



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0068869
(43) 공개일자 2013년06월26일

(51) 국제특허분류(Int. Cl.)

G10L 15/20 (2006.01) *G10L 21/02* (2006.01)

(21) 출원번호 10-2011-0136285

(22) 출원일자 2011년12월16일

심사청구일자 2011년12월16일

(71) 출원인

서강대학교산학협력단

서울특별시 마포구 백범로 35 (신수동, 서강대학교)

(72) 발명자

박형민

경기도 화성시 반송동 나루마을 월드반도아파트 638동 902호

오명우

서울특별시 마포구 염리동 상록아파트 103동 1207호

(74) 대리인

이지연

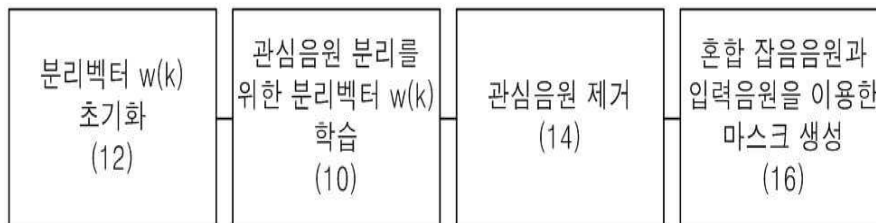
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 관심음원 제거방법 및 그에 따른 음성인식방법

(57) 요약

본 발명에 따르는 관심음원 제거방법은, 두 개의 마이크 각각으로부터의 입력 혼합신호를 제공받아 단구간 푸리에 변환하여 시간-주파수 영역으로 변환하는 단계; 상기 시간-주파수 영역의 입력 혼합신호들로부터 관심음원을 제거하기 위한 관심음원제거 벡터를 설정하는 단계; 상기 관심음원제거 벡터를 이용하여 입력 혼합신호에서 관심음원을 제거하여 혼합된 잡음신호를 생성하는 단계;를 구비함을 특징으로 한다.

대표도 - 도3



특허청구의 범위

청구항 1

관심음원 제거방법에 있어서,

두 개의 마이크 각각으로부터의 입력 혼합신호를 제공받아 단구간 푸리에 변환하여 시간-주파수 영역으로 변환하는 단계;

상기 시간-주파수 영역의 입력 혼합신호들로부터 관심음원신호를 제거하기 위한 벡터를 설정하는 단계;

상기 관심음원제거 벡터를 이용하여 입력 혼합신호에서 관심음원신호를 제거하여 혼합된 잡음신호를 생성하는 단계;를 구비함을 특징으로 하는 관심음원 제거방법.

청구항 2

제1항에 있어서,

상기 관심음원제거 벡터는 수학식 41 및 42에 따라 학습됨을 특징으로 하는 관심음원 제거방법.

수학식 41

$$\Delta \mathbf{w}(\mathbf{k}) \propto E\{ \mathbf{z}(\mathbf{k}, \tau) [((\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau))^* g(\sum |((\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau)|^2))] \}$$

수학식 42

$$\mathbf{w}(\mathbf{k}) \leftarrow \mathbf{w}(\mathbf{k}) / \|\mathbf{w}(\mathbf{k})\|$$

상기 수학식 41과 42에서 $\mathbf{w}$ 는 관심음원을 제거하기 위한 벡터, $\mathbf{z}$ 는 백색화된 입력 벡터, $\mathbf{k}$ 는 주파수 인덱스이다. 또한 g 는 비선형 함수 G 의 미분 결과 함수임.

청구항 3

제1항에 있어서,

상기 관심음원제거 벡터는 DUET 히스토그램에 따른 수학식 43에 따라 초기화됨을 특징으로 하는 관심음원 제거 방법.

수학식 43

$$\mathbf{w}(k) = \frac{1}{\sqrt{1 + \alpha^2(k)}} \begin{bmatrix} 1 \\ -A^{-1}(k) \end{bmatrix}$$

상기 수학식 43에서 $\mathbf{w}$ 는 관심음원을 제거하기 위한 벡터, α 와 δ 는 관심음원에 대한 감쇄 및 시간지연 파라미터, $A(k) = \alpha(k)e^{-j\omega_k \delta(k)}$ 이고, ω_k 는 k 번째 주파수 영역의 정규화 주파수임.

청구항 4

제3항에 있어서,

상기 DUET 히스토그램은 이차원 가우시안 원도우로 2차원 필터링됨을 특징으로 하는 관심음원 제거방법.

청구항 5

음성인식방법에 있어서,

두 개의 마이크 각각으로부터의 입력 혼합신호를 제공받아 단구간 푸리에 변환하여 시간-주파수 영역으로 변환하는 단계;

상기 시간-주파수 영역의 입력 혼합신호들로부터 관심음원신호를 제거하기 위한 벡터를 설정하는 단계;
 상기 관심음원제거 벡터를 이용하여 입력 혼합신호에서 관심음원신호를 제거하여 혼합된 잡음신호를 생성하는 단계;
 상기 혼합된 잡음신호와 입력 혼합신호 사이의 에너지 비에 따라 이진 마스크를 생성하는 단계;
 상기 이진 마스크와 상기 입력 혼합신호를 이용하여 관심음원에 대한 손실특징을 복원하는 단계;
 상기 관심음원에 대한 손실특징이 복원된 결과를 이용한 음성인식을 이행하는 단계;를 구비함을 특징으로 하는 음성인식방법.

청구항 6

제5항에 있어서,
 상기 관심음원제거 벡터는 수학식 44 및 45에 따라 학습됨을 특징으로 하는 음성인식방법.

수학식 44

$$\Delta \mathbf{w}(\mathbf{k}) \propto E\{ \mathbf{z}(\mathbf{k}, \tau) [((\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau))^* g(\sum |((\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau)|^2)] \}$$

수학식 45

$$\mathbf{w}(\mathbf{k}) \leftarrow \mathbf{w}(\mathbf{k}) / \|\mathbf{w}(\mathbf{k})\|$$

상기 수학식 44과 45에서 $\mathbf{w}$ 는 관심음원을 제거하기 위한 벡터, $\mathbf{z}$ 는 백색화된 입력 벡터, $\mathbf{k}$ 는 주파수 인덱스이다. 또한 g 는 비선형 함수 G 의 미분 결과 함수임.

청구항 7

제5항에 있어서,
 상기 관심음원제거 벡터는 DUEET 히스토그램에 따른 수학식 46에 따라 초기화됨을 특징으로 하는 음성인식방법.

수학식 46

$$\mathbf{w}(k) = \frac{1}{\sqrt{1 + \alpha^2(k)}} \begin{bmatrix} 1 \\ -A^{-1}(k) \end{bmatrix}$$

상기 수학식 46에서 $\mathbf{w}$ 는 관심음원을 제거하기 위한 벡터, α 와 δ 는 관심음원에 대한 감쇄 및 시간지연 파라미터, $A(k) = \alpha(k) e^{-i\omega_k \delta(k)}$ 이고, ω_k 는 k 번째 주파수 영역의 정규화 주파수임.

청구항 8

제7항에 있어서,
 상기 DUEET 히스토그램은 이차원 가우시안 윈도우로 2차원 필터링됨을 특징으로 하는 음성인식방법.

청구항 9

제5항에 있어서,
 상기 손실특징 복원은,
 수학식 47, 48, 49, 50에 따르는 클러스터 기반 손실특징 복원임을 특징으로 하는 음성인식방법.
 수학식 47

$$\hat{X}_u = \sum_{\nu} P(\nu | X_r, X_u \leq Y_u) \hat{X}_u^{\nu}$$

수학식 48

$$\hat{X}_u^\nu = \underset{X_u}{\operatorname{argmax}} P(X_u | X_r = Y_r, X_u \leq Y_u; \mu_\nu, \Theta_\nu)$$

수학식 49

$$P_\nu(X_r, X_u \leq Y_u) = \frac{c_\nu P(X_r, X_u \leq Y_u | \nu)}{\sum_j c_j P(X_r, X_u \leq Y_u | j)}$$

수학식 50

$$\begin{aligned} P(X_r, X_u \leq Y_u | \nu) &= P(X_r, -\infty \leq X_u \leq Y_u; \mu_\nu, \Theta_\nu) \\ &= \prod_{j \in R} \frac{1}{\sqrt{2\pi} \sigma_\nu(j)} e^{-\frac{(X(j) - \mu_\nu(j))^2}{2\sigma_\nu^2(j)}} \times \prod_{i \in U} \int_{-\infty}^{Y(i)} \frac{1}{\sqrt{2\pi} \sigma_\nu(i)} e^{-\frac{(x - \mu_\nu(i))^2}{2\sigma_\nu^2(i)}} dx \end{aligned}$$

