applied sciences

MDPI

Article

Automatic Classification of Synthetic Voices for Voice Banking Using Objective Measures

Agustin Alonso *, Víctor García, Inma Hernaez *, Eva Navas and Jon Sanchez

HiTZ Basque Center for Language Technologies—Aholab, University of the Basque Country UPV/EHU, 48013 Bilbao, Spain; victor.garcia@ehu.eus (V.G.); eva.navas@ehu.eus (E.N.); jon.sanchez@ehu.eus (J.S.)
* Correspondence: agustin@aholab.ehu.eus (A.A.); inma.hernaez@ehu.eus (I.H.)

Abstract: Speech is the most common way of communication among humans. People who cannot communicate through speech due to partial of total loss of the voice can benefit from Alternative and Augmentative Communication devices and Text to Speech technology. One problem of using these technologies is that the included synthetic voices might be impersonal and badly adapted to the user in terms of age, accent or even gender. In this context, the use of synthetic voices from voice banking systems is an attractive alternative. New voices can be obtained applying adaptation techniques using recordings from people with healthy voice (donors) or from the user himself/herself before losing his/her own voice. In this way, the goal is to offer a wide voice catalog to potential users. However, as there is no control over the recording or the adaptation processes, some method to control the final quality of the voice is needed. We present the work developed to automatically select the best synthetic voices using a set of objective measures and a subjective Mean Opinion Score evaluation. A prediction algorithm of the MOS has been build which correlates similarly to the most correlated individual measure.

Keywords: STOI; ESTOI; NISQA; SIIB; speech adaptation; voice banking

Citation: Alonso, A.; García, V.; Hernaez, I.; Navas, E.; Sanchez, J. Automatic Classification of Synthetic Voices for Voice Banking Using Objective Measures. *Appl. Sci.* **2022**, *12*, 2473. https://doi.org/10.3390/app12052473

Academic Editor: Douglas O'Shaughnessy

Received: 27 December 2021
Accepted: 24 February 2022
Published: 27 February 2022

1. Introduction

Speech is the most natural method that humans use to communicate with each other. When, due to an accident or illness, one person loses the ability to speak, technology can provide solutions to mitigate the impact of his or her disability. Text-to-speech (TTS) systems are a fundamental component of the so-called alternative and augmentative communication (AAC) devices, providing a synthetic voice to speak aloud the text that has been introduced through some kind of input device, such as a keyboard or an eye-gaze-controlled device. TTS systems have been available in the market for many years already, and nowadays synthetic voices are not only intelligible, but also have a high level of naturalness. However, it is often the case that the offered synthetic voices do not suit the user's preferences in terms of age, accent or even gender. Commercial voices are in general obtained from professional speakers, chosen precisely because of his or her pleasant voice, neutral accent and ability to keep a homogeneous speech style during long recording sessions, conditions which greatly help to obtain a high quality synthetic voice. Consequently, there are few (and for many languages, none) commercially available voices corresponding to an old person or to a child, or voices with regional accents. The orally disabled user must then use a synthetic voice unmatched to his or her own speaking characteristics. Speech is a fundamental part of the identity of a person. Synthetic voice customization tries to keep those hints of personality, nonexistent in a generic or commercial synthetic voice. Studies such as [1] show our tendency to form an impression on the personality of other people from their voice (as happens with other features, such as the face, or the color of the skin). Other studies demonstrate that using personalized voices can facilitate intellectual development for children with vision shortages [2]. Thus, it is our belief that the use of

personalized speech can help in reducing the social impact of using an electronic device for everyday communication.

Until the beginning of the last decade, the technologies applied for voice generation included concatenative synthesis with unit selection [3] and parametric statistical synthesis based on Markov models [4]. Concatenative unit selection systems generate voice by concatenating prerecorded fragments of a natural voice, chosen according to sophisticated selection criteria that include acoustic, phonetic, prosodic and linguistic aspects. These techniques produce very natural results in restricted application domains, but by expanding the domain, they obtain results of variable quality [5]. In addition, they have high memory, storage and processing requirements and reduced or null flexibility to generate new voices. Statistical–parametric systems create synthetic speech from averaged models of acoustically similar speech units. Speech is transformed into a series of parameters by means of a vocoder [6]. Typically, several types of parameters are used: spectral parameters related to the spectral envelope that include information about the energy in the different frequency bands; the fundamental frequency which is related to intonation; and finally some parameters related to the degree of voicing of the source, such as aperiodicities by bands (STRAIGHT [7] and WORLD [8]) or the maximum voiced frequency (AhoCoder [9]). Among the advantages of statistical–parametric systems are their consistency, flexibility, intelligibility and low need for storage. They allow the generation of new voices, using adaptation or interpolation techniques [10], and they produce a smooth voice of stable quality, although the use of vocoders reduces its naturalness. Its intelligibility is similar to that of natural speech and even better in noisy environments [11]. In short, these systems are very well suited to be used in the context of personalized TTS.

In more recent years, within the framework of parametric statistical synthesis, Markov models have been replaced by deep neural networks (DNNs) [12] with very good results in terms of quality of synthetic speech. DNNs can overcome some of the limitations of Gaussian models in representing the complex non-linear relationships that exist between the acoustic parameters of speech generation and the symbolic representation of speech. In [13], a very complete review of different possible strategies to use deep networks in the generation of the acoustic parameters of speech can be found.

