IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

AUTOMATIC SPEAKER ADAPTATION ASSESSMENT BASED ON OBJECTIVE MEASURES FOR VOICE BANKING DONORS

Agustin Alonso, Victor García, Inma Hernaez, Eva Navas, Jon Sanchez

HiTZ Basque Center for Language Technologies - Aholab, University of the Basque Country
UPV/EHU

agustin@aholab.ehu.eus, {victor.garcia, inma.hernaez, eva.navas, jon.sanchez}@ehu.eus

Abstract

Speech is the most common way of communication. People who have lost total or partially their ability to speak might benefit from the use of Alternative and Augmentative Communication (AAC) devices and the use of Text-to-Speech (TTS) technology. One problem that arouses is that the synthetic voices included in these devices might be impersonal and not accurate to the user terms of age, accent or even gender. Therefore, voice banking has become a good alternative to standard commercial voices. In our voice banking strategy, people with healthy voice (donors), or the user itself before losing his or her own voice, provide the recordings to obtain a new synthetic voice using adaptation techniques. In this way, a wide catalog of synthetic voices is provided to the potential user. However, because there is no control over the recording process, the final quality of the synthetic voice is very variable. In this paper, we propose a method to assess the result of the adaptation using objective measures. The results show that this strategy can be an alternative to subjective evaluation to select the best donated voices for the voice bank.

Index Terms: STOI, ESTOI, NISQA, speech adaptation, voice banking

1. Introduction

Speech is the most common and intuitive way to communicate between humans. Sadly, in some cases people lose total or partially their ability to speak due to an accident or illness. In these cases, speech technologies can help those with some speech impairment. One possibility is providing them with a Text-to-Speech (TTS) system, so entering a text as input the system can reproduce the output with a synthetic voice. In general, this solution leads to general and non-emotive voices. Furthermore, the default voices could not suit the user preferences in terms of age, accent or gender (e.g. a young woman could use a voice of an old man). A solution is providing them a personalized voice. With a small amount of samples of a healthy voice, a new voice can be adapted for the TTS system. In this context voice banking has become a popular solution. If the person who wants to use the system still preserves his/her voice, he/she can make the necessary recordings before loosing his/her voice (due a scheduled surgery or a degenerative illness) and use his/her own adapted voice as a backup. Another approach consists on altruistic donors recording their voice which is used to train new ones for impaired people. Then, the user can choose their favourite one from the voice bank catalog. In these cases voice banks can handle hundreds or even thousands of donors, so an important task is classifying the result of adapting all these donors. The quality of the adapted voice, understood as how clean and intelligible it sounds, depends on several factors (initial quality of the recordings, how accurate the segmentation done during the training is, etc.) and measuring it manually is a great effort.

In this paper we analyze a method for automatically assess the adapted synthetic voices of our voice bank based on objective measures. We use two objective measures typically used in speech enhancement STOI [1] and ESTOI [2] and an objective measure based on NISQA that tries to estimate the naturalness MOS (Mean Opinion Score) of synthetic speech [3].

The rest of the article is structured as follows: in section 2 we explain voice banks in general and ours in more detail. Section 3 describes the objective measures we use in our analysis. The proposed analysis method 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 Banks

Voice banks are an alternative to provide people with speech difficulties with a personalized voice. In these voice banks, a personalized synthetic voice can be generated using adaptation techniques applied to a small sample of a healthy voice. There are several voice banks like ModelTalker [4], Speak Unique [5], VocalID [6] and the Voice Keeper [7].

Our voice bank, ZureTTS [8], offers the possibility of obtaining a personalized voice in Spanish and Basque. It makes use of a statistical synthesis engine based on Hidden Markov Models (HMMs) [9]. Each user must record a total of 100 phonetically balanced sentences in the selected language. These are parameterized using ahocoder [10], a high-quality vocoder that extracts MCEP coefficients of order 39, log-f_0 and maximum voice frequency. These data are then used to adapt an average voice using state of the art adaptation techniques [11] 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 [12] database. It consists of 6800 phrases 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 [13]. This consists of 9 speakers (5 female and 4 male) all of which include 1 hour of speech except for two (female and male) which include 6 hours each. Currently, our voice bank has almost 9000 registered users.

