Voice Conversion Strategies for Parallel and Nonparallel Data Cases

Nirmesh J. Shah

Speech Research Lab, DA-IICT, Gandhinagar, India

nirmesh88_shah@daiict.ac.in

1. Introduction

Voice Conversion (VC) is a technique that converts the perceived speaker identity in a given speech signal from a source speaker to a target speaker without affecting the linguistic content of the utterance [1]. The basic framework of VC first extracts speaker-specific information (namely, *spectral* and *prosodic* features (especially speaking style)) from both the source and the target speakers. Learn the mapping function between corresponding feature pairs and predict the converted features. These features are again converted back to the speech signal using VOCODER. VC broadly can be categorized into parallel (if both the speakers have spoken the same utterances) and non-parallel cases (if both the speakers have spoken different utterances from a same language or different language) (for non-parallel corpus) [2].

Even if both the speakers have spoken the same utterances, the spectral features from both the source and the target speakers need to be aligned during training due to a speaking rate variation across the speakers (i.e., interspeaker variations) and a speech rate variations within the speaker (i.e., intraspeaker variations) [3, 4]. So is the case for non-parallel VC. One of the major focuses of this thesis is to identify issues related to the alignment on the quality of converted voices [5, 6]. In addition, we also identified the impact of outliers and hence, their removal on the quality of converted voices [7, 8]. Furthermore, we also proposed convergence theorem for one of the popular alignement techniques for non-parallel VC case, namely, **I**terative combination of **N**earest Neighbor search step and **C**onversion step **A**lignment (INCA) algorithm and its extension in the form of dynamic INCA algorithm [9, 10]. Moreover, we also proposed to use Vocal Tract Length Normalization-based warped features for avoiding the need of alignment stage in the case of non-parallel VC [11].

Apart from this, we proposed novel Amplitude Scaling (AS)-based method for BiLinear Frequency Warping (BLFW)-based mapping techniques for the VC [12]. Furthermore, novel Generative Adversarial Network (GAN)-based model with Minimum Mean Square Error (MMSE) Regularizer has also been proposed, which is used for the VC as well as Non-Audible Murmur-to-WHispered SPeech (NAM2WHSP) conversion task [13]. In my thesis, issues related to the objective evaluations were also identified and novel objective measure, based on acoustic-to-articulatory inversion technique was also proposed [14].

2. Major Contributions

Brief summary of each contribution along with its motivation and key results are presented in the next Sub-Section.

2.1. Impact of Alignment

Given a source and target speakers' parallel training speech database (in the parallel data VC case), first task is to align source and target speakers' spectral features at frame-level before learning the mapping function. The accuracy of alignment will affect the learning of mapping function and hence, the voice quality of converted voice in VC. The impact of alignment is not much explored in the VC literature [15]. Most of the alignment techniques try to align the acoustical features (namely, spectral features, such as Mel Cepstral Coefficients (MCC)). However, spectral features represents both speaker as well as speech-specific information. In this work, we have done analysis on the use of different speaker-independent features (namely, unsupervised posterior features, such as Gaussian Mixture Model (GMM)-based and Maximum A Posteriori (MAP) adapted from Universal Background Model (UBM), i.e., GMM-UBM-based posterior features) for the alignment task [5]. In addition, we proposed to use different metrics, such as symmetric Kullback-Leibler (KL) and cosine distances instead of Euclidean distance for the alignment [5]. Our analysis-based on % Phone Accuracy (PA) is correlating with the subjective scores of the developed VC systems with 0.98 Pearson correlation coefficient [5].

2.2. Outliers removal

Few corresponding pairs that are obtained after the alignment are inconsistent with the rest of the data called as *outliers*. These outliers shift the parameters of the mapping function from its true value and hence, affect the learning of the mapping function during the training phase of the VC task. We have proposed the effectiveness of the outliers removal as a pre-processing step in the VC [7]. The proposed method uses a score distance that is estimated using Robust Principal Component Analysis (ROBPCA) to detect the outliers. In particular, the outliers are determined using a fixed cut-off on the score distances, based on the degrees of freedom in a chi-squared distribution [16], which is speaker pair-independent. The fixed cut-off is due to the assumption that the score distances follow the normal (i.e., Gaussian) distribution, which is a weak statistical assumption even in the cases where quite a large number of samples are available [17]. Hence, we proposed to explore speaker pair-dependent cut-offs to detect the outliers [8]. We have presented our results on two state-of-the-art databases, namely, CMU-ARCTIC and Voice Conversion Challenge (VCC) 2016 by developing various state-of-the-art methods in the VC. In particular, we have presented effectiveness of outliers removal on GMM [18, 19, 20], Artificial Neural Netowrk (ANN) [21, 22], and Deep Neural Network (DNN)-based VC techniques [23, 24]. Furthermore, we have presented subjective and objective evaluations along with 95 % confidence interval to quote the statistical significance of the results. We obtained on an average 0.56 % relative reduction in Mel Cepstral Distortion (MCD) with the proposed outliers removal approach as a pre-processing step [8]. In particular, with the proposed speaker-dependent cut-offs for the outliers removal, we have observed relative improvement of 21.38 % and 21.99 % (on an average) in the speech quality and the Speaker Similarity (SS), respectively [8].

