Learning emotions latent representation with CVAE for Text-Driven Expressive AudioVisual Speech Synthesis

Sara Dahmani, Vincent Colotte, Valérian Girard, Slim Ouni

Learning emotions latent representation with CVAE for Text-Driven Expressive AudioVisual Speech Synthesis

Sara Dahmani, Vincent Colotte, Valérian Girard, Slim Ouni

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

Abstract

Great improvement has been made in the field of expressive audiovisual Text-to-Speech synthesis (EAVTTS) thanks to deep learning techniques. However, generating realistic speech is still an open issue and researchers in this area have been focusing lately on controlling the speech variability. In this paper, we use different neural architectures to synthesize emotional speech. We study the application of unsupervised learning techniques for emotional speech modeling as well as methods for restructuring emotions representation to make it continuous and more flexible. This manipulation of the emotional representation should allow us to generate new styles of speech by mixing emotions. We first present our expressive audiovisual corpus. We validate the emotional content of this corpus with three perceptual experiments using acoustic only, visual only and audiovisual stimuli. After that, we analyze the performance of a fully connected neural network in learning characteristics specific to different emotions for the phone duration aspect and the acoustic and visual modalities. We also study the contribution of a joint and separate training of the acoustic and visual modalities in the quality of the generated synthetic speech. In the second part of this paper, we use a conditional variational auto-encoder (CVAE) architecture to learn a latent representation of emotions. We applied this method in an unsupervised manner to generate features of expressive speech. We used a

Email addresses: sara.dahmani@loria.fr (Sara Dahmani), vincent.colotte@loria.fr (Vincent Colotte), valerian.girard@loria.fr (Valérian Girard), slim.ouni@loria.fr (Slim Ouni)

probabilistic metric to compute the overlapping degree between emotions latent clusters to choose the best parameters for the CVAE. By manipulating the latent vectors, we were able to generate nuances of a given emotion and to generate new emotions that do not exist in our database. For these new emotions, we obtain a coherent articulation. We conducted four perceptual experiments to evaluate our findings.

Keywords: Expressive audiovisual speech synthesis, conditional variational auto-encoder, Expressive talking avatar, emotion, facial expression, deep learning, bidirectional long short-term memory (BLSTM).

1. Introduction

Automatic animation of expressive virtual talking heads, or audiovisual speech synthesis, is constantly gaining attention due to its important impact on human machine interaction and its benefits to the fields of health and education for instance [1, 2, 3, 4, 5, 6, 7]. Expressiveness in speech synthesis systems has an added value where the interaction is more natural [8, 9]. Acoustic and visual parametric speech synthesis has improved in recent years, particularly in terms of intelligibility [10, 11]. This improvement happened thanks to statistical parametric techniques ranging from HMMs (Hidden Markov Models) to neural networks [12, 13]. In particular, Recurrent Neural Networks have proven to be very adaptable to text-to-speech thanks to their capability of taking into account the past and future information of a sequence [14, 11, 15]. These methods also followed the same evolution for the audiovisual speech synthesis (3D or photo-realistic domain) [11, 16].

Recently, end-to-end systems for acoustic speech synthesis emerged ([17], [18]). Those systems give state of the art synthesis results. Nevertheless, they need a large amount of data to be trained. This kind of corpus is difficult to find for expressive speech, especially in the case of audiovisual speech synthesis. One way to overcome this limitation is by taking advantage of the neutral data available and to link it with the emotional data. For instance, Li *et al.*

[19] 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 the synthesis of expressive talking avatar animations. In the same way, the network input can be augmented using emotion code [20]. Zhang *et al.* [21] used shared hidden layers across multiple emotions, while the output layers are emotion dependent and contains characteristics specific to each emotion. However, those methods can model only emotion categories present in the training set. Furthermore, emotion labels are not always available, and when available they are not completely reliable due to eventual errors of the annotators. Moreover, when emotions are grossly put into very large classes, the notion of nuances disappears and the natural variability in human speech will be lost.

On the other hand, the categorical emotion theory postulates that the affect system consists of six basic universal emotions (happiness, surprise, fear, sadness, anger, and disgust)[22]. But, 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 [23, 24, 25]. Hence, dimensional models regard affective experience as a continuum of non-extreme and highly interconnected states, similar to the spectrum of color [26, 27].

To be able to model emotions in a way that emulates the complexity of the human emotional system, our key insight is to learn, in an unsupervised manner, a latent representation of emotions that is independent of the textual content. This latent representation can be reshaped and manipulated to generate new emotions and speech styles, the same way we can mix primary colors to obtain a wide range of colors. In this work, we consider various aspects of speech. We use different neural networks to model speech phone duration, the acoustic and the visual modalities. after that, we focus more particularly on modeling emotions.

We start by studying the evolution of the quality of the synthesized speech

when training the acoustic and visual modalities separately then jointly. Then we make a cross validation to investigate the ability of the fully-connected architecture to learn characteristics that are specific to each emotion. This step is crucial to establish a baseline that will help us decide which parameters and neural layer type are better for the training of an EAVTTS system. Also, our aim was to ensure that our corpus was appropriate for the speech synthesis task.

As it is detailed in the following sections, our main contributions are the original application of CVAE to an audiovisual corpus and the usage of CVAE on this problem. We show in particular that CVAE can perform emotions interpolation using a large labeled corpus. This architecture learns a latent representation of the emotional space and we propose a method to find the value of a disentanglement coefficient (β parameter). We explain our procedure to reshape the learned latent space to make it malleable and easily manipulable to create new speech styles. Although CVAE has already been shown to be useful in interpolating speaking rate and pitch variation in an audio-only domain (see (Habib et al. 2020) [28]), it has not been shown that emotional interpolation could be done in practice, probably because there is no large corpus with emotion labels, as in our work. We finally present the result of the perceptive evaluation we made to validate our approach.

2. Related work

The first works in DNNs-based acoustic speech synthesis appeared in 2013 and used FeedForward DNNs to model the mapping between linguistic and acoustic features [12, 29, 30, 31]. Later, other studies worked on adding expressiveness to the synthesized voice [32, 20, 33, 34]. Regarding audiovisual speech, some works used DNNs to model emotion categories such as [35] and [16] who used FeedForward DNNs to synthesize expressive audiovisual speech. The two systems obtained satisfactory subjective results and showed that the quality of the results of DNN-based synthesis systems significantly exceeds that of HMMs systems.

Li *et al.* [19] used a recurrent network and compared several BLSTM architectures to adapt a model trained on a large neutral corpus with a small quantity of expressive data. The five proposed systems generate expressive visual animations from audio files. The results of objective and subjective experiments showed that using neutral corpus can improve the performance of an expressive talking avatar generation.

Some researches have been made to compare a joint and separate training of acoustic and visual models. Schabus *et al.* [36] trained an HMMs system for audiovisual speech modeling. This study showed that the joined modeling offers better synchronization between acoustic and visual modalities and that the quality of the predicted acoustic parameters does not undergo degradation compared to the acoustic model trained separately. In a similar study carried out on audiovisual data from a camera, Filntisis *et al.* [16] pointed out that there is no significant difference between the two DNNs models (joint and separate) regarding the realism results of the synthetic video. However, the acoustic results of the separate model were significantly more appreciated than those of the joint model, based on perceptual tests. In this study, our goal is to quantify the contribution of the quality when using a joint model, with objective measures. We note that, in this work, the visual information is 3D visual data acquired using a motion-capture system.

Different from the method cited above, which is able to generate only a specific number of emotion classes, some studies worked on modeling degrees and mixture of emotions. In the work of Hofer *et al.* [37], a unit selection system was considered to generate nuances of emotions using an annotated database with emotion degrees. In the rule-based emotional voice conversion system, Xue *et al.* [38] proposed a voice conversion system for emotional speech which utilized two-dimensional (valence and arousal) space to represent emotions in order to control their degrees. The conversion is done by modifying the acoustic features of neutral speech to create the different types of emotional speech. Henter *et al.* [39] and Zhu *et al.* [34] succeeded in creating nuances of emotions without using emotion degree annotations, nevertheless, this work still relies on

emotion labels as input.

In the second part of this paper, we address the problem of synthesizing expressive speech without relying on emotion labels, in contrast with the described methods above. Specifically, we explore the application of Variational Auto-Encoders (VAE) to Text-To-Expressive Audiovisual Speech Synthesis (TTEAVSS) and show the possibilities offered by the VAE that makes the blending between emotions possible. VAE was successfully used for extracting speakers specific characteristics from audio [40], in acoustic expressive speech synthesis [41, 42], for music generation [43] and to vary the prosody in speech synthesis [44]. The originality of this work is that it considers a Variational Auto-Encoder (VAE) for expressive text to audio-visual speech synthesis.

To improve the latent representation of emotions captured by VAE, we can introduce a parameter for weighting error terms in the loss function of the network [45, 46]. We explain this parameter in detail in the following sections. We can notice that in the work of Higgins *et al.* [45] this parameter was set with a visual inspection of the results, and a metric was proposed to calculate objective scores of dimension disentanglement. Wang *et al.* [47] used a β-VAE to obtain semantically significant and well clustered latent representations, and [48] used it for geologic image interpretation. However, in these studies, the choice of β was not justified. In the work of Alemi *et al.* [49] on image classification, the β parameter was chosen based on classification scores of the considered database. On their work on music synthesis and sounds interpolation, Roche *et al.* [50] experimented four values of β, claiming that the values have to be chosen wisely in order to find the best trade-off between the disentanglement of the latent dimensions and the reconstruction accuracy. However, the smallest value of β was selected ($\beta = 10e^{-6}$) and it is not clear if an even smaller value could have been used. In our work, we present a procedure to choose the appropriate β parameter for expressive speech synthesis and emotions interpolation.

