능을 향상시킬 수 있는 이점이 있다.

<65> 본 발명에 대해 상기 실시예를 참고하여 설명하였으나, 이는 예시적인 것에 불과하며, 본 발명에 속하는 기술분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면의 간단한 설명

<1> 도 1은 본 발명이 채용되는 음성인식시스템의 구성을 나타낸 블록도,

<2> 도 2는 본 발명에 따른 음성 특징벡터 추출장치의 일실시예의 구성을 나타낸 블럭도,

<3> 도 3은 도 2에 도시된 포먼트 강조부의 세부적인 구성을 나타낸 블록도,

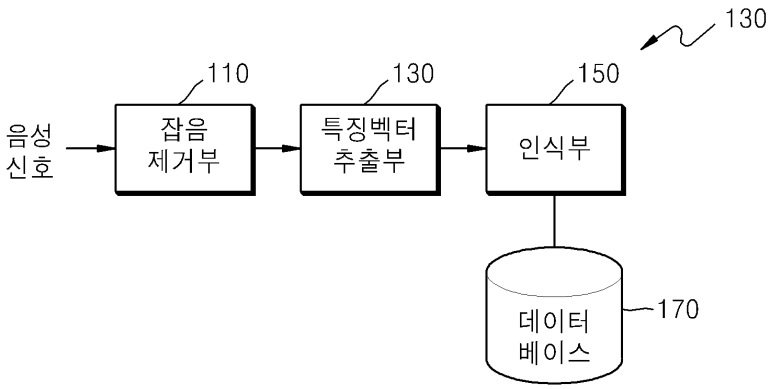
<4> 도 4a 및 도 4b는 본 발명과 종래기술간의 성능을 비교하기 위하여, 모음의 스펙트럼을 보여주는 도면,

<5> 도 5a 및 도 5b는 본 발명과 종래기술간의 성능을 비교하기 위하여, 한 문장의 스펙트로그램을 보여주는 도면, 및

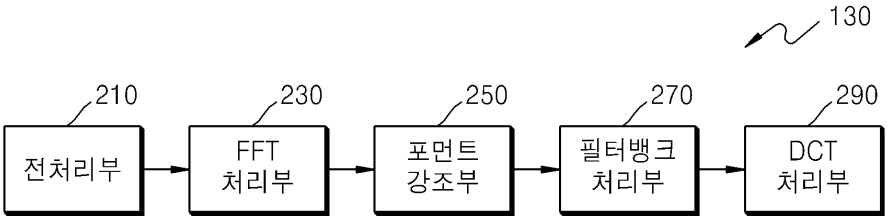
<6> 도 6a 및 도 6b는 본 발명과 종래기술간의 성능을 비교하기 위하여, 필터뱅크들의 스펙트로그램을 보여주는 도면이다.

