

On the potential of modular voice conversion for virtual agents

Silvan Mertes, Thomas Kiderle, Ruben Schlagowski, Florian Lingenfelser, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Mertes, Silvan, Thomas Kiderle, Ruben Schlagowski, Florian Lingenfelser, and Elisabeth André. 2021. "On the potential of modular voice conversion for virtual agents." In *2021 9th International Conference on Affective Computing and Intelligent Interaction, Workshops and Demos (ACIIW)*, 28 September – 1 October, 2021, Virtual Event, Nara, Japan, 1–7. Piscataway, NJ: IEEE.
<https://doi.org/10.1109/ACIIW52867.2021.9666349>.

Nutzungsbedingungen / Terms of use:

licgercopyright



On the Potential of Modular Voice Conversion for Virtual Agents

Silvan Mertes

Human-Centered Artificial Intelligence
Augsburg University
Germany
silvan.mertes@uni-a.de

Thomas Kiderle

Human-Centered Artificial Intelligence
Augsburg University
Germany
thomas.kiderle@uni-a.de

Ruben Schlagowski

Human-Centered Artificial Intelligence
Augsburg University
Germany
ruben.schlagowski@uni-a.de

Florian Lingenfelser

Human-Centered Artificial Intelligence
Augsburg University
Germany
florian.lingenfelser@uni-a.de

Elisabeth André

Human-Centered Artificial Intelligence
Augsburg University
Germany
elisabeth.andre@uni-a.de

Abstract—Virtual Agents are a way to give humans a familiar way to interact with the computer. An important component in the design of virtual agents is the voice with which they express themselves. The voice is not only a mere medium for information transfer, but also contains non-verbal functions such as the transmission of emotions. Additionally, in the context of virtual agents, it is important that the user accepts the voice of the agent and considers it consistent. To make this possible, it is necessary that such voices are highly customisable and adaptable. Current systems for generating speech from text are conceptually limited by the fact that a large part of their task is to model the semantics of what is spoken. Systems in the field of voice conversion, however, are decoupled from this, as they only need to model non-verbal features. Such systems become particularly efficient when they are limited to the transformation of dedicated, single characteristics.

This paper proposes that the use of such voice conversion systems, and furthermore the exploitation of the possibility to cascade them, can be an immense improvement for conventional Text-to-Speech systems for virtual agents.

Index Terms—voice conversion, virtual agents, affective speech, generative adversarial networks

I. INTRODUCTION

Today, people have the opportunity to be supported by computer-based processes and applications in various areas of life. This holds true not only in professional environments, where the advent of intelligent systems brings about a high degree of automation and facilitation of work. Also, in the private sphere, this opens up a whole lot of new opportunities. Here, various application contexts, such as virtual assistance systems [1]–[4], offer perspectives that put people at the centre and thus accompany the step towards a well-being society. Virtual agents, which often take the form of real, photorealistic visualizations of humans, serve the desire to raise interaction with the computer to such a familiar personal level. It is conceivable that the acceptance of such characters is a highly individual matter. As diverse as the potential areas of application of virtual agents are, as individual are

people’s ideas of what a perfect virtual agent for the respective context should be like. From a visual point of view, remarkable milestones like Epic Games’ *MetaHuman Creator*¹ have been reached recently that raise the individualisability of characters to a completely new level.

However, it is often neglected that in addition to the visual appearance of a virtual avatar, the individual needs for auditory components, especially the sound of the agent’s voice, might also play an immensely important role when it comes to acceptance of the agent, although there are only few studies that explicitly investigate the influence of an agent’s voice on users. For example, Esposito et al. [5] showed that an agent’s voice has high impact on the acceptance of such an agent when showed to elderly people. Ritschel et al. [6] demonstrated how individual users’ preferences of an agent’s voice are, although they did not use a virtual agent, but a physical robot. Further, they did not include linguistic aspects in their study, as the agent’s voice was expressed through musical melodies. Other studies revealed that users have preferences regarding the gender of an agent, which can also be expressed through voice [7]–[9]. Furthermore, Gong et al. found that the consistency of visual appearance of an agent and its voice is an additional crucial factor for the design of virtual agents, implying that the design of visual and auditive components of such agents can by no means be seen as independent and separable tasks [10].

