Deep Variational Metric Learning For Transfer Of Expressivity In Multispeaker Text To Speech

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Deep Variational Metric Learning For Transfer Of Expressivity In Multispeaker Text To Speech

Ajinkya Kulkarni
Université de Lorraine, CNRS, Inria, LORIA,
F-54000 Nancy, France.
ajinkya.kulkarni@loria.fr

Vincent Colotte
Université de Lorraine, CNRS, Inria, LORIA,
F-54000 Nancy, France.
vincent.colotte@loria.fr

Denis Jouvet
Université de Lorraine, CNRS, Inria, LORIA,
F-54000 Nancy, France.
denis.jouvet@inria.fr

***Abstract*—In this paper, we propose an approach relying on multiclass N-pair loss based deep metric learning in recurrent conditional variational autoencoder (RCVAE). We used RCVAE for implementation of multispeaker expressive text-to-speech (TTS) system. The proposed approach condition text-to-speech system on speaker embeddings, and leads to clustering the latent space representation with respect to emotion. The deep metric learning helps to reduce the intra-class variance and increase the inter-class variance in latent space. Thus, we present multiclass N-pair loss to enhance the meaningful representation of the latent space.**

For representing the speaker, we extracted speaker embeddings from the x-vector based speaker recognition model trained on speech data from many speakers. To predict the vocoder features, we used RCVAE for the acoustic modeling, in which the model is conditioned on the textual features as well as on the speaker embedding. We transferred the expressivity by using the mean of the latent variables for each emotion to generate expressive speech in different speaker's voices for which no expressive speech data is available. We compared the results with those of the RCVAE model without multiclass N-pair loss as baseline model. The performance measured by mean opinion score (MOS), speaker MOS, and expressive MOS shows that N-pair loss based deep metric learning significantly improves the transfer of expressivity in the target speaker's voice in synthesized speech.

***Index Terms*—text-to-speech, variational autoencoder, deep metric learning, expressivity**

I. Introduction

Evolution of text-to-speech (TTS) models have shown that parameterisation of waveform is still a critical step for achieving state-of-the-art performance. The development of end-to-end text-to-speech models are heavily relying on encoder-decoder attention based neural network architectures which map textual vector representation to sequence of frames of spectrograms [1- 4]. At present, the style of the produced speech signal is neutral, which results from the type of data used to train the models. Interacting with this generation of speech synthesis system for a long duration, makes it monotonous and less interactive. Multispeaker expressive speech synthesis is still an open problem due to limited availability of expressive speech corpora and time involved in collection and annotation of such corpora for a new speaker.

The work done by [11], proposed expressivity transplantation as an extension to speaker adaptation using Latent Hidden Unit Contribution (LHUC) units. Also, in [12] a layer adaptation technique is proposed, which is similar to domain adaptation. With reasonable results, the above approaches are limited to work in a single speaker text-to-speech framework.

Several approaches have been proposed previously to transfer expressivity either by controlling the prosody parameters in latent space for speech synthesis or by transferring the expressivity using interpolation of conditional embeddings of speaker identity and prosody embedding [5-10]. In [5, 6] authors proposed to adopt the variational autoencoder (VAE) framework within end-to-end TTS systems such as voiceLoop, and tacotron [1, 4], to transform parameterized speech into latent representation and disentangle the latent speech attributes such as prosody, and speaker. In [26], the author proposed a conditional VAE (CVAE) model for expressive audio-visual synthesis. The CVAE model is constrained for single speaker audio-visual synthesis. In our work, we present x-vector based speaker embedding, which paved the way to build a multi-speaker expressive TTS system. In our approach recurrent VAE based acoustic model is conditioned on textual features along with speaker embedding. With this conditioning, we expect to extract emotion information in latent space representation.

To transfer the expressivity, creating well clustered latent representation of emotions is a crucial factor for producing expressivity in synthesised speech. To enhance the latent space representation, we propose using deep metric learning framework along with variational inference [13]. Deep metric learning has been widely used for training discriminative models for computer vision applications [14, 15]. For augmenting deep metric learning along with variational inference, we proposed to use multi class N-pair loss as a deep metric learning.