Finally, more recently, DNNs are used not only for signal generation, but to carry out the entire TTS conversion chain. Deep Voice [14] was the first of these systems in which each stage of the TTS system was implemented using neural networks. Progressively, more and more end-to-end systems were devised (Deep Voice 2 [15], Deep Voice 3 [16] and Char2Wav [17]). Finally, totally end-to-end architectures have also been proposed, such as Tacotron [18], Tacotron 2 [19] and ClariNet [20] that generate spectrograms starting from the text. Subsequently, these spectrograms are converted into speech using WaveNet [21] or the Griffin–Lim algorithm [22]. These end-to-end systems are providing very good results in terms of the quality of the synthetic voice generated.

However, the amount of data required to build a robust TTS system based on neural networks is enormous. Although experiments have been performed to adapt voices and build systems with smaller amounts of data [23], in general, it has not been possible to produce quality voices with amounts of data comparable to those used in the production of personalized voices in parametric–statistical systems. To generate custom voices using DNNs, it is possible to embed the identity of the speaker in the model so that synthetic voices are generated with the characteristics of the voice of different speakers using only a few minutes of each voice. These representations of the speaker's voice (speaker embedding) have already been successfully used to generate different voices in Tacotron-based systems [15] but they only allow generating speech for those speakers seen during training. With the goal to generate quality synthetic voices for any speaker, strategies such as VoiceLoop [24] are more appropriate: it proposes an architecture based on a fixed-size buffer that can generate voice from speakers not seen in training. Another strategy suitable to generate voices for speakers not seen during training is knowledge transfer as proposed in [25]. In any case, in these strategies, there are quality differences in the results obtained

for speakers seen and not seen during the training [26] so that there are still important challenges to solve in order to generate the voice of any speaker.

In the work described here, the synthesis is based on a parametric statistical TTS system using adaptation techniques to obtain new voices with a small amount of recordings from a new speaker [9,27] and an average voice obtained with high quality recordings from professional speakers. This system has been implemented in a publicly available website [28], which allows users to record his or her own voice and download the resulting personalized synthetic voice to his or her device. At the same time, the synthetic voice remains in the system and is made available to orally disabled users through a 'Voice Bank' which acts as a repository of donated synthetic voices. The whole process is unsupervised and fully automated, which leads to a wide variability in the final quality of the synthetic voices. Indeed, the quality of the synthetic voice will strongly depend on the personal traits of the speaker or voice donor (speaking speed variability, degree of articulation, existence of a pathology in his or her voice, age, smoking habits, etc.). Another important factor is the similarity of the speaker's regional accent to that of the average voice. The acoustic conditions of the recording place as well as the recording equipment also play an important role and the best results are obtained when the recordings are performed in a professional environment. All these factors are of extreme importance, and lead to the fact that not all donors' synthetic voices are finally useful.

In this work, we describe a strategy to automatically select the synthetic voices with the best possible quality among thousands of donors' synthetic voices. The selected voices are candidates to populate the voice bank in an unsupervised manner. We extend the initial work described in [29] by evaluating four objective measures: short time objective intelligibility (STOI), enhanced short time objective intelligibility (ESTOI), non-intrusive speech quality assessment (NISQA) and speech intelligibility in bits (SIIB). A MOS evaluation is performed to validate the measures' scores. Using linear regression, the measures are also combined to produce an estimated MOS.

The rest of the article is structured as follows: in Section 2, we explain the purpose and characteristics of the developed voice bank. Section 3 describes the objective measures used in the analysis. The proposed method to score the voices is described in Section 4, and the experiments performed are presented in Section 5. Finally, in Section 6, the main conclusions of this work are drawn.

2. Voice Banking and Personalized Synthetic Voices

Voice banking is an alternative to provide people with speech difficulties with a personalized synthetic voice. In order to do that, the person has to record a number of sentences. Using those sentences, a synthetic voice with similarity to the recordings will be provided. Voice banking differs from message banking in that the last one just stores the recorded sentences to be played back when needed exactly as recorded. The recorded sentences will usually include expressions in which tone and emotion are of importance, as can be saying 'I love you', reading a bed-time story or even laughing. Instead, voice banking uses the recorded sentences to obtain a synthetic voice that can be used within a TTS system or more generally by an AAC device in daily communication. There are several voice banking providers for English (see Ref. [30–35]). In general, it is required that the person performs the recordings before the first symptoms of the disease are noticeable in the voice, because as commented in the introduction, this can condition the final quality of the synthetic voice. If this has already happened, some providers also offer the possibility to repair the synthetic voice, applying model surgery techniques [36–38]. On the other hand, the cited systems also differ in the number of sentences needed to be recorded (from a few hundred to several thousands) as well as on the operating systems, browsers or applications where the final user will be able to install it. Indeed, some AACs software providers do not allow the use of external voices, so some voice-banking providers also offer compatibility with the most popular commercial AACs software. Finally, also the cost of the provided services varies from one provider to another.

When it is impossible or very difficult for the user to make the recordings, normally because the disease already affects his or her speech, they can choose a voice donor to make the recordings for them. The donors are usually chosen from close family members or friends. In Ref. [31], they recommend to provide two donors and offer a final synthetic voice with mixed characteristics. In Ref. [32], a voice selected by the provider is offered to the user among a set of thousands of voices, with characteristics similar to the provided recordings.

To the best of our knowledge, the web-based voice-banking service in Ref. [28] is the only one provided for Spanish. In this portal, also Basque is offered, although the number of users is comparatively very small.

This portal offers the possibility of obtaining a personalized synthetic voice in Spanish and Basque. It makes use of a statistical synthesis engine based on hidden Markov models (HMMs) [4]. Each user must record a total of 100 phonetically balanced sentences in the selected language [27]. These are parameterized using ahocoder [9], a high-quality vocoder that extracts Mel-cepstral coefficients (MCEP) of order 39, log-f_0 and maximum voiced frequency. These data are then used to adapt an average voice using state-of-the-art adaptation techniques [39] based on Constrained maximum likelihood linear regression plus maximum a posteriori adaptation (CMLLR+MAP). For Spanish, the average voice was obtained with the subset 'phonetic' from the Albayzin [40] database. It consists of 6800 sentences from 204 different speakers in which each one has recorded 160, 50 or 25 sentences. For Basque, the average voice was obtained using the database described in [41]. This consists of two speakers (one female and one male) with four hours or speech each. Currently, our voice bank has almost 9000 registered users.

