Jian SUN[1], Xiongwei ZHANG[2]

Institute of Communication Engineering, PLA Univ. of Sci. & Tech. (1), Institute of Command Automation, PLA Univ. of Sci. & Tech. (2)

Analysis of State-Space Model based Voice Conversion

Abstract. *A new State-Space Model (SSM) based voice conversion method has been proposed recently which outperforms the traditional Gaussian Mixture Model (GMM) method. Although the implementation process of the new method has been elaborated, the theoretical essence of this method has not been analysed clearly. In this paper an exhaustive analysis of the SSM based method is given theoretically and experimentally. Through these analysis, much simpler equivalence form and performance upper bound of the new method are obtained. Finally possible improvements are discussed.*

Streszczenie. *Przedstawiono teoretyczna i eksperymentalną analizę nowego algorytm SSM przetwarzania sygnału mowy. (**Model SSM przetwarzania sygnałów głosowych**)*

Keywords: Voice conversion; State-Space Model (SSM); Linear Multivariate Regression (LMR); analysis
Słowa kluczowe: przetwarzanie głosu, model SSM.

1 Introduction

Voice Conversion (VC) aims to modify a speaker's voice to be perceived as uttered by another speaker, as shown in Fig.1. It can be applied to many areas. Text-To-Speech (TTS) would be one of the most important applications in which it hopes to synthesize different speaking style voices without a large utterance corpus [1]. By connecting VC module to the traditional TTS system, the output of TTS system can be converted into new voice with the target speaker's characteristic. In this way only a small amount of data is needed to train the module. Obviously the new plan is timesaving and easy to use.

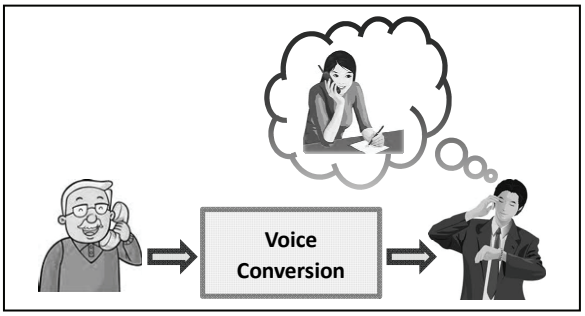

Fig.1. Schematic diagram of voice conversion

Since the VC problem was formulated in 1988 by Abe et al [2], several conversion methods have been proposed [3-6]. Despite many breakthroughs have been achieved in this area, the problem is far from fully solved.

In 2009 Xu et al proposed a new VC method using the SSM to model and transform the voice spectral features [7]. The main advantage of using SSM in voice conversion lies in its ability to model spectral parameter trajectory explicitly. The correlation between adjacent frames of speech is usually ignored in previous methods, especially in the frame-by-frame analysis framework like GMM based method. Additionally in this new method, the spectral parameters are decomposed into common information and differentia information. This decomposition property of SSM may provide a possible approach for separating the speaker's characteristics from the speech parameters as mentioned in [8].

Although the implementation process of the new method has been elaborated, the essence of this method has not yet been analysed clearly. In this paper the SSM based voice conversion method is analysed theoretically and experimentally. Through the analysis, much simpler equivalence form and performance upper bound of the new method are obtained. Besides, the decomposition property of SSM is analysed by well-designed experiments. And possible improvements of this method are also discussed.

The paper is organized as follows. In the next section, the SSM is briefly reviewed and the voice conversion method based on this model is summarized. In section 3 the theoretical analysis of the SSM based VC method is presented in detail. Furthermore the simpler equivalence form of this method is given. And the decomposition property of SSM on speech is discussed and testified by experiments. Finally, the conclusions and discussions are summarized in section 4.

2 SSM and Voice Conversion method

2.1 State-Space Model

SSM is a model that uses (possibly unobserved) state variables to describe a system by a set of first-order differential or difference equations. More specifically, the system we work with is discrete time linear dynamical system with gaussian noise. The basic model can be written as [9]:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\omega}_t \tag{1}$$

$$\mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \boldsymbol{\upsilon}_t \tag{2}$$

where $\mathbf{x}_t$ is a k-dimensional hidden variable, $\mathbf{y}_t$ is a p-dimensional observation vector. $\mathbf{A}$ and $\mathbf{B}$ are the state transformation and the observation matrices, respectively. Both $\boldsymbol{\omega}_t$ and $\boldsymbol{\upsilon}_t$ are zero mean Gaussian error terms with covariance matrix $\mathbf{Q}$ and $\mathbf{R}$, such that $\boldsymbol{\omega}_t \sim N(0,\mathbf{Q})$, $\boldsymbol{\upsilon}_t \sim N(0,\mathbf{R})$. The k–vector $\boldsymbol{\omega}_t$ and p-vector $\boldsymbol{\upsilon}_t$ represent the state evolution and observation noises respectively, which are independent of each other and independent of the values of $\mathbf{x}_t$ and $\mathbf{y}_t$. Equation (1) is called state equation and equation (2) is called observation equation.