상기 수학식 47~50에서 $\hat{X}_u$ 은 상기 알고리즘을 통해 찾는 신뢰성 없는 출력 특징의 성분이며, Y_r 과 Y_u 는 관찰 벡터 Y 에서 각각 신뢰성 있는 성분으로 구성된 벡터와 신뢰성 없는 성분으로 구성된 벡터를 나타내고, X_r 과 X_u 는 이에 대응되는 잡음의 영향이 없는 내재된 음성신호의 일부분임. R 과 U 는 신뢰성 있는 벡터 성분의 인덱스 집합과 신뢰성 없는 벡터 성분의 인덱스 집합을 나타내며, $X(j)$ 와 $Y(j)$ 는 각각 음성신호 벡터 X 와 관찰 벡터 Y 의 j 번째 성분을 나타냄. $c_\nu; \mu_\nu; \Theta_\nu$ 는 ν 번째 가우시안의 가중치, 평균벡터, 공분산행렬을 나타내며, $\sigma_\nu^2(j)$ 와 $\mu_\nu(j)$ 는 ν 번째 가우시안 분포에서 j 번째 인덱스의 분산과 평균을 나타냄. Y_r 과 Y_u 는 관찰 벡터 Y 에서 각각 신뢰성 있는 성분과 신뢰성 없는 성분을 나타냄. X_r 과 X_u 는 대응되는 성분에서 잡음의 영향이 없는 내재된 음성신호 부분을, X 는 X_r 과 X_u 를 결합한 내재된 음성신호 벡터를 나타냄. R 과 U 는 신뢰성있는 벡터 성분과 신뢰성없는 벡터 성분의 인덱스 집합을 나타내며, $\sigma_\nu^2(j)$ 와 $\mu_\nu(j)$ 는 현재 상태(state) ν 에 따른 가우시안 분포에서 j 번째 인덱스의 분산과 평균을 나타냄.

명세서

기술분야

[0001] 본 발명은 관심음원 제거기술에 관한 것으로, 더욱 상세하게는 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합신호에서 관심음원만을 제거하는 것을 통해 혼합된 잡음신호를 추정하는 관심음원 제거방법 및 그에따른 음성인식방법에 관한 것이다.

배경기술

[0002] 암묵 신호 분리(Blind Source Separation)기술은 두 개 이상의 마이크로부터 채집된 혼합신호로부터 관심음원을 분리하는 기술이다.

[0003] 상기한 암묵 신호 분리기술에 대한 종래 기술에 대해 설명한다.

[0004] [비정규분포의 정도(non-Gaussianity)에 기반한 독립성분분석(Independent Component Analysis)]

[0005] 먼저 비정규분포의 정도에 기반한 독립성분분석 기법에 대해 설명한다.

[0006] 일반적으로 독립성분분석은 다음과 같은 모델로 설명된다.

수학식 1

[0007] $\mathbf{x} = \mathbf{A}\mathbf{s}$

수학식 2

[0008] $\mathbf{y} = \hat{\mathbf{s}} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} \simeq \mathbf{A}^{-1}\mathbf{A}\mathbf{s}$

[0009] 상기 수학식 1 및 2에서 $\mathbf{y}$ 는 독립성분분석의 출력 벡터이고, $\mathbf{x}$ 는 마이크로폰의 입력 벡터, $\mathbf{s}$ 는 찾고자 하는 음원의 벡터이다.

[0010] 그러나 음원이 마이크로 들어오기까지의 혼합 행렬 $\mathbf{A}$ 를 알지 못하는 상태에서 그것의 의사 역행렬(Pseudo-inverse Matrix) $\mathbf{W}$ 를 찾아야 하는 것이 독립성분분석의 문제였다.

[0011] 상기한 독립성분분석의 문제는 Aapo Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Trans. on Neural Networks*, vol. 10, no. 3, 1999. 에 따라, 비정규분포의 정도(non-Gaussianity)를 측정함으로써 해결할 수 있으며, 이는 중심극한정리에 기초한다. 즉, 독립된 관심음원과 혼합된 잡음신호가 존재할 때 혼합된 잡음신호는 여러 개의 독립된 잡음신호들의 합성이므로 독립된 관심음원에 비해 상대적으로 정규분포를 띠게 된다. 이에따라 출력신호의 비정규분포의 정도를 최대화하면 각각의 독립성분을 분리해낼 수 있다는 것이다.

[0012] 도 1은 독립된 한 사람이 발성한 음원의 히스토그램과 많은 사람들이 한꺼번에 떠드는 소리인 배블잡음의 히스토그램을 예시한 것으로, 혼합된 잡음신호가 독립된 관심음원보다 상대적으로 정규분포에 더 가깝다.

[0013] [역엔트로피 최대화 기반 독립성분분석 기법]

[0014] 출력신호 y_i 의 비정규분포의 정도를 측정하기 위해서 수학식 3에 따른 역엔트로피(negentropy)의 정의를 활용한다.

수학식 3

[0015] $J(y_i) = H(y_i^{\text{gauss}}) - H(y_i)$

[0016] 여기에서, y_i^{gauss} 는 y_i 와 동일한 분산을 갖는 정규분포 확률변수이다. 그리고 확률 밀도 함수가 p_{y_i} 인 확률변수 y_i 의 엔트로피 H 는 수학식 4에 따라 표현된다.

수학식 4

[0017] $H(y_i) = - \int p_{y_i}(u) \log p_{y_i}(u) du$

[0018] 동일한 분산을 갖는 확률변수 중 가장 큰 엔트로피를 갖는 확률변수는 정규분포 확률변수이므로 역엔트로피를 최대화함으로써 추정되는 출력신호 y_i 에 대한 비정규분포의 정도를 최대화할 수 있고, 중심극한정리에 의해 이

렇게 추정된 출력신호 y_i 는 원래 음원 신호에 가까워진다. 상기한 역엔트로피는 직접적인 계산이 매우 복잡하기 때문에 대칭 분포를 갖는 확률변수에 대해 수학식 5와 같이 근사화할 수 있다.

수학식 5

$$J(y_i) \propto [E\{G(y_i)\} - E\{G(y_i^{\text{gauss}})\}]^2$$

이러한 근사화 방법은 고차원 중첩 근사화방법을 일반화한 것으로, 이는 출력신호 y_i 에 대한 2차가 아닌 비선형 함수(nonquadratic function) G 의 기대값을 이용한다. 이 비선형 함수 G 는 첨도(kurtosis) 기반일 때 $G(y) = y^4$ 로 근사화한다. 그리고 첨도 기반 외에도 다음과 같은 효과적인 비선형 함수가 존재한다.

수학식 6

$$G_1(y) = \frac{1}{a_1} \log \cosh a_1 y$$

수학식 7

$$G_2(y) = -\exp(-y^2/2)$$

상기 수학식 6 및 7에서 출력신호 $\mathbf{y}$ 는 평균은 0이고, 분산은 1로 가정한다. 이때, $\mathbf{x}$ 를 백색화(whitening)하면 평균은 0이고, 상관행렬(correlation matrix)은 단위행렬(identity matrix)이 되므로 유니테리행렬(unitary matrix)로 변환을 통해 평균이 0이고, 분산이 1인 $\mathbf{y}$ 를 추정할 수 있고 유니테리행렬의 특성을 이용하여 간략한 변환행렬 학습식을 유도할 수 있다. 백색화 변환 행렬을 $\mathbf{V}$ 라고 했을 때, 입력신호 $\mathbf{x}$ 에 대한 백색화 출력 $\mathbf{z}$ 는 수학식 8과 같이 표현된다.

수학식 8

$$\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x}$$

상기 수학식 8에서 $\mathbf{D} = \text{diag}(d_1 \dots d_n)$ 는 입력 공분산 행렬의 고유값들의 대각 행렬이고, $\mathbf{E}$ 는 입력 공분산 행렬의 고유벡터들의 모음 행렬이다.

상기한 수학식 8에 따라 백색화된 확률변수 $\mathbf{z}$ 를 이용한 역엔트로피 수식은 수학식 9와 같다.

수학식 9

[0027]
$$J_G(\mathbf{w}) = [E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}]^2$$

[0028] 상기 수학식 9에서 $\mathbf{w}$ 는 norm이 1인 벡터이며, v 는 평균은 0, 분산은 1인 정규분포 확률변수이다. 그리고 $E\{G(v)\}$ 보다 $E\{G(\mathbf{w}^T \mathbf{z})\}$ 이 항상 작기 때문에, 역엔트로피를 최대화하는 것은 $E\{G(\mathbf{w}^T \mathbf{z})\}$ 을 최소화하는 것과 동일하다. 상기 역엔트로피를 최대화하기 위하여 $\mathbf{w}$ 에 대한 최대경사법 알고리즘을 이끌어 낼 수 있다. 그에 따른 알고리즘은 수학식 10 및 수학식 11에 따른 학습 규칙을 갖게 된다.

