1999 年 度 大 韓 電 子 工 學 會 夏 季 綜 合 學 術 大 會 論 文 集 第 22 卷 第 1 號 99/6

고유영역을 이용한 문자독립형 화자인식에 관한 연구

함 철 배, 이 동 규, 이 두 수
한양대학교 전자공학과
전화 : (02) 2290-0358 / 팩스 : (02) 2298-1796

A Study On Text Independent Speaker Recognition Using Eigenspace

Chul Bae Ham, Dong Gyu Lee, Doo Soo Lee
Dept. of Electronic Engineering, Hanyang University
E-mail : isham@hymail.hanyang.ac.kr

Abstract

We report the new method for speaker recognition. Until now, many researchers have used HMM (Hidden Markov Model) with cepstral coefficient or neural network for speaker recognition. Here, we introduce the method of speaker recognition using eigenspace. This method can reduce the training and recognition time of speaker recognition system. In proposed method, we use the low rank model of the speech eigenspace. In experiment, we obtain good recognition result.

I. 서 론

사회가 복잡하고 다양해짐에 따라 개인의 신분을 확인해야하는 여러 가지 서비스들이 생겨났다. (신용카드 서비스와 같은 은행서비스, 전자자물쇠등이 이러한 예로 볼 수 있다.) 따라서, 개인의 신분을 증명하기 위한 여러 가지 방법이 제시 되었는데, 근래에는 신체의 특징을 이용한 홍채 인식, 지문 인식 및 음성을 통한 화자 인식(Speaker Recognition)등이 많이 시도되고 있다. 본 논문에서는 음성을 이용하여 화자를 인식하는 방법에 대해서 제시하였다.

화자 인식은 크게 특징벡터 추출 (Feature Extraction), 거리 측정 (Distance Measure), 화자 판단 (Decision)의 3가지단계로 나눌 수 있다[8]. 특징벡터로는 선형 예측 계수(LPC Coefficient)를 이용한 켑스트럼 계수와 반사 계수(Regression Coefficient)를 사용한 방법들이 연구되었고, 인식 방법으로는 은닉 마코프모델, 신경망, 벡터양자화가 많이 사용된다[4][6][7][8]. 기존의 방법들은 훈련시간(Training Time)과 인식시간(Recognition Time)이 많이 걸리고, 새로운 템플릿을 등록시킬 때 전체 템플릿에 대해 다시 훈련(Training)을 해야하는 단점이 있다. 본 논문에서는 이러한 단점을 보완한 고유값해석(Eigenvalue Analysis)을 이용한 화자 인식 방법을 제시한다. 제 2장에서는 화자 인식의 일반적인 방법에 대하여 소개를 하고, 제 3장에서는 고유벡터와 저차원 모델에 대하여 설명을 한다. 제 4장과 5장에서는 제안한 방법에 대한 자세한 설명과 모의실험결과를 보이고 제 6장에서는 결론을 맺겠다.

II. 화자 인식 시스템

화자 인식은 크게 화자 식별(Speaker Identification)과 화자 확인(Speaker Verification)의 2종류로 나눌 수가 있다[8]. 화자 식별은 이미 만들어져있는 여러개의 템플릿에서 입력음성이 어떤 사람의 음성인지를 식별해내는 과정이다. 반면에, 화자 확인의 경우는 이미 식별된 화자의 템플릿과 입력된 음성의 특징벡터가 어느 정도 일치하는지를 판정해 승인(Accept/Reject)하는 단계이다. 일반적으로 화자 인식 시스템은 이 두 가지를

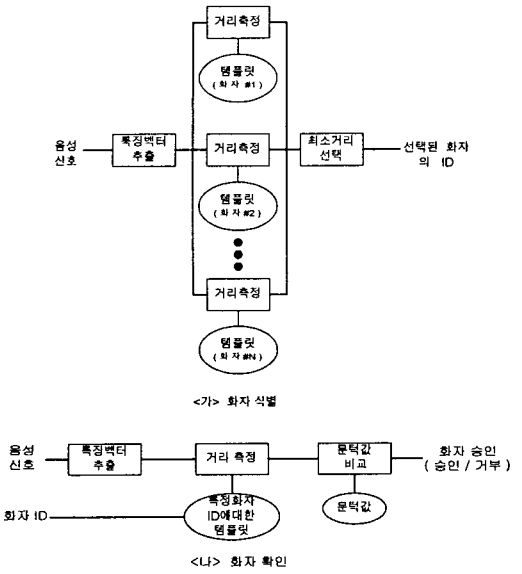

그림 1. 화자 인식에 대한 블럭도

묶어서 총칭한다.

