Multidim Syst Sign Process (2018) 29:385–403
https://doi.org/10.1007/s11045-017-0470-3

# Speech naturalness improvement via $\epsilon$-closed extended vectors sets in voice conversion systems

**Mohammad Javad Jannati**[1] · **Abolghasem Sayadiyan**[1] · **Abolfazl Razi**[2]

Received: 20 February 2016 / Revised: 6 November 2016 / Accepted: 4 January 2017 /
Published online: 12 January 2017

**Abstract** In conventional voice conversion methods, some features of a speech signal's spectrum envelope are first extracted. Then, these features are converted so as to best match a target speaker's speech by designing and using a set of conversions. Ultimately, the spectrum envelope of the target speaker's speech signal is reconstructed from the converted features. The spectrum envelope reconstructed from the converted features usually deviates from its natural form. This aberration from the natural form observed in cases such as over-smoothing, over-fitting, and widening of formants is partially caused by two factors: (1) there is an error in the reconstruction of spectrum envelope from the features, and (2) the set of features extracted from the spectrum envelope of the speech signal is not closed. A method is put forward to improve the naturalness of speech by means of $\epsilon$-closed sets of extended vectors in voice conversion systems. In this approach, $\epsilon$-closed sets to reconstruct the natural spectrum envelope of a signal in the synthesis phase are introduced. The elements of these sets are generated by forming a group of extended vectors of features and applying a quantization scheme on the features of a speech signal. The use of this method in speech synthesis leads to a noticeable reduction of error in spectrum reconstruction from the features. Furthermore, the final spectrum envelope extracted from voice conversions maintains its natural form and, consequently, the problems arising from the deviation of voice from its natural state are resolved. The above method can be generally used as one phase of speech synthesis. It is independent of the voice conversion technique used and its parallel or non-parallel training method, and can be applied to improve the naturalness of the generated speech signal in all

✉ Abolghasem Sayadiyan
sayadiyan@outlook.com

Mohammad Javad Jannati
mjannati@aut.ac.ir

Abolfazl Razi
abolfazl.razi@nau.edu

[1] Department of Electrical Engineering, Amirkabir University of Technology, P.O. Box 15875-4413, 424 Hafez Ave, Tehran, Iran

[2] Department of Electrical Engineering and Computer Science, Northern AZ University, Room 253, Eng. Bldg., 2112 S.Huffer Ln, Flagstaff, AZ, USA

common voice conversion methods. Moreover, this method can be used in other fields of speech processing like texts to speech systems and vocoders to improve the quality of the output signal in the synthesis step.

**Keywords** Voice conversion · Spectrum envelope reconstruction · $\epsilon$-Closed set of features · Speech synthesis

## 1 Introduction

Voice conversion is a branch of speech processing with a large variety of applications. Voice conversion refers to the process of generating the speech signal of a speaker (target speaker) from the speech signal of another speaker (source speaker). The goal of voice conversion is to change the nonlinguistic information of the uttered sentences, e.g., the identity of a speaker, while keeping the linguistic information intact (Toda et al. 2007).

In the available literature on the subject of voice conversion, extensive applications have been cited for voice conversion, ranging from medical and rehabilitation uses to educational, specialized, and even commercial applications (Lee 2007). Voice conversion can be used in dubbing, sound recording, add rhythm and melody to an ordinary speech and to turn it into singing (Sundermann et al. 2006; Saino et al. 2006). Voice conversion can be used in therapeutic procedures. It can be adopted in speech therapy practices and helps those who have lost their larynx or vocal chords as a result of cancer or other diseases (Ghorbandoost et al. 2015). Voice conversion can also be used in pronunciation correction/enhancement for those who want to properly speak a new language (Nakamura et al. 2010). One of the most complicated applications of voice conversion is speech translation from one language to another while preserving the message of the source speaker (Charlier et al. 2009). A specialized application of voice conversion is to increase the volume of data available in the training database (Eide and Picheny 2006).

One of the earliest works in voice conversion is a pioneer work in Childers et al. (1985). In the method presented in that work, the Linear Predictive Coding (LPC) coefficients of the speaker are converted into the LPC coefficients of the new speaker for a corresponding frame. This was a start in this field; however, the performance was poor due to the use of a global transformation function, and the generated speech did not resemble the voice of the target speaker.

Abe et al. (1988) presented the concept of the hard clustering of acoustic space by means of vector quantization. This method used one function for transformation; and because of the discretization of the acoustic space, the transformations generated low-quality voices. Four years later, Valbret et al. (1992) combined the vector quantized voices with linear multivariable regression (LMR) and alleviated the discreteness problem of the target speaker's acoustic space; however, the discreteness problem of the source speaker's acoustic space still remained unsolved.

In 1996, combining the Gaussian mixture model (GMM) and LMR, Stylianou (1996) achieved a great accomplishment in the field of voice conversion due to using soft-clustering instead of hard-clustering. Two years later, Kain and Macon (1998) modified the Stylianou's technique slightly and proposed a method based on Joint Density Gaussian Mixture Models (JDGMM). This method displayed a higher stability relative to the Stylianou's method.

In 2001, by combining the methods of JDGMM and Dynamic Frequency Warping (DFW), a system which performed better than each of these techniques is achieved (Toda et al. 2001).

Toda et al. (2005) made an improvement in the field of voice conversion by combining the JDGMM method and the maximum-likelihood parameter generation. They called their proposed method the Maximum Likelihood Estimation voice conversion (MLE-VC) method. This technique, in addition to using the stationary characteristics, exploits a set of dynamic features to solve the time discontinuity.

Toda et al. (2006) combined the MLE approach with one of the common Principal Component Analysis (PCA) based voice matching techniques called Eigenvoices, and established a system which performs fairly well, even by using only two training sentences from a target speaker. In 2007, Erro combined the methods of JDGMM and Frequency Warping and proposed a method called the Weighted Frequency Warping (WFW) which solves the over-smoothing problem of the JDGMM (Erro and Moreno 2007).

In 2009, the Artificial Neural Network (ANN) as a substitute for the MLE method is used (Desai et al. 2009). The drawback of this approach is its large computational load and its immense complexity. Nexr year, by combining the GMM and the Partial Least Square (PLS) regression, Helander et al. (2010) succeeded in achieving a better performance relevant to the JDGMM by using 10 training sentences.

Helander et al. (2012) combined Radial Basic Functions (RBF) and PLS together with the frame concatenation and came up with the Dynamic Kernel Partial Least Square (DKPLS) method. This is an effective nonlinear method which performs better than the MLE when there is a limited number of training sentences. In the same year, Erro et al. presented the Frequency Warping plus Amplitude Scaling (FW+AS) method as the parametric version of the WFW approach (Erro et al. 2012) and Takashima et al. (2012) put forward a voice conversion method using exemplars. In that method, parallel exemplars (dictionary) consisting of source and target exemplars were introduced, and the source signal was at first decomposed to signal and noise and weights were allocated to each. Voice conversion is then done using the weights of the source signal for target exemplars.

Further working on exemplars in Wu et al. (2013) presented a voice conversion technique based on non-negative spectrogram deconvolution. In that research, Exemplars, which are able to capture temporal context, are employed to generate converted speech spectrogram convolutely. Also, Erro et al. (2013) compared the FW+AS method with the method of MLE plus Global Variance (MLE+GV), and claimed that in most of the cases, these two methods perform similarly.

