Sub-Stream 기반의 Eigenvoice를 이용한 고속 화자적응

송화전(부산대), 이종석(보이스웨어), 김형순(부산대)

<차 례>

<Abstract>

Fast Speaker Adaptation Using Sub-Stream Based Eigenvoice

Hwa Jeon Song, Jong Seok Lee, Hyung Soon Kim

In this paper, sub-stream based eigenvoice method is proposed to overcome the weak points of conventional eigenvoice and dimensional eigenvoice. In the proposed method, sub-streams are automatically constructed by the statistical clustering analysis that uses the correlation information between dimensions. To obtain the reliable distance matrix from covariance matrix for dividing into optimal sub-streams, MAP adaptation technique is employed to the covariance matrix of training data and the sample covariance of adaptation data. According to our experiments, the proposed method shows 41% error rate reduction when the number of adaptation data is 50.

* Keywords: Fast speaker adaptation, Eigenvoice, Sub-stream based eigenvoice.

1. 서 론

훈련환경과 인식환경 사이의 불일치는 음성인식기의 성능을 하락시키며, 화자간의 차이 및 주위 환경에 의해 생성된 잡음 등이 훈련환경과 인식환경 사이의 불일치의 주원인으로 알려져 있다. 그 중에서 화자간의 차이를 극복하기 위한 기술로 화자적응 방법이 있으며, 화자적응 방법은 훈련과정에서 만들어진 모델에 대해 인식환경에서 훈련에 참여하지 않은 화자를 잘 표현하기 위해 적응 데이터를 이용하여 모델을 이동시킴으로써 인식기의 성능을 향상시키는 효과적인 방법이다. 특히 아주 적은 적응 데이터를 필요로 하는 고속 화자적응은 실제 환경에서 사용하기에 적합한 기술로서 최근 활발한 연구가 진행되고 있다.

Eigenvoice 화자적응 방법은 고속 화자적응에서 maximum a posteriori (MAP) 또는 maximum likelihood linear regression (MLLR) 방법에 비해 좋은 성능을 나타낸다[1]. 그러나, 적응 데이터 수가 많이 증가하더라도 인식성능이 추가적으로 향상되지 않는 단점을 가진다. 이를 해결하기 위해 여러 가지 방법이 제안되었으며, 차원별 eigenvoice 방법도 그 중 하나이다[2]. 그러나, 차원별 eigenvoice 방법에서는 추정할 파라미터 수가 eigenvoice에 비해 상당히 증가하게 되어 적응 데이터가 매우 적은 경우 MLLR과 마찬가지로 성능 하락을 나타낸다. 본 논문에서는 적응 데이터 수에 관계없이 고속 화자적응 성능을 향상시키기 위해 sub-stream 기반 eigenvoice 방법을 제안한다. 제안된 방법에서는 특징 벡터 stream을 차원 간에 상관성(correlation)이 큰 몇 개의 sub-stream으로 나눈 후 sub-stream별로 eigenvoice를 적용함으로써 성능 향상을 도모한다. 차원 간에 자동적으로 그룹화하기 위해 통계적 기법인 군집분석(clustering analysis) 방법을 이용하였고, 또한 적응데이터 수에 따라 군집화 문턱치를 자동 조정되도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 eigenvoice 방법에 대해 서술하고, 3장에서는 본 논문에서 제안한 방법인 sub-stream 기반 eigenvoice 화자적응 방법 및 자동으로 sub-stream을 구성하는 방법에 대해 서술한다. 4장에서는 실험 및 결과에 대해 살펴본 후 5장에서 결론을 맺는다.

