Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 00 (2015) 1–16

Speech Communication

Natural Spectrum Envelope Reconstruction via ϵ-Closed Extended Vectors Sets in Voice Conversion Systems

Mohammad Javad Jannati[1], Abolghasem Sayadiyan[1], Abolfazl Razi[2]

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

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

Abstract

In conventional voice conversion methods, first, some features of a speech signal's spectrum envelope are extracted. Then by designing and using a set of conversions, these features are converted so as to best match with a target speaker's speech. 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, which is observed in cases such as over-smoothing, over-fitting, and widening of formants, caused by two factors: 1) existence of error in the reconstruction of spectrum envelope from the features, and 2) the non-closure of the set of features extracted from the spectrum envelope of speech signal relative to the conversions used. In this research, a method to reconstruct a speech signal's natural spectrum envelope by means of ϵ-closed sets of extended vectors in voice conversion systems is introduced. In this approach, ϵ-closed sets for reconstructing 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. Also, 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 state, and can be applied to improve the naturalness of the generated speech signal in all the common voice conversion methods. Moreover, this method can be used in other fields of speech processing like text to speech systems and vocoders to improve the quality of the output signal in the synthesis step.

Keywords: Voice conversion; Synthesis based on vector quantization; ϵ-closed set; Extended vectors of features

1. Introduction

Following the use of gesture and pointing, speech is the first communication mechanism discovered and developed by humans. This means of communication, in addition to conveying information directly, can implicitly communicate information such as the feelings, particular expressions, character, and the other characteristics of a speaker. Examination and analysis of a speech signal, like the other acoustic signals, is an interesting subject for researchers. Speech processing concerns the analysis of the acoustic signals generated by the human vocal system. Voice conversion is a branch of speech processing with the most 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 [1]. In other words, based on proper and sufficient training, a voice conversion system able to establish an optimum mapping between a source speaker and a target speaker. In the course of this conversion, the linguistic information (i.e., speech context) should be preserved, and a quality speech should be generated as much as possible. Those familiar with a target speaker's speech should confirm the similarity of the converted speech to the target speaker's real speech.

1.1. Applications of voice conversion

The main application of voice conversion is in changing the character of a speaker while preserving the linguistic information [2]. With the help of voice conversion, the voice of the famous people or those who have passed away can be reproduced. In case of having sufficient recorded voice from these individuals, voice conversion techniques can be employed to generate their voices from a reference speaker. Voice conversion can be used in dubbing and sound recording [3]. Another application of voice conversion is to add rhythm and melody to an ordinary speech and to turn it into singing [4]. With the help of voice conversion, the speechs of the characters in a computer game can be generated, as the game is played, and based on the liking of the user; and thus the games becomes more attractive. Voice conversion can also be used in therapeutic procedures. For example, it can be employed in speech therapy practices or in teaching the correct form of speaking to children. Voice conversion helps those who have lost their larynx or vocal chords as a result of cancer or other diseases. These individuals can generate a voice by means of an electro larynx or non-audible murmur; however, this voice is of low quality and sounds mechanical. Voice conversion can turn this voice into a natural speech. [5]; and if sufficient amount of recordings from healthy vocal tracts exist, it can reproduce a similar speech. Voice conversion can also be used in pronunciation correction/enhancement for those who want to properly speak a new language [6]. One of the most complicated applications of voice conversion is translation of speech from one language to another, while preserving the message of the source speaker [7]. A specialized application of voice conversion is to increase the volume of data available in the training database. This is done by performing a regression between the data available in the database [8].

1.2. A review on voice conversion methods

One of the earliest works in voice conversion is a pioneer work in 1985 [9]. In the method presented in that work, the vocal tract lengths of the source and target speakers are calculated first by determining the formant frequencies from the training data, and by considering their ratios, the Linear Predictive Coding (LPC) coefficients of the speaker are converted to the LPC coefficients of the new speaker for a corresponding frame. The pitch frequency and the vocal tract excitation pulse shape are converted based on their average ratio. This was a good 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. In 1988, [10] 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. In 1992, [11] 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 et al. achieved a great accomplishment in the field of voice conversion due to using soft-clustering instead of hard-clustering [12]. In 1998, Kain and Macon modified the Stylianou's technique slightly and proposed a method based on jointly distributed Gaussian Mixture Models (JDGMM) [13]. This method displayed a higher stability relative to the Stylianou's method. The aforementioned GMM-based approaches, despite their wide use present two major drawbacks. Their first problem is over-smoothing, which stems from low model complexity. To reduce the degree of over-smoothing, the model complexity can be increased by increasing the degrees of freedom. However, the excessive increase of the degrees of freedom leads to another problem called over-fitting. In over-fitting, due to a model's high degrees of freedom, noise and other distortions or disorders of the model are also modeled, thereby reducing its quality. Thus, there should always be a tradeoff between over-smoothing and over-fitting. 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[14]. In 2005, Toda et al. made a great stride in the field of voice conversion by combining the JDGMM method and the speech parameter generation algorithm[15]. 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. In 2006, Toda et al. 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 acceptable, even by using only two training sentences from a target speaker [16]. 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 [17]. In 2009, the Artificial Neural Network (ANN) as a substitute for the MLE method is used [18]. The drawback of this approach is its large computational load and its immense complexity. In 2010, by combining the GMM and the Partial Least Square (PLS) regression,[19] succeeded in achieving a better performance relative to the JDGMM by using 10 training sentences . In 2012, [20] 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 [21]. Also, in 2013, [22] 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, [23] Achieved a performance superior than 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), [24] presented a method that they claimed performs better than the common GMM-based techniques. By combining the Mixture of Restricted Boltzmann Machines (MoRBM) and Mixture of Gaussian Bidirectional Associative Memories (MoGBAM), they were able to overcome problems such as the insufficiency of the JDGMM method in modeling the distribution of spectral envelope features and the loss of spectral details due to the use of high-level spectral features. 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 [25] [26].

2. Two drawbacks in voice conversion

In common voice conversion techniques, first, some features of a speech signal's spectrum envelope are extracted. Then using a set of conversions, these features are converted so as to have the best match with a target speaker's speech. Ultimately, the spectrum envelope of the target speaker's speech signal is reconstructed from the converted features. Generally, the spectrum envelope reconstructed from the converted features deviates from its natural form. This unnaturalness of speech, which is observed in cases like over-smoothing, over-fitting, and widening of formants, has been caused by two factors: 1) existence of error in the reconstruction of spectrum envelope from the extracted features, and 2) the non-closure of the set of features extracted from the spectrum envelope of speech signal relative to the conversions used. In this research, we first elaborate on these drawbacks in section 2.1 and 2.2. Then, we proposed a method based on using a set of extended vectors to tackle these problems as detailed in section 3 which is titled: "Reconstruction of natural spectrum envelope by using a set of extended vectors in voice conversion systems".

