TECNIACÚSTICA
CUENCA 2023

A COMPARATIVE STUDY OF NEURAL MODELS FOR EMOTIONAL VOICE CONVERSION

Clara Luzón Álvarez[1,2]
Christian Antoñanzas[1]
Máximo Cobos Serrano[2]

[1]Voicemod, S.L.
[2]Universitat de València

RESUMEN

La conversión de voz emocional es una tarea fundamental en el campo del procesamiento del habla, que permite la modificación del contenido emocional de un mensaje hablado mientras se preserva la identidad del hablante. Este artículo presenta un estudio comparativo preliminar de varios modelos neuronales aplicados a la conversión de voz emocional. Exploramos el rendimiento de modelos de última generación: seq2seq-EVC y CycleGAN-EVC. Para ello, evaluamos estos modelos en diversos conjuntos de datos emocionales y analizamos su capacidad para convertir con precisión emociones a lo largo de un espectro de estados afectivos. Nuestros experimentos preliminares revelan información sobre las fortalezas y limitaciones de cada modelo neuronal en la captura y transferencia de matices emocionales en el habla. Discutimos factores clave como la arquitectura del modelo, el tamaño del conjunto de datos y las estrategias de entrenamiento, arrojando luz sobre los compromisos entre la complejidad computacional y la calidad de la conversión. Además, evaluamos la calidad perceptual de la voz emocional convertida utilizando métricas subjetivas, ofreciendo una visión completa del rendimiento del modelo.

ABSTRACT

Emotional voice conversion (EVC) is a fundamental task in the field of speech processing, enabling the modification of the emotional content of a spoken message while preserving the speaker's identity. This article presents a preliminary comparative study of various neural models applied to emotional voice conversion. We explore the performance of state-of-the-art models: seq2seq-EVC and CycleGAN-EVC. To do so, we evaluate these models on various emotional datasets and analyze their ability to accurately convert emotions across a spectrum of affective states. Our preliminary experiments reveal insights into the strengths and limitations of each neural model in capturing and transferring emotional nuances in speech. We discuss key factors such as model architecture, dataset size, and training strategies, shedding light on the trade-offs between computational complexity and conversion quality. Additionally, we assess the perceptual quality of the converted emotional voice using subjective metrics, providing a complete view of model performance.

Palabras Clave— Emotional voice conversion, deep learning, speech processing

1. INTRODUCTION

During the last few years, more and more new models that accomplish the task of Voice Conversion (VC) have appeared [1,2,3]. These models have revolutionized our ability to transform the acoustic characteristics of speech, enabling personalized voice modification for various applications. However, a notable limitation persists within the majority of these models: the inability to alter the emotional cues conveyed in speech or, in some cases, the absence of emotional cues altogether. To address this critical issue, works have emerged, seeking to achieve Emotional Voice Conversion (EVC) [4]. This transformation represents an extraordinary evolution in the way voice is manipulated, as it introduces the capability to infuse synthesized speech with a variety of emotions that go from joy to sadness, anger to affection. This transformative approach not only opens up exciting prospects for enhancing virtual assistants [5], human-computer interfaces [6], and emotion-aware communication systems [7] but also has profound implications for the entertainment and therapeutic domains [8].

The essence of EVC lies in its capacity to imbue synthesized speech with the emotional characteristics desired by a user while retaining the speaker's inherent identity. Whether it is altering a stern, businesslike tone to convey warmth and empathy in a virtual assistant or transforming a somber narration into one teeming with enthusiasm for an audiobook, EVC offers a rich palette of possibilities. Beyond the realm of entertainment and user interface design, it finds applications in therapy and counseling, where individuals may seek to control and modulate the emotions conveyed in

[1] ***Autor de contacto****: clara.luzon@voicemod.net*

their speech to improve interpersonal interactions and convey empathy effectively.

As the demand for emotionally expressive synthetic speech surges, the role of neural models in driving innovation in this field has become increasingly prominent [9]. These models, which emerged from the deep learning revolution, have demonstrated remarkable capabilities in capturing complex patterns in speech data. In this work, we present a review of two relevant works concerning neural models for EVC, including a preliminary comparative study and assessment of the perceptual quality of their outputs via user-based metrics. More specifically, we focus on two different state-of-the-art models: Seq2Seq [10] and CycleGAN [11]. The chosen models represent two different technique trends, each with its unique strengths and attributes. Seq2Seq, with its sequence-to-sequence architecture, has proven itself as a workhorse in various natural language processing tasks, VC and speech synthesis, promising intriguing possibilities for EVC. On the other hand, CycleGAN, originally designed for image-to-image translation, has been adapted to the audio domain with good results, making it an intriguing contender in the pursuit of emotional voice synthesis.