It's assumed that the state of system at time t is specified by the state variable $\mathbf{x}_t$ which can't be observed directly. The output of the system $\mathbf{y}_t$ can be accessed and used to estimate the hidden states and learn the model parameters. For a given SSM, the model parameter set is defined by $\Theta=\{\mathbf{A},\mathbf{B},\mathbf{Q},\mathbf{R}\}$.

Given the observation $\mathbf{y}_t$, there are two common approaches to estimate the hidden states and model parameters: the iterative Expectation Maximum (EM) algorithm [10] and subspace method [11]. The result obtained by iterative EM method is more accurate in comparison with the subspace method which requires less computation time. The timesaving property of subspace method is obvious when the dimensions of state variable and observation vector are high. In [7] the EM algorithm is adopted.

2.2 Voice Conversion Based on SSM

The SSM has the desirable ability to explicitly model spectral parameter trajectory of speech signal which can alleviate the discontinuity of adjoining frames in converted speech. Using this property, a SSM based voice conversion method is proposed for the first time in [7].

In this method, the parameter set of SSM is divided into two parts: $\Theta_{comm}=\{\mathbf{A},\mathbf{Q}\}$, denoting the common information and $\Theta_{diff}=\{\mathbf{B},\mathbf{R}\}$, denoting the differentia information. The conversion process can be summurized as:

In the training phase:

1) Analysing the parallel training speech data of source and target speakers using STRAIGHT model [12], resulting in the spectral parameters. Then the spectral parameters are converted to Linear Spectral Frequencies (LSFs) and time-aligned by the Dynamic Time Warping (DTW) algorithm. The time-aligned LSFs of source and target speech can be denoted as $\mathbf{Y}_{train}^{source}$ and $\mathbf{Y}_{train}^{target}$ respectively. More specificly:

$$\mathbf{Y}_{train}^{source}=\left\{\mathbf{y}_{train\ 1}^{source},\mathbf{y}_{train\ 2}^{source}\cdots,\mathbf{y}_{train\ T}^{source}\right\} \tag{3}$$

$$\mathbf{Y}_{train}^{target}=\left\{\mathbf{y}_{train\ 1}^{target},\mathbf{y}_{train\ 2}^{target}\cdots,\mathbf{y}_{train\ T}^{target}\right\} \tag{4}$$

where T is the length of LSFs.

2) Estimating the parameter set and state series of source speaker using EM algrithm from $\mathbf{Y}_{train}^{source}$, resulting in:

$$\Theta^{source}=\left\{\mathbf{A}^{source},\mathbf{B}^{source},\mathbf{Q}^{source},\mathbf{R}^{source}\right\} \tag{5}$$

$$\mathbf{X}_{train}^{source}=\left\{\mathbf{x}_{train\ 1}^{source},\mathbf{x}_{train\ 2}^{source},\cdots,\mathbf{x}_{train\ T}^{source}\right\} \tag{6}$$

3) As the training data are parallel, the state variable series of the SSM trained from $\mathbf{Y}_{train}^{target}$ are assumed to be the same as that trained from $\mathbf{Y}_{train}^{source}$, which means:

$$\mathbf{X}_{train}^{source}=\mathbf{X}_{train}^{target} \tag{7}$$

And so is the Θ_{comm}^{target}, which is:

$$\Theta_{comm}^{target}=\left\{\mathbf{A}^{source},\mathbf{Q}^{source}\right\} \tag{8}$$

Then given $\mathbf{Y}_{train}^{target}$, $\mathbf{X}_{train}^{target}$ and Θ_{comm}^{target}, Θ_{diff}^{target} can be easily estimated, that is:

$$\Theta_{diff}^{target}=\left\{\mathbf{B}^{target},\mathbf{R}^{target}\right\} \tag{9}$$

In the conversion phase:

1) Analysing the new input source speech in the same way mentioned above resulting in the LSFs, denoted as $\mathbf{Y}_{convert}^{source}$.