3. Data

In the context of EAVTTS, the quality of the data used in training the models is correlated with the quality of the generated synthetic speech. Therefore, it is important to ensure that the emotions in the corpus are well perceived by humans. In addition, training a synthesis model requires a database of substantial size, containing at least a few hours of speech [51]. Existing expressive databases often only contain the acoustic modality (SynPaFlex, AlloSat, PAVOQUE, etc.). For audiovisual databases, for the most part, the visual modality is represented by 2D video recordings (GEMEP, CVSP-EAV, eNTERFACE'05, MSP-IMPROV, VAM-Video, SAVEE, MODALITY, etc.). Although they are easy and less expensive to record, in these recordings information about the depth of the scene is lost. As a result, certain speech-related gestures, such as the protrusion of the lips, cannot be tracked with precision. Fortunately, a few expressive audiovisual databases containing 3D data exist. For example, the AV-LASYN [52] database which contains a synchronous corpus of audio data and 3D facial marker trajectories, however, this database contains only one emotion and is dedicated to the audiovisual synthesis of laugh only. The IEMOCAP [53] database contains audiovisual sequences recorded with motion capture systems. This database contains recordings of ten actors and several emotions: neutral state, anger, joy, excitement, sadness, frustration, fear, surprise, etc. However, each speaker only recorded 30 minutes of scripted speech (all emotions combined) which is insufficient to train speech synthesis systems. In addition, the number of sentences per emotion is not balanced (neutral 28%, frustration 24%, excitement 17%, sadness 15%, anger 7%, joy 7%, surprise 2% , disgust 1%, the others ¡ 1%) which does not allow comparison of the performance of synthesis systems for the different classes of emotions. The Biwi 3D [54] database offers audiovisual recordings in the form of 3D scan sequences and audio. This corpus is very interesting because it provides complete information on the deformation of the entire face (and not just a selection of points), but it is also small (1109 sentences in total, 14 speakers, around 80 sentences per

speaker) and cannot be used in a synthesis process.

For all the reasons mentioned above, we decided to record our own corpus. In this work, we use the corpus presented in Dahmani *et al.* [55]. It was acted by a semi-professional actress expressing six emotions, in addition to the neutral state. Two thousand sentences were recorded for the neutral state (4 hours of speech). From these 2000 sentences, a subset of 500 sentences was selected for each of the six basic emotions: joy, sadness, anger, surprise, fear and disgust (between 55 min and 1h 53min of speech for each emotion). The linguistic content is identical for all the emotions. The sentences in this corpus have been considered in such a way that they offer a good phonetic coverage. The neutral corpus covers 92% of the French diphones and the sub-corpus of 500 sentences covers 52% of them.

3.1. Corpus validation

After defining the textual content of the corpus, it is important to ensure that the expressed emotions are well perceived and to assess the quality of the expressiveness of the corpus itself, before tackling the synthesis process. Actually, a corpus, containing wrong expressions or that is not sufficiently expressive, can impact the result quality of the synthesis. We performed three perceptual experiments to assess the quality of the expressive audiovisual corpus related to the visual and acoustic modalities.

3.1.1. Stimuli

In these experiments we use video sequences that we recorded in parallel with the 3D visual data. We chose 10 sentences with the most neutral linguistic content and we extracted the corresponding audiovisual sequences for each emotion. Three types of stimuli were presented to the participants for each emotion: 1) acoustic stimuli, 2) visual stimuli and 3) audiovisual stimuli.

3.1.2. Participants

The perceptual experiences counted more than thirty participants for each modality: 1) 34 participants (20 men and 14 women) for the acoustic modality,

2) 31 participants (20 men and 11 women) for the visual modality and 3) 35 participants (23 men and 12 women) for the audiovisual tests. The participants were not native French, but they were living in France during that period of time. Some participants took part in two or three experiments, in this case they had to respect the following order: 1) audio only, 2) visual only then 3) audiovisual experiment. For each experiment, each participant heard/saw 70 stimuli (10 for each emotion and for the neutral state). For the three experiments, the stimuli were presented in a random order to every participant.

3.1.3. Method

We have set up a web application in which participants can log in to perform the perceptual tests. A series of stimuli were presented one by one and participants had to choose from a list of seven choices (neutral, joy, surprise, fear, anger, sadness and disgust) the emotion expressed in the stimuli according to them. Participants had to select an answer and validate it to be able to see the next stimuli. They had the opportunity to replay the stimuli as many times as they wanted. They carried out the test with acoustic stimuli only then visual stimuli only before being able to participate in the audiovisual test.

3.1.4. Results

After collecting the results from all the participants, we computed the statistical significance levels using the p-values from the t-test and we corrected them using Holm Bonferoni method [56]. For each experiment, we used a degree of freedom equal to the number of participants minus one. We considered an alpha equal to 5% and a chance level of 14%. The results are presented in Tables 1, 2 and 3. We added an asterix symbol (*) to the statistically significant recognition rates and (-) for non-significant recognition rates.

We can see that the recognition rates are significant for all the emotions for the three experiments. This means that the emotions are as well carried by the acoustic modality as by visual one. We can notice some confusion between some emotions.

Table 1: The confusion matrices of the recognition rate of the 7 emotions with **the acoustic stimuli**. The columns represent the distribution of the answers given by the participants.

		Perceived emotion						
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**
Produced emotion	**Anger**	**73.24(*)**	8.82(-)	3.82(-)	5.00(-)	3.24(-)	2.35(-)	3.53(-)
	Disgust	4.41(-)	**48.82(*)**	8.53(-)	3.82(-)	17.35(-)	13.53(-)	3.53(-)
	Fear	10.00(-)	10.59(-)	**34.12(*)**	1.47(-)	16.76(-)	22.35(-)	4.71(-)
	Joy	15.59(-)	3.53(-)	5.00(-)	**50.00(*)**	7.35(-)	2.65(-)	15.88(-)
	Neutral	0.29(-)	1.76(-)	2.06(-)	2.94(-)	**81.18(*)**	8.53(-)	3.24(-)
	Sadness	1.76(-)	2.65(-)	13.24(-)	1.18(-)	2.94(-)	**77.06(*)**	1.18(-)
	Surprise	9.71(-)	2.06(-)	3.53(-)	9.71(-)	3.53(-)	2.06(-)	**69.41(*)**

Table 2: The confusion matrices of the recognition rate of the 7 emotions with **the visual stimuli**. The columns represent the distribution of the answers given by the participants.

		Perceived emotion						
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**
Produced emotion	**Anger**	**80.00(*)**	4.84(-)	6.13(-)	0.65(-)	1.94(-)	1.29(-)	5.16(-)
	Disgust	1.94(-)	**75.48(*)**	1.61(-)	2.58(-)	1.94(-)	15.81(-)	0.65(-)
	Fear	13.87(-)	1.94(-)	**63.55(*)**	0.00(-)	0.65(-)	0.32(-)	19.68(-)
	Joy	0.32(-)	0.32(-)	0.32(-)	**91.61(*)**	0.65(-)	0.00(-)	6.77(-)
	Neutral	0.00(-)	0.00(-)	1.29(-)	1.29(-)	**94.19(*)**	1.61(-)	1.61(-)
	Sadness	0.32(-)	0.97(-)	8.39(-)	2.90(-)	28.71(-)	**56.45(*)**	2.26(-)
	Surprise	19.68(-)	0.65(-)	7.74(-)	0.32(-)	1.61(-)	1.61(-)	**68.39(*)**

For the acoustic modality, fear sounded like sadness 22% of the time, while joy sounded like surprise and anger 15% of the time. For the visual modality, sadness is much more confused with neutral (2.94% confusion for acoustic modality vs 28.71% of confusion for visual modality), which indicated that sadness is more carried by the acoustic modality. Conversely, the other emotions all were better perceived with the visual modality, especially fear which became completely distinguishable from neutral and sadness, and joy from anger and surprise. Nevertheless, new confusions appeared between surprise, fear and anger. After statistically analyzing the rate of those confusions, we found that they are all statistically not significant.

Table 3: The confusion matrices of the recognition rate of the 7 emotions with **the audiovisual stimuli**. The columns represent the distribution of the answers given by the participants.

		Perceived emotion						
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**
Produced emotion	**Anger**	**92.57(*)**	2.00(-)	2.29(-)	0.00(-)	0.57(-)	0.29(-)	2.29(-)
	Disgust	1.14(-)	**89.43(*)**	2.00(-)	0.29(-)	1.71(-)	3.43(-)	2.00(-)
	Fear	5.43(-)	3.43(-)	**73.43(*)**	0.29(-)	1.71(-)	3.43(-)	12.29(-)
	Joy	0.29(-)	0.57(-)	0.00(-)	**95.14(*)**	1.14(-)	0.00(-)	2.86(-)
	Neutral	0.00(-)	0.00(-)	0.57(-)	0.00(-)	**97.43(*)**	2.00(-)	0.00(-)
	Sadness	0.57(-)	2.00(-)	4.86(-)	0.57(-)	1.14(-)	**90.86(*)**	0.00(-)
	Surprise	3.71(-)	0.86(-)	4.00(-)	0.86(-)	1.43(-)	0.29(-)	**88.86(*)**

Regarding the audiovisual test, the recognition rates are very high (over 73%) for all emotions and confirm that the acoustic and visual modalities are complementary. These results show that the majority of participants validate the performance of the actress and confirm the good quality of the expressive corpus produced. From these findings we can consider using this corpus for the purpose of expressive audiovisual speech synthesis.

3.2. Data formatting for neural network training

The textual, acoustic and visual data were automatically and phonetically aligned. The linguistic parameters (current phoneme, its previous and following phonemes) represent the input vector for training the three main models: duration, acoustic and visual. The duration of each phoneme is expressed in number of frames considering a rate of 5ms for each frame. For the acoustic parameters, we used the WORLD Vocoder [57] to extract 60 MFCC coefficients (Mel-Frequency Cepstral Coefficients), 5 BAP parameters (Band-Aperiodicity), the fundamental frequency with a logarithmic scale (log F0) and their dynamic parameters (Δ and $\Delta\Delta$) as well as a binary parameter for the voiced/unvoiced nature of the sound in each frame. These parameters were extracted from the audio files every 5 ms. They represent the output of the DNN which will be trained to generate acoustic parameters from the linguistic parameters. For the visual modality, we are focusing on the animation of the lower part of the face

for now. Of all the available data in the corpus, we select the sensors that cover the region related to speech articulation (lips, cheeks, jaw and chin). Similarly, the rate is 5ms for each frame. We have divided the corpus into three subsets: (1) the training set containing 80% of the data, (2) the validation set and (3) the test set, containing 10% of the data each.

