Perceptual Factor Analysis for Speech Enhancement

Chuan-Wei Ting and Jen-Tzung Chien

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
{motor,chien}@chien.csie.ncku.edu.tw

Abstract

This paper presents a new speech enhancement approach originated from factor analysis (FA) framework. FA is a data analysis model where the relevant common factors can be extracted from observations. A factor loading matrix is found and a resulting model error is introduced for each observation. Interestingly, FA is a subspace approach properly representing the noisy speech. This approach partitions the space of noisy speech into a principal subspace containing clean speech and a complimentary (minor) subspace containing the residual speech and noise. We show that FA is a generalized data model compared to signal subspace approach. To perform FA speech enhancement, we present a perceptual optimization procedure that minimizes the signal distortion subject to the energies of residual speech and noise under a specified level. Importantly, we present a hypothesis testing approach to optimally perform subspace decomposition. In the experiments, we implement perceptual FA speech enhancement using Aurora2 corpus. We find that proposed approach achieves desirable speech recognition rates especially when signal-to-noise ratio is lower than 5 dB.

1. Introduction

Automatic speech recognition (ASR) systems have been employed to many real-world applications. However, ASR systems are always degraded in presence of different noises in practical situations. To provide good speech quality for ASR systems, the speech enhancement is an important preprocessing procedure for noisy speech recognition. In the past decade, the researchers on speech enhancement for robust ASR have been attracting many people working on this issue. Spectral subtraction algorithm [2] is one of the most popular methods for speech enhancement. This algorithm has the drawbacks of producing speech distortion and "musical noise". The method in [11] was proposed to overcome "musical noise" problem by using human auditory models where the perceptual effect of "musical noise" was reduced under predefined threshold. Below the masking threshold, the residual noise becomes inaudible by human ear. Other researchers presented subspace approaches to balance the trade off between speech distortion and residual noise [5] [8].

The general concept of subspace approaches is originated from that the noisy speech signal can be projected onto two subspaces; one is the signal subspace in which clean speech signal and few noises are included, and the other is the noise subspace that only contains noise information. In [4], Ephraim and Van Trees proposed signal subspace approach to find optimal estimator or filter by

minimizing the speech distortion subject to the constraint of residual noise kept under a threshold. This work decomposed the noisy signal into signal subspace and noise subspace by using Karhunen-Loève transform (KLT). The noisy speech signal was accordingly enhanced by using inverse KLT. Rezayee and Gazor [8] used a diagonal matrix instead of the identity matrix for finding the linear time domain constrained estimator of clean speech. Hu and Loizou [5] estimated the optimal filter by using common matrix diagonalizing the covariance matrices of the clean and noise signals.

In this paper, we are presenting a FA speech enhancement using the perceptual optimization procedure. In general, FA is a data analysis model, which is popular in societies of social science and machine learning. FA is highly related to principal component analysis (PCA) developed for feature dimension reduction. One major difference is that PCA represents the covariance or correlation matrix using singular value decomposition (SVD), whereas FA incorporates a prior structure of the residual terms. Also, the common factors extracted by FA model are useful to represent the correlation between different features [1]. The full covariance matrix can be properly modeled. Although FA generative model is new in the society of speech technology, some researchers have successfully combined FA model and hidden Markov model (HMM) for building ASR system [9]. In this paper, we present a new perceptual FA model and solution to speech enhancement. The noisy speech signal is decomposed into principal factors and minor factors, or correspondingly projected onto two subspaces. The first subspace represents the clean speech and the other subspace is a residual subspace containing noise and residual speech. The decomposition can be fulfilled via eigen-analysis for covariance matrix of speech signal. However, in conventional signal subspace approach, the smaller eigenvalues were assumed to be zero for speech enhancement. When considering FA modeling of noisy speech, the residual covariance matrix is assumed to be a diagonal matrix, which is practical for speech enhancement in presence of colored noise [8]. Furthermore, we exploit the hypothesis testing for finding the optimal FA subspace decomposition. Correspondingly, the noisy speech signal can be enhanced. Experiments on Aurora2 corpus show that the proposed FA speech enhancement approach attains good recognition performance for different cases of signal-to-noise ratio (SNR).

2. Subspace Approaches

2.1. Signal Subspace (SS)

Signal subspace is a popular speech enhancement approach using a linear model assuming that *K*-dimensional noisy observation vector $\mathbf{z}$ is corrupted in a form of

$$\mathbf{z} = W_{\mathrm{SS}} \cdot \mathbf{x}_{\mathrm{SS}} + \mathbf{n}_{\mathrm{SS}} = \mathbf{y} + \mathbf{n}_{\mathrm{SS}}, \tag{1}$$

where W_{SS} is a $K \times M$ matrix of rank M ($M < K$) with column vectors consisting of bases of a subspace of Euclidean space R^K. This is a subspace of clean speech $\mathbf{y}$. $\mathbf{x}_{\mathrm{SS}}$ denotes the

coordinate vector and $\mathbf{n}_{\mathrm{SS}}$ denotes the noise signal. This model is established assuming that noise signal is additive and uncorrelated with clean speech. The covariance matrix of $\mathbf{y}$ with rank M is given by

$$R_y = E\{\mathbf{y}\mathbf{y}^T\} = W_{\mathrm{SS}} R_{x_{\mathrm{SS}}} W_{\mathrm{SS}}^T = W_y \Lambda_y W_y^T . \tag{2}$$

