

Neural network models for whisper to normal speech conversion

Cézar F. Yamamura

Academic Dept. of Electrical, Federal Technological University of Paraná, Cornélio Procópio, Brazil, (cezaryamamura@alunos.utfpr.edu.br) ORCID [0000-0002-4085-070X](https://orcid.org/0000-0002-4085-070X)

Paulo R. Scalassara

Academic Dept. of Electrical, Federal Technological University of Paraná, Cornélio Procópio, Brazil, (prscalassara@utfpr.edu.br) ORCID [0000-0001-7169-954X](https://orcid.org/0000-0001-7169-954X)

Marco A. Oliveira

Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 PORTO, Portugal, (marcoantoliveira@fe.up.pt) ORCID [0000-0002-3161-1109](https://orcid.org/0000-0002-3161-1109)

Aníbal J. S. Ferreira


Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 PORTO, Portugal, (ajf@fe.up.pt) ORCID [0000-0002-7278-6749](https://orcid.org/0000-0002-7278-6749)

Author Keywords

Whisper Speech, Multilayer
 Perceptron Network, Generative
 Adversarial Network.

Type: Research Article

 Open Access

 Peer Reviewed

 CC BY

Abstract

Whispers are common and essential for secondary communication. Nonetheless, individuals with aphonia, including laryngectomees, rely on whispers as their primary means of communication. Due to the distinct features between whispered and regular speech, debates have emerged in the field of speech recognition, highlighting the challenge of effectively converting between them. This study investigates the characteristics of whispered speech and proposes a system for converting whispered vowels into normal ones. The system is developed using multilayer perceptron networks and two types of generative adversarial networks. Three metrics are analyzed to evaluate the performance of the system: mel-cepstral distortion, root mean square error of the fundamental frequency, and accuracy with f1-score of a vowel classifier. Overall, the perceptron networks demonstrated better results, with no significant differences observed between male and female voices or the presence/absence of speech silence, except for improved accuracy in estimating the fundamental frequency during the conversion process.

1. Introduction

The human voice serves as a critical instrument for communication, enabling the transmission of information and ideas while fostering social connections. In certain contexts, whispering serves as an alternative mode of communication, typically utilized for private conversations in public settings via cell phone, or in quiet environments such as libraries, hospitals, and meeting rooms. In addition, individuals who suffer from conditions such as vocal fold paralysis, vocal nodules, and other related afflictions may experience difficulty producing normal speech as a result of partial or complete absence of vocal fold vibrations (i.e., voicing) ([Sharifzadeh et al. 2017](#)).

As a result, recent years have seen increased interest in the study of whispered speech. Furthermore, whispered speech typically exhibits a lower sound pressure level than normal speech, resulting in a lower signal-to-noise ratio (SNR). This characteristic poses challenges for speech processing, particularly for speech recognition, as whispered speech is generally more difficult to analyze and identify than normal speech (Grozdi and Jovii 2017).

Although the need for processing whispered speech is becoming increasingly important, the amount of research on this topic is relatively low. Traditional approach for speech analysis-synthesis such as MELP (Mixed Excited Linear Prediction), CELP (Code Excited Linear Prediction) and LPC (Linear Prediction Coding) have been commonly used in converting whisper to normal speech (Morris and Clements 2002; Sharifzadeh, McLoughlin, and Ahmadi 2010). However, an analysis-synthesis model requires fundamental frequency estimating from whispered speech. Unfortunately, whispered speech, characterized by the absence of vocal cord vibration, results in the absence of a fundamental frequency. As far as we currently know, there is no efficient method available for accurately estimating the fundamental frequency from whispered speech (Gao et al. 2021).

Recently, machine learning methods have also been used with great success, as in the case of multilayer perceptron network (Hinton et al. 2012). Among various MLP topologies, generative adversarial networks (GANs) (Goodfellow et al. 2020) have gained popularity in the field of speech processing (Wali et al. 2022), showing significant improvements in performance and quality for applications such as voice conversion (Dhar, Jana, and Das 2022), voice enhancement (Pascual, Serra, and Bonafonte 2019; Yu et al. 2021), voice synthesis (Saito, Takamichi, and Saruwatari 2018) and notably, the focus of this study, whisper-to-speech conversion (Gao et al. 2021; Shah et al. 2018; Ardaillon, Henrich, and Perrotin 2022).

Therefore, despite the limitations inherent in traditional approaches to converting whispered speech to normal speech, as previously discussed, the aim of this study is to develop a system that converts whispered vowels to normal speech using neural models approach, specifically MLP and GAN networks, the architecture proposed by referenced paper (Shah et al. 2018). To achieve this, we utilized the European Portuguese speech database from the Dysphonic to Natural Voice Reconstruction project (DyNaVoiceR). Some works related to this project are: manipulation of the fundamental frequency micro-Variations using a fully parametric and computationally efficient speech model (J. P. Silva et al. 2020) and flexible parametric implantation of voicing in whispered speech under scarce training data (J. Silva, M. Oliveira, and Ferreira 2021).

The remainder of this paper is structured as follows: Section 2 provides details about the speech database and methodology, which is divided into data pre-processing, neural network models, and performance metrics. Section 3 gives experimental results and discussion. Finally, Section 4 concludes this paper.

2. Methodology

In this paper, we first provide a detailed description of the speech database used in the study. We then present the proposed methodology, which is divided into speech signal feature extraction, neural network architectures, and performance metrics.

2.1. Database

The speech database used in this study was obtained from the DyNaVoiceR project, which focuses on advanced assistive technology to help patients affected by voice dysphonia, including temporary or permanent aphonia, to communicate more effectively and comfortably. This database was developed by European institutions, such as the Faculties of

Engineering and Medicine at the University of Porto The dataset is balanced, comprising recordings of 20 healthy speakers, evenly split between 10 males and 10 females, in .wav audio format at a sampling rate of 22,050 Hz. This balance is crucial for enabling fair and reliable gender-based performance comparisons in the results. For each speaker, recordings of sustained oral vowels were made in both normal and whispered speech. The database also includes manual phonetic annotation for each recording, allowing for the identification and localization of each phoneme. There are a total of 9 oral vowels used in European Portuguese, in sustained form (as shown in Table 1) (M. A. Oliveira 2020).