4. Audio-Visual Speech synthesis by classical fully-connected architecture

In this section we used a fully-connected architecture with two BLSTM layers to train the three models: acoustic, visual and duration.

4.1. Joint vs. separate modeling

In this experiment, we study the possible contribution of joint training of acoustic and visual modalities on the quality of audiovisual synthesis. We include the six categories of emotions in the learning process. The output vector for the joint model is the result of the concatenation of the acoustic and visual parameters.

Table 4 shows the results obtained with the two models. We note that the joint training of the two modalities degrades all the objective measures, whether it is for the acoustic or visual modality. By performing informal listening tests we found that the acoustic results of the joint model are more distorted with a slightly muffled sound, but for the visual results we didn't notice a humanly perceptible difference.

These observations join the results of [16] that showed, using perceptual tests, that the results of the separate models are considered to be slightly more realistic, but that no statistically significant difference was found between the audiovisual results of the two models. However, Filntisis *et al.* [16] acoustic results of the separate model were considered significantly more realistic than those generated by the joint model. For the visual data, no significant difference was found between the results of the two models.

Table 4: Results of the acoustic and visual parameters of the test subset generated by a model trained with acoustic and visual data separately and jointly.

	Separated models							Joint models (2048)						
	Neu	Joy	Sad	Ang	Sur	Fea	Dis	Neu	Joy	Sad	Ang	Sur	Fea	Dis
	Acoustic (1024)							Acoustic						
MCD (dB)	**4.863**	**5.738**	**5.288**	**5.262**	**5.699**	**5.226**	**5.431**	5.305	6.135	5.740	5.691	6.157	5.669	5.844
BAPD (dB)	**0.224**	**0.312**	**0.269**	**0.268**	**0.287**	**0.231**	**0.256**	0.265	0.359	0.304	0.322	0.335	0.269	0.304
F0-RMSE (Hz)	**26.172**	**46.723**	**32.074**	**39.514**	**32.203**	**40.617**	**35.972**	32.203	47.617	37.972	45.094	45.676	46.201	44.003
F0-Corr	**0.687**	**0.631**	**0.518**	**0.524**	**0.702**	**0.627**	**0.535**	0.683	0.627	0.514	0.513	0.683	0.488	0.518
V/N-V (%)	**6.900**	**10.167**	**7.692**	**8.082**	**9.874**	**7.711**	**9.137**	7.851	11.879	8.955	9.587	11.864	8.814	10.560
	Visual (1024)							Visual						
RMSE (mm)	**1.304**	**1.572**	**1.317**	**1.466**	**1.482**	**1.424**	**2.124**	1.309	1.581	1.320	1.475	1.504	1.429	2.132
Corr	**0.833**	**0.777**	**0.792**	**0.810**	**0.807**	**0.826**	**0.696**	0.829	0.776	0.790	0.808	0.803	0.825	0.689

In fact, we were expecting the model with joint models to perform better than the separate one but the results showed the contrary. One hypothesis behind the drop in the quality of EAVTTS with the joint modeling can be explained by the increase of the number of possible combinations of the input vector. This has the effect of reducing the number of examples for each class (combination) and making the training less effective.

4.2. Cross-validation

In this experiment we use separate acoustic and visual models. We analyze the ability of the models to learn characteristics specific to each emotion, using a cross-validation on duration, acoustic and visual modalities. We note here that the pronunciation of the sentences can change from one emotion to another (more or less pauses, suppression/addition of vowels). Pronunciation differences related to the emotional states are not studied in this work.

For the duration model, we use the linguistic information from the test set of a target emotion to generate the duration of all the other emotions and we calculate RMSE (in frames/phone) and Pearson correlation measures of all the emotions taking the original data of the target emotion as reference. For the acoustic and visual models, we considered the linguistic information as well as the duration of the original data of the target emotion test set. Using this information, we generate the acoustic and visual parameters corresponding to each emotion and calculate the different measurements. For visual modality we

Table 5: Results of RMSE in frames/phone and Pearson Corr of the cross-validation on test sub-set predictions generated by the duration model.

		Duration						
		Neutral	Joy	Sadness	Anger	Surprise	Fear	Disgust
Neutral	RMSE	**5.289**	6.110	6.001	5.917	5.652	6.378	13.280
	Corr	**0.831**	0.799	0.804	0.786	0.803	0.806	0.779
Joy	RMSE	7.346	**7.136**	7.385	7.272	7.206	7.708	14.703
	Corr	0.752	**0.774**	0.756	0.760	0.769	0.751	0.720
Sadness	RMSE	6.886	6.881	**6.606**	7.118	7.176	6.926	13.856
	Corr	0.770	0.765	**0.777**	0.755	0.754	0.770	0.747
Anger	RMSE	6.879	7.130	7.222	**6.463**	7.597	6.578	15.195
	Corr	0.720	0.737	0.728	**0.758**	0.729	0.744	0.686
Surprise	RMSE	6.394	6.905	7.134	6.471	**6.006**	7.532	14.582
	Corr	0.756	0.763	0.741	0.753	**0.781**	0.749	0.708
Fear	RMSE	7.573	7.468	7.287	7.760	7.789	**7.174**	13.578
	Corr	0.767	0.758	0.766	0.756	0.763	**0.781**	0.753
Disgust	RMSE	13.614	15.709	13.669	15.162	14.361	13.146	**9.311**
	Corr	0.728	0.716	0.723	0.693	0.712	0.721	**0.741**

compute RMSE (in millimeters) and Pearson correlation, in the case of acoustic features we compute the mel-cepstral distortion (MCD), the band-aperiodicity distortion (BAPD), the RMSE (F0-RMSE) and the correlation (F0-Correlation) of the F0 as well as the percentage of error on the prediction of voiced/unvoiced frames.

Tables 5, 6 and 7 present the results, where the rows present the result of each considered emotion. The results show that the three models are able to specialize in modeling the different emotions. For the duration model, disgust seems to be very different from the other emotions. This can be explained by specificity of this emotion in our corpus. In fact, this emotion was uttered by with a remarkably slow speech rate. The duration of the corpus of disgust (1h 53min) represents approximately twice the duration of the other emotions (between 55min and 1h 11min). The visual results show similarities between some emotions, in particular between sadness and the neutral state and between anger, fear and surprise. With regard to the acoustic model, we note that

Table 6: Results of RMSE and Pearson Correlation of the cross-validation on test sub-set visual trajectories generated by the visual model. Static column represents a face with a constant neutral expression with a closed mouth.

		Visual							
		Neutral	Joy	Sadness	Anger	Surprise	Fear	Disgust	Static
Neutral	RMSE	**1.304**	2.392	1.635	2.464	1.945	2.245	2.377	2.170
	Corr	**0.833**	0.77	0.801	0.769	0.782	0.805	0.739	—
Joy	RMSE	2.500	**1.572**	2.125	2.703	2.605	2.814	2.732	3.217
	Corr	0.727	**0.777**	0.734	0.722	0.736	0.734	0.712	—
Sadness	RMSE	1.655	2.092	**1.317**	2.378	2.241	2.221	2.325	2.364
	Corr	0.775	0.753	**0.792**	0.723	0.727	0.773	0.713	—
Anger	RMSE	2.604	2.564	2.439	**1.466**	2.100	1.688	3.124	3.308
	Corr	0.732	0.735	0.714	**0.810**	0.783	0.774	0.716	—
Surprise	RMSE	1.984	2.537	2.271	2.046	**1.482**	1.980	2.614	2.817
	Corr	0.750	0.744	0.723	0.785	**0.807**	0.771	0.727	—
Fear	RMSE	2.255	2.778	2.239	1.715	1.883	**1.424**	3.041	3.055
	Corr	0.794	0.772	0.791	0.795	0.790	**0.826**	0.748	—
Disgust	RMSE	2.823	3.160	2.822	3.414	3.063	3.460	**2.124**	3.530
	Corr	0.651	0.647	0.641	0.644	0.651	0.649	**0.696**	—

there is also a resemblance between the neutral state and sadness, between joy and surprise and between fear and disgust. Moreover joy and surprise are the emotions with the greatest difference of F0 compared to neutral and other emotions.

Those findings are in line with what we found in the perceptual study of the original corpus (ref 3.1), except for the similarity between fear and disgust for the acoustic modality in this study. Actually, when studying the recognition rate of the original corpus, the participants rated the whole acoustic performance of the actress which includes speech duration, but in this cross-validation study only the acoustic parameters are studied. As explained earlier, disgust emotion is remarkably characterized by a slow speech rate which may play a major role in distinguishing those two emotions in the original corpus.

This study allowed us to evaluate this baseline architecture for expressive audiovisual synthesis system. We found that this baseline is able to learn characteristics specific to each emotion that are in line with the perceptual results

Table 7: Results of the cross-validation of the test sub-set acoustic parameters of expressive data generated with the acoustic model.