2) Estimating state series $\mathbf{X}_{convert}^{source}$ of SSM from $\mathbf{Y}_{convert}^{source}$ under the constraint of Θ^{source} which is obtained in the training phase.

3) Converted LSFs can be obtained easily using equation (10):

$$\hat{\mathbf{Y}}_{convert}^{target}\simeq\mathbf{B}^{target}\mathbf{X}_{convert}^{target}=\mathbf{B}^{target}\mathbf{X}_{convert}^{source} \tag{10}$$

where the error term υ is elimiated as its expectation is zero.

The experimental results in [7] show that the SSM-based voice conversion method significantly outperforms the traditional GMM-based technique in the view of both speech quality and conversion accuracy of speaker individuality.

3 Analysis of SSM based VC method

Although the implementation process and experimental results are described clearly in [7], there still are some questions about this method unresolved as yet. For examples: as the SSM is a linear multivariate model, is there any relationship between the SSM based method and the classic Linear Multivariate Regression (LMR) method [13] for VC? What is the performance upper bound of this new method? Can the Θ_{comm} represent the charateristic of speaker and the Θ_{diff} and state series $\mathbf{X}$ represent the voice content information in speech? In order to find out the answers and get a better understanding of the SSM based method, a theoretical and experimental analysis is given in the following.

3.1 Equivalent form and the performance upper bound

From conversion procedure discribed above, it can be known that after SSMs being trained from the parallel LSFs, the spectral parameters can be fitted by the model parameter and state series, as:

$$\mathbf{y}_t^{source}=\mathbf{B}^{source}\mathbf{x}_t^{source}+\upsilon_t^{source} \tag{11}$$

$$\mathbf{y}_t^{target}=\mathbf{B}^{target}\mathbf{x}_t^{target}+\upsilon_t^{target} \tag{12}$$

where $\mathbf{x}_t^{target}=\mathbf{x}_t^{source}$ for t=1,$\cdots$,T since the training data is parallel. In practice, the dimension of state variable p is set smaller than that of observation vector k [7]. Besides, as the training data is sufficient, it can be assumed that $\mathbf{B}^{source}$ is full column rank. Then from equation (11) we have:

$$\mathbf{x}_t^{source}=\mathbf{B}^{source^-}\mathbf{y}_t^{source}-\mathbf{B}^{source^-}\upsilon_t^{source} \tag{13}$$

where $\mathbf{B}^{source^-}$ is the generalized inverse matrix of $\mathbf{B}^{source}$ which satisfied $\mathbf{B}^{source}\mathbf{B}^{source^-}=\mathbf{I}$ [14] and $\mathbf{I}$ is a $k\times k$ identity matrix. Combining equation (12) (13) we obtain:

$$\mathbf{y}_t^{target}=\mathbf{B}^{target}\mathbf{B}^{source^-}\mathbf{y}_t^{source}+\left(\upsilon_t^{target}-\mathbf{B}^{target}\mathbf{B}^{source^-}\upsilon_t^{source}\right) \tag{14}$$

where $\mathbf{B}^{target}\mathbf{B}^{source^-}$ is a $p\times p$ matrix. And recalling that the zero mean Gaussian random terms υ_t^{source}, υ_t^{target} are independent with each other, then $\upsilon_t^{target}-\mathbf{B}^{target}\mathbf{B}^{source^-}\upsilon_t^{source}$ is also zero mean Gaussian. For convenience define:

$$\mathbf{B}^{transform}=\mathbf{B}^{target}\mathbf{B}^{source^-} \tag{15}$$

$$\upsilon_t^{transform}=\upsilon_t^{target}-\mathbf{B}^{target}\mathbf{B}^{source^-}\upsilon_t^{source} \tag{16}$$

Then we can rewrite equation (14) as:

$$\mathbf{y}_t^{target}=\mathbf{B}^{transform}\mathbf{y}_t^{source}+\upsilon_t^{transfrom} \tag{17}$$

From equation (17) we know that the conversion equation (10) can be expressed in another way as:

$$\hat{\mathbf{Y}}_{convert}^{target}\simeq\mathbf{B}^{transform}\mathbf{Y}_{convert}^{source} \tag{18}$$

where the error term υ is elimiated the same as that in equation (10).