2.1. The problem of reconstructing the spectrum envelope from the spectral features

In voice conversion, after modifying a signal's spectrum envelope based on the newly generated parameters, the modified spectrum envelope and along with the other information such as pitch and gain are employed to synthesize the final speech signal. Hence, the quality of the synthesized speech is directly related to the quality of the generated spectrum envelope. The spectrum envelope of the output signal is estimated from spectral features such as Mel Frequency Cepstral Coefficients (MFCC), Line Spectral Frequencies (LSF) and etc. These features have already been extracted from the spectrum envelope of the original signal and modified by the voice conversion system. Depending on the features used, different methods are applied to estimate the spectrum envelope from the features. Irrespective of whether the features are modified or not, the reconstruction of the spectrum envelope from these features comes with error. The amount of this error depends on the type and also the number of feature used and the method employed to reconstruct the spectrum envelope from the features. Figure 1 shows the original and the reconstructed spectrum envelopes of one frame from the speech signal, from a relatively stable point, of vowel uttered by a male speaker from database mentioned in section 4.1. The speech signal has been recorded in 16-bit format, by using a sampling frequency of 16000 Hz and frame lenght is 5 ms. Different methods such as Sinusoidal Model (SM), Harmonic Stochastic Model (HSM), Harmonic Noise Model (HNM) and Fixed Dimension Modified Sinusoidal Model (FDMSM) [27] are reported for signal analysis, spectrum envelope extraction, and for speech synthesis. In this investigation, the spectrum envelope has been generated by means of the Speech Transformation and Representation

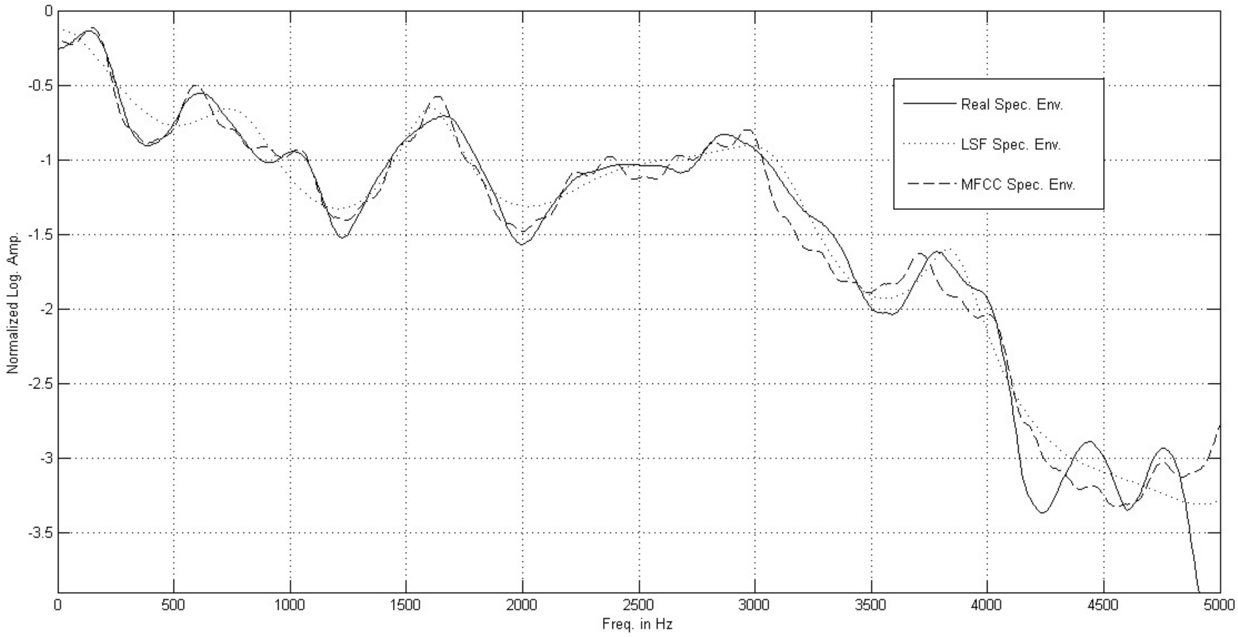

Figure 1. The original spectrum envelopes and the spectrum envelopes reconstructed from MFCC and LSF in the frequency range of 0-5 kHz

using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) method [28]. To evaluate the amount of error in the feature extraction process and the reconstruction of spectrum from the features, the features of LSF and MFCC have been extracted from the original spectrum envelope and again, without any conversion, the spectrum envelope has been reconstructed from these features. As is observed, although no conversion or processing has been carried out on the features, the original spectrum envelope and the reconstructed spectrum envelope are different from each other. The LSF performs better at certain frequencies, and the MFCC performs better at some other frequencies. The common criterion of Log-Spectral distortion error (LSD) for the comparison between the original and the reconstructed spectrums in the ith frame (in dB) is defined as [29] [30]

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

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. In [31], Paliwal has reported the maximum acceptable average value of LSD in a speech signal to be about 1 dB. Of course, the lower this value is, less error has occurred; however, for an average LSD value of about 1 dB, the resulting error (quality of reconstructed voice) will be acceptable. If the amount of LSD is determined from Expression 1 for the frame illustrated in Figure 1, then the values of 0.75 dB and 0.79 dB will be respectively obtained for the signal spectrum envelopes reconstructed from the MFCC and LSF. Thus, both signal reconstructions have errors; but comparatively, the signal spectrum envelope reconstruction from MFCC is more reliable (with equal number of features).

2.2. Non-closure of feature set with respect to the conversions useds

In conventional voice conversion methods, the voice conversion systems are normally trained by a set of features (MFCC, LSF andetc.) extracted from a considerable volume of speech signals used as the training data. Then the speech of the target speaker is fed into this system as the input and, depending on the input and the training mechanism of the system, the features of the converted speech are generated as the output of the voice conversion system. 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 (the mathematical definition of a closed set has been presented in Section 3.1). Conversely, 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, which the spectrum envelope extracted from this set results in the reconstruction of a synthetic, and not a natural, voice signal. Of course, as was mentioned in the previous section, many attempts have been made to bring this synthesized speech closer to natural voice.

3. Reconstruction of natural spectrum envelope by using a set of extended vectors

As mentioned earlier, two major drawbacks with the current spectrum envelope reconstruction techniquesare: 1) existence of reconstruction error, and 2) the non-closure of the features set. The method proposed in this research, titled Natural spectrum envelope reconstruction via ϵ-closed sets of extended vectors (NERVES) in voice conversion systems, by presenting a series of ϵ-closed sets of extended vectors for the reconstruction of spectrum from features,aims at reconstructing a speech signal's natural spectrum envelope and reducing the reconstruction error. The NERVES method first defines a number of extended vectors for each speaker. Then by presenting a quantization algorithm, this technique produces a reference ϵ-closed set of these extended vectors for each speaker. Finally, by using the reference ϵ-closed sets produced from the extended vectors, it reconstructs the spectrum envelope from the features. In the following section, first, the ϵ-closed set is defined. Then the extended vector is defined and the manner of its generation is explained. After that, the algorithm for extracting the ϵ-closed set from the reference extended vectors presented by the NERVES, the proposed method in this research, is introduced. Subsequently, the method of reconstructing a spectrum envelope via NERVES is explored. Finally, the adjustment procedure for the main parameters in the NERVES along with their description, roles and impacts is discussed.