Using eigen-decomposition, we obtain eigenvector matrix $W_y = [W_y^M \; W_y^{K-M}]$ and diagonal eigenvalue matrix Λ_y containing $K - M$ zero eigenvalues. The first M eigenvectors W_y^M span the same subspace as the clean speech subspace, i.e. $\mathrm{span}(W_{\mathrm{ss}}) = \mathrm{span}(W_y^M)$. To find the linear filtering for speech enhancement, it is popular to optimize the perceptually meaningful criterion, which is equivalent to minimize signal distortion while the residual noise energy is constrained under a predefined level. After solving a constrained optimization problem, we obtain the optimal solution to SS approach [4]

$$\hat{H}_{\mathrm{SS}} = R_y (R_y + \mu R_{n_{\mathrm{SS}}})^{-1} = W_y \Lambda_y (\Lambda_y + \mu W_y^T R_{n_{\mathrm{SS}}} W_y)^{-1} W_y , \tag{3}$$

where μ is the Lagrange parameter. In (3), we express the linear estimator $\hat{H}_{\mathrm{SS}}$ using eigen-decomposition of R_y.

2.2. Factor Analysis (FA)

On the other hand, FA is a general modeling approach to express an observed data vector [1]

$$\mathbf{z} = W_{\mathrm{FA}} \mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}} . \tag{4}$$

Here, the noisy speech signal $\mathbf{z}$ is considered with a preprocessing stage of mean removal. The basic idea of FA is to use a factor loading matrix W_{FA} and a common factor vector $\mathbf{x}_{\mathrm{FA}}$ to represent the observed data $\mathbf{z}$. Common factors are referred as the latent variables. The error term $\mathbf{n}_{\mathrm{FA}}$ is a specific factor representing the noise signal and/or residual speech signal. Different from principal component analysis (PCA) developed for dimension reduction, FA aims to extract the common factors for data modeling. Some properties have been specified to establish FA model. First, the observation, common factor and error term are assumed to be Gaussian distributed with zero mean $E[\mathbf{z}] = E[\mathbf{x}_{\mathrm{FA}}] = E[\mathbf{n}_{\mathrm{FA}}] = 0$. Also, common factor and error term are uncorrelated and their covariance matrices are diagonal, namely $E[\mathbf{x}_{\mathrm{FA}} \mathbf{n}_{\mathrm{FA}}^T] = 0$, $E[\mathbf{x}_{\mathrm{FA}} \mathbf{x}_{\mathrm{FA}}^T] = \mathbf{I}_M$ and

$E[\mathbf{n}_{\mathrm{FA}}\mathbf{n}_{\mathrm{FA}}^{T}]=\Psi$. For the case of isotropic noise, we have FA parameter $\Psi=\sigma^{2}\mathbf{I}_{K}$, where $\mathbf{I}_{K}$ is an $K\times K$ identity matrix. Typically, FA model in (4) is similar to the linear regression model. However, the estimation of FA and linear regression models is quite different. In linear regression model, only $\mathbf{x}_{\mathrm{LR}}$ is unknown (W_{LR} is known), whereas in FA model neither W_{FA} nor $\mathbf{x}_{\mathrm{FA}}$ are known. We should estimate FA parameters W_{FA}, $\mathbf{n}_{\mathrm{FA}}$ and later find $\mathbf{x}_{\mathrm{FA}}$. There are several approaches useful to estimate W_{FA}. One approach was derived from probabilistic PCA model [3] [10] using the maximum likelihood estimate. Nevertheless, W_{FA} can be estimated via eigen-decomposition of covariance matrix of $\mathbf{z}$

$$R_{z}=E[\mathbf{z}\mathbf{z}^{T}]=W_{\mathrm{FA}}W_{\mathrm{FA}}^{T}+\Psi=W_{z}\Lambda_{z}W_{z}^{T}= W_{z}^{M}\Lambda_{z}^{M\,1/2}\Lambda_{z}^{M\,1/2}W_{z}^{M\,T}+W_{z}^{K-M}\Lambda_{z}^{K-M}W_{z}^{K-M\,T}\quad. \tag{5}$$

where W_{z} and Λ_{z} are eigenvector and eigenvalue matrices, respectively. Through eigenvalue ordering, we obtain partitioned eigenvector matrix $W_{z}=[W_{z}^{M}\ W_{z}^{K-M}]$ and eigenvalue matrix $\Lambda_{z}=\mathrm{diag}[\Lambda_{z}^{M}\ \Lambda_{z}^{K-M}]$. Factor loading matrix W_{FA} is found using principal submatrix W_{z}^{M} and the preceding M eigenvalues in Λ_{z}. Or, we have $\mathrm{span}(W_{\mathrm{FA}})=\mathrm{span}(W_{z}^{M})$. The covariance matrix of error or noise term Ψ is generated using minor submatrix W_{z}^{K-M} and the last $K-M$ eigenvalues. Interestingly, FA parameters are estimated from two subspaces of $\mathbf{z}\in R^{K}$. FA can serve as SS approach. In what follows, we will explore the link between SS and FA for data modeling and find the solution to FA speech enhancement.