In 2014, by combining the GMM with the RBF Neural Networks, Chen and Zhang (2014) achieved a performance superior to that of the sole RBF approach, while reducing the volume of data needed for the network. In the same year, by employing the Deep Neural Network (DNN), Chen et al. (2014) presented a method that they claimed performs better than the common GMM-based techniques. Also, Wu et al. (2014) proposed a nonparametric framework exemplar-based sparse representation with residual compensation for voice conversion. In this framework, a spectrogram is reconstructed as a weighted linear combination of speech segments, which span multiple consecutive frames. The linear combination weights are constrained to be sparse to avoid over-smoothing, and high-resolution spectra are employed in the exemplars directly without dimensionality reduction to maintain spectral details. This paper claims to have obtained better results compared with PLS-based methods.

Next year, In 2015, by using the Conditional Restricted Boltzmann Machines (CRBM) based on objective criteria, a better performance than the existing methods based on GMM and ANN is achieved (Nakashika et al. 2015a, b).

In speech processing, the spectrum envelope of the output signal is estimated from spectral features, such as LPC coefficients Mel Frequency Cepstral Coefficients (MFCC), Line Spectral Frequencies (LSF), etc. (Ghorbandoost et al. 2015). These features were already extracted from the spectrum envelope of the original signal. Regardless of what features are used, the representation of speech signal spectrum envelope by a limited number of features involves loss. Whether or not the features are modified, the reconstruction of the spectrum envelope from these features involves error. The amount of this error depends on the type and also the number of features used and the method employed for reconstructing the spectrum envelope from the features. Therefore, the first factor that degrades speech naturalness is the lossy reconstruction of spectrum envelope from the extracted features.

Voice conversion methods (except certain methods based on unit selection that mainly have application in TTS) generally use linear/non-linear combinations on reference vectors at the conversion stage. For example, VQ-based methods at first cluster the vectorial space using different methods such as Linde Buzo Gray (LBG) and K-means and, then, select the reference vectors of the code book using various types of averaging. Additionally, exemplar-based voice conversion methods, which are currently considered to be state-of-the-art voice conversion methods, make use of sparse representation technique, select a number of basis atoms, and represent speech signal using a combination of the aforesaid basis atoms. This averaging and linear combination that may be seen in VQ-based and exemplar-based methods causes the signal spectrum envelope to lose its natural state. In some voice conversion methods, non-linear combinations may be adopted in addition to linear combinations. In view of the potentially nonlinear transformations used as well as the non-linearity imposed by the training and conversion methods, such as the use of neural networks, nonlinear mapping functions, GMM, etc., none of the existing methods is expected to retain the feature set closed with respect to the transformations.

From a mathematical point of view, a set under the effect of an operator is closed when the operator being applied to any member of the set would generate another member of that set. In voice conversion application, if a series of operations are applied to the set of natural speech signals for voice conversion, this set will be closed when a conversion being applied to any natural frame would result in another natural frame. If this speech database set remains closed to voice conversion operations, the voice conversion output will in fact remain natural, and the voice conversion operations does not cause the signal to lose its natural state. In practice, it may be observed by examining the existing voice conversion methods that the set of reference vectors adopted in these methods are not closed to voice conversion operations, and they lose their natural state once voice conversions are applied.

Therefore, the converted output signal has a distinguishable deviation from the natural signal which leads to cases such as over-smoothing, over-fitting, and widening of formants. Thus, the output of a voice conversion system consists of a set of features the spectrum envelope extracted from which results in the reconstruction of a synthetic, and not a natural, voice signal. This is the second factor that degrades speech naturalness. These two spots where degradation is created are, of course, to some extent related. Even though the degradation caused by loss-producing representation of spectrum envelope from features has a vaster scope and is not limited to voice conversion, the degradation resulting from conversions is exclusive to voice conversion or places where the features are modified between analysis and synthesis stages. As mentioned earlier, many attempts were made to bring this synthesized speech closer to a natural voice.

To mitigate the impacts of the two aforementioned problems in speech synthesis and conversion, Speech Naturalness Improvement via $\epsilon$-Closed sets of Extended vectors (NICE) technique is put forward in this research as a speech synthesis method and also a complement method to voice conversion methods. By providing its own algorithm, NICE forms perceptually (not mathematically) closed sets from basic vectors of features called $\epsilon$-closed sets and saves such information as spectrum envelope, pitch frequency, gain, and other natural information of signal. It then presents another algorithm to represent speech signal in a manner that the signal spectrum envelope would remain intact and natural. NICE method can be applied to the results of other methods at the stage prior to synthesis, leading to the naturalness of the spectrum envelope of the synthesized signals.

Further in Sect. 2, the scientific and mathematical foundation of NICE method is explained in detail. The subsequent section is allocated to the implementation of NICE technique, where NICE efficiency is initially investigated in natural spectrum envelope representation in the synthesis stage. To do so, the spectrum envelope represented by NICE is compared with that represented by features such as LSF and MFCC. On the other hand, to evaluate NICE efficiency in voice conversion, NICE efficiency is studied as a complement method in voice conversion by applying NICE to two standard voice conversion methods. The fourth and last section of the research presents the conclusion.

## 2 Speech naturalness improvement

By presenting a series of $\epsilon$-closed sets of extended vectors for the reconstruction of the spectrum from features, the method proposed in this research, titled "Speech Naturalness Improvement via $\epsilon$-Closed sets of Extended vectors (NICE)," aims at reconstructing a speech signal's natural spectrum envelope and reducing reconstruction error. NICE method first defines a number of extended vectors for each speaker. Then, by presenting a quantization algorithm, this technique produces a reference $\epsilon$-closed set of these extended vectors for each speaker. Finally, it reconstructs the spectrum envelope from the features using the reference $\epsilon$-closed sets produced from the extended vectors.

In the following section, the $\epsilon$-closed set is defined at first. Then, the extended vector is defined and the way it is generated is explained. After that, the algorithm for extracting the $\epsilon$-closed set from the reference extended vectors presented by NICE is introduced. Finally, the method of reconstructing a spectrum envelope via NICE is explored.

### 2.1 Defining the $\epsilon$-closed set

Mathematically, set $A$:$\{a_1, a_2, \ldots, a_n\}$ is considered to be closed with respect to the operator group $F$, if for $a_i$ being an arbitrary member of $A$, $F(a_i)$ is also a member of $A$ (Simovici and Djeraba 2014). In other words, if any member of $A$ is operated on by the operator group $F$, again a member of $A$ will be obtained. Thus, $A$ will be a closed set, if and only if

$$if\{a_i \in A\} \rightarrow \{F(a_i) \in A\} \tag{1}$$

On the other hand, if $a_k$ and $a_l$ are two arbitrary members of Set $A$, the closed segment specified by these two members will be defined as follows (Simovici and Djeraba 2014):

$$Closed\ Segment\ of\,(a_l, a_k) = \{\alpha a_l + (1-\alpha)a_k | 0 \leq \alpha \leq 1\} \tag{2}$$

If for any arbitrary $a_k$ and $a_l$ from Set $A$, their associated closed segment is also a member of $A$, Set $A$ is called a convex set (Simovici and Djeraba 2014).