3.1. Defining the ϵ-closed set

Mathematically, set A:$\{a_1, a_2, , a_n\}$ is considered closed with respect to the operator group F, if for a_i being an arbitrary member of A, $F(a_i)$ isl also be a member of A [32]; in other words, if any member of A is operated on by the operator group F, again a member of A be obtained. Thus, A will be a closed set, if and only if

$$if\{a_i \in A\} \rightarrow \{F(a_i) \in A\} \tag{2}$$

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 [32]:

$$Closed\,Segment\,of(a_l, a_k) = \{\alpha a_l + (1 - \alpha)a_k | 0 \leq \alpha \leq 1\} \tag{3}$$

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 [32]. Likewise, the ϵ-closed set is defined as follows. Set A:$\{a_1, a_2, , a_n\}$ will be an ϵ-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 ϵ. 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 ϵ from A (a member of A). Formally, Set A is ϵ-closed with respect to operator F, if

$$if\{a_i \in A\} \rightarrow \{\exists F(a_j) \in A : |F(a_i) - a_j| \leq \epsilon\} \tag{4}$$

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

$$if\{a_l, a_k \in A\} \rightarrow \{\exists F(a_j) \in A : |\alpha a_l + (1 - \alpha)a_k - a_j| \leq \epsilon\} \tag{5}$$

3.2. Generating the extended vectors

Provided that, there is sufficient speech signal 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 Figure 2. In this research, we use STRAIGHT method due to its superior quality of Spectrum envelope reconstruction. In the analysis phase, for each frame, the pitch frequency (F0), spectrum envelope (Spec), aperiodicity (AP) coefficients, Speech/non-speech flag, Speech/non-speech flag, and the sampling frequency are determined and recorded. In the

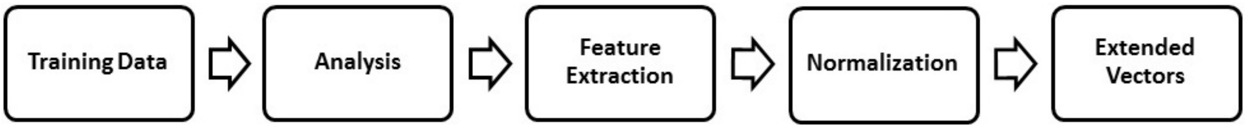

Figure 2. Steps of generating the extended vectors for the training database

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 Line Spectral Frequencies	24	Vector - 24 Double
Spec	Spectrum Envelope	1024	Vector - 1024 Double
Ap	Aperiodicity Coefficients	1024	Vector - 1024 Double

features extraction phase, for each frame, several features that are representative of the spectrum envelope are extracted from the spectrum envelope. In this research, 24 LSF coefficients have been 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 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 1^{st} LSF coefficient will have a range of values less than 400 Hz and the 24^{th} LSF coefficient will have a range of values between 7500-8000 Hz. Before any modification, it is necessary to normalize the ranges of coefficients. For this purpose, first, the mean (μ) and standard deviation (σ) values of the speech frame coefficients are computed for the entire database and for the 1^{st} through 24^{th} coefficients of LSF. Then by applying the following formula for all the database frames, all the feature vector values are normalized

$$NLSF = (LSF_i - \mu_i)/\delta_i \tag{6}$$

For each frame, NLSF$_i$ is the ith normalized LSF coefficient, LSF_i is the ith coefficient of LSF, μ_i is the mean value of the ith LSF coefficient, and σ_i is the standard deviation of the ith coefficient of LSF for all the speech frames of the database. By applying Expression 1, the dynamic ranges of all the 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 titles of the sub-parts of an extended vector have been presented in Table 1. In view of Table 1, in the present research, each frame of the database is represented by an extended vector that has 7 sub-parts and includes a total of 2076 parameters.

3.3. The proposed NERVES Method

The NERVES method comprises of two algorithms. The First one is for choosing an ϵ-closed set of extended vectors for an arbitrary speaker from all him/her extended vectors available in the database. The second one is for synthesis (reconstructing spectrum envelpe) of speech signal of the same speaker.

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

$$SV_i : \{ \underbrace{\{NLSF\}}_{First\ Part\ of Extended\ Vector} , \underbrace{\{F0\}, \{S/NS\}, \{V/UV\}, \{G\}, \{Spec\}, \{Ap\}}_{Second\ Patr\ of Extended\ Vector}\} \quad , i \in \{1, 2, \cdots, N\} \tag{7}$$

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 (paramerers necessary for synthesis). In generating the set of extended vectors (first algorithm), just first part of EV is used to elect the reference extended vectors between all sample extended vectors of a speaker. In this process, no alteration or transformation is applied on the second part and the reference extended vectors are selected based on the first part of the EV. The second part of the EV remains intact in all steps of the algorithms and will be used only in synthesis (second algorithm).

In the following, the algorithm used to extract M reference extended vectors associated with an arbitrary speaker comes (M is a parameter and value of it will be chosen in final step of algorithm). Then NERVES Algorithm for Reconstructing the spectrum envelope is described.

3.3.1. NERVES algorithm for extracting an ϵ-closed set from the reference extended vectors

Befor Starting the algorithm, the extended vectors related to the speech frames are separated. Next, depending on whether these extended vectors are voiced or un-voiced, they are divided into two groups. Then the following NERVES algorithm is independently and separately executed on the voiced and un-voiced groups of extended vectors. NERVES algorithm for extracting an ϵ-closed set from the reference extended vectors has three steps as follow

- Initialization step

  For $i = 1$

  1. EV$_1$ is selected as first reference EV.

  2. The variable for counting the number of repetitions of first reference EV (EV_No_1), is set to 1.

  3. The variable for the total number of reference EVs ($Total\_EV$), is set to 1.

- Recursion step

  For i from 2 to N, the following steps are taken:

  1. 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 8) .

  2. If at least one or more than one of these distances is less than the threshold value δ,

  EV$_i$ will be assigned to the first reference EV whose distance from EV$_i$ is less than δ.

  The number of repetitions of that reference EV will be raised by one (1).

  3. If the condition of Step 2 is not met,

  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

  1. The reference EVs are separated.

  2. Every reference EV whose repetition is less than 2 is omitted (as outliers).

  3. The M remaining reference EVs are chosen as members of the ϵ-closed set of EVs.

The distance used in the above algorithm can be defined based on the type of features used. For determining the distance between the feature vectors, different approaches such as the Euclidian distance and the perceptual methods [33] have been introduced. Since NLSF coefficients have been used in this research, the distance between the two extended vectors EV$_k$ and EV$_j$ is computed from the following relation [34]:

$$Dist(EV_k, EV_j) = \sum_{i=1}^{K} |\omega_i(NLSF_k(i) - NLSF_j(i))|^2 \tag{8}$$

In the above equation, NLSF$_k$ are the feature coefficients of extended vector EV$_k$, NLSF$_j$ are the feature coefficients of 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). ω_i is a weight factor. In this research ω_i has been considered as normalizd mel-scaled mean of LSF(i) in total database.

As it mentioned in expression 8, The distance between two EV is equal to the distance between their NLSFs (first part of EV). By using the above algorithm, a set of extended vectors can be generated for any arbitrary database. The manner of extracting the natural spectrum envelope from the features by means of such a set of extended features will be described in the following section. It should be pointed out that the spectrum envelope extracted from features by using the method proposed in this paper is always natural, but it may come with an error. The amount of this error depends on the variety, degree of covering and the comprehensiveness of the database used. With more diverse and extensive speechs and with a higher number of speakers in the database, the spectrum envelope will become more natural and the generated voice will have a higher quality.

3.3.2. NERVES Algorithm for Reconstructing the spectrum envelope

To reconstruct the spectrum envelope from the features by the NERVES method, the following procedure is implemented.

Reconstruction process:

1. The features of the considered frame (NLSF in this research) are obtained by any arbitrary method.

2. For each Frame, it is determined whether the frame is voiced or un-voiced; and accordingly, an appropriate extended vectors set is used.