도면

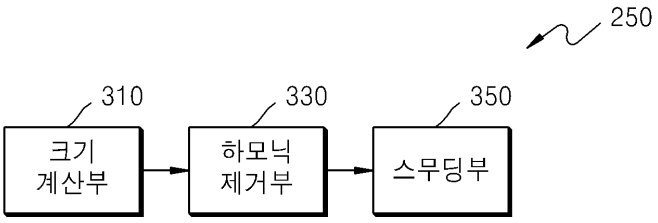
도면1



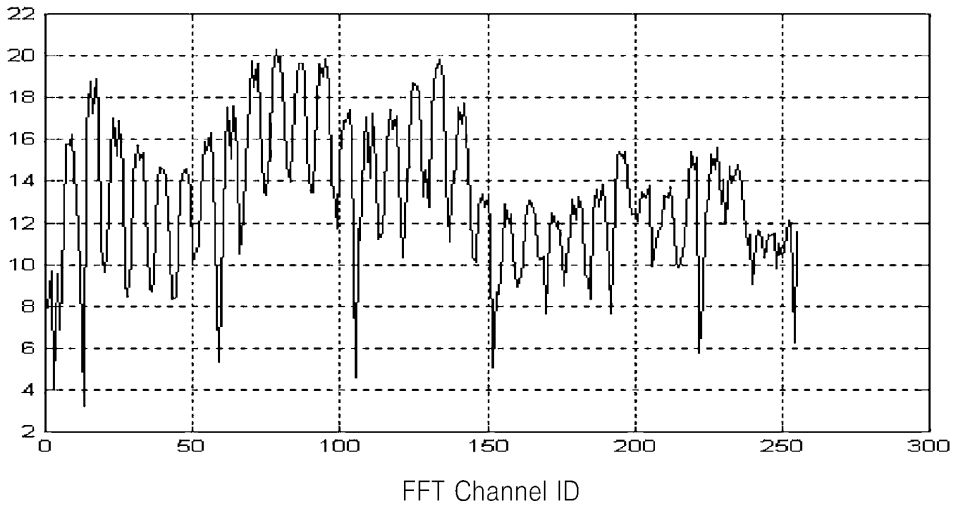
도면2


130
210
230
250
270
290
전처리부
FFT
처리부
포먼트
강조부
필터뱅크
처리부
DCT
처리부

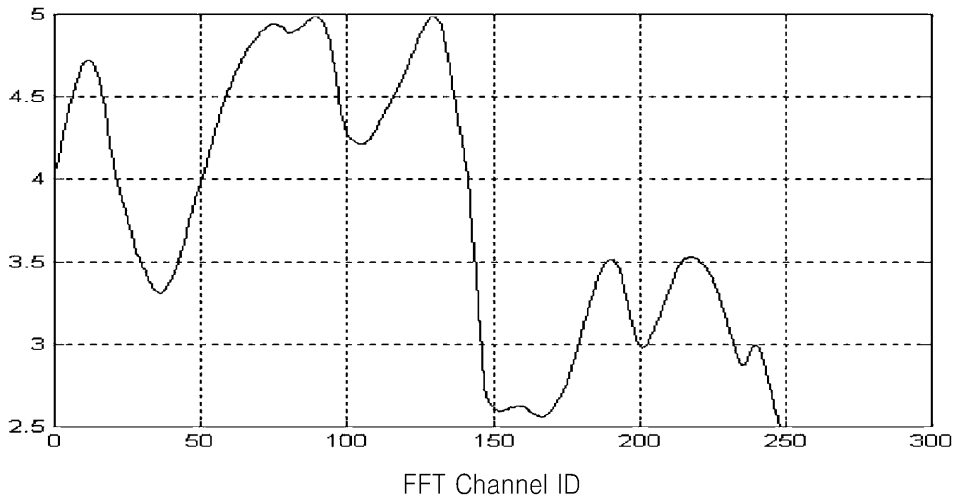
도면3


250
310
330
350
크기
계산부
하모닉
제거부
스무딩부

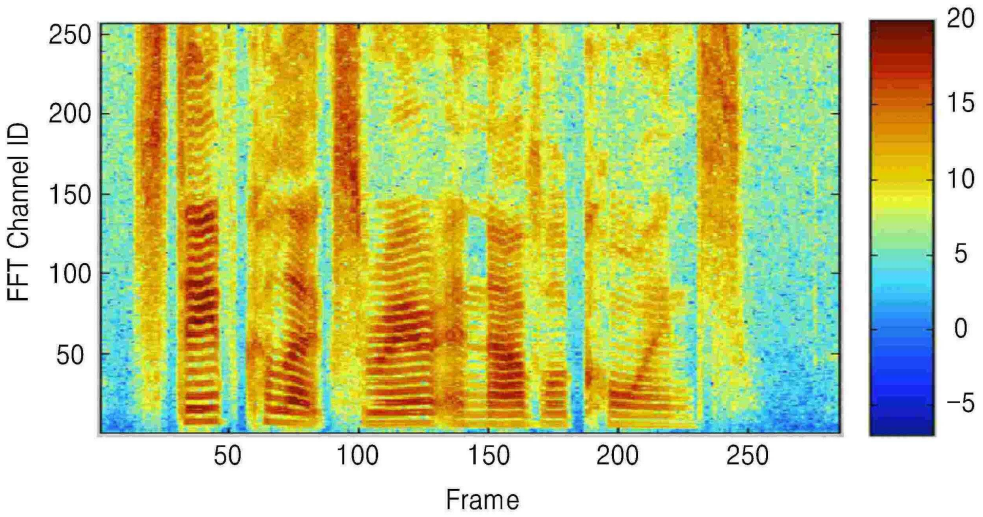
도면4a


22
20
18
16
14
12
10
8
6
4
2
0
50
100
150
200
250
300
FFT Channel ID

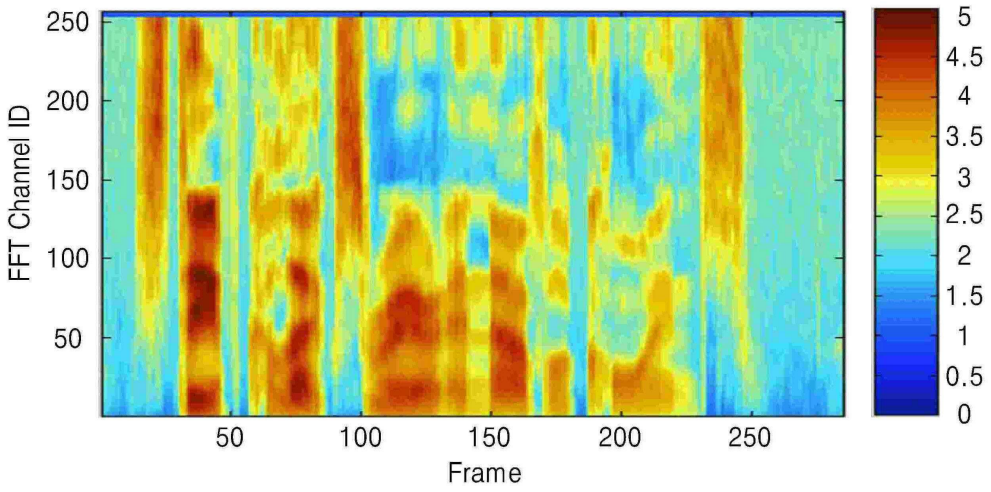