수학식 10

[0029]
$$\Delta \mathbf{w} \propto \gamma E\{\mathbf{z}g(\mathbf{w}^T \mathbf{z})\}$$

수학식 11

[0030]
$$\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|$$

[0031] 여기에서 $\gamma = E\{G(\mathbf{w}^T \mathbf{z})\} - E\{G(v)\}$ 이다. 미분 과정에서 $E\{G(v)\}$ 가 사라진 것은 평균이 0,

분산이 1인 정규분포 확률변수에 대한 항의 값이 norm이 1인 $\mathbf{w}$ 에 대해서 고정이기 때문이다. 상기 γ 의 부호는 학습 과정에서 안정성에 영향을 주기 때문에, 이를 고정하는 것으로 알고리즘을 더 단순화할 수 있다. 특히 독립 성분의 선형적 정보를 통해 γ 의 부호를 정할 수 있는데, 예를 들어 음성 신호의 경우, 수퍼가우시안 분포를 띄므로 $g(E)$ 가 $\tanh(E)$ 일 때 γ 를 -1로 고정하여 음원 신호를 찾을 수 있다. 따라서 역엔트로피를 최대화하는 것 대신에 $E\{G(\mathbf{w}^T \mathbf{z})\}$ 를 최소화함으로써 특정 음원 신호를 복원할 수 있다.

[0032] 상기한 종래 방식들은 주파수 평면에서의 분석을 요하기 때문에, 분리 벡터 $\mathbf{w}$ 및 입력 신호, 출력신호들이 모두 복소수이다. 비용함수가 복소수의 형태를 띌면 그것의 크기를 정의할 수 없고, 그에 따른 비용함수의 최소화가 불가능하기 때문에, $\mathbf{w}$ 에 대한 비용함수를 다음과 같이 $\mathbf{w}^H \mathbf{z}$ 의 절대값의 제곱의 형태로 표현한다.

수학식 12

[0033]
$$J_G'(\mathbf{w}) = E\{G(|\mathbf{w}^H \mathbf{z}|^2)\}$$

[0034] 여기에서 G 는 실수 평면에서의 G 와 다른, 아래의 함수들을 사용한다.

수학식 13

[0035] $G_1(y) = \sqrt{a_1 + y}$

수학식 14

[0036] $G_2(y) = \log(a_2 + y)$

수학식 15

[0037] $G_3(y) = \frac{1}{2}y^2$

[0038] 상기의 비용함수를 미분하여, $\mathbf{w}$ 에 대한 최대경사법 알고리즘을 이끌어 낼 수 있다. 그에 따른 알고리즘은 수학식 16 및 수학식 17에 따른 학습 규칙을 갖게 된다.

수학식 16

[0039] $\Delta \mathbf{w} \propto -E\{\mathbf{z}(\mathbf{w}^H \mathbf{z})^* g(|\mathbf{w}^H \mathbf{z}|^2)\}$

수학식 17

[0040] $\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|$

[0041] [독립벡터분석(Independent Vector Analysis)을 적용한 역엔트로피 최대화 기반 독립성분분석 기법]

[0042] 그리고 독립벡터분석 알고리즘은 독립성분분석의 개념을 주파수 평면으로의 다변수 성분으로 확장한 것이다. 상기 독립벡터분석은 내재된 독립 성분과 관측된 신호 각각이 모두 다변수 신호, 즉 벡터 신호인 독립 성분 분석 문제로 간주될 수 있다.

[0043] 독립벡터분석 모델에서 각각의 신호원 벡터들은 서로 다른 소스 벡터들과 확률적으로 독립이라고 가정되어 있는 반면, 각 신호원 벡터 내부의 성분들은 서로 확률적으로 독립이 아니고 서로 연관되어 있다고 가정한다.

[0044] 이 가정을 주파수 평면으로의 알고리즘에 대응하면, 각각의 신호원 벡터들은 주파수 축으로의 벡터이고, 벡터 내부의 성분들, 즉 주파수 간의 성분들끼리 연관성을 가진다고 설명할 수 있다.

[0045] 도 2는 2채널 입출력 주파수 신호에 대한 주파수 평면 독립성분분석과 독립벡터분석의 비교 모식도를 도시하였다.

[0046] 따라서 위 독립성분분석에서, 비용함수에 포함된 비선형 함수 G 가 주파수 축으로의 다변수 벡터를 인수로 받게 되며, 이에 따라 변화한 $\mathbf{w}$ 에 대한 비용함수는 수학식 18과 같다.

수학식 18

$$J_G''(\mathbf{w}(\mathbf{k})) = E\{G(\sum_k |(\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau)|^2)\}$$

[0047]

[0048] 상기 수학식 18에서, k 와 $\mathcal{T}$ 는 각각 주파수와 시간 프레임의 인덱스를 의미한다. 상기 수학식 18에 따르면 독립 벡터분해에 의해 비선형 함수 G 의 인수가 다변수 벡터로 변화한 것을 알 수 있다. 이를 미분하여, $\mathbf{w}(\bar{k})$ 에 대한 최대경사법 알고리즘을 이끌어 낼 수 있다. 그에 따른 알고리즘은 수학식 19 및 20과 같은 학습 규칙을 갖게 된다.

수학식 19

$$\Delta \mathbf{w}(\mathbf{k}) \propto -E\{\mathbf{z}(\mathbf{k}, \tau)[(\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau)]^* g(\sum_k |(\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau)|^2)\}$$

[0049]

수학식 20

$$\mathbf{w}(\mathbf{k}) \leftarrow \mathbf{w}(\mathbf{k}) / \|\mathbf{w}(\mathbf{k})\|$$

[0050]

[0051] [관심음원 분리 알고리즘]

[0052] 상술한 종래 기술에 따른 관심음원 분리 알고리즘을 통해 출력 신호의 역엔트로피를 최대화하여 효과적으로 관심 음원을 추정할 수 있으나, 이를 위해서는 이론적으로 혼합된 음원의 개수만큼 마이크 입력이 존재해야만 한다.

[0053] 그러나 실제 환경에서 혼합된 음원의 개수만큼 마이크를 준비하는 것은 불가능할 뿐만 아니라 준비한다고 해도 추정할 파라미터가 너무 많아지는 문제가 있었다.

[0054] 상술한 음원 분리에 대한 종래기술로는 [Aapo Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Trans. on Neural Networks*, vol. 10, no. 3, 1999.], [E. Bingham and A. Hyvarinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals", *International Journal of Neural Systems*, vol. 10, no. 1, 2000.], [I. Lee, T. Kim, and T. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation", *Signal Processing*, vol. 87, Issue 8, 2007.] 등이 있다.

발명의 내용

해결하려는 과제

[0055] 본 발명은 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합 신호에서 관심음원을 제거하는 것을 통해 혼합된 잡음신호를 추정할 수 있는 관심음원 제거 방법을 제공하는 것을 그 목적으로 한다.

[0056] 또한 본 발명의 다른 목적은 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합신호에서 관심음원을 제거하는 것을 통해 혼합된 잡음신호를 추정하고, 이를 마이크를 통해 제공되는 혼합신호와 비교하여 혼합신호에서 잡음신호가 지배적인 영역을 구분한 후 이 영역에 대해 손실특징 복원 기술을 적용하여 특징값을 복원한 음성인식방법을 제공하는 것이다.

과제의 해결 수단

[0057] 상기한 목적을 달성하기 위한 본 발명에 따르는 본 발명에 따르는 관심음원 제거방법은, 두 개의 마이크 각각으로부터의 입력 혼합신호를 제공받아 단구간 푸리에 변환하여 시간-주파수 영역으로 변환하는 단계; 상기 시간-주파수 영역의 입력 혼합신호들로부터 관심음원을 제거하기 위한 관심음원제거 벡터를 설정하는 단계; 상기 관심음원제거 벡터를 이용하여 입력 혼합신호에서 관심음원을 제거하여 혼합된 잡음신호를 생성하는 단계;를 구비함을 특징으로 한다.

발명의 효과

[0058] 본 발명은 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합 신호에서 관심음원을 제거하는 것을 통해 혼합된 잡음신호를 효과적으로 추정할 수 있다.

[0059] 또한 본 발명은 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합 신호에서 관심음원을 제거하는 것을 통해 혼합된 잡음신호를 추정하고, 이를 마이크를 통해 제공되는 혼합 신호와 비교하여 혼합신호에서 잡음신호가 지배적인 영역을 구분한 후 이 영역에 대해 손실특징 복원 기술을 적용하여 특징값을 복원함으로써 음성인식 시에 음성인식율을 향상시키는 효과를 야기한다.

도면의 간단한 설명

[0060] 도 1은 한 사람의 발성음원과 배틀잡음의 히스토그램을 도시한 도면.
 도 2는 주파수 도메인의 독립성분분석과 독립벡터분석의 차이를 도시한 모식도.
 도 3은 본 발명의 바람직한 실시예에 따른 관심음원 제거방법의 절차도.
 도 4는 DUET 알고리즘에서 감쇄-시간지연 히스토그램을 예시한 도면.
 도 5는 본 발명의 바람직한 실시예에 따른 음성인식장치의 구성도.
 도 6은 본 발명의 바람직한 실시예에 따른 관심음원 제거 실험을 위한 실내 환경 모식도.
 도 7은 관심음원과 잡음신호와 관심음원제거 출력신호를 도시한 도면.
 도 8은 다양한 입력신호의 신호대 잡음비 및 반향시간에 대한 관심음원제거 출력신호의 신호대 잡음비 결과를 나타내는 표.
 도 9는 다양한 입력신호의 신호대 잡음비 및 반향시간에 대한 음성인식 결과를 나타내는 표.