3. By using Expression 8, the distance of each frame from all reference EVs of the ϵ-closed set of EVs obtained in the previous section, is calculated.

4. The reference EV with the least distance (δ_{min}) from the each frame is selected.

5. For each frame, the spectrum envelope of the selected EV (second part of EV) is chosen as the spectrum envelope of the frame.

3.4. Selecting the values of ϵ and δ

Paliwal [31] has stated that if the average LSD between the frame spectrum envelopes of two speech signals is about 1 dB, human ear recognizes these two voices to be acceptably similar. 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 amount of error will be acceptable. Therefore, if the set of extended vectors produced in this research by the proposed method is ϵ-closed and if the amount of ϵ is about 1 dB, every arbitrary vector from this set will have a maximum distance of ϵ from one the reference vectors and, consequently, its difference with the mentioned reference vector can be ignored. As is observed in the presented algorithm, δ is the threshold distance of an ordinary extended vector from the reference extended vectors. The amount of δ must be selected so that the set remains ϵ-closed; and therefore, δ must be chosen such that the average LSD value does not exceed 1 dB. Considering the abovementioned points, for every speaker, the value of δ must be chosen so that the average LSD is equal to 1 dB, and so that the set of extended vectors for every speaker, for an $\epsilon = 1$ dB, remains ϵ-closed. Two values of 0.05 and 0.001 for δ have been examined in [34]. Of course, considering the fact that in the coding process, the number of bits used for each coded symbol is considered as a limitation, in [31], [34] and other similar papers in the context of coding and data compression, limits have been placed on the number of bits used for each symbol. This issue influences the effective value of δ 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 in the applications of this work; and therefore, no limitation has been imposed on δ in this regard. So, the voices recorded from 4 speakers (detailed descriptions about these speakers have been provided in Section 3.4) were analyzed for three δ values of 0.01, 0.05 and 0.001, without considering any constraints on the number of bits. Using 50, 75 and 100 training sentences, the NERVES algorithm was applied for each speaker and each of the above values to generate an ϵ-closed set of extended vectors. Then, 20 sentences that were not applied in the training were used in the test phase.

Table 2 and 3 show the obtained results for each speaker and for different δ values. As is observed, 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 sentences than when they increase from 75 to 100 sentences. This means that the rate of reduction of LSD declines as the the number of training sentences grows. For 100 training sentences and for $\delta = 0.01$, the mean LSD of

Table 2. Mean (μ) and standard deviation (σ) of LSD (in dB) obtained for δ values of 0.001, 0.05 and 0.01 by using 50 and 75 training sentences

Speaker	50 Train Sentences						75 Train Sentences					
	δ =0.001		δ =0.05		δ =0.01		δ =0.001		δ =0.05		δ =0.01	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
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 (μ) and standard deviation (σ) of LSD (in dB) obtained for δ values of 0.001, 0.05 and 0.01 by using 100 training sentences

Speaker	100 Train Sentences					
	δ =0.001		δ =0.05		δ =0.01	
	μ	σ	μ	σ	μ	σ
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

the test sentences for all 4 speakers is about 1 dB. So, considering the fact that more than 100 sentences are available from each speaker, in case of using 100 or more training sentences, for $\delta = 0.01$, the NERVES algorithm can present an ϵ-closed set of extended vectors for each speaker, whose total LSD in the test phase will be about 1 dB. Thus, a δ value of 0.01 was used in this research.

Table 4 shows the results of the ϵ-convex test for the set of reference extended vectors for each speaker. In this test, for the 5 values of α, all the possible double combinations of the members of the extended vectors set for each speaker were selected, and the closed segments that these combinations produce ($\alpha a_l + (1 - \alpha)a_k$) were obtained. Then the distance of the obtained expression to the nearest reference vector was determined based on the LSD criterion. The mean and standard deviation values obtained for all the double combinations have been listed in Table 4 for different speakers and α values. As is observed, the average LSD values are about 1 dB, which are acceptable, considering the criterion presented in [31]. Thus, for $\delta = 0.01$, the sets of reference extended vectors obtained for each speaker are ϵ-closed and ϵ-convex, with regards to Expressions 4 and 5, respectively.

4. Implementing the NERVES method and analyzing its results

Since, in reconstructing the spectrum envelopes, the NERVES algorithm only uses the spectrum envelopes of the reference extended vectors for synthesis, the possible output spectrum envelopes are limited to the spectrum envelopes of the extended vectors generated by the NERVES. 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) irrespective of what signal, and with what types of features, is synthesized, all the output spectrum envelopes belong to an ϵ-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, we will have a spectrum envelope in the output, 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; and so no spectrum envelope outside this set will exist in the output. Thus, the set of spectrum envelopes will remain intact and unaltered. This improvement in naturalness and intactness has some consequence costs. First cost is the need of more data storage capacity for saving EVs (insted of spectral features). Although this can be mentioned as a cost but with the large capacity of today hard disk drives it is not a serious problem. Second is the computational cost for extracting the ϵ-closed sets of EVs. As was mentioned in previous parts, this operation is offline and just one time must be done for each database. Therefore it can be ignored. Third cost of improvement in naturalness is the introduction of error in the 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 3 of the algorithm for reconstructing the spectrum envelope) and using its spectrum envelope instead of the converted

Table 4. Mean (μ) and standard deviation (σ) of LSD (in dB) for $\delta = 0.01$ and for α values ranging from 0.1 to 0.5, for all the dual combinations of reference vectors for each speaker

	$\alpha = 0.1$		$\alpha = 0.2$		$\alpha = 0.3$		$\alpha = 0.4$		$\alpha = 0.5$	
Speaker	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
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

spectrum envelope. The longer this nearest distance (δ_{min}) is, the higher the reconstruction error of that frame will be. The value of δ_{min} is inversely related 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, the δ_{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 speakers and the extent and diversity of the existing voices in the database. So, as was previously mentioned, 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 used database is introduced and the test setup is described. Then the spectrum envelope reconstruction results obtained by the NERVES approach are compared with those of other methods. Finally, the effect of the NERVES method on converted voice, in conjunction with the application of two common and standard voice conversion techniques, is investigated.

4.1. The test setup

To evaluate the effect of NERVES algorithm on the results of the two methods mentioned in the previous section, the objective test of Perceptual Evaluation of Speech Quality (PESQ) [35] and the subjective test of Mean opinion score (MOS) [36] and also the preference test were used. For this purpose, the STRAIGHT method with a frame shift of 5 ms was employed for data analysis and synthesis. 24 features of LSF and MFCC were chosen for each frame. To produce the extended vectors, the data of the CMU ARCTIC database [37] established by the Carnegie Mellon University's speech group were used, at a sampling frequency of 16 kHz. A full description of this database has been given in [37]. 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 based on these sentences. The objective test of PESQ and the subjective test of MOS were used to compare the qualities of the generated speech signals. In the PESQ test, the method and the code presented in [35] were used. 8 native speakers were employed for the MOS test, in order to determine the quality of the produced voices. Also, the preference test was used to check the naturalness of the generated voices.