화자 인식 시스템은 문장독립형(Text Independent)과 문장종속형(Text Dependent)의 2가지방법으로 구현이 가능하다. 문장독립형은 템플릿의 문장과 입력 문장의 일치여부에 관계없이 화자를 인식하는 방법이고, 문장종속형은 템플릿의 문장과 입력문장을 일치시킨 경우이다. 화자 식별과 화자 확인에 대한 기본적인 순서도는 그림 1과 같이 나타낼 수 있다.

위의 블럭도는 크게 특징벡터 추출, 거리 측정, 화자 판단의 3가지 단계로 나눌 수가 있다. 특징벡터추출 단계는 선형 예측 계수(LPC Coefficient)를 이용한 캡스트럼 계수(Cepstral Coefficient)가 주로 사용된다. 거리 측정은 각 패턴에 대한 실제적인 인식단계고 확률적인 방법에 기초한 은닉 마코프모델(HMM), 신경망(Neural Network), 벡터양자화(Vector Quantization)등이 많이 사용된다[6][7][8]. 이 때, 거리란 각 템플릿과 입력음성의 특징벡터와의 차이로 유클이디안(Euclidean)을 사용한다. 화자 판단에서는 화자 식별의 경우, 거리가 가장 작은 템플릿을 화자로 인식하고, 화자 확인의 경우에는 일정 문턱치(Threshold)를 기준으로 화자를 승인(Acceptance)하거나 거부(Rejection)한다.

III. 고유벡터와 저차원 모델

입력 신호 $\boldsymbol{x} = [x_0, x_1, \ldots, x_n]^T$에대한 고유벡터와 고유값은 다음과 같은 식으로 나타낼 수 있다.

$$(\boldsymbol{R} - \lambda \boldsymbol{I})\boldsymbol{q} = 0 \quad where \ \boldsymbol{R} = E[\boldsymbol{x}\,\boldsymbol{x}^T]$$

여기에서 $\boldsymbol{R}$은 $\boldsymbol{x}$에대한 자기상관행렬(Autocerrelation Matrix), λ는 고유값(Eigenvalue), $\boldsymbol{q}$는 고유벡터(Eigenvector)이다. 만약, $\boldsymbol{R}$이 $M \times M$ 행렬이라고 할 때, 고유값 λ는 M개가 나온다. 그리고, 각 λ에 따른 고유벡터 $\boldsymbol{q}$가 나온다. 이때, 입력 $\boldsymbol{x}$는 다음 식과 같이 고유벡터 $\boldsymbol{q}$의 선형조합으로 나타낼 수 있다.

$$\boldsymbol{x} = \sum_{i=1}^{M} c_i \boldsymbol{q}_i$$

고유값 λ를 크기에 따라 다음과 같이 쓸 수가 있다.

$\lambda_1 > \lambda_2 > \ldots > \lambda_{p-1} > \lambda_p > \lambda_{p+1} > \ldots > \lambda_{M-1} > \lambda_M$

여기서 고유값이 큰 p개의 영역은 신호영역, 나머지 M- p개는 잡음에 대한 영역을 나타낸다[1]. 따라서, p개의 고유값에 대한 고유벡터로 다음과 같이 신호를 재구성할 수 있다.

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{p} c_i \boldsymbol{q}_i \quad , \quad p < M$$

$\hat{\boldsymbol{x}}$는 신호 $\boldsymbol{x}$의 저차원 모델(Low Rank Model)이라고 한다[5]. 만약, 여러 다른 사람들의 음성 데이터베이스에서 추출된 주된 고유벡터(Eigenvector)들이 있다고 할 때, 이 고유벡터들은 각 사람의 음성에 대한 고유영역(Eigenspace)을 만든다. 본 논문에서는 고유벡터들의 저차원 모델을 이용하여 화자를 인식하는 방법을 제안한다.

IV. 제안한 방법

본 논문에서는 음성의 파워스펙트럼의 자기상관행렬(Autocorrelation Matrix)을 구하고, 이 자기상관행렬의 고유벡터를 이용해서 화자를 인식하는 방법을 제시한다. 제시한 방법에 대한 기본적인 순서도는 그림 2와 같다.

1. 스펙트럼 조정(Spectrum Adjustment)

음성은 크게 자음과 모음으로 나뉘어 진다. 우리가 어떤 사람의 목소리를 인식할 때, 성도의 특성을 반영하는 모음과 그 사람만의 독특한 음색으로 인식을 한다. 따라서, 성도의 영향을 반영하는 모음과 음색을 반영하는 자음의 영향을 고르게 반영하기 위해서 스펙트럼조정(Spectrum Adjustment)을 해주게 된다[2]. 음성의 파워스펙트럼을 Φ라고 할 때, 조정식은 다음과

같이 나타낼 수 있다.

$$\Phi' = \alpha \log_{10}(\beta \Phi + 1.0)$$

여기서 α와 β는 실험적으로 결정된 상수이다.

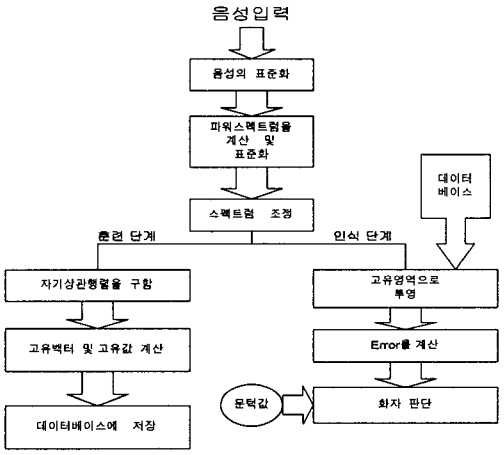

그림 2 . 제안한 방법의 순서도

2. 자기상관행렬(Autocorrelation Matrix)과 고유벡터

한 사람의 N개의 음성에 대한 파워스펙트럼 평균을 Φ_A, 각 음성의 파워스펙트럼을 Φ_n이라고 할 때, 자기상관행렬 R는 다음과 같다[6][7].

$$R = E[(\Phi_n - \Phi_A)(\Phi_n - \Phi_A)^T], \quad 1 \le n \le N$$

이 때, $E[(\Phi_n - \Phi_A)] = 0$를 만족한다.

R에 대한 고유벡터와 고유값을 각각 q와 λ로 나타낼 때, 이들간의 식은 다음과 같이 나타낼 수 있다.

$$(R - \lambda I)q = 0$$

앞의 II장의 내용에 따라, 신호영역에 해당하는 p개의 큰 고유값에 대한 고유벡터를 사용하여 인식에 이용할 저차원 모델을 구성한다. p는 최소최대법칙(Minimax Theorem)을 이용해 구한다[5].

3. 화자 판단

입력음성 x가 들어왔을 때, x의 파워스펙트럼 P와 데이터베이스 안의 평균값 Φ_A간의 차이를 p개의 고유벡터로 각각 투영을 시킨다. 이때, 투영된 값 w_k를 다음과 같이 나타낼 수 있다.

$$\omega_k = (P - \Phi_A)^T \cdot q_k \ , \ k = 1,2,3,..,p$$

따라서, 입력음성과 템플릿간의 거리는 다음과 같이 계산된다.

$$\varepsilon = (P - \Phi_A) - \sum_{i=1}^{p} \omega_k q_k$$

이 때, ε이 다음 조건을 만족하면, 화자로 인식을 한다.

$$\|\varepsilon\| < \zeta \ , \quad \zeta : \textit{threshold value}$$

문턱값 ζ는 실험적으로 결정한다.

V. 모의 실험

템플릿 구성을 위한 데이터베이스로는 각 화자당 10분의 음성 데이터를 사용하였다. 각 데이터는 8kHz로 표본화(Sampling)를 하였다. 각 음성은 화자로 하여금 1-2일 간격으로 1분씩 책을 읽게 한 데이터이다. 일반적으로 화자의 특성을 나타내기 위해서는 30초이상 데이터의 앙상블을 요구한다[3]. 따라서, 본 실험에서는 10분의 데이터를 각 1분씩 10구간으로 나누고 이들의 파워스펙트럼의 자기상관행렬을 사용하였다. 또, 각 구간은 해밍창을 이용해 20ms(160 Sample)의 프레임으로 나누고, 이들 파워스펙트럼의 평균을 그 구간을 대표하는 파워스펙트럼으로 사용하였다[6]. 스펙트럼조정을 위한 α와 β값은 각각 1과 10000을 사용하였다.

그림 3은 화자 A와 화자 B에대한 파워스펙트럼을 나타낸다. 그림 3에서 스펙트럼조정을 한 경우 고주파성분의 변화가 확연하게 나타난다. 따라서, 모음과 자음의 효과를 고르게 반영할 수 있다. 그림 4에서는 화자 A와 화자 B에대한 고유값과 최소최대법칙으로 구한 p값을 보여주고 있다. p는 23을 사용하였다. 실험에서 인식을 위해 사용한 음성데이터는 사람마다 5초의 데이터 30개를 이용하였다.

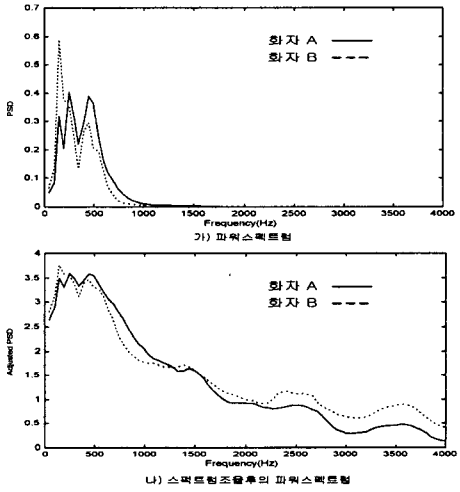

그림 3 . 화자 A와 화자 B에 대한 파워스펙트럼

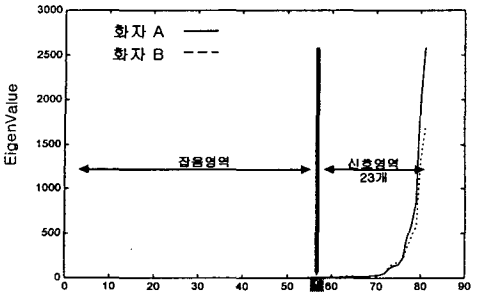

그림 4 . 화자 A와 화자 B에대한 고유값

화자 판단을 위한 문턱값(ζ)은 실험적으로 FA와 FR이 같아질 때의 값을 사용하였다[4]. 이 때, FA와 FR은 화자 확인 시스템에 대한 성능평가를 위해 사용하는 오차율(Error Rate)이다. FA(False Acceptance)는 잘못된 화자를 옳다고 인식한 경우를, FR(False Rejection)는 옳은 화자를 틀리게 인식한 경우를 나타낸다[3][4]. 실험에 대한 결과는 표 1과 그림 5에서 나타내었다. 표 1.에서 보는바와같이 화자 A와 화자 B에 대해서 각각 96.7%와 93.3%의 인식률을 얻었다.

	FA (%)	FR (%)
화자 A	3.33	3.33
화자 B	6.67	6.67
평 균	5.00	5.00

표 1 . 화자 A와 화자 B에대한 오차율

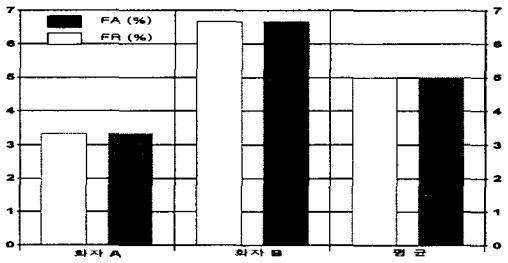

그림 5 . 화자 A와 화자 B의 인식결과 그래프

VI. 결 론

본 논문에서는 고유값해석법(Eigenvalue Analysis)을 이용하여 화자를 인식하는 방법을 제안하였다. 기존의 은닉 마코프모델이나 신경망모델과 비교해 볼 때, 훈련시간(Training Time)과 인식시간(Recognition Time)이 단축된다. 또, 새로운 사용자를 등록하려면, 기존의 방법은 전체 템플릿을 다시 훈련시켜야 한다. 반면에, 제안한 방법에서는 새로운 사용자의 저차원 고유영역만 추가해주면 되므로 보다 효율적이다. 제 V장의 모의 실험결과에서 2명의 화자에 대한 화자 확인에 대하여 평균 95%의 인식률을 얻었다. 앞으로의 과제는 저차원 모델 계수 p를 결정하는 방법과 오차를 계산하는 방법에 있어서 좀 더 많은 연구가 필요하다.

참고문헌

[1] Futoshi ASANO and Satoru HAYAMIZU, "Speech Enhancement Using Array Signal Processing Based on the Coherent-Subspace Method", *IEICE TRANS. FUNDAMENTALS.* VOL. F80 A. NO. 11, pp. 2276-2285, NOVEMBER 1997

[2] Huadong Wu, Mel Siegel and Pradeep Khosla,"Vehicle sound Signature Recognition by Frequency Vector Principal Component Analysis", *IEEE Instrumentation and Measurement Technology Conference. St. Paul, Minnesota, USA*, May 18-20,1998

[3] JOHN D.MARKER and STEVEN B.DAVIS, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time - Spaced Data Base", *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, VOL. ASSP-27, NO. 1, pp. 74-82, FEBRUARY 1979

[4] SADAOKI FURUI, "Recent advances in speaker recognition", *ELSEVIER, PATTERN RECOGNITION LETTERS*, VOL. 18, pp. 859-872, 1997

[5] Simon Haykin, "Adaptive Filter Theory", *Prentice Hall International Editions*, 1996

[6] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of Speech Recognition", *Prentice-Hall International*, Inc. , 1993

[7] John R. Deller. Jr, John G. Proakis, John H. L. Hansen, "Discrete-Time Processing of Speech Signals", *Macmillan Publishing Company*, 1993

[8] Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, "Automatic Speech And Speaker Recognition", *Kluwer Academic Publishers*, 1996