발명을 실시하기 위한 구체적인 내용

[0061] 본 발명은 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합 신호에서 관심음원을 제거하는 것을 통해 혼합된 잡음신호를 추정한다.

[0062] 또한 본 발명은 다수 개의 음원으로부터의 음향신호가 혼합되는 환경에 설치된 두 개의 마이크를 통해 제공되는 혼합 신호에서 관심음원을 제거하는 것을 통해 혼합된 잡음신호를 추정하고, 이를 마이크를 통해 제공되는 혼합 신호와 비교하여 혼합신호에서 잡음신호가 지배적인 영역을 구분한 후 이 영역에 대해 손실특징 복원 기술을 적용하여 특징값을 복원함으로써 음성인식 시에 음성인식율을 향상시키는 효과를 야기한다.

[0063] [관심음원 제거]

[0064] 상기한 본 발명의 바람직한 실시예에 따른 관심음원 제거절차를 도 3을 참조하여 상세히 설명한다.

[0065] 상기 관심음원 제거는 두 개의 마이크로부터의 혼합신호에 대해 주파수별로 하나의 널(null)을 형성하여 혼합신호에서 관심음원만을 제거하는 것을 통해 혼합된 잡음신호를 예측한다.

[0066] 먼저 혼합신호에서 관심음원을 제거하기 위한 벡터 $\mathbf{w}^{(k)}$ 를 학습한다(10단계).

[0067] 상기 벡터 $\mathbf{w}^{(k)}$ 의 학습과정을 설명한다.

[0068] 혼합된 잡음신호 전체를 독립된 잡음음원 여러 개가 혼합된 형태로 가정할 때, 중심극한정리에 의해 혼합된 잡음신호는 관심음원보다 정규분포에 더 가깝다고 할 수 있다. 이와같이 정규분포에 더 가까운 잡음신호만을 추출하려면 역엔트로피를 최소화해야 하는데, 이를 위해서 최대경사법을 이용한 관심음원 분리 알고리즘과 반대 부호로 학습한다.

[0069] 상기 벡터 $\mathbf{w}^{(k)}$ 의 학습 규칙은 수학식 21 및 22와 같다.

수학식 21

$$\Delta \mathbf{w}(\mathbf{k}) \propto E\{ \mathbf{z}(\mathbf{k}, \tau) [((\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau))^* g(\sum_k |((\mathbf{w}(\mathbf{k}))^H \mathbf{z}(\mathbf{k}, \tau)|^2))] \}$$

[0070]

수학식 22

$$\mathbf{w}(\mathbf{k}) \leftarrow \mathbf{w}(\mathbf{k}) / \|\mathbf{w}(\mathbf{k})\|$$

[0071]

[0072] 상기 수학식 21과 22에서 $\mathbf{w}$ 는 관심음원을 제거하기 위한 벡터, $\mathbf{z}$ 는 백색화된 입력 벡터, k 는 주파수 인덱스이다. 또한 g 는 비선형 함수 G 의 미분 결과 함수이다.

[0073] [DUET(Degenerate Unmixing Estimation Technique) 히스토그램을 통한 벡터 초기화]

[0074] 상기한 벡터 $\mathbf{w}^{(k)}$ 의 초기화는 DUET(Degenerate Unmixing Estimation Technique) 히스토그램을 이용한다(12단계).

[0075] 반향 환경을 고려한 DUET 알고리즘에 있어서 두 개의 마이크에 입력된 혼합 신호는 다음과 같이 나타낼 수 있다.

수학식 23

$$x_1(t) = \sum_{j=1}^N h_{1j}(t) * s_j(t)$$

$$x_2(t) = \sum_{j=1}^N h_{2j}(t) * s_j(t)$$

[0076]

수학식 24

$$\begin{aligned} \begin{bmatrix} X_1(k, \tau) \\ X_2(k, \tau) \end{bmatrix} &= \begin{bmatrix} H_{11}(k) \cdots H_{1N}(k) \\ H_{21}(k) \cdots H_{2N}(k) \end{bmatrix} \begin{bmatrix} S_1(k, \tau) \\ \vdots \\ S_N(k, \tau) \end{bmatrix} \\ &= \begin{bmatrix} 1 & \cdots & 1 \\ \frac{H_{21}(k)}{H_{11}(k)} & \cdots & \frac{H_{2N}(k)}{H_{1N}(k)} \end{bmatrix} \begin{bmatrix} S_1(k, \tau) \\ \vdots \\ S_N(k, \tau) \end{bmatrix} \\ &= \begin{bmatrix} 1 & \cdots & 1 \\ a_1(k)e^{-i\omega_k\delta_1(k)} & \cdots & a_N(k)e^{-i\omega_k\delta_N(k)} \end{bmatrix} \begin{bmatrix} S_1(k, \tau) \\ \vdots \\ S_N(k, \tau) \end{bmatrix} \end{aligned}$$

[0077]

[0078] 수학식 23의 $x_1(t)$ 과 $x_2(t)$ 는 각 마이크의 입력 신호이고, $h_{ij}(t)$ 는 음원과 마이크 사이의 채널 필터이다. 수학식 24의 $X_1(k, \tau)$ 과 $X_2(k, \tau)$ 는 단구간 푸리에 변환을 통한 $x_1(t)$ 과 $x_2(t)$ 의 시간-주파수 영역 성분이고, k 와 τ 는 각 시간-주파수 영역의 주파수와 시간 프레임 인덱스이며, ω_k 는 k 번째 주파수 영역의 정규화 주파수이다. $H_{ij}(k)$ 는 푸리에 변환을 통한 h_{ij} 의 주파수 영역 성분이고, a 와 δ 는 각 음원 신호에 관계된 파라미터로서 마이크의 제1채널 입력과 제2채널 입력의 세기 차이를 나타내는 감쇄 변수 및 제 1채널 입력과 제 2채널 입력의 도달 지연 시간의 차이를 나타내는 시간 지연 변수를 각각 나타낸다.

[0079] 상기 DUET 알고리즘은 특정 시간 주파수 영역에서는 지배적인 음원 신호가 최대 하나까지 존재한다는 W-DO(W-Disjoint Orthogonality) 가정에 기반하고 있다.

[0080] 상기 W-DO 가정에 의해 특정 시간-주파수 (k, τ) 에서 관심 음원 신호 $S(k, \tau)$ 가 지배적일 경우에 수학식 24는 수학식 25로 근사화된다.

수학식 25

$$\begin{bmatrix} X_1(k, \tau) \\ X_2(k, \tau) \end{bmatrix} \approx \begin{bmatrix} 1 \\ A(k) \end{bmatrix} S(k, \tau)$$

[0081]

[0082] 상기 수학식 25에서 $A(k)$ 는 관심음원 $S(k, \tau)$ 에 대한 파라미터로 얻어진 복소수 $a(k)e^{-i\omega_k\delta(k)}$ 이고, $S(k, \tau)$ 는 관심음원이고, $X_1(k, \tau)$ 와 $X_2(k, \tau)$ 는 각 마이크 입력의 시간-주파수 영역의 값이며, ω_k 는 k 번째 주파수 영역의 정규화 주파수이다.

[0083] 상기 시간 주파수 (k, τ) 에서 관심 음원 신호에 대한 파라미터 a 와 δ 는 다음의 수식을 통해 추정된다.

수학식 26

$$(\hat{a}(k), \hat{\delta}(k)) \approx \left(\frac{|X_2(k, \tau)|}{|X_1(k, \tau)|}, \frac{1}{\omega_k} \text{Im} \left\{ \ln \left(\frac{X_1(k, \tau)}{X_2(k, \tau)} \right) \right\} \right)$$

[0084]

[0085] 상기 수학식 26에서 $X_1(k, \tau)$ 와 $X_2(k, \tau)$ 는 각 마이크 입력의 시간-주파수 영역의 값이고, $\hat{a}(k)$ 는 관심 음원 신호에 대한 마이크 입력 간 시간-주파수 영역의 감쇄 값이고, $\hat{\delta}(k)$ 는 관심 음원 신호에 대한 마이크 입력 간 시간-주파수 영역의 시간 지연 값이다.

[0086] 두 음원 신호가 혼합된 마이크 입력을 이용하여 특정 주파수의 모든 시간 프레임에서 수학식 26의 우변식을 이용하여 감쇄 및 시간지연 파라미터를 추정하고 이 값으로 히스토그램을 생성한 것을 도시한 도 4를 참조하면, 두 음원에 해당되는 봉우리가 형성된다. 관심 음원 신호가 다른 여러 잡음원 신호에 비하여 전반적으로 우세한 에너지를 가지고 있다고 가정하면 히스토그램에서 가장 높은 봉우리는 관심 음원 신호에 관계된 파라미터에 대응될 것이다.

[0087] 따라서 특정 주파수의 모든 시간 프레임에서 수학식 26의 우변식을 이용하여 감쇄 및 시간지연 파라미터를 추정하고 이 값으로 히스토그램을 생성한 후 가장 높은 봉우리에 해당하는 감쇄 및 시간지연 값을 관심 음원에 대한 파라미터로 가정한다. 만약 각 주파수별로 히스토그램을 생성함에 있어서 시간 프레임의 개수가 충분하지 않을 경우 DUET 히스토그램을 이차원 가우시안 윈도우로 2차원 필터링하여 스무딩(smoothing)한 후 봉우리에 해당하는 파라미터를 추정한다.

