# Maximum mutual information estimation을 이용한 linear spectral transformation 기반의 adaptation

유봉수, 김동현, 육동석
고려대학교 컴퓨터학과 음성정보처리연구실

# Maximum mutual information estimation linear spectral transform based adaptation

Bongsoo Yoo, Dong Hyun Kim, Dongsuk Yook
Speech Information Processing Lab., Department of Computer Science and Engineering,
Korea University

{skylive, kaizer, yook}@voice.korea.ac.kr

## Abstract

In this paper, we propose a transformation based robust adaptation technique that uses the maximum mutual information(MMI) estimation for the objective function and the linear spectral transformation(LST) for adaptation. LST is an adaptation method that deals with environmental noises in the linear spectral domain, so that a small number of parameters can be used for fast adaptation The proposed technique is called MMI-LST, and evaluated on TIMIT and FFMTIMIT corpora to show that it is advantageous when only a small amount of adaptation speech is used.

## Ⅰ. 서론

일반적으로 HMM을 기반으로 하는 음성인식 시스템에서 모델 파라미터들은 maximum likelihood estimation(MLE)을 이용하여 estimation되어진다. 하지만 MLE가 최적의 성능을 발휘하기 위해서는 실제 음성이 correct한 HMM에 의해서 가정된 통계치를 따라야 하고 무한한 학습 데이터가 주어져야만 한다. 하지만, 실제 estimation에서는 이 두 가정을 충족시킬 수 없기 때문에 MLE는 estimation방법으로써, 많은 한계를 갖는다. 이런 MLE의 대안으로써, 최근 MLE에 비해 좋은 성능을 보이고 있는 discriminative estimation방법들의 적용을 고려해 볼 수 있다[1][2].

MLE가 학습 데이터에 해당하는 모델들의 확률만을 증가시키는 반면, discriminative estimation방법들은 학습 데이터에 해당하는 모델들의 확률을 증가시킬 뿐만 아니라, MLE가 고려하지 않는 인식 에러로 인식될 가능성이 있는 다른 모델들의 확률은 감소 시켜준다. 따라서 기존 MLE보다 복잡하지만, 좋은 성능을 내어 내준다 Discriminative estimation방법들은 maximum a posterior(MAP), maximum classification error (MCE)와 maximum mutual information(MMI) 등이 있다. 이들 모두 MLE에 비해 훨씬 많은 계산량을 요구한다. 하지만, MMI방식은 confuse한 데이터들에 대한 정보를 효율적으로 표현할 수 있는 latttice를 이용하여 많은 계산량을 요구하는 대용량 어휘 인식 시스템상의 적용에도 MLE에 비해 괄목할 만한 성능향상을 보여주었다[2]. 또한, 최근 linear regression방법들에 discriminative 방법을 적용한 MMI linear regerssion (MMILR)[3], conditional maximum likelihood linear regression(CMLLR)[4] 등이 좋은 성

능을 보여 주고 있다.

본 논문에서는 적은 양의 데이터로도 빠른 적응이 가능한 방법인 linear spectral transformation (LST)[5] 알고리즘에 성능 향상을 위해 기존 MLE방법 대신 discriminative 방법 중에 하나인 MMIE을 적용하여 additive 잡음과 convolution 잡음에 대한 효과적인 적응방법을 기술한다.

논문의 순서는 다음 장에서 적응기법으로써 적용하고자 하는 linear spectral transformation에 대해 기술하고, 다음 III장에서는 이를 MMIE 목적함수를 이용하여 잡음 매개변수를 추정할 수 있는 MMI-LST방법에 대해서 기술이 이어진다. IV장에서는 성능 평가를 위한 실험 결과를 기술하고, 마지막 V장에서 결론을 맺는다

## II. Linear spectral transformation

본 논문에서는 [5]에서와 같이 배경잡음들로 인한 additive 잡음과 마이크의 채널차이로 인한 convolutional 잡음을 적응 대상으로 한다. 이것들을 식으로 표현하면 다음과 같다.