2. EMOTIONAL VOICE CONVERSION

The realm of EVC has undergone a remarkable evolution over the past few decades, transitioning from classical signal processing techniques to the transformative era of deep learning. This journey has not only broadened the horizons of what is achievable in speech processing but has also significantly enhanced the quality and expressiveness of synthetic emotional speech.

In the early stages of emotional voice conversion, classical signal processing techniques held sway. These approaches primarily relied on the manipulation of spectral features, prosody, and other acoustic attributes to convey emotional variations in speech. Techniques such as pitch modification, time-stretching, and formant shifting were commonly employed to alter the acoustic characteristics of speech. While these methods achieved some level of success, they often faced challenges in preserving the naturalness of the converted speech and maintaining the speaker's identity.

The advent of statistical parametric models marked a significant step forward in the field of emotional voice conversion. Techniques like Gaussian Mixture Models (GMMs) [12] and Hidden Markov Models (HMMs) [13] allowed for the statistical modeling of speech and emotion, providing a framework for capturing and synthesizing emotional prosody. These models could learn statistical relationships between emotional features and acoustic parameters, making them a valuable tool for voice conversion tasks. However, these models had limitations in capturing the intricate nuances of emotional content, often resulting in a somewhat artificial-sounding output. The landscape of emotional voice conversion underwent a seismic shift with the rise of deep learning techniques. Neural networks, particularly deep neural networks (DNNs) [14], such as convolutional neural networks (CNNs) [15] or recurrent neural networks (RNNs) [16] revolutionized the field. These models offered the capacity to automatically learn complex patterns in speech data, enabling the capture and synthesis of emotional content in a more natural and expressive manner. The integration of deep learning techniques into emotional voice conversion has not only improved the perceptual quality of synthetic speech but has also expanded the horizons of what is achievable. With the ability to capture and synthesize emotional nuances, these models are poised to revolutionize various domains, from human-computer interaction to entertainment and therapy.

3. NEURAL MODELS FOR EMOTIONAL VOICE CONVERSION

In the interest of maintaining the paper's conciseness and focus on our experimental findings, we have provided condensed descriptions of the models we compare. Detailed technical specifications of each model can be found in their respective original papers, ensuring that interested readers can explore their architectures, training methods, and nuances in greater depth.

3.1. Seq2Seq-EVC

The Seq2Seq model [10] initially emerged as a solution for machine translation tasks and swiftly demonstrated its effectiveness in speech synthesis and voice conversion challenges. When integrated with attention mechanisms, it not only improves its ability to acquire feature mapping and alignment but also enables the conversion process to concentrate on emotion-relevant regions, marking a significant advancement. Another important advance accomplished with this model is the ability to train using non-parallel and limited data. This EVC framework encompasses five key components, each with distinct functions and architectures:

- *Text encoder* (Et). A convolutional layer stack, followed by a bidirectional LSTM and a fully connected layer, is orchestrated to adeptly transform textual inputs into linguistic embeddings.
- *Seq2seq automatic speech recognition (ASR) encoder*. (Er). Tasked with utilizing acoustic feature sequences for phoneme sequence prediction and automatic alignment. This encoder employs two pyramid bidirectional LSTM layers, while its attention-based decoder features a single-layer LSTM.

[1] ***Autor de contacto****: clara.luzon@voicemod.net*

- *Style encoder* (Es). Responsible for transmuting input acoustic feature sequences into style embeddings, this component features a configuration of stacked bidirectional LSTM layers, average pooling, and a fully connected layer.
- *Auxiliary classifier* (Cs). The auxiliary classifier plays a pivotal role in the framework, primarily focused on adversarial training to selectively remove speaker or emotional information from the linguistic space. In the author's implementation, this classifier adopts a deep neural network (DNN) architecture, meticulously designed to predict outcomes for each input embedding vector.
- *Seq2seq decoder* (Da). Crafted in resonance with the decoder architecture delineated in the Tacotron model, as referenced in [17], this component skillfully reconstructs acoustic sequences by harnessing the combined power of linguistic embeddings and style embeddings.