[0088] 그리고 상기 관심음원에 대한 파라미터로부터 얻어진 복소수 $A(k) = a(k)e^{-i\omega_k \delta(k)}$ 의 역수를 취하여 관심 음원 제거를 위한 벡터 $\mathbf{w}(k)$ 를 아래 수학식 27과 같이 초기화할 수 있다. 단, ω_k 는 k 번째 주파수 영역의 정규화 주파수이다.

수학식 27

$$\mathbf{w}(k) = \frac{1}{\sqrt{1 + a^2(k)}} \begin{bmatrix} 1 \\ -A^{-1}(k) \end{bmatrix}$$

[0089]

[0090] [관심음원 제거]

[0091] 상기한 바와 같이 초기화 및 학습된 관심 음원 제거를 위한 벡터 $\mathbf{w}(k)$ 를 토대로 입력 혼합신호로부터 관심음원 신호를 제거하여 혼합된 잡음신호를 생성한다(14단계). 여기서, 두 개의 마이크 각각으로부터의 입력 혼합신호는 단구간 푸리에 변환되어 시간-주파수 영역으로 변환된 후에 벡터 $\mathbf{w}(k)$ 에 의해 관심음원신호가 제거되어 혼합된 잡음신호를 생성한다.

[0092] [혼합된 잡음신호와 입력 혼합신호를 이용한 마스크 생성]

[0093] 상기한 바와 같이 혼합된 잡음신호가 생성되면, 이 신호와 입력 혼합신호를 시간-주파수 영역에서 비교하여 마스크를 생성한다(16단계). 상기 입력 혼합신호와 혼합된 잡음신호를 각각 단구간 푸리에 변환을 통해 시간-주파수 영역에서 표현하고 각 시간-주파수 (k, τ) 에서 입출력 에너지 비율(OIR: Output-to-Input energy Ratio)을 수학식 28과 같이 구한다.

수학식 28

$$\text{OIR}(k, \tau) = \frac{|U(k, \tau)|^2}{|X_1(k, \tau)|^2}$$

[0094]

[0095] 상기 수학식 28에서, $\text{OIR}(k, \tau)$ 는 각 시간-주파수 (k, τ) 에서 입출력 에너지 비율을 나타내며, $X_1(k, \tau)$ 는 입력 혼합신호이고, $U(k, \tau) = \mathbf{w}^H(\mathbf{k}) \mathbf{z}(k, \tau)$ 는 전체 혼합신호에서 관심음원 성분을 제거한 출력신호를 나타낸다.

[0096] 실제 입출력 에너지 비율은 전체 혼합 신호 $X_1(k, \tau)$ 의 에너지 중 관심음원 성분을 제거한 신호 $U(k, \tau)$ 의 에너지 비율을 나타내기 때문에, 이 값이 크다면 $X_1(k, \tau)$ 에 관심 음원 성분이 적게 포함되어 있음을 의미하고, 이 값이 작다면 관심 음원 성분이 많이 포함되어 있음을 의미한다.

[0097] 따라서 각 시간-주파수에서 입출력 에너지 비율을 계산하여 적절한 문턱 값 $\theta(k)$ 를 설정하면 수학식 29과 같이 이진 마스크를 구성할 수 있다.

수학식 29

$$\Omega(k, \tau) = \begin{cases} 1, & \text{OIR}(k, \tau) \leq \theta(k) \\ 0, & \text{otherwise} \end{cases}$$

[0098]

[0099] 상기 수학식 29에는 $\Omega(k, \tau)$ 는 이진 마스크이고, $\text{OIR}(k, \tau)$ 는 각 시간-주파수 (k, τ) 에서 입출력 에너지 비율이고, $\theta(k)$ 는 임의로 설정한 문턱 값이다.

[0100] [입력 혼합신호와 이진 마스크를 이용한 음성인식 과정]

[0101] 상기한 바와 같이 획득된 음원 신호에 대해 음성 인식하는 과정을 도 5의 음성인식장치의 구성도를 참조하여 설명한다.

[0102] 상기 음성인식장치는 제1 및 제2마이크(100, 102) 및 신호처리장치(104)로 구성된다.

[0103] 상기 제1 및 제2마이크(100, 102) 각각은 미리 정해진 거리만큼 이격된 상태에서 다수의 음원으로부터의 혼합신호를 입력받아 신호처리장치(104)의 입력신호로 제공한다.

[0104] 상기 신호처리장치(104)는 관심음원 성분 구분을 위한 마스크 생성부(106)와 클러스터 기반 손실특징 복원부(108)와 MFCC 변환부(110)와 음성인식부(112)로 구성된다.

[0105] 상기 관심음원 성분 구분을 위한 마스크 생성부(106)는 본 발명에 따라 상기 제1 및 제2마이크(100, 102)로부터의 입력신호를 입력받아, 입력신호에서 관심음원 성분을 제거하여 혼합된 잡음신호를 생성하고, 그 혼합된 잡음신호와 입력신호 사이의 에너지 비에 따른 이진 마스크를 생성한다.

[0106] [클러스터 기반의 손실 특징 복원기법]

[0107] 상기 클러스터 기반 손실특징 복원부(108)의 동작을 설명한다.

[0108] 일반적으로 잡음의 영향을 받은 음성신호는 시간-주파수 영역에서 로그 스펙트럼 신호로 나타낼 때 다음과 같이 나타낼 수 있다.

수학식 31

[0109]
$$Y(l, k) = X(l, k) + N(l, k)$$

[0110] 상기 수학식 31에서 $Y(l, k)$, $X(l, k)$ 와 $N(l, k)$ 는 각각 l 번째 프레임에서 k 번째 주파수 밴드의 관측 신호, 잡음의 영향이 없는 음성 신호와 잡음신호의 로그 스펙트럼 성분을 나타낸다.

[0111] 여기서, 상기 $Y(l, k)$ 의 값이 잡음에 의한 영향이 강해서 신뢰성 없는(unreliable) 성분이라고 판단되는 벡터를 $Y_u(l, k)$ 로 나타내고, 잡음에 의한 영향이 작아서 신뢰성 있는(reliable) 성분이라고 판단되는 벡터를 $Y_r(l, k)$ 로 나타낸다고 가정한다.

[0112] 잡음의 영향이 적은 $Y_r(l, k)$ 이 주어졌을 때, 이 성분 속에 포함된 $X_r(l, k)$ 에 대한 적절한 근사치는 바로 $Y_r(l, k)$ 라고 볼 수 있다. 하지만 $Y_u(l, k)$ 에 포함된 $X_u(l, k)$ 는 $N_u(l, k)$ 의 영향이 매우 크기 때문에 쉽게 추정할 수 없다. 그러나 로그 스펙트럼 성분이라는 점을 고려하여 수학식 32와 같은 제한 조건을 가정할 수 있다.

수학식 32

[0113]
$$\begin{aligned} X_r(l, k) &= Y_r(l, k) \\ X_u(l, k) &\leq Y_u(l, k) \end{aligned}$$

[0114] 이러한 특성을 이용하여 손실 특징 복원 기술을 적용하는 방법은 크게 두 가지로 분류할 수 있다.

[0115] 첫 번째 방법은 인식 시스템 수정의 대표적인 접근 방법인 마지널라이제이션(marginalization)이 있는데, 이는 HMM(Hidden Markov Model) 기반의 음성인식 시스템 내부에서 신뢰성 없는 성분에 대한 확률계산을 주변(marginal) 확률분포 값으로 추정하여 인식 수행을 시도하는 것이다. 즉, 임의의 프레임에서 잡음의 영향을 받은 관찰 벡터 Y 가 주어졌을 때, 주어진 제한 조건을 따르는 신뢰성 있는 성분 X_r 과 신뢰성 없는 성분 X_u 로 구성되는 음성에 대한 벡터 X 가 내재되어 있다고 가정하면, 음성 인식 시스템에서 현재 상태(state) s 에 대한 주어진 음성에 대한 관찰 벡터 X 의 출력 확률은 수학식 33로 나타낼 수 있다.

수학식 33

[0116]
$$\begin{aligned} P(X|s) &= P(X_r = Y_r, -\infty \leq X_u \leq Y_u | s) \\ &= \prod_{j \in R} \frac{1}{\sqrt{2\pi} \sigma_s(j)} e^{-\frac{(X(j) - \mu_s(j))^2}{2\sigma_s^2(j)}} \times \prod_{i \in U} \int_{-\infty}^{Y(i)} \frac{1}{\sqrt{2\pi} \sigma_s(i)} e^{-\frac{(x - \mu_s(i))^2}{2\sigma_s^2(i)}} dx \end{aligned}$$

[0117] 상기 수학식 33에서, Y_r 과 Y_u 는 관찰 벡터 Y 에서 각각 신뢰성 있는 성분으로 구성된 벡터와 신뢰성 없는 성분으로 구성된 벡터를 나타내고, X_r 과 X_u 는 이에 대응되는 잡음의 영향이 없는 내재된 음성신호의 일부분이며, X 는 X_r 과 X_u 를 결합한 내재된 음성신호 벡터를 나타낸다. R 과 U 는 신뢰성 있는 벡터 성분의 인덱스 집합과 신뢰성 없는 벡터 성분의 인덱스 집합을 나타내며, $Y(j)$ 는 관찰 벡터 Y 의 j 번째 성분을 나타내고, $\sigma_s^2(j)$ 와 $\mu_s(j)$ 는 현재 상태(state) s 에 따른 가우시안 분포에서 j 번째 인덱스의 분산과 평균을 나타낸다.]