2. Eigenvoice 기반 고속 화자적응

2.1 Eigenvoice

우선, T 개의 잘 훈련된 화자 종속(speaker-dependent, SD) 모델 파라미터들을 각각 차원 L의 벡터로 구성한다. 이러한 벡터를 "supervector"라고 하며, HMM 파라미터가 supervector에 저장되는 순서는 상관없지만, T 개의 supervector가 저장되

는 순서는 동일해야 한다. 그 다음, 주성분 분석법(Principal Component Analysis, PCA)과 같은 방법을 적용해서 차원 L을 가지는 T 개의 eigenvector를 얻을 수 있으며 이 eigenvector를 "eigenvoice"라고 한다. 최초 몇 개의 eigenvoice들이 주어진 데이터가 가진 변동의 대부분을 표현하기 때문에 T 개의 eigenvoice 중 최초의 K 개, 즉, e(1), ..., e(K)만으로 전체 변동을 대표할 수 있다($K < T << L$). 이와 같이 선택된 K 개의 eigenvoice는 K-space를 생성한다. 적응데이터가 주어지면 새로운 화자는 다음 식과 같이 K 개의 eigenvoice로 나타낼 수 있다.

$$\hat{\mu} = e(0) + \sum_{k=1}^{K} w(k)e(k) \tag{1}$$

여기서 e(0)는 T명의 SD 모델의 평균을 나타낸다. 그리고, 가중치 $w(k)$는 maximum likelihood eigen-decomposition (MLED) 방법[1]을 통해 구한다.

2.2 차원별(Dimensional) Eigenvoice

Eigenvoice 적응 방법은 아주 적은 적응 데이터 수에서 baseline이나 MLLR 적응방법보다 좋은 성능을 보이지만 반면에 적응 데이터 수가 증가하더라도 성능이 추가적으로 향상되지 않는 단점이 있다. 이를 보완하고자 [2]에서 eigenvoice의 가중치를 음성 특징 벡터의 각 차원별로 추정하는 방법을 제안하였고, 이를 차원별 eigenvoice라고 명명하였다.

3. Sub-Stream 기반 Eigenvoice 고속 화자 적응

3.1 Sub-Stream 기반 Eigenvoice

기존의 eigenvoice의 단점을 보완하기 위해 제안된 차원별 eigenvoice의 경우 적응 데이터가 아주 적은 경우에는 그 성능을 보장할 수가 없었다[2]. 이는 적응 데이터로부터 추정할 파라미터 수가 eigenvoice보다 훨씬 증가하기 때문이다. 본 논문에서는 기존의 eigenvoice 방법과 차원별 eigenvoice 방법의 단점을 보완하여 적응 데이터 수에 관계없이 고속 화자적응에서 높은 성능을 얻기 위해, 차원 간에 상관성이 높은 몇 개의 sub-stream으로 나누어서 eigenvoice를 적응시키는 방법을 도입하였다. 식 (2)는 D 차원의 평균벡터를 임의의 차원을 가진 3개의 sub-stream으로 나눈 한 가지 예이다.

$$\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} = w^{(1)}(1)\begin{bmatrix}e_1(1)\\ e_2(1)\end{bmatrix} + \cdots + w^{(1)}(K)\begin{bmatrix}e_1(K)\\ e_2(K)\end{bmatrix}$$
$$\begin{bmatrix}\mu_3\\ \mu_4\\ \mu_5\\ \mu_6\end{bmatrix} = w^{(2)}(1)\begin{bmatrix}e_3(1)\\ e_4(1)\\ e_5(1)\\ e_6(1)\end{bmatrix} + \cdots + w^{(2)}(K)\begin{bmatrix}e_3(K)\\ e_4(K)\\ e_5(K)\\ e_6(K)\end{bmatrix}$$
$$\begin{bmatrix}\mu_7\\ \mu_8\\ \vdots\\ \mu_D\end{bmatrix} = w^{(3)}(1)\begin{bmatrix}e_7(1)\\ e_8(1)\\ \vdots\\ e_D(1)\end{bmatrix} + \cdots + w^{(3)}(K)\begin{bmatrix}e_7(K)\\ e_8(K)\\ \vdots\\ e_D(K)\end{bmatrix} \quad (2)$$

Eigenvoice에 기반한 제안한 방식들에 대해 수식을 일반화하기 위해 먼저 다음과 같이 사용되는 용어들을 정리한다.

1) N_{SS} : sub-stream의 총수

2) $\{C_1, C_2, \cdots, C_r, \cdots, C_{N_{ss}}\}$: N_{SS} 개의 sub-stream 집합

여기서, $C_r \cap C_s = \phi$, $r \neq s\ (r,s = 1,\cdots,N_{SS})$ 이고, ϕ는 공집합을 나타내며, 각각의 sub-stream은 군집화 방법을 사용하여 $(d \in C_r,\ 1 \le d \le D,\ 1 \le r \le N_{SS})$ 군집화된 차원들의 집합이다.