In [7-9] authors proposed to use reference encoder models to learn disentangling in latent space, the speaker embedding which is then used to derive speech signal for the desired speaker in a tacotron based speech synthesis system. Contrary to this approach, we propose a transfer learning approach where speaker information is derived from x-vector embeddings extracted from pertained speaker recognition model. The x-vector embedding maps variable length utterances to text independent fixed dimensional embeddings, which are

trained using deep neural network that discriminate between speakers [16]. The extracted x-vector embeddings are used to build a speaker encoder network that will produce speaker representation for the multi-speaker TTS task.

In this paper, we present the deep variational metric learning for the acoustic model in section II. We discuss speaker embeddings created for French speaker using pretrained speaker recognition model in section III. We provide details about speech and text data preprocessing for training multi-speaker expressive TTS system in section IV. Afterwards, we present the experimentation set up in section V and section VI discusses the results obtained using the RCVAE model with and without multiclass N-pair loss to show impact in the transfer of expressivity.

II. Model

Variational autoencoders were introduced in 2013 by Kingma and Rezende independently [17, 18]. Variational autoencoders have components such as encoder, decoder and loss function. The loss function corresponds to the reconstruction loss plus a regularization term defined with a Kullback-Leibler (KL) divergence. The reconstruction loss represents the expectation over the reconstruction of input, $logP(x|z,c)$. The KL divergence measure indicates how close the learned distribution $Q(z|x,c)$ is to the true prior distribution $P(z|c)$. We applied a λ factor to the KL divergence term as explained in VAE architecture [19] (as mentioned in Eq 1.). This facilitates the VAE model to avoid the sudden drop in KL loss term and it also helps to enhance the disentangled latent space representation with good interpretability of the latent variable.

$$Loss = E_z[logP(x|z,c)] + \lambda KL[Q(z|x,c)\|P(z|c)] + log(1 + \sum_{i=1}^{N-1} exp(z^\top z_i^- - z^\top z^+)) \quad (1)$$

For the RCVAE architecture, we implemented bidirectional long short term memory (BLSTM) based encoder network, presented with a sequence of input data, x along with condition c. In the encoder network, the last hidden state of BLSTM network is given to feedforward layers to estimate both mean and variance to describe the encoder's latent space distribution. Similar to the encoder network, the decoder network also have bidirectional long short term memory layers, which takes as input a latent variable z and the condition c, to generate the output sequence x. During the inference, we sample, z from the latent space and give it to the decoder network along with the condition c, as shown in Fig 1.

The focus of the presented work is on implementation of acoustic model with multi class N-pair loss training using RCVAE architecture. In this framework, we used explicit duration model to predict the number of acoustic feature frames required to synthesize the speech for a given text. Thus, for predicting duration for each phoneme, we used BLSTM based neural network as explained in [27].

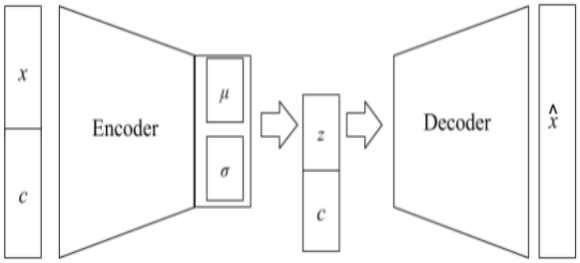

Fig. 1. RCVAE architecture used for acoustic model. Here, x is a sequence of input features to be reconstructed as $\hat{x}$, c is condition μ and σ are mean and standard deviation parameters provided by the encoder network, and used to generate the latent representation z.

For the acoustic model, the condition c corresponds to textual features, duration information and speaker embedding and $\hat{x}$ represent the predicted acoustic features. The proposed RCVAE architecture integrates the textual and speaker information in the condition c, and we observed emotion in the latent space representation as shown in Fig 2. In inference phase, we used mean value of latent variables to synthesize expressive speech. Therefore, choosing an appropriate latent variable is a crucial factor in generating appropriate expressivity in synthesized speech.