Likewise, in this research, the $\epsilon$-closed set is defined as follows. Set $A$:$\{a_1, a_2, \ldots, a_n\}$ will be an $\epsilon$-closed set with respect to operator F, if for $a_i$ being an arbitrary member of A, there exists a member of $A$ whose distance from $F(a_i\ )$ is less than a predefined threshold $\epsilon$. In other words, $F(a_i\ )$ may not be a member of $A$ and may not lie within $A$; but $F(a_i\ )$ lies at a maximum distance of $\epsilon$ from $A$ (a member of $A$). Formally, Set A is $\epsilon$-closed with respect to operator $F$, if

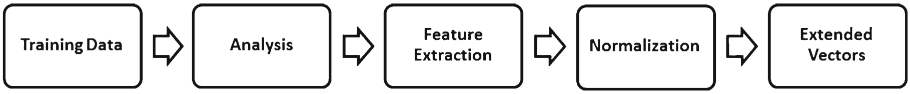

**Fig. 1** Steps of generating the extended vectors for the training database

$$if\{a_i \in A\} \rightarrow \{\exists F(a_j) \in A : |F(a_i) - a_j| \le \epsilon\} \tag{3}$$

Also, if $a_k$ and $a_l$ are two arbitrary members of Set $A$, $A$ will be an $\epsilon$-convex set, if

$$if\{a_l, a_k \in A\} \rightarrow \{\exists F(a_j) \in A : |\alpha a_l + (1 - \alpha)a_k - a_j| \le \epsilon\}. \tag{4}$$

## 2.2 Generating the extended vectors

Provided that there are sufficient speech signals available from each of the speakers of the training database. The first step is to form the extended vectors employing one of the existing speech analysis methods, as depicted in Fig. 1. Different methods such as Sinusoidal Model (SM), Harmonic Stochastic Model (HSM), Harmonic Noise Model (HNM), and Fixed Dimension Modified Sinusoidal Model (FDMSM) were reported for analysis, spectrum envelope extraction, and synthesis (Mowlaee et al. 2009). In this study, the spectrum envelope was generated by means of Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) method (Kawahara et al. 1999). In the analysis phase, for each frame, pitch frequency (F0), spectrum envelope (Spec), aperiodicity (AP) coefficients and speech/non-speech flag are determined and recorded. In the features extraction phase, several features representing the spectrum envelope are extracted from the spectrum envelope for each frame. A total of 24 LSF coefficients were extracted from the spectrum envelope for each frame. Note that, based on the type of features used, each element of a feature vector may have a different dynamic range. For example, if the sampling frequency of the speech signal is 16000 Hz and the spectrum envelope of each speech frame is represented by 24 LSF coefficients, then the 1st LSF coefficient will have a range of values less than 400 Hz and the 24th LSF coefficient will have a range of values between 7500 and 8000 Hz.

Before any modification, it is necessary to normalize the ranges of coefficients. For this purpose, the mean ($\mu$) and standard deviation ($\sigma$) values of the speech frame coefficients are at first computed for the entire database and for the 1st through 24th the coefficient of LSF. Then, by applying the following formula for all database frames, all the feature vector values are normalized.

$$NLSF_i = (LSF_i - \mu_i)/\sigma_i \tag{5}$$

For each frame, NLSF$_i$ is the ith normalized LSF coefficient, $LSF_i$ is the ith coefficient of $LSF$, $\mu_i$ is the mean value of the ith LSF coefficient, and $\sigma_i$ is the standard deviation of the ith coefficient of LSF for all speech frames of the database. By applying expression 5, the dynamic ranges of all coefficients become identical. Following the normalization procedure, an extended vector is formed for each frame which consists of a number of scalars and a number of vectors. The names and descriptions of the sub-parts of an extended vector are presented in Table 1. In view of Table 1, each frame of the database is represented by an extended vector that has 7 subparts and includes a total of 2076 parameters.

**Table 1** Names and descriptions of the sub-parts of the extended vector associated with an arbitrary frame

| Sub part | Description | Number of parameters | Form and size |
|---|---|---|---|
| F0 | Pitch period | 1 | Scalar-1 double |
| S/NS | Speech or nonspeech | 1 | Scalar-1 bit |
| V/UV | Voiced unvoiced | 1 | Scalar-1 bit |
| G | Frame gain | 1 | Scalar-1 double |
| NLSF | Normalized LSFs | 24 | Vector-24 double |
| Spec | Spectrum envelope | 1024 | Vector-1024 double |
| Ap | Aperiodicity coefficients | 1024 | Vector-1024 double |

## 2.3 The proposed NICE method

NICE method comprises two algorithms. The first one is for choosing an $\epsilon$-closed set of extended vectors for an arbitrary speaker from all his/her extended vectors available in the database. The second one is for synthesis (reconstructing spectrum envelope) of the speech signal of the same speaker.

To form the $\epsilon$-closed set of extended vectors for a speaker, $M$ reference extended vectors should be selected from the extended vectors available in the database for that speaker ($M$ is not the same for all speakers and may vary). Suppose that the database contains $N$ frames and, thus, $N$ extended vectors from this speaker, sequentially designated as $EV_1$, ..., $EV_N$. Each $EV_i$ is an extended vector with the following vector structure.

$$EV_i : \left\{ \underbrace{\{NLSF\}}_{First\ Part\ of\ Extended\ Vector}, \underbrace{\{F0\}, \{S/NS\}, \{V/UV\}, \{G\}, \{Spec\}, \{Ap\}}_{Second\ Patr\ of\ Extended\ Vector} \right\} \quad (6)$$

Each extended vector (EV) comprises two parts. The first part contains the features extracted from spectrum envelope of the ith frame (NLSF in this research). The second part includes the natural parameters of analysis, especially the spectrum envelope parameters of real speech (parameters necessary for synthesis). In generating the set of extended vectors (first algorithm), only the first part of EV is used for selecting the reference extended vectors among all sample extended vectors of a speaker. In this process, no alteration or transformation is applied to the second part and the reference extended vectors are selected according to the first part of the EV. The second part of the EV remains intact in all steps of the algorithm and will be used only in synthesis (second algorithm).

In what follows, the algorithm employed to extract $M$ reference extended vectors associated with an arbitrary speaker is explained ($M$ is a parameter and its value will be chosen in the final step of the algorithm). Then, NICE algorithm for reconstructing the spectrum envelope is described.

### *2.3.1 NICE algorithm for extracting an $\epsilon$-closed set*

Before starting the algorithm, the extended vectors relevant to the speech frames are separated. Next, depending on whether these extended vectors are voiced or unvoiced, they are divided into two groups. The following NICE algorithm is then independently and separately executed on the voiced and unvoiced groups of extended vectors. The NICE algorithm has three steps for extracting an $\epsilon$-closed set from the reference extended vectors as follows in Algorithm 1.