3. Objective Measures Overview

In this section, we briefly describe the selected objective measures: two intrusive objective measures typically used in speech enhancement, STOI [42] and ESTOI [43]; one also intrusive intelligibility measure based on information theory, SIIB [44]; and one measure based on NISQA that estimates the mean opinion score (MOS) of the naturalness of synthetic speech [45].

3.1. STOI and ESTOI

In the field of measuring the intelligibility of speech, several algorithms have been proposed with the aim of replacing expensive subjective listening tests. Among them, STOI [42] has proven to be good for evaluating intelligibility in signals to which time-frequency weighting is applied. The method requires both the signal to be evaluated and a clean time-aligned reference. It calculates the time frequency (TF) representation of both signals with a discrete Fourier transform (DFT) of the windowed frames and, using a one-third octave analysis, it groups the bins of the DFT into 15 bands and computes the norm of each one, which is called the TF-unit. It uses an intermediate measure of intelligibility for each TF-unit, which depends on N consecutive TF-units of both the signal to be evaluated and the reference. Typically the value of N is such that the intermediate measure depends on speech information from the last $\approx$ 400 ms. To calculate the global intelligibility measure, the average of the intermediate intelligibility measurements between frames and frequency bands is computed. This operation implies an independent contribution to the global measure of each band.

One evolution of STOI is ESTOI [43], which unlike STOI does not assume independence between frequency bands. This feature allows to better capture the effect of time-modulated noise maskers. Both STOI and ESTOI evaluate the signals between 0 and 1.

The success of these measures has led to propose their use in several areas. STOI has proven to predict intelligibility quite accurately in mobile phone output [46]; noisy speech processed by ideal time-frequency masking and single channel speech enhancement algorithms [47]; and speech processed by cochlear implants [48]. Additionally, it is also robust against different types of languages, such as Mandarin [49], Danish [47] or Dutch [50]. STOI has been also used to evaluate synthetic speech, showing high correlation with

MOS [51,52]. In Ref. [53], it was used to evaluate the intelligibility of dysarthric speech. In this case, as a time-aligned reference signal was not available, so an utterance-dependent reference signal was generated from several healthy speakers and then dynamic time warping (DTW) was used to align the pathological signal and the reference signal.

3.2. SIIB

SIIB [44] estimates the amount of information shared between a talker and a listener in bits per second. This measure is motivated by information theory, and suggests that the speech process can be understood as the transmission of a message from a talker to a listener. The message $\{M\}$ can be thought of as a sequence of sentences or phonemes. The talker encodes the message into a speech signal $\{X\}$ and sends the signal through a communication channel that may distort it and create a degraded speech signal $\{Y\}$. So, the whole communication process is described by a Markov chain:

$$\{M\} \rightarrow \{X\} \rightarrow \{Y\} \quad (1)$$

where $\{M\} \rightarrow \{X\}$ is the speech production channel and $\{X\} \rightarrow \{Y\}$ in the environmental channel.

This metric is based on the assumption that intelligibility is a function of the mutual information rate between the message $\{M\}$ and the degraded speech $\{Y\}$, so it needs the signal before the distortion as reference. The authors state that this measure works better in general conditions than other measures designed with a heuristic motivation and with an specific distortion or dataset in mind. SIIB estimates the intelligibility in an open scale, where an optimal signal obtains a score between 150 and 180 b/s.

3.3. NISQA-TTS

An important aspect to evaluate in synthetic voices is naturalness. In [45], a method based on NISQA [54] is proposed to measure the naturalness of synthetic voices without the need for a reference signal. The proposed prediction model is based on a convolutional neural network long short-term memory (CNN-LSTM) network architecture. It was trained using 16 databases with 12 different languages, so it is language independent and can be used to evaluate naturalness in any TTS. The model is publicly available at [55], so it can be used directly. As it estimates the naturalness in a MOS scale, its estimations can vary between 1 and 5. For the rest of the paper, this method will be referred to simply as NISQA.

4. Proposed Methodology

In this section, we describe the method followed to obtain the objective measures. For STOI, ESTOI and SIIB, the original recordings for each speaker are used as clean references and aligned with the corresponding synthetic sentences. To obtain the NISQA score, no processing of the synthetic signals is required.

4.1. STOI-ESTOI

To perform the alignment of the synthetic and reference signals, the first step is to obtain the phonetic segmentation of the recordings. This is done by forced alignment using Montreal forced aligner (MFA) [56]. This step will also provide the actual positions for the pauses made by the speaker during the recordings, and the synthetic signals will be generated using this information, thus with pauses at the same locations. In this way, a parallel corpus of recordings and synthetic signals is available. However, the signals will have different durations; therefore it will be necessary to align them. To do this, although they could be aligned at sentence level, in our system, we perform an alignment with DTW at the phoneme level. For the alignment, the cepstral distances between reference and 'target' are calculated. The cepstral coefficients for the synthetic signals are obtained directly from the adapted voice model, while for the reference signals, they must be calculated. These cepstral coefficients were obtained using Ahocoder [9]. Additionally, it

was necessary to synchronize the different frame rates used by the synthesis system and STOI/ESTOI algorithms.

After these steps, a score can be obtained with the STOI and ESTOI algorithms for each sentence. The final score for each speaker is the average of the scores obtained for all available sentences.