$$\hat{X} = N_{\otimes}[X + N_{\oplus}] \tag{1}$$

X는 잡음이 없는 음성이고 $\hat{X}$는 잡음이 포함된 음성을 나타내고, $N_{\otimes}$와 $N_{\oplus}$는 각각 convolutional 잡음과 additive 잡음을 나타낸다. Convolution 잡음은 특정 상수 값으로 가정하고, additive 잡음은 random한 값으로 가정하면, (1)식에 대한 기댓값은 다음식과 같이 나타낼 수 있다

$$E[X] = N_{\otimes}E[X] + N_{\otimes}E[N_{\oplus}] \tag{2}$$

이어서, linear spectral transform의 linear spectral 잡음 평균 벡터 $\tilde{\boldsymbol{\mu}}^s$는 다음과 같은 식으로 표현될 수 있다.

$$\tilde{\boldsymbol{\mu}}^s = \mathbf{A}\cdot\boldsymbol{\mu}^s + \mathbf{b} \tag{3}$$

**A**는 convolutional 잡음 파라미터들의 diagonal 행렬을 나타내고, $\boldsymbol{\mu}^s$는 잡음이 없는 linear spectral 평균벡터, **b**는 additive 잡음 파라미터들의 벡터이다. 마찬가지 비슷 방법으로 linear spectral 잡음 분산 $\tilde{\boldsymbol{\Sigma}}^s$는 다음 식과 같이 표현 할 수 있다.

$$\tilde{\boldsymbol{\Sigma}}^s = \mathbf{A}\cdot\boldsymbol{\Sigma}^s\cdot\mathbf{A}^{\mathrm{T}} + \mathbf{D} \tag{4}$$

**D**는 additive 잡음 분산과 convolutional 잡음 파라미터들의 곱이다. **D**는 additive 잡음이 각 주파수 bin에서 독립적이라고 가정한 diagonal 행렬로 간략화 될 수 있다. 식 (3), (4)는 평균 벡터들과 분산 벡터들의 linear spectral transform을 형성한다. 그러나 음성 waveform은 일반적으로 mel frequency cepstral coefficients(MFCC)으로 변환 되고, 특징 벡터들의 통계치들이 cepstral domain에서 얻어지며, 음성 인식 시스템은 cepstral domain에서 표현된 모델들을 사용한다. 로그 정규 분포를 가정하고[6], 잡음이 없는 linear spectral 평균과 공분산을 다음과 같이 cepstral domain 통계치들로 표현할 수 있다.

$$\mu_n^s = \exp\left[\sum_k c_{nk}^{-1}\mu_k + \frac{1}{2}\sum_k\sum_l c_{nl}^{-1}\Sigma_{lk}c_{nk}^{-1}\right], \tag{5}$$

$$\Sigma_{nm}^s = \mu_n\mu_m\left[\exp\left[\sum_k\sum_l c_{nl}^{-1}\Sigma_{lk}c_{mk}^{-1}\right] - 1\right], \tag{6}$$

$c_{nk}$는 discrete cosine transformation(DCT)행렬의 요소이며, $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$는 cepstrum의 평균 벡터와 공분산 행렬을 나타낸다. 다음으로 잡음이 포함된 cepstral mean, $\tilde{\mu}_n$,은 식 (3)과 (7)을 적용하여 얻을 수 있다

$$\tilde{\mu}_n = \sum_k c_{nk}\left[\log(a_n\mu_n^s + b_n) - \frac{1}{2}\log\left(\frac{a_n^2\Sigma_{nn}^s + d_{nn}}{(a_n\mu_n^s + b_n)^2} + 1\right)\right]. \tag{7}$$

$a_{nn}$, $b_n$ 과 $d_{nn}$ 은 **A**, **b**와 **D**의 $n$번째 구성요소들이다. 비슷한 방법으로 잡음 cepstral 공분산, $\tilde{\Sigma}_{nm}$,은 식 (4)와 (8)을 적용하여 얻을 수 있다.

