Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis

Sara Dahmani, Vincent Colotte, Valérian Girard, Slim Ouni

Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis

Sara Dahmani[1], *Vincent Colotte*[1], *Valérian Girard*[1], *Slim Ouni*[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

sara.dahmani@loria.fr, vincent.colotte@loria.fr, valerian.girard@loria.fr, slim.ouni@loria.fr

Abstract

In recent years, the performance of speech synthesis systems has been improved thanks to deep learning-based models, but generating expressive audiovisual speech is still an open issue. The variational auto-encoders (VAE)s are recently proposed to learn latent representations of data. In this paper, we present a system for expressive text-to-audiovisual speech synthesis that learns a latent embedding space of emotions using a conditional generative model based on the variational auto-encoder framework. When conditioned on textual input, the VAE is able to learn an embedded representation that captures emotion characteristics from the signal, while being invariant to the phonetic content of the utterances. We applied this method in an unsupervised manner to generate duration, acoustic and visual features of speech. This conditional variational auto-encoder (CVAE) has been used to blend emotions together. This model was able to generate nuances of a given emotion or to generate new emotions that do not exist in our database. We conducted three perceptive experiments to evaluate our findings.

Index Terms: Expressive audiovisual speech synthesis, conditional variational auto-encoder, Expressive talking avatar, emotion, facial expression, deep bidirectional long short-term memory (DBLSTM)

1. Introduction

Automatically animated 3D Virtual talking heads are gaining great attention lately and are coveted in numerous fields [1, 2, 3, 4]. Expressiveness in speech synthesis systems is increasingly required since it enhances the user experience and makes the interaction more natural [5, 6]. The domain of video games, animation movies as well as educational and medical domains can profit from those advances. Animation of characters nowadays are either created manually by animators, working frame by frame to generate complete animations or by capturing actor's performance with motion capture systems. Those two methods are very expensive and time consuming. Some methods for automatic expressive 3D character animation have emerged taking advantages of the progress in the deep learning area. Xu Li et al. [7] used recurrent network (DBLSTM) to generate audiovisual animation from audio by simply retraining the model with emotion-specific data. Their experiments showed that using neutral corpus can improve the performance of expressive talking avatar generation. Shumin et al. [8] augmented the network input using emotion codes. Zhang et al. [9] used shared hidden layers across multiple emotions, while the output layers are emotion dependent and each head represents a specific emotion characteristics. However, those methods can model only emotion categories present in the training set. Moreover, emotion labels are not always available, and when available they are not completely reliable due to eventual errors of the annotators. More than that, when the emotions are grossly put into very large classes, the notion of nuances disappears and the natural variability in human speech will be lost.

On another hand, the categorical emotion theory postulates that the affect system consists of six basic universal emotions (happiness, surprise, fear, sadness, anger, and disgust)[10]. Yet, the diversity of the human emotions can generate many complex and subtle affective states such as disapproval, depression and contempt that cannot be covered by these basic emotion categories. Furthermore, some research confirms that affective states are not isolated entities, but they are rather systematically connected [11], [12], [13]. Hence, dimensional models regard affective experience as a continuum of non-extreme and highly interconnected states, similar to the spectrum of color [14], [15].

In the work of Hofer et al. [16], a unit selection system was considered to generate nuances of emotions using a database annotated with emotion degrees. In the rule-based emotional voice conversion system, Xue et al. [17] proposes a voice conversion system for emotional speech which utilized two-dimensional (valence and arousal) space to represent emotion in order to control the degree of emotion. The conversion is done by parameterizing and replacing acoustic features related to emotion. Henter et al. [18] succeeded in creating emotion degrees without emotion degree annotations, yet, this work still relies on emotion labels as input.

Different from these methods, this paper addresses the problem of synthesizing expressive speech without relying on emotion labels. Specifically, this paper explores the application of Variational AutoEncoders (VAE) to Text To Expressive Audiovisual Speech Synthesis (TTEAVSS) and shows the possibilities offered by the VAE that makes the blending between emotions possible. VAE was successfully used for extracting speakers specific characteristics from audio [19], in acoustic expressive speech synthesis [20], for emotion representation from audio features [21] and for music generation [22]. To the extent of our knowledge, this is the first time that a Variational Autoencoder (VAE) has been considered for expressive text to audio-visual speech synthesis.