Vowel	Example
/i/	il <u>h</u> a
/ê/	p <u>e</u> so
/é/	<u>e</u> la
/á/	<u>á</u> gua
/â/	<u>a</u> marelo
/ó/	<u>ó</u> culos
/ô/	<u>o</u> vo
/u/	<u>u</u> va
/e/	sed <u>e</u>

Table 1: Sustained phonemes in European Portuguese from the DyNaVoiceR project database.

Table 2 presents the separation of the 9 oral vowels into four groups used in the experiments to evaluate the performance of the conversion networks from whispered to normal speech. Each group was augmented with a variety of vowels, arbitrarily chosen, to verify the system's performance. By categorizing the vowels into groups and introducing this diversity, we aim to evaluate the effectiveness of the conversion networks when dealing with vowels characterized by diverse phonetic and articulatory characteristics. Essentially, the goal is to determine whether the conversion process performs equally well on vowels that exhibit different speech sounds and vocal tract configurations. This experiment serves as a means to assess how effectively the conversion networks can accommodate these variations in speech characteristics when transforming whispered speech into a more normal, audible form.

Group	Variety amount	Vowels
1	1	/á/ /â/
2	2	/á/ /â/ /ê/ /é/
3	3	/á/ /â/ /ê/ /é/ /ó/ /ô/
4	5	/á/ /â/ /ê/ /é/ /ó/ /ô/ /i/ /u/ /ú/

Table 2: Groupings of vowel phonemes.

In addition to separating the database by vowels, it is also divided by gender (men and women) and signal speech without silence. The main difference between female and male voices is the fundamental frequency. Adult men typically have larger and thicker vocal folds that vibrate between 80 and 150 Hz, while women generally have smaller and thinner vocal folds that vibrate between 150 and 250 Hz (Behlau 2001).

2.2. Feature Extraction

Firstly, silent intervals in the database are removed to study their impact on the performance of the conversion system. These intervals are manually removed using the annotations provided in the database.

For the extraction of acoustic features, a parametric vocoder named AHOCODER (Erro et al. 2011) was utilized. The sampling rate of the speech signals was reduced from 22.050Hz to 16.000Hz to enable the use of this tool. The 40-dimensional Mel Frequency Cepstral Coefficients (MFCCs), which include the 0_{th} coefficient, and the logarithm of the fundamental frequency, $\log(F_0)$, were extracted from the whispered and normal speeches using a 25ms window and 5ms frameshift.

The $\log(F_0)$ plot are shown in Figure 1. It can be noticed that the $\log(F_0)$ in the whispered signal has some impulses but most of the time it is zero, because the lack of vocal folds vibration causes absence of the fundamental frequency and its harmonics. After feature extraction, these features will be used to train and evaluate a system that converts whispered speech to normal speech.

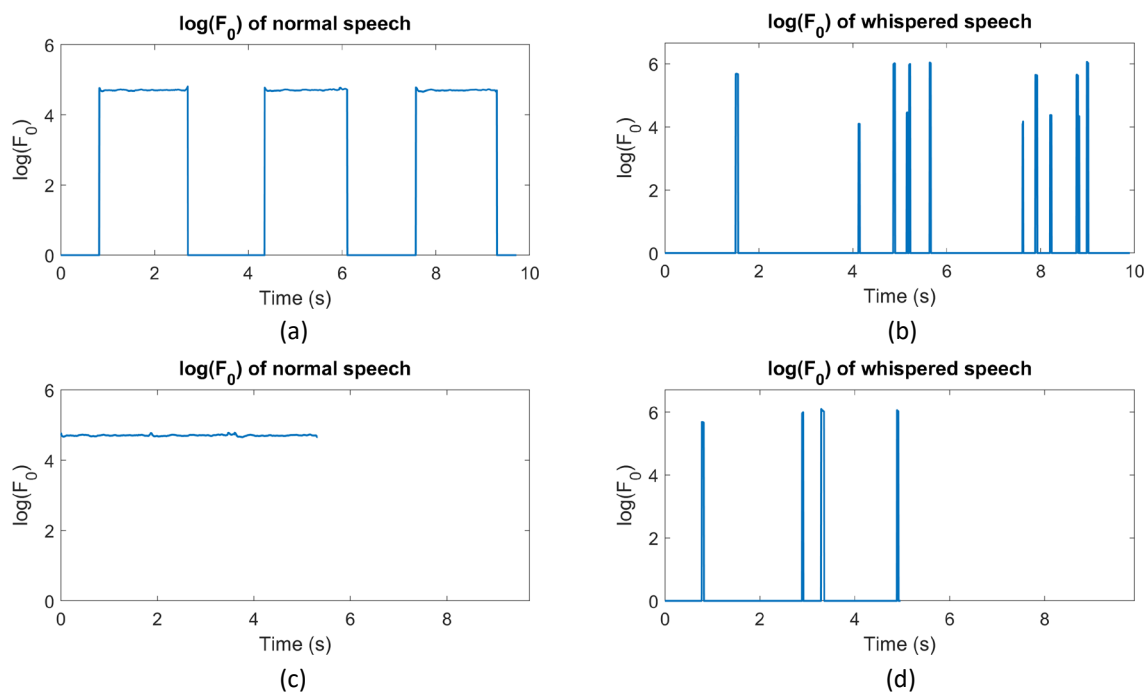


Figure 1: The logarithm of the fundamental frequency of normal and whispered speech: (a)-(b) without silence removed and (c)-(d) with silence removed.