$$\tilde{\Sigma}_{nm} = \sum_k\sum_l c_{nl}\log\left[\frac{a_{ll}a_{kk}\Sigma_{lk}^s + d_{lk}}{(a_{ll}\mu_l^s + b_l)(a_{kk}\mu_k^s + b_k)} + 1\right]c_{mk}. \tag{8}$$

## III. Maximum mutual information linear spectral transformation

MMIE는 학습 데이터$\{X_1, X_2, \ldots, X_j, \ldots, X_R\}$와 해당하는 모델$\{W_1, W_2, \ldots, W_j, \ldots W_R\}$들의 사후 확률

$\sum_{j=1}^{R} P(W_j \mid X_j)$을 최대화하기 위해서 아래 (9)식과 같은 목적함수를 이용한다. ($\bar{\Theta}$는 language model과 관련된 파라미터 집합을 나타낸다.)

$$\mathrm{MMI} = \max_{\Theta} \log \frac{P(\mathbf{X}_j \mid W_j, \Theta) P(W_j \mid \Theta)}{\sum_{l=1}^{l} P(X_j \mid W_l, \Theta_l) P(W_l \mid \bar{\Theta})} \qquad (9)$$

목적함수의 numerator에서는 correct한 단어 열에 대한 likelihood를 높여주고, denominator에서는 모든 가능한 단어 열들에 대한 likelihood들의 합을 낮춰줌으로써, 목적함수의 전체 확률을 최대화 시킬 수 있는 모델 파라미터 집합 $\Theta$ 를 구할 수 있다.

[4]에서는 (9)의 목적함수를 이산 밀도 approximation [7]을 사용하지 않고 연속 밀도 re-estimation 식을 [8]의 inequality를 이용하여 간략하게 유도하였다. 이에 따른 보조 함수를 나타내면 다음 식과 같다.

$$Q(\Theta \mid \Theta^0)$$

$$= -\frac{1}{2} \sum_j p(W_j, X_j \mid \Theta^0) \{ \sum_g \sum_t \gamma_{j\,g}^{t\,N} - \gamma_{j\,g}^{t\,D} ) [ n \log(2\pi) + \log \mid \Sigma_g \mid + (\mathbf{X}_j^t - \mu_g)' \Sigma_g^{-1} (\mathbf{X}_j^t - \boldsymbol{\mu}_g) ] + \tilde{C} \}, \qquad (10)$$

$$\tilde{C} = \sum_g d'(q_j) \int_{X_j} p(\mathbf{X}_j^t \mid q_j^t = g, \Theta_j^0) \log b_g(\mathbf{X}_j^t \mid \Theta_{j,g}) d\mathbf{X}_j^t \qquad (11)$$

$\gamma_{j\,g}^{t\,N}$는 학습 데이터와 correct한 단어 열이 주어졌을 때, 시간 $t$에서 Gaussian g의 점유 확률을 나타내고, $\gamma_{j\,g}^{t\,D}$는 학습 데이터와 모든 가능한 단어 열이 주어졌을 때, 시간 $t$에서 Gaussian g의 점유 확률을 나타낸다. 또, $\boldsymbol{\mu}_g$ 와 $\Sigma_g$는 각각 Gaussian g의 평균벡터와 공분산 행렬을 나타내고, $d'(q_j)$는 수렴 상수이고, $b(\cdot)$는 output 확률 분포 함수이다. 또, $X_j^t$는 $j$번째 학습 데이터의 시간 $t$에 해당하는 값이고, $q_j$는 $j$번째 학습 데이터의 상태 열, $\Theta^0$는 이전 파라미터 집합, $\Theta_j^0$는 $j$번째 학습 데이터에 해당하는 파라미터 집합, $\Theta_{jg}$는 $j$번째 학습 데이터의 Gaussian g에 해당하는 파라미터 집합을 나타낸다.