3. Objective Measures Overview

3.1. STOI and ESTOI

In the field of intelligibility, several algorithms have been proposed that try to replace expensive subjective listening tests. Among them, STOI (Short Time Objective Intelligibility) [1] has proven to be good for evaluating intelligibility in signals

10.21437/IberSPEECH.2021-45

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 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 that so the intermediate measure depends on speech information from the last $\approx 400ms$. 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. STOI has proven to predict intelligibility quite accurately in different situations, such as mobile phone output [14], noisy speech processed by ideal time-frequency masking and single channel speech enhancement algorithms [15] and speech processed by cochlear implants [16] and it is also robust against different types of languages like Mandarin [17], Danish [15] or Dutch [18].

One evolution of STOI is Extended STOI (ESTOI) [2], which unlike STOI does not assume independence between frequency bands. This feature allows to better capture the effect of time-modulated noise maskers.

The success of these measures has led to propose using them in several areas, for example, to evaluate the intelligibility of dysarthric speech [19]. In this work, STOI - ESTOI could not be applied directly since the time-aligned reference signal was not available. To overcome this problem, an utterance-dependent reference signal was generated from several healthy speakers and Dynamic Time Warping (DTW) was used to align the pathological signal and the reference signal.

3.2. NISQA

An important aspect to evaluate in synthetic voices is naturalness. In [3] a method based on Non-Intrusive Speech Quality Assessment (NISQA) [20] is proposed to measure the naturalness of synthetic voices without the need for a reference signal. The prediction model they propose is based on a CNN-LSTM network architecture with transfer learning domain knowledge from a speech quality database. It has been trained using 16 databases with 12 different languages, so it is presented as a language independent method that can be used in any TTS. Furthermore, the model has been trained using signals with different speech levels, to be able to be used with different types of signals. This makes it suitable for our voice banking case, were the adapted models are obtained with recordings without any control about speech level. The model is publicly available at [21] so it can be used directly.

4. Proposed Methodology

In this section we describe the method followed to obtain the objective measures. For STOI and ESTOI, the original recordings for each speaker are set as clean references and the synthetic signals are generated using the adapted voice model for each speaker. To obtain the NISQA score no processing of the synthetic signals is required. Once the predicted scores are obtained, we performed a clustering to identify the representative speakers, which will be included in the subjective evaluation.

4.1. STOI-ESTOI

The first step is to obtain the phonetic segmentation of the recordings. This is done by forced alignment using Montreal Forced Aligner (MFA) [22]. This step will be useful for two main reasons:

- It will provide the actual positions for the pauses made by the speaker during the recordings. The synthetic signals will be generated using this information, thus with the pauses at the same locations. We must consider that announcers can skip pauses indicated by spelling signs, or on the contrary, make pauses that are not in the text.
- Have the segmentation available for later use in the alignment between the reference and the signal whose intelligibility is going to be evaluated.

Next, the synthetic signals corresponding to each speaker are generated, with the previously detected pauses. In this way, a parallel corpus of recordings-synthetic signals is available. However, these signals will have different duration, and 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 phoneme level. For the alignment, the cepstral distances between reference and 'target' are calculated. The cepstral coefficients are obtained directly from the adapted voice model for synthetic signals, while for reference signals, they are obtained using Ahocoder [10]. It has also been necessary to adapt the frame rate of the synthesis system to the one used by the STOI / ESTOI algorithm.

After these steps, a score can be obtained with the STOI and ESTOI algorithm for each sentence. The final score for each speaker will be the average of the scores obtained for the 100 available sentences.

4.2. 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.