2.3. Networks architectures

A Generative Adversarial Network (GAN) is a machine learning framework introduced by Ian Goodfellow, where two neural networks engage in a competitive, zero-sum game, with one network's success leading to the other network's failure (Goodfellow et al. 2020). Figure 2 illustrates the GAN architecture. A whispered speech signal is used as input to the Generator network, which aims to produce, in its output, a signal similar to normal speech. The Discriminator network tries to distinguish between the two speech signals, real and fake speech, with its response used for parameter adjustments of the networks. Thus, the Generator network learns to generate a reconstructed speech close to normal speech. However, training GANs can encounter challenges such as mode collapse, where the

Generator produces limited or repetitive outputs, and training instability, where the networks fail to converge or exhibit oscillatory behavior.

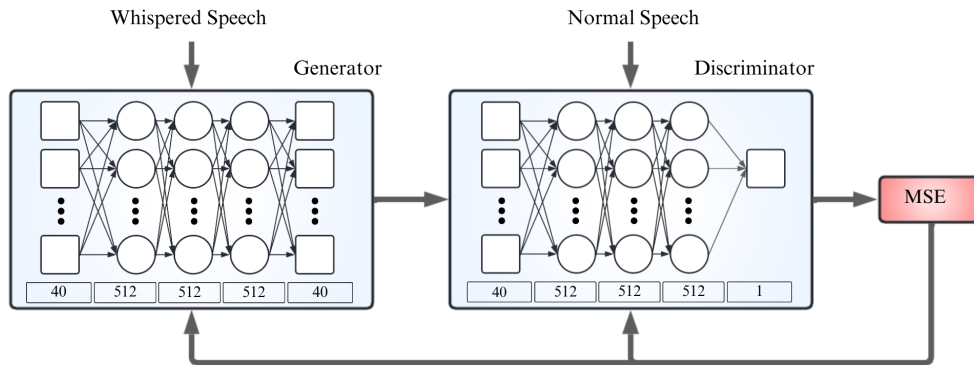


Figure 2: GAN architecture for Whispered Speech Reconstruction.

To address the challenges of mode collapse and instability commonly encountered during GAN training, we also used DiscoGAN, the architecture proposed by (Shah et al. 2018) and illustrated in Figure 3. In the context of two domains, whispered speech (W) and normal speech (S), features are mapped (X_S and X_W) for the two different speech types. The model uses two generators (G_{WS} and G_{SW}) along with two discriminators (D_W and D_S). G_{WS} transforms X_W into X_{WS} (converted features of normal speech) in such a way that X_{WS} becomes indistinguishable from the authentic samples X_S . Similarly, G_{SW} accomplishes this task. The discriminator D_S attempts to distinguish between X_S and X_{WS} while D_W performs a similar function for X_W .

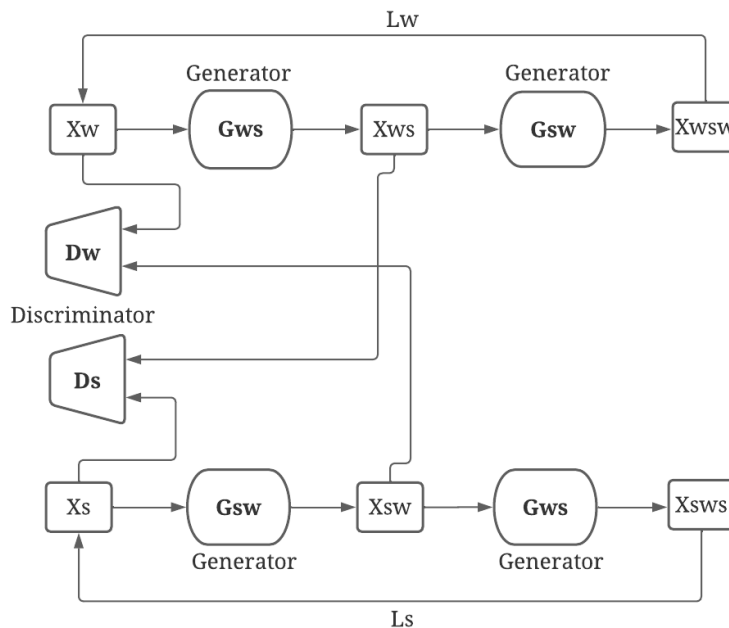


Figure 3: DiscoGAN architecture for discriminators outputs.

2.4. Conversion system

For the whispered to normal speech conversion, we used two neural networks as shown in Figure 4. The first network maps the normal MFCC (MFCCn) to the corresponding whispered MFCC (MFCCw), taking into account the temporal differences between them. To handle these temporal differences, we used Dynamic Time Warping (DTW), which is especially important under parallel training scenarios for the network to generate a normal MFCC estimate (MFCCn') in the inference phase.