[0118] 인식 시스템 수정 방법은 잡음 음성에 있어서 로그 스펙트럼 모델을 이용한 강인한 음성 인식을 수행하기 위해 직관적이고 그럴 듯한 방법이 될 수 있지만, 인식 시스템과 특징이 직접적으로 연관되어야 하므로 cepstrum 특징을 사용하기 어려워 기본적인 인식 성능이 떨어지는 단점이 있다.

[0119] 두 번째 방법으로는 특징 보상의 대표적인 접근 방법인 클러스터 기반 손실 특징 복원 방법이 있는데, 이의 기본 가정은 시간-주파수 영역에서 각 프레임의 스펙트럼 벡터가 서로 독립이라는 것이다. 이러한 가정 하에 하나의 스펙트럼 벡터 내에서 성분들 간의 연관성을 이용하여 신뢰성 없는 성분을 복원하게 된다.

[0120] 이를 위해 먼저 잡음이 없는 환경에서 얻어진 학습데이터로 수학식 34와 같은 가우시안 혼합 모델 분포를 얻어낸다.

수학식 34

$$P(\mathbf{X}) = \sum_{\nu} c_{\nu} \frac{1}{\sqrt{(2\pi)^d |\Theta_{\nu}|}} \exp\left(-\frac{1}{2}(\mathbf{X} - \mu_{\nu})^T \Theta_{\nu}^{-1} (\mathbf{X} - \mu_{\nu})\right)$$

[0121]

[0122] 상기 수학식 34에서, $P(\mathbf{X})$ 는 가우시안 혼합 모델 분포이고, d 는 벡터의 차원이고, $c_{\nu}, \mu_{\nu}, \Theta_{\nu}$ 는 ν 번째 가우시안의 가중치, 평균벡터, 공분산행렬을 나타낸다. 이러한 분포에 대한 파라미터들은 학습데이터로부터 EM(Expectation Maximization) 알고리즘을 통해 학습된다.

[0123] 이러한 가우시안 혼합 모델이 결정되면, 신뢰성 없는 성분들이 포함된 스펙트럼 벡터 $\mathbf{Y}$ 가 주어졌을 때, 우리가 찾고자 하는 잡음의 영향이 없는 내재된 음성신호 벡터 $\mathbf{X}$ 를 찾기 위해서 다음과 같은 계산과정을 수행한다. 먼저 $\mathbf{X}_r$ 은 신뢰성 있는 성분들 $\mathbf{Y}_r$ 과 같다고 설정하고, 신뢰성 없는 성분들 $\mathbf{Y}_u$ 에 대응되는 $\mathbf{X}_u$ 에 대한 추정 값은 바운디드(bounded) MAP 추정 방법을 이용하여 수학식 35에 따라 구한다.

수학식 35

$$\hat{\mathbf{X}}_u = \underset{\mathbf{X}_u}{\operatorname{argmax}} P(\mathbf{X}_u | \mathbf{X}_r = \mathbf{Y}_r, \mathbf{X}_u \leq \mathbf{Y}_u)$$

[0124]

[0125] 이를 구하기 위한 반복적인 과정은 다음과 같다.

[0126] 1) 먼저 추정할 $\bar{\mathbf{X}}(k)$ 를 $\bar{\mathbf{X}}(k) = \mathbf{Y}(k), \forall k$ 로 초기화한다.

[0127] 2) 각각의 $\mathbf{X}(i), i \in U$ 을 수학식 36에 따라 산출한다.

수학식 36

$$\begin{aligned} \tilde{\mathbf{X}}(k) &= \underset{\mathbf{X}(k)}{\operatorname{argmax}} P(\mathbf{X}(k) | \mathbf{X}(i) = \bar{\mathbf{X}}(i), i \neq k) \\ &= \mu(k) + \frac{1}{\sigma^2(k)} \Theta_{\mathbf{X}(k), \bar{\mathbf{X}}} (\bar{\mathbf{X}} - \bar{\mu}) \\ \bar{\mathbf{X}}(k) &= \min(\tilde{\mathbf{X}}(k), \mathbf{Y}(k)) \end{aligned}$$

[0128]

[0129] 상기 수학식 36에서 $\bar{\mathbf{X}}$ 는 $\mathbf{X}(i), i \in U$ 를 성분으로 갖는 벡터를 나타내며, $\bar{\mu}$ 는 $\bar{\mathbf{X}}$ 의 평균을 나타낸다. 그리고 $\mu(k), \sigma^2(k)$ 은 $\mathbf{X}(k)$ 에 대한 평균 및 분산을, $\Theta_{\mathbf{X}(k), \bar{\mathbf{X}}}$ 은 $\mathbf{X}(k)$ 와 $\bar{\mathbf{X}}$ 간의 상호 공분산을 나타내는 행벡터이다.

[0130] 3) $\bar{X}^{(k)}$ 가 수렴할 때까지 상기 2)단계를 반복하며, 최종적으로 $\hat{X}_u$ 는 수렴된 $\bar{X}(i), i \in U$ 가 된다.

[0131] 그러나 가우시안 혼합 모델로부터 $\hat{X}_u$ 를 직접 추정하는 것은 어려운 일이므로, 각 가우시안에 대한 조건부 바운디드(bounded) MAP 추정 값의 선형 조합으로 근사화한다.

수학식 37

[0132]
$$\hat{X}_u = \sum_{\nu} P(\nu | X_r, X_u \leq Y_u) \hat{X}_u^{\nu}$$

수학식 38

[0133]
$$\hat{X}_u^{\nu} = \underset{X_u}{\operatorname{argmax}} P(X_u | X_r = Y_r, X_u \leq Y_u; \mu_{\nu}, \Theta_{\nu})$$

수학식 39

[0134]
$$P(\nu | X_r, X_u \leq Y_u) = \frac{c_{\nu} P(X_r, X_u \leq Y_u | \nu)}{\sum_j c_j P(X_r, X_u \leq Y_u | j)}$$

수학식 40

[0135]
$$P(X_r, X_u \leq Y_u | \nu) = P(X_r, -\infty \leq X_u \leq Y_u; \mu_{\nu}, \Theta_{\nu})$$

$$= \prod_{j \in R} \frac{1}{\sqrt{2\pi} \sigma_{\nu}(j)} e^{-\frac{(X(j) - \mu_{\nu}(j))^2}{2\sigma_{\nu}^2(j)}} \times \prod_{i \in U} \int_{-\infty}^{Y(i)} \frac{1}{\sqrt{2\pi} \sigma_{\nu}(i)} e^{-\frac{(x - \mu_{\nu}(i))^2}{2\sigma_{\nu}^2(i)}} dx$$

[0136] 상기 수학식 37~40에서 $\hat{X}_u$ 은 상기 알고리즘을 통해 찾는 신뢰성 없는 출력 특징의 성분이며, Y_r 과 Y_u 는 관찰 벡터 Y 에서 각각 신뢰성 있는 성분으로 구성된 벡터와 신뢰성 없는 성분으로 구성된 벡터를 나타내고, X_r 과 X_u 는 이에 대응되는 잡음의 영향이 없는 내재된 음성신호의 일부분임. R 과 U 는 신뢰성 있는 벡터 성분의 인덱스 집합과 신뢰성 없는 벡터 성분의 인덱스 집합을 나타내며, $X(j)$ 와 $Y(j)$ 는 각각 음성신호 벡터 X 와 관찰 벡터 Y 의 j 번째 성분을 나타내고, $c_{\nu}, \mu_{\nu}, \Theta_{\nu}$ 는 ν 번째 가우시안의 가중치, 평균벡터, 공분산행렬을 나타내며, $\sigma_{\nu}^2(j)$ 와 $\mu_{\nu}(j)$ 는 ν 번째 가우시안 분포에서 j 번째 인덱스의 분산과 평균을 나타냄. Y_r 과 Y_u 는 관찰 벡터 Y 에서 각각 신뢰성 있는 성분과 신뢰성 없는 성분을 나타냄. X_r 과 X_u 는 대응되는 성분에서 잡음의 영향이 없는 내재된 음성신호 부분을, X 는 X_r 과 X_u 를 결합한 내재된 음성신호 벡터를 나타내고, R 과 U 는 신뢰성있는 벡터 성분과 신뢰성없는 벡터 성분의 인덱스 집합을 나타내며, $\sigma_{\nu}^2(j)$ 와 $\mu_{\nu}(j)$ 는 현재 상태(state) ν 에 따른 가우시안 분포에서 j 번째 인덱스의 분산과 평균을 나타낸다.

