

Chapter 3

A Voice Conversion Method Based on the Separation of Speaker-Specific Characteristics

Zhen Ma, Xiongwei Zhang and Jibin Yang

Abstract This paper aims to study independent and complete characterization of speaker-specific voice characteristics. Thus, the authors conduct a method on the separation between voice characteristics and linguistic content in speech and carry out voice conversion from the point of information separation. In this paper, authors take full account of the K-means singular value decomposition (K-SVD) algorithm which can train the dictionary to contain the personal characteristics and inter-frame correlation of voice. With this feature, the dictionary which contains the personal characteristics is extracted from training data through the K-SVD algorithm. Then the authors use the trained dictionary and other content information to reconstruct the target speech. Compared to traditional methods, the personal characteristics can be better preserved based on the proposed method through the sparse nature of voice and can easily solve the problems encountered in feature mapping methods and the voice conversion improvements are to be expected. Experimental results using objective evaluations show that the proposed method outperforms the Gaussian Mixture Model and Artificial Neural Network based methods in the view of both speech quality and conversion similarity to the target.

Keywords Voice conversion · Speaker-specific characteristics · Information separation · K-SVD

Z. Ma (✉) · X. Zhang · J. Yang
Institute of Communication Engineering, PLA University of Science and Technology,
Nanjing, China
e-mail: mazhen1989@126.com

3.1 Introduction

Voice conversion systems modify a speaker's voice to be perceived as uttered by another speaker [1]. Voice conversion technology can be applied to many areas. Definitely, one of the most important applications is the use in the area of text-to-speech synthesis (TTS) where it hopes to synthesize different speaking style voices without a large utterance corpus. Besides speech synthesis, however, voice conversion has other potential applications in areas like entertainment, security and speaker individuality for interpreting telephony and voice restoration.

Many conversion methods have been proposed since the problem was first formulated in 1988 by Abe et al. [2]. Generally speaking, the main methods such as Gaussian Mixture Model (GMM) [3], Hidden Markov Model (HMM) [4], Frequency Warping (FW) [5] and Artificial Neural Network (ANN) [6] based methods has gained similarity in converted speech together with quality declines dramatically. Though many improved methods like bring Global Variance and Parameter Trajectory in GMM based method has been put forward over the last few years, and all upgraded in quality to some extent, the conversion results is not satisfactory for practical application. Therefore, we need to pursue new conversion methods to achieving high-quality converted speech.

It's clear that the information separation of speech signals is going to be a new approach for speech processing. Among the information, the meaning of message being uttered is of prime importance, and the information of speaker identity also plays an important role in oral communication. We define these two types of information as content and style of speech respectively. General knowledge tells us that human auditory perception system has the ability to identify the meaning of message and the speaker identity simultaneously, i.e. to separate content and style factors sophisticatedly from speech signals.

In Popa Victor's work, the bilinear model was used to model speech, and two sets of speech parameters, indicating style and content respectively, are achieved by singular value decomposition (SVD) on speech observations [7]. With the separation result, voice conversion is realized by replacing the source style with the target one, while preserving the initial content, i.e. the converted speech is reproduced based on source content and target style. Based on this literature, in this paper we use a new method to separate the information of speaker identity.

K-means singular value decomposition (K-SVD) is a signal decomposition method aims at sparse representation of signals. This method has been successfully used in image denoising, character extracting and so on. Because of the sparsity nature and other characteristics, we can use K-SVD to decomposing the vocal tract spectrum into a dictionary which conveys the personal identity and corresponding sparse matrix which contains the content information. Taking this into account, we can use the dictionary to achieve the separation of speaker-specific characteristics and speech conversion based on substitution of the dictionary.

Combining the identity of speech with K-SVD, we implement voice conversion using style replacing technique as proposed in Popa Victor's work. Experimental

results show that the proposed voice conversion approach outperforms the traditional GMM and ANN method in terms of both similarity and naturalness, especially in the case of small size of training dataset.