본 논문에서는 단지 평균 벡터들만을 적응할 것이므로 평균 벡터들과 관련이 없는 값들은 상수로 치환함으로써, 식 (10)은 다음과 같이 간략화 할 수 있다.

$$Q(\Theta \mid \Theta^0)$$

$$= -\phi - \eta \sum_j \sum_g \left[ \left[ \sum_t \sum_n (\gamma_{j\,g}^{t\,N} - \gamma_{j\,g}^{t\,D}) \frac{(x_{j\,n}^t - \mu_{g\,n})^2}{\sigma_{g\,n}^2} \right] + \left[ d'(q_j) \sum_n \frac{(\sigma_{g\,n}^0)^2 + (\mu_{g\,n}^0 - \mu_{g\,n})^2}{\sigma_{g\,n}^2} \right] \right], \qquad (12)$$

$\phi$와 $\eta$은 상수, $x_{j\,n}^t$는 $j$번째 학습 데이터의 $n$번째 dimension의 시간 $t$에 해당하는 값, $\mu_{g\,n}^0$는 현재의 평균, $\sigma_{g\,n}^0$는 현재의 표준 편차를 나타낸다.

평균 벡터에 대해 간략화된 목적함수인 식 (12)에 LST의 잡음이 포함된 평균 벡터 식 (7)을 적용해 다음과 같은 잡음이 포함된 평균 벡터에 대한 보조함수 식을 유도할 수 있다.

$$Q_{\tilde{\mu}} = \sum_j \sum_g \sum_t \sum_n \left( \sigma_{g,n}^0 \right)^{-2} \cdot \left[ \left( \gamma_{j,g}^{t,N} - \gamma_{j,g}^{t,D} \right) \left( x_{j,n}^t - \sum_m c_{nm} \log \frac{(a_m \mu_{g\,m} + b_m)^2}{\sqrt{(a_m \mu_{g,m} + b_m)^2 + (a_m^2 \sigma_{g\,m}^2 + d_m)}} \right)^2 + d'(q_j) \cdot \left\{ (\sigma_{g\,n}^0)^2 + \left( \mu_{g\,n}^0 - \sum_m c_{nm} \log \frac{(a_m \mu_{g\,m} + b_m)^2}{\sqrt{(a_m \mu_{g,m} + b_m)^2 + (a_m^2 \sigma_{g\,m}^2 + d_m)}} \right)^2 \right\} \right] \qquad (14)$$

$$d'(q_j) = K \sum_t p(q_j^t = g \mid X_j, \Theta_j^0) = K \sum_t \gamma_{j\,g}^{t\,D} \qquad (15)$$

$m$은 fillterbank index를 나타내고, $\mu_{gm}$ 은 Gaussina g의 $m$번째 filterbank에 해당하는 평균이다.

위 목적함수의 $d'(q_j)$는 수렴 속도와 더불어 성능을 좌우하는 중요한 요소로써, 그 중 $K$값의 설정에는 다음 두 가지 방법이 있다.

① 상수 값으로 설정 ($K$=1 or 2)
② 음소 혹은 Gaussian 단위로 각각 다른 값 설정

첫 번째 상수로 설정하는 경우, $K$ 값이 너무 크게 설정 될 경우 안정적인 estimation을 할 수 있지만 수렴이 길어지는 단점이 있고, $K$ 값이 너무 작게 설정되면 반대로 수렴은 빠르지만 적절한 estimation이 이루어지지 않는다. 두 번째 방법이, 각 단위별로 각각 적합한 값을 계산하여 설정하는 방식이기 때문에 첫 번째보다 안정적인 estimation과 더불어 수렴시간을 앞당길 수 있다. 또 두 번째 방법 중 음소 별로 설정하는 것 보다 Gaussian별로 설정하는 것이 더 수렴속도를 빠르게

한다[4]. 따라서 본 논문에서는 Gaussian별로 다른 값을 설정하는 방법을 사용한다.

위 (14)식을 최종 목적함수로 설정하고 이를 최대화하는 각 평균 벡터에 적용 되어 잡음을 표현하게 될 매개변수 a, b, c 값을 추정하여 잡음이 포함된 평균값을 얻는다. 구해진 값을 모델에 갱신함으로써 주어진 잡음에 대한 적응모델을 구할 수 있다. 여기서 매개변수 a, b, c 값에 대한 추정은 modified Powell's 알고리즘[9]을 이용하여 이루어진다.

