Experimental investigation on the efficacy of Affine-DTW in the quality of voice conversion

Gaku Kotani, Hitoshi Suda, Daisuke Saito and Nobuaki Minematsu
Graduate School of Engineering, The University of Tokyo, Japan
E-mail: {kotani, hitoshi, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—In this paper, the performance of Affine-DTW, which performs appropriate time alignment between source and target features in voice conversion (VC), is experimentally and thoroughly investigated. In traditional VC, parallel data are often required to train a mapping model between source and target features. While VC with non-parallel data is also studied to avoid collecting parallel data, the quality of its converted speech is still inferior to the traditional one with parallel data. One approach to further progress in VC is exploiting both parallel and non-parallel data, the former of which is pre-stored and the latter of which is assumed to be easily collected. In this case, it is still worthwhile to study time-alignment techniques to obtain appropriate alignment of parallel data. Affine-DTW is a technique in which dynamic time warping (DTW) and coarse conversion based on affine transformation are iteratively performed. In Affine-DTW, time alignment and parameters of affine transformation can be analytically calculated so that it can be easily adopted as pre-processing in VC. However, the influence on the performance of trained models based on the obtained alignments has not been well investigated experimentally. Hence, this paper investigates the performance of Affine-DTW in terms of quality improvement of converted speech in traditional VC methods based on Gaussian mixture models, non-negative matrix factorization and neural networks. Experimental results show that Affine-DTW obtains appropriate alignments and the naturalness improvement of converted speech in subjective assessments is observed in trained models based on the alignments.**

I. Introduction

Voice conversion (VC) is a technique of modifying non-linguistic information of a speech utterance while maintaining its linguistic information [1]. The modification technique can be applied to various applications such as converting speaker identity of output speech of text-to-speech synthesis and so on [2], [3], [4].

In many traditional VC, parallel data is required to learn a mapping model from source to target features, which is constructed from pairs of utterances with the same linguistic content from both source and target speakers. For the mapping models, Gaussian mixture models (GMM), deep neural networks (DNN) and non-negative matrix factorization (NMF) are widely used [2], [5], [6]. In the case of frame-by-frame mapping, alignment between source and target sequences is needed and usually it is obtained by dynamic time warping (DTW) [1]. VC with parallel data has an advantage that the statistical model only needs to focusing on learning the mapping between different speaker identities. On the other hand, VC with non-parallel data is studied to avoid collecting parallel data and mismatch of frame pairs in alignment [7], [8], [9]. In VC with non-parallel data, however, the quality of its converted speech is still inferior to the traditional one with parallel data. One approach to further progress in VC is exploiting both parallel and non-parallel data, the former of which is pre-stored and the latter of which is assumed to be easily collected. In this case, it is still worthwhile to study time-alignment techniques to obtain appropriate alignment of parallel data.

In VC with parallel data, DTW is most widely used for time alignment. DTW minimizes the mean square error between source and target feature vector sequences. That is, the difference in the mel-cepstral coefficients is the criterion for alignment. Naive DTW assumes that the distance between two frames containing the same linguistic information is close, but not in practice, especially in the case of cross-gender conversion. Since mel-cepstral coefficients represent not only linguistic information but also speaker identity, the obtained alignment can be affected by these represented information. To overcome this defect, iterative alignment techniques or implicit alignment techniques are studied [10], [11], [12]. Affine-DTW is one of the iterative alignment techniques [10]. Affine-DTW performs as following five steps: (1) performing general DTW, (2) estimating the parameters of affine transformation based on the obtained alignment, (3) applying the estimated affine transformation to source features, (4) performing DTW between the transformed source and the original target feature vector sequences and (5) the three steps (2–4) are iterated. Given the influence on alignment by the difference of speaker identities, it can be natural to iterate alignment between converted features and training models based on the alignment. If we use GMM or DNN for the conversion model, however, it does not work well. In other words, the more complex mapping model the more easily over-fitting based on the rough alignment.