In the first section of this paper, we present a high quality expressive corpus that we created based on linguistic analysis and with motion capture system. After that, we discuss the architecture we choose for the three speech aspects: duration, acoustic and visual. We finally present the result of the perceptive evaluation we made to validate our system and finish with a discussion and a conclusion.

2. Acquisition

In this section we present the protocol we followed to create a high quality expressive audiovisual corpus and the post-processing phases we went through to prepare our data for

neural network training. We recorded a semi-professional actress reading 2000 sentences in a neutral mode (no emotion expressed). From the 2000 sentences we selected 500 sentences to form 6 emotion specific mini-corpus (joy, sadness, fear, surprise, disgust and anger).

2.1. Linguistic analysis

The purpose of this phase is to create a textual corpus with a maximum phonetic coverage while keeping a reasonable amount of sentences. First, we collected more than 7000 French sentences from internal-to-team corpus and open source ones. Then, we used a Greedy algorithm to extract 2000 sentences that have the best coverage rate of French diphones (succession of two phonemes). This corpus was used for neutral speech and a smaller corpus of 500 sentences was created for the six basic emotions to lighten the post-processing phase. The neutral corpus covers 92% of the French diphones and the emotion mini-corpus covers 52% of them. Both covered 100% of French phones.

2.2. Acquisition protocol

As we planed to record the corpus in multiple sessions, it was mandatory to ensure a constant emotion performance for each mini-corpus. To do that, we provided the actress with the desired definition of each emotion. The actress was also provided with a set of scenarios. Those scenarios were important in order to use the acting technique of Stanislavski [23]. The Stanislavski technique allows the actress to dig into her own affects to create an emotion, taking advantage of her emotional memory. The method is often shown as particularly naturalist, as opposed to a more figurative performance [24].

In this work, we presented to the actress three scenarios for each emotion. The actress picked up the scenario that felt the closest to her affective experience. The emotional scenarios (in French) given to the actress were taken from the GEMEP (GEneva Multimodal Emotion Portrayals) Corpus [25].

In the beginning of each session, we glued 65 retro-reflective markers on the actress's face in the same positions (using a 3D printed mask with 65 holes). We put 5 markers on the hat to keep track of the head pose. The actress was seating in front of the cameras while reading clearly the sentences prompted on the screen. The microphone and the chair were placed in a fixed position during all the acquisition sessions to preserve the same acoustic amplitude. In this work, we used only 44 sensors, corresponding to the lower part of the actress's face, since the movements of the upper part of the face are not correlated with speech sounds (phone labels).

2.3. Post-processing

The post-processing task consists of computing the absolute 3D spatial positions of the reflective markers for each frame, after removing the head movement. Concerning the textual corpus, we used Kaldi (a toolkit for speech recognition) to generate accurate phonetic alignment with audio. We have used an in-house alignment Kaldi model trained over 500 hours of French acoustic speech of the ESTER database [26].

3. Neural architecture

In this section, we first introduce VAE and Conditional VAE, which form the basis of this work, and then we present the architecture we use for TTEAVSS.

3.1. Variational Autoencoder

The standard Autoencoder [27] consists of an encoder and a decoder. It learns a latent representation z for a set of input data x by reducing the difference between the generated outputs $\tilde{x}$ of the Autoencoder and the inputs x. Besides the condition of reducing the reconstruction error between x and $\tilde{x}$, VAE [28] introduces an additional condition that forces the latent representation z to follow a Gaussian distribution. The loss of the Variational Autoencoder is as follows:

$$Loss = RE + KL \tag{1}$$

The first term RE is the reconstruction error between x and $\tilde{x}$, it encourages the decoder to learn to reconstruct the data. The second term KL represents the Kullback-Leibler divergence between the encoder's distribution and a standard Normal distribution with mean zero and variance one (the detailed formulas can be find in [28]). It acts as a regularizer that forces the latent distribution to be a normal, which has as effect to bring the latent data clusters closer to each other while maximizing their variance. This behavior encourages a maximum coverage of the latent space and makes it smoother by removing eventual dead zones which makes blending between the different latent vectors possible. In the scope of this work, Variational Autoencoder consists of two neural networks:

1. Emotion embedding network (encoder): neural network that maps input x to the latent representation z to approximate the intractable posterior distribution of the input data.
2. Generative prediction network (decoder): neural network that reconstructs the input variable x from the latent representation z.