Table 1 serves as an overview of the structural details for each component within the framework, offering a clear and concise reference for further examination.

Table 1. Summary of the seq2seq framework arquitecture.

Text Encoder	Et	Conv1D-5-512-BN-ReLU-Dropout(0.5) ×3 → 1 layer BLSTM, 256 cells each direction → FC-512-Tanh → Ht
Recognition Encoder	Er encoder	2 layer Pyramid BLSTM , 256 cells each direction, i.e. reducing the sequence time resolution by factor 2
Recognition Encoder	Er decoder	1 layer LSTM, 512 cells with location-aware attention → FC-512-Tanh → Hr
Speaker Encoder	Es	2 layer BLSTM, 128 cells each direction → average pooling → FC-128-Tanh → hs
Auxiliary Classifier	Cs	FC-512-BN-LeakyReLU ×3 → FC-99-Softmax → Pˆs
Seq2Seq Decoder	Da Encoder	1 layer BLSTM, 256 cells each direction
Seq2Seq Decoder	Da PreNet	FC-256-ReLU-Dropout(0.5) ×2
Seq2Seq Decoder	Da decoder	2 layer LSTM, 512 cells with forward attention , 2 frames are predicted each decoder step
Seq2Seq Decoder	Da PostNet	Conv1D-5-512-BN-ReLU-Dropout(0.5) ×5 → Conv1D-5-80, with residual connection from the input to output

Concerning the training strategy, the framework adopts a two-stage training process, summarized in Figure 1. The initial stage, termed "*style initialization*," involves the model's learning to disentangle speaking style from linguistic content. This stage utilizes the VCTK dataset [18], a multi-speaker TTS corpus. The subsequent phase of training, referred to as "emotion training," fine-tunes all components initialized in the first stage using emotional speech data.
During the first stage, the model takes acoustic features (80-dimensional Mel-spectograms) and phoneme sequences as inputs. The text encoder and ASR encoder collaborate to predict linguistic embeddings from these inputs. Finally, the decoder reconstructs the data from the combined style and linguistic embeddings. Notably, the style encoder in this stage learns speaker-dependent information while effectively excluding linguistic information from the acoustic features, aided by an auxiliary classifier.

In the second stage, the previously initialized style encoder assumes the role of an emotion encoder. It embeds the acoustic features into an emotion vector. Simultaneously, the auxiliary classifier, now functioning as an emotion classifier, aids in the removal of emotion-related information from the linguistic space. This dual-stage training approach enables the model to effectively handle both style and emotion aspects, contributing to its overall performance and adaptability.