3. FA Speech Enhancement

3.1. Relation between SS and FA

Actually, the underlying concept of FA is similar to SS. Both methods decompose the signal space into two subspaces. Using FA model, the principal subspace $\mathrm{span}(W_{\mathrm{FA}})$ or $\mathrm{span}(W_{z}^{M})$ is used to represent all observed clean and noisy data. The minor subspace $\mathrm{span}(W_{z}^{K-M})$ contains the information of residual speech and noise. However, in SS approach, the signal subspace and noise subspace represent clean speech and noise signal, respectively. The linear models of SS in (1) and FA in (4) look similar. Typically, FA model is desirable for modeling full covariance or correlation matrix of observed data. After eigen-decomposition, the first M common factors have high energy. They are used for representing clean speech signal. The correlation between corresponding feature components is significant. But, the last $K-M$ common factors contain residual speech and noise signal with small energy. In SS model, the last $K-M$ eigenvectors span the noise subspace. This is the key difference between FA and SS models. To explain this property, let us use the same factor loading matrix W_{FA} and common factor $\mathbf{x}_{\mathrm{FA}}$ to express the corresponding clean speech

$$\mathbf{y}=W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}}+\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}. \tag{6}$$

The term $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$ means the error due to residual speech. Then, the observed noisy speech has the form

$$\mathbf{z} = W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}} + \mathbf{n}_{\mathrm{FA}}^{\mathrm{n}}. \tag{7}$$

Here, the residual speech $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$ and noise signal $\mathbf{n}_{\mathrm{FA}}^{\mathrm{n}}$ are summed up to denote the error term of noisy speech, i.e. $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}} + \mathbf{n}_{\mathrm{FA}}^{\mathrm{n}} = \mathbf{n}_{\mathrm{FA}}$. Accordingly, the covariance matrix of noisy speech turns out to be

$$\begin{aligned} R_z &= E[(W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}})(W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}})^T] \\ &= W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}} + R_{\mathrm{n}} \end{aligned}. \tag{8}$$

Two covariance matrices R_{rs} and R_{n} corresponding to error variables $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$ and $\mathbf{n}_{\mathrm{FA}}^{\mathrm{n}}$ are produced, respectively. Basically, FA is a generalized data modeling approach compared to SS.

3.2. Perceptual Criterion for Speech Enhancement

We have explained how FA is used to model noisy speech data. Under this data modeling framework, we would like to develop speech enhancement approach. Similar to SS speech enhancement, we should adopt an objective function to be optimized to estimate the clean speech signal $\hat{\mathbf{y}}$. A $K \times K$ matrix H_{FA} serves as a linear estimator or filter for speech enhancement $\hat{\mathbf{y}} = H_{\mathrm{FA}}\mathbf{z}$. The residual speech signal ε due to this estimation becomes

$$\varepsilon = \hat{\mathbf{y}} - \mathbf{y} = (H_{\mathrm{FA}} - \mathbf{I}_K)\mathbf{y} + H_{\mathrm{FA}}\mathbf{n}_{\mathrm{FA}}^{\mathrm{n}} = \varepsilon_y + \varepsilon_n, \tag{9}$$

where ε_y is the speech distortion and ε_n is the residual noise. The energies of signal distortion and residual noise are obtained by

$$\bar{\varepsilon}_y^2 = \mathrm{tr}E[\varepsilon_y^T \varepsilon_y] = \mathrm{tr}[(H_{\mathrm{FA}} - \mathbf{I}_K)R_y(H_{\mathrm{FA}} - \mathbf{I}_K)^T], \tag{10}$$

$$\bar{\varepsilon}_n^2 = \mathrm{tr}E[\varepsilon_n^T \varepsilon_n] = \mathrm{tr}[H_{\mathrm{FA}} R_{\mathrm{n}} H_{\mathrm{FA}}^T]. \tag{11}$$

Also, from (6), we calculate the covariance matrix of clean speech $\mathbf{y}$ as

$$R_y = W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}}. \tag{12}$$

Notably, there is an additional term in FA model due to the residual speech. When finding FA speech enhancement solution, such generalized model should be better for estimation of clean speech. In this

study, we also take into account the auditory effects [6][7] while estimating the optimal filter. In human's auditory perception system, frequency masking is a phenomenon under which one sound can't be perceived if another sound close in frequency has a high enough level. Based on the masking effects, the residual noise is constrained to be smaller than a masking threshold rather than subtracting all noise in the noisy speech. Additionally, human is more sensitive to the distorted sound. There is a tradeoff between signal distortion and residual noise. Less residual noise will causes larger signal distortion, and the optimal filter will become an identity matrix if we enhance speech signal without distortion. According to these two properties, we adopt perceptual criterion for FA speech enhancement. Namely, we minimize the energy of speech distortion by considering the masking effect that the energy of residual noise should be controlled under a specific threshold. The objective function and constraint are given by