A simple example shows how individual the perception of speech is: while reading this work, the reader will probably “hear” the words in his or her own voice, or even in a voice that he or she mentally attributes to the author, even though he or she has never heard him. No matter what the exact nature of this specific imagination is, it is highly likely that the imagination will be different for each reader [11]. This idea highlights the urgency that virtual para-linguistic constructs

¹<https://www.unrealengine.com/en-US/digital-humans>

should follow the guiding principle of individualisation for a consistent experience.

In order to give the virtual agent a voice, modern methods such as Text-to-Speech (TTS) systems are a common mechanism to generate speech for virtual agents [12], increasing the flexibility of agents when compared to using pre-recorded speech samples. However, customisability, both when using recorded speech and when using TTS systems, is still an unsolved problem. It should be noted that the perceived integrity of virtual agents depends not only on the user-relevant individualisation and the diversity required for this, but also on the agent being able to express individual content in a differentiated way. In short, the agent's speech should not be limited to serving as a medium for content, but should also be a means of emphasising the agent's emotional and affective characteristics [13]. Further, the agent's personality is a big part of designing virtual agents [14]–[17]. Hereby, synthesized speech is an excellent tool for modeling personality [18]–[20].

While research in the field of affective speech generation is making steady and remarkable progress in terms of controllable speech with regards to non-verbal characteristics, it has still not been possible to develop a system that offers both the ability to handle emotional speech and the customisability with respect to other speech properties to a satisfactory and interpretable degree. This is not surprising when one considers the fact that a TTS system generates all the properties of the speech material at once, in a single transformation from the domain *text* to the domain *speech*. Properties that satisfy the individual wishes of users (such as identity, gender or age specific sound characteristics) as well as emotion specific factors are determined immutable in one step, and are just one part of the superordinate problem, as a huge part of the speech generation process is to address the modeling of language, i.e., the verbal characteristics. As can be easily seen, this immutability is a limitation. Although there are TTS systems dedicated to the goal of generating speech output in a controlled modifiable way [21]–[24], the current state of the art is still far from a satisfactory solution due to the complexity of the problem and interference between different features to be controlled.

A concept that opens up new possibilities in this context is the principle of voice conversion, which is an active field of ongoing research, traditionally focusing on transforming speech of a specific source speaker to a target speaker [25], but recently also covering speech transformations that do not only tackle speaker identity conversions but different characteristics such as emotion conversion [26]–[28]. Emotional voice conversion, for example, addresses the task of converting speech of a certain emotion to speech of another emotion. While TTS systems perform a domain transfer in the technical sense (i.e., *text* domain to *audio* domain), voice conversion procedures transfer speech to speech, thus remaining in the same technical domain. By maintaining this technical domain *audio*, voice conversion enables, at least on a conceptual level, the application of principles of cascading and recursion. This implies that the restriction prevailing in TTS systems

that all features must be made controllable in a single system is eliminated. Cascading of different voice conversion components enables the distribution of responsibility among different components. For example, the use of such systems allows dedicated transformations to be implemented that adjust only the emotional content of the speech, without having to worry about the basic identity of the voice. The identity of the voice, in turn, can be implemented through a further, cascaded voice conversion step and, moreover, can be staggered at any granular level. Especially in the design of virtual agents, where the freedom in shaping different characteristics is particularly important, this kind of modularity offers advantages.