In Affine-DTW, time alignment and parameters of affine transformation are analytically calculated so that it can be easily adopted as pre-processing in VC. However, the influence on the performance of trained models based on the obtained alignments has not been well investigated experimentally. Hence, this paper investigates the performance of Affine-DTW in terms of quality improvement of converted speech. Experimental results show that Affine-DTW obtains appropriate alignments and the naturalness improvement of converted speech in subjective assessments is observed in trained models based on the obtained alignments.

II. Related works

In traditional VC with parallel data, time alignment between source and target utterances is needed. One of the major techniques for time alignment is DTW which obtains frame-by-frame time-alignment between them by dynamic programming [1]. DTW assumes that the distance between two feature vectors containing the same linguistic information is close, even if they are derived from different speakers. The assumption, however, is not in practice so that some approaches have been proposed to overcome the issue. One approach is to avoid utilizing parallel data, i.e. VC with non-parallel data. In non-parallel VC without additional transcribed data, iterative adaptation or representation learning are studied [7], [9]. They are technically important but the quality of their converted speech is still inferior to VC with parallel data. With additional transcribed data, phonetic posteriorgrams obtained from a recognizer are often used as speaker-independent representation of features [13], [12]. Another approach is not to align frame by frame externally, i.e. sequence-to-sequence modeling [11], [14]. This approach can avoid frame mismatch but requires complex procedure of training models or much training data. DTW is superior to the above methods in its easy implementation and computation, and obtaining appropriate time alignment is important to exploit parallel data efficiently. To obtain appropriate alignment by DTW, Affine-DTW has been proposed, which iterates DTW and coarse conversion based on affine transformation between source and target features [10]. In the next section, we explain Affine-DTW in detail.

III. Affine-DTW

Affine-DTW assumes that the conversion between source and target mel-cepstrum coefficients can be coarsely represented as global affine transformation. The coarse conversion modeling has an affinity with the fact that the difference in vocal tract length is represented as linear transformation in cepstral space [15]. In addition, parameters of affine transformation are analytically calculated so that it is convenient as pre-processing of VC. Affine-DTW performs as following five steps: (1) performing general DTW, (2) estimating the parameters of affine transformation based on the obtained alignment, (3) applying the estimated affine transformation to source features, (4) performing DTW between the transformed source and the original target feature vector sequences and (5) the three steps (2–4) are iterated. A sequence of feature vectors from an utterance of a source speaker is defined as $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_M]$, while $\boldsymbol{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N]$ represents that from an utterance of a target speaker. In naive DTW, a warping path $\boldsymbol{W} = \boldsymbol{w}_1, .., \boldsymbol{w}_K$ aligns between source and target feature vectors. That is, $\boldsymbol{W}$ is a sequence of grid points, where $\boldsymbol{w}_k$ corresponds to a point (m, n), the m and n of which indicate frame indices in source and target sequences, respectively. The distance between two frames from source and target sequences is represented as $\delta(m, n) = \delta(\boldsymbol{w}_k)$, and then the warping path obtained by naive DTW is shown as follows

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{k=1}^{K} \delta(\boldsymbol{w}_k). \tag{1}$$

Note that the constraints of transition of grid points and the weights of local path are applied.

In Affine-DTW, parameters of affine transformation $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{b}\}$ are introduced and the distance is represented as follows

$$\delta_{\boldsymbol{\theta}}(m, n) = (\boldsymbol{y}_n - (\boldsymbol{A}\boldsymbol{x}_m + \boldsymbol{b}))^2, \tag{2}$$

In the step of coarse conversion of Affine-DTW, the parameters are calculated based on the criterion of minimization of the mel-cepstrum distortion between source and target sequences, which is shown as follow

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{K} \sum_{k=1}^{K} \delta_{\boldsymbol{\theta}}(\boldsymbol{w}_k). \tag{3}$$

Note that the estimation of the parameters is not performed from an utterance but the whole training data set. In the step of DTW, except for consideration of weights of local paths, DTW minimizes almost the same distortion, which is shown as follows

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{k=1}^{K} \delta_{\boldsymbol{\theta}}(\boldsymbol{w}_k). \tag{4}$$

This is, the distortion between source and target sequences is expected to decrease along with the number of iterations in Affine-DTW. In this paper, we confirm experimentally that it decreases and converges and Affine-DTW improves the quality of VC based on GMM, NMF and DNN.