We propose to add a multi class N-pair loss criteria as a deep metric learning to variational inference along with reconstruction loss and KL loss. Multiclass N-pair loss has shown superior performance compared to triplet loss or contrastive loss by considering one positive sample and $N-1$ negative samples for N classes [15]. This loss criteria increases the intercluster distance from $N-1$ negative samples and decreases the intracluster distance between positive samples and training example. In this case, positive samples refer to latent variables from same emotion class and negative samples corresponds to examples of different emotion classes. For N classes, z^+ is a positive sample and $\{z_i^-\}_1^{N-1}$samples are from negative classes as stated in Eq. 1. For sampling positive and negative examples, we used precomputed mean of latent variables for each emotion.

To show that the addition of multiclass N-pair loss improves the overall performance of multi-speaker TTS and transfer of expressivity to the target speaker's voice, we plotted a t-SNE plot using latent variables z generated using encoder network in RCVAE. The t-SNE plots in Fig. 2 shows that without using N-pair loss, the latent representations of the emotions overlap (left side). However, including N-pair loss in the training process leads to rather well separated clusters associated to each emotion. In the following experimentation, we compared the performance of the RCVAE acoustic model without multiclass N-pair loss as a baseline model and with multiclass N-pair loss as the proposed model.

III. Speaker Embedding

The encoder-decoder network in RCVAE is explicitly conditioned on the speaker to disentangle the expressive information

in latent space. To capture this speaker's information, we propose to create speaker embedding from reference speech samples. These embeddings should represent only characteristics of speaker identity irrespective of the textual content in the reference speech sample. For generating such embeddings, we proposed to develop a speaker encoder using x-vector embeddings extracted from a pretrained speaker recognition model.

The x-vector embeddings are deep neural network based embeddings trained on time-delay neural networks with a statistical pooling layer trained for the speaker recognition task [16]. Firstly, we extracted x-vector embeddings from the pretrained speaker recognition model trained on the voxceleb corpus available in the Kaldi tool [20, 21]. To adapt the speaker embeddings to French speakers, we used extracted x-vector embeddings to train a feedforward neural network based speaker recognition model for discriminating between the speakers of our French speech synthesis corpora. Also, we reduced the dimension of speaker embeddings by taking the output of the last hidden layer of feedforward neural networks in the French speaker recognition model.

IV. Data preparation

We worked with 4 speech corpora, namely Lisa [12], a French female neutral corpus (approx. 3 hrs), Caroline [26], a French female expressive corpus (approx. 9hrs), SIWIS [23], a French female neutral corpus (approx. 3 hrs), and Tundra [24], A French male neutral corpus (approx. 2hrs). Caroline expressive speech corpus consists of several emotions, namely joy, surprise, fear, anger, sadness, and disgust (approx. 1hr for each emotion and 3hrs for neutral). For each emotion, there are approximately 500 utterances for a total of 1hr duration. All the speech signals were used at a sampling rate of 16 kHz.

We parameterized speech using the WORLD vocoder [22] with 187 acoustic features computed every 5 milliseconds, namely 180 spectral features as Mel generalized cepstrum coefficients (mgc), 3 log fundamental frequencies (lf0), 3 band-aperiodicities (bap) and 1 value for voiced-unvoiced information (vuv). Based on the mean and standard deviation values, the acoustic features extracted from the WORLD vocoder were z-normalized. We used the front-end test processor from SOJA-TTS (developed internally in our team) for converting French text to linguistic features also known as context labels (dimension 180) which include pentaphone information.

V. Experimentation

For implementing RCVAE architecture, we incorporated 2 layers of BLSTM network of 256 hidden units for both encoder network and decoder network with a latent variable of 50 dimensions, a learning rate of 0.0001, Adam optimizer was initialized with default parameters and a batch size of 10 and lambda factor of 0.001. The model wass trained until the 100th epoch. To ensure better convergence of model parameters multi-class N-pair loss was activated only after the first 5 epochs. Also, for the baseline model, we trained the RCVAE model without deep metric learning.