II. THE CHALLENGE OF SYNTHESIZING VIRTUAL AGENTS' SPEECH

Human speech is a complex and multifaceted construct. Even in contexts apart from virtual agents or other visually supported environments, mapping these multiple facets is a highly challenging task. In these contexts, however, a simplifying factor can be that the design of the voice is a blank sheet of paper, i.e., the voice can be placed in the centre without being disfigured by interference with visual components. The combination of a virtual avatar with a corresponding ability to speak limits this freedom to a great extent. Here, speech can no longer be seen as a uni-modal interface to humans, but must work in symbiosis with the visual representation. Since these visual components are usually already very narrowly defined by the context of the respective applications, speech synthesis must be capable of not only supporting these restrictions, but rather actively contributing to the consistency of the overall impression. For this purpose, it is necessary to make synthesised speech adaptable to the highest degree of detail, which is impossible with current TTS systems. A major focus of TTS systems, which currently represent the state-of-the-art when it comes to synthesising agent speech, is still the linguistic and verbal modelling of speech. Expressive TTS systems can therefore, if at all, only take into account few non-verbal or prosodic voice modalities, since the task to be solved is already very complex. No systems exist that allow for a targeted and comprehensive adaptation of all relevant speech features to an adequate extent, simply because this problem is too demanding and difficult. It is an obvious and natural solution to break complex problems into sub-problems. This possibility is limited within TTS systems, since verbal and non-verbal features inevitably happen all at once when jumping from the text to the speech domain. This is precisely where voice conversion systems can act as an extremely useful complement to TTS systems. The advantage of voice conversion systems is not that they have a better quality than state-of-the-art TTS systems (in fact, this would be a very strong assumption that not necessarily has to be true), but rather that the goal pursued by voice conversion systems is a much simpler one than that of TTS systems. While TTS systems perform a transformation in the verbal and non-verbal space, this verbal level becomes obsolete with voice conversion systems. As a result, voice conversion systems are able to structure and simplify problems

more easily. They give developers of virtual agents the ability to split a large problem into several sub-problems, as will be elaborated on in the following section.

III. THE MODULARITY OF VOICE CONVERSION

TTS systems address the problem of generating speech from textual information. Such systems can therefore be seen as domain transformers that transfer information from the text domain to the speech domain. To solve this problem, modern TTS algorithms work with deep learning methods. Thus, during the training process, the underlying neural network is shown numerous example pairs of text and corresponding speech output while the algorithm attempts to model the relation of these modalities. However, the fact that the speech domain is by no means homogeneous is problematic. The sound of speech is defined by a myriad of properties, such as pitch, intonation, speech rate, and so on. Different combinations of these properties result in the voice being automatically classified into certain clusters by humans. For example, voices with a lower pitch are intuitively assigned to the *Male* cluster, while voices with a higher pitch are perceived as *Female*. Conceptually, such different clusters can also be seen as domains, i.e. *Domain of male speech*, *Domain of happy speech*, *Domain of speech of very old people*, *Domain of the voice of the author of this paper*. It is a triviality that these domains are by no means disjoint, but rather result from the definition of different domain models. There is an infinite number of domain models that can be defined that can be relevant for the design of virtual agents and virtual agents' voice, where each domain model can be seen as one capability that the agents' voice could have. A few domain models were already mentioned in this work, but the list could be continued indefinitely. While the domain model of *Gender* could, among others, include speech subdomains as *Male Speech* and *Female Speech*, one could also define the domain model of *Personality*, which could be defined by speech domains that each contain speech of a certain personality. If needed, the agent's voice should be able to be shifted to one of these domains, i.e., the designer of the virtual agent's voice should have the ability to give the voice one particular personality. In that sense, the domain model *Personality* could, if desired, even be split up into more detailed domain models, such as (inspired by the *Big Five* model [29], [30]) the domain models of *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*. Here, the designer of the virtual agent's voice should for example be able to shift the agent's voice to be part of a domain that contains speech of a low extraversion degree.

In TTS, however, this paradigm is not taken into account. The entirety of the speech is, at least predominantly, considered as one domain. If one wants to obtain a TTS system that produces speech of a very specific sub-domain, the training data set may traditionally only consist of this sub-domain. Addressing different domain models significantly reduces the number of available training data, which has a substantially negative effect on the performance of such neural network-based systems. The authors are aware that

some approaches exist that address this problem. For example, there are approaches to process training data already annotated according to emotions in the training process in order to specifically obtain different emotional speech colourings [31], [32]. However, already during the design of the dataset, a certain modelling is committed, which cannot be changed afterwards, and again only one domain model is taken into account (in that case, *Emotions*). Further, it has to be noted, that the main focus of all TTS systems is to model the verbal characteristics, i.e., the language itself. It is therefore only natural to divide up the complex task of feature-based speech generation so that TTS systems can focus on what they are incredibly good at: Converting text into the higher-level, general domain of *Speech*.

For the controllability of individual non-verbal features, such as emotion, age, gender or voice identity, downstream methods of voice conversion that can be attached to the output of TTS systems are a good choice. Voice conversion offers the important advantage that the technical domain *speech* is retained. More generally spoken, voice conversion algorithms are a sort of domain transfer systems, which are able to shift voice from one speech sub-domain to another. An important point is, that by staying in the same technical domain, voice conversion algorithms can be stacked on top of each other, giving the possibility to build cascaded pipelines of voice conversion systems. This allows domains and domain models to be redefined at any point in time, and to move from sub-domains to other sub-domains iteratively, and, which is the crucial advantage, the overlapping of different domains does not have a *cascading* influence on the training data set of these voice conversion algorithms, i.e., only the domains of one single voice conversion step have to be disjoint, while they can overlap with the domain definitions of the following voice conversion step. Thus, speech can be modified as often as desired, where each modification is done by transforming the speech to be part of another domain. The key point is, that each single Voice Conversion module can model another domain model.

So let's say you have an existing TTS model that works perfectly, but only models a single male speaker and does not include any emotional colouring. It is now possible to run a voice conversion algorithm that converts this emotionally neutral speech into sad speech. Thus, the domains that we consider relevant for this particular step are the domain of *neutral speech* and the domain of *sad speech*. For the training of this voice conversion algorithm, only some data of neutral speech and some data of sad speech are necessary, without this speech having to fulfil other feature restrictions. If we want the sad speech to sound feminine, we can apply another voice conversion algorithm to the result of the emotional colouring, which has been trained to convert from masculine to feminine speech. Again, to train this voice conversion algorithm, one only needs data from male and female speech, without any other restrictions. These cascaded voice conversion steps can now, in theory, be performed many more times in different domains, and by doing so the resulting voice can be fine-tuned

to one's own preferences. The remarkable thing is that the individual voice conversion models are independent of each other. This implies that they can be trained independently, but also that whole frameworks can always be extended with new voice conversion modules. Another voice conversion algorithm can be trained at any time, which converts desired features into each other. The result is a modular system in which a single TTS system provides the entry point to unlimited adaptability through nesting and cascading of voice conversion algorithms.

IV. ADVERSARIAL VOICE CONVERSION

An exemplary concept for the implementation of such voice conversion methods that is very promising is the class of algorithms that can be aligned to the field of *Adversarial Voice Conversion*.²

Adversarial Voice Conversion is based on the use of Generative Adversarial Networks (GANs), which were first introduced by Goodfellow et al. [33]. The basic principle of GANs is that two neural networks, the so-called *generator* and the *discriminator*, learn opposing tasks and thus improve each other (hence the name *adversarial*). The generator attempts to modify a random noise distribution in order to produce new, highly realistic data that looks or sounds as if it came from a specific training data set. The discriminator, on the other hand, tries to expose the data generated by the generator as fake data. In other words, it tries to distinguish between real data from the training data set and fake data. The trained generator can then be used to generate deceptively real examples from random noise vectors. From the perspective of speech research, these real examples can have the technical structure of audio. The objective of a GAN can mathematically be formulated as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))], \quad (1)$$

where G is the generator network, D is the discriminator, and z is the input noise. The scheme of such a basic GAN is depicted in Figure 1a.

Obviously, these original GANs cannot be called voice conversion algorithms, since they do not convert audio to audio, but rather unstructured noise to audio. Extensions of the original GANs address this by replacing the noise input with real data. The learning goal thus changes in a way that new data is not to be constructed from noise vectors, but new data from existing data, which is already quite close to the concept of voice conversion. Approaches that implement the aforementioned learning objective in the presented form can be found primarily in the field of image processing [34].

One problem that arises, however, is that the relation between input and output data pairs must be known during training. This means that during the training of the network

we have to give concrete data pairs to the network, which differ only by the desired feature to be converted. If we take the example of emotion conversion, this would mean that we would need a large number of data pairs that do not differ in any feature except the emotion. For example, the data pairs would have to contain the same sentence, be spoken by the same speaker, be recorded in the same setting, to name just a few restrictions, making the overriding objective of simple extensibility of speech for virtual agents absurd.

Fortunately, further modifications to the original GAN framework have been developed that overcome this problem. A prominent example of this are the so-called *Cycle-consistent GANs* (CycleGANs) [35]. CycleGANs consist of two individual GANs, which in turn play against each other. Let's assume generically that we want to carry out a conversion between domain A (DN_A) and domain B (DN_B). The goal of the first GAN (G_1) would then be to convert data from DN_A to DN_B , while the second GAN (G_2) converts data from DN_B to DN_A . If we now take a sample S_{DA} belonging to DN_A and transform it with the help of G_1 , we obtain $G_1(S_{DA})$, which should ideally be perceived as belonging to DN_B . If we now convert this result with the help of G_2 , then $G_2(G_1(S_{DA}))$ should in turn be part of DN_A . This is the key factor for the functionality of CycleGANs, because now $G_2(G_1(S_{DA}))$ can be compared with S_{DA} . Ideally, the two conversions have now only converted the desired features back-and-forth, while the uninvolved features have remained the same, which is why the difference between these two data points can be used directly to teach the CycleGAN. This so-called *cycle consistency loss* can be defined mathematically as follows:

$$\mathcal{L}_{cycle}(G_1, G_2) = \mathbb{E}_{x \sim p_{data}(x)} [G_2(G_1(x)) - x_1] + \mathbb{E}_{y \sim p_{data}(y)} [G_1(G_2(y)) - y_1], \quad (2)$$

where G_1 and G_2 are the generators of the both GANs, and x and y are data from the domains DN_A and DN_B respectively.

This mechanism causes the two networks to allow conversion between domains without the need to use a paired dataset. Combined with the formulations for the two incorporated GANs, the loss function of a CycleGAN is defined as:

$$\mathcal{L}(G_1, G_2, D_A, D_B) = \mathcal{L}_{GAN}(G, D_B, DN_A, DN_B) + \mathcal{L}_{GAN}(G_2, D_A, DN_B, DN_A) + \lambda \mathcal{L}_{cycle}(G_1, G_2), \quad (3)$$

where D_A and D_B are the discriminators of the both GANs, and λ is a factor to weight the impact of the cycle-consistency loss. $\mathcal{L}_{GAN}$ is the adversarial loss that can be derived from the original GAN objective formulated in (1). The basic scheme of a CycleGAN is depicted in Figure 1b.

Kaneko et al. and Fang et al. already successfully applied CycleGANs to the task of voice conversion, although their work has been limited to the problem of speaker identity conversion [36]–[40]. Yook et al. extended the CycleGAN architecture with a conditional component in order to be

²Exemplary sound samples of speech that was converted with techniques of Adversarial Voice Conversion can be found at <https://github.com/SilvanMertes/VoiceConversion>.

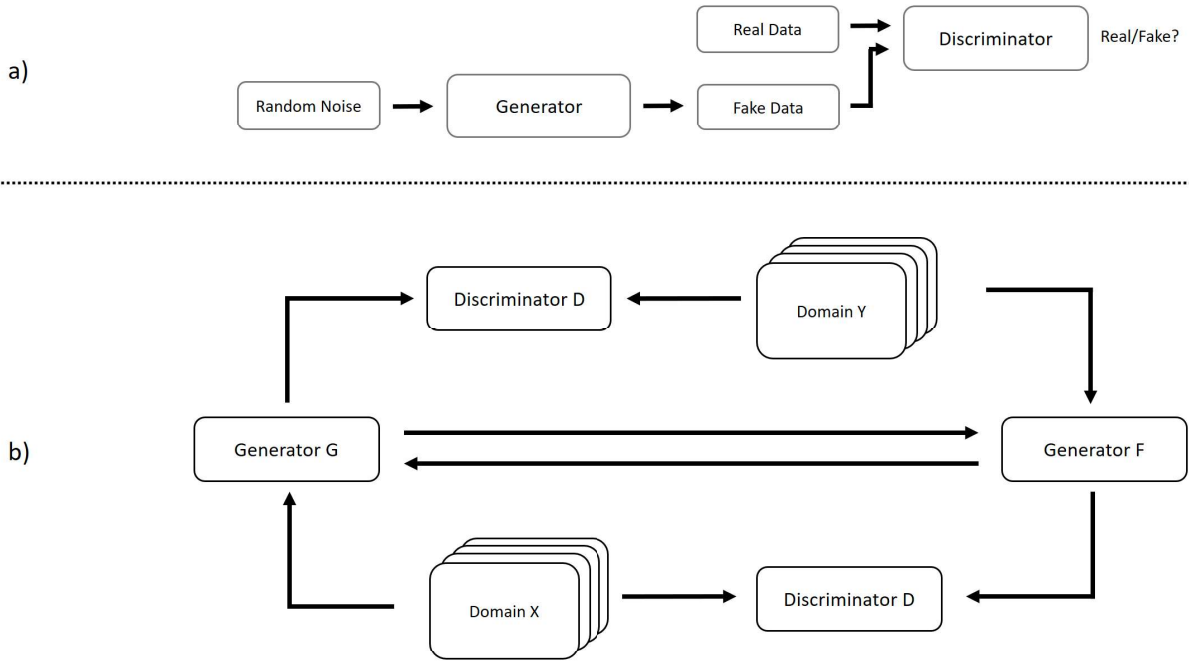


Fig. 1. a) Basic scheme of a Generative Adversarial Network. b) Basic scheme of a CycleGAN.

able to model a multi-speaker conversion with one single CycleGAN [41]. Liu et al. were the first to use CycleGANs for emotional voice conversion [42].

Since the operating principle of CycleGANs forces that only speech features are changed that are important for the actual domain transfer, it offers the optimal basis for building modular voice conversion systems for virtual agents. This becomes particularly apparent when one realises that CycleGAN is designed for binary domain transfer. This means that the complete learning problem is limited to two single domains, for example the conversion between two emotions. There exist other algorithms from the field of adversarial voice conversion, such as methods based on the *Stargan* architecture, which enables a domain transfer from one domain to several other domains [43]. Such methods have also been successfully used for voice conversion tasks [44]–[47]. In terms of modularity, however, it can be argued that in binary transformation problems the entire capacity of the learning process is focused on a small, defined part of the overall problem and is thus simplified and more efficient. This is a highly relevant consideration, especially with regard to training effort and training data requirements. It is quite conceivable that the concept of modular, cascaded, CycleGAN-based voice conversion represents a practical option to realise a fine-granular and well extendable adaptation and individualisation of virtual agents.

V. CURRENT LIMITATIONS OF VOICE CONVERSION

The modular use of voice conversion systems offers a promising horizon for the auditory design of virtual agents. However, the goal of using these techniques in real-world applications and providing developers and consumers with the benefits of modular voice conversion undoubtedly still faces a number of obstacles.

Probably the most significant inhibiting factor is the quality of the results of current voice conversion systems. Although algorithms that can be assigned to the field of adversarial voice conversion in particular achieve impressive results today, there is still a long way to go to make such systems arbitrarily cascable. Cascading voice conversion algorithms inevitably implies that any loss of quality in the speech is also passed on to the next link in the chain. The smallest disturbances, such as the emergence of noise or similar, add up due to the principles of cascading. This results in a steady decline in speech quality, which would lead to unacceptable overall results when using current state-of-the-art algorithms. Much research and development work still needs to be invested in the improvement and optimisation of such methods until operational systems can be constructed.

In addition to the quality of current voice conversion systems, the choice of domains to be converted into each other also leaves a lot of room for manoeuvre. The majority of existing voice conversion systems still limit themselves to the task of speaker identity conversion. Fortunately, other conversion tasks, e.g. emotion conversion, have become in-

creasingly popular in recent years, but the immense advantage of modular voice conversion can only be realised reasonably when a large number of further domain models are mapped. For example, the authors are not aware of any work that performs voice conversion in terms of the *Big Five* based personality of the voice. Even intuitively simpler problems such as the conversion of the gender of speech have, to the authors' knowledge, not yet been addressed explicitly.

Furthermore, it must be considered that, especially in the context of virtual agents, speech must by no means be seen in isolation from visual design possibilities. The visual appearance of virtual agents must be designed consistently with auditory components, and the adaptation of speech with the help of voice conversion also implies that the associated potentials in the graphic development of characters must be considered.