$$\min_{H_{\mathrm{FA}}} \bar{\varepsilon}_y^2$$

$$\text{subject to: } \bar{\varepsilon}_n^2 \le \gamma \sigma_n^2, \tag{13}$$

where $\gamma\sigma_n^2$ denotes the permissible residual noise level, σ_n^2 is a predefined noise energy, γ is an adjustable parameter which controls the masking level and that is restricted between the range of 0 and 1. The optimum linear estimator can be solved through Lagrange optimization procedure. By introducing the Lagrange multiplier μ, we can find the solution to FA speech enhancement

$$\hat{H}_{\mathrm{FA}} = (W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}})(W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}} + \mu R_{\mathrm{n}})^{-1}. \tag{14}$$

The key difference between SS solution in (3) and FA solution in (14) lies in the models of clean speech $\mathbf{y}$ in (1) and (6). Either $K \times M$ matrix W_{SS} or W_{FA} is not sufficient to represent clean speech signal. With an additional residual speech term $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$, we are able to achieve precise data model contributed by the last $K - M$ eigenvectors. If we neglect the residual speech in FA, clean speech becomes $\mathbf{y} = W_{\mathrm{FA}} \mathbf{x}_{\mathrm{FA}}$. The covariance matrix R_{n} disappears in the solution. The FA solution is reduced to SS solution.

3.3. Optimal Subspace Decomposition

Using either FA or SS, it is critical to determine the partition of principal factors (or signal subspace) and minor factors (or noise subspace). This partition is controlled by the parameter of noise threshold σ_n^2. To significantly perform subspace decomposition, in this study, we employ hypothesis test principle to estimate optimal σ_n^2 instead of empirically assigning a value using SS approach.

Accordingly, we are not only able to determine the dimension of principal factors but also the parameters σ_n^2 without using the additional empirical parameter γ. We are testing the null hypothesis [1] that the last $K-M$ eigenvalues are equal $H_0 : \lambda_{M+1} = \lambda_{M+2} = \cdots = \lambda_K$ against the alternative hypothesis H_1 that at least two of the last $K-M$ eigenvalues are different. Assuming that eigenvalues are Gaussian distributed, we can represent the likelihood under null hypothesis as

$$L(H_0) = (2\pi)^{-\frac{N(K-M)}{2}} \left|\Lambda_2\right|^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \Delta\mathbf{x}_i \Lambda_2^{-1} \Delta\mathbf{x}_i^T\right\}. \tag{15}$$

where $\Delta\mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ denotes the *i*th row of $\Delta\mathbf{X}$ with $\sum_{i=1}^{N} \Delta\mathbf{x}_i \Delta\mathbf{x}_i^T = \mathrm{tr}\left[\Delta\mathbf{X}^T \Delta\mathbf{X}\right]$. Λ_2 is a diagonal matrix with its diagonal elements equal to the last $K-M$ eigenvalues and *N* is the number of training observations. $L(H_0)$ can be arranged as

$$\begin{aligned} L(H_0) &= (2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} (\Delta\mathbf{x}_i \Delta\mathbf{x}_i^T) \Lambda_2^{-1}\right\} \\ &= (2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{N}{2}\mathrm{tr}\left[\left(\frac{1}{N}\Delta\mathbf{X}^T \Delta\mathbf{X}\right)\Lambda_2^{-1}\right]\right\}. \\ &= (2\pi)^{-\frac{N(K-M)}{2}} \left[\left(\frac{1}{K-M}\sum_{k=M+1}^{K} \lambda_k\right)\right]^{-\frac{N}{2}} \cdot \exp\left\{-\frac{N}{2}\mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1}\right)\right\} \end{aligned} \tag{16}$$

Similarly, the likelihood under alternative hypothesis is yielded by

$$\begin{aligned} L(H_1) &= (2\pi)^{-\frac{N(K-M)}{2}} \left|\Lambda_2\right|^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{N} (\Delta\mathbf{x}_i \Delta\mathbf{x}_i^T) \Lambda_2^{-1}\right]\right\} \\ &= (2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{N}{2}\mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1}\right)\right\} \end{aligned}. \tag{17}$$

We evaluate the likelihood ratio q of $L(H_0)$ to $L(H_1)$. The resulting test statistic q has the form of

$$
\begin{aligned}
q &= \frac{L(H_0)}{L(H_1)} \\
&= \frac{(2\pi)^{-\frac{N(K-M)}{2}} \left[\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M} \right]^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{N}{2} \mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1} \right) \right\}}{(2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k \right)^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{N}{2} \mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1} \right) \right\}} . \qquad (18) \\
&= \left[\frac{\prod_{k=M+1}^{K} \lambda_k}{\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M}} \right]^{\frac{N}{2}}
\end{aligned}
$$