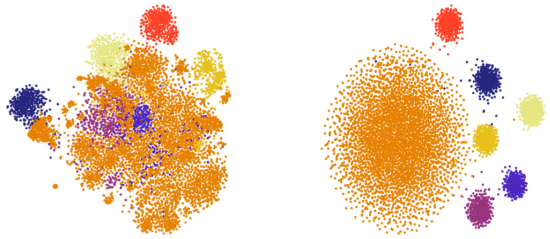

Fig. 2. t-SNE plot of latent representation of RCVAE acoustic model (left side) and RCVAE acoustic model with npair loss (right side). Each color represents the emotion.

In the inference phase, we used the mean of latent variables constituting given emotions as latent variable to synthesize a particular emotion. As mentioned before, we implemented a duration model explicitly for each speaker using the BLSTM network of 512 hidden units with the same configuration of batch size, learning rate, and optimizer as of RCVAE architecture.

With respect to speaker embeddings, we extracted 512-dimensional x-vector from the speaker recognition model trained on the voxceleb corpus [20] for all speech samples in the corpus. Thereafter, we implemented 5 layers of feed-forward neural network trained to classify 4 French speakers (corresponding to our speech synthesis corpora) with (512-256-128-64-16) hidden units, using cross-entropy loss criteria, Adam optimizer, and 50 epochs of training. We extracted speaker embedding for each speech sample by taking the output of the last hidden layer of dimension 16.

VI. Results

We carried out the Mean Opinion Score (MOS) [25] perception test for evaluating our multi-speaker expressive text-to-speech synthesis system. For the MOS perception test, each listener had to score the synthesized speech stimuli from 1 to 5, where 1 is bad and 5 is excellent, considering intelligibility, naturalness, and quality of the speech stimuli. With this test, we evaluated TTS for all speakers and all the emotions present in the Caroline corpus. 12 French listeners participated in the perception test; each listener had to score 5 stimuli for each speaker-emotion pair randomly chosen from the test set. The results of the MOS test are shown in Table 1 with an associated 95% confidence interval. The MOS scores for all speakers except Caroline have comparably similar results, in which Lisa speaker received the highest score for both the models trained with and without deep metric learning. The average MOS score for Caroline's voice is 2.4 for RCVAE model and 2.9 for RCVAE with N-pair loss. The MOS score presented for Caroline speaker in Table I represents the average score obtained for Caroline's neutral voice and for all Caroline emotions (with associated confidence interval). Due to limited training data (1hr) for each emotion for Caroline's voice, MOS score performance on Caroline's speech synthesis is lower

TABLE I
MOS SCORE FOR EVALUATION OF MULTI-SPEAKER TTS SYSTEM

Baseline MOS	***Caroline***	***Lisa***	***Siwis***	***Tundra***
RCVAE	2.4 ± 0.3	2.8 ± 0.7	2.6 ± 0.8	2.7 ± 0.2
RCVAE+N-pair	2.9 ± 0.2	3.1 ± 0.6	3.0 ± 0.5	2.9 ± 0.4

TABLE II
SPEAKER MOS SCORE FOR EVALUATION OF TRANSFER OF SPEAKER CHARACTERISTICS

Speaker MOS	***Lisa***	***Siwis***	***Tundra***
RCVAE	2.3 ± 0.2	2.2 ± 0.1	2.7 ± 0.3
RCVAE+N-pair	3.0 ± 0.1	2.7 ± 0.3	2.9 ± 0.2

TABLE III
EXPRESSIVE MOS SCORE FOR EVALUATION OF TRANSFER OF EXPRESSIVITY

Expressive MOS	***Lisa***	***Siwis***	***Tundra***
RCVAE	1.4 ± 0.4	1.5 ± 0.3	1.7 ± 0.5
RCVAE+N-pair	1.9 ± 0.3	1.9 ± 0.4	2.0 ± 0.2