3) $D_1, D_2, \cdots, D_r, \cdots, D_{N_{ss}}$: 각각의 sub-stream의 차원

4) 특징벡터 차원(D)와 각 sub-stream 차원 사이의 관계

$$D = \sum_{r=1}^{N_{ss}} D_r \quad (3)$$

5) $w^{(r)}(k)$: r번째 sub-stream에 대한 k번째 eigenvoice의 가중치.

상태 s와 mixture m에서의 r번째 sub-stream의 평균 벡터는 다음과 같이 나타낼 수 있다.

$$\left\langle \hat{\boldsymbol{\mu}}_m^{(s)} \right\rangle_{C_r} = \left\langle \mathbf{e}_m^{(s)}(0) \right\rangle_{C_r} + \sum_{k=1}^{K} w^{(r)}(k) \left\langle \mathbf{e}_m^{(s)}(k) \right\rangle_{C_r}, \quad 1 \le r \le NSS \quad (4)$$

식 (4)에서 N_{ss} = 1이면 일반적인 eigenvoice가 되고, N_{ss} = D이면 차원별 eigenvoice가 됨을 알 수 있다.

적응데이터가 적은 경우에는 sub-stream수가 적은 경우가 성능향상에 유리하고 적응데이터가 많은 경우에는 sub-stream 수가 많아지는 것이 유리하다. 따라서, 적

응데이터 수에 따라 sub-stream 수가 자동적으로 정해지는 방법이 필요하다.

본 논문에서는 자동적으로 sub-stream을 나누기 위해 통계적 군집분석(clustering analysis) 방법[3]을 도입하였으며, 그 중 single linkage와 average linkage방법을 사용하여 실험을 수행하였다. <그림 1>에 적응데이터를 사용하여 군집분석을 수행하는 과정을 나타내었다. 먼저 적응데이터에 대해 상관계수 행렬(correlation coefficient matrix)를 구하고 Gower 방법[3]을 통해 거리 행렬(distance matrix)로 변환한다. 구해진 거리 행렬을 기본으로 하여 군집분석을 통해 차원을 군집화한다.

$$d_{ik} = \sqrt{2(1 - s_{ik})} \tag{5}$$

여기서 d_{ik} 와 s_{ik} 는 i 와 k 번째 차원 사이의 거리와 상관계수를 나타낸다.

3.2 군집분석 방법[3]

여러 가지 통계적 군집분석(clustering analysis) 방법[3] 중에 본 논문에서는 agglomerative hierarchical procedure 중 linkage 방법을 사용하였으며, 아래에 N개의 개체를 군집화하는 방법을 간략하게 설명하였다.

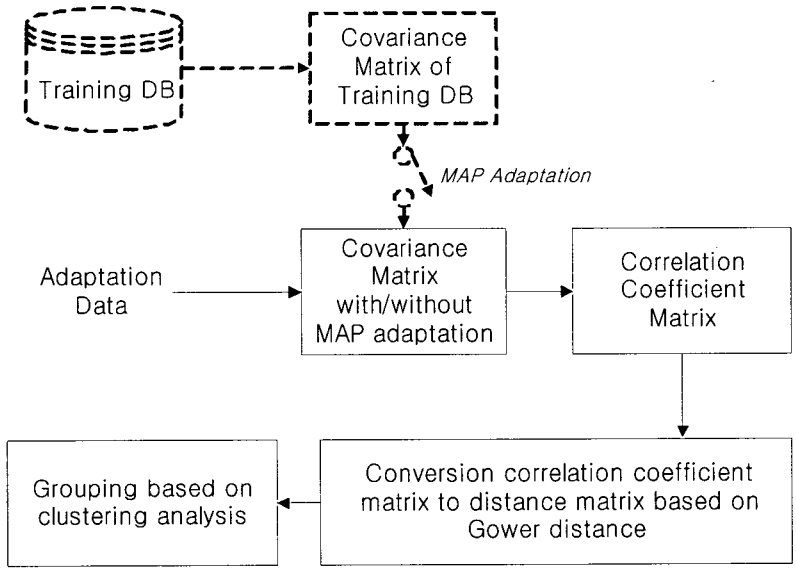
단계 1 : N 개체를 이용하여 식 (5)를 이용하여 N x N 거리 행렬 **D**=$\{d_{ik}\}$를 구성한다.

