포르만트 주파수를 이용한 한국어 음성의 자동 인식에 관한 연구

83306

김 순 협*
광운대학 전자계산기공학과 교수

박 규 태**
연세대학교 전자계산기 공학과 교수

A Study on the Automatic Recognition of Korean Speech by Formant Frequency.

(Abstract)

In Speech signal processing, ARMA spectral estimation method is used. It has been demonstrated that the ARMA model provides better spectral estimation then the more specialized AR model and MA model.

Dynamic program is used to achieve time alignment. Speech sound similarity is defined to be proportional to the distance seperating to sound in a Vector space defined by ARMA model.

As a result, the recognition rate of 97.3% for three speaker is obtained.

<본문>

(1) 본 연구의 목적은 자동회귀 이동 평균 모형 (Autoregressive Moving Average model ; 이하 'ARMA' 모형 라라함) 방식을 이용하여 한국어의 음성을 분석하고, 단독 숫자음을 대상으로한 인식 방법과 인식 시스템을 제시하는데 있다.

(2) 시간 영역에서의 음성 신호 분석 방법은 영교차율과 대수 에너지를 이용하였다.

영교차율은 분석 구간 프레임내의 파형이 영점(zero) 축과 교차하는 회수를 말하며 일반적은 $Z = \sum_{m=1}^{N} [1 - sgn(X_{m+1}) \cdot sgn(X_m)] / 2$ ..... (1)

이고, 여기서 $sgn[X(m)] = \begin{cases} 1, & X(m) \geq 0 \\ -1, & X(m) < 0 \end{cases}$

으로 표시한다.

대수 에너지는 이산계(discrete system)에서 $E = \sum_{m=-\infty}^{\infty} X^2(m)$ 으로 표현되며 무성음으로 부터 유성음을 구분하는데 유용하다.

(3) DP 정합법의 기본 원리는 시험 패턴(Test pattern)과 표준 패턴(Reference pattern)을 비교하여 두 패턴들 사이의 유사성을 결정하는 것이며 유사도는 패턴 사이의 시간축 정규화와 거리 계산등으로 이루어 진다.

한편, 근본적인 음성 패턴의 시간적 구조에는 연속성, 단조증가, 그리고 음성학 라라이라의 변화 속도에 대한 조건이 따른다.

이러한 조건들은 warping 함수에 대한 ① 끝점 제약 또는 축 경계조건 ② 부분적 경로의 연속성 제약 ③ 전체 경로의 제약으로 만족될 수 있고, DTW 알고리즘의 최적 warping 경로를 위한 거리 측정(distance measurement)은 다음과 같은 식의 일반적 형태를 갖는다.

$$D(i(k), j(k)) = \frac{\sum_{k=1}^{K} d[i(k), j(k)] \cdot w(k)}{N(w)} \quad \cdots (3)$$

(4) 음성 신호의 PSD 추정 알고리즘에서 음성 생성 과정에 대한 정확한 모델을 ARMA 모형으로 표시하면

$$y_k = -\sum_{i=1}^{p} a_i y_{k-i} + \sum_{j=0}^{q} b_j u_{k-j} \quad \cdots\cdots (4)$$

이고 이식은 $y(m) = \sum_{i=0}^{q} b_i u(m-i) - \sum_{i=1}^{p} a_i y(m-i) \cdots (5)$

로 변형 가능하다. 여기서 출력 샘플링 값이 $y(1), y(2), \cdots, y(m)$ 인 시계열의 자동상관 함수는 $r_y(m) = E\{y(k)\cdot y^*(m-k)\} \cdots (6)$

로 되며 이식을 푸리에 변환 시키면 PSD를 추정 할 수 있다.

즉, $P_y(w) = \sum_{m=-\infty}^{\infty} r_y(m) e^{-jwm} \quad \cdots\cdots (7)$ 이다.

이것을 유리화 스펙트럼 모델로 표시하면

$$P_y(w) = \left| \frac{b_0 + b_1 e^{-jw} + \cdots + b_q e^{-jqw}}{1 + a_1 e^{-jw} + \cdots + a_p e^{-jpw}} \right|^2 \quad \cdots (8)$$

이 된다. 여기서 $b_0 = G$ 이며 이 식으로 부터 a_k와 C_k의 계수를 결정하여 음성 신호의 PSD를 추정 했다.

(5) 본 연구에서 사용된 실험 방법은 음성 신호를 4.8 [kHz]의 저역 통과 여과기(Low pass Filter)에 통과 시킨 다음 12 bit A/D 변환기의 입력 정격 전압이 되도록 증폭기를 구성하여 콤퓨터에 연결하여 인식시켰다.

이상의 방법으로 성인 화자 3인에 의해 발성된 숫자음에 대해 97.3%의 인식율을 얻었다.

앞으로는 화자 수와 어휘 수에 상관없는 음성 인식을 위해서 음성의 음소 분석과 음소 인식 실험이 반드시 선행되어야 할 것이다.