A new term β, as shown in equation (2), was initially introduced by Higgins et al. [29] to encourage latent space dimensions disentanglement. It was then used in [30] to balance regularization and reconstruction accuracy. High β values foster regularization at the expense of reconstruction accuracy. In this work the value of this parameter was empirically chosen for each aspect of speech (see section 3.3).

$$Loss = RE + \boldsymbol{\beta}\, KL \tag{2}$$

3.2. Conditional Variational Autoencoder

The conditional VAE (CVAE) is a variant of the VAE that is conditioned on an additional feature $\boldsymbol{c}$. In this work the condition $\boldsymbol{c}$ represents the phone labels corresponding to the input x. We believe that adding this conditional input to the decoder network forces the latent representation to be independent from the textual input. The network should learn to represent features that are not contained in the textual input since the decoder receives this information in addition to the latent representation as inputs.

3.3. Proposed architecture

We use a CVAE to predict : 1) duration 2) acoustic 3) visual data (see Figure 1). We used Merlin TTS system [31] as a basic toolkit for acoustic speech synthesis. We augmented Merlin with a visual synthesis module and a CVAE architecture. In this work we use an asymmetrical CVAE. Since the decoder is not only decompressing the encoder output (latent vector), yet, it computes a more complex non-linear prediction task, we use a

deeper network for the decoder part. No dropout or specific regularization was used to train the three models. Different architectures and β values were used for each model. The encoder and decoder neural networks were trained jointly. For all the models we used a 50 nodes dense layer with linear activation function for the latent variables.

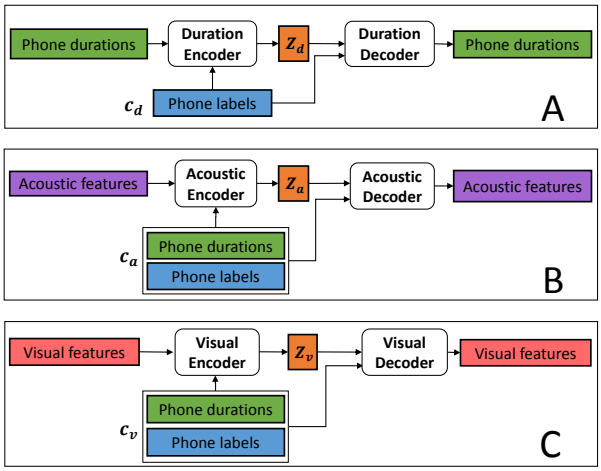

Figure 1: *The encoder-decoder architecture of the three models. A: The duration model is conditioned on phone labels only (c_d). B and C: the encoder-decoder architecture of the acoustic and visual models respectively are conditioned on labels and phone duration (c_a and c_v).*

3.3.1. Duration

This module learns to predict the duration of the phones. One input parameter was given to the network corresponding to the length of the phone. We concatenate this parameter with phone labels to feed the encoder. A single BLSTM layer of 1024 nodes was used as an encoder. The decoder has a single layer of 256 nodes with 'TANH' as activation function followed by a linear output layer. A learning rate of $5 \times 10e^{-4}$ was used, with $\beta = 2 \times 10e^{-5}$.

3.3.2. Acoustic

We extract the recommended Merlin acoustic features, concatenate them with phone labels to feed the encoder. The encoder is a single layer BLSTM network of size 1024. The decoder has two BLSTM layers of 1500 nodes followed by a linear output layer. A learning rate of $10e^{-4}$ was used, with $\beta = 5 \times 10e^{-3}$.

3.3.3. Visual

This module learns to predict 3D (x,y,z) sensors trajectories from phone labels. We give an input of size 132 (44 sensors with x, y and z coordinates) to the encoder with the phone labels. The encoder is a single layer BLSTM network of size 1024. The decoder has two BLSTM layers of 1024 nodes and a linear output layer. We used a learning rate of $5 \times 10e^{-5}$ and $\beta = 0.1$.