**Algorithm 1** : NICE algorithm for extracting an $\epsilon$-closed set from the reference extended vectors

```
Initialization step ($i = 1$)
  – $EV_i$ ($EV_1$) is selected as first reference EV.
  – The variable for counting the number of repetitions of first reference EV ($EV\_No_1$), is set to 1.
  – The variable for the total number of reference EVs ($Total\_EV$), is set to 1.
Recursion step ($i$ from 2 to $N$)
  • The distances of the $NLSF_i$ (related to $EV_i$), from every single NLSF of the previous EVs that have
    been chosen as reference EVs are calculated (by Expression 7) .
  • If at least one or more than one of these distances is less than the threshold value $\delta$,
    $EV_i$ will be assigned to the first reference EV whose distance from $EV_i$ is less than $\delta$.
    The number of repetitions of that reference EV will be raised by one (1).
  • If all of these distances are more than the threshold value $\delta$,
    The $Total\_EV$ count will be raised by one (1).
    $EV_i$ will be selected as the $Total\_EV$th reference EV.
    The variable for counting the number of repetitions of that reference EV will be set to 1.
Final step
  – The reference EVs are separated.
  – Every reference EV whose repetition is less than 2 is omitted (as outliers).
  – The $M$ remaining reference EVs are chosen as members of the $\epsilon$-closed set of EVs.
```

The distance used in Algorithm 1 can be defined according to the type of features used. To determine the distance between the feature vectors, different approaches, such as the Euclidian distance and perceptual methods (Doost et al. 2009) were introduced. Since NLSF coefficients are used in this research, the distance between the two extended vectors $EV_k$ and $EV_j$ is computed from the following relation (Linde et al. 1980):

$$Dist(EV_k, EV_j) = \sum_{i=1}^{K} |\omega_i (NLSF_k(i) - NLSF_j(i))|^2 \tag{7}$$

In the above equation, $NLSF_k$ are the feature coefficients of the extended vector $EV_k$, $NLSF_j$ are the feature coefficients of the extended vector $EV_j$, NLSF(i) is the ith member of the feature vectors, and K is the number of features (24 in this research). $\omega_i$ is a weight factor. $\omega_i$ was considered as the normalized mel-scaled mean of LSF(i) in total database.

As mentioned in expression 7, the distance between the two EVs is equal to the distance between their NLSFs (the first part of the EV). By using the above algorithm, a set of extended vectors can be generated for any arbitrary database. The way the natural spectrum envelope is extracted from the features by means of such a set of extended features will be described in the following subsection.

### *2.3.2 NICE algorithm for reconstructing the spectrum envelope*

To reconstruct the spectrum envelope from the features using NICE method, the following procedure (Algorithm 2) is implemented.

Any voice conversion and representation technique utilizes a different concatenation method. This concatenation method is applied to the results improved by NICE instead

**Algorithm 2** : NICE algorithm for reconstructing the spectrum envelope

**Reconstruction process**

- The features of the considered frame (NLSF in this research) are obtained by any arbitrary method.
- For each Frame, it is determined whether the frame is voiced or un-voiced; and accordingly, an appropriate extended vectors set is used.
- By using Expression 7, the distance of each frame from all reference EVs of the $\epsilon$-closed set of EVs obtained in the previous section, is calculated.
- The reference EV with the least distance ($\delta_{min}$) from the each frame is selected.
- For each frame, the spectrum envelope of the selected EV (second part of EV) is chosen as the spectrum envelope of the frame.

of the main frames. NICE aids signal representation methods to replace frequency content that is out of normal conditions as a result of conversions by natural frequency content. It does not cause any change to the process and stages of conversion or representation including concatenation. NICE complies with the method to which it is applied for improving results. NICE technique is independent from the type of synthesis adopted. It operates regardless of the method used for frame reconstruction and concatenation. That is, whether single-frame or piecewise methods are employed, NICE extracts and replaces the basic vectors commensurate with each reconstructed frame. In other words, once the signal is represented via an arbitrary method whether in piecewise or in single-frame form, NICE method is applied to the results, leading to improved quality and increased naturalness of the signal. Of course, regarding the methods that conduct signal representation in piecewise form, basic vectors of NICE technique should also be designed to be piecewise [Matrix Quantization (MQ) is adopted instead of VQ]. As a result, NICE method is a complement method to the existing signal representation techniques.

## 3 Implementation and results

Since NICE algorithm only uses the spectrum envelopes of the reference extended vectors for synthesis in reconstructing the spectrum envelopes, the possible output spectrum envelopes are limited to the spectrum envelopes of the extended vectors generated by NICE. This leads to two conclusions: (1) since the spectrum envelopes of the extended vectors are unaltered and natural, the output spectrum envelopes will also be unaltered and natural and, (2) regardless of what signal, or what type of features is synthesized, all output spectrum envelopes belong to an $\epsilon$-closed set of the extended vectors and other than the set of spectrum envelopes related to extended vectors, nothing else will be present in the output. By applying this method, a spectrum envelope in the output will be obtained which is not only completely natural but is also a definite member of the set of spectrum envelopes belonging to the reference extended vectors set. Hence, no spectrum envelope outside this set will exist in the output. Thus, the set of spectrum envelopes will remain intact and unaltered. The consequence of this improvement in naturalness and intactness is certain costs. The first cost is the need of more data storage capacity for saving EVs instead of spectral features (in this research about 80 times). Although this can be mentioned as a cost, concerning the large capacity of todays hard disk drives, it is not a serious problem. The second one is the computational cost to extract the $\epsilon$-closed sets of EVs. As mentioned in previous parts, this operation is offline and it must be done only once for each database. Therefore, it can be ignored. The third cost of the improvement in naturalness is the introduction of error in reconstruction. The error is generated by selecting the nearest extended vector to the frame being synthesized (and not the frame itself) in the selection phase (step 4 of Algorithm 2) and using its spectrum envelope instead of the converted spectrum envelope. The longer this nearest distance ($\delta_{min}$) is, the higher the reconstruction error of that frame will be. The value of $\delta_{min}$ is inversely related to the degree of covering of the extended vectors set. This means that as the extent of covering of the existing spectrum envelopes set in the reference extended vectors set increases, $\delta_{min}$ distance gets shorter and, thus the spectrum envelope reconstruction error diminishes. On the other hand, the degree of covering of the existing spectrum envelopes in the reference extended vectors set is directly related to the variety and number of the speakers and the extent and diversity of the existing voices in the database. Hence, as mentioned earlier, as the number and variety of speakers and the extent and diversity of the voices recorded in the database increase, the error in the reconstruction of spectrum envelope diminishes. In the following sections, the adjustment procedure for the main parameters in NICE along with their description, roles, and impacts is discussed, then the test setup is described. Afterward, the adopted database, test setup and results are presented. The spectrum envelope reconstruction results obtained by NICE approach are compared with those of other methods. Finally, the effect of NICE method on the converted voice, in conjunction with the application of two common and standard voice conversion techniques will be investigated.

### 3.1 Selecting the values of $\epsilon$ and $\delta$

Paliwal and Atal (1993) has claimed that if the average Log-Spectral distortion error (LSD) between the frame spectrum envelopes of two speech signals is about 1 dB, human ear recognizes these two voices to be acceptably similar. The common criterion of LSD for a comparison between the original and the reconstructed spectrums in the i-th frame (in dB) is defined as (Ramachandran and Mammone 1995; Naylor and Gaubitch 2010)

$$LSD(i) = \sqrt{\frac{1}{K}\sum_{j=1}^{K} |20\log(S_r(i,j)) - 20\log(S_c(i,j))|^2} \tag{8}$$

where $S_r$ is the amplitude of the original spectrum envelope, $S_c$ is the amplitude of the spectrum envelope reconstructed from the feature, and K is the number of Fast Fourier Transform (FFT) points. Paliwal and Atal (1993) reported the maximum acceptable average value of LSD in a speech signal to be about 1 dB. Of course, the lower this value is, the fewer errors will occur; however, for an average LSD value of about 1 dB, the resulting error (quality of reconstructed voice) will be acceptable.