The paper is organized as follows. In the next section, we briefly introduce the K-SVD theory. In Sect. 3.3, we describe the voice conversion scheme based on K-SVD. Next, we make experiments to evaluate the proposed method, and the results demonstrate the benefits of K-SVD comparing with GMM and ANN in voice conversion. Finally, we make remarks on the proposed method, and some potential future research directions are also presented in section V.

3.2 The K-SVD Algorithm

We have witnessed a growing interest in the use of sparse representations for signals in recent years. Using an overcomplete dictionary matrix $D \in \mathbb{R}^{n \times K}$ that contains K atoms, $\{d_j\}_{j=1}^K$, as its columns, it is assumed that a signal $Y \in \mathbb{R}^n (n \ll K)$ can be represented as a sparse linear combination of these atoms. The representation of Y may be approximate, $Y \approx DX$, satisfying $\|Y - DX\|_2 \leq \varepsilon$. The vector $X \in \mathbb{R}^K$ contains the representation coefficients of the signal Y . This sparsest representation is

$$(P_{0,\varepsilon}) \min_x \|x\|_0 \text{ subject to } \|Y - DX\|_2 \leq \varepsilon \quad (3.1)$$

In the K-SVD algorithm [8] we solve (1) iteratively, using two stages, parallel to those in K-Means. In the sparse coding stage, we compute the coefficients matrix X , using any pursuit method, and allowing each coefficient vector to have no more than T_0 non-zero elements. Then, we update each dictionary element sequentially, changing its content, and the values of its coefficients, to better represent the signals that use it. This is markedly different from the K-Means generalizations that were proposed previously, e.g., since these methods freeze X while finding a better D , while we change the columns of D sequentially, and allow changing the relevant coefficients as well. This difference results in a Gauss-Seidel-like acceleration, since the subsequent columns to consider for updating are based on more relevant coefficients. We also note that the computational complexity of the K-SVD is equal to the previously reported algorithms for this task.

We now describe the process of updating each atom d_k and its corresponding coefficients, which are located in the k -th row of the coefficient matrix X , denoted as x_k . We first find the matrix of residuals, and restrict this matrix only to the columns that correspond to the signals that initially use the currently improved atom. Let r be the set of indices of these signals, similarly, denote E_k^r as the restricted residual matrix, which we would now like to approximate using a multiplication of the two updated vectors d_k and x_k , i.e. The equation converted to:

$$\|Y - DX\|_F^2 = \|y - \sum_{j=1}^N d_j x_j^r\| = \|(y - d_j x_j^r) - d_k x_k^r\|_F^2 = \|E_k - d_k x_k^r\|_F^2 \quad (3.2)$$

We seek for a rank-one approximation. Clearly, this approximation is based on the singular value decomposition (SVD), taking the first left and right singular vectors, together with the first singular value. So we can get the sparse representation of speech signals as well as the dictionary.

We can suppose that the personal characteristics are contained in the dictionary while the content is conveyed in coding matrix due to the dictionary open out the main identity and inner structure of speech. And for this hypothesis, we will demonstrate it at the experiment phase.

3.3 Voice Conversion Based On K-SVD

So far, we know that a speech signal can be decomposed to a dictionary conveys the information related to personal identity and a corresponding sparse matrix contains the information related to content. Naturally we can replace the source speaker's dictionary by the target speaker's dictionary and then multiply with the source sparse matrix to implement voice conversion. While this scheme is easy to realize, there is a cute problem that the result trained by K-SVD is of multiplicity which means that a vocal track spectral matrix may have different combination of dictionary and sparse matrix. To settle this problem, we excerpt the means put forward by XU [9] to introduce partial limitation for the vocal track spectral conversion which is shown in Fig. 3.1.

There are two basic modes in the voice conversion system. In training mode, we process the data of source and target speaker (part 1 in Fig. 3.1) and extract the dictionary based on K-SVD (part 2 in Fig. 3.1). In converting mode, we implement voice conversion based on dictionary replacement (part 3 in Fig. 3.1). Next, we will introduce the process of part 1, 2 and 3 concretely.