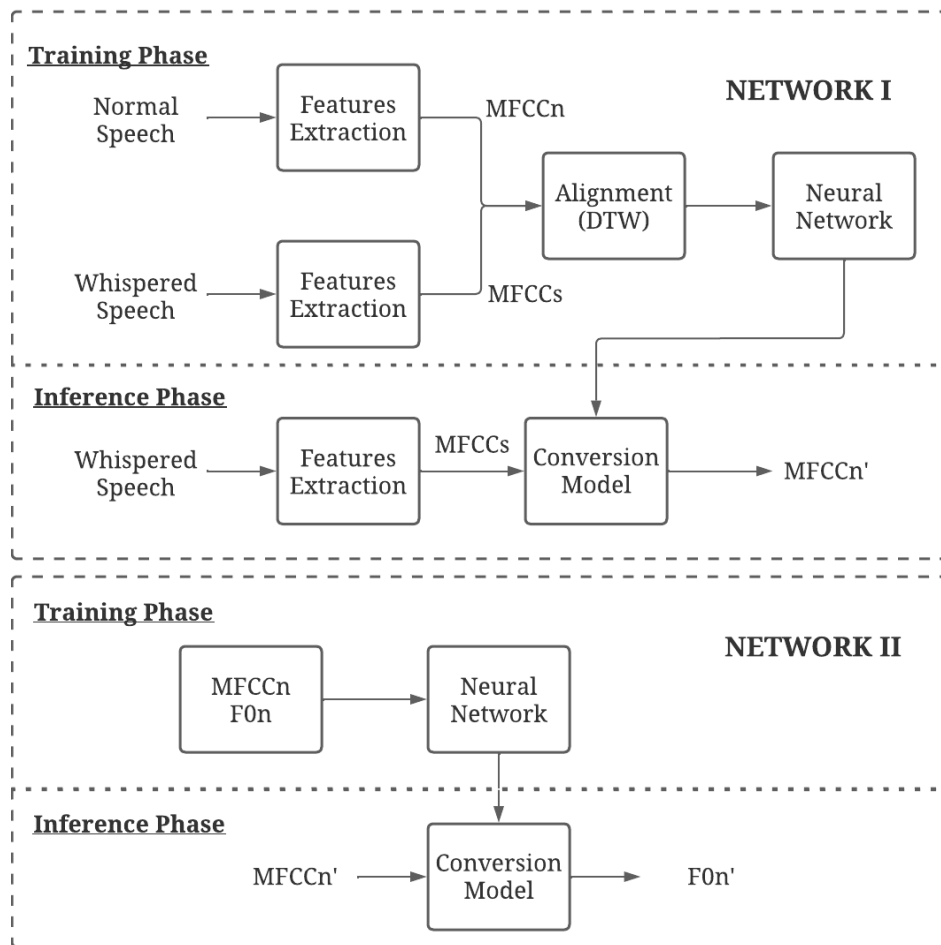


Figure 4: Block diagram of (a) MFCCs to MFCCn' Conversion Network I and (b) MFCCn' to F_{0n} Conversion Network II.

The second network uses both MFCCn and F_{0n} (normal F_0) as inputs during its training phase. In its inference phase, the output of the first network, MFCCn', is used as input to generate the F_0 estimation (F_{0n}').

Therefore, three different architectures of neural networks were utilized: Multilayer Perceptron Networks (MLP), GAN Networks, and DiscoGAN Networks.

These three networks follow an identical architecture with three hidden layers. Having uniform architecture aids in more equitably analyzing the advantages of adversarial training. Each hidden layer contains 512 neurons with Rectified Linear Unit (ReLU) activation, while the output layer of the first network has a linear activation function. The discriminators of the GAN, DiscoGAN, and the second network also have three hidden layers with ReLU activation. Their output layer uses a sigmoid activation function since it has only one output.

All networks are developed in Python and are trained for 100 epochs, using batch sizes of 1000 samples, as suggested in (Shah et al. 2018). The parameters are optimized using Adam optimization, with a learning rate of 0.0001.

2.5. Objective metrics

We have applied Mel Cepstral Distortion (MCD), Root Mean Square Error (RMSE) and a neural vowel classifier to evaluate the effectiveness of whisper to normal speech conversion system.

The traditional MCD measure is used here which is given by (Parmar et al. 2019):

$$MCD = \frac{10}{\ln(10)} \sqrt{2 \sum_{i=1}^N (MFCC_{n_i} - MFCC_{c_i})^2} \quad (1)$$

where $MFCC_{n_i}$ and $MFCC_{c_i}$ are the i^{th} MFCCs of the reference and converted normal speech, respectively, and N represents the dimension of the cepstral coefficient feature. The MCD measures the mean squared difference between the two sets of MFCCs and is a commonly used metric with respect to the spectral features to evaluate the performance of speech conversion systems. A lower MCD value indicates a better system performance, as it suggests that the converted speech is more similar to the reference speech.

To measure the RMSE of $\log(F_0)$, the actual reference speech and the converted speech signals are first aligned in time using the Dynamic Time Warping (DTW) algorithm. This alignment results in pairs of voiced-voiced, voiced-unvoiced, unvoiced-voiced, and unvoiced-unvoiced segments. We only consider the voiced-voiced pairs for computing the RMSE of the $\log(F_0)$. The RMSE of the $\log(F_0)$ is calculated as follows (Parmar et al. 2019):

$$RMSE(\log(F_0)) = \sqrt{\sum_{i=1}^K [\log(F_{0n_i}) - \log(F_{0c_i})]^2} \quad (2)$$

where F_{0n_i} and F_{0c_i} represent the reference and converted $\log(F_0)$ values at time frame i , respectively, and K is the total number of voiced frames. A lower RMSE value indicates a better performance of the conversion system.

The final metric used to evaluate the whispered to normal speech conversion system is the vowel classification performance, depicted in Figure 5. The classifier is a multilayer perceptron (MLP) with three hidden layers, each containing 128 neurons with Rectified Linear Unit (ReLU) activation. The network is trained with both MFCCn and F_{0n} , resulting in a 41-neuron input layer. The output layer employs a sigmoid activation function and has a number of neurons corresponding to the number of vowels in Table 2. The performance of the system is evaluated based on the classification accuracy and f1-score achieved by the vowel classifier.