Of course, for an LSD value lower than 1 dB, the mismatch error is even lower, but for an average LSD of about 1 dB, the number of errors will be acceptable. Therefore, if the set of extended vectors produced by the proposed method is $\epsilon$-closed and if the amount of $\epsilon$ is about 1 dB, every arbitrary vector from this set will have a maximum distance of $\epsilon$ from one of the reference vectors and, consequently, its difference with the mentioned reference vector can be ignored. As observed in the presented algorithm, $\delta$ is the threshold distance of an ordinary extended vector from the reference extended vectors. The amount of $\delta$ must be selected in a way that the set would remain $\epsilon$-closed and, therefore, $\delta$ must be chosen such that the average LSD value would not exceed 1 dB. Considering the abovementioned points, for every speaker, the value of $\delta$ must be chosen in a way that the average LSD would be equal to 1 dB, and the set of extended vectors for every speaker, for an $\epsilon = 1$ dB, would remain $\epsilon$-closed. Two values of 0.05 and 0.001 for $\delta$ have been examined in Linde et al. (1980). Of course, regarding the fact that in the coding process, the number of bits used for

**Table 2** Mean ($\mu$) and standard deviation ($\sigma$) of LSD (in dB) obtained for $\delta$ values of 0.001, 0.05 and 0.01 using 50 and 75 training sentences

| Speaker | 50 Train sentences | | | | | | 75 Train sentences | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\delta = 0.001$ | | $\delta = 0.05$ | | $\delta = 0.01$ | | $\delta = 0.001$ | | $\delta = 0.05$ | | $\delta = 0.01$ | |
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| BDL | 1.81 | 0.31 | 1.93 | 0.41 | 2.58 | 0.51 | 1.19 | 0.09 | 1.21 | 0.11 | 1.31 | 0.12 |
| RMS | 1.63 | 0.33 | 1.68 | 0.47 | 2.31 | 0.55 | 1.14 | 0.11 | 1.18 | 0.12 | 1.33 | 0.13 |
| CLB | 1.66 | 0.32 | 1.78 | 0.4 | 2.25 | 0.49 | 1.15 | 0.08 | 1.19 | 0.10 | 1.2 | 0.11 |
| SLT | 1.81 | 0.37 | 2.00 | 0.44 | 2.61 | 0.52 | 1.20 | 0.10 | 1.21 | 0.11 | 1.25 | 0.12 |

**Table 3** Mean ($\mu$) and standard deviation ($\sigma$) of LSD (in dB) obtained for $\delta$ values of 0.001, 0.05 and 0.01 using 100 training sentences

| Speaker | 100 Train sentences | | | | | |
|---|---|---|---|---|---|---|
| | $\delta = 0.001$ | | $\delta = 0.05$ | | $\delta = 0.01$ | |
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| BDL | 0.89 | 0.08 | 0.93 | 0.09 | 1.10 | 0.10 |
| RMS | 0.82 | 0.08 | 0.88 | 0.08 | 1.12 | 0.11 |
| CLB | 0.81 | 0.07 | 0.88 | 0.08 | 0.98 | 0.10 |
| SLT | 0.90 | 0.09 | 0.96 | 0.10 | 1.01 | 0.09 |

each coded symbol is considered as a limitation, in Paliwal and Atal (1993), Linde et al. (1980), and other similar papers in the context of coding and data compression, limits were placed on the number of bits used for each symbol. This issue influences the effective value of $\delta$ used. Since the aim of the present research is to improve the speech synthesis quality through spectral features and voice conversion, this limitation regarding the number of bits used in the coding or compression of symbols is not necessary for the applications of this work; therefore, no limitation was imposed on $\delta$. Hence, the voices of four speakers from the database were analyzed for three $\delta$ values of 0.01, 0.05, and 0.001, without considering any constraints on the number of bits (in Sect. 3.2 the databese is introduced). Using 50, 75, and 100 training sentences, NICE algorithm was applied for each speaker and each of the above values for generating an $\epsilon$-closed set of extended vectors. Then, 20 sentences which, were not applied in the training were used in the test phase.

Tables 2 and 3 show the obtained results for each speaker and for different $\delta$ values. As can be seen, with an increase in the number of training sentences which leads to an increase in the number and variety of training data, the average value of LSD diminishes. The rate of this reduction is higher when the training sentences increase from 50 to 75 than when they increase from 75 to 100 sentences. This means that the rate of reduction of LSD declines as the number of training sentences grows. For 100 training sentences and $\delta = 0.01$, the mean LSD of the test sentences for all four speakers is about 1 dB. Therefore, considering the fact that more than 100 sentences are available from each speaker, in the case of using 100 or more training sentences, for $\delta = 0.01$, NICE algorithm can present an $\epsilon$-closed set of extended vectors for each speaker whose total LSD in the test phase will be about 1 dB. Thus, a $\delta$ value of 0.01 was used in this research.

Table 4 shows the results of the $\epsilon$-convex test for the set of reference extended vectors for each speaker. In this test, for the 5 values of $\alpha$, all possible double combinations of the members of the extended vectors set for each speaker were selected, and the closed segments

**Table 4** Mean ($\mu$) and standard deviation ($\sigma$) of LSD (in dB) for $\delta = 0.01$ and for $\alpha$ values ranging from 0.1 to 0.5, for all dual combinations of reference vectors for each speaker

| Speaker | $\alpha = 0.1$ | | $\alpha = 0.2$ | | $\alpha = 0.3$ | | $\alpha = 0.4$ | | $\alpha = 0.5$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| BDL | 1.11 | 0.12 | 1.01 | 0.11 | 0.99 | 0.09 | 1.12 | 0.11 | 0.98 | 0.10 |
| RMS | 1.15 | 0.16 | 1.14 | 0.13 | 0.96 | 0.09 | 1.07 | 0.11 | 0.99 | 0.09 |
| CLB | 0.98 | 0.11 | 1.03 | 0.09 | 0.98 | 0.10 | 1.07 | 0.09 | 1.03 | 0.09 |
| SLT | 0.97 | 0.09 | 0.97 | 0.08 | 1.13 | 0.11 | 1.00 | 0.10 | 1.02 | 0.10 |

produced by these combinations ($\alpha a_l + (1 - \alpha)a_k$) were obtained. Then the distance of the obtained expression to the nearest reference vector was determined as per LSD criterion. The mean and standard deviation values obtained for all double combinations are listed in Table 4 for different speakers and $\alpha$ values. As can be observed, the average LSD values are about 1 dB which are acceptable considering the criterion presented in Paliwal and Atal (1993). Thus, for $\delta = 0.01$, the sets of reference extended vectors obtained for each speaker are $\epsilon$-closed and $\epsilon$-convex, with regard to expressions 3 and 4, respectively.

## 3.2 Database, tests and results

To evaluate the effect of NICE algorithm on the results of the two methods mentioned in the previous section, the objective test of Perceptual Evaluation of Speech Quality (PESQ) (Hu and Loizou 2008), the subjective test of Mean opinion score (MOS) (Streijl et al. 2016), and also the preference test were used. The STRAIGHT method was adopted for analysis and synthesis of speech with a frame shift of 5 ms. A total of 24 features of LSF and MFCC were chosen for each frame. To produce the extended vectors, the data of CMU ARCTIC database established by Carnegie Mellon University's speech group were used, at a sampling frequency of 16 kHz. A full description of this database was given in Kominek and Black (2004). To train the conversions, the voices of two male speakers (BDL and RMS) and two female speakers (CLB and SLT) from the same database were used. For the tests, 10 random sentences were selected, and the tests were performed according to these sentences. The objective test of PESQ and the subjective test of MOS were adopted for comparing the qualities of the generated speech signals. In PESQ test, the method and the code presented in Hu and Loizou (2008) were used. A total of 8 native speakers were employed for MOS test in order to determine the quality of the produced voices. The preference test was also used for checking the naturalness of the generated voices.