IV. Experimental evaluations

A. Experimental conditions

To evaluate the performance of Affine-DTW, subjective evaluations were carried out. We compared the quality of converted speech of each trained model based on an alignment which was obtained by i-th iteration of Affine-DTW ($i = 0, 1, ..., 10$). For the comparison, preference tests about naturalness and speaker identity of converted speech were conducted, in which converted speech of two trained models based on each alignment obtained from i-th and j-th iterations of Affine-DTW were compared, e.g. 0-th (non-affine) vs. 1-st iterations or 2-nd vs. 4-th iterations.

For mapping models, three models were adopted, which were GMM, DNN and NMF. For GMM-based VC, a joint density GMM with cross-diagonal covariance was used [2]. The number of mixtures was 512 and training of GMM was performed until convergence of its likelihood. For DNN-based VC, feed-forward NN was used [5]. The number of hidden layers was 4, and each of them had 1024 units. The activation function of hidden and output layers were LeakyReLU and identity mapping, respectively. As an optimization method of feed-forward NN, AMSGrad with a learning rate of 0.01 was

used [16]. The training epochs were repeated 3 times, and the batch size is 512 in each epoch. In both GMM-based and DNN-based VC, input and output features were mel-cepstrum coefficients from the 1-st to 39-th and dynamic features were exploited. For the the generation of the static features, the Maximum Likelihood Parameter Generation (MLPG) algorithm was used [17]. In NMF-based VC, the conversion is implemented in spectral space, i.e. NMF is trained to map between source and target spectral sequences based on time alignment in cepstral space [18]. The size of bases was 200 and parameters of NMF was iteratively updated until convergence. As spectral features, spectra from the 1-st to 512-th bins were used. In addition, fundamental frequency was converted based on global linear transformation in all the cases.

The ATR Japanese speech dataset B-set were used as two source and target pairs, which were male-to-male and male-to-female pairs (MHT to MMY and MHT to FKS) [19]. From the dataset, subset I and J of phoneme-balanced sentences were used for validation and testing, respectively. For training, subset A and B were used. Each subset had about 50 sentences. Speech signals were sampled at 20 kHz. Feature vectors were extracted with a 1-ms shift and the feature vector consisted of spectra of 513 bins or the 0-th through 39-th mel-cepstrums, which were derived from WORLD analysis [20] (D4C edition [21]). In Affine-DTW, parameters of affine transformation were analytically calculated only with training set.

In the preference tests, the numbers of subjects were 24–27 for each test. For each naturalness evaluation, AB test was conducted in which 20 or 30 pairs of two converted speech were suggested and each subject chose one which sounded more natural speech. For each speaker identity evaluation, in a similar manner to naturalness one, ABX test was conducted. In the test, two converted speech and reference speech were suggested and each subject chose converted one which sounded more similar to the reference one in terms of speaker identity.

B. Experimental results

The results show that Affine-DTW obtains appropriate alignments and the naturalness improvement of converted speech is observed in trained models based on the alignments. Fig. 1 shows mel-cepstrum distortion (MCD) between converted features from source ones based on affine transformation and original target ones, along with the number of iterations of Affine-DTW. In the figure, MCD is getting to convergence along with the number of iterations. The alignments which are obtained by each iteration of Affine-DTW are shown in Fig. 2. In both male-to-male (M2M) and male-to-female (M2F) cases, alignments obtained by 0-th and 1-st iterations of Affine DTW are obviously different but ones obtained by i-th and j-th ($i, j > 1$) are not so different. From this result, it may be said that 1-st iteration of Affine-DTW is sufficient, but it is not in subjective evaluations. The results of subjective evaluations are shown in Fig. 3. In GMM-based VC, alignment obtained by 1-st iteration of Affine-DTW (Affine-1) is superior to one obtained by naive DTW (the 0-th iteration of Affine-DTW) in both M2M and M2F cases.