단계 2 : **D**에서 각 개체 중 가장 적은 거리를 갖는 요소를 찾은 후 가장 가까운 개체인 U와 V 사이의 거리를 d_{UV}라고 한다.

단계 3 : U와 V 개체를 병합하고 개체 UV라고 한다. 또한 **D**에서 U와 V에 해당되는 행과 열을 삭제하고 병합된 UV 개체와 남은 개체사이의 거리를 linkage 방법으로 계산하여 행과 열이 삭제된 **D**에 다시 추가한다.

단계 4 : 단계 2와 3을 N-1 번 반복한다. 위의 병합된 결과들을 기록한다.

단계 3에서 사용되는 linkage 방법으로 single, complete, 그리고 average linkage 방법이 대표적이다.



<그림 1> Sub-stream 자동 병합 방법

4. 실험 및 결과

4.1 실험환경

본 논문에서 제안한 방법의 성능을 비교 평가하기 위해 [2]와 동일한 환경에서 실험을 실시하였다. 사용한 음성 특징 파라미터로는 20ms Hamming window를 10ms씩 이동시키면서 12차 MFCC, delta 및 delta-delta를 구하여 총 36차의 파라미터를 사용하였고, tree-based clustering(TBC)를 사용한 triphone을 기본 모델로 사용하였으며 모델 당 상태 수는 3개로 정하였다. 또한 훈련을 위하여 POW 음성 데이터베이스[5] 중에서 남성 40명분의 음성 데이터베이스를 이용해서 모델을 훈련시켰다.

그리고, 화자적응 및 인식 실험을 위해서는 훈련용 POW 3,848 음성 데이터베이스와는 어휘 내용이 다른 452 균일 음소 분포 단어(Phonetically Balanced Words, PBW) 데이터베이스[6]의 일부를 사용하였다. 남성 화자 10명의 1회 발성분에 대해서 처음 50개 단어 수를 늘려가면서 적응에 사용하였고, 나머지 중 400개 단어를 성능 평가에 사용하였다.

Eigenvoice를 생성시키기 위해 먼저 POW DB를 사용하여 화자 독립(speaker-independent, SI) 모델을 구성한 후 40명의 각각의 화자에 대해 MAP 적응 방식을 사용하여 40개의 SD 모델을 구성하였다. 각각을 supervector로 만든 후

PCA를 통하여 40개의 eigenvoice를 구성하였다. 본 논문에서 사용한 tied state 수는 4050개이며, supervector의 총 차원은 상태 당 mixture 가 1개인 경우 L = 145800 (=4050*1*36) 이다. 그림 2에 eigenvoice 방법과 차원별 eigenvoice 방법에 대한 성능을 나타내었다[2]. 그림에서 나타나 바와 같이 적응 데이터가 적은 경우에는 eigenvoice가 유리하고 적응데이터가 증가하는 경우에는 차원별 eigenvoice가 유리함을 알 수 있다. 또한 차원별 eigenvoice의 경우에는 적응데이터가 아주 적은 경우에 성능이 급격히 떨어짐을 볼 수 있다.

4.2 실험결과

<그림 2>에 본 논문에서 제안한 군집분석 방법을 사용한 sub-stream 기반 eigenvoice의 성능을 나타내었다. 그림에 나타낸 결과는 적응 및 평가에 사용한 10명의 남성화자의 적응데이터 수에 따른 인식성능의 평균이다. 군집화시 사용한 문턱치(TH1)는 다음과 같이 군집화시 사용한 거리의 평균과 표준편차의 값을 이용하였다.