3.3.1 Data Processing

The source and target speech to be trained are denoted as x and y , in which the contents are the same. As for the extraction of vocal tract spectral, we take use of STRAIGHT model which can separate the excitation to vocal tract better to analyze the speech. The trained speech data analyzed by the STRAIGHT model and we get the STRAIGHT spectrum of source and target speech are denoted as S_x and S_y .

Before training, the S_x and S_y must be aligned in time to ensure the coherence of the coding matrix decomposed at the next step. While the common method

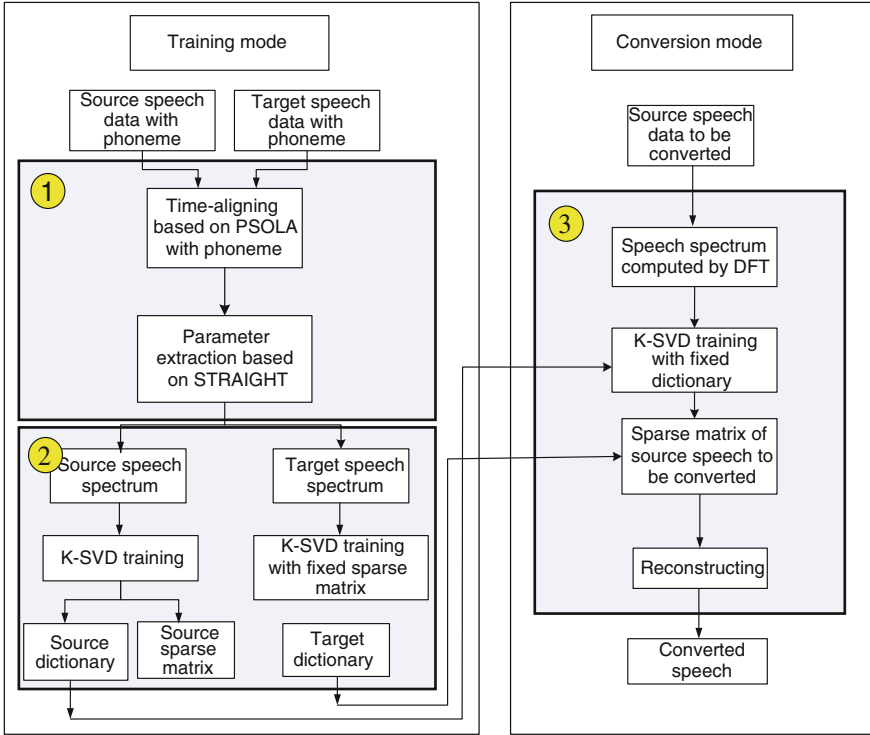


Fig. 3.1 Voice conversion system based on K-SVD

Dynamic Time Warping (DTW) [2] may lead to the quality declines dramatically because of having not considered the continuity of the inter-frame when inserting or removing the sequential frames, we adopt the time-aligned method with phoneme label information put forward in [10]. S'_x and S'_y denote time-aligned source and target STRAIGHT spectrum respectively.

3.3.2 Dictionary Extraction

So far, we have got the time-aligned STRAIGHT spectrum, and then we will extract the dictionary used for conversion.

As illustrated in part 2 of Fig. 3.1, using K-SVD to decomposing S'_x , we can get the dictionary D_x and sparse matrix H_x of source speech. If the target speaker's STRAIGHT spectrum is analyzed in the same way, it is difficult to ensure the coherence of the dictionary of the target with the source. Considering the hypothesis that the sparse matrix is determined by the content information, it is clearly that the sparse matrix H_x of S'_y is the same with H_x . Therefore, set $H_y = H_x$

when decomposing S'_y to get corresponding dictionary of the target speaker denoted as D_y .