Now the SSM based VC procedure can be reinterpret as follows:

In training phase, the parameter $\mathbf{B}^{transform}$ is estimated from parallel training data using EM algorithm. In the conversion phase, the converted LSFs are obtained by equation (18) using $\mathbf{B}^{transform}$ estimated in training phase.

Obviously this procedure is structurally equivalent to the LMR based method which has been used in [13].

For LMR, the coefficient $\mathbf{B}^{transform}$ is trained by Least Square (LS) algorithm minimizing the sum of $\left\|\upsilon_t^{transform}\right\|^2$ over all the training data. As the error term $\upsilon_t^{transform}$ is Independently Identically Distributed (IID) and obeys normal distribution for all t, the result estimated by LS method is also the ML (Maximum Likelihood) estimate. But for $\mathbf{B}^{transform}$ which is trained in the SSM based method through EM algorithm, the result is trying to approach the ML estimate but hardly can be. In other words, the result obtained in LMR based method should be the performance upper bound of SSM based method. In summary, we can conclude that the SSM based method is structurally equivalent to the LMR method while its conversion performance upper bound is that of LMR method using LS algorithm.

In order to demonstrate the conclusions made above, comparison experiments are carried out using the two methods for voice conversion. The results are evaluated both objectively and subjectively. The experiments presented here are using the parallel SLT (US Female) and BDL (US Male) of CMU ARCTIC databases [15] for male to female and female to male conversion tasks. The first 300 sentences are chosen for training with 100 sentences selected randomly from the remaining database to be used for conversion. LSFs order is set to 16 and state order is set to 9 according to [7].

The speech data is first preprocessed to get the time aligned LSFs according to the procedure in SSM based method. Then SSM and LMR based methods are used for spectral conversion respectively. In the evaluation process, the Average Spectral Distortion Measure (ASDM) [16] is used to evaluate the objective performance of the two methods:

$$\varepsilon_{ASDM} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{\pi}\int_0^{\pi}\left[10\log_{10}\mathbf{s}_t(\omega)-10\log_{10}\hat{\mathbf{s}}_t(\omega)\right]^2 d\omega} \tag{19}$$

where $\mathbf{s}_t$ and $\hat{\mathbf{s}}_t$ represent the target and converted power spectra of the tth frame in speech. The evaluation results are shown in Fig.2.

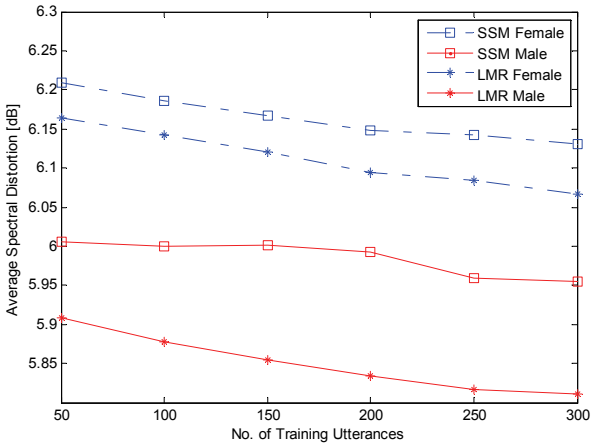

Fig.2. ASDM for SSM and LMR based VC methods against number of training data

As shown in Fig.2, it's obvious that the LMR based method outperforms the SSM based method.

In the subjective evaluation, a similarity listening test is conducted. 5 listeners give scores between 1 and 5 for measuring the similarity of the converted speeches produced by the two methods. 1 means totally dissimilar and 5 means identical. The averaged results are shown in Table 1 which means the outputs of the two methods are almost identical.

Table 1. Subjective similarity evaluation of the two methods

	Averaged Scores
Female to Male conversion	5
Male to Female conversion	5

From the subjective experiment, we know that the conversion results produced by the two methods are almost identical which validates the structurally equivalence of the two methods. And the objective evaluation result demonstrates the conclusion that the performance of LMR based VC method trained by LS algorithm is the upper bound of SSM based method.

3.2 Analysis of the decomposition property

By recalling the VC procedure in SSM method, we know that the conversion process is implemented by substituting the $\mathbf{B}^{source}$ with $\mathbf{B}^{target}$ in equation (10) to produce the converted LSFs. Besides the state series $\mathbf{X}$ is assumed to be the the same between parallel speech data. And the parameter set Θ of one speaker is considered to be unchanged across the whole conversion process.