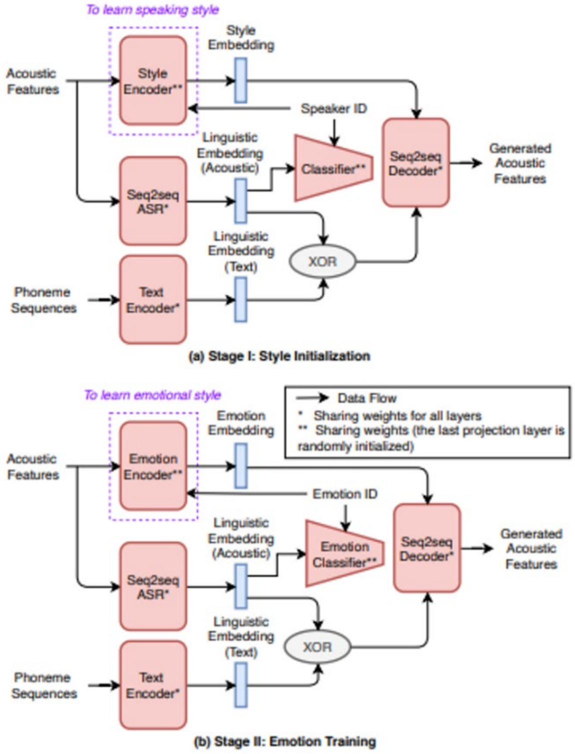

To learn speaking style
Style Embedding
Acoustic Features
Style Encoder**
Speaker ID
Linguistic Embedding (Acoustic)
Classifier**
Seq2seq Decoder*
Generated Acoustic Features
Seq2seq ASR*
Linguistic Embedding (Text)
XOR
Phoneme Sequences
Text Encoder*
(a) Stage I: Style Initialization
To learn emotional style
Data Flow
* Sharing weights for all layers
** Sharing weights (the last projection layer is randomly initialized)
Emotion Embedding
Acoustic Features
Emotion Encoder**
Emotion ID
Linguistic Embedding (Acoustic)
Emotion Classifier**
Seq2seq Decoder*
Generated Acoustic Features
Seq2seq ASR*
Linguistic Embedding (Text)
XOR
Phoneme Sequences
Text Encoder*
(b) Stage II: Emotion Training

Figure 1. Seq2Seq training stages. Extracted from [10].

[1] ***Autor de contacto**: clara.luzon@voicemod.net*

3.2. CycleGan-EVC

CycleGAN operates as a generative adversarial network (GAN), which consists of two neural networks, the generator (G) and the discriminator (D), engaging in a competitive learning process [11]. The generator attempts to synthesize emotionally expressive speech from one emotion domain to another, while the discriminator seeks to distinguish between real and generated emotional speech. Through adversarial training, the generator becomes increasingly proficient at converting emotions in speech, creating highly convincing emotional voice transformations.
The network architecture is concisely presented in Table 2. For the discriminator, a 2D CNN architecture is employed. In contrast, the generator was designed using a one-dimensional (1D) CNN, tailored to capture relationships among overall features while preserving the temporal structure of the speech. The network architecture incorporates a combination of downsampling, residual, and upsampling layers, along with the integration of instance normalization techniques. To achieve upsampling, the authors opted for the pixel shuffler method, known for its effectiveness in high-resolution image regeneration tasks.

The CycleGAN model incorporates three crucial loss functions, each with specific purposes that collectively guide the learning process for both forward and inverse mappings between the source and target domains, ensuring effective and consistent transformations between these two domains:

Table 2. CycleGan architecture summary.

Discriminator	D	Conv2D-[3x3]-128x2 → 1 layer glu → Conv2d - Instance Normalization- Conv2d – Instace normalization- glu X3
Generator	G	Conv1D-15-128 x 2 -> 1 layer glu → conv1d - instance Normalization – conv1d- instance Normalization -glu x2-> (Conv1d -Instance Norm – GLU-Conv1d -Instance Norm) with residual connection x6 → Conv1d-PixelShuffler-Instance Norm- GLU x2 → Conv 1d

- *Adversarial Loss:* This loss function measures the distinguishability between the data distribution of the generated data and the distribution of the source or target data. The aim is to make the generated data as indistinguishable as possible from real data. A smaller adversarial loss indicates a closer match between the generated and real data distributions, leading to more convincing results.
- *Cycle-Consistency Loss*: The cycle-consistency loss enforces a vital property in CycleGAN. It ensures that if we apply the mapping function from one domain to the other and then reverse the process, we should obtain the original data. This loss encourages the model to maintain consistency and fidelity in the transformations it performs, preventing undesirable artifacts.
- *Identity-Mapping Loss*: The objective of the identity-mapping loss is to preserve linguistic information during the conversion process without introducing external alterations. It aims to ensure that the core linguistic content remains unchanged in the converted data, emphasizing the importance of retaining the original meaning and semantics.

The training phase of the CycleGAN EVC framework is depicted in Figure 2. In this framework, 24-dimensional Mel-cepstral features are employed for spectrum conversion training. Additionally, for prosody training, 10-dimensional F0 (fundamental frequency) features are utilized for speech frames. Notably, the prosody features are computed using the Continuous Wavelet Transform (CWT), which decomposes the F0 signal into various variations across multiple time scales. This modeling approach allows for the comprehensive representation of F0, capturing nuances from micro-prosody levels to the entirety of the utterance. During the training stage, it is important to note that both the target and source speeches originate from the same speaker. However, the emotions and linguistic content differ between these two sets of data. This distinction is a critical aspect of the training process, as it facilitates the model's ability to convert emotional characteristics while preserving the speaker's identity. Furthermore, the inclusion of the WORLD vocoder [19] (D4C edition [20]) within the framework is critical for fundamental computations. It performs spectral feature analysis, F0 estimation, and aperiodicity calculation for the input utterance. Additionally, it plays a central role in the final step of speech resynthesis following feature conversion.