4.2. SIIB

The SIIB measure also needs a clean time-aligned reference. For this reason, we used a similar approach as with STOI–ESTOI. Thus, the same alignment was used and again the frame rate of the synthesis system and the one used in SIIB algorithm were synchronized. As SIIB authors warn that the score is not reliable for sentences shorter than 20 s, we concatenate several sentences after DTW to ensure that all the stimuli to be evaluated are long enough.

Finally, the SIIB score is obtained as the average score of all concatenated bunches.

4.3. NISQA

As this measurement does not require any reference, to calculate the donor's NISQA score, the score of all the previously generated synthetic sentences is calculated and averaged.

5. Experiments

In this section, we describe the details of the experiments performed to score the donors' voices. In order to establish an upper limit for the scores, several synthetic voices, obtained using similar synthesis techniques but generated from high quality professional recordings, are also included in the scoring process. We call these voices 'Standard HTS voices' and the obtained scores are described first in this section. Then the scores for all donors' voices are obtained, and the best of each objective measure are selected as potential good quality voices. Finally, the MOS evaluation to find the combination of objective measures that brings the best results is described.

5.1. Evaluation of Standard HTS Voices

In order to set an upper limit for the scores, we obtained the scores of the four objective measures for several high quality synthetic voices. These voices were trained using about four hours of high quality recordings from professional speakers. The training was a standard HTS training [57]. Six synthetic voices were used for this evaluation, obtained from three professional speakers—two females and one male—each speaker recorded both in Spanish (ES) and Basque (EU). The objective measures are language independent, so the language should not affect the results. The same sentences used to train the voices were used to obtain the scores because no other recordings were available from the same speakers. The number of available sentences for each speaker is shown in Table 1.

Table 1. Standard HTS voices training corpus.

Voice	Speaker	Language	Gender	# of Sentences
F1-ES	F1	ES	F	3994
F1-EU	F1	EU	F	3797
F2-ES	F2	ES	F	3712
F2-EU	F2	EU	F	3831
M1-ES	M1	ES	M	3995
M1-EU	M1	EU	M	3799

Using the six synthetic voices, the objective measures were computed by averaging the scores obtained over all the sentences. Notice that except for NISQA, the original recordings are used as reference to calculate the score. The mean values and the standard

deviations obtained for all the voices and for each measure are shown in Figure 1. As it can be seen, the synthetic voices obtain similar scores for a given measure, without significant differences among them. However, we can find some exceptions: for STOI and ESTOI, voice M1-ES stands out for its high scores; for SIIB, there are no significant differences; and finally for NISQA voice F2-ES has significantly higher score than the others, and M1-ES has the lowest score. When looking at the data from the language perspective, it is remarkable that the two voices from speaker M1 obtain significantly different values for the two languages, showing lower scores for Spanish (ES) than for Basque (EU) only in NISQA, and the other way round for STOI and ESTOI.

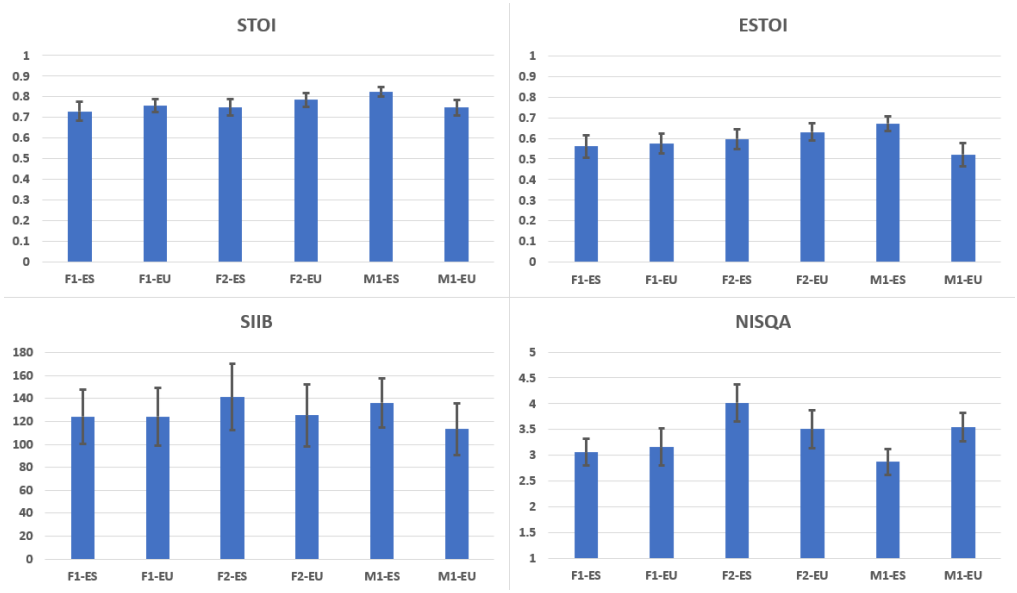

Figure 1. Mean values and standard deviation obtained for the objective measures for the Standard HTS voices.

5.2. Objective Evaluation of the Donors' Voices

The four objective measures were computed for the adapted synthetic voices available in the voice bank [58]. To simplify the subjective evaluation (described in Section 5.3) only Spanish was considered in the evaluation of donors' voices. A total of 1090 donors' voices was used. As the adaptation is performed using 100 recorded sentences from the donor, the same set of sentences was synthesized with each synthetic voice, and the original recordings were used as reference.