3.3.3 Voice Conversion Based on Dictionary Replacement

As shown in part 3 of Fig. 3.1, in converting mode we should first extract the STRAIGHT spectrum of the source speech. And then we shall use OMP algorithm to decompose this spectrum with D_x fixed to get the corresponding sparse matrix $H_x^{convert}$. So we can achieve the converted STRAIGHT spectrum based on $H_x^{convert}$ and the dictionary D_y accepted at the training mode according to the Eq. (3.3).

$$S_y^{convert} = D_y \cdot H_x^{convert} \quad (3.3)$$

3.4 Experimental Evaluations

3.4.1 Experimental Conditions

We conducted two experimental evaluations. Firstly, in order to demonstrate the dictionary conveys the speaker-specific characteristics and the sparse matrix conveys the content information, we use K-SVD to separate the speaker-specific characteristics with content information. Secondly, in order to demonstrate the effectiveness of the proposed conversion method, we compared the method against the GMM and ANN based methods, which are the most popular conversion techniques in the past few years.

In our experiments, all the data are selected from the CMU ARCTIC databases. It consists of four different database named SLT, BDL, JMK and AWB uttered by four different speakers.

The two parameters may affect the conversion result for the K-SVD; they are the numbers of atoms (K) in the dictionary and the sparsity (t) of the sparse matrix. It is clear that the reconstruction error can be decreased with the increasing of the two parameters. However, the result is not always better as the parameters increase too much. We have found that when the two parameters fill Eq. (3.4)

$$K = 40, t = 12 \quad (3.4)$$

The converted voice is the best in quality as well as similarity through many experiments. Therefore, we set the two parameters as Eq. (3.4) at the next experiments.

3.4.2 Speaker-Specific Characteristics Separation

To demonstrating the speaker-specific characteristics are included in the dictionary, we do the work as follow.

Firstly, we select the preceding 10 sentences in BDL and SLT and compute the STRAIGHT spectrum respectively, the dimensionality of each row vector is 256. Secondly, getting the dictionary D_{BDL} and D_{SLT} through K-SVD. Then, we choose the succedent 5 sentences in BDL and SLT to calculate the STRAIGHT spectrum. At last, computing the reconstructed error with the dictionary settled as D_{BDL} and D_{SLT} . The reconstructed error is calculated using,

$$e = 10 \log_{10} \left(\frac{1}{N_S} \sum_{i=1}^{N_S} \|\mathbf{s}_i - \hat{\mathbf{s}}_i\|^2 \bigg/ \frac{1}{N_S} \sum_{i=1}^{N_S} \|\mathbf{s}_i\|^2 \right) \quad (3.5)$$

where N_S denotes the number of rows in STRAIGHT spectrum, \mathbf{s}_i denotes the i -th row of the source spectrum, $\hat{\mathbf{s}}_i$ denotes the i -th row of the reconstructed spectrum. The results are summarized in Table 3.1.

It is clear that the reconstructed error of dictionary suited to the voice is much less than the other when analyzing by K-SVD. Therefore, we can assure that the dictionary trained by K-SVD is certain to containing the speaker-specific characteristics, and we can separate the speaker-specific characteristics effectively. And according to this, we can implement voice conversion by replacing the dictionary.

3.4.3 Effectiveness of the Proposed Method

In Fig. 3.2, we can see the source, target and converted speech directly. To contrasting the converted result, two subjective listening tests were carried out. One test is Mean Opinion Score (MOS), which aims to give a score between 1 and 5 to evaluate speech naturalness, and the other ABX which uses the correct rate of speaker recognition to evaluate the successfulness of individuality conversion. All tests were performed by five listeners who have been engaged on speech processing research for many years. The results are given in Table 3.2.

From the experimental results shown in Table 3.2, it is easy to see that the proposed method for voice conversion based on style and content separation clearly outperform traditional GMM and ANN algorithm based on parameter

Table 3.1 Results of reconstructed error (db)