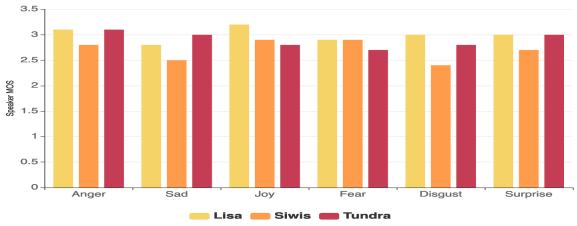

Fig. 3. RCVAE model with N-pair loss, Speaker MOS score for all emotions and speakers

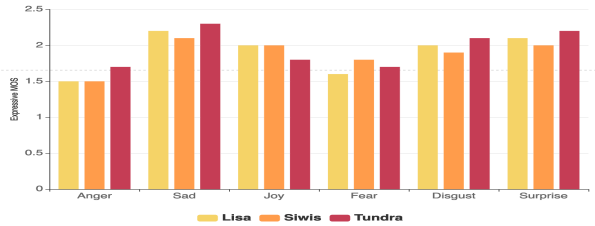

Fig. 4. RCVAE model with N-pair loss, Expressive MOS score for all emotions and speakers

compared to other speakers. The results presented in Table I show that deep variational learning approach leads to better results, this is also in line with the better seperation between emotions in the latent space, as observed from the t-SNE plots in Fig 2.

We used speaker MOS and expressive MOS to evaluate the performance of the proposed architecture transferring expressivity onto other speaker voices. In the speaker MOS perception test, we instructed listeners to provide a score about the similarity between the original speaker speech stimuli and synthesized expressive speech in a range of 1 (bad) to 5 (excellent). Likewise, we also directed listeners to score expressivity observed in the synthesized expressive speech stimuli on a scale of 1 (bad) to 5 (excellent) depending on the closeness of expressive characteristics in speech stimuli compared to original expressive speech stimuli. 12 French listeners participated in a perception test, each listener scored 3 sets of stimuli for each target speaker-emotion pair. The results of expressive MOS and speaker MOS are shown in Table II and Table III, with associated 95% confidence interval. Figure 3 and 4 display the speaker MOS and expressive MOS scores relating to the transfer of expressivity respectively. Figure 3 and 4 show that 3 speakers have similar results for all emotion, Siwis speaker got slightly lower score compared to other speakers.

The obtained results showed that the addition of deep metric learning certainly improves the representation of expressivity. Consequently, this improves the performance of transfer of expressivity. The presented work is the first approach that uses deep metric learning in a variational inference framework to improve the performance of latent space representation in a multi-speaker expressive TTS system.

Furthermore, speaker MOS showed that while transferring expressive knowledge, addition of N-pair loss to architecture improve retainment of the speaker characteristics. Also results from Table III showed that the system is able to equally transfer the expressivity not only from female (Caroline) to female (Lisa, Siwis) speakers but also from female (Caroline) to male (Tundra) speaker. In addition, our proposed approach shows better performance for Lisa speaker than previous layer adaptation [12] approach for both speaker MOS and expressive MOS. From Fig. 3 and 4, Tundra speaker shows that sad and surprise are the emotions perceived as close to expressive characteristics with respect to the original reference speech provided in evaluation. While anger is the least perceive emotion for all speakers. On the other hand, transferring anger emotion to target speakers received higher speaker MOS scores.

VII. CONCLUSION

We presented variational autoencoder architecture for transferring expressivity characteristics in a multispeaker text-to-speech synthesis system. To enhance the disentanglement of information in the latent space, we have included a multiclass N-pair loss component as deep metric learning. In our approach, the deep variational metric learning helped to enforce the better clustering of emotions in latent space representation.

In addition to this, we presented a novel way to represent the speaker's characteristics by encoding speaker information using x-vector embedding as input of a speaker recognition neural network model trained on the speakers of our French synthesis corpora; the last hidden layer provides the speaker embedding for the TTS system. The perception tests conducted show that the proposed approach retains the target speaker's voice while transferring the expressivity. In the future, we would like to adopt a similar RCVAE based deep variational metric learning in an end-to-end TTS system.

REFERENCES

[1] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., Saurous, R. A.. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. Journal: CoRR, arxiv.org, volume: abs/1703.10135, 2017.