Table 2 shows the first five candidate voices ranked by each scoring algorithm (the speakers were arbitrarily named). Only the first five positions are shown for clarity reasons. As it can be seen in the table, 5 voices (out of the 20 possible candidates) obtained the best scores for more than one measure, so there are finally 15 final voice candidates occupying the first 5 positions. Indeed, the only measure that shows a unique set of voices (i.e., it has chosen five voices different from those chosen by the other measures in the first five positions) is SIIB. The other three measures share some voices among them.

Considering that there is overlapping in the ranked voices, we also computed the correlation coefficient ρ between the scores obtained for the different measures. The results in Table 3 show that although STOI and ESTOI are highly correlated, this is not the case for the other measures. The fact that STOI and ESTOI use a similar approach to estimate intelligibility explains the high correlation between them. Additionally, considering that NISQA measures naturalness and not intelligibility, a low correlation of this measure with the other measures can be expected. SIIB, a measure of intelligibility, does not correlate with any other measure, as corroborated by the selected set of speakers.

Table 2. Best voices sorted by objective measure.

-	STOI	ESTOI	SIIB	NISQA
1st	SPK01	SPK02	SPK03	SPK04
2nd	SPK04	SPK04	SPK05	SPK06
3rd	SPK07	SPK08	SPK09	SPK02
4th	SPK02	SPK10	SPK11	SPK12
5th	SPK10	SPK13	SPK14	SPK15

Table 3. Correlation ρ between different objective measures.

-	STOI	ESTOI	SIIB
ESTOI	0.912	-	-
SIIB	0.363	0.292	-
NISQA	0.449	0.476	0.258

5.3. MOS Evaluation

We performed an evaluation to determine which are the objective measures that lead to the best speakers according to people's preferences. For each measure, we selected the 5 best ranked synthetic voices, i.e., the 15 voices shown in Table 2. Then, 10 synthetic sentences from the corpus used to compute the objective scores were selected for each voice (some samples used in the evaluation can be found at https://aholab.ehu.eus/users/agustin/demos/mdpi21/ accessed on 10 February 2022). Each participant scored the quality of 5 randomly selected sentences per donor's voice in a MOS scale (a total of 75 sentences). The participants were asked to score the sentences from 1 to 5 according to how suitable they found the voice to be used as a communication voice by people with speech impairments. A total of 25 people took part in the evaluation.

Figure 2 shows the MOS scores obtained by each voice, together with the 95% confidence interval. The voices are ordered left-to-right from the highest to the lowest MOS. For each voice, the measure (STOI, ESTOI, SIIB or NISQA) with that voice among the best five is also shown. For example, SPK02 was among the first five positions for STOI, ESTOI, and NISQA.

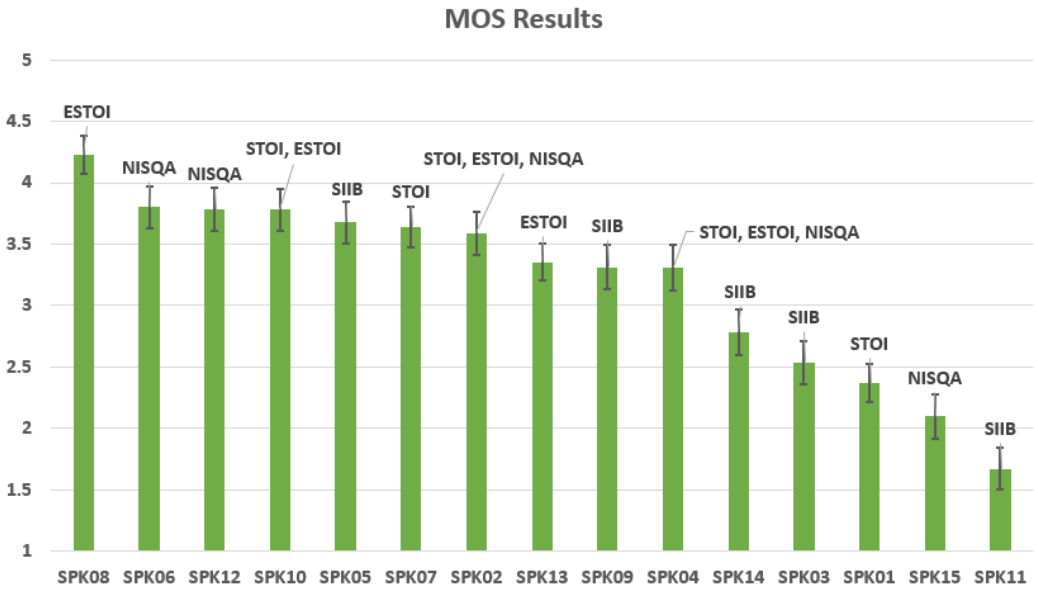

Figure 2. MOS results with 95% confidence interval.

As it can be seen, the best voices according to MOS are selected by different objective measures. Table 4 shows how the scores obtained by the objective measures correlate with the evaluated MOS. The most correlated measure is ESTOI, which has its selected best voices among the top seven positions in Figure 2. It is followed by STOI with its 5 selected voices among the 13 top positions, NISQA with its 5 selected voices among the top 14, and finally SIIB among the top 15.

Table 4. Correlations between MOS and objective measures.

STOI	ESTOI	NISQA	SIIB
0.452	0.584	0.356	−0.199

Considering that each objective measure uses different criteria, we have investigated whether a combination of the measures can produce a better estimation than the individual measures. The method and results are described in the next subsection.

5.4. MOS Prediction

Using the available objective measures and the MOS scores obtained in the subjective evaluation, we have built an estimator of the MOS based on linear regression. This predictor may then be applied to the thousands of synthetic voices from the voice bank in order to automatically select the ones with best MOS estimations to populate the voice catalog offered to the voice bank users.