2.3. INCA and D-INCA

Non-parallel Voice Conversion (VC) has gained significant attention since last one decade. Obtaining corresponding speech frames from both the source and target speakers before learning the mapping function in the non-parallel VC is a key step in the standalone VC task. Obtaining such corresponding pairs, is more challenging due to the fact that both the speakers may have uttered different utterances from same or the different languages. Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) and its variant Temporal Context (TC)-INCA are popular unsupervised alignment algorithms [25, 26]. INCA algorithm was shown to converge empirically, however, its theoretical proof has not been discussed in detail in the VC literature. We have presented that the INCA algorithm will converge monotonically to a local minimum in mean square error (MSE) sense [9]. In addition, we also present the reason of convergence in MSE sense in the context of VC task. The INCA and TC-INCA iteratively learn the mapping function after getting the Nearest Neighbor (NN) aligned pairs from the intermediate converted and the target spectral features. In this work, we propose to use dynamic features along with static features to calculate the NN aligned pairs in both the INCA and TC-INCA algorithms (since the dynamic features are known to play a key role to differentiate major phonetic categories [27, 28, 29, 30]). We obtained on an average relative improvement of 13.75 % and 5.39 % with our proposed Dynamic INCA and Dynamic TC-INCA on the CMu-ARCTIC database, respectively [10]. This improvement is also positively reflected in the quality of converted voices.

2.4. VTLN-based features for non-parallel VC

In the non-parallel VC with the INCA algorithm, the occurrence of one-to-many and many-to-one pairs in the training data will deteriorate the performance of the standalone VC system. The work on handling these pairs during the training is less explored. In this work, we establish the relationship via intermediate speaker-independent posteriorgram representation, instead of directly mapping the source spectrum to the target spectrum. To that effect, a DNN is used to map the source spectrum to posteriorgram representation and another DNN is used to map this posteriorgram representation to the target speaker's spectrum. In this work, we propose to use unsupervised Vocal Tract Length Normalization (VTLN)-based warped Gaussian posteriorgram features as the speaker-independent representations. We performed experiments on a small subset of VCC 2016 database [11]. We obtain the lower Mel Cepstral Distortion (MCD) values with the proposed approach compared to the baseline as well as the supervised phonetic posteriorgram feature-based speaker-independent representations. Furthermore, subjective evaluation gave relative improvement of 13.3 % with the proposed approach in terms of Speaker Similarity (SS) [11].

2.5. BLFW + novel AS techniques

In Frequency Warping (FW)-based VC, the source spectrum is modified to match the frequency-axis of the target spectrum followed by an Amplitude Scaling (AS) to compensate the amplitude differences between the warped spectrum and the actual target spectrum [31, 32, 33]. We propose a novel AS technique which linearly transfers the amplitude of the frequency-warped spectrum using the knowledge of a GMM-based converted spectrum without adding any spurious peaks. The novelty of the proposed approach lies in avoiding a perceptual impression of wrong formant location (due to perfect match assumption between the warped spectrum and the actual target spectrum in state-of-the-art AS method) leading to deterioration in converted voice quality. From subjective analysis, it is observed that the proposed system has been preferred *33.81 %* and *12.37 %* times more compared to the GMM and state-of-the-art AS method for voice quality, respectively [12]. Similar to the quality conversion trade-offs observed by other studies in the literature, speaker identity conversion was *0.73 %* times more and *9.09 %* times less preferred over GMM and state-of-the-art AS-based method, respectively [12].

2.6. Novel MMSE GAN

The murmur produced by the speaker and captured by the Non-Audible Murmur (NAM)-one of the Silent Speech Interface (SSI) technique, suffers from the speech quality degradation. This is due to the lack of radiation effect at the lips and lowpass nature of the soft tissue, which attenuates the high frequency-related information. A novel method for NAM2WHSP conversion incorporating GAN is proposed. The GAN minimizes the distributional divergence between the whispered speech and the generated speech parameters (through adversarial optimization) [34, 35]. The objective and subjective evaluation performed on the proposed system, justifies the ability of adversarial optimization over ML-based optimization networks, such as a DNN, in preserving and improving the speech quality and intelligibility. The adversarial optimization learns the mapping function with 54.2 % relative improvement in MOS and 29.83 % absolute reduction in % WER w.r.t. the state-of-the-art mapping techniques [13]. Furthermore, we evaluated the proposed framework by analyzing the level of contextual information and the number of training utterances required for optimizing the network parameters, for the given task and database [13].