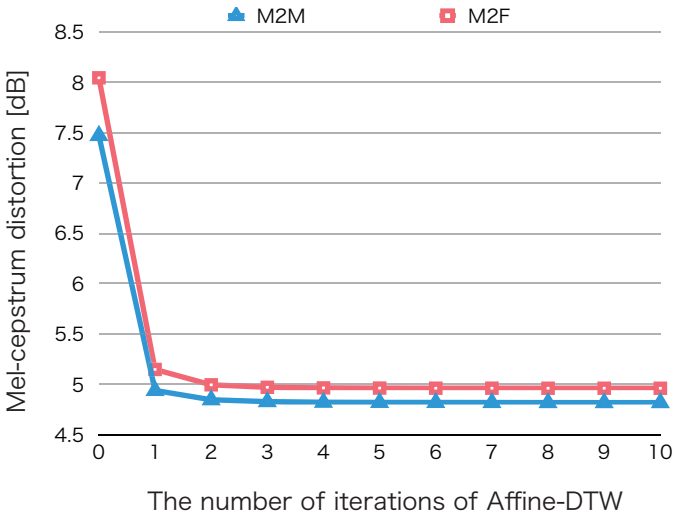

Fig. 1. Mel-cepstrum distortion (MCD) between converted features from source ones based on affine transformation and original target ones, along with the number of iterations of Affine-DTW. In both male-to-male (M2M) and male-to-female (M2F) cases, the convergence of MCD along with the number of iterations is observed.

Although Affine-i and Affine-j ($i, j > 1$) are comparable in both cases, in the comparison between Affine-1 and Affine-10, in case of M2F Affine-10 has naturalness improvement and they are comparable in the other case. The result can indicate that Affine-DTW improves the quality of GMM-based VC until the convergence of its alignment.

The other results, in cases of NMF-based and DNN-based VC, also show almost the same but there is some room for discussion. In case of M2M-NMF, the result is interesting because Affine-2 is inferior to Affine-1 while Affine-1 is superior to naive DTW, it is not in GMM-based VC. The reason is probably the qualitative difference between the two methods. NMF-based VC is briefly regarded as the replacement of spectral bases based on alignment. In other words, it can be said that NMF-based VC is sensitive to the difference of alignments. In the process of iteration of Affine-DTW, each updated alignment may not be always more appropriate than previous one, and the alignment in the middle of the process to the convergence can affect the performance of NMF. The result of M2M-DNN is also interesting. In the result, Affine-1 and naive DTW are comparable while in the other cases Affine-1 is superior to naive DTW. Additionally, in the case of M2M-DNN Affine-10 is superior to Affine-0, while in the other methods the number of iteration does not effect on the results of M2M. One reason for this incomprehensible result can be the problem of tuning hyper-parameters of DNN. In our experiments, the hyper-parameters of DNN and random seed of its implementation are fixed. The difference of alignments, however, change the size of training dataset so that the training procedure is different among each trained DNN. In addition, the optimal number of iterations of Affine-DTW can depend on mapping models because the coarse conversion in Affine-DTW is regarded as GMM-based VC