도면4b


5
4.5
4
3.5
3
2.5
0
50
100
150
200
250
300
FFT Channel ID

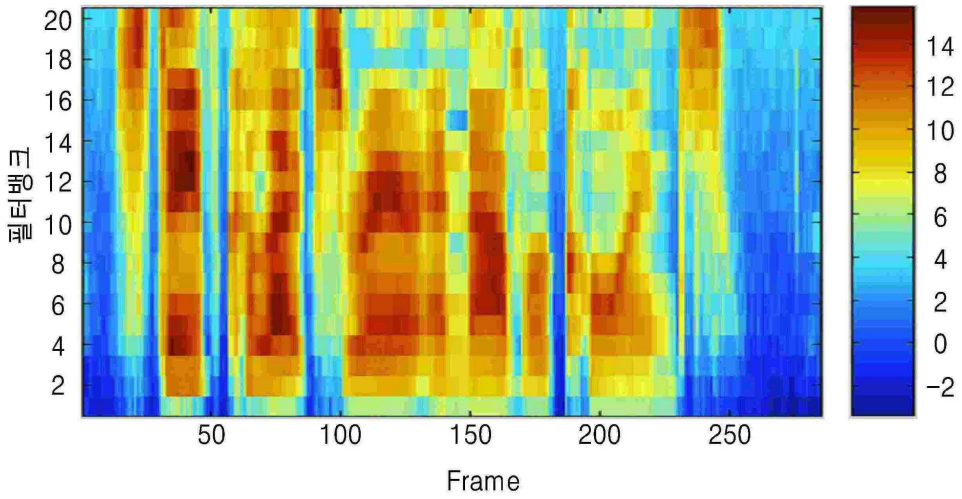
도면5a


250
200
150
100
50
FFT Channel ID
50
100
150
200
250
Frame
20
15
10
5
0
−5

도면5b


FFT Channel ID
250
200
150
100
50
50
100
150
200
250
Frame
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0

도면6a


필터뱅크
20
18
16
14
12
10
8
6
4
2
50
100
150
200
250
Frame
14
12
10
8
6
4
2
0
−2

도면6b