4.3. Clustering

We have three objective measures for each donor: STOI, ESTOI and NISQA. Since the scale of the three is different, the z-score of each one is calculated by normalizing by mean and variance so that all of them have the same importance in the clustering process. The donors are then grouped by applying the k-means clustering algorithm using the normalized objective measures.

5. Experiments

5.1. Experimental Setup

The proposed method has been tested using the recordings and the adapted synthetic voices in Spanish from ZureTTS voice bank [23]. A total of 1090 voices have been used.

Considering that users evaluate the quality of the synthetic voices using a 5 grade scale, we have arbitrarily set a final number of 5 clusters. In order to find out which objective measure is more important for the users when evaluating their preference for one synthetic voice or another, the clustering has been done in two different ways.

A) Taking into account only intelligibility related measures, i.e. STOI and ESTOI

B) Taking into account the three measures

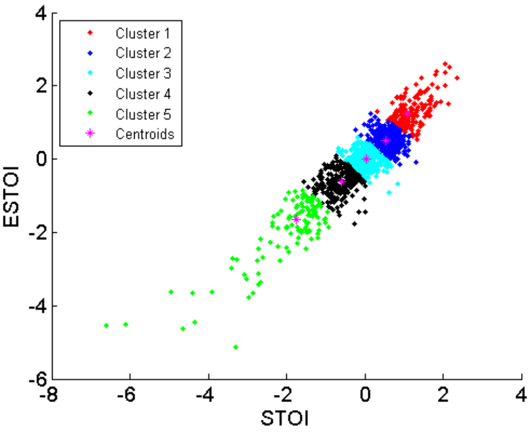


Figure 1: *Clustering using normalized STOI and ESTOI averaged scores.*

The scatter plots of figures 1 and 2 show the computed normalized scores and the resulting clusters. Figure 1 corresponds to the clustering performed when only STOI and ESTOI measures are used. As expected both measures are highly correlated ($\rho = 0.912$) and the clustering shows clear linear boundaries. When the naturalness score is incorporated (Figure 2) the final clusters are modified. The changes mainly affect to clusters 2 and 3, which are separated in the NISQA dimension (figures 2b and 2c), but mixed in the STOI and ESTOI dimensions (figure 2a).

As reference donors to represent each cluster, the centroids are chosen. The objective measures values for these centroids are shown in table 1. Only the measures used for the clustering are shown.

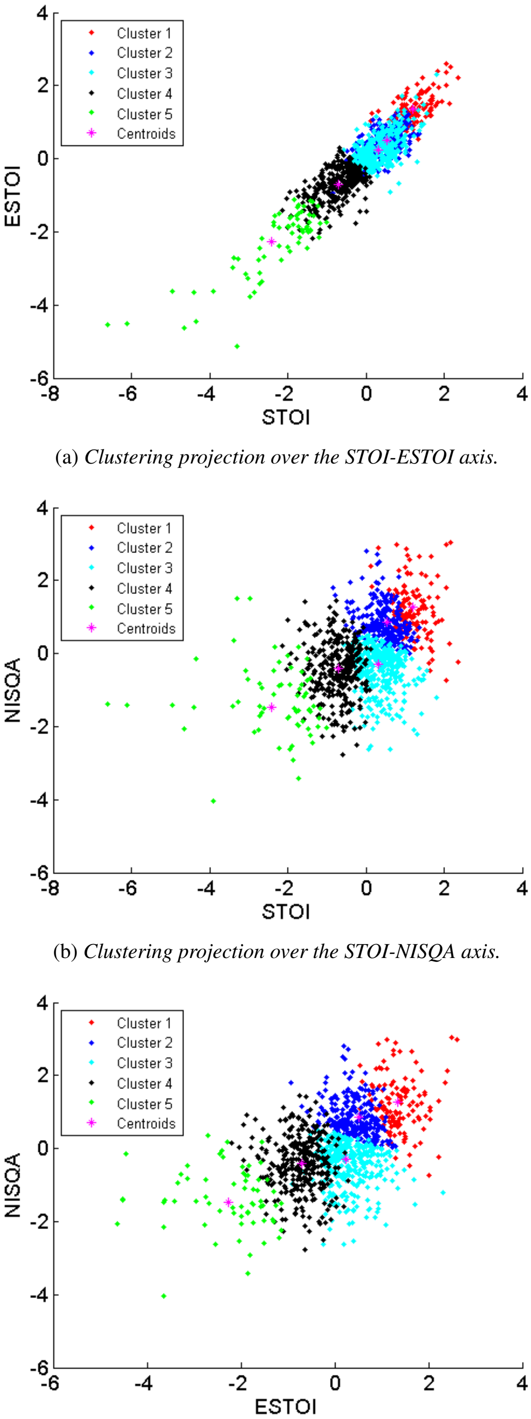
Table 1: *Objective measures for representative donors*