4. Synthesis

As shown in Figure 3, at the synthesis phase, the encoders are not used. We choose a vector z_d from the duration latent space, and we give to the duration decoder along with the phone labels to predict their duration. We choose z_a/z_v from the acoustic/visual latent space and with the predicted duration and the phone labels the acoustic/visual data are predicted by the acoustic/visual decoder. The acoustic and visual generated data are synchronized since they are based on the same phone duration. The visual data trajectories are decomposed into blendshape weights to animate a 3D character.

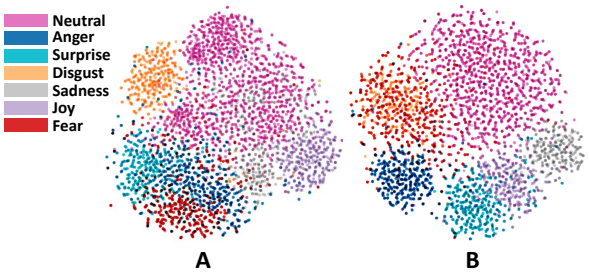

Figure 2: *t-SNE plot [32] of the seven clusters of the latent representations formed by data distribution corresponding to the six emotions and the neutral state. The closest points in the higher dimensional space (latent variables size is 50) are the closest in the projection 2D space. The regularization term pushes data samples to gather around zero meanwhile maximizing their variance. The data samples were clustered differently depending on the nature of data (A: visual and B: acoustic).*

5. Evaluation

To evaluate our system, we conducted three perceptual experiments to validate different results of the CVAE. For each experiment, the generated duration, acoustic and visual data were used to create audiovisual animations of a 3D avatar . Since we animate only the lower part of the avatar's face, we deliberately blurred the upper part of its face to eliminate any unintentional bias caused by its lack of expressiveness. For the three experiments, and for each speech aspect (duration, acoustic and visual) we choose the average z vector of each emotion cluster (ref. Fig 2) to be the representation of the six emotions and the neutral state. We copy-synthesized the original audio files with the same vocoder (WORLD [33]) used for generating synthetic audio files. This is to eliminate bias due by the quality drop caused by the vocoder.

5.1. Generating basic emotions

In this first experiment, we evaluated the ability of our system to generate recognizable emotions. To do that, we choose the center of the each emotion's cluster to generate duration, acoustic and visual features of speech. We presented to 12 participants 10 generated synthetic animations and 10 animations created from original data for each emotion in a random order (total of 140 animations). The participants were asked to choose the emotion corresponding to the animation from a list of seven choices. The results are shown in Table 1.

Table 1: *The diagonal of the confusion matrix for the original (orig) and the synthetic (synt) animations for the six emotions and the neutral state. The values represent the percentages of the correct recognition answers.*

	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Orig	97	67	42	69	77	57	72
Synt	71	83	11	71	92	26	73

5.2. Generating nuances of emotions

The aim of this second experiment was to evaluate the ability of our system to generate nuances of a given emotion. We used a