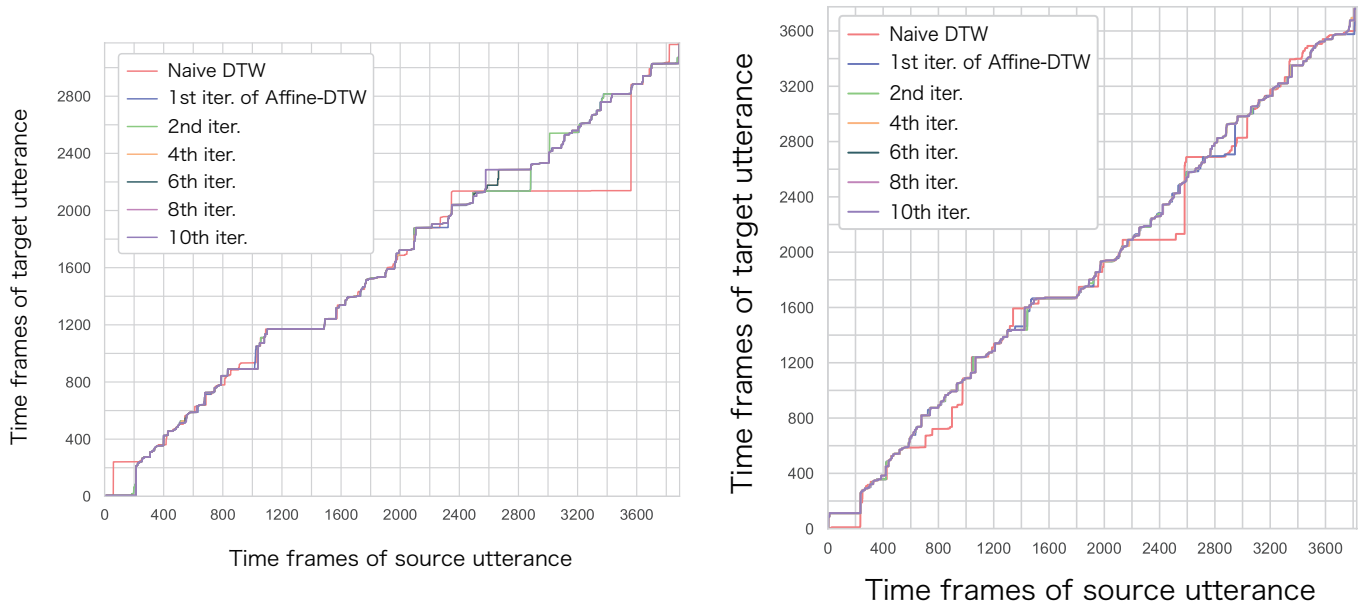

Fig. 2. Visualizations of examples of time alignments obtained by naive DTW and Affine-DTW. The left figure depicts time alignments between utterances of male and male, the other depicts that of male and female. In both figures, the difference of non-affine alignment and the other alignments obtained by Affine-DTW is clearly shown, while the differences among ones by Affine-DTW are small. Although the differences can not be confirmed in the figures, the alignments are changed along with the number of iterations of Affine-DTW.