In other words, under the assumption of SSM method the parallel speech data from different speakers can have the same state series $\mathbf{X}$ and each speaker has his/her own invariant Θ_{diff}. The converison is accomplished by combining the $\mathbf{X}$ of source speech and the target speaker's $\mathbf{B}$ of his/her Θ by equation (10).

Then we can infer that the parameter $\mathbf{B}$ contains speaker's charateristic information of the speech otherwise the voice converted by this procedure can't sound like the tagart speaker's. Further more, another inference can be made that the SSM trained by EM algorithm using LSFs has the ability to decompose the speech parameters into speaker's characteristic information which is contained in Θ_{diff} and voice content related information which is represented by $\mathbf{X}$.

If these inferences held, SSM based VC method would pioneer a path for seperating the speaker's characteristic information and voice content related information of speech. The successful seperation of these two parts can not only be beneficial to VC process, but also be helpful to other areas in speech processing. But whether or not the SSM based VC method really accomplished, these should to be verified.

As we know, the EM algorithm which is used for training the SSM model is a method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. And the EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the latent variables, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. So the results obtained by the EM algorithm is an approximation to the ML estimate. As mentioned above, the error term $\mathbf{v}_t^{transform}$ in SSM is IID and obeys normal distribution for all t, then its ML estimate is also the LS result, which is:

$$\{\mathbf{B},\mathbf{R},\mathbf{X}\} = \arg\min_{\{\mathbf{B},\mathbf{R},\mathbf{X}\}} \sum_t \left\|\mathbf{y}_t - \mathbf{B}\mathbf{x}_t\right\|^2 \tag{20}$$

From this, it's known that the parameters esitimated by EM algorithm are trying to minimize the sum of the squared error in observation equation over all t iteratively. Guided only by this direction, definitely no guarantee can be made to insure that the algorithm can seperate the speech parameters into speaker's characteristic information and voice content related information. In the followings, a simple expriment is carried out to testify this conclusion.

Through the experiment, the invariability of paremeter $\mathbf{B}$ is examined. No doubt, if Θ_{diff} contains the speaker's characteristic information of speech, the parameter $\mathbf{B}$ in Θ_{diff} should be almost invariant when the training data is different and sufficient while collected from the same people under the same acquisition environment. In the experiment, the SLT (US Female) and BDL (US Male) of CMU ARCTIC databases are used as the training data. Each time 300 sentences are selected randomly from one person's speech corpus as training data. The SSM for each person is trained 5 times, then the difference of $\mathbf{B}$ for each speaker and the difference of $\mathbf{B}$ between the two speakers are compared. The difference for each speaker is obtained by:

(21) $$d_{inter} = \frac{1}{I}\sum_{i=1}^{I} s\left(\left|\mathbf{B}_i - \bar{\mathbf{B}}\right|\right)$$

where $\bar{\mathbf{B}}$ is the mean matrix of $\mathbf{B}_i$ for i=1,⋯,I and I is the number of training times for each speaker. Function s calculates the sum of the matrix's elements.

The difference for the two speakers is given by:

(22) $$d_{exter} = s\left(\left|\bar{\mathbf{B}}_{SLT} - \bar{\mathbf{B}}_{BDL}\right|\right)$$

The results are shown in Table 2. The SLT and BDL are acquired from people of different sexes, so the characteristic difference between them should be quite obvious. However, from the results we know that the differences of $\mathbf{B}$ inside each speaker and between the two speakers aren't big enough to lead us to the invariability conclusion about $\mathbf{B}$. Thus it can be seen that the parameter $\mathbf{B}$ can't be said representing the speaker's characteristic information of speech. In addition the inference made about the SSM method's seperation ability is faulty too.

Table 2. The comparisons of the differences of $\mathbf{B}$

	SLT	BDL	SLT vs. BDL
The difference of $\mathbf{B}$	8966	5044	10934

4. Conclusions and discussions

From the theoretical analysis and experimental results, the following conclusions are reached: 1) the SSM based VC method is structurally equivalent to the LMR based method; 2) the conversion results of LMR based method trained by LS algorithm is the performance upper bound of that obtained by SSM based method; 3) the SSM method doesn't have the ability to separate speaker's characteristic information and voice content related information from speech parameters.

The reason why speaker's characteristics can be transformed in the SSM method lies in the fact that it embeds an affine transformation, as shown by equation (18). The merit of this method which makes it outperform over GMM method maybe attribute to its ability to model spectral envelope evolution of speech. By this mean the discontinuity between adjacent frames in GMM method can be alleviated.