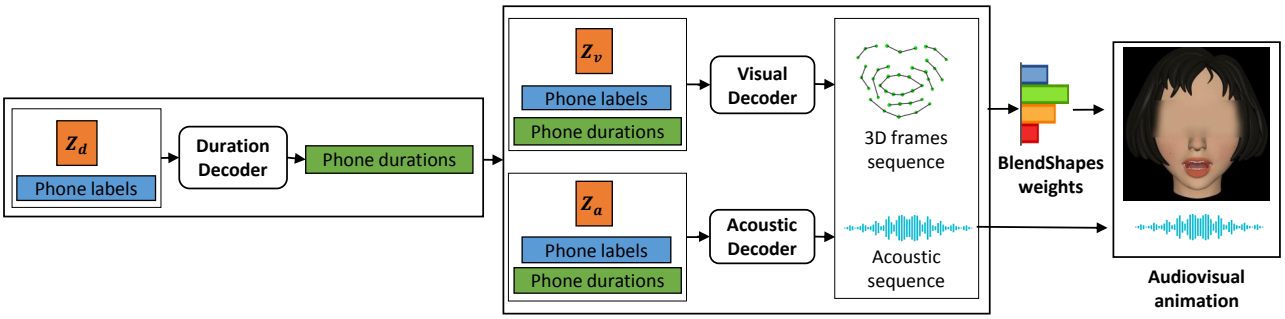

Figure 3: *The architecture of the audiovisual animation system at synthesis phase. Only the decoder part is useful at this stage. The phone labels and the chosen latent vector z_d are given to the duration decoder to predict phones duration. Phone labels, duration as well as the latent vectors z_a and z_v from acoustic and visual latent spaces are passed to the acoustic and visual decoders respectively to generate synchronized audiovisual animation. The upper part of the avatar's face was intentionally blurred.*

latent vector z corresponding to a linear combination between the center of the neutral cluster and the center of the other six emotions. We generate nuances at 33% and 67% of each emotion. We presented a set of animations from a same emotion with different emotion degrees, two by two, to 10 participants and we asked them to choose the animation that was the most expressive according to them. For the six emotions we generated 5 examples, each example results in 6 comparisons (total of 180 comparisons). The results are presented in Table 2.

Table 2: *Percentages of correct answers when comparing emotion nuances two by two for the six emotions. The emotion degrees compared are 100% neutral (represented by 0), 33%, 67% and 100% of a given emotion.*

	0/33	**0/67**	**0/100**	**33/67**	**33/100**	**67/100**
Anger	82	94	90	94	96	88
Disgust	52	80	82	92	86	70
Fear	58	56	80	66	72	80
Joy	74	92	96	90	90	90
Sadness	56	70	88	74	76	86
Surprise	78	92	92	90	94	86
Average	***66***	***80***	***88***	***84***	***85***	***83***

5.3. Generating blended emotions

In this third experiment we evaluated the ability of our system to generate mixtures of emotions by blending emotions together. We showed animations of original and synthetic data at 100% of emotion degree and animations corresponding to blended emotions (50% of $emotion_1$ and 50% of $emotion_2$) in a random order to 12 participants. We asked the participants to estimate the contribution of the blended emotions on a slider having $emotion_1$ and $emotion_2$ as extremities. We generated 5 examples for 4 blending scenarios. Each scenario contains 5 animations (for a total of 100 animations). The results of this experiment are shown in Figure 4.

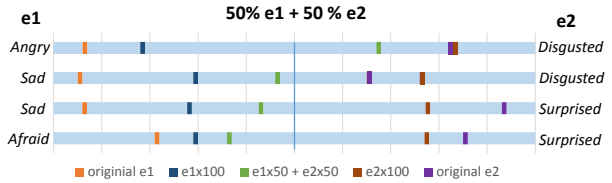

Figure 4: *The generated blended emotion (in green) was perceived as an intermediate emotion between e_1 and e_2 for the four blending scenarios.*

6. Discussion

The results of the first experiment confirm that the synthetic audiovisual animations were highly recognizable for almost all the emotions with more than 71% of recognition rate. Sadness and fear were the hardest to recognize, even for the original animations. This result was expected, since the upper part of the face is crucial for recognizing these emotions [34], [35]. Some synthetic emotions were better recognized than original ones (disgust, joy and slightly surprise). We think this is due to the use of the same latent vector z for all the animations of a given emotion. The participants were able to detect the pattern related to the chosen z and identify more easily the synthetic emotions. It also shows that the latent representation has well captured the specificity of emotions. Recall here, that no label of emotion was used in the learning phase. The emotion label was just used to identify the targeted cluster in the synthesis phase. For the second experiment, in average, the nuances were well identified (more than 80%) for 5 on 6 compared degrees. The subtle nuance (33%) compared with neutral is under 70% mainly due to fear, sadness and disgust low scores. The high recognition scores of different degrees shows that the latent representation gives a good clustering to express nuances by combining neutral with a given emotion. The results of the third experiment show that our system succeeded in creating blended emotions that were correctly perceived as intermediate emotions in the four considered blending scenarios. The choice of the four combinations of these emotions was based on the Plutchick wheel of emotions [36] to obtain coherent combinations (for instance anger and disgust results in contempt).