		Acoustic						
		Neutral	Joy	Sadness	Anger	Surprise	Fear	Disgust
Neutral	MCD (dB)	**4.863**	6.409	5.390	5.784	6.548	5.327	5.539
	BAPD (dB)	**0.224**	0.304	0.243	0.268	0.263	0.229	0.232
	F0-RMSE (Hz)	**26.172**	97.521	33.810	42.063	108.220	35.640	38.839
	F0-Corr	**0.687**	0.610	0.598	0.546	0.404	0.558	0.604
	V/N-V (%)	**6.900**	7.565	7.594	7.512	7.612	7.154	7.486
Joy	MCD (dB)	7.010	**5.738**	6.696	6.417	6.132	7.227	7.045
	BAPD (dB)	0.367	**0.312**	0.347	0.330	0.334	0.377	0.371
	F0-RMSE (Hz)	103.444	**46.723**	85.568	97.438	58.942	113.167	109.806
	F0-Corr	0.586	**0.631**	0.552	0.547	0.455	0.526	0.540
	V/N-V (%)	11.015	**10.167**	10.792	10.751	10.547	10.812	10.908
Sadness	MCD (dB)	5.688	6.442	**5.288**	5.825	6.642	5.660	5.817
	BAPD (dB)	0.271	0.301	**0.269**	0.284	0.284	0.271	0.271
	F0-RMSE (Hz)	33.943	82.658	**32.074**	39.353	96.357	47.107	44.815
	F0-Corr	0.503	0.476	**0.518**	0.496	0.284	0.509	0.514
	V/N-V (%)	8.107	8.246	**7.692**	8.167	8.258	7.984	8.023
Anger	MCD (dB)	6.177	6.069	5.793	**5.262**	6.171	6.039	6.040
	BAPD (dB)	0.303	0.287	0.290	**0.268**	0.291	0.303	0.304
	F0-RMSE (Hz)	43.043	89.370	41.357	**39.514**	97.935	44.919	43.601
	F0-Corr	0.440	0.454	0.497	**0.524**	0.357	0.505	0.491
	V/N-V (%)	8.705	8.615	8.515	**8.082**	8.807	8.347	8.495
Surprise	MCD (dB)	6.806	5.916	6.616	6.277	**5.699**	6.951	6.911
	BAPD (dB)	0.305	0.30	0.302	0.301	**0.287**	0.311	0.311
	F0-RMSE (Hz)	102.248	51.066	86.765	97.706	**32.203**	112.539	109.363
	F0-Corr	0.449	0.564	0.394	0.444	**0.702**	0.382	0.391
	V/N-V (%)	10.176	9.769	9.967	10.078	**9.874**	10.007	10.105
Fear	MCD (dB)	5.730	7.015	5.729	6.097	7.075	**5.252**	5.226
	BAPD (dB)	0.246	0.334	0.262	0.297	0.286	**0.234**	0.231
	F0-RMSE (Hz)	37.586	116.012	48.980	41.281	128.206	**32.505**	35.090
	F0-Corr	0.435	0.404	0.487	0.477	0.242	**0.494**	0.627
	V/N-V (%)	7.951	8.254	8.307	8.258	8.352	**7.649**	7.711
Disgust	MCD (dB)	5.995	7.124	5.949	6.250	7.269	5.641	**5.431**
	BAPD (dB)	0.272	0.338	0.279	0.328	0.296	0.265	**0.256**
	F0-RMSE (Hz)	38.890	117.422	46.779	42.528	132.273	36.477	**35.972**
	F0-Corr	0.473	0.439	0.512	0.502	0.299	0.516	**0.535**
	V/N-V (%)	9.350	9.840	9.446	9.837	9.828	9.245	**9.137**

of the original corpus. In the next section we will use the same data with an enhanced architecture to learn a latent representation of emotions.

5. Speech synthesis approach with β-CVAE architecture

In this section, we propose a different synthesis system based on encoder-decoder architecture. The aim of using this architecture is to have control over the internal representation of emotions learned by the network. Being able to control this internal representation, also called latent representation, allow us to access regions of the latent space that remains inaccessible with fully-connected architecture. Also, this architecture allows us to generate new speech styles by mixing available emotions latent vectors. We first introduce VAE, CVAE and β-CVAE, and then we present the architecture we use for our TTS expressive audiovisual speech synthesis system (TTEAVSS).

5.1. Variational Auto-Encoder

The standard Auto-Encoder [58] 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 Auto-Encoder and the inputs x. Besides the condition of reducing the reconstruction error between x and $\tilde{x}$, VAE [59] introduces an additional condition that forces the latent representation z to follow a Gaussian distribution. The loss of the Variational Auto-Encoder 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 found in Kingma *et al.* [59]). 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 Auto-Encoder 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.* [45] to encourage latent space dimensions disentanglement. It was then used in [50] to balance regularization and reconstruction accuracy. High β values foster regularization at the expense of reconstruction accuracy. In this work, we explain the procedure we used to choose the value of this parameter in section 5.4.

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

5.2. Conditional Variational Auto-Encoder

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. Anatov *et al.* [60] showed that conditioning a network on a variable $\boldsymbol{c}$ makes the latent representation independent from this variable. In the work of Skerry *et al.* [61], authors succeeded in transferring prosody from one utterance to another, by isolating prosody from the other speech variations. To do so, they conditioned their network on linguistic content, speaker identity and channel effects (i.e. the recording environment). In a similar way, by adding the conditional input to the decoder network we force 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. By doing so, we can get a latent representation containing feature relative only to the emotional states.

5.3. β-CVAE

5.3.1. Proposed architecture

We use a CVAE to predict: (1) duration, (2) acoustic, and (3) visual data. Figure 1 shows the CVAE architecture for the free models. The duration model conditioned on linguistic data c_d only. The acoustic and visual models are conditioned on the linguistic and on the duration data c_{a_v}. The implementation details of these models are presented in the next section (Section 5.3.2).

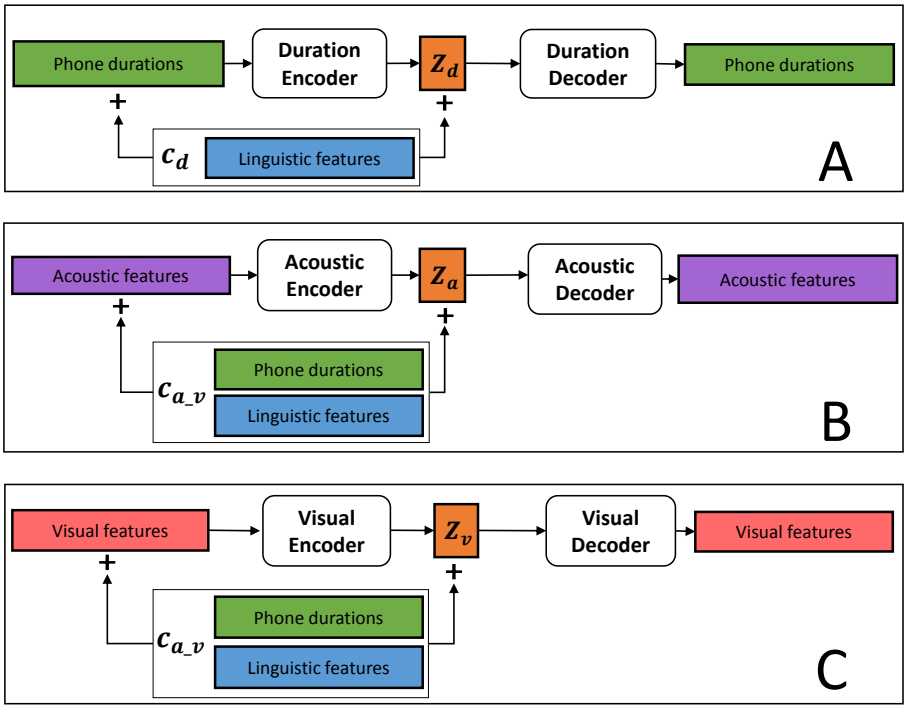

Figure 1: *The CVAE architecture of the three models. A: The duration model conditioned on linguistic data c_d only. B and C: CVAE architecture of the acoustic and visual models respectively, they are conditioned on the linguistic and on the duration data c_{a_v}.*

We also trained a CVAE with audiovisual data to analyze its latent space. It uses c_{a_v} as a condition. The role of the encoder is to extract a compressed latent representation of the input data. In fact, the encoder performs a dimensionality reduction task similar to a PCA. However, in the case of DNNs based encoder, this task is performed in a non-linear way. As we saw above, the encoder is able to encode the information contained in the input data while ignoring the variations contained in the condition $\boldsymbol{c}$. The figure 2 shows the evolution of the training error of the β-CVAE with visual data. We show the result of a configuration where we use the emotion labels as input data, and in the second case without using them.

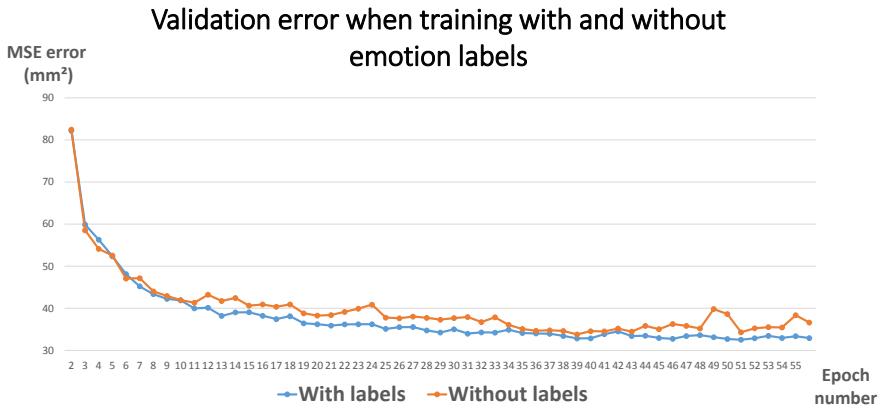

Figure 2: *The impact of removing emotion labels on the training process of the visual model.*

We can see that the network reaches a comparable error value after a certain number of epochs (37 epochs) for the two configurations. Although, training the network with emotion labels makes the learning process more stable. We found similar behavior for acoustic and duration models as well. Those findings are very interesting because they allow us to train the models without providing the explicit information about the emotion categories (labels). We can therefore adopt this unsupervised learning approach to overcome the problem related to annotations. In this work, emotion labels are used only to obtain more explainable graphical plots and to enable data analysis, nevertheless, they are

not used during the training step it-self. In the next sections, we will present the details of the configuration and the results we obtained with this architecture.

5.3.2. Implementation details

We used Merlin TTS system [62] 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 (where the encoder and the decoder have different layers number and size). Since the decoder is not only decompressing the encoder output (latent vector), but, 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. 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. Different architectures and β values were independently chosen for each model. Just below, we describe the final used architecture (and β value). In Section 5.4, we will explain how we have chosen β parameter.

Duration. - This model learns to predict the duration of the phones. One input parameter was given to the network corresponding to the length of the phone (number of frames). 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 $\beta = 2 \times 10e^{-5}$.