### *3.2.1 Results of spectrum envelope reconstruction by NICE method*

It was previously mentioned that the process of reconstructing a speech signal's spectrum envelope from the features would involve error. For a comprehensive comparison of the reconstruction errors in the direct reconstruction of spectrum envelope by MFCC and LSF with the reconstruction error obtained in NICE approach, 10 sentences were randomly selected. These sentences were analyzed and the spectral features were extracted from their spectrum envelopes. To compare the qualities of the synthetic signals, without doing any conversion on them, the speech spectrum envelopes were extracted from these features. By performing PESQ and MOS tests, the qualities of the speeches synthesized by MFCC, LSF, and NICE

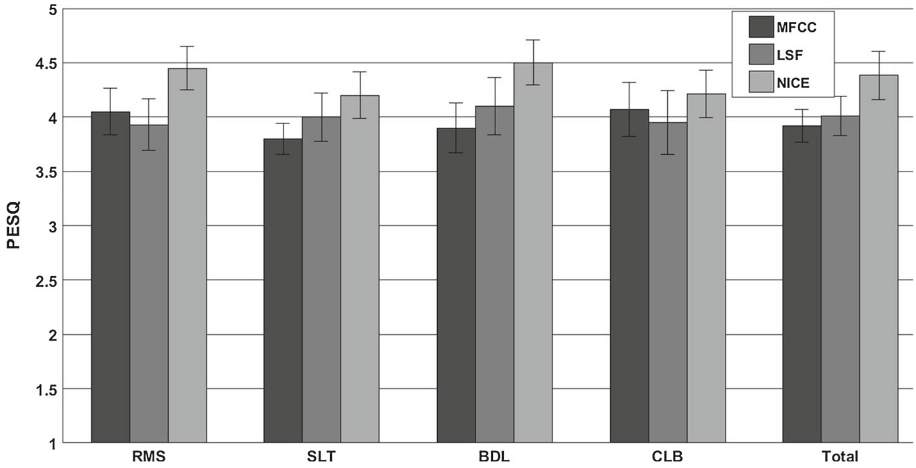

**Fig. 2** Results of PESQ test for the reconstruction of spectrum envelopes from features directly and also using NICE method

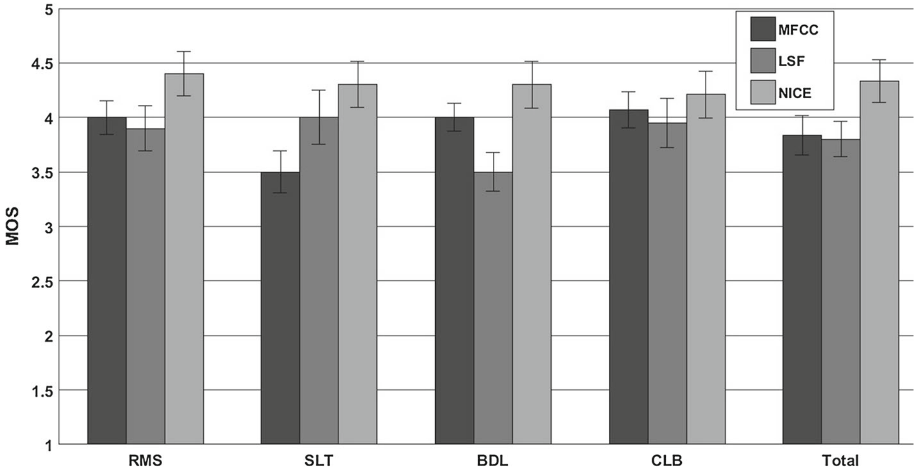

**Fig. 3** Results of the MOS test for the reconstruction of spectrum envelope from features directly and also by using the NICE method

methods were compared with each other. Figures 2 and 3 illustrate the results of PESQ and MOS tests, respectively, and the results of these two tests match one another and indicate the superiority of NICE method. As can be seen, NICE method can reconstruct the natural spectrum envelope from arbitrary features and synthesize the speech signal. This shows the quantitative superiority of NICE approach. Of course, producing a natural and intact spectrum envelope is the other special advantage of NICE method. This characteristic does not exist in MFCC and LSF, and it is impossible for these two methods to achieve such a quality. Hence, NICE approach was able to reduce the spectrum envelope reconstruction error and it could successfully reduce the effects of the first cause of speech unnaturalness (i.e., the existence of an error in reconstructing a signal's spectrum envelope). Therefore, NICE can be used in other fields of speech processing like the text to speech systems (TTS) and vocoders for improving the quality and naturalness of the generated speech signal in the synthesis step, and its application is not confined to voice conversion.

### *3.2.2 Converted voice quality and naturalness improvement by NICE method*

As stated earlier, the second most important cause of making a voice signal unnatural in the voice conversion systems is the non-closure of the features space with respect to voice transformations. This means that the applied transformations may alter the features in a way that the spectrum envelope extracted from them does not exist in the natural voice signal and the generated voice clearly sounds synthetic. This problem manifests itself in the shifting and widening of formants and in the over-smoothing and over-fitting of the generated voice signal. Since NICE method preserves a signal's natural spectrum envelope, it is not hampered by the above-mentioned problems and is a suitable choice for generating a signal's spectrum envelope from the converted features in a way that the closest natural spectrum envelope is selected as the output. As Sect. 3.2.1 indicates, in NICE algorithm, the estimation of spectrum envelope from features is independent of the way the features are produced or estimated and, thus, the synthesized speech signal is independent of the voice conversion method used. Hence, by applying NICE on two known and common voice conversion methods, i.e., JDGMM (Kain and Macon 1998) and DKPLS (Helander et al. 2010), the performances of these two techniques solely and in conjunction with NICE method, are compared. To observe the effect of NICE method on the quality of the generated voices, all possible voice transformations among the speakers were taken into consideration. Hence, the voice of each speaker was converted into the voices of the other three speakers and, consequently, 12 conversions were made among the speakers. Using MOS test and the preference test, the voice signals generated with and without the use of NICE approach were compared with the standard methods of JDGMM and DKPLS. The results of MOS test for 4 groups of transformations [male-to-male (M_M), male-to-female (M_F), female-to-male (F_M) and female-to-male (F_F)] are illustrated in Fig. 4. As can be seen, the performances of both standard methods have improved. Figure 5 shows the results of the naturalness preference test for JDGMM by itself and JDGMM + NICE. According to this figure, the naturalness

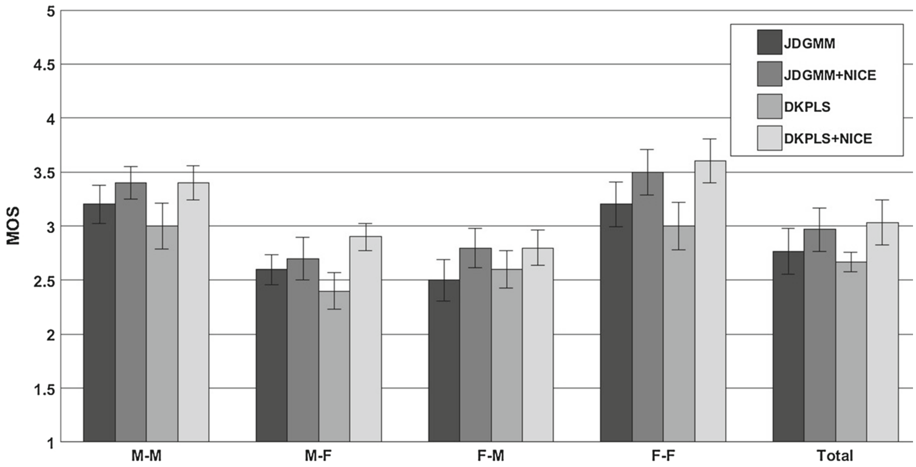

**Fig. 4** Comparing the results of MOS test regarding the performances of JDGMM and DKPLS methods with and without the application of NICE