[0137] 상술한 바와 같이 학습 데이터들로부터 EM 알고리즘으로 학습된 가우시안 혼합 모델의 파라미터로 수학식 37과 같이 바운디드 MAP 추정 방법을 이용하여 각 가우시안 모델의 사후 확률을 구해 그것을 중첩하여 손실 특징을 복원하게 된다.

- [0138] 상기한 클러스터 기반 손실 특징 복원 방법은 공분산 기반 손실 특징 복원 방법에 비하여 모델이 단순하여 적은 학습 데이터를 필요로 하면서도 효과적인 복원 성능을 나타낸다.
- [0139] 또한 이러한 특징 복원 방법은 인식 시스템 수정 방법과 달리 인식 시스템과 별개로 구현이 가능하므로 인식 시스템을 수정하는 부담없이 캡스트림을 자유롭게 활용할 수 있어서, 기본적인 인식 성능 저하없이 강인한 인식 성능을 나타낸다.
- [0140] [MFCC 변환]
- [0141] 상기한 바와 같이 클러스터 기반 손실 특징이 복원된 관심음원은 MFCC 변환부(110)에 입력된다.
- [0142] 상기 MFCC 변환부(110)는 입력된 스펙트럼 영역에서의 관심음원의 특징정보를 캡스트림 변환하여 MFCC로 변환한다.
- [0143] [음성인식]
- [0144] 상기 MFCC 변환된 특징정보는 음성인식부(112)로 입력된다.
- [0145] 상기 음성인식부(112)는 상기 MFCC 변환된 특징정보에 따른 음성인식을 수행한다.
- [0146] 상기한 바와 같은 본 발명에 따른 성능평가에 대해 설명한다.
- [0147] [실험 환경 및 성능 평가 기준]
- [0148] 도 6는 본 발명의 평가를 위한 세트의 개략적인 구조도이다.
- [0149] 본 발명의 평가를 위해서 RM 데이터베이스의 음성인식 세트를 이용하였다. 시뮬레이션은 실내 충격 응답 모델을 사용하여 데이터베이스의 테스트 음원과 잡음을 섞어서 얻었다. 잡음 신호는 NoiseX의 많은 사람이 떠드는 소리인 배블 잡음을 사용하였다. 실내의 크기는 가로 4m, 세로 5m, 높이 3m이며 마이크의 위치는 가로 2.5m, 세로 2m에 20cm간격으로 위치하며 M으로 표시하였다. 음원의 위치는 두 마이크의 중심에서 왼쪽으로 30도, 1미터 떨어진 위치에 위치하며 S로 표시하였다. 잡음의 위치는 두 마이크의 중심에서 각각 0도, 60도, 120도, 180도, 240도, 300도 위치에 1m 떨어진 곳에 위치하며 N으로 표시하였다. 각 잡음은 모두 다른 형태의 배블 잡음이며 각 음원마다 무작위로 다른 배블 잡음을 섞었다. 이렇게 섞은 입력 신호를 여러 방법으로 테스트했는데, 각 입력 신호마다 0dB, 5dB, 10dB, 20dB의 4가지 SNR로 테스트하고, 반향 시간 RT60를 0.1초, 0.3초, 0.5초의 3가지로 테스트하였다.
- [0150] [관심음원 제거 실험]
- [0151] 도 7은 테스트 신호 중 입력 SNR이 5dB인 데이터 중 하나를 도시한 것이다. 도 7(a)는 음원신호의 파형이고, 도 7(b)는 잡음신호의 파형이다. 도 7(c)는 파란색 파형은 입력신호의 파형이고 붉은색 파형은 출력신호의 파형이다.
- [0152] 도 7(c)에서 도시한 바와 같이 출력파형은 입력파형에서 음원신호가 제거된, 잡음신호로 구성된다.
- [0153] 도 8은 입력 SNR 별, 반향 시간 별 출력 SNR의 값의 평균을 나타낸 표를 도시한 것이다. 상기 도 8의 표에 도시한 바와 같이 입력 SNR에 비해 출력 SNR이 상당량 줄어든 것을 볼 수 있는데, 이것은 출력 신호에서 음원의 세기가 크게 줄어, 잡음 신호만이 출력되었다는 것을 의미한다. 그러므로 본 발명에 의해 관심음원을 제거하는 것이 가능하다.
- [0154] [음성인식 실험]
- [0155] 도 9는 본 발명에 따라 관심음원 제거 알고리즘을 통해 얻어진 출력신호와, 입력신호를 통해 얻은 OIR 마스크에

입력신호를 통과시킨 출력을 클러스터 기반의 손실 특징 복원을 수행하여 HMM 기반의 음성 인식기를 이용해 음성인식을 수행한 결과를 나타낸 표를 도시한 것이다.

[0156] 베이스 라인(Baseline)은 본 발명에 따른 관심음원 제거 알고리즘을 사용하지 않고, 입력 신호만을 이용해 음성인식을 수행한 결과이다. 그리고 본 발명에 따른 관심음원 제거 알고리즘을 통해 테스트한 어떤 입력 SNR이나 어떤 반향 시간에서도 음성인식 결과가 향상되었다. 이를 통해 실제 환경에서 관심음원 제거 알고리즘을 통해 음성인식의 성능을 향상시킬 수 있음을 알 수 있다.

[0157] [결론]

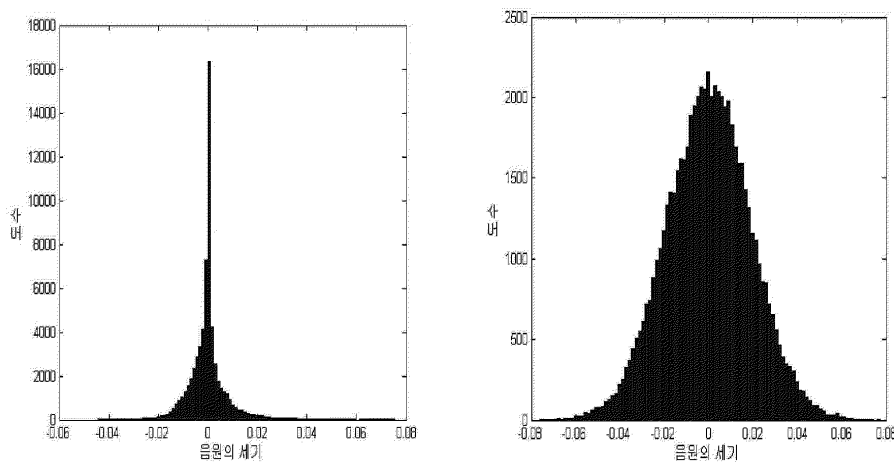
[0158] 상기한 바와 같은 실험 결과, 본 발명에 따른 독립벡터분석 기반의 관심음원 제거 알고리즘과 OIR 마스크, 그리고 이것을 이용한 클러스터 기반의 손실 특징 복원기법을 이용하여 관심음원을 제거하거나 관심음원 신호를 인식하는 알고리즘은 다양한 신호 대 잡음비의 환경과 다양한 반향 환경에서, 그리고 마이크보다 많은 잡음원들이 존재하는 환경에서도 단지 마이크 두 개만으로도 효과적으로 관심음원을 제거해낼 수 있다. 또한 이를 이용하여 음성인식을 수행했을 때 더 좋은 성능의 음성인식 결과를 보인다.