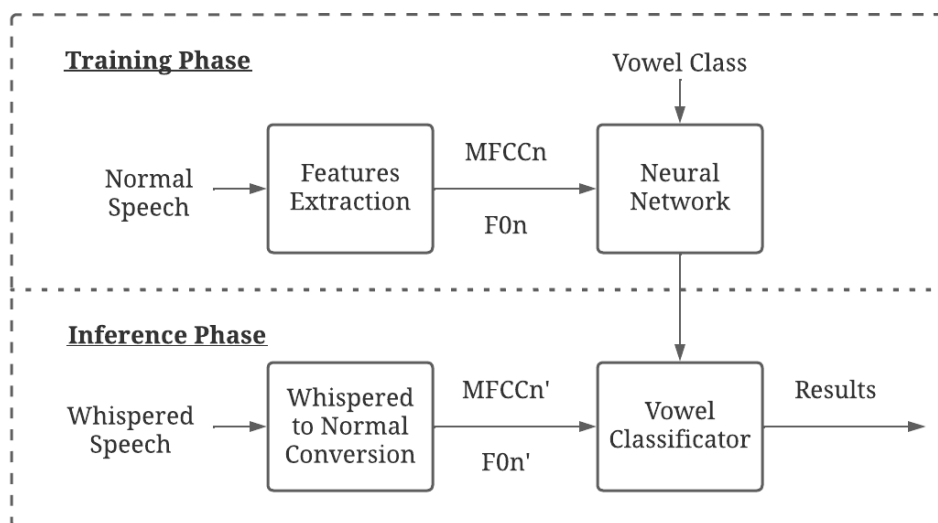


Figure 5: Block diagram of vowels classifier.

3. Results and discussion

3.1. MCD analysis

MCD measures the distance between the converted and reference cepstral features, where a lower MCD value indicates better performance of the system. Table 3 and Table 4 present the mean and standard deviation of MCD for the database without and with silence removed, respectively.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	4.58 ± 0.39	5.60 ± 0.38	5.26 ± 0.48	4.83 ± 0.61	5.05 ± 0.74	5.91 ± 0.78
2	3.93 ± 0.30	5.31 ± 0.83	5.24 ± 0.56	4.03 ± 0.45	5.81 ± 0.75	5.96 ± 0.85
3	4.70 ± 0.47	5.58 ± 0.60	6.78 ± 0.54	4.87 ± 0.64	9.20 ± 1.01	5.99 ± 0.66
4	4.72 ± 0.57	6.52 ± 0.56	5.82 ± 0.59	4.99 ± 0.66	6.69 ± 0.93	5.94 ± 0.73
mean	4.48 ± 0.43	5.75 ± 0.59	5.77 ± 0.54	4.68 ± 0.59	6.69 ± 0.85	5.95 ± 0.75

Table 3: The mean values and standard deviations of MCD without silence removed.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	3.93 ± 0.48	7.91 ± 1.78	4.74 ± 0.73	4.21 ± 0.66	6.41 ± 0.83	6.04 ± 1.01
2	4.43 ± 0.75	5.11 ± 0.80	5.12 ± 0.71	4.84 ± 0.61	6.31 ± 0.58	6.76 ± 0.78
3	4.58 ± 0.75	5.17 ± 0.80	5.22 ± 0.85	4.52 ± 0.70	5.05 ± 0.81	5.16 ± 0.72
4	4.71 ± 0.79	7.39 ± 0.92	5.27 ± 0.78	4.84 ± 0.71	12.01 ± 2.61	5.33 ± 0.77
mean	4.41 ± 0.69	6.39 ± 1.08	5.08 ± 0.77	4.60 ± 0.67	7.45 ± 1.20	5.82 ± 0.82

Table 4: The mean values and standard deviations of MCD with silence removed.

We compared the results of three neural models, and found that the mean MCD of the MLP model was better than that of the GAN and DiscoGAN models, with differences of 36% and 28% respectively. Furthermore, there was no statistically significant difference in the results for male and female voices and no significant advantage in excluding silent segments from the speech signals.

3.2. RMSE analysis

The second metric is the RMSE of $\log(F_0)$, which measures the error between the converted and reference F_0 values after they are time-aligned using the Dynamic Time Wrapping (DTW) algorithm. Table 5 and Table 6 present the mean RMSE values and standard deviations for the database without and with the silent parts removed, respectively.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	13.90 ± 5.24	13.68 ± 5.97	10.59 ± 4.05	21.33 ± 8.10	17.72 ± 9.86	21.80 ± 13.93
2	16.00 ± 9.27	10.70 ± 7.16	12.95 ± 7.41	17.52 ± 8.91	14.56 ± 8.97	13.67 ± 8.24
3	15.59 ± 7.31	17.44 ± 10.59	17.44 ± 11.38	22.86 ± 10.20	20.98 ± 9.92	20.65 ± 10.77
4	17.06 ± 9.32	19.52 ± 9.49	16.75 ± 10.24	23.40 ± 11.55	18.15 ± 8.57	21.61 ± 12.53
mean	15.63 ± 7.78	15.36 ± 8.30	14.43 ± 8.27	21.27 ± 9.69	17.85 ± 9.33	19.43 ± 11.36

Table 5: The mean values and standard deviations of RMSE of $\log(F_0)$ without silence removed.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	6.43 ± 3.67	10.10 ± 7.45	6.69 ± 7.27	7.55 ± 3.05	7.07 ± 3.35	8.70 ± 6.79
2	6.17 ± 2.23	6.59 ± 2.42	8.04 ± 3.52	19.56 ± 8.30	16.50 ± 9.21	21.73 ± 11.51
3	6.87 ± 3.54	6.72 ± 3.79	7.24 ± 4.29	9.04 ± 3.34	7.29 ± 3.48	7.43 ± 3.01
4	7.83 ± 3.61	9.38 ± 5.12	8.13 ± 4.39	10.02 ± 3.77	9.38 ± 7.19	7.20 ± 3.43
mean	6.82 ± 3.26	8.20 ± 4.69	7.52 ± 4.86	11.54 ± 4.61	10.06 ± 5.80	11.26 ± 6.18

Table 6: The mean values and standard deviations of RMSE of $\log(F_0)$ with silence removed.

Using this metric, we observed that the database with silence obtained higher values, which is undesirable, while the database without silence had better results. Therefore, the presence of silence makes it difficult to calculate the $\log(F_0)$ estimate. An example of the generated F_0 contour using the various developed systems are shown in Figure 6.