Acoustic. - We extract the acoustic features presented in Section 3.2 and 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

$\beta = 5 \times 10e^{-3}$.

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 $\beta = 0.1$.

Audiovisual. - This module learns to predict audiovisual features from phone labels. We concatenate the acoustic and visual and linguistic features to construct the input of the encoder. The encoder is a single layer BLSTM network of size 1500. The decoder had two BLSTM layers of 2048 nodes and a linear output layer. We used a learning rate of $10e^{-4}$ and $\beta = 0$.

5.4. Choice of β based on clusters overlapping optimization

As discussed in Section 5.1, the choice of the parameter β is crucial to obtain a well-structured latent space. A very small β parameter does not allow the clusters of emotions to be close enough to be able to interpolate existing latent vectors. However, a very large value of β will result in a very large overlap of clusters, and prevent the network from learning to reconstruct the data of different emotions correctly. Figure 3 shows the impact of increasing the value of β on the quality of the reconstruction of visual data.

In this section, we propose a procedure to choose the β value for expressive audiovisual speech synthesis. The main goal is to obtain a set of clusters sufficiently close, with a slight overlapping as presented in figure 4 to manage to mix emotions or building different degrees of emotion.

In fact, the different dimensionality reduction and projection techniques are not reliable to judge objectively the state of the latent space. When using different techniques (principal component analysis (PCA), t-SNE, U-map) the

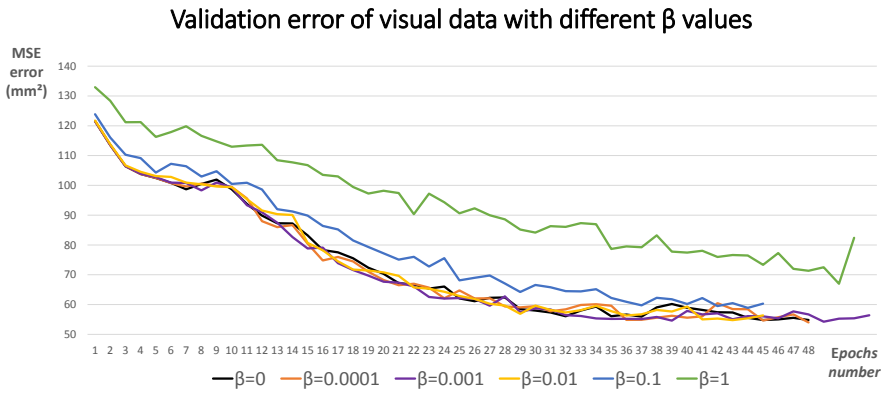

Figure 3: The impact of the gradual introduction of the regularization term (see equation 2) on the quality of the reconstruction of the visual data from the validation set.

results of the projections are different. Moreover, using the same projection technique, the choice of the parameters (number of neighbors of each vector, learning rate, epochs number,...) is crucial and has a significant impact on the structure of the projection of the resulted clusters. Furthermore, when using the PCA technique, for the acoustic modality for instance, the variance that can be seen on a 3D plot corresponds only to 22.9% of the total of the variation. Thus, it is imperative to use a numerical quantification technique of the overlap rate which takes into account all the dimensions of the latent space. The projections can be then used only as visual aid/accompaniment of the numerical results which will allow them to be interpreted.

Inspired by the ecology domain, we suggest the use of the probabilistic method of Swanson *et al.* [64] used actively in ecology, and mainly used to compute the degree of overlapping between species niches and evaluate coexisting and competitive species [65, 66, 67]. This method provides directional estimates of overlap between niches and produces unique projections of multivariate data. Although three-dimensional data were used in this article, the authors explain that the method has no constraints on the number of dimensions considered.

In this work, we want to use this method to compute and quantify the

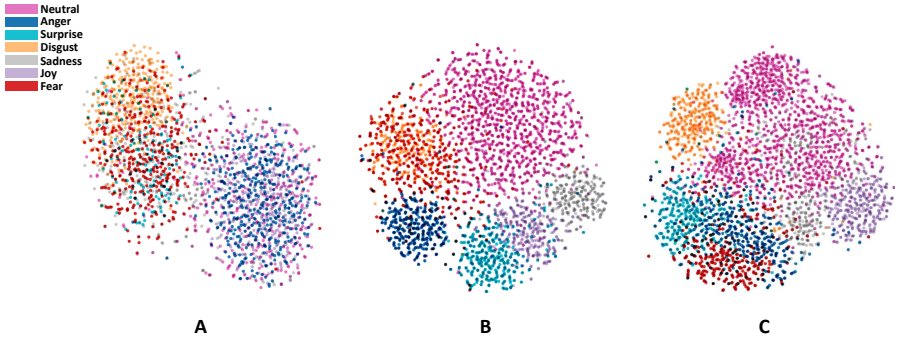

Figure 4: t-SNE plot [63] of the seven clusters of the latent representations formed by data distribution corresponding to the six emotions and the neutral state. The closest points in the higher dimensional space (latent variables size is 50) are the closest in the projection 2D space. The regularization 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: duration, B: acoustic and C: visual).

degree of overlapping between the emotions clusters independently from the projection technique chosen for the visualization. This method computes the intervals of overlapping between the clusters distribution in high dimensions. The overlapping is defined as the probability that an individual from species A to be found in the region specific to species B. In the original work, this method was applied to 3D data. In our work, we apply it to 50-dimensional data (the size of the latent vectors). This method requires a single hyperparameter which is the desired confidence interval of the overlapping, we used the value of 95% in this work. We start with a value of $\beta = 0$ and we increase it gradually until our clusters start overlapping. We stop the procedure when the overlapping metrics show that there is no isolated clusters or subgroup of clusters. To compute the overlapping metric, Swanson's method needs to know the cluster label of each latent vector. In this work we gave the real cluster labels to enable future analysis of the similarity between emotions (considering that more similar emotions will tend to overlap the most). However, when the cluster labels are not known, which is the case of unsupervised learning, it is possible to use the t-SNE plots to identify distant clusters and to give them arbitrary names. In

fact, the final aim of using the Swanson method is not to study the overlapping between specific emotions, but it is to bring distant latent vectors closer to cover dead zones in the latent space and make it smoother.

The Figures 5, 6, 7 and 8 show the distribution of the probabilistic emotion overlapping metric of the different models and represent the probability of emotions displayed in columns overlapping onto those displayed in rows. The distribution means and 95% credible intervals are displayed with continuous and dashed lines respectively.

For visual data, we see in Figure 5 that with $\beta = 0.05$, the latent clusters are not overlapping, which indicates that there may be discontinuities in the latent space. These discontinuities, or dead zones, compromise the possibility of interpolation between the clusters. To fix that, we increase the value of β to push the clusters to be closer. We can see that with $\beta = 0.1$ the clusters are sufficiently close to start overlapping slightly. We can also see that no emotion clusters or subgroup of clusters are isolated and that the different emotion clusters tend to gather around neutral latent cluster.

Concerning acoustic data (Figure 6), a smaller value ($\beta = 5 \times 10e^{-3}$) was enough to obtain sufficient overlapping, especially for the surprise latent cluster that seems to be only connected with joy latent cluster.

When analyzing duration latent space (Figure 7), we could have kept $\beta = 0$, but disgust emotion was isolated from all other emotion clusters, thus, we introduced small β values ($1 \times 10e^{-5}$ and $\beta = 2 \times 10e^{-5}$) until the disgust cluster started overlapping.

In the case of audiovisual data (Figure 8), the clusters are already overlapping with $\beta = 0$, in this case, there is no need to increase its value. Those results are coherent with what we got with the fully-connected architecture. Actually, for the audiovisual model, the high-overlapping degree between the latent clusters without introducing the regularization term ($\beta = 0$), explain the poor performance of the jointly trained model that we noticed in the previous section.

For the other models, we can notice a similarity between the results of the

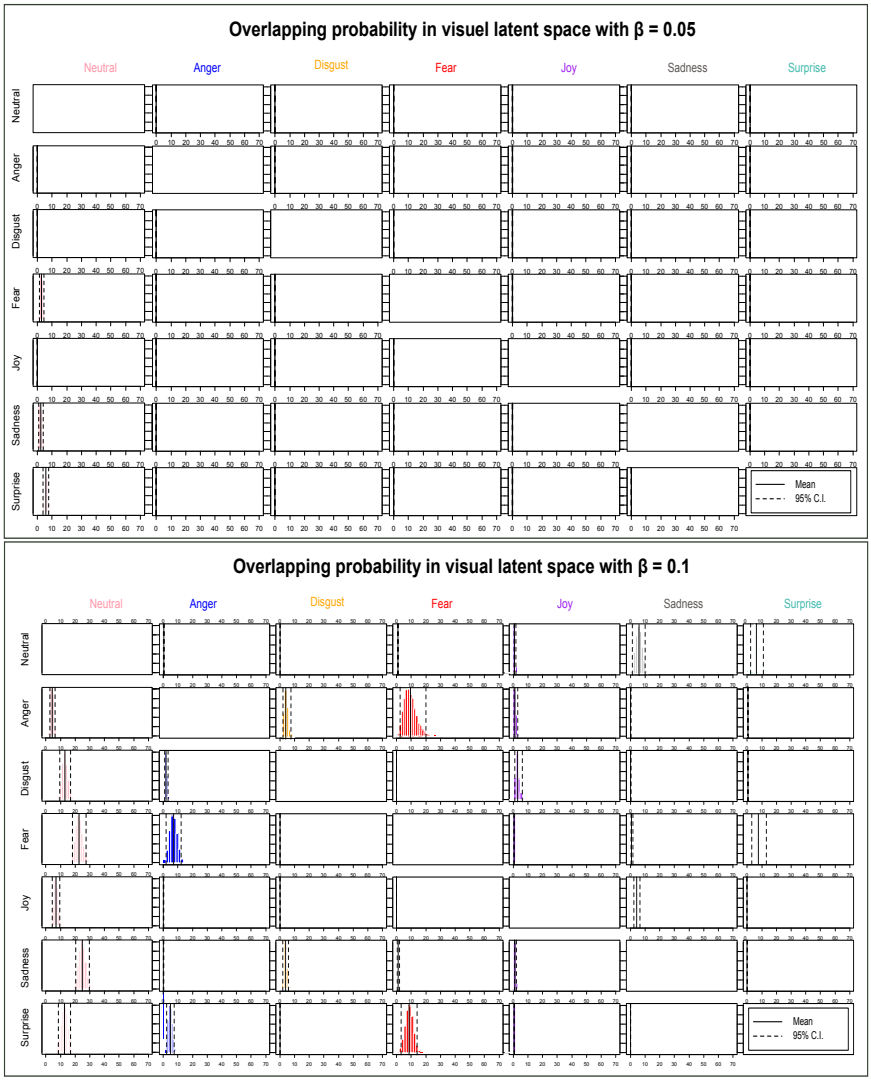

Figure 5: Distribution of the probabilistic emotion overlapping metric for visual modality. The matrix of overlapping distributions are presented for two values of β, 0.05 and 0.1.