Donor	STOI	ESTOI	NISQA
SPK1	0.6832	0.5045	-
SPK2	0.6689	0.4689	-
SPK3	0.6155	0.4220	-
SPK4	0.5747	0.3790	-
SPK5	0.5015	0.3093	-
SPK6	0.6895	0.5122	3.0036
SPK2	0.6689	0.4689	2.8210
SPK7	0.6343	0.4368	2.5261
SPK8	0.5691	0.3732	2.4926
SPK9	0.4601	0.2679	1.1778

Speakers SP1 to SPK5 are the centroids corresponding to clustering A, while speakers SPK6 to SPK9 correspond to the centroids of clustering B. SPK2 coincided as a centroid in both clusterings.

5.2. MOS Evaluation

To check how the clustering results agree with people's preferences, a subjective evaluation was carried out. Using the adapted models from the 9 representative donors 10 new short sentences were synthesized, i.e. 90 sentences in total. 15 people took part in the evaluation, 5 of them were experts in speech technologies. Each one scored 5 randomly selected sentences from each donor, so each evaluator evaluated 45 sentences. The sentences were presented in a simple web interface where for each case they had to score on a MOS scale of 1 to 5 a single question: "Would you use this synthetic voice in a TTS system?" where 1 meant "No, I do not like it and I would not use



(a) *Clustering projection over the STOI-ESTOI axis.*

(b) *Clustering projection over the STOI-NISQA axis.*

(c) *Clustering projection over the ESTOI-NISQA axis.*

Figure 2: *Clustering using normalized STOI, ESTOI and NISQA averaged scores.*

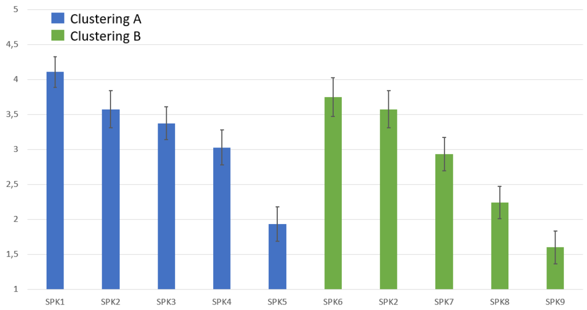


Figure 3: *MOS results with 95% confidence interval.*

it" and 5 meant "Yes, I like how it sounds and I would to use it". Some evaluation samples can be listened here[1].

The results of the subjective evaluation are shown in figure 3.

As it can be observed in the figure, the ordering of the speakers is the same for the two considered classification systems (Clustering A and Clustering B). In fact, the three objective measures, considered independently, are highly correlated with the obtained subjective MOS scores. Table 2 shows the correlation coefficient between the obtained final scores, using the nine donors for STOI and ESTOI and the five donors of Clustering B for NISQA. We can see that intelligibility scores correlation with MOS is higher than for naturalness. We can also observe from figure 3 that the set of donors from clustering A have obtained better scores than those from Clustering B. A possible interpretation is that when choosing a synthetic voice, the features measured by the NISQA algorithm are not as relevant as those measured by the intelligibility measures STOI and ESTOI for the evaluators. Also, naturalness is not relevant when intelligibility is not guaranteed, as is the case for some adapted voices in our system.