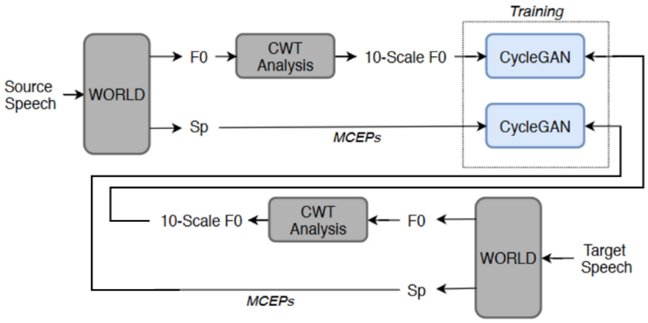


Figure 2. Training phase of CycleGAN. Extracted from [11].

4. DATASETS

In the context of EVC, the availability and quality of datasets hold a paramount position in shaping the performance and capabilities of the models. Some well-known datasets are RAVDESS [21], IEMOCAP [22], ESD [23], CREMA-D [24] and EmoV-DB [25]. Following a comprehensive review of existing emotional speech databases, we made the decision to

[1] ***Autor de contacto***: *clara.luzon@voicemod.net*

proceed with the Emotional Speech Databases (ESD) for the development of our comparative study. ESD is designed to address several limitations found in existing emotional speech dataset. It is characterized by its multilingual nature, containing parallel and acted emotional speech recordings. The dataset is thoughtfully constructed, featuring contributions from 10 native English speakers and 10 native Chinese speakers. Each of these speakers has provided 350 utterances, collectively covering five distinct emotion categories: Neutral, Happy, Angry, Sad, and Surprise. All the speech data in ESD database is recorded in a typical indoor environment with an SNR of above 20 dB and a sampling frequency of 16 kHz, ensuring suitability for building state-of-the-art EVC frameworks. ESD's unique combination of attributes, including its multilingual nature, diverse emotion categories, and meticulous construction, aligns perfectly with the objectives of our research.

6. EXPERIMENTS

In our evaluation, we have focused on comparing the performance of the two considered models, Seq2Seq and CycleGAN, for specific emotional conversions. Seq2Seq has been trained for the neutral-to-happy and neutral-to-sad conversions, while CycleGAN has been trained for the neutral-to-happy and neutral-to-surprise transformations. We recognize that this divergence in training pairs presents a challenge in directly comparing the two models in all aspects. The selection of these emotional transitions was motivated by the desire to assess their practical significance. Happy and sad emotional states are commonly encountered in a wide range of applications, from human-computer interaction to entertainment, while surprise and neutral serve as valuable control states for comparison. This approach allows us to directly compare Seq2Seq and CycleGAN for the neutral-to-happy conversion, offering insights into their relative performance. Additionally, we provide supplementary evaluations, showcasing Seq2Seq's capabilities in the neutral-to-sad conversion and CycleGAN's performance in the neutral-to-surprise conversion. These indirect comparisons may not facilitate a direct model-to-model assessment, but they provide valuable data on the models' adaptability to diverse emotional transitions and their ability to preserve speaker identity. While we acknowledge this limitation, our comprehensive evaluation across multiple emotional states contributes to a more holistic understanding of the strengths and limitations of these models in practical applications.

Unlike some tasks in speech processing, where a clear ground-truth or reference exists for evaluating correctness, emotional voice conversion lacks a definitive benchmark. This is due to the inherent subjectivity of emotional expression in speech. Consequently, our assessment places a particular emphasis on subjective evaluation by human listeners, who can provide valuable insights into the naturalness and emotional expressiveness of the converted voice. While we recognize the intricacies of evaluating emotional voice conversion, our approach aims to provide a comprehensive and perceptually meaningful evaluation, acknowledging the subjective nature of the task.

To address the challenge of assessing the subjective quality of emotional voice conversion, we have devised a rigorous evaluation methodology. In our subjective evaluation, we will engage a panel of 10 human subjects, carefully selected for their diverse backgrounds and listening experiences. These individuals will play a pivotal role in providing valuable perceptual judgments of the converted speech. To ensure robust and reliable evaluations, we will employ a widely accepted metric known as the Mean Opinion Score (MOS). The MOS allows our subjects to assign numerical ratings to the quality of the converted speech, effectively quantifying their subjective perceptions of naturalness, emotional expressiveness, and overall quality. This comprehensive and well-established evaluation method, coupled with a diverse panel of evaluators, enables us to capture a broad spectrum of human perceptions, thus enhancing the reliability and validity of our subjective assessment.