부호의 설명

[0159] 100 : 제1마이크
102 : 제2마이크
104 : 신호처리장치

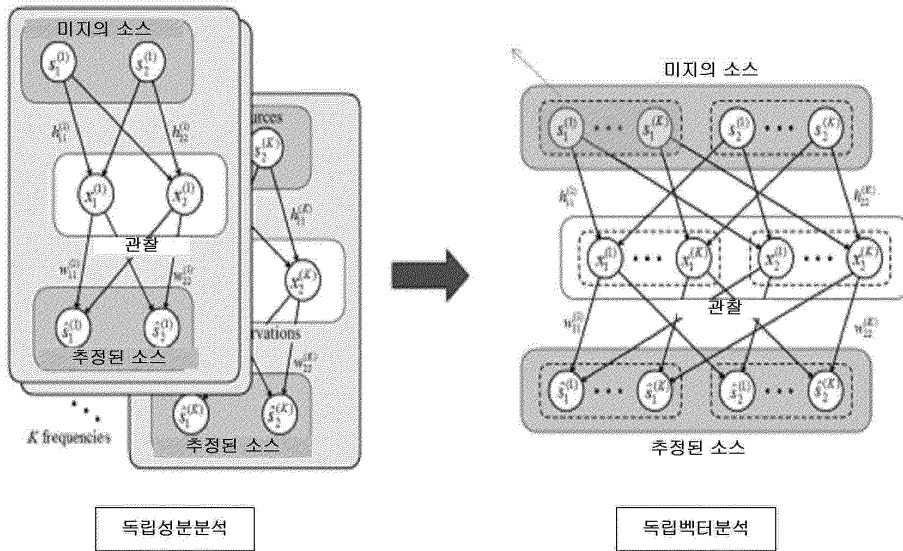
도면

도면1

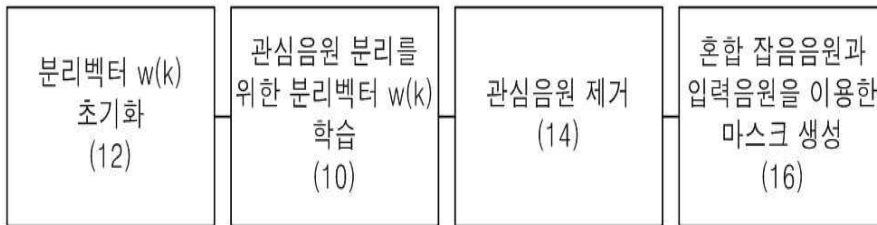


사람의 발성 음원과 많은 사람들이 한꺼번에 떠드는 소리인
배경잡음의 히스토그램

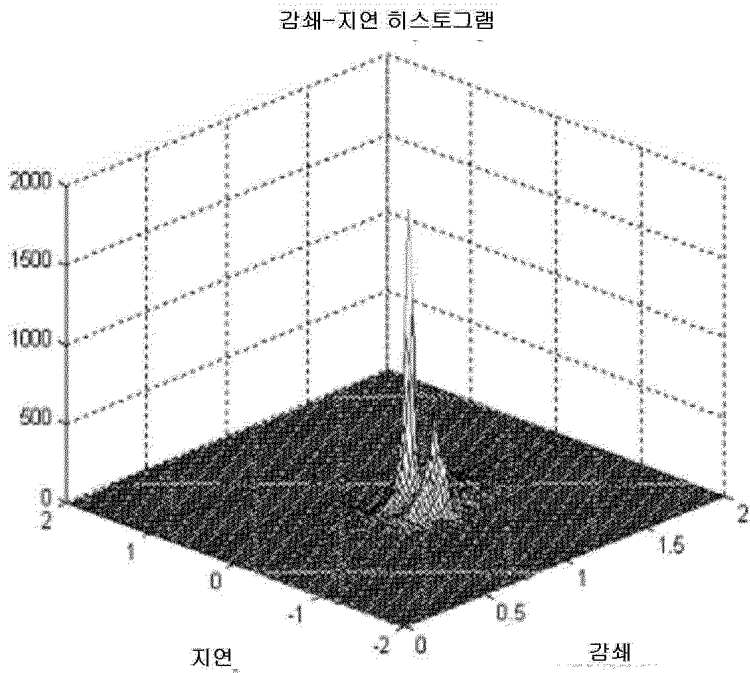
도면2



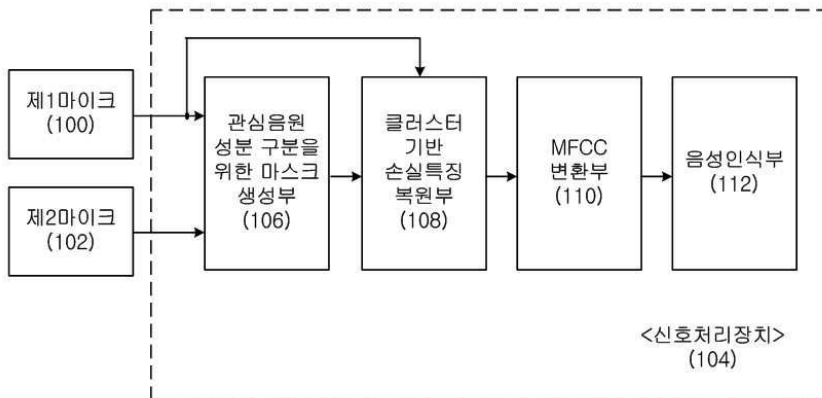
도면3



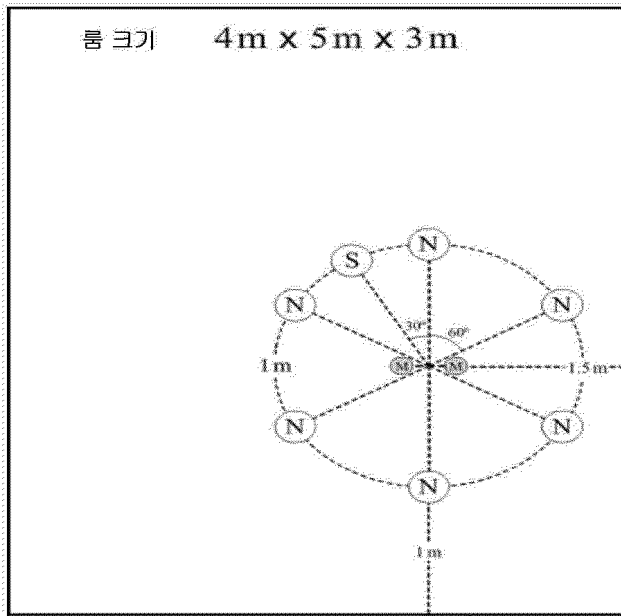
도면4



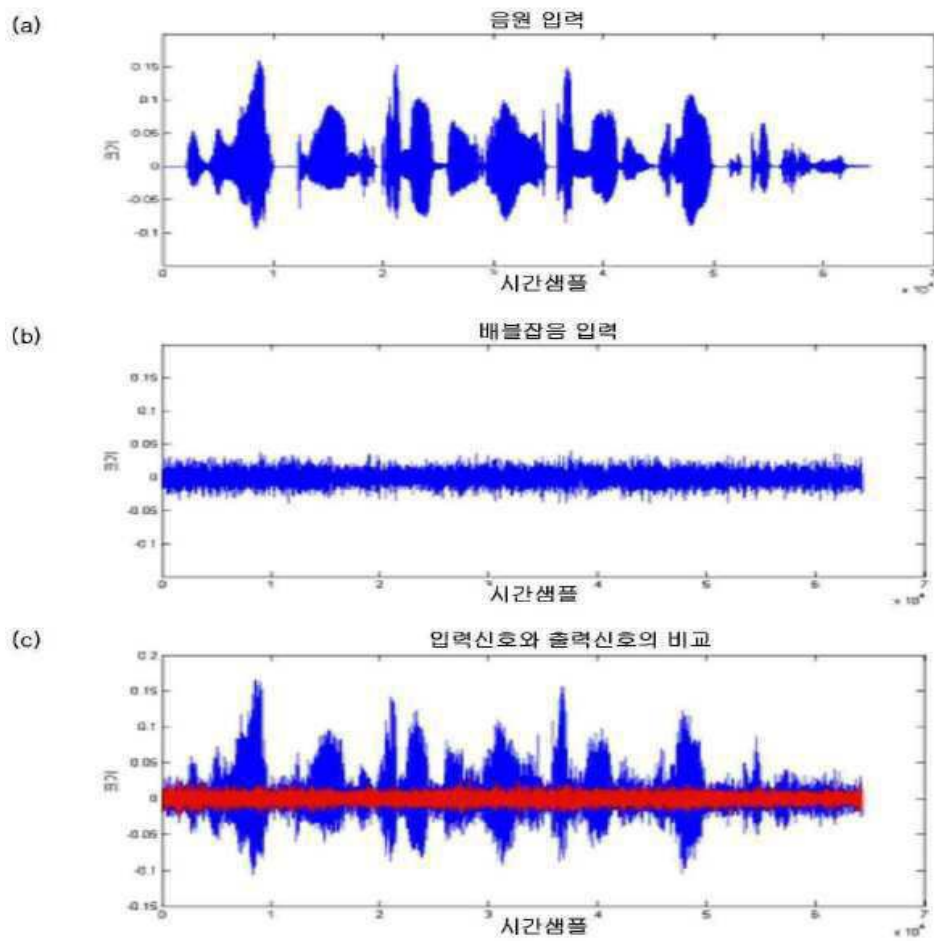
도면5



도면6



도면7



도면8

(dB)	SNR _{in} = 0dB		SNR _{in} = 5dB		SNR _{in} = 10dB		SNR _{in} = 20dB	
	SNR _{out}	SNR _{out} - SNR _{in}	SNR _{out}	SNR _{out} - SNR _{in}	SNR _{out}	SNR _{out} - SNR _{in}	SNR _{out}	SNR _{out} - SNR _{in}
RT60=0.1s	-11.67	-11.67	-9.86	-14.86	-7.37	-17.37	-0.43	-20.43
RT60=0.3s	-6.17	-6.17	-1.92	-6.92	3.05	-6.95	14.34	-5.66
RT60=0.5s	-3.51	-3.51	1.08	-3.92	6.12	-3.88	17.03	-2.97

도면9

(%)	SNR _{in} = 0dB		SNR _{in} = 5dB		SNR _{in} = 10dB		SNR _{in} = 20dB	
	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
RT60=0.1s	5.26	15.38	18.43	43.97	52.71	70.64	89.69	93.44
RT60=0.3s	4.75	9.81	13.86	29.32	42.37	58.65	84.38	88.60
RT60=0.5s	5.02	7.38	8.84	17.69	25.50	40.73	67.04	75.75