7. Conclusion

In this paper we applied CVAE to Text To Expressive Audio-Visual Speech Synthesis. We acquired a high quality emotional audiovisual corpus based on a fine linguistic analysis, motion-capture system and naturalist theater techniques. We explored the CVAE architecture for generating duration, acoustic and visual aspects of speech without using emotion labels. The results of our system were validated by three perceptual experiments that confirmed the capacity of our system to generate recognizable emotions. More than that, the generative nature of the CVAE allowed us to generate well detected nuances of the six emotions and to blend different emotions together.

8. Acknowledgements

This work was supported in part by Contrat de plan État / Région Lorraine - LCHN. We thank Grid'5000 platform for providing GPU resources to train our models [37].

9. References

[1] L. Sproull, M. Subramani, S. Kiesler, J. H. Walker, and K. Waters, "When the interface is a face," *Human-Computer Interaction*, vol. 11, no. 2, pp. 97–124, 1996.

[2] I. S. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," *The visual computer*, vol. 15, no. 7-8, pp. 330–340, 1999.

[3] D. M. Dehn and S. Van Mulken, "The impact of animated interface agents: a review of empirical research," *International journal of human-computer studies*, vol. 52, no. 1, pp. 1–22, 2000.

[4] J. Ostermann and D. Millen, "Talking heads and synthetic speech: An architecture for supporting electronic commerce," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 1. IEEE, 2000, pp. 71–74.

[5] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4009–4012.

[6] M. Charfuelan and I. Steiner, "Expressive speech synthesis in MARY TTS using audiobook data and emotionML." in *INTERSPEECH*, 2013, pp. 1564–1568.

[7] X. Li, Z. Wu, H. M. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data." in *INTERSPEECH*, 2016, pp. 1477–1481.

[8] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1613–1616.

[9] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4990–4994.

[10] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[11] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[12] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

[13] R. J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion." 1992.

[14] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[15] J. A. Russell and B. Fehr, "Fuzzy concepts in a fuzzy hierarchy: Varieties of anger." *Journal of personality and social psychology*, vol. 67, no. 2, p. 186, 1994.

[16] G. O. Hofer, K. Richmond, and R. A. Clark, "Informed blending of databases for emotional speech synthesis," 2005.

[17] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[18] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Principles for learning controllable TTS from annotated and latent variation." in *INTERSPEECH*, 2017, pp. 3956–3960.

[19] J. Chorowski, R. J. Weiss, S. Bengio, and A. v. d. Oord, "Unsupervised speech representation learning using wavenet autoencoders," *arXiv preprint arXiv:1901.08810*, 2019.

[20] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," *arXiv preprint arXiv:1804.02135*, 2018.

[21] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[22] E. Çakir and T. Virtanen, "Musical instrument synthesis and morphing in multidimensional latent space using variational, convolutional recurrent autoencoders," in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[23] S. Moore, *The Stanislavski system: The professional training of an actor*. Penguin, 1984.

[24] K. S. Stanislavski and J. Vilar, *La formation de l'acteur*. Payot Paris, 1963.

[25] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[26] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri, "The ester evaluation campaign for the rich transcription of french broadcast news." in *LREC*, 2004.

[27] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[29] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[30] F. Roche, T. Hueber, S. Limier, and L. Girin, "Autoencoders for music sound synthesis: a comparison of linear, shallow, deep and variational models," *arXiv preprint arXiv:1806.04096*, 2018.

[31] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *SSW*, 2016, pp. 202–207.

[32] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[33] 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.

[34] J. N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face." *Journal of personality and social psychology*, vol. 37, no. 11, p. 2049, 1979.

[35] E. Costantini, F. Pianesi, and M. Prete, "Recognising emotions in human and synthetic faces: the role of the upper and lower parts of the face," in *Proceedings of the 10th international conference on Intelligent user interfaces*. ACM, 2005, pp. 20–27.

[36] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[37] D. Balouek, A. C. Amarie, G. Charrier, F. Desprez, E. Jeannot, E. Jeanvoine, A. Lèbre, D. Margery, N. Niclausse, L. Nussbaum *et al.*, "Adding virtualization capabilities to the Grid'5000 testbed," in *International Conference on Cloud Computing and Services Science*. Springer, 2012, pp. 3–20.