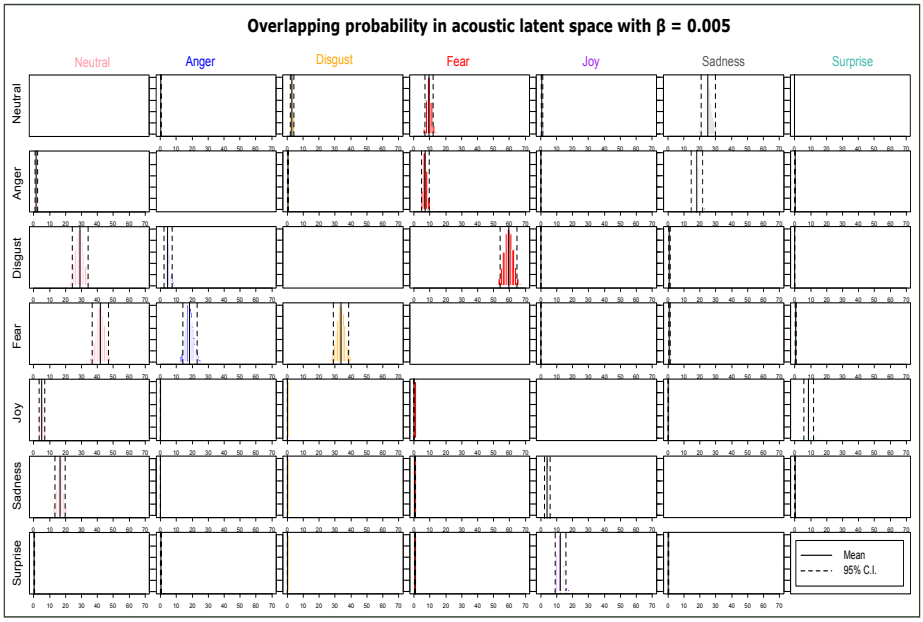

Figure 6: Distribution of the probabilistic emotion overlapping metric for acoustic modality.

cross-validation tables and the overlapping metric results. Actually, for visual data, Table 6 and Figure 5 report both similarities between sadness and neutral, between fear and anger and between surprise and fear. For acoustic data Table 7 and Figure 6 show both similarities between fear and disgust, between neutral and sadness and between surprise and joy. Concerning duration data, Figure 7 confirms that duration for all emotions are very similar, only disgust emotion stands out with a very low speech rate. Thus, disgust latent cluster is barely overlapping with the other emotion clusters.

Our method allows choosing one of the best β values among those selected. Unfortunately, we cannot formally guarantee that it is optimal. Indeed, even if we have run several training for several β values, the very long duration of each training does not allow trying an exhaustive list of β values (for information, about 1 week of computation is needed for 1 model and 1 β value - without optimizing the code). However, we conducted a perceptual experiment for our "best" value of β but it is extremely difficult to multiply this type of experiment

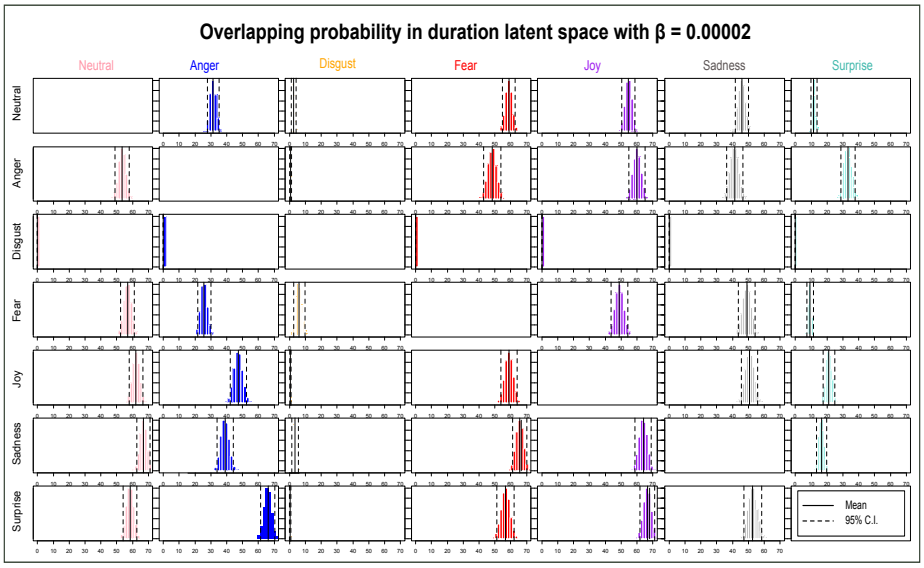

Figure 7: Distribution of the probabilistic emotion overlapping metric for the duration model.

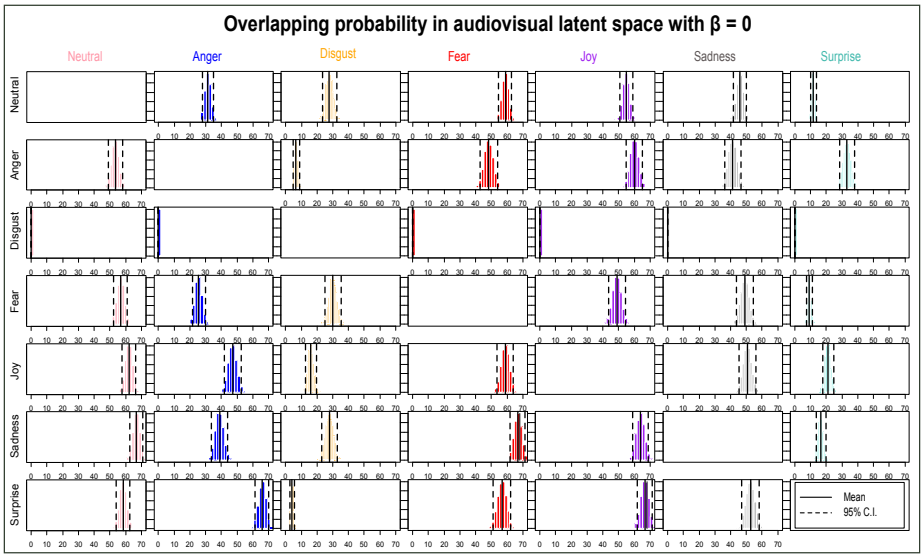

Figure 8: Distribution of the probabilistic emotion overlapping metric for the audiovisual model.

for several values of β. It should be noted that it is clear that the choice of an optimal β requires further investigation.

6. Synthesis: speech generating process

As shown in Figure 9, at the synthesis phase, the encoders are not used. Only the decoder part is useful at this stage. We choose a vector z_d from the duration latent space, and we give it to the duration decoder along with the phone labels to predict their duration. We recall that for training we didn't use the emotion label. The clusters were built automatically. In the synthesis stage, and in order to choose z_d from a cluster, we just need to know few true labels in each cluster. Afterwards, we can choose any z_d of this cluster. That means that we do not need to use a fully labeled corpus, just few labels of the corpus need to be known. z_d is for us, the centroid of the cluster (more explanation can be found in section 7). 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. The upper part of the avatar's face was intentionally blurred to avoid any bias caused by its static state (the upper part is not animated in this work) and to help the participants to focus only on the lower part.

7. Evaluation

To evaluate our system, we conducted four perceptual experiments to validate different results of the CVAE. For each experiment, the generated duration, the acoustic and the 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 four experiments, and for

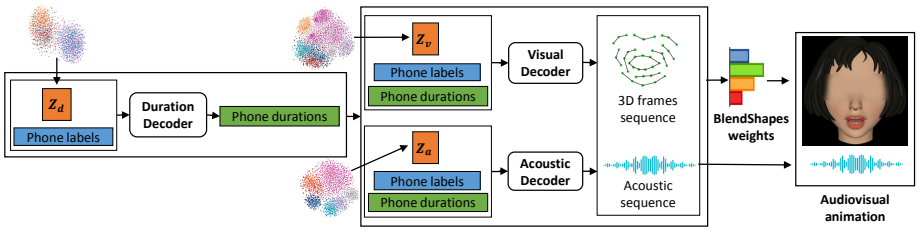

Figure 9: The architecture of the audiovisual animation system at synthesis phase. Phone labels and z_d are given to the duration decoder to predict phones duration. Phone labels, duration as well as z_a and z_v from acoustic and visual latent spaces are passed to the acoustic and visual decoders respectively to generate synchronized audiovisual animation.

each speech aspect (duration, acoustic and visual) we choose the average z vector of each emotion cluster (ref. Figure 4) to be the representation of the six emotions and the neutral state. We copy-synthesized the original audio files with the same vocoder (WORLD [57]) used for generating synthetic audio files. This is to eliminate bias due to the quality drop caused by the vocoder. After collecting the results from all the participants, we computed the statistical significance levels using the p-values from the t-test and we corrected them using Holm Bonferoni method [56].

7.1. Generating basic emotions

In this first experiment, we evaluated the ability of our system to generate recognizable emotions. To do that, we choose the centroid of 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 8 and 9. We added an (*) symbol to statistically significant recognition rates and (-) for non-significant recognition rates in Tables 8 and 9.