4.2. Results of reconstructing the spectrum envelope from the feature vectors by the NERVES method

It was previously mentioned that error will accompany the process of reconstructing a speech signal's spectrum envelope from the features. Figure 3 shows the spectrum envelope reconstructed by the method of using the ϵ-closed set of extended vectors in comparison with the spectrum envelopes obtained from LSF and MFCC at the frequency range of 0-5 kHz (similar to Fig. 1). As is observed, at the range of 0-1400 Hz, the best performance belongs to NERVES. The LSF performs better from 1500 to 1700 Hz. From 1700 to 3300 Hz, NERVES has a better performance. From 3300 to 4000 Hz, LSF excels again. At frequencies higher than 4000 Hz, NERVES performs much better. The MFCC doesn't have the best performance at any frequency range; however, since it also never has the worst performance, on the average, it performs better than LSF. LSF performs better at the hills than valleys. NERVES has a good performance at high frequencies, which contain the details and play an effective role in voice quality. The value of the LSD measure for reconstructing the spectrum envelope from the features in the considered frame in Figure 3, obtained for the NERVES method by using Expression 1, is 0.64 dB. The LSD values for producing the spectrum envelope solely from MFCC and LSF are 0.75 and 0.79 dB, respectively. This shows the quantitative superiority of the NERVES approach. Of course, the other special advantage of the NERVES method is that the spectrum envelope produced by it is natural and intact; this characteristic doesn't exist in the MFCC and LSF, and

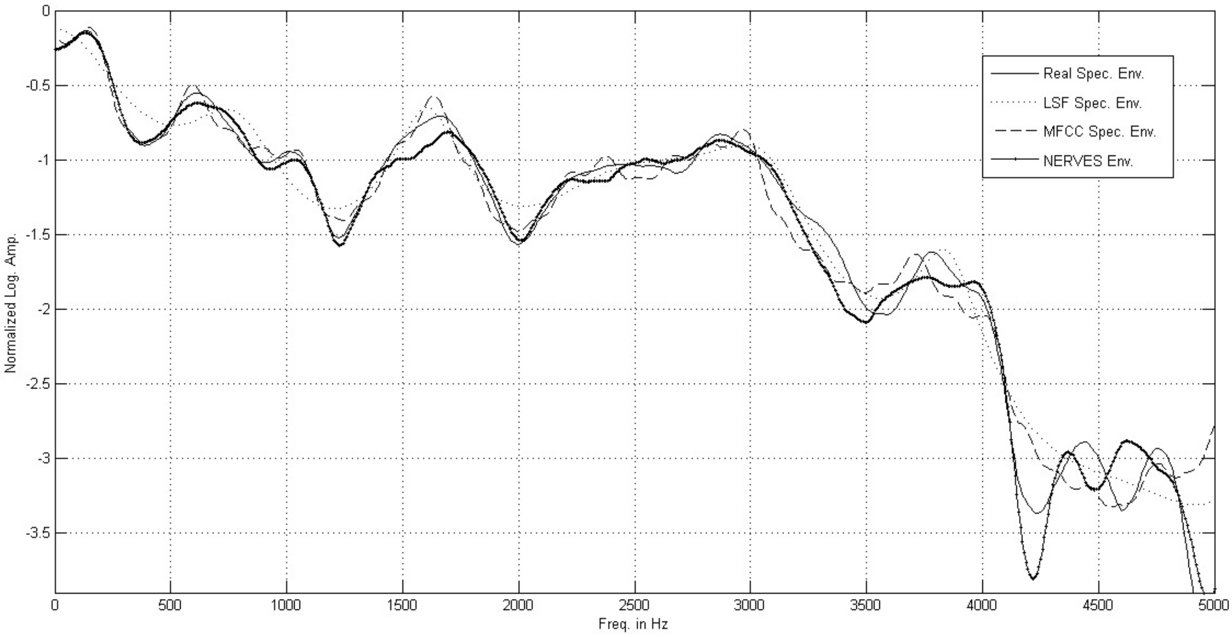

Figure 3. The original spectrum envelopes and the spectrum envelopes reconstructed by MFCC, LSF and NERVES at the frequency range of 0-5 kHz

it is impossible for these two methods to achieve such a quality. So the NERVES approach was able to reduce the spectrum envelope reconstruction error in the frame depicted in the figure, and it could successfully reduce the effects of the first cause of speech unnaturalness (i.e., existence of error in reconstructing a signal's spectrum envelope). This result is not exclusive to the illustrated frame. For a more comprehensive comparison of the reconstruction errors in the direct reconstruction of spectrum envelope by the MFCC and LSF with the reconstruction error obtained in the NERVES 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 the PESQ and MOS tests, the qualities of the speechs synthesized by the MFCC, LSF and NERVES methods were compared with each other. Figures 4 and 5 illustrate the results of the PESQ and MOS tests, respectively and the results of these two tests match each other and indicate the superiority of the NERVES method. As is observed, the NERVES method can reconstruct the natural specrtum envelope from arbitrary features and synthesis the speech signal. Therefore, NERVES can be used in other fields of speech processing like text to speech systems (TTS) and vocoders to improve the quality and naturalness of the generated speech signal in the synthesis step and its usage in not limited to voice conversion.

4.3. Improving the quality and the naturalness of the converted voice by means of the NERVES method

As was previously mentioned, 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 the NERVES method preserves a signal's natural spectrum envelope, it is not hampered by the abovementioned problems and is a suitable choice for generating a signal's spectrum envelope from the converted features; so that the closest natural spectrum envelope is selected as the output. As Section 3.3 indicates, in the NERVES 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 the NERVES on

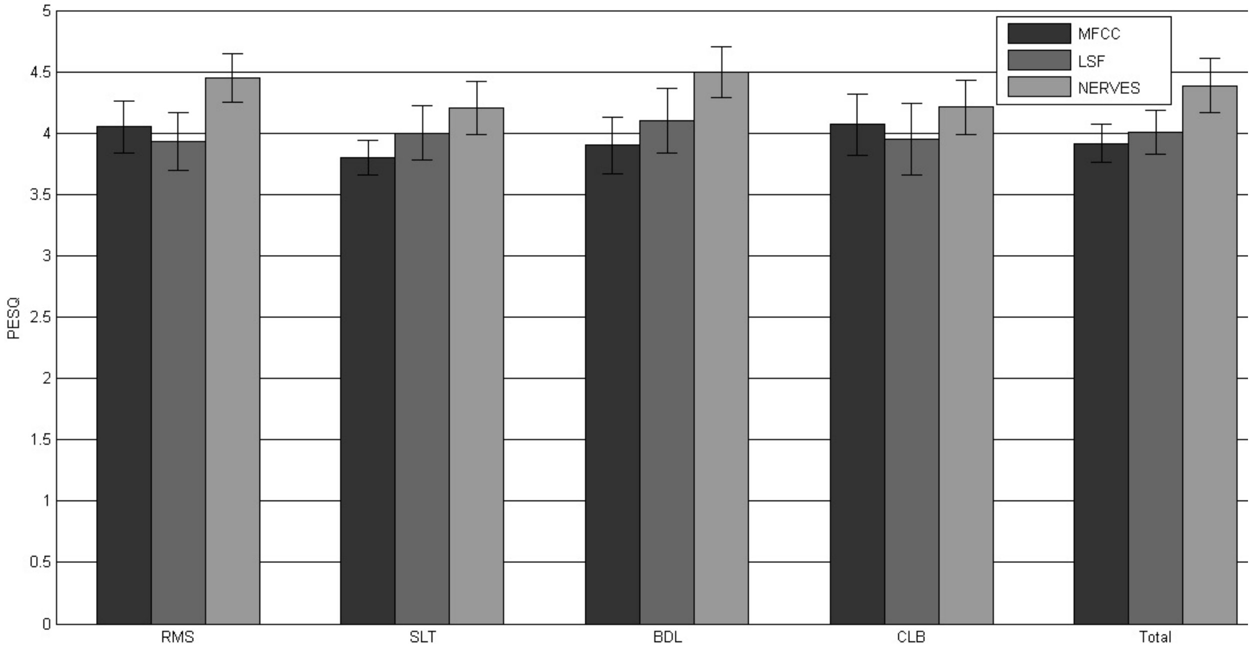

Figure 4. Results of the PESQ test for the reconstruction of spectrum envelopes from features directly and also by using the NERVES method