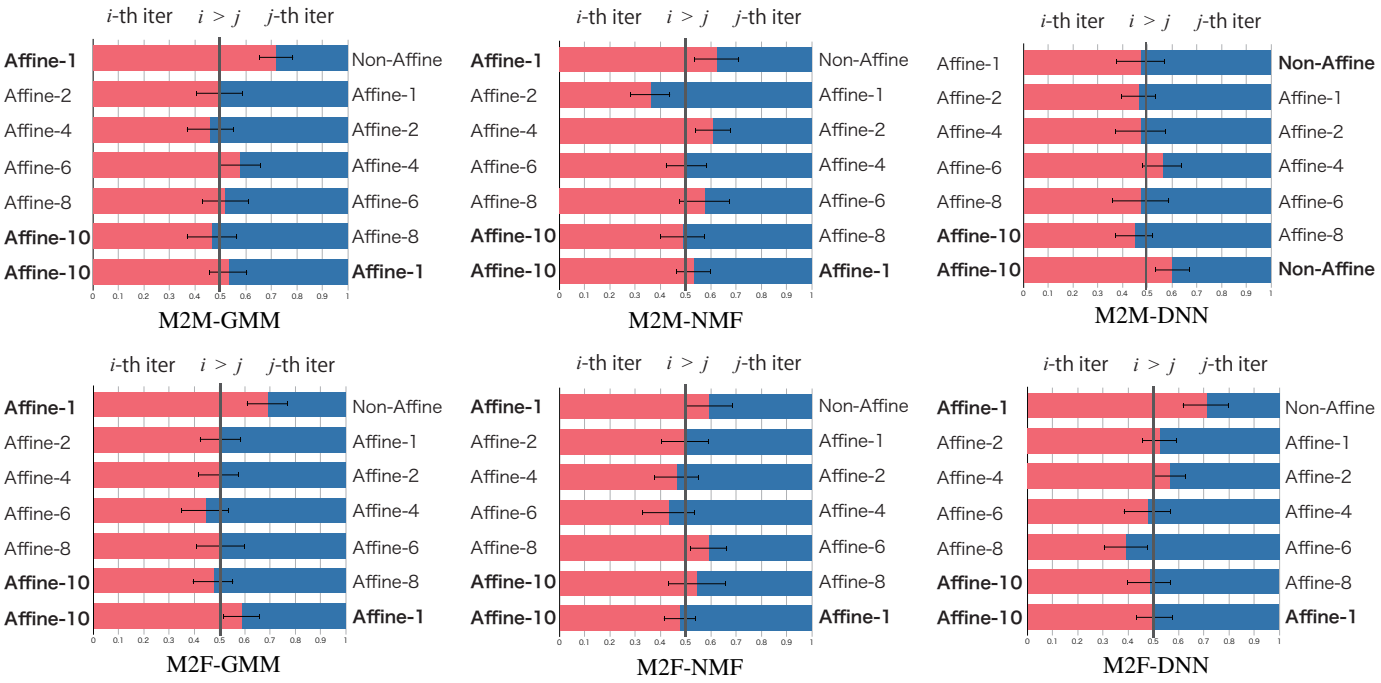

Fig. 3. Results of subjective evaluations on naturalness of converted speech of GMM-based, NMF-based and DNN-based VC, which are trained with the alignments obtained by Affine-DTW of each number of iterations. Note that only in M2M-DNN Affine-1 was comparable to None-Affine so the comparison between Affine-10 and None-Affine was performed.