## V. 실험

MMI-LST방법에 대한 실험은 Viterbi decoding algorithm에 기반한 자동 음성인식 시스템 상에서 이루어졌다. 기본 모델은 화자독립, tied state, crossword, state당 10개의 Gaussian 분포를 갖는 triphone HMM이다. TIMIT 데이터로부터 3,696개의 문장들이 학습에 사용되었다. 그리고, 평균 10dB의 SNR과 35dB의 additive 잡음을 갖고 있는 FFMTIMIT 데이터로부터 1,296개의 문장들이 테스트를 위해 사용되었다. 각 화자들에 대해, 적응 데이터의 문장을 이루는 11개의 단어 각각을 적응을 하는데 사용하고 동일 화자의 다른 문장들에 대해 테스트를 하였다.

표 1: 약 0.25초 길이의 한 단어에 대한 MMI-LST adaptation 평균 음소 에러율(%).

| Estimation 방법 | 에러율(%) |
|---|---|
| TIMIT(ML training) | 26.4 |
| FFMTIMIT | 47.0 |
| MLLR | 51.2 |
| ML-LST | 42.9 |
| MMI-LST | 42.2 |

(TIMIT: TIMIT으로 학습 후 TIMIT으로 테스트, FFMTIMIT: TIMIT으로 학습 후 FFMTIMIT(SNR 10dB)으로 테스트)

표 1은 다양한 적응 기법들의 성능을 보여준다. MLLR은 39차의 변환 파라미터를 사용한 반면, ML-LST와 MMI-LST는 단지 24차 잡음 파라미터만을 사용하였다. MLLR 적응 기법은 한 단어만을 이용한 적응이 이루어 지지 않고 오히려 성능이 낮아진 반면, ML-LST와 MMI-LST는 적응을 위해 한 단어만을 이용하더라도 적응 전보다 5%정도의 에러율을 감소시킴으로써, 빠른 적응에 효과적인 적응 기법임을 알 수 있다. 또, 간소한 차이이긴 하지만 MMI-LST가 ML-LST보다 나은 성능을 보여주고 있다.

## VI. 결론

본 논문은 maximum mutual information linear spectral transformation 방법을 이용하여 잡음환경에서 강인한 적응 알고리즘을 제안했다 MMI-LST는 적은 transformation 파라미터들을 사용하고, discriminative estimation 방법인 MMI를 estimation 방법으로 사용하여, 기존 적응 기법들에 비해 적은 양의 적응 데이터의 사용으로도 좋은 성능을 발휘한다.

## 참고문헌


[1] Valtchev, V., *Discriminative methods in HMM based speech recognition*, Ph.D. thesis, Cambrige university, 1995

[2] Woodland, P. C. and Povey, D.,"Large scale discriminative training of hidden Markov models for speech recognition", *Computer Speech and Language*, vol. 16, pp 25-47, 2002

[3] Uebel, L.F. and Woodland, P. C.,"Improvements in linear transform based speaker adaptation", in *Proc.* of *ICASSP*, pp 49-52, 2001

[4] Ganawaradana, A. and Byrne, W.,"Discriminative speaker adaptation with conditional maximum likelihood linear regression", in *Proc.* of *EuroSpeech*, pp 1203-1206, 2001

[5] Kim, D. and Yook, D.,"Fast channel adaptation for continuous density HMMs using maximum likelihood spectral transform", *IEE Electronics Letters*, vol. 40, no. 10, pp. 632-633, May 13, 2004.

[6] Gales, M.J.F., *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, Cambridge University, 1995

[7] Normandin, Y., Cardin, R. and Mori, R. D.,"High-Performance connected digit recognition using maximum mutual information estimation", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 299-311, April 1994.

[8] Gopalakrishnan, P. S., Kanevsky, D., Nadas, A., and Nahamoo, D.,"A generalization of the Baum algorithm to rational objective functions", *in Proc. of ICASSP*, pp. 631-634, 1989.

[9] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B., *Numerical recipes in C++*, pp. 398-460, Cambridge University Press, 2002.