$$TH_1 = \overline{m} + \gamma \cdot \overline{\sigma} \tag{6}$$

여기서

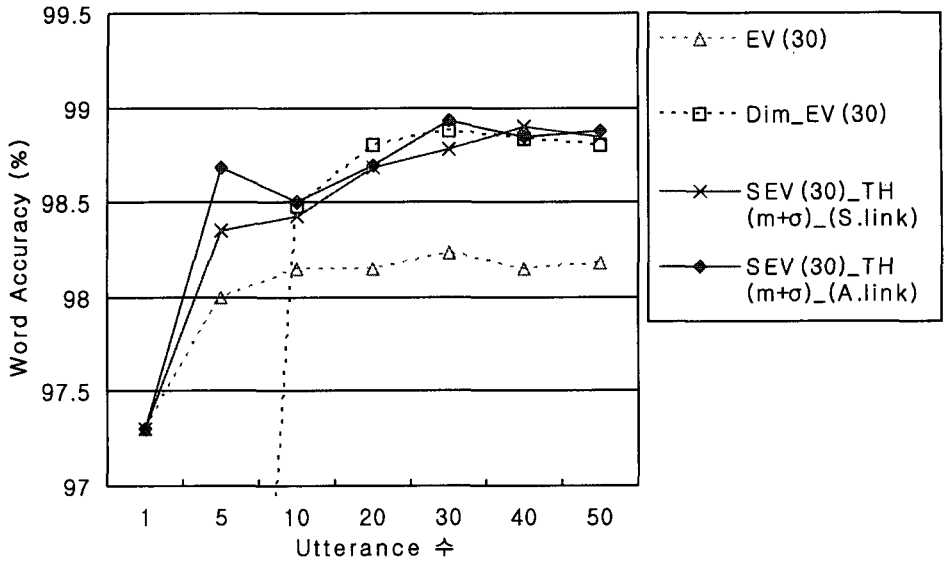
$$\overline{m} = \frac{1}{D-1}\sum_{n=1}^{D-1} d^n_{(UV)W} \tag{7}$$

$$\overline{\sigma}^2 = \frac{1}{D-1}\sum_{n=1}^{D-1} (d^n_{(UV)W} - \overline{m})^2 \tag{8}$$

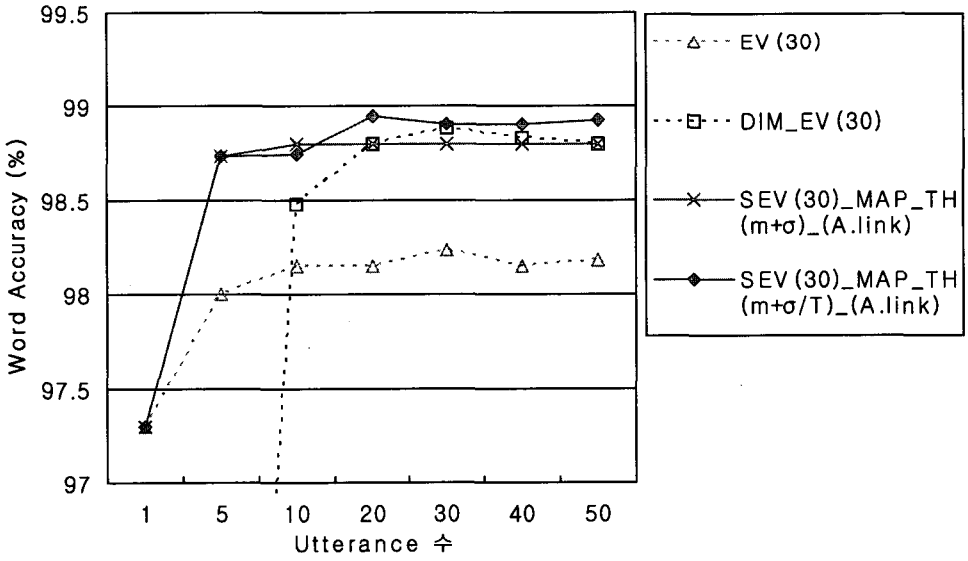
이고, D는 총 차원 수를 뜻하며 $d^n_{(UV)W}$는 cluster (UV) 내의 개체와 cluster W 내의 개체사이의 거리를 나타내며, $d^n_{(UV)W}$에서 n은 n번째 군집화 단계를 뜻한다.

식 (6)에서 γ 값에 따라 표준편차를 크게 더하는 것은 sub-stream을 더 작게 나누고자 하는 의미이다. Single linkage 알고리즘보다 average linkage 방법을 사용하는 것이 적은 적응 데이터 수를 사용하였을 때 좋은 성능을 보여준다.