	D_{BDL}	D_{SLT}
BDL	-18.02	-15.65
SLT	-13.23	-19.21

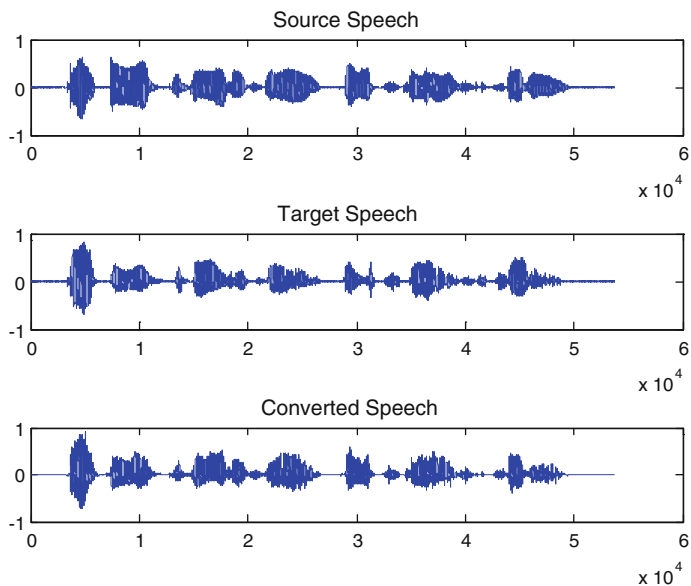


Fig. 3.2 Source, target and converted speech

Table 3.2 Results of subjective listening tests

Subjective test	Voice conversion methods		
	Proposed method	GMM	ANN
Naturalness (MOS)	2.96	2.81	2.77
Individuality (ABX)	0.88	0.74	0.80

extraction and mapping. So we can make conclusion that the method would work with only small size of training data, while GMM and ANN methods does not because of unreliable estimation of GMM parameters due to insufficient data.

The result also shows that although the proposed method is able to implement voice conversion in the framework of style and content separation, and has superior performance. We may conclude that K-SVD has the advantage of preserving the speaker-specific characteristics and taking the inter-frame correlation into account for separation and conversion.

3.5 Conclusion

This paper presents a voice conversion method based on style and content separation, which is solved by K-SVD. And on the basis of successful separating speaker-specific characteristics, the authors invent a novel method for voice

conversion using style replacement. And they also have solved the multiplicity of the decomposition result of K-SVD in the conversion method. Experimental results show that the proposed method for voice conversion has superior performance than traditional GMM and ANN based algorithm in view of both speech quality and conversion similarity and naturalness.

The authors present a novel way to realize voice conversion and improve the conversion performance dramatically which is still a linear combination of style and content while the complex relationship between observed data and style and content is not able to be fully described only by linearity. Therefore, in future studies, researchers will extend the K-SVD to non-linearity to improve the performance of speech style and content separation technique. In addition, the style and content separation technique will immensely promote the development of technique in other speech signal processing domain, such as speech recognition, speaker identification, and low-rate speech coding, etc.

References

1. Stylianou, Y.: Voice transformation: a survey. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3585–3588 (2009)
2. Abe, M., Nakamura, S., Shikano, K., Kuwabara, H.: Voice conversion through vector quantization. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 655–658 (1988)
3. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.* **6**(2), 131–142 (1998)
4. Yamagishi, J., Kobayashi, T., Nakano, Y., et al.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Audio Speech Lang. Process.* **17**(1), 66–83 (2009)
5. Erro, D., Moreno, A., Bonafonte, A.: Voice conversion based on weighted frequency warping. *IEEE Trans. Audio Speech Lang. Process.* **18**(5), 922–931 (2010)
6. Desai, S., Black, A.W., Yegnanarayana, B., et al.: Spectral mapping using artificial neural networks for voice conversion. *IEEE Trans. Audio Speech Lang. Process.* **18**(5), 954–964 (2010)
7. Popa, V., Nurminen, J., Gabbouj M.: A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, pp. 6–10. Brighton, U.K. (2009)
8. Michal, A., Michael, E., Alfred, B.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006)
9. Xu, N., Yang, Z., Zhang, L.H., et al.: Voice conversion based on state-space model for modelling spectral trajectory. *Electron. Lett.* **45**(14), 763–764 (2009)
10. Jian, S., Xiongwei, Z., Tiejong, C. et al. Voice conversion based on convolutive non negative matrix factorization. *Data Collect. Process.* **28**(3), 285–390 (2012)