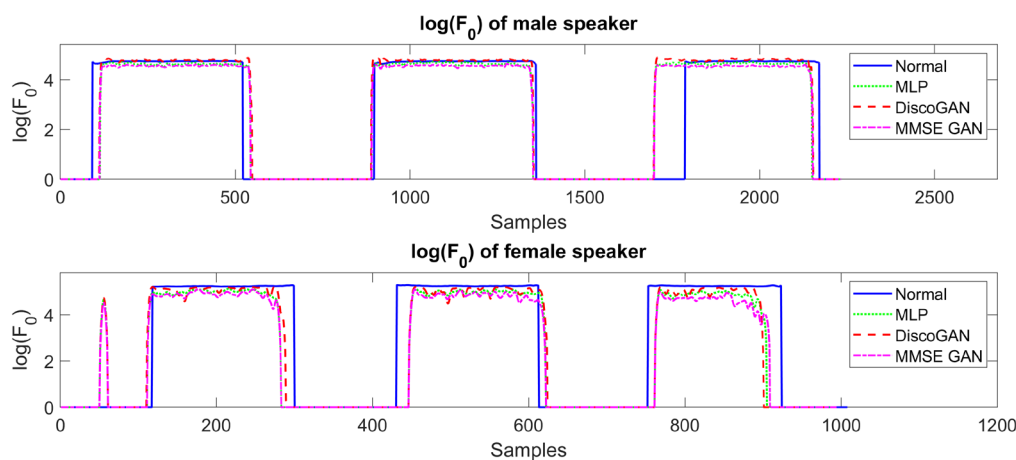


Figure 6: The $\log(F_0)$ predicted using the corresponding natural speech signal, MLP, GAN, and DiscoGAN for male and female speakers.

In Figure 6, we can observe that all the conversion systems, whether using MLP, GAN, or DiscoGAN, closely follow the distribution of the F_0 corresponding to the natural speech, with some oscillation. Tables 5 and 6 show, comparing the neural models, GAN and DiscoGAN outperform MLP. However, considering the RMSE standard deviation, the three systems perform similarly. We concluded that the presence of silence affects the results, with an improvement of 46% for the system trained with the database without silence. This can be explained by the challenges neural networks face when estimating F_0 in the presence of silence. In regions of silence, the F_0 value is effectively zero, and the transition between zero and the actual pitch value can create inconsistencies that are difficult for the model to learn. However, silence removal did not have a significant impact on the MCD or vowel classification performance. Therefore, silence removal only contributed to improving the F_0 estimation, but it did not have a notable effect on the overall performance of vowel conversion or classification tasks.

3.3. Vowel classifier

In the final analysis, we evaluated the performance of a vowel classifier based on a MLP with three hidden layers. The network was trained on normal speech signals and tested on the converted signals to assess its ability to distinguish between vowels. To obtain reliable results, we adopted the K-fold cross-validation method with $k=10$, dividing the database into ten equal parts.

Table 7 and Table 8 present the accuracy of the vowel classifier network without and with silence removed, respectively. Group 1, which contains only the vowel /a/, was excluded from the analysis as it was used to classify between the vowel /a/ and silence. Hence, Table 8 did not consider group 1.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
2	92.8%	93.5%	91.2%	85.4%	90.4%	92.4%
3	94.8%	93.1%	88.8%	89.4%	74.1%	84.8%
4	88.3%	83.4%	85.5%	79.8%	81.9%	77.0%
mean	93.3%	91.6%	90.7%	86.8%	84.5%	84.9%

Table 7: Accuracy of the vowel classifier for signals without silence removed.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
2	99.5%	99.4%	99.3%	80.3%	82.6%	83.7%
3	97.6%	97.4%	94.6%	93.9%	94.5%	95.7%
4	87.7%	58.3%	77.6%	84.7%	21.0%	84.2%
mean	94.9%	85.0%	90.5%	86.3%	66.0%	87.8%

Table 8: Accuracy of the vowel classifier for signals with silence removed.

Tables 7 and 8 show, in general, MLP had a better performance than GAN and DiscoGAN. It was also noticed that the accuracy decreases as the number of vowels in the database increases. However, this is understandable because the network must classify more variables. We also assessed the F1-score for all vowels, using only group 4 for this purpose. The F1 score blends precision and recall using their harmonic mean. Maximizing for the F1 score implies simultaneously maximizing for both precision and recall. Tables 9 and 10 present the F1-score for each vowel, without and with silence removed, respectively.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
silence	93.4%	93.7%	93.0%	87.4%	94.4%	86.5%
i	83.1%	80.2%	77.3%	71.6%	76.4%	65.0%
e	90.1%	85.0%	84.6%	84.3%	81%	78.9%
a	91.1%	83.7%	87.6%	72.7%	28.0%	69.0%
o	82.3%	73.8%	77.6%	66.7%	42.5%	64.8%
u	71.2%	46.1%	60.5%	65.7%	71.2%	58.4%
mean	85.2%	77.1%	80.1%	74.7%	65.6%	70.4%

Table 9: F1-score without silence removed.

Group	Male			Female		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
i	93.1%	71.7%	72.3%	69.8%	0%	77.0%
e	92.0%	60.6%	86.3%	89.3%	0%	90.3%
a	93.4%	79.1%	89.1%	89.1%	0%	89.4%
o	85.5%	28.5%	77.8%	80.3%	34.6%	82.1%
u	76.3%	51.9%	51.3%	77.4%	1.8%	77.3%
mean	88.1%	58.4%	75.4%	81.2%	0%	83.2%

Table 10: F1-score with silence removed.

Once again, we observed that the MLP network outperformed the other networks. Regarding the vowels, /a/ yielded the best result, while the worst was the vowel /u/. By examining the

confusion matrix of the best average F1-score results, as shown in Figure 7, we notice that the vowel /u/ indeed exhibited the poorest performance and was mainly confused with the vowel /o/.