Then, the distribution of statistic $-2\log q$ turns out to be a χ^2 distribution

$$
\begin{aligned}
-2\log q &= -2\log \left[\frac{\prod_{k=M+1}^{K} \lambda_k}{\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M}} \right]^{\frac{N}{2}} \\
&= -N \cdot \log \left[\frac{\prod_{k=M+1}^{K} \lambda_k}{\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M}} \right] \\
&= -N \cdot \left[\log \prod_{k=M+1}^{K} \lambda_k - \log \left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M} \right] \qquad . \qquad (19) \\
&= N \cdot \left[(K-M) \log \left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right) - \sum_{k=M+1}^{K} \log \lambda_k \right] \\
&= N \cdot \left[(K-M) \log \bar{\lambda} - \sum_{k=M+1}^{K} \log \lambda_k \right] \sim \chi^2_{(\nu)}
\end{aligned}
$$

Finally, we find that null hypothesis H_0 is rejected at a significance level α if

$$
N \cdot \left[(K-M) \log \bar{\lambda} - \sum_{k=M+1}^{K} \log \lambda_k \right] \geq \chi^2_{\nu;\alpha} . \qquad (20)
$$

In (20), $\bar{\lambda}$ is a sample mean of eigenvalues, and ν is the degree of freedom of χ^2 distribution.

4. Experiments

4.1. Speech Database and Experimental Setup

We performed speech recognition and SNR calculation using Aurora2 database for evaluating performance of proposed speech enhancement methods. Aurora2 database consisted of English digits and English alphabet-sequence in the presence of additive noise and linear convolutional distortion. There were three test sets in the corpus. Set A had four noise types (subway, babble, car and exhibition hall) that were similar to those in the training data, and set B contained four noise types (restaurant, street, airport and station noise) different from those in the training data. An additional convolutional channel was used in set C. All these three test sets consisted of six SNR conditions (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB) and clean condition. Acoustic models in clean and multi-conditional noise conditions were estimated for comparison. There were 8,440 clean training utterances. The multi-conditional training data consisted of the same utterances artificially added with four different noise types (subway, babble, car and exhibition hall) in five different environmental conditions (5 dB, 10 dB, 15 dB, 20 dB and clean).

Speech features consisted of 13 MFCC coefficients and energy along with the delta and acceleration coefficients. There were 39 features extracted for each frame. In this paper, we estimated continuous-density hidden Markov models (HMM's) and built the recognizer using HTK toolkit package [12]. Some parameters were used: 1) 16 states per word; 2) 3 mixture components of Gaussian density per state; 3) only the variances of all acoustic coefficients are used; 4) optimal subspace decomposition was done by performing hypothesis testing frame by frame and significance level α was set to be 0.05 in multi-condition training and 0.02 in clean training. In speech enhancement procedure, we used 40 sampling point for a frame. The filter of 40×40 matrix was estimated. When computing the covariance matrix, a window of 9 frames was used. The control parameter μ was dynamically specified according to the SNR measured in each frame. Larger μ corresponded to smaller residual noise and larger signal distortion. In the experiments, we preset the range $\mu = 0 \sim 4$.

4.2. Evaluation of SNR Performance

In this subsection, we collected test sets containing six SNR conditions in Aurora2 database for evaluation of SNR's when applying FA speech enhancement. Assuming clean speech and noise signals are independent, SNR formula is calculated by

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{t=1}^{T}\sum_{k=1}^{K} y_t^2(k)}{\sum_{t=1}^{T}\sum_{k=1}^{K}\left(z_t(k)-y_t(k)\right)^2}\times 100\%. \tag{21}$$

where y is clean speech signal and z is noisy speech signal. In Figure 1, the SNR's defined in Aurora2 are similar to those calculated by (21). When applying FA speech enhancement, we find that SNR's are significantly improved for different SNR conditions. The SNR evaluation shows that FA

approach does suppress the noise level. Such suppression does not assure small distortion of speech signal itself. Namely, over suppression of noise in noisy speech will cause the series distortion of speech signal at the same time. To verify the effectiveness of using FA approach, we further conduct experimental comparison for the application of noisy speech recognition using Aurora 2 database.