In addition, ways and concepts must be worked out to make corresponding systems technically accessible. It is not enough to create dedicated algorithms and models if they cannot actually be used by practitioners to integrate the corresponding procedures into real-world applications.

VI. CONCLUSION & OUTLOOK

Enabling virtual agents with flexible speech is an important means of promoting the acceptance and integrity of such systems.

While modern TTS systems are improving and are already delivering impressive results, they are conceptually limited by the fact that a large part of their task involves the semantic conversion of text to speech, while prosodic enrichment in terms of various characteristics, such as emotions or personality, further increases the complexity of the task to be solved. In this paper, the authors have discussed why it may be an appropriate option to outsource this second part of the task by replacing it by the principles of modular voice conversion. Not only is it conceivable that by simplifying the overall problem for TTS systems, these can thereby achieve better performance, but also the use of downstream and cascaded voice conversion is a way towards more modularity and extensibility of speech design. By addressing individual domain transfers in a dedicated way, modular voice conversion at a conceptual level enables a variety of voice properties to be successively adapted, even if these individual properties are not independent of each other but depend on the same features. The research community is encouraged to jointly address the improvement and optimisation of voice conversion algorithms in order to be able to use them in cascable frameworks in real-world scenarios.

In the future, we plan to implement a framework that allows the use of CycleGAN-based voice conversion systems in a cascable manner for the personalised design of speech for virtual agents.

ACKNOWLEDGMENT

This work has been funded by the European Union Horizon 2020 research and innovation programme, grant agreement 856879.

REFERENCES

- [1] Fiol-Roig, Gabriel, et al. "The intelligent butler: A virtual agent for disabled and elderly people assistance." International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI 2008). Springer, Berlin, Heidelberg, 2009.
- [2] Wanner, Leo, et al. "Kristina: A knowledge-based virtual conversation agent." International conference on practical applications of agents and multi-agent systems. Springer, Cham, 2017.
- [3] Hasler, Béatrice S., Peleg Tuchman, and Doron Friedman. "Virtual research assistants: Replacing human interviewers by automated avatars in virtual worlds." Computers in Human Behavior 29.4 (2013): 1608-1616.
- [4] Weitz, Katharina, et al. "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design." Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019.
- [5] Esposito, Anna, et al. "The Dependability of Voice on Elders' Acceptance of Humanoid Agents." INTERSPEECH. 2019.
- [6] Ritschel, Hannes, et al. "Personalized synthesis of intentional and emotional non-verbal sounds for social robots." 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019.
- [7] Esposito, Anna, et al. "Seniors' acceptance of virtual humanoid agents." Italian forum of ambient assisted living. Springer, Cham, 2018.
- [8] Payne, Jeunese, et al. "Gendering the machine: Preferred virtual assistant gender and realism in self-service." International Workshop on Intelligent Virtual Agents. Springer, Berlin, Heidelberg, 2013.
- [9] Shang, Xiumin, Marcelo Kallmann, and Ahmed Sabbir Arif. "Effects of Virtual Agent Gender on User Performance and Preference in a VR Training Program." Future of Information and Communication Conference. Springer, Cham, 2019.
- [10] Gong, Li, and Clifford Nass. "When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference." Human communication research 33.2 (2007): 163-193.
- [11] Vilhauer, Ruvanee P. "Inner reading voices: An overlooked form of inner speech." Psychosis 8.1 (2016): 37-47.
- [12] Stan, Adriana, and Beáta Lőrincz. "Generating the Voice of the Interactive Virtual Assistant." Virtual Assistant. IntechOpen, 2021.
- [13] Ghafurian, Moojan, Neil Budnarain, and Jesse Hoey. "Role of emotions in perception of humanness of virtual agents." Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. 2019.
- [14] Liu, Kris, et al. "Two techniques for assessing virtual agent personality." IEEE Transactions on Affective Computing 7.1 (2015): 94-105.
- [15] McRorie, Margaret, et al. "Evaluation of four designed virtual agent personalities." IEEE Transactions on Affective Computing 3.3 (2011): 311-322.
- [16] Siddique, Farhad Bin, et al. "Zara Returns: Improved Personality Induction and Adaptation by an Empathetic Virtual Agent." ACL (System Demonstrations). 2017.
- [17] Cerekovic, Aleksandra, Oya Aran, and Daniel Gatica-Perez. "How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits." International Workshop on Human Behavior Understanding. Springer, Cham, 2014.
- [18] Nass, Clifford, and Kwan Min Lee. "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction." Journal of experimental psychology: applied 7.3 (2001): 171.
- [19] Aylett, Matthew P., Alessandro Vinciarelli, and Mirjam Wester. "Speech synthesis for the generation of artificial personality." IEEE transactions on affective computing 11.2 (2017): 361-372.
- [20] Aylett, Matthew P., Yolanda Vazquez-Alvarez, and Skaiste Butkute. "Creating robot personality: effects of mixing speech and semantic free utterances." Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. 2020.
- [21] Wang, Yuxuan, et al. "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis." International Conference on Machine Learning. PMLR, 2018.
- [22] Valle, Rafael, et al. "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