Table 2: *Correlation coefficient between subjective MOS and objective measures*

	STOI	**ESTOI**	**NISQA**
ρ	0,9484	0,9500	0,7858

6. Conclusions and Future Work

We have seen how the use of objective measures on the result of the adaptation in our voice bank can be used to make a first classification in the voice catalog. The subjective evaluation using the representative donors of each cluster confirms that the better the objective measures, the better the acceptance of the synthetic voice by the users. We can use the result of the clustering to do an initial categorization and set the donors from the first cluster as voice bank catalog. Clustering performed without using NISQA naturalness measure has resulted in more MOS correlated scores, so calculation of only STOI and ESTOI seems enough to estimated people's perception. However, as future work, we plan to consider more measures such as DAU [24] or GP (Glimpse Proportion) [25] [26] to the study, trying to improve clustering and thus a more accurate automatic classification of adapted voices.

[1] https://aholab.ehu.eus/users/agustin/demos/ib20/

7. Acknowledgements

This work has been funded by the Basque Government under the project ref. PIBA 2018-035 and IT-1355-19.

8. References

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[2] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[3] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *Proc. Interspeech 2020*, pp. 1748–1752, 2020.

[4] *Model Talker*. [Online]. Available: https://www.modeltalker.org/

[5] *Speak Unique*. [Online]. Available: https://www.speakunique.co.uk/

[6] *VocalID*. [Online]. Available: https://vocalid.ai/

[7] *The Voice Keeper*. [Online]. Available: https://thevoicekeeper.com/

[8] D. Erro, I. Hernáez, E. Navas, A. Alonso, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños, R. Pérez-Ramón, M. Sulír, X. Tian, X. Wang, and J. Ye, "ZureTTS: Online platform for obtaining personalized synthetic voices," in *Proc. eNTERFACE'14*, 2014.

[9] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[11] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[12] M. A. Moreno Bilbao, D. Poig, A. Bonafonte Cávez, E. Lleida, J. Llisterri, J. B. Mariño Acebal, and C. Nadeu Camprubí, "Albayzin speech database: Design of the phonetic corpus," in *EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993*. . EUROSPEECH, 1993, pp. 175–178.

[13] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sanchez, I. Saratxaga, and I. Odriozola, "Versatile speech databases for high quality synthesis for basque." in *LREC*. Citeseer, 2012, pp. 3308–3312.

[14] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, no. 5, pp. 1016–1025, 2015.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[16] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[17] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4465–4468.

[18] J. Jensen and C. H. Taal, “Speech intelligibility prediction based on mutual information,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 2, pp. 430–440, 2014.

[19] P. Janbakhshi, I. Kodrasi, and H. Bourlard, “Pathological speech intelligibility assessment based on the short-time objective intelligibility measure,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6405–6409.

[20] G. Mittag and S. Möller, “Non-intrusive speech quality assessment for super-wideband speech communication networks,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7125–7129.

[21] *NISQA*. [Online]. Available: https://github.com/gabrielmittag/NISQA

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: trainable text-speech alignment using kaldi,” in *Proceedings of INTERSPEECH*, 2017, pp. 498–502.

[23] D. Erro, I. Hernaez, A. Alonso, D. García-Lorenzo, E. Navas, J. Ye, H. Arzelus, I. Jauk, N. Q. Hy, C. Magariños *et al.*, “Personalized synthetic voices for speaking impaired: Website and app,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] C. Christiansen, M. S. Pedersen, and T. Dau, “Prediction of speech intelligibility based on an auditory preprocessing model,” *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, 2010.

[25] M. Cooke, “A glimpsing model of speech perception in noise,” *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[26] Y. Tang, M. Cooke *et al.*, “Glimpse-based metrics for predicting speech intelligibility in additive noise conditions.” in *Interspeech*, 2016, pp. 2488–2492.