Although the SSM based method doesn't bring a breakthrough for VC, it does present some new ideas for guiding future research on VC. On one hand, it introduces a new model for analyzing the speech data which can alleviate the discontinuity between adjacent frames by modelling spectral parameter trajectory. In the current work, only the most general linear SSM is used which underestimates the complexity of the relationship between hidden variables and observation results in speech signal. In the future work, nonlinear SSM can be tried to model this relationship in order to get a better fitting results surpassing the perform-ance upper bound of currunt work.

On the other hand, there is a brand new idea embedded in the current work: seperating the common and the differentia information of speech parameters and implem-enting the VC process based on the seperation. Although in [7] the two parts don't be named as speaker characteristic information and voice content related information, this underlying meaning can be infered from the conversion process. Whatever the seperation result is, the way by which VC was performed is very illuminating. In the future work, the restrictions which can guide the direction of the seperation in SSM method rather than towards the LS result should be explored. Maybe under some suitable directions, the speaker characteristic information and voice content related information can be properly seperated from speech in the framework of SSM. Then the performance of SSM based method will be improved fundamentally.

Acknowledgement

The authors wish to acknowledge the financial support of Natural Science Foundation of Jiangsu Province in 2009(BK2009059) and Pre-research Foundation of PLA University of Sci. & Tech. in 2009(2009TX08).

REFERENCES

[1] Stylianou Y., (2008). Voice Transformation: A survey, Springer Handbook of Speech Processing, *ICASSP*, (2009), 3585-3588

[2] Abe M., Nakamura S., Shikano K., Kuwabara H., Voice conversion through vector quantization, *ICASSP*, (1988), 655-658

[3] Stylianou Y., Cappe O., Moulines E., Continuous probabilistic transform for voice conversion, *Speech and Audio Processing, IEEE Transactions on*, 6(1998), No. 2, 131-142

[4] Shuang Z., Bakis R., Qin Y., IBM Voice Conversion Systems for 2007 TC-STAR Evaluation, *Tsinghua Science & Technology*, 13(2008), No. 4, 510-514

[5] Erro D., Moreno A., Bonafonte A., Voice Conversion based on Weighted Frequency Warping, *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2010), No. 5, 922-931

[6] Narendranath M., Murthy H. A., Rajendran, S., Yegnanarayana B., Transformation of formants for voice conversion using artificial neural networks, *Speech Commun.*, 16(1995), No. 2, 207-216

[7] Xu N., Yang Z., Zhang L. H., Zhu W. P., Bao, J. Y., Voice conversion based on state-space model for modelling spectral trajectory, *Electronics Letters*, 45(2009), No. 14, 763-764

[8] Popa V., Nurminen J., Gabbouj M., A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models, *the Proc. of the 10th Annual Conference of the International Speech Communication Association*, (2009), 2655-2658

[9] Roweis S., Ghahramani Z., A unifying review of linear Gaussian models, *Neural Comput.*, 11(1999), No. 2, 305-345

[10] Tanizaki H., Nonlinear Filters: Estimation and Applications (2nd edn.), Berlin: Springer-Verlag, (1996)

[11] Ljung L., System identification: theory for the user (2nd edn.), Upper Saddle River, NJ: Prentice Hall PTR, (1999)

[12] Kawahara H., Masuda-Katsuse I., de Cheveign, A., Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, *Speech Communication*, 27(1999), No. 3-4, 187-207

[13] Valbret H., Moulines E., Tubach J. P., Voice transformation using PSOLA technique, *Speech Communication*, 11(1992), No. 2-3, 175-187

[14] Ben-Israel A., Greville T. N. E., Generalized inverses theory and applications (2nd edn.), New York: Springer, (2003)

[15] Kominek J., Black A. W., CMU ARCTIC databases for speech synthesis, *5th ISCA Speech Synthesis Workshop*, (2003), 223-224

[16] Xydeas C. S., Papanastasiou, C., Split matrix quantization of LPC parameters, *Speech and Audio Processing, IEEE Transactions on*, 7(1999), No. 2, 113-125

***Authors**:*

Jian Sun, Postgraduate Team 2, Institute of Communications Engineering, Biaoyin 2, Yudao Street, Nanjing, China, 210007, Email: sunjian001@gmail.com;

Xiongwei Zhang, Institute of Command Automation, PLA Univ. of Sci. & Tech.