- [23] Valle, Rafael, et al. "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis." arXiv preprint arXiv:2005.05957 (2020).
- [24] van Rijn, Pol, et al. "Exploring emotional prototypes in a high dimensional TTS latent space." INTERSPEECH. 2021.
- [25] Mohammadi, Seyed Hamidreza, and Alexander Kain. "An overview of voice conversion systems." *Speech Communication* 88 (2017): 65-82.
- [26] Kawanami, Hiromichi, et al. "GMM-based voice conversion applied to emotional speech synthesis." (2003).
- [27] Aihara, Ryo, et al. "GMM-based emotional voice conversion using spectrum and prosody features." *American Journal of Signal Processing* 2.5 (2012): 134-138.
- [28] Zhou, Kun, et al. "Converting anyone's emotion: Towards speaker-independent emotional voice conversion." arXiv preprint arXiv:2005.07025 (2020).
- [29] McCrae, R.R. and John, O.P., "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, 1992.
- [30] Mehrabian, A., "Analysis of the big-five personality factors in terms of the PAD temperament model", *Australian Journal of Psychology*, vol. 48, no. 2, August 1996, pp. 86-92.
- [31] Lorenzo-Trueba, Jaime, et al. "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis." *Speech Communication* 99 (2018): 135-143.
- [32] Lee, Younggun, Azam Rabiee, and Soo-Young Lee. "Emotional end-to-end neural speech synthesizer." arXiv preprint arXiv:1711.05447 (2017).
- [33] Goodfellow, Ian J., et al. "Generative adversarial networks." arXiv preprint arXiv:1406.2661 (2014).
- [34] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [35] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [36] Kaneko, Takuhiro, and Hirokazu Kameoka. "CycleGAN-vc: Non-parallel voice conversion using cycle-consistent adversarial networks." 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018.
- [37] Kaneko, Takuhiro, et al. "CycleGAN-vc2: Improved cycleGAN-based non-parallel voice conversion." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- [38] Kaneko, Takuhiro, et al. "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion." arXiv preprint arXiv:2010.11672 (2020).
- [39] Kaneko, Takuhiro, and Hirokazu Kameoka. "Parallel-data-free voice conversion using cycle-consistent adversarial networks." arXiv preprint arXiv:1711.11293 (2017).
- [40] Fang, Fuming, et al. "High-quality nonparallel voice conversion based on cycle-consistent adversarial network." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
- [41] Yook, Dongsuk, In-Chul Yoo, and Seungho Yoo. "Voice conversion using conditional CycleGAN." 2018 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2018.
- [42] Liu, Songxiang, Yuewen Cao, and Helen Meng. "Emotional Voice Conversion With Cycle-consistent Adversarial Network." arXiv preprint arXiv:2004.03781 (2020).
- [43] Choi, Yunje, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [44] Kameoka, Hirokazu, et al. "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks." 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.
- [45] Kaneko, Takuhiro, et al. "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion." arXiv preprint arXiv:1907.12279 (2019).
- [46] Rizos, Georgios, et al. "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
- [47] Wang, Ruobai, et al. "One-Shot Voice Conversion Using Star-Gan." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.