2.7. Quality Evaluation in VC

We propose a novel application of the acoustic- to- articulatory inversion (AAI) towards a quality assessment of the voice converted speech. The ability of humans to speak effortlessly requires the coordinated movements of various articulators, muscles, etc. This effortless movement contributes towards a naturalness, intelligibility and speakers identity (which is partially present in voice converted speech). Hence, during VC, the information related to the speech production is lost. We propose to quantify this loss by showing an increase in RMSE error for a male voice (up to 12.7 % in tongue tip) for voice converted speech followed by showing a decrease in mutual information (I) (by 8.7 %) [14]. Similar results are obtained in the case of a female voice. This observation is extended by showing that the articulatory features can be used as an objective measure. The effectiveness of the proposed measure over MCD is illustrated by comparing their correlation with a MOS [14]. Moreover, the preference score of MCD contradicted ABX test by 100 %, whereas the proposed measure supported ABX test by 45.8 % and 16.7 % in the case of female-to-male and male-to-female VC, respectively [14].

3. Acknowledgements

I would like to thank my supervisor Prof. (Dr.) Hemant A. Patil for his guidance and support. In addition, I acknowledge the authorities of DA-IICT, Gandhinagar and the MietY, Govt. of India, for their kind support to carryout this research work.

4. References

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

[2] S. H. Mohammadi and A. Kain, “An overview of voice conversion systems,” *Speech Communication*, vol. 88, pp. 65–82, 2017.

[3] P. K. Ghosh and S. S. Narayanan, “Closure duration analysis of incomplete stop consonants due to stop-stop interaction,” *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 126, no. 1, pp. EL1–EL7, 2009.

[4] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice. 1^{st} Edition, Chapter 2, pp. 94*. Pearson Education India, 2006.

[5] N. J. Shah and H. A. Patil, *Analysis of features and metrics for alignment in text-dependent voice conversion*. B. Uma Shankar et. al. (Eds), Lecture Notes in Computer Science (LNCS), Springer, PReMI, vol. 10597, pp. 299–307, 2017.

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

[7] S. V. Rao, N. J. Shah, and H. A. Patil, “Novel pre-processing using outlier removal in voice conversion,” in 9^{th} *ISCA Speech Synthesis Workshop (SSW)*, Sunnyvale, CA, USA, 2016, pp. 147–152.

[8] N. J. Shah and H. A. Patil, “Novel outliers removal approach for voice conversion,” *under revision Computer Speech and Language, Elsevier*, 2018.

[9] N. Shah and H. A. Patil, “On the convergence of INCA algorithm,” in *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*. Kuala Lumpur, Malaysia: IEEE, 2017, pp. 1–4.

[10] N. J. Shah and H. A. Patil, “Effectiveness of dynamic features in INCA and temporal context-INCA,” to appear in *INTERSPEECH*, Hyderabad, India, 2018.

[11] N. J. Shah, M. Madhavi, and H. A. Patil, “Unsupervised vocal tract length warped posterior features for non-parallel voice conversion,” to appear in *INTERSPEECH*, Hyderabad, India, 2018.

[12] N. J. Shah and H. A. Patil, “Novel amplitude scaling method for bilinear frequency warping based voice conversion,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5520–5524.

[13] N. Shah, N. J. Shah, and H. A. Patil, “Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion,” to appear in *INTERSPEECH*, Hyderabad, India, 2018.

[14] A. Rajpal, N. J. Shah, M. Zaki, and H. A. Patil, “Quality assessment of voice converted speech using articulatory features,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5515–5519.

[15] 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. 1453–1456.

[16] P. J. Rousseeuw and M. Hubert, “Robust statistics for outlier detection,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[17] J. Hardin and D. M. Rocke, “The distribution of robust distances,” *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, pp. 928–946, 2005.

[18] 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 Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

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

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

[21] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3893–3896.

[22] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[23] 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.

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

[25] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from nonparallel corpora,” *IEEE Transactions on Audio, Speech and Lang. Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[26] H. Benisty, D. Malah, and K. Crammer, “Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7909–7913.

[27] G. Fant, *Speech Sounds and Features*. The MIT Press, 1973.

[28] B. Delgutte, “Auditory neural processing of speech,” *The handbook of Phonetic Sciences*, pp. 507–538, 1997.

[29] J. W. Schnupp, I. Nelken, and A. J. King, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, First Edition, 2012.

[30] S. Furui, “On the role of spectral transition for speech perception,” *The Journal of the Acoust. Soc. of Amer. (JASA)*, vol. 80, no. 4, pp. 1016–1025, 1986.

[31] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001, pp. 841–844.

[32] D. Erro, E. Navas, and I. Hernaez, “Parametric voice conversion based on bilinear frequency warping plus amplitude scaling,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 21, no. 3, pp. 556–566, 2013.

[33] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 20, no. 4, pp. 1313–1323, 2012.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems (NIPS)*, Montral, Canada, 2014, pp. 2672–2680.

[35] M. H. Soni, N. Shah, and H. A. Patil, “Time-frequency masking-based speech enhancement using generative adversarial network,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5039–5043.