To be able to estimate the prediction ability of the linear regression a leave-one-out strategy was applied: 15 different linear regression polynomials were built using in each case data from 14 speakers to train and the remaining 1 to test. SIIB presents a special difficulty in being included in the regression, as it is not possible to obtain one SIIB value per sentence due to its minimum length requirement of 20 s to be reliably calculated. As commented in Section 4.2, several sentences were concatenated to be able to reliably calculate SIIB, and consequently there is no available value for each sentence. Taking this issue into account, and considering also that SIIB correlates negatively with the MOS obtained in the subjective evaluation and that three of the five voices selected according to this measure have less than 3 in MOS, we decided not to include SIIB in the regression.

Each regression polynomial was evaluated over the 10 sentences from the speaker not included in the training set. The mean value of the predicted MOS together with the corresponding 95% confidence interval are shown in Figure 3.

We can see that the prediction is accurate with no significant differences for 8 out of the 15 evaluated voices (voices SPK06, SPK12, SPK10, SPK02, SPK13, SPK09, SPK03, and SPK11). For voices SPK08, SPK05 and SPK07, the MOS was underestimated, while for SPK01 and SPK15, it was overestimated.

The correlation between the predicted MOS and the actual MOS obtained in the evaluation is $\rho = 0.546$. Therefore, it is higher than the correlation obtained with all objective measures individually (see Table 4), except for ESTOI, which has a slightly higher correlation. This suggests that although using only ESTOI can produce a useful first selection, combining it with other objective measures can help in considering other high-quality voices that would otherwise be left out. Applying a suitable threshold over the predicted MOS (higher than 3.5 for instance) removes most of the poor quality voices and would populate the synthetic voice catalog mostly with voices of acceptable quality.

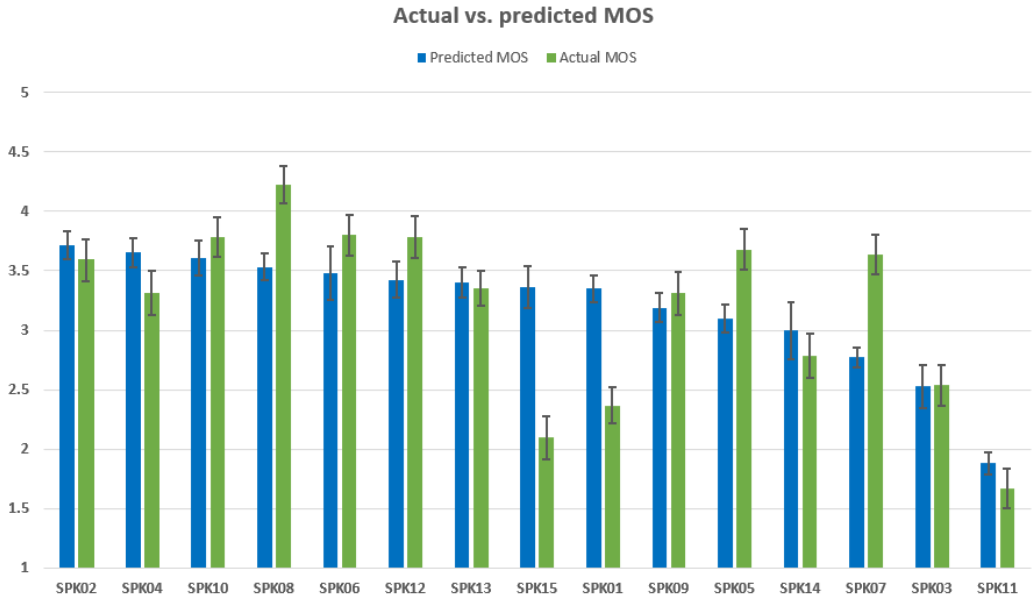

Figure 3. Actual MOS vs. predicted MOS, mean and 95% confidence interval.

6. Conclusions and Future Work

The investigation of methods to objectively evaluate synthetic voices has gained increased interest in the last years. Data-driven methods using deep neural networks have also gained popularity. In this paper, we describe the experiments developed with the aim of finding a method to select the best synthetic voices among a high number of voices of unknown quality with a reduced set of sentences. Using several existing objective measures of intelligibility and naturalness to rank the voices, we analyzed the performance of each measure in relation to the ranking obtained with a MOS evaluation. The evaluated measures were STOI, ESTOI, NISQA and SIIB as described in the paper.

The evaluated synthetic voices were obtained with statistical parametric synthesis methods based on HMMs. The selected objective measures, however, were designed to be applied in applications such as evaluation of speech in noise, speech enhancement algorithms and others. In general, a reference signal is needed to obtain the measure (i.e., they are intrusive measures), which is not directly available in TTS. The non-intrusive method NISQA was developed using data extracted from challenges where the participants competed with very high quality synthetic voices, while the quality of the voices here presented can be described as being of low-to-medium quality. On the other hand, some measures aim at evaluating naturalness (NISQA), while others (STOI, ESTOI and SIIB) are designed to measure intelligibility. Additionally, some measures (SIIB) need several sentences of speech to be reliably obtained. The decision of which is the most suitable measure is not straightforward, and this was the main motivation of the work presented in this paper.

The first conclusion that can be drawn from the experiments is that each measure provides a different set of best voices candidates. The measure that best correlates with the performed MOS evaluation scores is ESTOI. To test if a combination of measures would provide a better result than a single measure, a simple linear regression algorithm was built to predict MOS, considering all the individual measures, except SIIB. The correlation with actual MOS scores is slightly smaller than the one obtained by ESTOI, but higher than the one obtained with the rest of individual measures. A more elaborate prediction algorithm could provide more accurate results and is left for future work.

