프레임레벨유사도정규화를 적용한 문맥독립화자식별시스템의 구현

Realization a Text Independent Speaker Identification System with Frame Level Likelihood Normalization

김민정, 석수영, 김광수, 정현열
Min-Jung Kim, Su-Young Suk, Kwang-Soo Kim, Hyun-Yeol Chung

영남대학교 정보통신공학과
Department of Information and Communication Eng., Yeungnam University
kmj@orgio.net, lera@orgio.net, kks@speech.yeungnam.ac.kr, hychung@ynucc.yeungnam.ac.kr

요 약

본 논문에서는 Gaussian mixture model을 이용한 실시간 문맥독립화자식별시스템을 구현하여 인식실험을 수행하였으며, 인식시스템의 성능을 향상시키기 위하여 화자검증시스템에서 좋은 결과를 보인 유사도 정규화(Likelihood normalization)방법을 적용하여 인식실험을 하였다. 시스템은 크게 전처리단과 화자모델생성단, 화자식별단으로 나누어진다. 전처리단에서는 화자의 발성변화를 고려하여 CMN(Cepstral mean normalization)과 Silence removal 방법을 적용하였다. 화자모델생성단에서는, 화자발성의 음향학적 특징을 잘 표현할 수 있는 GMM(Gaussian mixture model)을 이용하여 화자모델을 작성하였으며, GMM의 파라미터를 최적화하기 위하여 MLE(Maximum likelihood estimation)방법을 사용하였다. 화자식별단에서는 학습된 데이터와 테스트용 데이터로부터 ML(Maximum likelihood)을 이용하여 유사도를 계산하였으며, 이 과정에서 유사도 정규화를 적용한 경우에는 프레임단위로 유사도를 계산하게 된다. 계산된 유사도는 스코어(S_C)로 표현하였고, 가장 높은 스코어를 가지는 화자가 인식화자로 결정된다. 화자인식에서 발성의 종류로는 문맥독립 문장을 사용하였다. 인식실험을 위해서는 ETRI445 DB와 KLE452 DB를 사용하였으며, 특징파라미터로서는 켑스트럼계수 및 회귀계수값만을 사용하였다. 인식실험에서는 등록화자의 수를 달리하여 일반적인 화자식별방법과 프레임단위유사도정규화방법으로 각각 인식실험을 하였다. 인식실험결과, 프레임단위유사도정규화방법이 인식화자수가 많아지는 경우에 일반적인방법보다 향상된 인식률을 얻을수 있었다.

Abstract

In this paper, we realized a real-time text-independent speaker recognition system using gaussian mixture model, and applied frame level likelihood normalization method which shows its effects in verification system. The system has three parts as front-end, training, recognition. In front-end part, cepstral mean normalization and silence removal method were applied to consider speaker's speaking variations. In training, gaussian mixture model was used for speaker's acoustic feature modeling, and maximum likelihood estimation was used for GMM parameter optimization. In recognition, likelihood score was calculated with speaker models and test data at frame level. As test sentences, we used text-independent sentences. ETRI 445 and KLE 452 database were used for training and test, and cepstrum coefficient and regressive coefficient were used as feature parameters. The experiment results show that the frame-level likelihood method's recognition result is higher than conventional method's, independently the number of registered speakers.

Key words : Speaker identification, Frame level, Likelihood normalization, GMM, CMN, Silence removal

Ⅰ. 서 론

화자식별(Speaker identification)이란 여러 명의 등록화자 중 발성화자를 식별하는 것을 말한다. 이러한 화자식

별 기술은 개인의 음성 특징이 유일하다는 사실을 근거로 하고 있으며 최근의 인터넷 기술의 발전과 더불어 보안을 위한 인증방법으로 각광을 받고 있다. 화자식별시스템은 발성의 종류에 따라 문맥종속 및 문맥독립화자인식으로 나눌 수 있는데, 문맥독립화자인식의 경우 보안성이 높아 이에 관해 많은 연구가 진행중이다. 문맥독립화자인식방법으로서는 장시간(Long-term)통계에 기반한 방법[1], VQ(Vector - quantization)에 기반한 방법[2], HMM과 GMM(Gaussian mixture model)에 기반한 방법[3] 등이 연구되고 있으며 이러한 접근방법들 중 화자특성변화의 표현에 있어서나 화자인식률 면에서 좋은 결과를 나타내고 있는 GMM에 의한 접근방법이 가장 유리한 것으로 알려져 있다[4]. 이에 본 연구에서는 GMM을 이용하여 시스템을 구성하였다.

화자검증시스템에서 유사도정규화는 문장의 변화에 따른 입력발성의 변화를 최소화 하기 때문에 시스템의 성능에 중요한 역할을 하였다. 하지만, 이러한 유사도정규화방법은, 인식화자를 결정하는 스코어가 단일발성에서 계산되어지는 화자식별시스템에서는 의미가 없다고 볼수 있다[5]. 하지만, 화자식별시스템에서 이러한 유사도정규화를 적용하는 대상이 단일벡터나 단일 프레임이라면 유사도정규화방법의 효과를 기대할수 있을것이다. 따라서 본 논문에서는 프레임단위에 유사도정규화방법도 적용하여 시스템을 구성하여 실험을 수행하였다.

2장에서는 전처리단에서 사용된 CMN과 잠음제거에 대해 살펴보고, 3장에서 Gaussian mixture model에 대해서 설명한후, 4장에서 일반적인 화자식별방법과 프레임단위 유사도정규화방법을 적용한것에 대해서 설명한다. 5장에서는 시스템의 구성 및 실험결과에 대해서 기술한 다음 마지막으로 6장에서 결론을 맺도록 한다.

II. Cepstral mean normalization and Silence removal

화자인식을 어렵게 하는 것 중 하나는 화자특성의 변화이다. 이 변화는 화자 자신의 특성, 녹음환경, 전송환경, 잡음 등이 시간에 따라 변하기 때문에 나타나는 현상으로 화자는 매 발성마다 정확히 꼭 같은 발성을 반복할 수 없기 때문이다. 이와 같은 여러 요인으로부터 발생하는 변화를 보상하기 위해서, 각 관측 켑스트럼벡터로부터 평균을 차감해줌으로서 이러한 변화를 보상하는 방법이 많이 이용되고 있다[6].

$X=\{x_1,\ldots,x_N\}$가 켑스트럼벡터열이라면, 장구간 켑스트럼 평균은 다음과 같이 구할 수 있다.

$$\bar{x}=\frac{1}{N}\sum_{t=1}^{N}x_t \qquad (1)$$

각 벡터 x_t로부터 $\bar{x}$의 차감은 다음과 같다.

$$x_t{}^n = x_t - \bar{x} \qquad (2)$$

연속음성이 화자인식에 이용될 경우에는 묵음을 제거(Silence removal)하는 방법을 고려할 수 있다. 연속음성 발성 중 화자가 무엇을 생각할 때, 또는 숨을 고르는 등의 경우에 묵음이 발성에 포함된다. 이러한 묵음은 화자에 대한 어떠한 정보도 제공하지 않기 때문에 모델링할 필요가 없으므로 테스트 데이터나 학습데이터로부터 묵음 구간을 제거하는 것이 유리하다. 묵음 구간을 제거하는 방법에는 여러 가지가 있지만, 본 논문에서는 낮은 에너지를 가지는 프레임을 제거하는 방법을 사용하였다.

III. Gaussian mixture model

GMM(Gaussian mixture model)은 출력확률밀도함수가 가우시안밀도혼합(Gaussian density mixture)인 1개의 상태만으로 구성된 CHMM(Continuous HMM)의 한 형태이다.

화자인식에 GMM을 사용하는 이유로 두 가지를 들 수 있다. 첫째로, GMM은 음향학적클래스(Acoustic class)의 집합을 모델링할 수 있다는 것이다. 화자의 목소리에 대응되는 음향 공간은 모음이나 비음, 파찰음과 같은 음소를 표현하는 음향학적클래스의 집합으로 표현될 수 있는데, 이러한 음향학적클래스는 화자를 구별하는데 이용되는 화자의 성도에 대한 정보를 가지고 있다[7]. i^{th} 음향학적클래스의 스펙트럼 형태는 i^{th} component 밀도의 평균 μ_i으로 표현되고, 평균 스펙트럼형태의 변화는 공분산 행렬 $\sum_i$로 표현된다. 모든 학습 및 테스트의 음성은 레이블되지 않기 때문에, 음향학적클래스는 hidden으로 볼 수 있다. 독립특징벡터를 가정하면, 이러한 hidden 음향학적클래스로부터 추출된 특징벡터의 관측밀도가 Gaussian mixture이다.

두 번째로, Gaussian basis 함수의 선형조합은 샘플분포(Sample distribution)의 클래스를 표현할 수 있다는 것이다[8]. GMM의 성질 중 하나가 임의의 형태를 가지는 밀도를 부드러운 형태로 근사시키는 것이다. unimodal 가우시안 화자모델은 평균벡터(Mean vector)와 공분산(Covariance)으로 화자의 특징분포를 표현하고, VQ-distortion 모델은 특징벡터의 이산집합으로 화자분포를 표현한다. 이와 같은 점을 고려하여 구성된 GMM은 가우시안 함수의 이산집합을 사용하고, 각각의 평균과 공분산을 가지게 함으로써 이들 두 모델의 특징을 혼합한 형태이다[9].

가우시안 혼합밀도는 M component 밀도의 가중합계이며, 다음의 식에 의해 얻어진다[3].

$$p(x|\lambda) = \sum_{i=1}^{M} c_i N(x, \mu_i, \sum_i) \qquad (3)$$

여기서, x는 d-차원랜덤벡터이며, $b_i(x), i=1,\ldots,M$은 component 밀도이고, $c_i, i=1,\ldots,M$은 mixture weight이다.

각 component 밀도는 평균 μ_i과 공분산 $\sum_i$을 가지는 d-variate Gaussian 함수이다.

$$N(x, \mu_i, \sum_i) = \qquad (4)$$
$$\frac{1}{(2\pi)^{\frac{d}{2}} |\sum_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_i)' \sum_i^{-1}(x-\mu_i)\right\}$$

여기에서, mixture weight는

$$\sum_{i=1}^{M} c_i = 1 \qquad (5)$$

로 제한한다.

Gaussian mixture 밀도는 모든 component밀도의 mixture weight와 공분산행렬, 평균벡터로 구성된다.

$$\lambda = \{c_i, \mu_i, \sum_i\} \quad i=1,\ldots,M \quad (6)$$

화자모델학습은 주어진 학습음성으로부터 학습특징벡터의 분포와 가장 잘 맞는 GMM, λ 파라미터를 추정하는 것이다. GMM의 파라미터를 추정하는 방법에는 여러 가지가 있으나, 가장 잘 알려진 방법으로는 MLE(maximum likelihood estimation)이 있다. MLE는 주어진 학습데이터에서 GMM의 유사도를 최대화하는 모델 파라미터를 찾는데 사용된다.

T 학습벡터 $X = x_1, x_2, \ldots, x_T$의 열에서, GMM 유사도는 다음과 같고,

$$P(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \qquad (7)$$

이를 로그영역에서 표현하면 다음과 같다.

$$L(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda) \qquad (8)$$

Ⅳ. 화자식별

1. 화자식별의 일반적인 방법

일반적인 방법에서의 화자식별은 Bayes' rule 따라, N명의 화자 중 사후확률 $P(\lambda_i|X)$, $1 \le i \le N$ 를 최대화하는 모델 λ_{i^*}의 화자 i^* 를 찾는 것이다.

$$P(\lambda_i|X) = \frac{p(X|\lambda_i)P(\lambda_i)}{p(X)} \qquad (9)$$

여기에서, 사전정보가 없기 때문에, 사전확률 $P(\lambda_i)$는 다음과 같이 표현할 수 있다.

$$P(\lambda_i) = \frac{1}{N}, \quad 1 \le i \le N \qquad (10)$$

$\max_i p(X|\lambda_i)$로 사후확률은 최대가 되고, 식별화자는 다음으로 결정된다.

$$i^* = \arg \max_i p(X|\lambda_i) \qquad (11)$$

이러한 일반적인 화자식별 방법을 그림으로 나타내면 그림 1과 같다. 입력된 음성은 전처리단을 거치면서 벡터열 X로 변환되고, 각 화자모델들과의 유사도가 계산되어지고, 계산된 유사도중 가장높은 유사도를 가지는 화자가 인식화자로 결정된다.

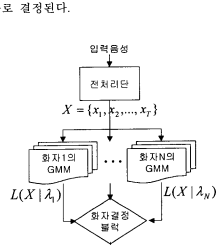

그림 1. 일반적인 화자식별방법
Fig 1. Conventional speaker identification method

2. 프레임단위유사도정규화 (Frame level likelihood normalization)

화자검증시스템에서는 유사도정규화기법을 적용함으로

서 시스템의 성능을 향상시킬 수 있었다[5,10,11]. 화자검증의 일반적인 방법은 요구된 화자 모델 λ_c를 이용하여 입력발성 $X=x_1, x_2, \ldots, x_T$에 Likelihood ratio test[12]를 적용하는것이다.

$$l(X)=\frac{p(\lambda_c|X)}{p(\bar{\lambda}_c|X)} \quad (12)$$

여기에 Bayes' rule를 적용하고, 사전확률이 동일하다고 가정한다면, 로그 영역에서의 Likelihood ratio는 다음으로 표현할수 있다.

$$\Lambda(X)=\log P(X|\lambda_c)-\log P(X|\bar{\lambda}_c) \quad (13)$$

여기에서, $\bar{\lambda}_c$는 모든 다른 가능한 화자를 나타낸다.

유사도 $P(X|\lambda_c)$는 식(8)로부터,

$$\log P(X|\lambda_c)=\frac{1}{T}\sum_{t=1}^{T}\log p(x_t|\lambda_c) \quad (14)$$

로 계산할수 있다.

유사도 $P(X|\bar{\lambda}_c)$는 백그라운드 화자들의 모델을 사용하여 계산되어지며, B개의 백그라운드 화자모델을 $\{\lambda_1, \ldots, \lambda_B\}$라고 하면, 백그라운드 화자들의 로그 유사도는 다음과 같이 계산할수 있다.

$$\log P(X|\bar{\lambda}_c)=\log\left\{\frac{1}{B}\sum_{b=1}^{B}P(X|\lambda_b)\right\} \quad (15)$$

백그라운드 모델에 의한 유사도 정규화는 발성문장의 변화에 따른 변화를 최소화 할수 있기 때문에 화자검증 시스템의 성능을 향상시킬수 있었다. 화자식별시스템에서는 식별화자를 결정하는데 있어서 단일 발성에서 계산되어진 유사도가 사용되기 때문에 정규화 과정이 필요없다[5]. 하지만, 화자식별시스템에서 이러한 정규화 과정을 단일 벡터나 단일 프레임에 적용시킨다면 그 의미가 달라질 것이며, 유사도 정규화는 다음과 같이 적용시킬수 있다.

$$\cdot\ p_{norm}(x_t|\lambda_i)=\frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B}p(x_t|\lambda_b)} \quad (16)$$

모든 벡터 $x_t,\ t=1,2,\ldots,T$에서, 계산되어진 유사도의 전체 합계를 구하면 각 화자 모델 i에 대한 새로운 스코어가 계산되고,

$$Sc_i(X|\lambda_i)=\frac{1}{T}\sum_{t=1}^{T}\log p_{norm}(x_t|\lambda_i) \quad (17)$$

인식화자는 가장 높은 스코어 $Sc_i(X|\lambda_i)$를 가지는 화자로 결정된다.

그림2는 일반적인 화자식별방법에 프레임단위유사도정규화를 적용한 방법을 그림으로 나타낸 것이다. 입력된 음성은 전처리단을 거치면서 벡터열 X로 변환되고 모든 화자모델들과의 유사도 $p(x_t|\lambda_i),\ i=1,2,\ldots,N$가 각각 계산되어진후, 계산된 유사도는 다음단인 유사도 및 스코어계산에서 $t=1,2,\ldots,T$에 대한 각각의 합계가 이루어지며 결과로서 $Sc(X|\lambda_i)$를 출력한다. 인식화자는 가장높은 스코어 $Sc(X|\lambda_i)$를 가지는 화자로 결정된다.

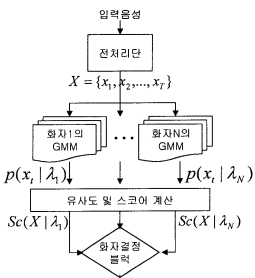

그림 2 프레임단위유사도정규화를 적용한 화자식별방법

Fig 2. Speaker identification method with frame level likelihood normalization

V. 시스템구성 및 인식실험

1. 데이터베이스 및 전처리

실험에는 KLE452데이터베이스와 ETRI445데이터베이

스를 사용하였다. KLE452데이터베이스는 남성38명, 여성 31명으로 구성되어 있으며, 각 화자 당 발성단어는 452개의 PBW단어로 구성되어있다. ETRI445데이터의 경우는 남성화자20명의 발성에 의한 445개의 PBW 단어로 이루어져 있다.

모델작성에는 각 데이터베이스로부터 화자별 200개의 단어를 사용하였으며, 테스트를 위해서는 모델작성에 사용되지 않은 나머지 단어 중에서 무작위로 선택한 10개의 단어를 사용하였다. 전처리단에서 사용된 분석조건은 표1과 같다.

표 1. 전처리단의 분석조건
Table 1. Analysis-condition of preprocessor block

Sampling Rate	16 kHz
Pre-emphasis coefficient	0.98
Hamming Windows	yes
Frame length	256 points
Frame Shift	120 ppoints
Cepstram vector dimmension	10

2. 시스템의 구성

시스템의 전체 구성은 그림3과 그림4에 나타내었다. 시스템은 크게 화자모델생성 및 화자식별의 단계로 구성하였다. 화자모델생성단계는 다시 전처리, 코드북 생성, GMM 생성단계로 구성하였다. 화자식별단계에서는 입력된 음성으로부터 전처리단계를 거쳐 얻어진 벡터열이 각 화자 모델과의 유사도가 계산된후 각 화자별 유사도의 합계를 구해지고 가장 높은 스코어 $Sc_M(X)$를 가지는 화자를 인식화자로 결정하게 된다.

인식실험에서 GMM의 Mixture의 수는 계산량을 고려하여 16으로 고정하였으며, 특징파라미터는 켑스트럼계수와 회귀계수 값만을 사용하였다. 인식실험결과를 표2에 나타내었다.

표 2. 특징파라미터 및 화자식별인식실험결과
Table 2. Feature parameter and result of speaker identification experience

Number of Mixture : 16		
Parameter : CEP + Δ CEP		
화자수	Baseline	프레임단위 유사도정규화
ETRI 445(20명)	95.0 %	95.0 %
KLE 452(31명)	93.5 %	93.5 %
ETRI + KLE(51명)	94.1 %	96.1 %
ETRI+KLE(89명)	94.3 %	95.5 %

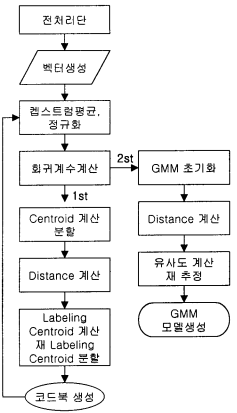

그림 3. 모델 생성단
Fig 3. Model training block

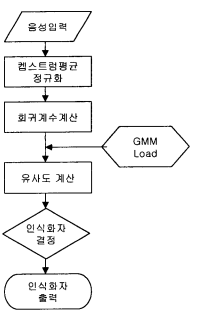

그림 4. 화자 식별단
Fig 4. Speaker identification block

Ⅵ. 결 론

본 논문에서는 실시간문맥독립화자식별시스템을 구현하여 인식실험을 수행하였으며, 이 시스템에 화자검증에서 좋은 성능을 나타내는 유사도정규화 방법을 적용하여 인식실험을 수행하였다. 전처리단에서는 화자발성의 변화를 고려하여 CMN과 잡음제거 방법을 사용하였으며, 화자모델로서는 화자의 특성을 잘 표현할 수 있는 GMM(Gaussian Mixture Model)을 이용하였다. 생성된 화자모델의 학습을 위해서는 MLE (maximum likelihood estimation)법을 적용하였고, 화자인식알고리즘으로는 ML(Maximum likelihood)을 이용하였다. 인식실험에서는 두방법 모두에서 등록화자의 수를 달리하여 실험하였으며, 일반적인 화자식별방법과 프레임단위유사도정규화방법의 인식실험결과를 비교, 검토하였다. 실험결과, 프레임단위유사도정규화방법이 인식화자수가 많아지는 경우에도 인식률 향상이 있었으며, 향후 프레임별 유사도에 따라 가중치를 부여하여 비슷한 유사도를 가지는 화자사이에서 분별력을 높이는 방법을 고려할 계획이다.

접수일자 : 2001. 11. 14 수정완료 : 2002. 1. 28

참고문헌

[1] S. Furui, F. Itakura, and S. Saito, "Talker recognition by longtime averaged speech spectrum," Trans. IECE, Vol. 55-A, No. 1, pp. 549-556, 1972..

[2] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent models," Computer Speech and Language, Vol. 2, pp. 143-157, 1987.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independentspeaker identification using Gaussian mixture speaker models," IEEE Trans. on SAP, Vol. 3, No. 1, pp. 72-83, 1995.

[4] S. Furui, "An overview of speaker recognition technology," in Acoustic speech and speaker recognition(C.-H. Lee, F. K. Soong, and K. K. Paliwal, eds.), Ch. 2, pp. 31-56, Kluwer Acad. Pub., 1996.

[5] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication , Vol. 17, No.1-2,pp.91-108, 1995.

[6] A. Rosenberg, C. LEe, and F.Soong, "Cepstral channel normalization techniques for Hmm-based speaker verification," in Proc. ICSLP, pp. 1835-1838, 1994

[7] H. Matsumoto and H. Wakita, "Vowel normalization by frequency warped spectral matching," Speech Communication, Vol. 5, No. 2, pp. 239-251, 1986.

[8] K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, Inc., second ed., 1990.

[9] H. Gish and M. Schmidt, "Text-independent speaker identification," IEEE Signal Processing Magazine, pp. 18-32, Oct. 1994.

[10] A. Rosenberg, J. DeLong, C.Lee, B.Juang and F. Soong, "The use of cohort normalized scores for speaker verification", proc. ICSLP, pp.599-602, 1992.

[11] T. Matsui and S. Furui, ""Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," Speech Communication, Vol. 17, pp. 109-116, Aug. 1995.

[12] K. Fukunaga, "Introduction to statistical pattern recognition", Academic Press Inc., 1990.

김민정(Min-Joung KIM)
準會員
1999년 영남대학교 일반대학원
멀티미디어 통신공학과
(공학석사)
1999년-현재 영남대학교 일반대학원
정보통신공학과
(박사수료)
관심분야 : 디지털신호처리, 음성처리, 음성인식,
화자 인식

석수영(Soo-Young SUK)
準會員
1998년 계명대학교 물리학과
(이학사)
2000년 영남대학교 일반대학원
멀티미디어 통신공학과
(공학석사)
2000년 3월-현재 영남대학교 일반대학원
정보통신공학과 박사과정
관심분야 : 디지털신호처리, 문자인식, 음성인식

김광수(Kwang-Soo Kim)
正會員
1994년 경남대학교 전자공학과(공학사)
1998년 영남대학교 일반대학원
전자공학과(공학석사)
1998년-현재 영남대학교 일반대학원
전자공학과(박사수료)
2001년 3월-현재 경운대학교 컴퓨터전자정보공학부
전임강사
관심분야 : 음성분석 및 인식, 음성 및 오디오 신호처리,
음질평가

정현열(Hyun-Youl JUNG)
正會員
1975년 영남대학교 전자공학과
(공학사)
1989년 일본 동북대학교 정보공학과
(공학박사)
1989년 3월-현재 영남대학교
전자정보공학부 교수
1992년 7월-1993년 7월 미국 CMU Robotics 연구소
객원연구원
1994년 12월-1995년 2월 일본 토요하시기술과학대학
외국인 연구자
2000년 6월-2000년 8월 미국 Qualcomm Inc.
수석 엔지니어
관심분야 : 음성인식, 화자인식, 음성합성 및
DSP 응용분야