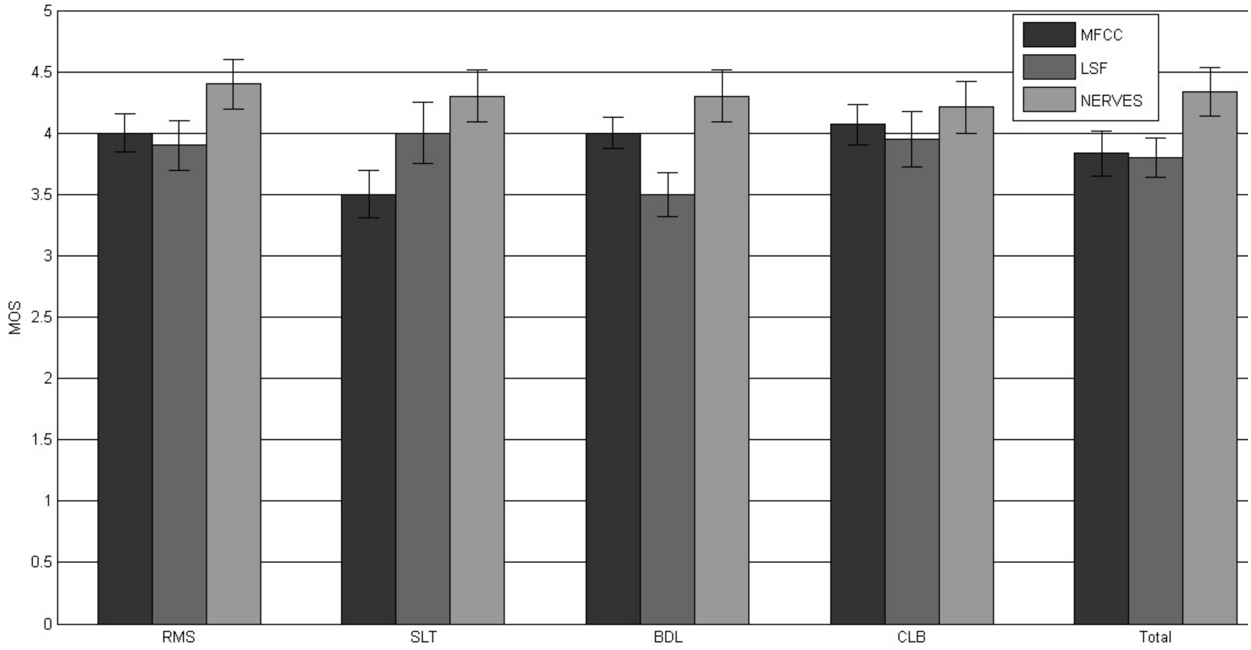

Figure 5. Results of the MOS test for the reconstruction of spectrum envelope from features directly and also by using the NERVES method

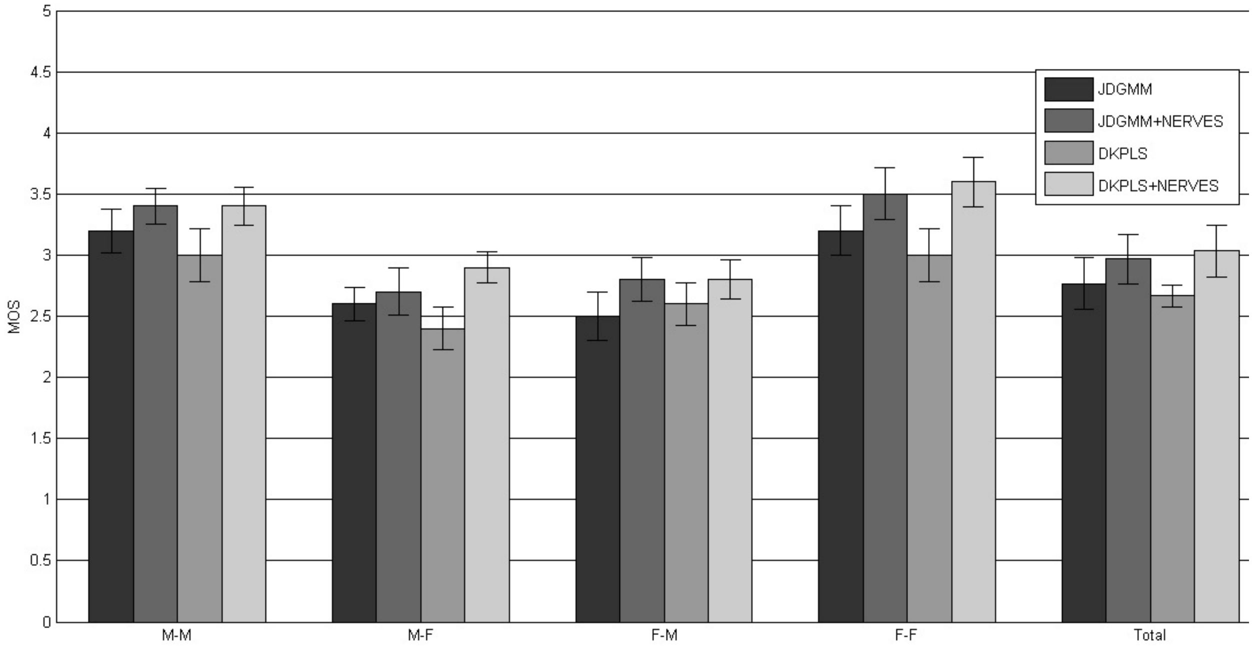

Figure 6. Comparing the results of the MOS test regarding the performances of the JDGMM and DKPLS methods with and without the application of NERVES

two known and common voice conversion methods, i.e., JDGMM [13] and DKPLS [19], the performances of these two techniques, solely and in conjunction with the NERVES method, are compared. To observe the effect of the NERVES method on the quality of the generated voices, all the possible voice transformations between the speakers were taken into consideration. So the voice of each speaker was converted to the voices of the other three speakers; and consequently, 12 conversions were made among the speakers. Using the MOS test and the preference test, the voice signals generated with, and without, the use of NERVES approach were compared for the standard methods of JDGMM and DKPLS. The results of the 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)) have been illustrated in Figure 6. As is observed, the performances of both standard methods have improved. Figure 7 shows the results of the naturalness preference test for JDGMM by itself and JDGMM+NERVES. According to this figure, the naturalness of the converted voice has improved. Figure 8 illustrates the results of the same test for the sole DKPLS and DKPLS+NERVES. 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, the PESQ criterion is not usually used for comparing the qualities of converted voices. In the comparisons between converted voices, the PESQ approach does not yield a reliable answer, and so it is not regularly used in the papers dealing with voice conversion. Thus, in evaluating the effect of NERVES on the two standard voice conversion methods, the PESQ criterion was not used. As the results indicate, the application of NERVES in conjunction with the 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 the NERVES approach. This means that, by using real spectrum envelopes, NERVES produces higher quality voices that have a natural spectrum envelope. Since NERVES is not dependent on the type of transformation used, it can be effectively used in combination with other voice conversion methods.

All the mentioned tests were also performed on two male and two female Farsi speakers from the database established by the Information Processing Research Laboratory of Amirkabir University in Iran, and similar results were obtained. These findings confirm the independency of the NERVES method from the language spoken by a speaker. The mentioned database includes the speech signals of 50 male and 50 female native Farsi speakers (without a particular accent). From each speaker in this database, 232 and 40 sentences are available for training and testing, respectively. These sentences have been adjusted so as to achieve a uniform use of language phonemes. The speech

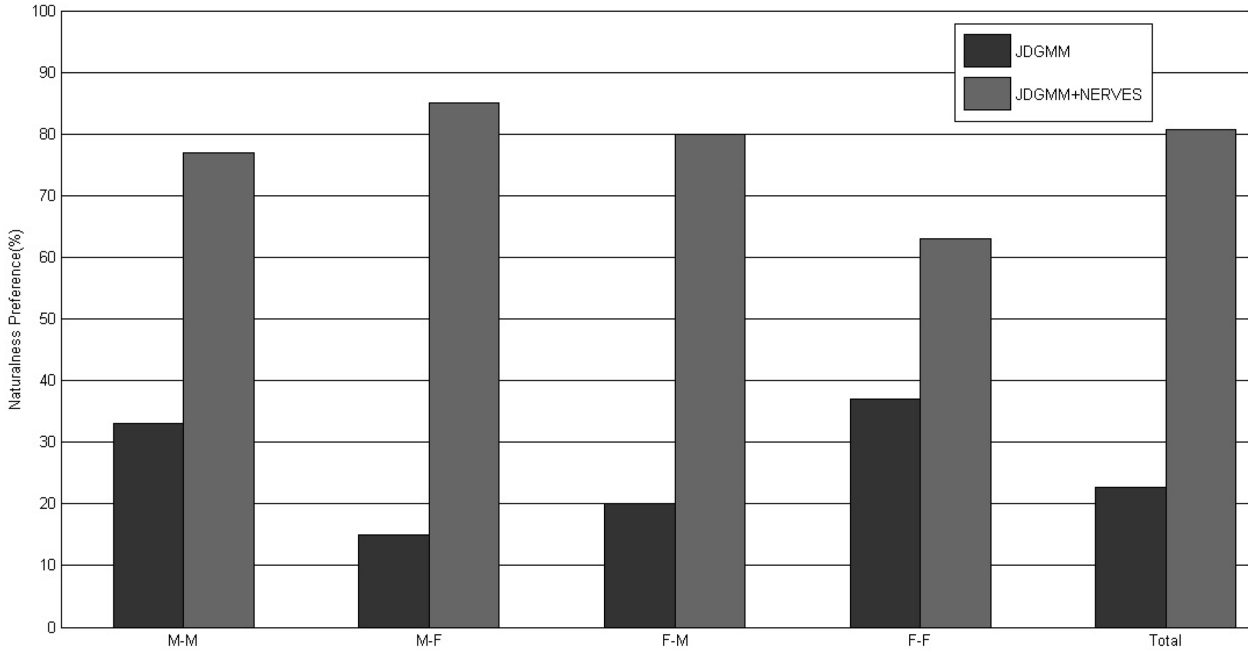

Figure 7. Comparing the naturalness qualities of the voices converted by the JDGMM method with and without the application of NERVES

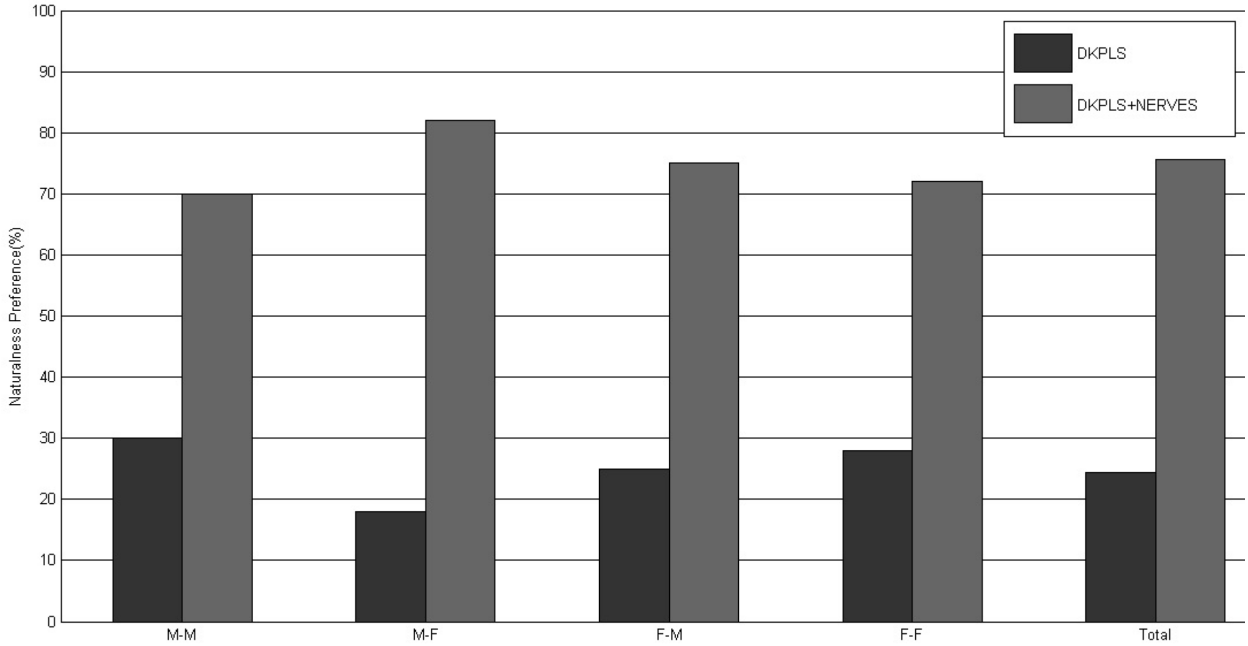

Figure 8. Comparing the naturalness qualities of the voices converted by the DKPLS method with and without the application of NERVES

samples have been obtained at a sampling frequency of 16000 Hz and have a good quality. However, since this database has not been officially published, the test results for these speakers have not been presented and referred to in this research.

5. 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 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. In this research, the NERVES method was introduced for solving these two problems by producing ϵ-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 ϵ-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, the MOS test indicated that the NERVES 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. Also, the results of the preference test on the voice signals converted by the JDGMM and DKPLS methods with, and without, the application of NERVES indicated that the use of NERVES in conjunction with the mentioned methods improves the naturalness of the generated voices. The NERVES 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. The NERVES approach is not dependent on the method of voice conversion and whether it is parallel or non-parallel. Moreover, NERVES can be used in other fields of speech processing like text to speech systems (TTS) and vocoders to improve the quality and naturalness of the generated speech signal in the synthesis step and its usage in not limited to voice conversion.

[1] T. Toda, A. W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, IEEE Transactions on Audio, Speech and Language Processing 15 (8) (2007) 2222–2235. doi:10.1109/tasl.2007.907344.

[2] K. Lee, Statistical approach for voice personality transformation, IEEE Transactions on Audio, Speech and Language Processing 15 (2) (2007) 641–651. doi:10.1109/tasl.2006.876760.

[3] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, S. Narayanan, Text-independent voice conversion based on unit selection, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '06, Vol. 1, 2006, pp. 81–84. doi:10.1109/ICASSP.2006.1659962.

[4] K. Saino, H. Zen, Y. Nankaku, A. Lee, K. Tokuda, An HMM-based singing voice synthesis system, in: Ninth International Conference on Spoken Language Processing, INTERSPEECH '06, Pittsburgh, PA, USA, September 17-21, 2006.

[5] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, H. Sheikhzadeh, A. S. Shahrebabaki, J. Amini, Voice conversion based on feature combination with limited training data, Speech Communication 67 (2015) 113–128. doi:10.1016/j.specom.2014.12.004.

[6] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Evaluation of extremely small sound source signals used in speaking-aid system with statistical voice conversion, IEICE Transactions on Information and Systems 93 (2010) 1909–1917. doi:10.1587/transinf.E93.D.1909.

[7] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, T. Dutoit, Cross-language voice conversion based on eigenvoices, in: 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009, pp. 1635–1638.

[8] E. Eide, M. Picheny, Towards pooled-speaker concatenative text-to-speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '06, Vol. 1, 2006, pp. 73–76. doi:10.1109/ICASSP.2006.1659960.

[9] D. Childers, B. Yegnanarayana, K. Wu, Voice conversion: Factors responsible for quality, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '85., Vol. 10, 1985, pp. 748–751. doi:10.1109/ICASSP.1985.1168479.

[10] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, Voice conversion through vector quantization, in: International Conference on Acoustics, Speech, and Signal Processing, ICASSP-88., 1988, pp. 655–658 vol.1. doi:10.1109/ICASSP.1988.196671.

[11] H. Valbret, E. Moulines, J. Tubach, Voice transformation using PSOLA technique, Speech Communication 11 (1992) 175 – 187. doi:10.1016/0167-6393(92)90012-V.

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

[13] A. Kain, M. Macon, Spectral voice conversion for text-to-speech synthesis, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Vol. 1, 1998, pp. 285–288 vol.1. doi:10.1109/ICASSP.1998.674423.

[14] T. Toda, H. Saruwatari, K. Shikano, 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, 2001, pp. 841–844 vol.2. doi:10.1109/ICASSP.2001.941046.

[15] T. Toda, A. Black, K. Tokuda, 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, 2005, pp. 9–12. doi:10.1109/ICASSP.2005.1415037.

[16] T. Toda, Y. Ohtani, K. Shikano, Eigenvoice conversion based on gaussian mixture model, in: Ninth International Conference on Spoken Language Processing, INTERSPEECH '06, Pittsburgh, PA, USA, September 17-21, 2006, 2006.

[17] D. Erro, A. Moreno, Weighted frequency warping for voice conversion, in: Annual Conference of the International Speech Communication Association, InterSpeech '07, 2007.

[18] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, K. Prahallad, Voice conversion using artificial neural networks, IEEE International Conference on Acoustics, Speech, and Signal Processing 0 (2009) 3893–3896. doi:10.1109/ICASSP.2009.4960478.

[19] E. Helander, T. Virtanen, J. Nurminen, M. Gabbouj, Voice conversion using partial least squares regression, IEEE Transactions on Audio, Speech, and Language Processing 18 (5) (2010) 912–921. doi:10.1109/TASL.2010.2041699.

[20] E. Helander, H. Silen, T. Virtanen, M. Gabbouj, Voice conversion using dynamic kernel partial least squares regression, IEEE Transactions on Audio, Speech, and Language Processing 20 (3) (2012) 806–817. doi:10.1109/tasl.2011.2165944.

[21] D. Erro, E. Navas, I. Hernáez, 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, 2012, pp. 86–89.

[22] D. Erro, E. Navas, I. Hernaez, Parametric voice conversion based on bilinear frequency warping plus amplitude scaling, IEEE Transactions on Audio, Speech, and Language Processing 21 (3) (2013) 556–566. doi:10.1109/tasl.2012.2227735.

[23] X. Chen, L. Zhang, 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) (2014) 68–75. doi:10.1016/s1005-8885(14)60333-2.

[24] L. Chen, Z. Ling, L. Liu, L. Dai, Voice conversion using deep neural networks with layer-wise generative training, IEEE ACM Transactions on Audio, Speech, and Language Processing 22 (12) (2014) 1859–1872. doi:10.1109/taslp.2014.2353991.

[25] T. Nakashika, T. Takiguchi, Y. Ariki, Voice conversion using RNN pre-trained by recurrent temporal restricted boltzmann machines, IEEE ACM Transactions on Audio, Speech, and Language Processing 23 (3) (2015) 580–587. doi:10.1109/taslp.2014.2379589.

[26] T. Nakashika, T. Takiguchi, Y. Ariki, Voice conversion using speaker-dependent conditional restricted boltzmann machine, EURASIP Journal on Audio, Speech, and Music Processing 2015 (1) (2015) 8. doi:10.1186/s13636-014-0044-3.

[27] P. Mowlaee, A. Sayadiyan, H. Sheikhzadeh, FDMSM robust signal representation for speech mixtures and noise corrupted audio signals, IEICE Electronics Express 6 (15) (2009) 1077–1083. doi:10.1587/elex.6.1077.

[28] H. Kawahara, I. Masuda-Katsuse, A. de Cheveign, 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) 187 – 207. doi:10.1016/S0167-6393(98)00085-5.

[29] R. Ramachandran, R. Mammone, Modern Methods of Speech Processing, 1st Edition, 0893-3405, Springer US, 1995. doi:10.1007/978-1-4615-2281-2.

[30] P. A. Naylor, N. D. Gaubitch, Speech Dereverberation, 1st Edition, Springer-Verlag London, 2010. doi:10.1007/978-1-84996-056-4.

[31] K. Paliwal, B. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, IEEE Transactions on Speech and Audio Processing 1 (1) (1993) 3–14. doi:10.1109/89.221363.

[32] D. A.Simovici, C. Djeraba, Mathematical Tools for Data Mining, 2nd Edition, Springer-Verlag London, 2014. doi:10.1007/978-1-4471-6407-4.

[33] R. Doost, A. Sayadiyan, H. Shamsi, A new perceptually weighted distance measure for vector quantization of the STFT amplitudes in the speech application, IEICE Electronics Express 6 (12) (2009) 824–830. doi:10.1587/elex.6.824.

[34] Y. Linde, A. Buzo, R. Gray, An algorithm for vector quantizer design, IEEE Transactions on Communications 28 (1) (1980) 84–95. doi:10.1109/TCOM.1980.1094577.

[35] Y. Hu, P. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on Audio, Speech, and Language Processing 16 (1) (2008) 229–238. doi:10.1109/TASL.2007.911054.

[36] R. C. Streijl, S. Winkler, D. S. Hands, Mean opinion score revisited: methods and applications, limitations and alternatives, Multimedia Systemsdoi:10.1007/s00530-014-0446-1.

[37] J. Kominek, A. W. Black, The CMU arctic speexh databases, in: Fifth ISCA Workshop on Speech Synthesis, 2004.