One limitation of the work is the fact that only those voices automatically selected and ranked by the objective measures were evaluated with a MOS evaluation. In addition, we arbitrarily set a boundary on the top five positions, in order to limit the number of candidates to be subjectively evaluated. It is of course possible that voices which are further

away from that boundary could obtain better MOS scores. However, adding more positions would have increased the number of voices to evaluate, which in turns requires a higher number of participants. A crowd sourcing based evaluation could help with this issue. We plan to use such a technique to check the performance of the predictor over the entire set of available voices.

The voice bank portal is being updated to use neural network based synthesis techniques. The developed method can then be applied to the new obtained adapted voices so a new catalog is generated.

Author Contributions: Conceptualization, A.A. and I.H.; methodology, A.A., V.G., I.H. and E.N.; software, A.A., V.G. and J.S.; validation, A.A., I.H. and E.N.; formal analysis, A.A., I.H. and E.N.; investigation, A.A. and I.H.; resources, J.S.; data curation, J.S; writing—original draft preparation, A.A. and I.H.; writing—review and editing, V.G., I.H., E.N. and J.S.; visualization, A.A.; supervision, I.H. and E.N.; project administration, I.H. and E.N.; funding acquisition, I.H. and E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been funded by the Basque Government under the project ref. PIBA 2018-035 and IT-1355-19. This work is part of the project Grant PID 2019-108040RB-C21 funded by MCIN/AEI/10.13039/501100011033.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to Europe GDPR.

Acknowledgments: The voice bank could not exist without all the donors who have donated their voices in an altruistic way so other people can recover theirs. We deeply thank all of them.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lavan, N.; Mileva, M.; McGettigan, C. How does familiarity with a voice affect trait judgements? *Br. J. Psychol.* **2021**, *112*, 282–300. [CrossRef] [PubMed]
2. Pucher, M.; Zillinger, B.; Toman, M.; Schabus, D.; Valentini-Botinhao, C.; Yamagishi, J.; Schmid, E.; Woltron, T. Influence of speaker familiarity on blind and visually impaired children's and young adults' perception of synthetic voices. *Comput. Speech Lang.* **2017**, *46*, 179–195. [CrossRef]
3. Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; Volume 1, pp. 373–376.
4. Zen, H.; Tokuda, K.; Black, A. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [CrossRef]
5. Pollet, V.; Breen, A. Synthesis by generation and concatenation of multiform segments. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
6. Dudley, H. Remaking speech. *J. Acoust. Soc. Am.* **1939**, *11*, 169–177. [CrossRef]
7. Kawahara, H.; Masuda-Katsuse, I.; De Cheveigne, A. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.* **1999**, *27*, 187–207. [CrossRef]
8. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
9. Erro, D.; Sainz, I.; Navas, E.; Hernaez, I. Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis. *IEEE J. Sel. Top. Signal Process.* **2014**, *8*, 184–194. [CrossRef]
10. Yamagishi, J.; Usabaev, B.; King, S.; Watts, O.; Dines, J.; Tian, J.; Guan, Y.; Hu, R.; Oura, K.; Wu, Y.J.; et al. Thousands of voices for HMM-based speech synthesis—Analysis and application of TTS systems built on various ASR corpora. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 984–1004. [CrossRef]
11. Suni, A.; Raitio, T.; Vainio, M.; Alku, P. The GlottHMM speech synthesis entry for Blizzard Challenge 2010. In *The Blizzard Challenge 2010 Workshop*; Language Technologies Institute LTI: Pittsburgh, PA, USA, 2010; pp. 1–6.
12. Ze, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.