5.1. Subjective Evaluation

The MOS is a standardized and widely-used method for assessing the perceived quality of audio or speech signals through subjective human evaluation. It typically employs a numerical scale, with 1 being the lowest score (indicating poor quality) and 5 being the highest (indicating excellent quality). Listeners are asked to assign a MOS rating to each converted voice sample, reflecting their judgment of the overall quality and naturalness of the speech. During the evaluation process, our panel of 10 listeners will individually assess a set of converted voice samples. Each listener will listen to multiple converted speech samples generated by the models under study, encompassing different emotional conversions. After listening to each sample, they will assign a numerical MOS rating based on their perception of the voice's quality. These ratings will be collected for each sample. To ensure consistent and reliable evaluations, our listeners have undergone a training phase. During this phase, they were familiarized with the task and the MOS scale, ensuring that they understand the range and meaning of the scores. Additionally, they were exposed to a set of reference samples representing different emotional expressions, allowing them to calibrate their perceptions and align their assessments with a common reference point. Listener training is a critical step in minimizing inter-listener variability and ensuring that evaluations are as objective as possible.

During the actual evaluation experiments, listeners were asked to assess the quality of converted speech samples. They

[1] ***Autor de contacto**: clara.luzon@voicemod.net*

were provided with clear and concise instructions, emphasizing that they should focus on evaluating the naturalness and emotional expressiveness of the converted speech. Specifically, they were instructed to consider aspects such as how well the emotional content was conveyed while preserving the identity of the speaker. Listeners were encouraged to provide honest and unbiased assessments, using the full range of the MOS scale to reflect their judgments accurately.

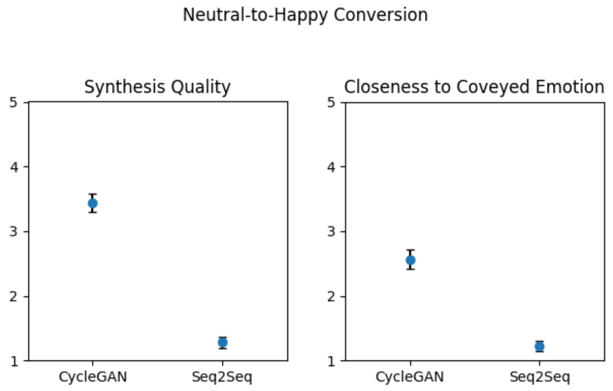
5.2. Direct Model-to-Model Evaluation: Conveying Happiness and Positive Emotions

In our evaluation, we have directed particular attention to the direct comparison of two key models, Seq2Seq and CycleGAN, with a specific focus on the neutral-to-happy emotional conversion. Both Seq2Seq and CycleGAN have undergone training for this transformation, enabling a direct model-to-model assessment. This particular emotional conversion holds significant practical relevance, as it frequently appears in applications such as enhancing virtual assistants, video games, and human-computer interfaces, where conveying happiness and positive emotions is essential. By concentrating our evaluation on this shared conversion, we gain valuable insights into the relative performance of Seq2Seq and CycleGAN, enabling a direct and meaningful comparison of their capabilities in conveying positive emotional states while preserving the speaker's identity. This focused evaluation serves as a key benchmark for assessing the models' proficiency in a commonly encountered emotional transformation, providing essential insights into their respective strengths and limitations in practical applications.

The findings from our analysis are presented in Figure 3, where the mean and standard error of the mean are used to represent the results. The graph on the right showcases the results pertaining to the synthesis quality of the two models, while the graph on the left illustrates the results related to the closeness to the conveyed emotion.

When it comes to synthesis quality, CycleGAN surpasses the Seq2Seq model achieving a more natural-sounding speech. We attribute this improvement to CycleGAN's utilization of the Continuous Wavelet Transform (CWT) applied to the fundamental frequency (f0) and Mel-cepstral features as acoustic characteristics, in conjunction with the WORLD vocoder for audio synthesis. In contrast, the Seq2Seq model relies solely on mel-spectrograms as acoustic features and employs the Griffin-Lim algorithm for audio synthesis.