[2] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Omer Arik, Ajay Kannan Sharan Narang, Jonathan Raiman and John Miller. Deep Voice 3: 2000-Speaker Neural Text-to-Speech, CoRR, arxiv.org, volume abs/1710.07654, 2017.

[3] Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron C. Courville, Yoshua Bengio: Char2Wav: End-to-End Speech Synthesis. ICLR (Workshop), Toulon, France, 2017.

[4] Taigman, Y., Wolf, L., Polyak, A., Nachmani, E. VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. Journal: CoRR, arxiv.org, volume: abs/1707.06588, 2017.

[5] Ya-Jie Zhang, Shifeng Pan, Lei He and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. CoRR, volume, 1812.04342, 2018.

[6] Akuzawa K., Yusuke, I., and Yutaka, M. Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder. Interspeech, Hyderabad, India, 2018.

[7] Hsu W. N., Zhang Y., R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang. Hierarchical Generative Modeling for Controllable Speech Synthesis. In Proc. International Conference on Learning Representations (ICLR), New Orleans, USA, May 2019.

[8] Wang, Yuxuan, Daisy Stanton, Yu Lin Zhang, Skerry-Ryan R. J., Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren and Rif A. Saurous. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis." ArXiv abs/1803.09017 2018.

[9] Skerry-Ryan, R. J., Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark and Rif A. Saurous. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. ArXiv abs/1803.09047 2018.

[10] Lee, Younggun and Taesu Kim. Robust and Fine-grained Prosody Control of End-to-end Speech Synthesis. IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[11] Parker, J., Stylianou, Y., Cipolla, R. Adaptation of an Expressive Single Speaker Deep Neural Network Speech Synthesis System. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5309-5313, 2018.

[12] Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet. Layer adaptation for transfer of expressivity in speech synthesis. LTC'19 - 9th Language Technology Conference, Poznan, Poland, 2019.

[13] Lin, Xudong, Duan, Yueqi, Dong, Qiyuan, Lu, Jiwen and Zhou, Jie. Deep Variational Metric Learning.The European Conference on Computer Vision (ECCV) pp. 689-704, 2018.

[14] Kaya, Mahmut and BİLGE, Hasan Şakir, Deep Metric Learning: A Survey, Symmetry, volume 11 ISSN, 2073-8994 2019.

[15] Sohn, Kihyuk. Improved Deep Metric Learning with Multi-class N-pair Loss Objective., NIPS 2016.

[16] Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey and Sanjeev Khudanpur. X-Vectors: Robust DNN Embeddings for Speaker Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, 5329-5333, 2018.

[17] Kingma, Diederik P. and Max Welling. Auto-Encoding Variational Bayes. CoRR, arxiv.org, abs/1312.6114 2013.

[18] Rezende, Danilo Jimenez, Shakir Mohamed and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ICML, Beijing 2014.

[19] Bowman S. R., L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, Generating Sentences from a Continuous Space. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, 2016.

[20] Chung, Joon Son, Arsha Nagrani and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. Interspeech 2018.

[21] Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukás Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer and Karel Veselý. The Kaldi Speech Recognition Toolkit. IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

[22] Morise, Masanori, Fumiya Yokomori and Kenji Ozawa. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Transactions 2016.

[23] Yamagishi, Junichi, Pierre-Edouard Honnet, Philip Neil Garner and Alexandros Lazaridis. The SIWIS French Speech Synthesis Database. 2017.

[24] Stan, Adriana, Oliver Watts, Yoshitaka Mamiya, Mircea Giurgiu, Robert A. J. Clark, Junichi Yamagishi and Simon King. TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. Interspeech, 2013.

[25] Streijl, Robert, C., Winkler, S., Hands, D. S. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. Multimedia System. Volume 22.2, 2016.

[26] Dahmani, Sara, Vincent Colotte, Valérian Girard and Slim Ouni. Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis. Interspeech 2019.

[27] Zhizheng Wu, Oliver Watts, Simon King. Merlin: An Open Source Neural Network Speech Synthesis System. In Proceedings 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, USA, 2016.