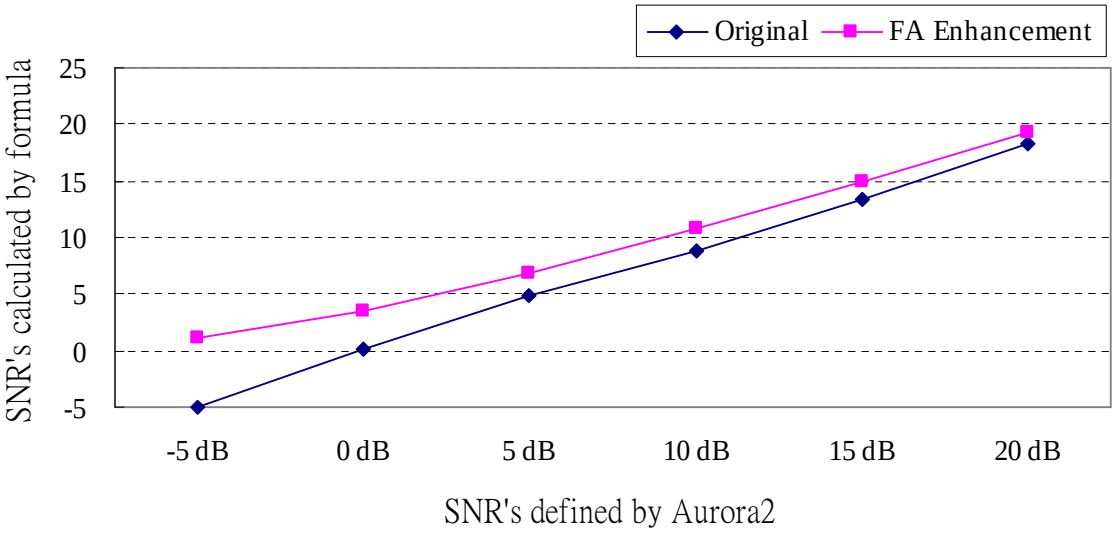

Figure 1. SNR improvement using FA speech enhancement

4.3. Evaluation of Speech Recognition Performance

Table 1 reports the baseline word accuracy (%) results of using clean training set in Aurora2, Tables 2 and 3 show the results after enhancement using signal subspace (SS) and factor analysis (FA) enhancement. On average, in Tables 1 and 3, the relative word error rate is improved by 14.88%, specifically in 0 dB (29.01%), -5 dB (24.88%), and 5 dB (18.18%). Error reduction is achieved for cases of Car (26.01%), Subway (19.47%), and Station (19.34%) environments. Figure 1 shows the performances for baseline system and two enhancement approaches in different environments.

Table 1. Baseline results for clean training

Aurora 2 Clean Training - Results (Baseline)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	98.83	98.61	98.78	99.01	98.81	99.29	99.24	99.25	99.48	99.32	99.42	99.00	99.21	99.09
20 dB	97.61	97.97	98.51	97.35	97.86	98.43	97.52	98.24	98.95	98.29	94.11	94.89	94.50	97.36
15 dB	95.09	94.20	95.71	94.82	94.96	95.36	94.35	95.32	95.68	95.18	87.96	89.45	88.71	93.79
10 dB	84.83	81.38	80.55	84.70	82.87	85.63	82.16	83.21	82.66	83.42	75.74	76.45	76.10	81.73
5 dB	63.77	59.37	48.97	56.09	57.05	63.31	54.23	60.72	54.89	58.29	54.59	51.72	53.16	56.77
0 dB	35.12	35.04	22.70	25.15	29.50	37.21	28.48	35.88	26.94	32.13	29.44	26.90	28.17	30.29
-5 dB	15.08	19.17	11.09	12.47	14.45	18.36	15.30	18.19	14.56	16.60	14.09	13.75	13.92	15.21
average	70.05	69.39	65.19	67.08	67.93	71.08	67.33	70.12	67.59	69.03	65.05	64.59	64.82	67.75

Table 2. Signal subspace enhancement for clean training

Aurora 2 Clean Training - Results (SS Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	99.26	99.06	99.14	99.51	99.24	99.26	99.24	99.14	99.51	99.29	99.32	99.00	99.16	99.24
20 dB	98.28	97.40	98.09	97.81	97.90	97.97	97.67	97.55	97.55	97.69	97.42	96.55	96.99	97.63
15 dB	95.52	92.87	96.69	94.48	94.89	94.32	93.23	94.18	94.18	93.98	95.33	93.74	94.54	94.45
10 dB	87.69	81.89	90.34	85.34	86.32	84.07	83.92	84.46	84.46	84.23	87.81	83.65	85.73	85.36
5 dB	74.70	64.06	74.59	62.94	69.07	65.18	67.02	67.52	67.52	66.81	73.07	65.63	69.35	68.22
0 dB	52.16	39.84	45.45	38.29	43.94	41.23	39.02	43.66	43.66	41.89	46.05	39.21	42.63	42.86
-5 dB	28.31	21.49	20.34	16.11	21.56	17.93	16.84	21.47	21.47	19.43	19.62	16.14	17.88	19.97
average	76.56	70.94	74.95	70.64	73.27	71.42	70.99	72.57	72.62	71.90	74.09	70.56	72.32	72.53

Table 3. Factor analysis enhancement for clean training

Aurora 2 Clean Training - Results (FA Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	99.32	99.03	99.19	99.48	99.26	99.32	99.21	99.22	99.54	99.32	99.39	98.94	99.17	99.26
20 dB	98.31	97.88	98.45	97.84	98.12	98.25	97.85	98.06	98.80	98.24	97.54	96.80	97.17	97.98
15 dB	95.95	93.86	96.72	95.62	95.54	95.15	94.77	95.14	95.99	95.26	95.30	93.86	94.58	95.24
10 dB	89.04	83.83	91.47	88.98	88.33	86.06	85.73	85.45	86.76	86.00	88.21	83.49	85.85	86.90
5 dB	77.04	66.08	76.53	68.22	71.97	68.25	69.80	69.79	72.72	70.14	74.74	66.81	70.78	71.00
0 dB	55.76	42.38	47.27	43.26	47.17	44.34	41.05	44.23	41.59	42.80	47.44	40.93	44.19	44.83
-5 dB	28.80	23.88	18.67	17.49	22.21	23.18	19.14	24.31	20.55	21.80	20.36	17.32	18.84	21.37
Average	77.75	72.42	75.47	72.98	74.66	73.51	72.51	73.74	73.71	73.37	74.71	71.16	72.94	73.80

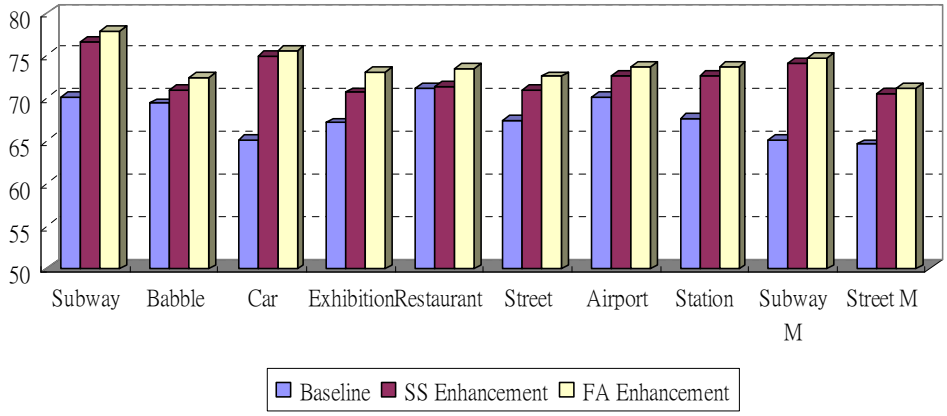

Figure 2. Comparison of different environments for clean training

In Figure 2, we find that the proposed FA enhancement performs better than SS enhancement especially in presence of larger background human voices, e.g. the noise conditions of exhibition and restaurant. Experimental results using multi-condition training for baseline system, signal subspace and factor analysis enhancement are also reported in Tables 4, 5 and 6, respectively.

Table 4. Baseline results for multi-condition training

Aurora 2 Multicondition Training - Results (Baseline)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	99.02	98.73	98.88	99.01	98.91	99.02	98.91	98.87	99.01	98.95	98.86	98.85	98.86	98.92
20 dB	98.43	98.04	98.27	97.99	98.18	98.71	98.46	98.63	98.83	98.66	97.76	97.64	97.70	98.28
15 dB	97.42	97.82	97.88	97.16	97.57	98.43	97.28	98.21	98.46	98.10	96.96	96.31	96.64	97.59
10 dB	95.00	96.43	95.97	94.17	95.39	96.90	95.89	96.96	97.04	96.70	94.20	93.68	93.94	95.62
5 dB	89.28	89.48	88.07	88.06	88.72	91.31	88.88	92.48	89.63	90.58	82.65	83.22	82.94	88.31
0 dB	67.39	65.69	53.50	64.89	62.87	71.26	66.60	72.08	64.49	68.61	47.44	56.65	52.05	63.00
-5 dB	25.15	30.14	19.59	24.56	24.86	37.95	30.59	35.73	27.21	32.87	18.36	26.00	22.18	27.53
average	81.67	82.33	78.88	80.83	80.93	84.80	82.37	84.71	82.10	83.49	76.60	78.91	77.76	81.32

Table 5. Signal subspace enhancement for multi-condition training

Aurora 2 Multicondition Training - Results (SS Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	98.86	98.85	98.84	99.04	98.90	98.86	98.97	98.84	99.04	98.93	98.68	98.94	98.81	98.89
20 dB	98.77	98.25	98.60	98.09	98.43	98.80	98.43	98.72	98.95	98.73	98.00	98.04	98.02	98.47
15 dB	97.45	97.73	98.18	97.25	97.65	98.25	97.46	98.24	98.52	98.12	96.99	96.40	96.70	97.65
10 dB	96.28	95.37	96.48	95.34	95.87	95.43	95.77	96.60	96.82	96.16	95.52	94.07	94.80	95.77
5 dB	90.54	89.45	91.74	90.03	90.44	86.61	88.94	91.23	90.74	89.38	87.96	86.19	87.08	89.34
0 dB	71.72	66.75	74.89	67.23	70.15	70.03	73.43	76.44	78.46	74.59	72.46	60.37	66.42	71.18
-5 dB	49.19	35.07	50.40	47.82	45.62	41.93	43.77	44.94	45.82	44.12	41.20	36.67	38.94	43.68
average	86.12	83.07	87.02	84.97	85.29	84.27	85.25	86.43	86.91	85.72	84.40	81.53	82.96	85.00

Table 6. Factor Analysis enhancement for multi-condition training

Aurora 2 Multicondition Training - Results (FA Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	98.99	98.70	98.87	99.04	98.90	98.99	99.03	98.87	99.07	98.99	98.86	98.97	98.92	98.94
20 dB	98.53	98.22	98.36	98.06	98.29	98.80	98.52	98.72	99.01	98.76	98.13	98.04	98.09	98.44
15 dB	97.85	97.52	98.30	97.19	97.72	98.40	97.52	98.33	98.58	98.21	96.93	96.70	96.82	97.73
10 dB	96.84	96.16	96.99	95.28	96.32	96.53	96.13	96.99	97.44	96.77	95.73	94.56	95.15	96.27
5 dB	91.00	90.60	92.54	89.95	91.02	89.59	90.57	92.25	92.22	91.16	88.30	87.30	87.80	90.43
0 dB	77.03	68.26	79.51	67.48	73.07	73.75	75.18	78.97	78.68	76.65	73.75	64.72	69.24	73.73
-5 dB	50.51	41.38	50.43	47.95	47.57	44.89	45.50	48.23	47.36	46.50	41.97	37.36	39.67	45.56
Average	87.25	84.41	87.86	84.99	86.13	85.85	86.06	87.48	87.48	86.72	84.81	82.52	83.67	85.87

When looking at Tables 4 and 6, we find that the relative word error rate is improved by 22.15%. The performances are improved in 5 dB (32.92%), 10 dB (28.30%), 20 dB (23.47%) and 15 dB (23.24%). The error reduction is obtained for Car (29.08%), Subway (28.72%), and Exhibition (23.85%) environments. The relative improvement percentage at -5 dB SNR in clean training (7.27%) is much less than that in multi-condition training (24.88%). This is because that well-trained clean models could not predict unknown noise influence. Performances for baseline and two enhancement approaches in different environments using multi-condition training are also shown in Figure 2.

From the experiments, we find that in noise environments of subway, car, and station, the speech recognition improvements were larger compared to other noise environments. This is because that machine noises are quite different from human voice noises. More human sound in background noise obtains fewer improvement using proposed enhancement methods. In this work, we evaluate the performances of FA enhancement using SNR measures and speech recognition rates. The experiments

show that FA enhancement is better than SS enhancement. This implies that a generalized model is desirable for representing signals. Moreover, the optimal subspace decomposition by hypothesis testing is used frame by frame to find the optimal filter with suitable dimensions of residual subspace. Such technique is also beneficial to obtain better performance than SS enhancement.

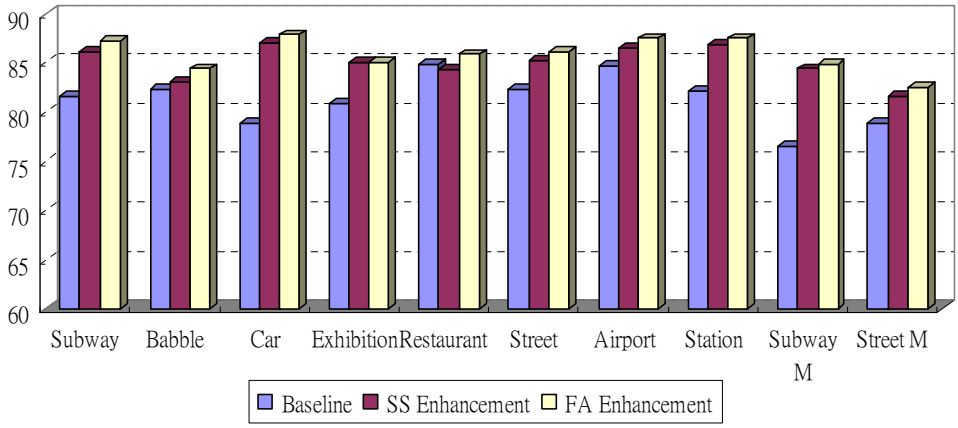

Figure 3. Comparisons for different environments (multi-condition training)

5. Conclusions

In this paper, we have presented FA enhancement of noisy speech signal for application of speech recognition. Interestingly, we built the bridge between FA and signal subspace approaches. Experimental results showed that the proposed approach improved the performance of ASR systems especially in low SNR environments. Compared to other subspace approaches, we presented a novel hypothesis testing approach to optimally perform subspace decomposition. In the future, we will extend this speech enhancement approach by considering the phase effect. Also, we will also improve FA framework through developing new estimation criteria for factor loading matrix. Additionally, we will also explore FA approaching to other speech related applications, e.g. speaker adaptation and acoustic modeling.

6. References

[1] B. Alexander, *Statistical factor analysis and related methods*, John Wiley & Sons, Inc., 1994.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[3] J.-T. Chien and C.-W. Ting, "Speaker identification using probabilistic PCA model selection", *Proc. of ICSLP*, vol. 3, pp. 1785-1788, 2004.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, 1995.

[5] Y. Hu, and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, 2003.

[6] F. Jabloun and B. Champagne, "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise", *Proc. of ICASSP*, vol. 1, pp. 569-572, 2002.

[7] F. Jabloun and B. Champagne, "On the use of masking properties of the human ear in the signal subspace speech enhancement approach," *Proc. Int. Workshop Acoust. Echo Noise Control, Darmstadt, Germany*, pp. 199–202, 2001.

[8] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, 2001.

[9] L. K. Saul, M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 115-125, 2000.

[10] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers", *Neural Computation*, vol. 11, pp. 443-482, 1999.

[11] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Transactions on Speech and Audio Processing*, vol. 7, no.2, pp. 126-137, 1999.

[12] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, The HTK Book, Cambridge University Speech Group, 2000.