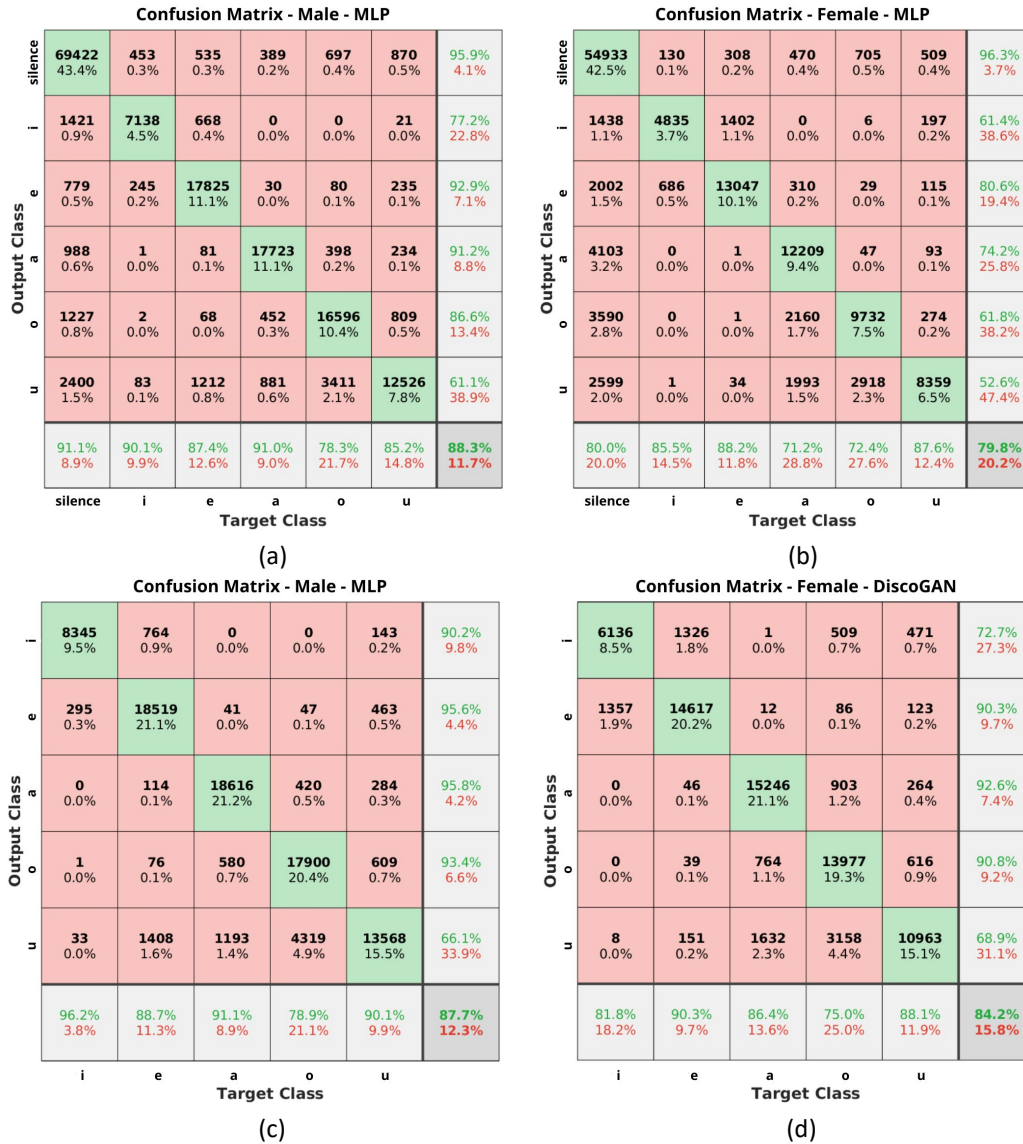


Figure 7: Confusion matrix of the best F1-score results: (a)-(b) without silence removed and (c)-(d) with silence removed.

Furthermore, we observed that the vowel classifier encountered greater challenges in classifying female voices compared to male ones. This discrepancy was evident in both accuracy and F1-score, particularly when dealing with signals generated by the GAN, which exhibited a lower performance of 58.3% for male voices and 21.0% for female voices. Upon analyzing the F1-score, it was observed that the vowels contributing most to this difficulty were /o/ and /u/, likely due to their similar acoustic and articulatory properties, as highlighted in studies such as Chl dkov  and Escudero (2012). These vowels are inherently confusable, particularly in whispered speech where distinctions are less pronounced. Regarding the silence removed, it did not contribute at all to the improvement of the result.

4. Conclusions

This study presents an implementation of whisper to normal European Portuguese vowel conversion system using neural network models. One of the contributions of this study is the use of the European Portuguese database provided by the collaboration between FEUP and UTFPR for the development of the speech conversion systems.

To validate the methodology of the speech conversion system, we implemented three neural networks: i) MLP, ii) GAN, and iii) DiscoGAN. We partitioned the database by gender, groups of different vowels, and speech signal with and without silence removed for all networks. We evaluated the performance of each system using three metrics: i) MCD, ii) RMSE of $\log(F_0)$, and iii) Vowel classification performance.

During the analysis, it was observed that the RMSE of $\log(F_0)$ was improved when using the database with the silence removed. However, silence removal did not have a significant impact on the MCD or vowel classification performance. Therefore, silence removal only contributed to improving the F_0 estimation, but it did not have a notable effect on the overall performance of vowel conversion or classification tasks.

In the gender-based performance comparison, female voices exhibited greater variation in certain metrics, particularly in RMSE of $\log(F_0)$. This variation can be attributed to several factors, including anatomical differences in vocal tract structures and fundamental frequency ranges between males and females. Additionally, the acoustic features of female voices might be more sensitive to small changes in speech characteristics, which could contribute to the observed higher variability in RMSE.