13. Ling, Z.H.; Kang, S.Y.; Zen, H.; Senior, A.; Schuster, M.; Qian, X.J.; Meng, H.M.; Deng, L. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Process. Mag.* **2015**, *32*, 35–52. [CrossRef]
14. Arık, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J.; et al. Deep voice: Real-time neural text-to-speech. In Proceedings of the International Conference on Machine Learning (PMLR), Sydney, NSW, Australia, 6–11 August 2017; pp. 195–204.
15. Arik, S.; Diamos, G.; Gibiansky, A.; Miller, J.; Peng, K.; Ping, W.; Raiman, J.; Zhou, Y. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv* **2017**, arXiv:1705.08947.
16. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.Ö.; Kannan, A.; Narang, S.; Raiman, J.; Miller, J. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. 2017. *arXiv* **2017**, arXiv:1710.07654.
17. Sotelo, J.; Mehri, S.; Kumar, K.; Santos, J.F.; Kastner, K.; Courville, A.; Bengio, Y. Char2wav: End-to-End Speech Synthesis. 2017. Available online: https://mila.quebec/wp-content/uploads/2017/02/end-end-speech.pdf (accessed on 10 February 2022).
18. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
19. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
20. Ping, W.; Peng, K.; Chen, J. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv* **2018**, arXiv:1807.07281.
21. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
22. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
23. Toman, M.; Meltzner, G.S.; Patel, R. Data Requirements, Selection and Augmentation for DNN-based Speech Synthesis from Crowdsourced Data. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 2878–2882.
24. Taigman, Y.; Wolf, L.; Polyak, A.; Nachmani, E. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv* **2017**, arXiv:1707.06588.
25. Jia, Y.; Zhang, Y.; Weiss, R.J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Moreno, I.L.; et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv* **2018**, arXiv:1806.04558.
26. Cooper, E.; Lai, C.I.; Yasuda, Y.; Fang, F.; Wang, X.; Chen, N.; Yamagishi, J. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6184–6188.
27. Erro, D.; Hernáez, I.; Navas, E.; Alonso, A.; Arzelus, H.; Jauk, I.; Hy, N.Q.; Magariños, C.; Pérez-Ramón, R.; Sulír, M.; et al. ZureTTS: Online Platform for Obtaining Personalized Synthetic Voices. In Proceedings of the eNTERFACE'14, Bilbao, Spain, 9 June–5 July 2014.
28. Laboratory, A.S.P. ZureTTS. Available online: https://aholab.ehu.eus/ahomytts/ (accessed on 10 February 2022).
29. Alonso, A.; García, V.; Hernaez, I.; Navas, E.; Sanchez, J. Automatic Speaker Adaptation Assessment Based on Objective Measures for Voice Banking Donors. In Proceedings of the IBERSPEECH 2020, Valladolid, Spain, 24–26 March 2021; pp. 210–214.
30. Model Talker. Available online: https://www.modeltalker.org/ (accessed on 10 February 2022).
31. Speak Unique. Available online: https://www.speakunique.co.uk/ (accessed on 10 February 2022).
32. VocalID. Available online: https://vocalid.ai/ (accessed on 10 February 2022).
33. The Voice Keeper. Available online: https://thevoicekeeper.com/ (accessed on 10 February 2022).
34. Acapela Group. Available online: http://www.acapela-group.com/solutions/my-own-voice/ (accessed on 10 February 2022).
35. CereVoice. Available online: https://www.cereproc.com/en/products/cerevoiceme (accessed on 10 February 2022).
36. Yamagishi, J.; Veaux, C.; King, S.; Renals, S. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoust. Sci. Technol.* **2012**, *33*, 1–5. [CrossRef]
37. Creer, S.; Cunningham, S.; Green, P.; Yamagishi, J. Building personalised synthetic voices for individuals with severe speech impairment. *Comput. Speech Lang.* **2013**, *27*, 1178–1193. [CrossRef]
38. Pierard, A.; Erro, D.; Hernaez, I.; Navas, E.; Dutoit, T. Surgery of speech synthesis models to overcome the scarcity of training data. In Proceedings of the International Conference on Advances in Speech and Language Technologies for Iberian Languages, Lisbon, Portugal, 23–25 November 2016; pp. 73–83.
39. Yamagishi, J.; Nose, T.; Zen, H.; Ling, Z.H.; Toda, T.; Tokuda, K.; King, S.; Renals, S. Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. *IEEE Trans. Audio, Speech Lang. Process.* **2009**, *17*, 1208–1230. [CrossRef]
40. Moreno Bilbao, M.A.; Poig, D.; Bonafonte Cávez, A.; Lleida, E.; Llisterri, J.; Mariño Acebal, J.B.; Nadeu Camprubí, C. Albayzin speech database: Design of the phonetic corpus. In Proceedings of the EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology, Berlin, Germany, 22–25 September 1993; pp. 175–178.
41. Sainz, I.; Erro, D.; Navas, E.; Hernáez, I.; Sanchez, J.; Saratxaga, I.; Odriozola, I. Versatile Speech Databases for High Quality Synthesis for Basque. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 23–25 May 2012; pp. 3308–3312.

42. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4214–4217.
43. Jensen, J.; Taal, C.H. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio Speech, Lang. Process.* **2016**, *24*, 2009–2022. [CrossRef]
44. Van Kuyk, S.; Kleijn, W.B.; Hendriks, R.C. An instrumental intelligibility metric based on information theory. *IEEE Signal Process. Lett.* **2017**, *25*, 115–119. [CrossRef]
45. Mittag, G.; Möller, S. Deep Learning Based Assessment of Synthetic Speech Naturalness. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 1748–1752.
46. Jørgensen, S.; Cubick, J.; Dau, T. Speech intelligibility evaluation for mobile phones. *Acta Acust. United Acust.* **2015**, *101*, 1016–1025. [CrossRef]
47. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time—Frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
48. Falk, T.H.; Parsa, V.; Santos, J.F.; Arehart, K.; Hazrati, O.; Huber, R.; Kates, J.M.; Scollie, S. Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. *IEEE Signal Process. Mag.* **2015**, *32*, 114–124. [CrossRef]
49. Xia, R.; Li, J.; Akagi, M.; Yan, Y. Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4465–4468.
50. Jensen, J.; Taal, C.H. Speech intelligibility prediction based on mutual information. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 430–440. [CrossRef]
51. Schlittenlacher, J.; Baer, T. Text-to-speech for the hearing impaired. *arXiv* **2020**, arXiv:2012.02174.
52. Ayllón, D.; Sánchez-Hevia, H.A.; Figueroa, C.; Lanchantin, P. Investigating the Effects of Noisy and Reverberant Speech in Text-to-Speech Systems. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019; pp. 1511–1515.
53. Janbakhshi, P.; Kodrasi, I.; Bourlard, H. Pathological Speech Intelligibility Assessment Based on the Short-time Objective Intelligibility Measure. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6405–6409. [CrossRef]
54. Mittag, G.; Möller, S. Non-intrusive speech quality assessment for super-wideband speech communication networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7125–7129.
55. NISQA. Available online: https://github.com/gabrielmittag/NISQA (accessed on 10 February 2022).
56. McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; Sonderegger, M. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 498–502.
57. Tokuda, K.; Zen, H.; Yamagishi, J.; Black, A.; Masuko, T.; Sako, S.; Oura, K.; Hasimoto, K.; Sawada, K. Yoshimura, T. The HMM-based speech synthesis system (HTS). Available online: http://hts.sp.nitech.ac.jp (accessed on 10 February 2022).
58. Erro, D.; Hernaez, I.; Alonso, A.; García-Lorenzo, D.; Navas, E.; Ye, J.; Arzelus, H.; Jauk, I.; Hy, N.Q.; Magariños, C.; et al. Personalized synthetic voices for speaking impaired: Website and app. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.