적응 데이터만을 사용하여 구한 공분산 행렬의 경우에는 신뢰성이 떨어지므로 본 논문에서는 <그림 1>과 같이 상관행렬을 구할 때 훈련 DB의 공분산 행렬과 적응 데이터의 공분산 행렬을 가중합하는 방법을 취하였다. 상관 행렬을 구할 때 POW DB의 공분산 행렬과 적응데이터의 공분산 행렬에 대해 MAP 적응방법을 사용하였고, <그림 3>에 그 성능을 나타내었다. 이때 average linkage 방법을 사용하였고, 또한 다음과 같이 적응데이터 수에 따라 문턱치값(TH2)이 조정되도록 한 실험결과도 함께 나타내었다.



<그림 2> 자동 차원 병합을 이용한 sub-stream 기반 eigenvoice(SEV)의 성능



<그림 3> MAP 적응 방법을 이용한 sub-stream 기반 eigenvoice(SEV)방법의 성능

$$TH_2 = \overline{m} + \frac{\gamma \cdot \overline{\sigma}}{T} \tag{9}$$

여기서 T는 적응 데이터 수를 뜻한다. 적응 데이터가 적은 경우에도 차원별 eigenvoice의 성능보다 상당히 향상됨을 알 수 있으며, 기존의 eigenvoice 화자적응 방법[2]에 비교하였을 때 적응 데이터 수가 50개일 경우 최대 41%의 단어 오인식률 감소를 얻었다.

이상의 sub-stream 기반 eigenvoice 실험 모두 적응 데이터 수가 1개인 경우에서는 성능을 제대로 얻지 못하였다. 이에 따라, 적응 데이터가 1개인 경우에는 식 (4)에서 Nss = 1이 되도록 sub-stream의 수에 제약조건을 부여하였다.

5. 결 론

본 논문에서는 sub-stream 기반의 eigenvoice를 제안하였으며, 이를 통해 eigenvoice 및 eigenvoice 기반으로 기존에 제안한 방법인 차원별 eigenvoice등을 일반화 하였다. 또한 자동적으로 sub-stream을 구성하기 위해 군집분석을 도입하여 고속 화자적응 성능을 향상시켰다. 앞으로 본 논문에서 도입한 군집분석 방법을 기반으로 각 차원별 상관관계뿐만 아니라 이웃하는 차원에 대한 가중치 등을 고려한 좀 더 효과적인 군집화 방법에 대한 연구를 진행시킬 예정이다.

참고문헌

[1] R. Kuhn, P. Nguyen, J. C. Jungua, L. Goldwasser, N. Niedzielski, S. Finche, K. Field and M. Contolini, "Eigenvoices for speaker adaptation", *in Proc. ICSLP*, pp.1771-1774, 1998.

[2] 송화전, 김형순, "차원별 Eigenvoice 적응방법과 적응모드선택에 기반한 고속 화자적응 성능향상", *제 15회 음성통신 및 신호처리 학술대회*, pp.79-82, 2002.

[3] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, 1998.

[4] 유재원, *연속음성인식을 위한 음성 단위 발음사전 구성방법 연구*, 위탁과제 최종연구보고서, *한국전자통신연구소*, 1995.

[5] Yeonja Lim and Youngjik Lee, "Implementation of the POW(Phonetically Optimized Words) algorithm for speech database", *In Proc. ICASSP95*, pp.89-91, 1995.

[6] 이용주, 김봉완 외, "음성 DB 용 PBW에 관한 검토", *제 12회 음성통신 신호처리 워크샵 논문집*, pp.310-314, 1995.

접수일자 : 2005년 8월 22일
게재결정 : 2005년 9월 20일

▶ 송화전(Hwa Jeon Song)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: hwajeon@pusan.ac.kr

▶ 이종석 (Jong Seok Lee)
주소: 135-500 서울시 강남구 대치동 1024번지 나산빌딩 5층
소속: (주)보이스웨어
전화: 02) 3016-8500
E-mail: jslee@voiceware.co.kr

▶ 김형순(Hyung Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr