Novel Pre-processing using Outlier Removal in Voice Conversion

Sushant V. Rao, Nirmesh J. Shah, Hemant A. Patil

Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar
{sushant_rao,nirmesh88_shah,hemant_patil}@daiict.ac.in

Abstract

Voice conversion (VC) technique modifies the speech utterance spoken by a source speaker to make it sound like a target speaker is speaking. Gaussian Mixture Model (GMM)-based VC is a state-of-the-art method. It finds the mapping function by modeling the joint density of source and target speakers using GMM to convert spectral features framewise. As with any real dataset, the spectral parameters contain a few points that are inconsistent with the rest of the data, called *outliers*. Until now, there has been very few literature regarding the effect of outliers in voice conversion. In this paper, we have explored the effect of outliers in voice conversion, as a pre-processing step. In order to remove these outliers, we have used the score distance, which uses the scores estimated using Robust Principal Component Analysis (ROBPCA). The outliers are determined by using a cut-off value based on the degrees of freedom in a chi-squared distribution. They are then removed from the training dataset and a GMM is trained based on the least outlying points. This pre-processing step can be applied to various methods. Experimental results indicate that there is a clear improvement in both, the objective (*8 %*) as well as the subjective (*4 %* for MOS and *5 %* for XAB) results.

Index Terms: Voice conversion, outliers, Gaussian mixture model, robust principal component analysis.

1. Introduction

Voice Conversion (VC) is a technology that is used to modify the voice of one speaker (i.e., source speaker) to sound like that of another speaker (i.e., target speaker). This is done by modifying *prosodic* and *spectral* features of the source speaker [1]. The general framework of VC comprises of mainly two phases, namely, training and conversion. Depending on the training data available from the source and target speakers, VC can be broadly classified into text-dependent VC (for parallel data case [2]) and text-independent VC (for non-parallel or cross-lingual [2]). For text-dependent VC, first task is to align spectral features extracted from the source and target speakers' parallel utterances. It has been proved experimentally that alignment accuracy will impact the quality of speech in speech synthesis [3], [4] as well as in VC [5]. In the case of parallel data, Dynamic Time Warping (DTW) algorithm is used for alignment. However, in the presence of silent frames, the performance of DTW deteriorates [5]. In addition, the non-linear warping function of DTW will generate one-to-many and many-to-one mapping which will produce pairs that are undesirable for training. These pairs deviate away from the regular trend of the dataset (i.e., outliers). Outliers are the data points which are inconsistent with the dataset.

In this work, we explore the impact of outliers on the quality of VC. There are several methods for outlier detection. Earlier, statistical methods (such as Grubbs' method, etc.) were popular [6]. However, these methods increase the processing time where the data transformations are complex [6]. Proximity-based outlier detection methods include calculating nearest neighbours using the Euclidean or the Mahalanobis distance. They are easy to implement and need no initial assumptions about the data. The proximity-based outlier detection methods are suitable for a parallel dataset. Since, we have used a source and target parallel dataset, these methods would be ideal here. In particular, the Mahalanobis distance-based outlier detection has been implemented in order to detect and remove outliers in this paper.

Using the Mahalanobis distance on a high-dimensional dataset increases the computational cost and time [7]. To reduce this, we perform dimension reduction of the joint data matrix to a optimal number of components. Principal Component Analysis (PCA) is used for this purpose. The problem in performing the standard PCA is that it does not take into consideration the effect of outliers [7]. Hence, again, we have to perform PCA which is robust to such points. This is done by applying a Robust PCA (ROBPCA) instead of the standard (classical) one. The ROBPCA algorithm and its functioning will be explained in the later Sections. Performing PCA gives us the score matrix, which explains the high-dimensional data in a lower-dimensioned subspace. Using this, algorithm, we find the robust PCA scores and using these, we find the score distance of the data having a reduced dimension. This enables us to save computational time and cost of the analysis. After discarding the outlying points, mapping function can be learned for VC.

Several methods have been proposed for converting the spectral features of source speaker to target speaker. Among various statistical methods available such as Vector Quantization (VQ)-based VC [8], GMM-based VC [9], [10], [11], Partial Least Squares regression-based VC [12], [13], and Deep Neural Network (DNN)-based VC [14], [15], [16], etc., GMM-based VC is considered as the state-of-the-art method. On the other hand, dictionary-based methods, namely, Exemplar-based [17], [18] and Non-Negative Matrix Factorization [19], are also used. Source and target dictionaries are created based on the exemplars of speech. Here, dictionary mapping of the source and target is done.

Joint-density GMM (JDGMM) was first introduced in voice conversion by Kain *et. al.* in 1998 [9]. GMM-based VC method well transforms the overall gross spectral characteristics by means of weighted combination of linear transforms. However, finer details are not well transformed due to statistical averaging leading to deterioration of speech quality which is called the *oversmoothing* problem in a VC [20]. To overcome this problem, use of dynamic features and global variance (GV) enhancement techniques was proposed in [11]. Our proposed method will still suffer from *oversmoothing* and *over-*

fitting. However, the removal of outliers will provide a better statistical transformation of the source and target spectral features.

The JDGMM approach uses the mean and covariance between both the speakers' features to train the system. The problem here is that a dataset will contain outliers [21]. These few outlying points cause the mean and scatter to be shifted away from the ideal values as the GMM method relies mostly on mean and covariance in the conversion function, we might get errors in the converted voice. Hence, the need to robustify our mean and covariance matrix against such points is the motivation and objective of this paper. Using these robust estimates, we remove the outlying points from the training process by calculating the Mahalanobis distance of the scores and comparing each point with a threshold value [7]. Then, we build the classical GMM on the source and target joint feature vectors using the robust training data.

Rest of the paper is organized as follows: in Section 2, the proposed pre-processing technique is discussed in detail and covers the summary of the state-of-the-art joint density GMM method. Experimental setup is explained in Section 3 whereas Section 4 discusses the objective and subjective results obtained followed by the conclusion is Section 5.

2. Pre-processing using Outlier Detection

The GMM-based VC systems do not consider the effect of outliers while clustering the source or joint feature vectors into the specified number of components. Outliers are data points that deviate away from the general trend of the dataset [21]. These outliers affect the performance of the system as they seem to be inconsistent with the dataset. They tend to shift the mean and scatter of the data away from their ideal values, which affects the performance of the VC system. The removal of outlying observations helps in improving the quality of training. Outlier removal essentially reduces noise in the dataset, which might occur due to mechanical faults, changes in system behaviour, human error, instrument error, etc. [6]. Essentially, outlying points can be detected by using the Mahalanobis distance when the number of variables are more than two. Although this distance helps to detect outliers, it is susceptible to the masking effect in multivariate datasets. This effect occurs when a considerable number of outliers skew the mean and variance towards the cluster that results in the distance of the outlying points in the cluster to be below a threshold value [22]. Hence, we need robust estimators of location and scatter to apply in multivariate methods like PCA, regression and discriminant analysis.

Various methods are used to detect these outliers. Statistical and proximity-based techniques are typically used to detect the outlying observations [6]. With increasing dimensionality, statistical methods required more processing time along with the spreading of the convex hull. Hence, even though statistical methods were popular earlier, they suffered the problem [6]. Proximity-based techniques are much simpler to implement as compared to the statistical ones. The issue of large processing time still remains, however, the computational complexity is reduced. Proximity-based techniques detect the outliers based on distances. Either Euclidean or Mahalanobis distance is used for detection. The Mahalanobis distance gives the distance of a point to the centroid of the data. This method is more accurate and efficient for high-dimensional datasets than its Euclidean counterpart. Hence, we have used the Mahalanobis distance which is given as,

$$D_x = \sqrt{(x_i - \mu)^T C^{-1} (x_i - \mu)}, \tag{1}$$

where x_i is the i^{th} observation and C is the covariance matrix.

As dimensionality of the data increases, calculating the distance using Eq. 1 becomes expensive in terms of computational time. In addition, the convex hull will be very complicated [6]. Hence, dimensionality reduction is used to reduce the computation time, expense and to compact the convex hull too. We have used ROBPCA in place of the classical PCA since we need our system to be robust to the effect of outliers.

2.1. Robust Principal Component Analysis (ROBPCA)

Principal Component Analysis (PCA) is a statistical method used to extract relevant information in a high-dimensional dataset by using dimension reduction. The dimension reduction is done by looking at the covariance between the variables and projecting the information on less number of dimensions having the maximum covariances between them. As a result, PCA is often used as a first stage in the statistical analysis of multivariate data [22].

The standard PCA method first calculates the eigenvector having the largest eigenvalue of the covariance of the data. This is done because the dominant eigenvector shows the direction in which the data has the largest variance. After projecting the data on the first component, we decompose the covariance matrix. This matrix is used to calculate the dominant eigenvector for maximizing the covariance for finding the second component. The second component is *orthogonal* to the first component. This procedure continues till we project the data on the pre-selected number of components. The optimum number of components can be calculated using cross-validation techniques or selected manually. As mentioned earlier, data contains outlying points. The first components are found by maximizing the covariance and hence, they are more susceptible to include outlying points and fail to capture variance of the regular data. To avoid such limitations, we consider the data obtained, after removing outliers, to calculate the principal components. This method is called Robust Principal Component Analysis (ROBPCA) [7].

In this paper, we have used ROBPCA for reduction of dimension. ROBPCA is performed on the joint data $Z_{n,m} = (X_{n,p}, Y_{n,q})$. Here, n is the number of frames and p and q are the number of variables of the source and target speakers', respectively. ROBPCA gives the score and loading matrices. First, we apply Singular Value Decomposition (SVD) to find the affine transformation of the data. SVD is performed on the mean-centred data matrix. Here, the dominant eigenvector of the covariance matrix is not retained since this would imply we were performing the standard PCA. We find h least outlying points and use their covariance matrix to obtain a k_0 dimension subspace. The number of h least outlying points is calculated as $h = max\{[\alpha n], [(n + k_{max} + 1)/2]\}$, where the default value of k_{max} is *24* (since we have taken *24*-D MFCCs) and the parameter α*=0.75* [7]. Using this set of *h* data points, we find the location and scatter (i.e., mean and variance) of the data using the Minimum Covariance Determinant estimator. The objective of this estimator is to find a certain number of observations that have the lowest determinant of their covariance matrix. ROBPCA uses a computationally fast variant of the Minimum Covariance Determinant estimator, called the Fast Minimum Covariance Determinant algorithm [22]. The eigenvalues and their corresponding eigenvectors of the final robust scatter

matrix are calculated. The first k dominant eigenvectors are retained and they form the k-dimensional loading matrix $P_{p,k}$. This loading matrix, along with the final robust mean obtained earlier, is used to compute the scores using the formula,

$$T_n, k = (Z_{n,d} - 1_n \hat{\mu}')P_{n,k}. \quad (2)$$

The score matrix given in eq. 2 is used to calculate the score distances which further helps us in determining the outliers. It can be calculated as,

$$SD_i^{(k)} = \sqrt{(t_i^z - \hat{\mu}^z)^T (S^z)^{-1} (t_i^z - \hat{\mu}^z)}, \quad (3)$$

where $L_{k,k}^z$ is given by eigenvalues of $Z_{n,m}$ and $SD_{i(k)}$ is the score distance of the i^{th} frame when k is the number of components in ROBPCA [7].

To determine the outlying observations, we compare the score distance with the cut-off value $\sqrt{\chi^2_{k,0.975}}$ [7]. The frames that have distance more than this value are termed as outliers and hence, omitted from the training process. Since we assume our data to be normally distributed, the scores are also distributed normally [7]. The squared Mahalanobis distance of normally distributed data follow the χ^2_k distribution approximately. Figure 1 shows the scatter plot of the score distance calculated using eq.(3) and the cut-off line. The points having a score distance above this cut-off line are termed as outliers and removed from the dataset to be used for training. After the removal of outlier frames from the training dataset, we now fit a Gaussian Mixture Model (GMM) with a certain number of components on the outlier-free frames. We will use the state-of-the-art joint density GMM [9], for mapping of source and target speaker feature vectors.

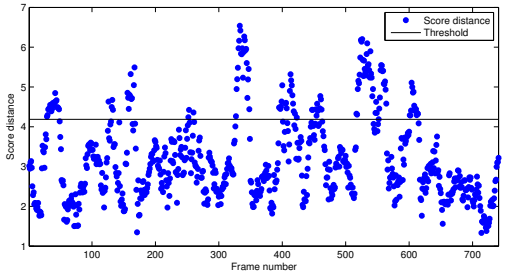

Figure 1: Outliers in male-to-male conversion.

2.2. Joint-density Gaussian Mixture Model (JD-GMM)

The main objective of Joint Density (JD) GMM-based VC technique is to find mapping function between source and target speakers' feature vectors. Joint density GMM is one of the most followed, state-of-the-art methods in VC. It has two phases, namely, training and testing (conversion). In the training phase, joint feature vectors, $z_i = [x_i; y_i]$ are used. Here, x_i and y_i are the time-aligned source and target feature vectors of the i^{th} observation [9], [11]. The joint probability density function (pdf) of the source and target is given by,

$$p(X, Y) = p(Z) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(z; \mu_m, \Sigma_m), \quad (4)$$

where α_m is weight of the m^{th} GMM component and $\mathcal{N}(z; \mu_m, \Sigma_m)$ is the normal distribution of the joint matrix Z with mean $\mu_m = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}$ and $\Sigma_m = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$ as the covariance matrix. The weights α_m are such that $\sum_{m=1}^{M} w_m = 1$ constraint for total probability. The conversion function for the joint density GMM method is given by,

$$\hat{y}_i = \sum_{m=1}^{M} w_m(x_i)(\mu_m^{(y)} + \Sigma_m^{(yx)}(\Sigma_m^{(xx)})^{-1}(x_i - \mu_m^{(x)})), \quad (5)$$

where $w_m = \frac{\alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx})}{\sum_{l=1}^{M} \alpha_m \mathcal{N}(x; \mu_l^x, \Sigma_l^{xx})}$ and x_i and $\hat{y}_i$ are the i^{th} frame of the source and converted feature vector, respectively. This conversion mapping function is obtained by using the minimum mean square error (MMSE). While training the system, the model parameters $\alpha_m, \mu_m^{(z)}, \Sigma_m^{(z)}$ are estimated using the Expectation Maximization (EM) algorithm which fits the data to a GMM [23].

3. Experimental Setup

The experiments were conducted on the joint-density GMM (JD-GMM) model, both with and without outlier removal. The number of components *(i.e., k=16)* for ROBPCA was estimated using the eigenvalue outlyingness method (gives scatter of the data) [7]. The least number of components which explain over *90 %* of the variance in the data was chosen as the required k value. The outlyingness for the number of GMM components used in experimentation reached *90 %* for *15*, *16* and *17* ROBPCA components variably. Hence, we have set the value of k to *16*. JD-GMM with *8, 16, 32, 64* and *128* components was performed and evaluated objectively. JD-GMM with-*32* components gave the least errors. Hence, it was used to test all the four systems.

The CMU-ARCTIC database was used for experimentation (in particular, *bdl and rms* for male and *clb and slt* for female were used) [24]. The source and spectral features were extracted using AHOCODER [25] analysis-synthesis framework. The prosodic features were log (F_0) and spectral features were *24* order Mel Frequency Cepstral Coefficients (MFCCs). Before training, the feature vectors of the source and target were time-aligned using the Dynamic Time Warping (DTW) algorithm. The number of frames used for training were *9,342, 16880* and *32,159* from *10, 20* and *40* training sentences, respectively. Testing was performed on four speaker pairs, namely, male-to-male (M-M), male-to female (M-F), female-to-male (F-M) and female-to-female (F-F)

The prosodic features (i.e., F_0 here) was converted using the mean-variance method. The conversion formula is given as,

$$\hat{y}_t = \frac{\sigma^y}{\sigma^x}(x_t - \mu^x) + \mu^y, \quad (6)$$

where μ and σ are the mean and standard deviation of the source and target log-scaled F_0 values.

4. Experimental Results

In order to evaluate, both objective and subjective measures were used. For the objective evaluation, we used Mel Cepstral Distortion (MCD) whereas MOS and ABX tests were conducted to evaluate the voice conversion systems subjectively.

Objective evaluation: We calculated the MCD between the

original target and the converted sentences, which is given by,

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{24}(c_i - \hat{c_i})^2}, \quad (7)$$

where c_i and $\hat{c_i}$ are the i^{th} coefficients of the original target and converted MFCC features. Using *10, 20* and *40* training sentences, *50* testing sentences were converted using *8, 16, 32, 64* and *128* GMM components for the objective evaluation. The testing sentences were those which were not used in the training process.

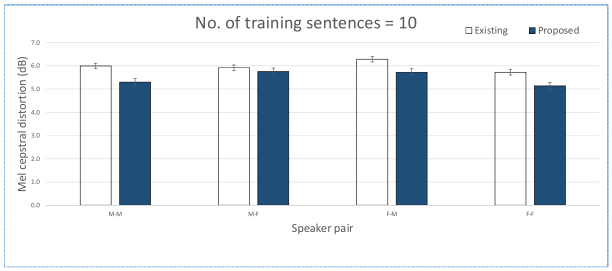

Figure 2: MCD analysis with *95 %* confidence interval for *10* training utterances.

Figure 2, Figure 3 and Figure 4 shows the comparison of the existing joint-density GMM method without outlier removal and the proposed method with outlier removal in the same GMM technique. It shows an improvement in the Mel cepstral distortion values as compared to the baseline approach across all the cases. Only in the case of F-M conversion, the proposed approach has higher distortion value than the existing one. In figure 2, we have the plot for the MCD values of the

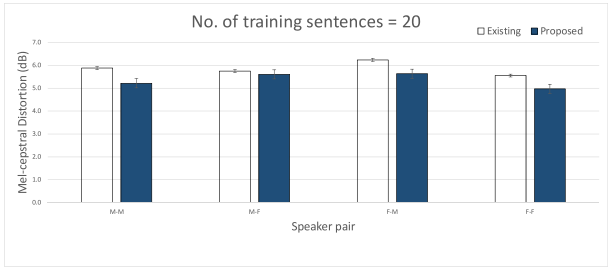

Figure 3: MCD analysis with *95 %* confidence interval for *20* training utterances.

four speaker pair systems using 10 sentences for training. Removing the outliers seem to have a positive impact on the MCD. The male-to-male and female-to-female voice conversions have shown almost a *10%* improvement as against GMM without any preprocessing. In the case of female-to-male voice conversion, a rise in the MCD value has been observed.

Even when the training sentences are increased to *20*, we get the results similar to what we obtained when *10* training sentences were used. This is evident from figure 3. Here, the same-gender voice converison performs better objectively, whereas without the pre-processing step, the cross-gender voice conversion has lower MCD values. From figure 2 and figure 3, it is observed that using *10* and *20* sentences to train the system we get over a *10 %* improvement for the M-M and M-F systems,

Figure 4: MCD analysis with *95 %* confidence interval for *40* training utterances.

while the results are comparable when we increase the number of sentences to *40* as evident from Figure 4. Although we get a comparable result (*2 %* increase) in the M-F system, the MCD in the F-M conversion decreases in the proposed approach, irrespective of the number of training sentences. Only in the case of F-F, we observe that the improvement is consistent over all the training sets.

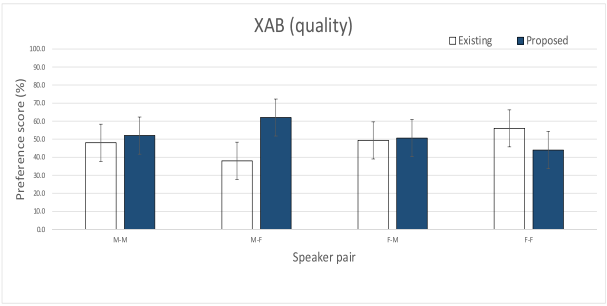

Figure 5: XAB test for quality with *95 %* confidence interval.

Subjective evaluation: Listening tests were conducted to evaluate the converted speech qualitatively. *30* listeners participated in the subjective evaluation. There were *18* male and *12* female listeners (having age between 22-28 years) who have no hearing impairment. Two parameters were evaluated while performing the tests, namely, speech quality and speaker similarity. To evaluate the speech quality, we presented converted sentences to listeners in a random order from both the methods and asked them which of the sentences sounded more natural.

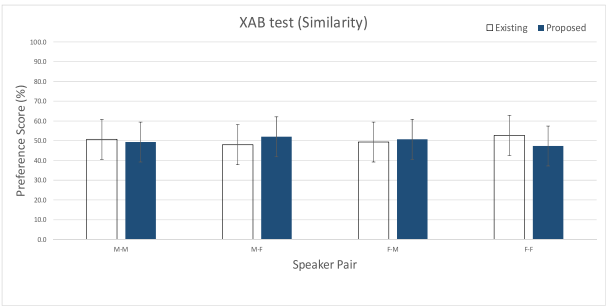

Figure 6: XAB test for similarity with *95 %* confidence interval.

Speaker similarity was tested using XAB test. Here, X was presented as the analysis-synthesized target speech and speech

converted using the baseline and the proposed approach were presented randomly as A or B. In this test, listeners were asked to provide a preference the converted speech signal, A or B, that sounded similar to X. The number of mixture components was set to *32* as it provided with the least error. We also performed the Mean Optimum Score (MOS) test to evaluate both, quality and similarity parameters, of the converted speech sounds [26]. In this test, the listeners were asked to rate the voice converted speech for naturalness on a scale of *1* to *5*. An average of these scores was then calculated for all the four speaker pairs. The speaker similarity parameter was also tested using the MOS test. Here, the listeners had to compare the voice converted sound to a reference (analysis-synthesized) sound and rate on a scale of *1* to *5*. The standard values for the MOS test are **1** - bad, **2** - poor, **3** - fair, **4** - good, **5** - excellent.

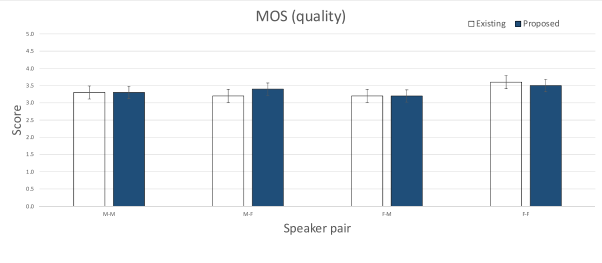

Figure 7: MOS test for quality with *95 %* confidence interval.

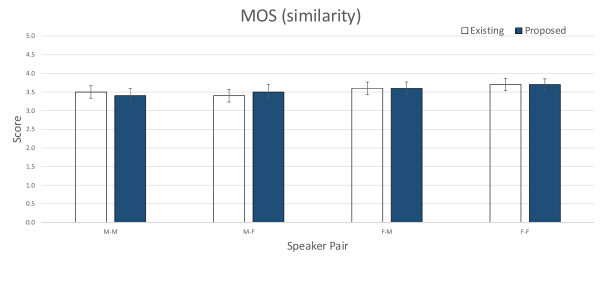

Figure 8: MOS test for similarity with *95 %* confidence interval.

The results of the XAB and MOS test for quality and similarity are shown in figure 5, 6, 7 and 8, respectively. We observed that our proposed approach was better in terms of both speech quality and similarity in the cross-gender conversion, while it showed comparable performance when the source and target speaker gender was the same. In contrast, the objective results showed a significant improvement in the same-gender voice conversion systems while, the MCD for F-M voice conversion was higher when the pre-processing step of outlier removal was used. In Figure *7* and Figure *8*, the proposed pre-processing step has a considerably higher performance for the male-to-female and female-to-male voice conversions. As opposed to the objective results, the male-to-male and female-to-female have a lower or comparable preference score.

5. Summary and Conclusions

We have proposed a pre-processing method that helps to remove inconsistent data points before training. We saw improvement in the objective as well as subjective evaluation when outliers were discarded from the dataset over the joint density method without removing the outliers. To save computational time, cost and increase the efficiency, we used ROBPCA for dimension reduction and calculated the corresponding score distance using the robust (free from outlier effect) mean and scatter of the data. This pre-processing step is very effective, since it eliminates the unwanted data points so that the system is trained in a way that it understands the true regularity of the dataset and is not influenced by a few irregular observations. This pre-processing will be useful in many statistical methods, such as multivariate regression, used in voice conversion. Furthermore, we can also incorporate the orthogonal distances along with the score distances to remove those points which lie outside the threshold of both the distances. This will assist in eliminating the masking effect of outliers that occurs while using the Mahalanobis distance *30* listeners (*18* male and *12* female listeners who have no hearing impairment) participated in the subjective evaluation. We can also extend this proposed pre-processing method to technologies beyond voice conversion. It can be used in speech and speaker recognition systems where statistical methods are often used.

6. Acknowledgments

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India, for sponsored project, Development of Text-to-Speech (TTS) System in Indian Languages (Phase-II) and the authorities of DA-IICT Gandhinagar. We also thank all the participants who took part in subjective evaluations.

7. References

[1] Y. Stylianou, "Voice transformation: a survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei,Taiwan, 2009, pp. 3585–3588.

[2] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A first step towards text-independent voice conversion," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, 2004.

[3] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, "Effectiveness of PLP-based phonetic segmentation for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014, pp. 270–274.

[4] M. Zaki, J. N. Shah, and H. A. Patil, "Effectiveness of multiscale fractal dimension-based phonetic segmentation in speech synthesis for low resource language," in *International Conference on Asian Language Processing (IALP)*, Kuching, Borneo Malaysia, 2014, pp. 103–106.

[5] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 1–5.

[6] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[7] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. New York, NY, USA: IEEE, 1988, pp. 655–658.

[9] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, 1998, pp. 285–288.

[10] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[11] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[12] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[13] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 20, no. 3, pp. 806–817, 2012.

[14] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[15] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *IEEE Spoken Language Technology Workshop (SLT)*, Nevada, USA, 2014, pp. 19–23.

[16] S. H. Mohammadi and A.Kain, "Semi-supervised training of a voice conversion mapping function using a joint-autoencoder," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 1–5.

[17] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, 2013, pp. 201–206.

[18] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[19] R. AIHARA, T. TAKIGUCHI, and Y. ARIKI, "Semi-non-negative matrix factorization using alternating direction method of multipliers for voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5170–5174.

[20] T. Toda, H. Saruwatari, and 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)*, Salt Lake City, UT, USA, 2001, pp. 841–844.

[21] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Springer, 2005, vol. 2.

[22] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[24] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[25] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved hnm-based vocoder for statistical synthesizers." in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.

[26] I. Rec, "P. 85. a method for subjective performance assessment of the quality of speech voice output devices," *International Telecommunication Union (ITU), Geneva.*, Available Online:{https://www.itu.int/rec/T-REC-P.85-199406-I/en} Last Accessed{July 26, 2016}.