The results of this experiment confirm that the synthetic audiovisual animations were highly recognizable for almost all the emotions with more than

Table 8: Confusion matrix for the original animations for the six emotions and the neutral state. The values represent the percentages of the correct recognition answers.

		Perceived emotion						
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**
Emotion produced	**Anger**	**97.50(*)**	0.00(-)	0.00(-)	0.00(-)	0.00(-)	0.00(-)	2.50(-)
	Disgust	0.83(-)	**67.50(*)**	8.33(-)	0.00(-)	0.83(-)	22.50(-)	0.00(-)
	Fear	15.00(-)	5.00(-)	**42.50(*)**	0.00(-)	12.50(-)	22.50(-)	2.50(-)
	Joy	18.33(-)	0.00(-)	0.00(-)	**69.17(*)**	1.67(-)	0.83(-)	10.00(-)
	Neutral	0.00(-)	0.00(-)	4.17(-)	12.50(-)	**77.50(*)**	4.17(-)	1.67(-)
	Sadness	2.50(-)	0.00(-)	32.50(-)	5.00(-)	0.83(-)	**57.50(*)**	1.67(-)
	Surprise	16.67(-)	0.00(-)	0.83(-)	10.00(-)	0.00(-)	0.00(-)	**72.50(*)**

Table 9: Confusion matrix for the synthetic animations for the six emotions and the neutral state. The values represent the percentages of the correct recognition answers.

		Perceived emotion						
		Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**
Emotion produced	**Anger**	**71.67(*)**	15.83(-)	5.00(-)	0.00(-)	5.00(-)	0.83(-)	1.67(-)
	Disgust	1.67(-)	**83.33(*)**	1.67(-)	0.00(-)	3.33(-)	10.00(-)	0.00(-)
	Fear	8.33(-)	11.67(-)	**11.67(-)**	0.83(-)	42.50(-)	20.83(-)	4.17(-)
	Joy	5.00(-)	0.00(-)	3.33(-)	**71.67(*)**	5.00(-)	4.17(-)	10.83(-)
	Neutral	0.00(-)	0.00(-)	1.67(-)	0.83(-)	**92.50(*)**	5.00(-)	0.00(-)
	Sadness	5.00(-)	7.50(-)	15.00(-)	0.00(-)	45.00(-)	**26.67(-)**	0.83(-)
	Surprise	5.00(-)	0.00(-)	4.17(-)	9.17(-)	8.33(-)	0.00(-)	**73.33(*)**

605 71% of the 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 [68], [69]. Some synthetic emotions were better recognized than the original ones (disgust, joy and slightly surprise). We think this is due to the use of the same latent vector z for all the
610 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.

We can also notice the same confusion tendencies between original and synthetic emotions. Confusion was detected between fear and sadness, between joy and surprise, also fear and sadness were seen as neutral state which explain their low recognition rate.

7.2. Assessing generated speech quality

In this experiment, we evaluated the ability of our system to generate coherent articulatory sounds and gestures. To do that, we used the 140 animations generated in the previous experiment. Using a web application, we presented to 19 French-speaking participants, the 140 animations in a random order. For each animation we asked the participants to note the degree of correspondence between the sounds pronounced and the movements of the lips of the avatar. The participants were asked to put the cursor on the adequate position on a slider containing the following coherence degrees: 1) never (0%), 2) rarely (25%), 3) moderately (50%), 4) often (75%) and 5) all the time (100%).

Table 10: The degree of coherence between the pronounced sounds and the movements of the avatar's lips for original and synthetic animations considering a scale of 0% (never) to 100% (all the time).

	Anger	**Disgust**	**Fear**	**Joy**	**Neutral**	**Sadness**	**Surprise**
Original	72.53%	72.76%	77.53%	68.03%	75.44%	72.67%	69.67%
Synthetic	76.57%	71.04%	73.11%	69.86%	78.69%	74.86%	72.60%

The results of this experiment are presented in the tables 10. We present the results for each emotion of the original and the synthetic animations. Those results show that the original and the synthetic animations contain coherent sounds and lip movements. Moreover, using a t-test to compare the results of the two groups of animation (original and synthetic), we found that there is no significant difference between them and this is for all the different emotions. Those findings confirm the quality of the articulation of the synthetic samples generated with our method and confirm that, besides expressing correctly the

different emotions, our system is able to generate coherent sounds and lip movements in an expressive context.

7.3. 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 latent vector z corresponding to a linear combination between the centroid of the neutral cluster and the centroid of the other six emotions. We generate nuances at 33% and 67% of each emotion. We presented a set of animations from the 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 7 comparisons (total of 210 comparisons). The results are presented in Table 11, we added (-) symbol to statistically non-significant scores.

Table 11: Percentages of correct answers when comparing emotion nuances two by two. The emotion degrees compared are 100% neutral (represented by 0), 33%, 67% and 100% and the original animation of a given emotion.

	0/33	**0/67**	**0/100**	**33/67**	**33/100**	**67/100**	**100/original**
Anger	82	94	90	94	96	88	82
Disgust	52 (-)	80	82	92	86	70	86
Fear	58 (-)	56 (-)	80	66	72	80	88
Joy	74	92	96	90	90	90	91
Sadness	56 (-)	70	88	74	76	86	95
Surprise	78	92	92	90	94	86	98
Average	***66***	***80***	***88***	***84***	***85***	***83***	***90***

For this experiment, on average, the nuances order was well respected (66% for 0/33 and >80% for the other comparisons). The high scores of comparisons between neutral and the different nuances (0/33, 0/67, 0/100), show that the emotions are mainly well perceived and easily detectable, especially for the 0/100 comparison. The subtle nuance (33%) compared with neutral is under 70% mainly due to fear, sadness and disgust low scores. Concerning the comparison of the generated nuances between them (33/67, 33/100 and 67/100),

the scores (>80%) show that the graduation of the emotions is successfully represented by our linear combination of latent vectors. The participants were able to perceive the difference between the different emotion nuances and correctly identify the less/more expressive animation. Those results are very interesting, since they prove that we succeeded in restructuring the latent space and making it continuous. Actually, the vectors used to generate the different nuances do not correspond to any previously seen real data. They are in fact newly generated vectors and the nuances of emotions that we generated by linear combination are completely invented. On the other hand, the original animations were seen as more expressive than synthetic ones (at 100%). This result can be explained by the fact that by putting the two animations side by side, the imperfections of the synthetic data become obvious and easily identifiable. Especially since some fine details of the voice are lost during the learning process (trembling, cracking of the voice). Also the original animations have a richer prosody, containing more variability within the same sentence, while our duration model seems to average the duration of the phonemes and results in a more monotonous speech.

7.4. 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 10.

The results of this 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 [70] to obtain coherent

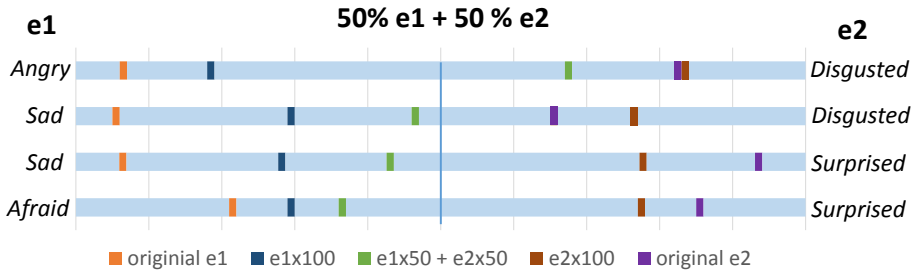

Figure 10: The generated blended emotion (in green) was perceived as an intermediate emotion between e_1 and e_2 for the four blending scenarios.

combinations (for instance anger and disgust results in contempt). As we said in the previous experiment, the vectors that we created by linearly combining latent vectors do not correspond to any real data in our corpus. This is the strong aspect of the VAE as a generative model, since it is able to generate a coherent output from fictive/invented latent vectors. Moreover, we point out that even the linearly combined vectors (the centroids) are themselves non-existent in the original corpus. This result validates again the continuity of the reconstructed latent space. Regarding the original animations, we can see that they were mainly perceived as closer to the emotion definition that the synthetic ones at 100%. The only exception is *disgust*, since the synthetic animations were seen as more disgusted than the original ones. Those results confirm our findings in the first experiment (7.1 Generating basic emotions).

8. Conclusion

In this paper, we studied different neural architecture for Text To Expressive Audio-Visual Speech Synthesis. We first validated the emotional content of our audiovisual corpus with three perceptual experiments. In the first part of this paper, we used a fully connected architecture to study the ability of the network to learn characteristics specific to each emotion. The results of the cross-validation confirm that the baseline architecture is able to learn emotion-specific features that are in line with the recognition rates of the original corpus.

We also found that a joint training of acoustic and visual data degrades the performance of the model. This result is in line with what can be found in the literature.

In the second part of this paper, we applied CVAE to Text To Expressive Audio-Visual Speech Synthesis. We explored the CVAE architecture for generating duration, acoustic and visual aspects of speech without using emotion labels. Inspired by the ecology field, we applied a probabilistic method to compute the overlapping degree between the emotions high dimensional latent clusters. This probabilistic metric allowed us to successfully choose the β parameter of the CVAE. The results of our system were validated by four perceptual experiments that confirmed the capacity of our system to generate recognizable emotions with a coherent articulation. 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. Those results show that we succeeded in well structuring the latent space of the CVAE and making it completely malleable. The latent space became particularly robust, since we were able to generate coherent outputs from new vectors that were created with linearly interpolating real latent vectors.

9. Acknowledgements

This work was supported in part by Contrat de plan État / Région Lorraine - LCHN. We thank Grid'5000 for providing GPUs to train our models [71].

References

[1] L. Sproull, M. Subramani, S. Kiesler, J. H. Walker, K. Waters, When the interface is a face, Human-Computer Interaction 11 (2) (1996) 97–124.

[2] I. S. Pandzic, J. Ostermann, D. Millen, User evaluation: Synthetic talking faces for interactive services, The visual computer 15 (7-8) (1999) 330–340.