Furthermore, the performance evaluation of the three networks, MLP, GAN, and DiscoGAN, indicated that MLP yields better results for MCD and vowel classification performance, while GAN outperformed the other models in terms of RMSE of $\log(F_0)$.

References

- Ardailon, L., N. Henrich, e O. Perrotin. 2022. "Voicing Decision Based on Phonemes Classification and Spectral Moments for Whisper-to-Speech Conversion." In *Proceedings of Interspeech 2022*, 2253–2257. <https://doi.org/10.21437/Interspeech.2022-10675>.
- Behlau, M. 2001. *Voz: O Livro do Especialista*. São Paulo: Revinter. <https://ria.ufrn.br/jspui/handle/123456789/2886>.
- Chládková, K., e P. Escudero. 2012. "Comparing Vowel Perception and Production in Spanish and Portuguese: European versus Latin American Dialects." *Journal of the Acoustical Society of America* 131 (2): EL119–EL125. <https://doi.org/10.1121/1.3674991>.
- Dhar, S., N. D. Jana, e S. Das. 2022. "An Adaptive Learning-Based Generative Adversarial Network for One-to-One Voice Conversion." *IEEE Transactions on Artificial Intelligence*, in press. <https://doi.org/10.1109/TAI.2022.3149858>.
- Erro, D., I. Sainz, E. Navas, e I. Hernáez. 2011. "Improved HNM-Based Vocoder for Statistical Synthesizers." In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 1809–1812. Florence, Italy. <https://aholab.ehu.eus/papers/IS110663.pdf>.
- Gao, T., J. Zhou, H. Wang, L. Tao, e H. K. Kwan. 2021. "Attention-Guided Generative Adversarial Network for Whisper to Normal Speech Conversion." *arXiv preprint arXiv:2111.01342*. <https://doi.org/10.48550/arXiv.2111.01342>.

- Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, e Y. Bengio. 2020. "Generative Adversarial Networks." *Communications of the ACM* 63 (11): 139–144. <https://doi.org/10.1145/3422622>.
- Grozdić, D. T., e S. T. Jović. 2017. "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (12): 2313–2322. <https://doi.org/10.1109/TASLP.2017.2738559>.
- Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, et al. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29 (6): 82–97. <https://doi.org/10.1109/MSP.2012.2205597>.
- Morris, R. W., e M. A. Clements. 2002. "Reconstruction of Speech from Whispers." *Medical Engineering & Physics* 24 (7): 515–520. [https://doi.org/10.1016/S1350-4533\(02\)00060-7](https://doi.org/10.1016/S1350-4533(02)00060-7).
- Oliveira, M. A. M. 2020. *Modelização de Filtro de Trato Vocal para Reconstrução de Voz Disfónica*. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal. <https://repositorio-aberto.up.pt/bitstream/10216/126255/2/386486.pdf>.
- Parmar, M., S. Doshi, N. J. Shah, M. Patel, e H. A. Patil. 2019. "Effectiveness of Cross-Domain Architectures for Whisper-to-Normal Speech Conversion." In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 1–5. <https://doi.org/10.23919/EUSIPCO.2019.8902961>.
- Pascual, S., J. Serra, e A. Bonafonte. 2019. "Time-Domain Speech Enhancement Using Generative Adversarial Networks." *Speech Communication* 114: 10–21. <https://doi.org/10.1016/j.specom.2019.09.001>.
- Saito, Y., S. Takamichi, e H. Saruwatari. 2018. "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (1): 84–96. <https://doi.org/10.1109/TASLP.2017.2761547>.
- Shah, N. J., M. Parmar, N. Shah, e H. A. Patil. 2018. "Novel MMSE DiscoGAN for Cross-Domain Whisper-to-Speech Conversion." In *Machine Learning in Speech and Language Processing (MLSLP) Workshop*, 1–3. Google Office.
- Sharifzadeh, H. R., A. HajiRassouliha, I. V. McLoughlin, I. T. Ardekani, J. E. Allen, e A. Sarrafzadeh. 2017. "A Training-Based Speech Regeneration Approach with Cascading Mapping Models." *Computers & Electrical Engineering* 62: 601–611. <https://doi.org/10.1016/j.compeleceng.2017.06.007>.
- Sharifzadeh, H. R., I. V. McLoughlin, e F. Ahmadi. 2010. "Reconstruction of Normal-Sounding Speech for Laryngectomy Patients Through a Modified CELP Codec." *IEEE Transactions on Biomedical Engineering* 57 (10): 2448–2458. <https://doi.org/10.1109/TBME.2010.2053369>.
- Silva, J., M. Oliveira, e A. Ferreira. 2021. "Flexible Parametric Implantation of Voicing in Whispered Speech Under Scarce Training Data." In *Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, 416–420. <https://doi.org/10.23919/Eusipco47968.2020.9287684>.
- Silva, J. P., M. A. Oliveira, C. F. Cardoso, e A. J. Ferreira. 2020. "Manipulation of the Fundamental Frequency Micro-Variations Using a Fully Parametric and Computationally Efficient Speech Model." In *Proceedings of the 2020 IEEE Workshop on Signal Processing Systems (SiPS)*, 1–6. <https://doi.org/10.1109/SiPS50750.2020.9195214>.

Wali, A., Z. Alamgir, S. Karim, A. Fawaz, M. B. Ali, M. Adan, e M. Mujtaba. 2022. “Generative Adversarial Networks for Speech Processing: A Review.” *Computer Speech & Language* 72: 101308. <https://doi.org/10.1016/j.csl.2021.101308>.

Yu, G., Y. Wang, H. Wang, Q. Zhang, e C. Zheng. 2021. “A Two-Stage Complex Network Using Cycle-Consistent Generative Adversarial Networks for Speech Enhancement.” *Speech Communication* 134: 42–54. <https://doi.org/10.1016/j.specom.2021.09.010>.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors also thank the Federal University of Technology - Paraná, Brazil.