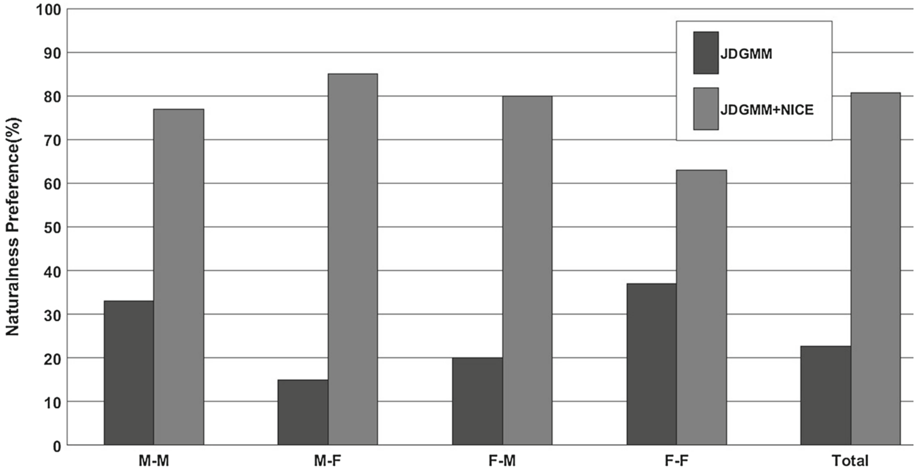

**Fig. 5** Comparing the naturalness qualities of the voices converted by JDGMM method with and without the application of NICE

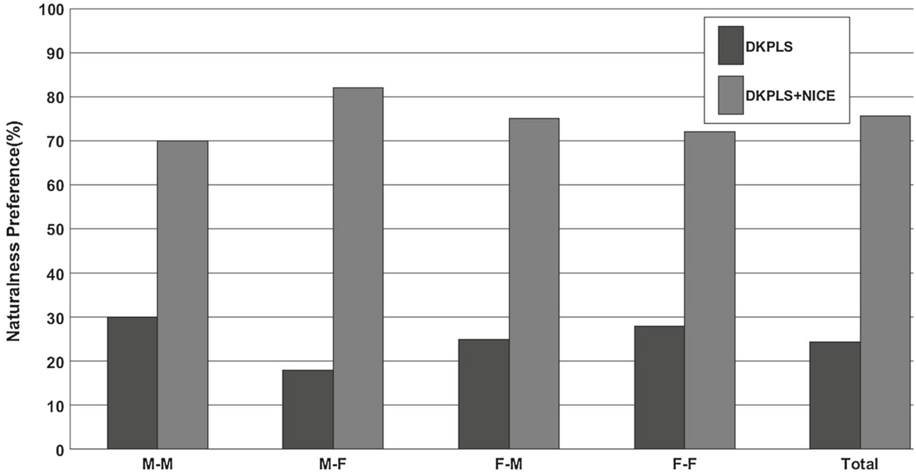

**Fig. 6** Comparing the naturalness qualities of the voices converted by DKPLS method with and without the application of NICE

of the converted voice has improved. Figure 6 illustrates the results of the same test for sole DKPLS and DKPLS + NICE. Again, the converted voice displays an improvement in terms of naturalness. Because of the phase variations resulting from voice conversion in most of the voice conversion techniques, PESQ criterion is not usually used for comparing the qualities of converted voices. In the comparisons between converted voices, PESQ approach does not yield a reliable answer, and therefore it is not regularly used in the papers dealing with voice conversion. Thus, in evaluating the effect of NICE on the two standard voice conversion methods, PESQ criterion was not used. As the results indicate, the application of NICE in conjunction with JDGMM and DKPLS methods has had a positive effect on the improvement of voice quality, especially the naturalness of voice. The reason is the naturalness of the spectrum envelope generated by NICE approach. This means that by using real spectrum envelopes, NICE produces higher quality voices that have a natural spectrum envelope. Since NICE is not dependent on the type of transformation used, it can be effectively used in combination with other voice conversion methods.

## 4 Conclusion

In common voice conversion systems, the generated voices normally deviate from their natural form, and this unnaturalness can be detected by a listener. The unnaturalness of voice manifests itself in cases such as the shifting and widening of formants, and in the over-smoothing and over-fitting of the model used. The underlying causes include the existence of an error in the reconstruction of spectrum envelopes from features and also the non-closure of the spectral features set with respect to the applied voice transformations. These two factors cause the spectrum envelope of a speech signal to deviate from its natural form and, thus, reduce the naturalness of the voice. NICE method was introduced to solve these two problems by producing $\epsilon$-closed sets of extended vectors. The results obtained from the objective test of PESQ and subjective test of MOS confirm that this method, by forming $\epsilon$-closed sets of extended vectors for speakers and applying the presented algorithm as well as the given search algorithm, is able to improve the quality of the synthesized speech and reduce the spectrum envelope reconstruction error in comparison with the methods of direct envelope reconstruction from feature vectors in MFCC and LSF. Moreover, MOS test indicated that NICE approach improves the naturalness of the converted voices for the two standard voice conversion methods of JDGMM and DKPLS. The reason is the use of the extended vectors set for the reconstruction of spectrum envelope and the use of a signal's natural spectrum envelope instead of the converted spectrum envelope in the synthesis phase. The results of the preference test on the voice signals converted by JDGMM and DKPLS methods with and without the application of NICE also indicated that the use of NICE in conjunction with the mentioned methods improves the naturalness of the generated voices. NICE approach is independent of the type of voice conversion method used and, therefore, by incorporating it during the reconstruction of spectrum envelopes from the converted features, regardless of the conversion technique used, the naturalness of the converted voice can be improved. NICE approach is not dependent on the method of voice conversion and whether it is parallel or non-parallel. Moreover, NICE can be used in other fields of speech processing like the text to speech systems and vocoders to improve the quality and naturalness of the generated speech signal in the synthesis step and its usage is not limited to voice conversion.

**Acknowledgements** The authors would like to thank the reviewers for their relevant and constructive remarks, which have enabled us to significantly improve the quality of the paper.

## References

Abe, M., Nakamura, S., Shikano, K., & Kuwabara, H. (1988). Voice conversion through vector quantization. In *International conference on acoustics, speech, and signal processing, ICASSP-88* (Vol. 1, pp. 655–658).

Charlier, M., Ohtani, Y., Toda, T., Moinet, A., & Dutoit, T. (2009). Cross-language voice conversion based on eigenvoices. In *10th Annual conference of the international speech communication association*, Brighton, UK, September 6–10, pp. 1635–1638.

Chen, L., Ling, Z., Liu, L., & Dai, L. (2014). Voice conversion using deep neural networks with layer-wise generative training. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1859–1872.

Chen, X., & Zhang, L. (2014). High-quality voice conversion system based on GMM statistical parameters and RBF neural network. *The Journal of China Universities of Posts and Telecommunications*, *21*(5), 68–75.