[3] D. M. Dehn, S. Van Mulken, The impact of animated interface agents: a review of empirical research, International journal of human-computer studies 52 (1) (2000) 1–22.

[4] J. Ostermann, 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] M. Dworkin, A. Chakraborty, S. Lee, C. Monahan, L. Hightow-Weidman, R. Garofalo, D. Qato, A. Jimenez, A realistic talking human embodied agent mobile phone intervention to promote hiv medication adherence and retention in care in young hiv-positive african american men who have sex with men: qualitative study, JMIR mHealth and uHealth 6 (7) (2018) e10211.

[6] C. J. Falconer, E. B. Davies, R. Grist, P. Stallard, Innovations in practice: Avatar-based virtual reality in camhs talking therapy: two exploratory case studies, Child and Adolescent Mental Health 24 (3) (2019) 283–287.

[7] J. Beskow, On talking heads, social robots and what they can teach us, in: International Congress of Phonetic Sciences ICPhS 2019, 2019.

[8] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, 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.

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

[10] S. King, Measuring a decade of progress in text-to-speech, Loquens 1 (1) (2014) 006.

[11] B. Fan, L. Wang, F. K. Soong, L. Xie, Photo-real talking head with deep bidirectional lstm, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4884–4888.

[12] H. Ze, A. Senior, M. Schuster, Statistical parametric speech synthesis using deep neural networks, in: 2013 ieee international conference on acoustics, speech and signal processing, IEEE, 2013, pp. 7962–7966.

[13] H. Zen, A. Senior, Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis, in: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2014, pp. 3844–3848.

[14] Y. Fan, Y. Qian, F.-L. Xie, F. K. Soong, Tts synthesis with bidirectional lstm based recurrent neural networks, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[15] V. Klimkov, A. Moinet, A. Nadolski, T. Drugman, Parameter generation algorithms for text-to-speech synthesis with recurrent neural networks, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 626–631.

[16] P. P. Filntisis, A. Katsamanis, P. Tsiakoulis, P. Maragos, Video-realistic expressive audio-visual speech synthesis for the greek language, Speech Communication 95 (2017) 137–152.

[17] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards end-to-end speech synthesis, arXiv preprint arXiv:1703.10135.

[18] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499.

[19] X. Li, Z. Wu, H. M. Meng, J. Jia, X. Lou, L. Cai, Expressive speech driven talking avatar synthesis with dblstm using limited amount of emotional bimodal data., in: Interspeech, 2016, pp. 1477–1481.

[20] S. An, Z. Ling, 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.

[21] Y. Zhang, Y. Liu, F. Weninger, 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.

[22] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (3-4) (1992) 169–200.

[23] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (6) (1980) 1161.

[24] R. Plutchik, Emotions: A general psychoevolutionary theory, Approaches to emotion 1984 (1984) 197–219.

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

[26] J. Posner, J. A. Russell, B. S. Peterson, The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Development and psychopathology 17 (3) (2005) 715–734.

[27] J. A. Russell, B. Fehr, Fuzzy concepts in a fuzzy hierarchy: Varieties of anger., Journal of personality and social psychology 67 (2) (1994) 186.

[28] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. Kao, T. Bagby, Semi-supervised generative modeling for controllable speech synthesis, in: International Conference on Learning

Representations - ICLR2020, 2020.
URL https://iclr.cc/virtual_2020/poster_rJeqeCEtvH.html

[29] H. Lu, S. King, O. Watts, Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis, in: Eighth ISCA Workshop on Speech Synthesis, 2013.

[30] Y. Qian, Y. Fan, W. Hu, F. K. Soong, On the training aspects of deep neural network (dnn) for parametric tts synthesis, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 3829–3833.

[31] Z. Wu, C. Valentini-Botinhao, O. Watts, S. King, Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 4460–4464.

[32] K. Inoue, S. Hara, M. Abe, N. Hojo, Y. Ijima, An investigation to transplant emotional expressions in dnn-based tts synthesis, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 1253–1258.

[33] Y. Lee, T. Kim, Robust and fine-grained prosody control of end-to-end speech synthesis, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5911–5915.

[34] X. Zhu, L. Xue, Building a controllable expressive speech synthesis system with multiple emotion strengths, Cognitive Systems Research 59 (2020) 151–159.

[35] J. Parker, R. Maia, Y. Stylianou, R. Cipolla, Expressive visual text to speech and expression adaptation using deep neural networks, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4920–4924.

[36] D. Schabus, M. Pucher, G. Hofer, Joint audiovisual hidden semi-markov model-based speech synthesis, IEEE Journal of Selected Topics in Signal Processing 8 (2) (2013) 336–347.

[37] G. O. Hofer, K. Richmond, R. A. Clark, Informed blending of databases for emotional speech synthesis., in: Interspeech, International Speech Communication Association, 2005, pp. 501–504.

[38] Y. Xue, Y. Hamada, M. Akagi, Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space, Speech Communication 102 (2018) 54–67.

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

[40] J. Chorowski, R. J. Weiss, S. Bengio, A. van den Oord, Unsupervised speech representation learning using wavenet autoencoders, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (12) (2019) 2041–2053.

[41] K. Akuzawa, Y. Iwasawa, Y. Matsuo, Expressive speech synthesis via modeling expressions with variational autoencoder, in: Proc. Interspeech 2018, 2018, pp. 3067–3071. doi:10.21437/Interspeech.2018-1113.
URL http://dx.doi.org/10.21437/Interspeech.2018-1113

[42] S. Latif, R. Rana, J. Qadir, J. Epps, Variational autoencoders for learning latent representations of speech emotion: A preliminary study, in: B. Yegnanarayana (Ed.), Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018, ISCA, 2018, pp. 3107–3111. doi:10.21437/Interspeech.2018-1568.
URL https://doi.org/10.21437/Interspeech.2018-1568

[43] E. Çakir, T. Virtanen, Musical instrument synthesis and morphing in multidimensional latent space using variational, convolutional recurrent autoen-

coders, in: Audio Engineering Society Convention 145, Audio Engineering Society, 2018.

[44] T. Kenter, V. Wan, C.-A. Chan, R. Clark, J. Vit, Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network, in: International Conference on Machine Learning, PMLR, 2019, pp. 3331–3340.

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

[46] J. Yang, G. Lee, S. Chang, N. Kwak, Towards governing agent's efficacy: Action-conditional beta-vae for deep transparent reinforcement learning, in: Asian Conference on Machine Learning, 2019, pp. 32–47.

[47] A. Wang, N. Blair, S. Belkhale, Encouraging categorical meaning in the latent space of a vae.

[48] S. Sen, S. Kainkaryam, C. Ong, A. Sharma, Saltseg: A β-variational autoencoder constrained encoder-decoder architecture for accurate geologic interpretation, in: SEG Technical Program Expanded Abstracts 2019, Society of Exploration Geophysicists, 2019, pp. 2493–2497.

[49] A. A. Alemi, I. Fischer, J. V. Dillon, K. Murphy, Deep variational information bottleneck, ICLR17.

[50] F. Roche, T. Hueber, S. Limier, L. Girin, Autoencoders for music sound synthesis: a comparison of linear, shallow, deep and variational models, IEEE SMC 2019.

[51] D. Guennec, Study of unit selection text-to-speech synthesis algorithms, Ph.D. thesis (2016).

[52] H. Cakmak, J. Urbain, T. Dutoit, J. Tilmanne, The av-lasyn database: A synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis., in: LREC, 2014, pp. 3398–3403.

[53] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (4) (2008) 335.

[54] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, L. Van Gool, Acquisition of a 3d audio-visual corpus of affective speech, IEEE Transactions on Multimedia 12 (6) (2010) 591–598.

[55] S. Dahmani, V. Colotte, V. Girard, S. Ouni, Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis, Proc. Interspeech 2019 (2019) 2598–2602.

[56] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian journal of statistics (1979) 65–70.

[57] M. Morise, F. Yokomori, K. Ozawa, World: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE TRANSACTIONS on Information and Systems 99 (7) (2016) 1877–1884.

[58] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1798–1828.

[59] D. P. Kingma, M. Welling, Auto-encoding variational bayes, ICLR 2014.

[60] A. Atanov, A. Volokhova, A. Ashukha, I. Sosnovik, D. Vetrov, Semi-conditional normalizing flows for semi-supervised learning, arXiv preprint arXiv:1905.00505.

[61] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, R. A. Saurous, Towards end-to-end prosody

transfer for expressive speech synthesis with tacotron, arXiv preprint arXiv:1803.09047.

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

[63] L. v. d. Maaten, G. Hinton, Visualizing data using t-SNE, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

[64] H. K. Swanson, M. Lysy, M. Power, A. D. Stasko, J. D. Johnson, J. D. Reist, A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap, Ecology 96 (2) (2015) 318–324.

[65] L. Chavarie, K. L. Howland, L. N. Harris, M. J. Hansen, C. P. Gallagher, W. J. Harford, W. M. Tonn, A. M. Muir, C. C. Krueger, Habitat overlap of juvenile and adult lake trout of great bear lake: Evidence for lack of a predation gradient?, Ecology of Freshwater Fish 28 (3) (2019) 485–498.

[66] M. C. Jackson, D. J. Woodford, T. A. Bellingan, O. L. Weyl, M. J. Potgieter, N. A. Rivers-Moore, B. R. Ellender, H. E. Fourie, C. T. Chimimba, Trophic overlap between fish and riparian spiders: potential impacts of an invasive fish on terrestrial consumers, Ecology and evolution 6 (6) (2016) 1745–1752.

[67] D. McNicholl, G. Davoren, A. Majewski, J. Reist, Isotopic niche overlap between co-occurring capelin (mallotus villosus) and polar cod (boreogadus saida) and the effect of lipid extraction on stable isotope ratios, Polar Biology 41 (3) (2018) 423–432.

[68] 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 37 (11) (1979) 2049.

[69] E. Costantini, F. Pianesi, M. Prete, Recognising emotions in human and synthetic faces: the role of the upper and lower parts of the face, in: Pro-

ceedings of the 10th international conference on Intelligent user interfaces, ACM, 2005, pp. 20–27.

[70] 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 89 (4) (2001) 344–350.

[71] 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.