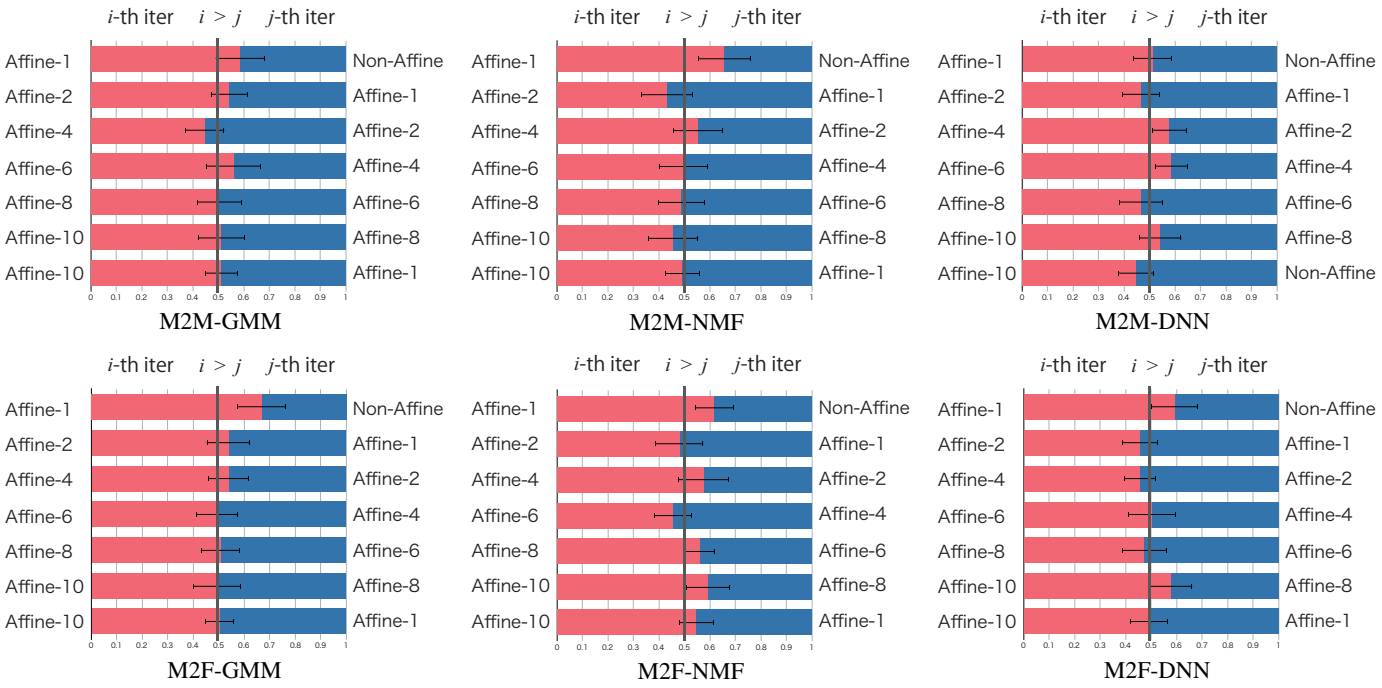

Fig. 4. Results of subjective evaluations on speaker identity of converted speech of GMM-based, NMF-based and DNN-based VC, which are trained with the alignments obtained by Affine-DTW of each number of iterations.

which has a single component of Gaussian. In other words, the assumption that the conversion between source and target mel-cepstrum coefficients can be coarsely represented as global affine transformation is advantageous for GMM-based VC.

In summary, Affine-DTW improves the quality of VC based on GMM, NMF and DNN with parallel data, while the optimal number of its iterations depends on pairs of source and target speakers and possibly mapping models, but the distortion between converted from source and original target features in Affine-DTW decreases along with the number of iterations.

V. Conclusions

This paper has experimentally investigated the performance of Affine-DTW which is a technique to obtain appropriate time alignment, in the quality of voice conversion with parallel data. Affine-DTW iterates DTW and coarse conversion based on affine transformation. In our experiments, the performance of Affine-DTW has been investigated in subjective evaluations with traditional GMM-based, NMF-based and DNN-based methods. The results have shown that Affine-DTW improves the quality of VC, while the optimal number of its iterations depends on pairs of source and target speakers and possibly mapping models. Affine-DTW is a simple pre-processing of VC so that it can be easily adopted in many VC methods with parallel data. One of our future works is to obtain more appropriate time alignment, e.g. by utilizing more complex conversion model in the conversion step of Affine-DTW. In this case, we must care about the relevance of over-fitting of conversion models and appropriateness of alignment. Finally, we claim that in the comparison between methods with parallel and non-parallel data, we should be careful not to neglect the minimum effort to obtain appropriate alignment.

Acknowledgment

This research and development work was supported by the MIC/SCOPE #182103104.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 655–658.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 285–288.

[3] H. Kawanami, Y. Iwami, T. Toda, H. Hiroshi, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proceedings of European Conference on Speech Communication and Technology*, 2003, pp. 1–4.

[4] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1999, pp. 837–840.

[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3893–3896.

[6] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proceedings of ISCA Workshop on Speech Synthesis*, 2013, pp. 201–206.

[7] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[8] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[9] C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–6.

[10] H. Suda, G. Kotani, S. Takamichi, and D. Saito, "A revisit to feature handling for high-quality voice conversion based on Gaussian mixture model," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018, pp. 816–822.

[11] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, "Spectral conversion based on statistical models including time-sequence matching," in *Proceedings of ISCA Workshop on Speech Synthesis*, 2007, pp. 333–338.

[12] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proceedings of International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[13] L. M. Arslan and D. Talkin, "Speaker transformation using sentence hmm based alignments and detailed prosody modification," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 1, 1998, pp. 289–292.

[14] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6805–6809.

[15] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[16] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proceedings of International Conference on Learning Representations*, 2018.

[17] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 1877–1884, 2007.

[18] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proceedings of Spoken Language Technology Workshop*, 2012, pp. 313–317.

[19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357 – 363, 1990.

[20] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[21] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57 – 65, 2016.