When it comes to capturing the conveyed emotion, once again, the CycleGAN algorithm outperforms the Seq2Seq model. However, it's worth noting mentioning that neither of these models appears to effectively convey the happy emotion in a readily recognizable manner.



5.3. Diverse Emotional Transformations: Assessing Model Performance Across Varied Expressions

In our evaluation, we extend our assessment to encompass scenarios where Seq2Seq is employed for the neutral-to-sad transformation, while CycleGAN is utilized for the neutral-to-surprise emotional conversion. These selected emotional transitions, though distinct from the direct comparison mentioned earlier, are equally essential in practical applications. Surprise is a common emotional state encountered in various contexts, including virtual assistants and gaming, where eliciting user engagement and responses are paramount. Conversely, the neutral-to-sad transformation finds relevance in applications such as narrative storytelling and therapeutic interventions, where conveying empathy and solemnity is crucial. While this evaluation does not permit a direct model-to-model comparison, it provides valuable insights into the adaptability and effectiveness of Seq2Seq and CycleGAN across diverse emotional transitions. By assessing their capabilities in conveying surprise and sadness while maintaining speaker identity, we gain a nuanced understanding of the models' performance in handling a range of emotional expressions, further enriching our comprehensive evaluation of their practical applicability.

The outcomes of our analysis are showcased in Figure 4, following a similar format to our previous analysis, with the utilization of mean values and standard errors of the mean. The graph on the left illustrates the results obtained from the CycleGAN model when transitioning from a neutral state to a state of surprise, while the graph on the right portrays the results from the Seq2Seq model when executing the transformation from a neutral state to a state of sadness.

The results for the CycleGAN model in terms of synthesis quality show similarities when transforming to a happy emotion compared to the surprise emotion. However, the model appears to represent the surprise emotion more effectively. We attribute this improvement to the variations in the fundamental frequency (f0) associated with different emotions. When transitioning to happiness, the f0 changes more rapidly, leading to higher-pitched speech, which can introduce artifacts that may mask the emotion. In contrast, the

[1] ***Autor de contacto**: clara.luzon@voicemod.net*

surprise emotion, while still exhibiting some of these aspects, is not as intense in terms of f0 variations, contributing to a more accurate representation.

Similarly, for the Seq2Seq model, we observed similar results. The synthesis quality resembles that of transforming to a happy emotion, but it seems to better represent the sad emotion. This can be attributed to the same reasoning as before: sad speech typically lacks rapid fundamental frequency (f0) variations, and the f0 values tend to be lower compared to the happy emotion. As a result, the emotion is less likely to be masked by artifacts stemming from these characteristics, leading to a clearer representation of sadness.

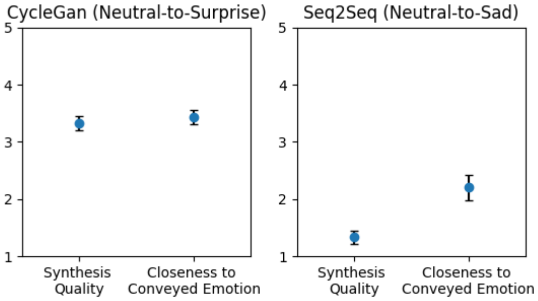


Figure 4 shows the MOS, where bars indicate the standard error of the mean, for the diverse emotional transformation evaluated by the listeners: CycleGan Neutral-to-Surprise transformation and Seq2Seq Neutral-to-Sad transformation.

6. CONCLUSIONS

In summary, our study has explored the capabilities of Seq2Seq and CycleGAN models in the context of emotional voice conversion (EVC). Through a comprehensive evaluation, we have gained valuable insights into their performance across various emotional transformations. In a direct comparison for the neutral-to-happy conversion, CycleGAN exhibited superior ability to imbue natural and emotionally expressive qualities into the converted speech. Conversely, Seq2Seq demonstrated suboptimal synthesis quality, suggesting room for improvement. Our supplementary evaluations CycleGAN's commendable adaptability in the neutral-to-surprise transformation and revealed improvements in Seq2Seq's effectiveness in the neutral-to-sad conversion. While these indirect comparisons did not facilitate direct model-to-model assessments, they underscored the models' versatility in handling diverse emotional transitions.

However, it's crucial to emphasize that both models, when employed appropriately, offer promising avenues for enhancing applications spanning human-computer interaction, entertainment, and therapy. Notably, the absence of a ground-truth for emotional voice conversion remains a challenge, underscoring the subjective nature of quality assessment in this field. Our employment of a Mean Opinion Score (MOS) approach, bolstered by a diverse panel of evaluators, serves as a commendable effort to address this challenge. Despite the inherent subjectivity, MOS evaluations provided robust insights into perceptual quality. As we conclude, our findings contribute to a holistic understanding of the strengths and limitations of Seq2Seq and CycleGAN in practical EVC applications. They also emphasize the need for continued research in the evolving field of emotional voice conversion, where innovation holds the promise of more emotionally expressive and identity-preserving synthetic voices.

ACKNOWLEDGEMENTS

Work partially funded by project CIAEST/2021/10 from Generalitat Valenciana and projects TED2021-131003B-C21 and PID2022-137048OB-C41 from the Spanish Ministry of Science and Innovation.

12. REFERENCIAS

[1] B. Sisman, J Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 132-157. 2020.
[2] Z. Wu and H. Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and*
Information Processing, vol. 3, pp. e17, 2014.
[3] S. H. Mohammadi and A.r Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.
[4] K. Zhou, B. Sisman. R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD." *Speech Communication*, 137, 1-18, 2022.
[5] H. J. Do, and W. T. Fu, "Empathic virual assistant for healthcare information with positive emotional experience," in 2016 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 318-318). IEEE, 2016.
[6] S. Brave, and C. Nass, "Emotion in human-computer interaction." *Human-computer interaction fundamentals*, 20094635, 53-68, 2009.
[7] M. Chen, Y. Zhang, Y. Li, S. Mao and V. C. Leung, "EMC: Emotion-aware mobile cloud computing in 5G." *IEEE Network*, 29(2), 32-38, 2015.
[8] K. Vijayan, H. Li and T. Toda, "Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes." *IEEE Signal Processing Magazine*, 36(1), 95-102, 2018.
[9] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino and Y. Ochiai, "Investigating different representations for modeling and controlling multiple

[1] ***Autor de contacto****: clara.luzon@voicemod.net*

emotions in DNN-based speech synthesis." *Speech Communication*, 99, 135-143, 2018.

[10] K. Zhou, B. Sisman and H. Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training". arXiv preprint arXiv:2103.16809, 2021.

[11] K. Zhou, B. Sisman and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data." arXiv, 24 de octubre de 2020. Accedido: 12 de abril de 2023. Disponible en: http://arxiv.org/abs/2002.00198, 2020.

[12] R. Aihara, R. Takashima, T. Takiguchi and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features." *American Journal of Signal Processing*, 2(5), 134-138, 2012.

[13] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English." *Speech Communication*, 51(3), 268-283, 2009.

[14] Z. Luo, T. Takiguchi and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-5). IEEE, 2016.

[15] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for cnn-based speech synthesis." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6950-6954), 2019.

[16] H. Ming, D. Y. Huang, L., Xie, J. Wu, M. Dong and H. Li. "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion." In Interspeech (pp. 2453-2457), 2016.

[17] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.

[18] Veaux, Christophe; Yamagishi, Junichi; MacDonald, Kirsten. (2017). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/1994.

[19] Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Transactions on Information and Systems, E99.D(7), 1877–1884. https://doi.org/10.1587/transinf.2015edp7457

[20] Morise, M. (2016, November). D4C, a band-aperiodicity estimator for high-quality speech synthesis. Speech Communication, 84, 57–65. https://doi.org/10.1016/j.specom.2016.09.001

[21] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

[22] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. Language Resources and Evaluation, 42(4), 335-359. https://doi.org/10.1007/s10579-008-9076-6

[23] Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional Voice Conversion: Theory, Databases and ESD. Speech Communication, 137, 1-18. https://doi.org/10.1016/j.specom.2021.11.006

[24] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5(4), 377-390. https://doi.org/10.1109/taffc.2014.2336244

[25] Adigwe, A., Tits, N., Haddad, K., Ostadabbas, S., & Dutoit, T. (2018). The emotional voices database: Towards controlling the emotion dimension in voice generation systems. arXiv preprint arXiv:1806.09514.

[1] ***Autor de contacto***: *clara.luzon@voicemod.net*