Childers, D., Yegnanarayana, B., & Wu, K. (1985) Voice conversion: Factors responsible for quality. In *IEEE international conference on acoustics, speech, and signal processing, ICASSP '85* (Vol. 10, pp. 748–751).

Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., & Prahallad, K. (2009). Voice conversion using artificial neural networks. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 3893–3896).

Doost, R., Sayadiyan, A., & Shamsi, H. (2009). A new perceptually weighted distance measure for vector quantization of the STFT amplitudes in the speech application. *IEICE Electronics Express*, *6*(12), 824–830.

Eide, E., & Picheny, M. (2006). Towards pooled-speaker concatenative text-to-speech. In *IEEE International conference on acoustics, speech and signal processing, ICASSP '06* (Vol. 1, pp. 73–76).

Erro, D., & Moreno, A. (2007). Weighted frequency warping for voice conversion. In *Annual conference of the international speech communication association, InterSpeech '07*.

Erro, D., Navas, E., & Hernáez, I. (2012). Iterative MMSE estimation of vocal tract length normalization factors for voice transformation. In *13th Annual conference of the international speech communication association, INTERSPEECH '12*, Portland, Oregon, USA, September 9–13, pp. 86–89.

Erro, D., Navas, E., & Hernaez, I. (2013). Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(3), 556–566.

Ghorbandoost, M., Sayadiyan, A., Ahangar, M., Sheikhzadeh, H., Shahrebabaki, A. S., & Amini, J. (2015). Voice conversion based on feature combination with limited training data. *Speech Communication*, *67*, 113–128.

Helander, E., Silen, H., Virtanen, T., & Gabbouj, M. (2012). Voice conversion using dynamic kernel partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 806–817.

Helander, E., Virtanen, T., Nurminen, J., & Gabbouj, M. (2010). Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(5), 912–921.

Hu, Y., & Loizou, P. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238.

Kain, A., & Macon, M. (1998). Spectral voice conversion for text-to-speech synthesis. In *IEEE international conference on acoustics, speech and signal processing, ICASSP '98* (Vol. 1, pp. 285–288).

Kawahara, H., Masuda-Katsuse, I., & de Cheveign, A. (1999). 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*, 187–207.

Kominek, J., & Black, A. W. (2004). The CMU Arctic speech databases. In *Fifth ISCA workshop on speech synthesis*.

Lee, K. (2007). Statistical approach for voice personality transformation. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(2), 641–651.

Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, *28*(1), 84–95.

Mowlaee, P., Sayadiyan, A., & Sheikhzadeh, H. (2009). FDMSM robust signal representation for speech mixtures and noise corrupted audio signals. *IEICE Electronics Express*, *6*(15), 1077–1083.

Nakamura, K., Toda, T., Saruwatari, H., & Shikano, K. (2010). Evaluation of extremely small sound source signals used in speaking-aid system with statistical voice conversion. *IEICE Transactions on Information and Systems*, *93*, 1909–1917.

Nakashika, T., Takiguchi, T., & Ariki, Y. (2015a). Voice conversion using RNN pre-trained by recurrent temporal restricted boltzmann machines. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, *23*(3), 580–587.

Nakashika, T., Takiguchi, T., & Ariki, Y. (2015). Voice conversion using speaker-dependent conditional restricted boltzmann machine. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 8.

Naylor, P. A., & Gaubitch, N. D. (2010). *Speech dereverberation* (1st ed.). London: Springer.

Paliwal, K., & Atal, B. (1993). Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Transactions on Speech and Audio Processing*, *1*(1), 3–14.

Ramachandran, R., & Mammone, R. (1995). *Modern methods of speech processing* (1st ed.). New York: Springer. ISSN: 0893-3405.

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). An HMM-based singing voice synthesis system. In *Ninth international conference on spoken language processing, INTERSPEECH '06*, Pittsburgh, PA, USA, September 17–21.

Simovici, D. A., & Djeraba, C. (2014). *Mathematical tools for data mining* (2nd ed.). London: Springer.

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score revisited: Methods and applications, limitations and alternatives. *Multimedia Systems*.

Stylianou, I. (1996). Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications.

Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A., & Narayanan, S. (2006). Text-independent voice conversion based on unit selection. In *IEEE international conference on acoustics, speech and signal processing, ICASSP '06* (Vol. 1, pp. 81–84).

Takashima, R., Takiguchi, T., & Ariki, Y. (2012). Exemplar based voice conversion in noisy environment. In *Spoken language technology workshop (SLT)* (pp. 313–317).

Toda, T., Black, A., & Tokuda, K. (2005). Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *IEEE international conference on acoustics, speech, and signal processing, ICASSP '05* (Vol. 1, pp. 9–12).

Toda, T., Black, A. W., & Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(8), 2222–2235.

Toda, T., Ohtani, Y., & Shikano, K. (2006). Eigenvoice conversion based on Gaussian mixture model. In *Ninth international conference on spoken language processing, INTERSPEECH '06*, Pittsburgh, PA, USA, September 17–21, 2006.

Toda, T., Saruwatari, H., & Shikano, K. (2001). Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum. In *IEEE international conference on acoustics, speech, and signal processing, ICASSP '01* (Vol. 2, pp. 841–844).

Valbret, H., Moulines, E., & Tubach, J. (1992). Voice transformation using PSOLA technique. *Speech Communication*, *11*, 175–187.

Wu, Z., Virtanen, T., Chng, E. S., & Li, H. (2014). Exemplar based sparse representation with residual compensation for voice conversion. *IEEE Transactions on Speech and Audio Processing*, *22*(10), 1506–1521.

Wu, Z., Virtanen, T., Kinnunen, T., Chng, E. S., & Li, H. (2013). Exemplarbased voice conversion using nonnegative spectrogram deconvolution. In *8th ISCA speech synthesis workshop*.

**Mohammad Javad Jannati** is a Ph.D. student of electrical engineering at Amirkabir University of Technology. He received his B.S. and M.S. degrees all in Electrical Engineering, respectively, from Shahed University (2007), and Iran University of Science and Technology (2010). His Ph.D. research is centered on nonparallel voice conversion methods. His M.S. research was about Under Water Acoustic Communications. In B.S. he works on image compression based on wavelet transform. Now, he is a member of the Information Processing Research Laboratory of Amirkabir University of Technology.

**Abolghasem Sayadiyan** is an Associate Professor of Information Processing in Department of Electrical Engineering at the Amirkabir University of Technology since 1997. His research and teaching interest areas include Audio and speech processing, image and video processing, coding and inception methods and blind and sparse methods. He is the head of the Information Processing Research Laboratory of Amirkabir University of Technology.

**Abolfazl Razi** is an Assistant Professor of Informatics, Computing and Cyber-security at Northern Arizona University since September 2015. He received his B.S. and M.S. and Ph.D. degrees all in Electrical Engineering, respectively, from Sharif University (1998), Amirkabir University (2001), and the University of Maine (2013). His Ph.D. research was centered on distributed algorithm design for hybrid wireless networks with active and passive nodes. During his postdoctoral appointments at Duke University (2014) and Case Western Reserve University (2015), he developed data-driven predictive models in order to solve problems in biomedical signal processing and cancer genomics. He served as PACE chair of IEEE Maine Section in 2010–2012. He is the recipient of several academic awards including the best graduate research assistant of year from College of Engineering, University of Maine (2011) and the best paper award from IEEE/CANEUS fly by wireless workshop in Montreal, Canada (2011). His current research interests are in the intersection of wireless networking, high-dimensional biomedical signal processing and predictive modeling.