The 35th Canadian Conference on Artificial Intelligence
DOI: 0

Voice Cloning Applied to Voice Disorders: a Study of Extreme Phonetic Content in Speaker Embeddings

Lily Wadoux[†,*], Nelly Barbot[†,*], Jonathan Chevelu[†,*], Damien Lolive[†,*]
[†] Univ Rennes, CNRS, IRISA

Abstract

Organic dysphonia can lead to vocal impairments. Recording patients' impaired voice could allow them to use voice cloning systems. In the domain of speech synthesis, voice cloning is the process of producing speech matching a target speaker voice, given textual input and an audio sample from the speaker. It can achieve high-quality speech with only few data from the target speaker. However, dysphonic patients may only produce speech with specific or limited phonetic content. To our knowledge, the impact of such constraints on a voice cloning system remains to be studied. This article presents the results of preliminary experiments on the matter, along with specifications about the models and datasets used.

Keywords: voice cloning, speaker encoder, Text-to-Speech, x-vector, voice disorders

1. Introduction

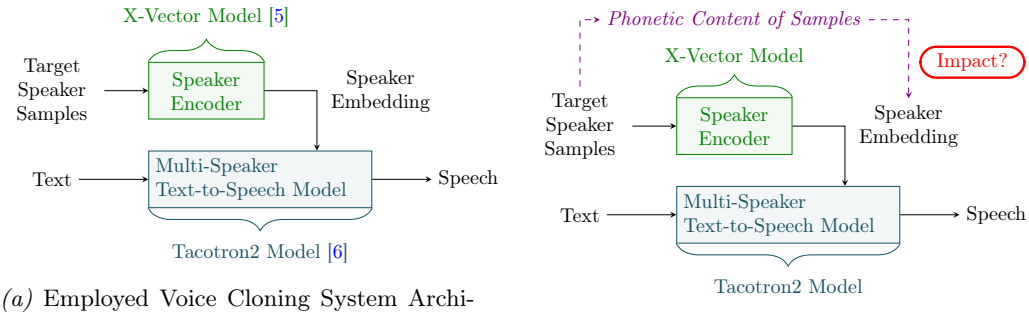
Organic dysphonia can lead to serious vocal damage [1]. As it deteriorates communication, this disability can cause social isolation. Besides, as the voice is a personal way of expression, it can be considered as part of a person's identity. This is why it would be an interesting possibility to use speech synthesis devices, fed by patients' voice data, to improve their speech intelligibility. However, patients' health condition presents a number of constraints which can impact voice recording. Long recording sessions can prove very tiring, inducing more vocal instability and the pathology can highly restrict the phonetic coverage. In a context of speech synthesis, this medical application would require a study on the impact of the patient's vocal corpus' content and duration on the synthesized speech.

Neural-network based Text-to-Speech (TTS) systems produce speech given textual input. They are trained on aligned text and audio samples, organized in a corpus, containing data from one or several speakers. Speech matching a target speaker voice can be produced with a Text-to-Speech system. However, it requires target speaker samples in the training corpus.

Voice cloning methods, such as speaker adaptation and speaker encoding, offer more flexibility and can generate speech from speakers unseen during training [2–4]. Speaker adaptation relies on a second training step when the pre-trained multi-speaker TTS model is specialised, or fine-tuned, to produce only the target speaker voice. To generate speech matching another speaker, the pre-trained model must be fine-tuned for the new speaker. For speaker encoding, on the contrary, no fine-tuning step is required. Instead, a second model, called speaker encoder, outputs to the TTS model a vectorial representation of speaker features, called speaker embedding, as illustrated in figure 1a. To match another speaker, new audio samples are simply given as input to this encoder.

Both approaches need relatively few data from the target speaker, with very good results from ten minutes of speech, and good results even from ten seconds samples [4]. In this study, we use the speaker encoder approach, and the x-vector model in particular. Indeed, despite slightly lower results compared to speaker adaptation [3], it only needs one training phase,

[*] {lily.wadoux, nelly.barbot, jonathan.chevelu, damien.lolive}@irisa.fr



(a) Employed Voice Cloning System Architecture

(b) Context of the Considered Study

Figure 1. Employed Voice Cloning System Architecture (left) and Context Precision for the Considered Study (right).

facilitating its generalisation to new speakers. This fits with the idea of being accessible, in the long term, for as many patients as possible.

This article presents a preliminary study on the impact of target speaker corpus' phonetic content on speaker embeddings. The considered medical application is detailed in section 2. Experimental protocol is defined in section 3, training settings and data in section 4. Last, results are discussed in section 5.

2. Medical Application

Dysphonia can be defined as an alteration of the voice timber, but also, more broadly, as a momentary or lasting disorder of the vocal function, felt as such by the subject or relatives [1]. Dysphonia can be of organic or functional origin, depending whether it is maintained principally by vocal gesture disturbances or caused by organic disorders. Here, only organic dysphonia are considered for voice cloning. Indeed, dysfunctional dysphonia symptoms, in most cases, can be greatly improved through speech therapy, and uttered speech remains understandable. As for organic dysphonia, most symptoms can also be improved by speech therapy or surgery. However, especially for degenerative organic dysphonia such as pharynx and larynx cancers [7] and amyotrophic lateral sclerosis, sometimes comes a time when uttered speech is very damaged and difficult to understand. We believe voice cloning could be useful for such patients. Moreover, some pathologies such as stenosis, Riegel and Gerhardt syndromes, larynx and pharynx cancers, can require surgery. Post-surgery speech can be impossible, or difficult to utter and understand, with possible improvements or lasting effects. Voice cloning could be a useful communication tool here, while it should not substitute with speech therapy and a regular use of natural speech.

Damaged voice can take various forms depending on pathologies, patients, and degrees of evolution. Common symptoms can include alterations of timbre or pitch, vocal irregularities, intermittent rhythm, and articulation disorders like a non-differentiation of consonants and vowels and even a disappearance of consonants. The variety and range of symptoms make it difficult to thoroughly simulate pathological voices from a healthy voice. For this study, extreme phonetic content strategies, described section 4, are considered to simulate some of the symptoms: *MSW* and *Phn* which can be linked to random, intermittent speech, *Vowels* corresponding to a voice with consonant disappearance and *Phn-A* which is closer to a voice with consonant disappearance and vowel non-differentiation. The remaining strategy, *Sentences*, represents a speech baseline. Further studies could include actual dysphonia samples, once difficulties regarding their availability and protection are overcome.

3. Experimental Protocol

For this first investigation of the impact of the target speaker corpus' phonetic content on voice cloning, experiments focus on the speaker encoder, as illustrated figure 1b, to determine if the produced speaker embeddings are influenced. Several sets of samples are tested as input for the speaker encoder, with different duration and extreme phonetic content. If these experiments were to highlight an impact with a restricted field and extreme voices, then it would be legitimate to wonder about such a phenomenon for more usual phonetic contents with a complete voice cloning system. Yet if the speaker encoder turns out independent from the phonetic content, then it could be possible to directly offer voice cloning systems to patients with voice disorders. Otherwise, solutions such as sample preprocessing or system modifications could be considered to adapt existing voice cloning models to patients.

3.1. Considered Models

The considered approach, illustrated figure 1a, relies on two models, a speaker encoder model transmitting a speaker embedding to a multi-speaker TTS model. Their training can be separated in two phases, as in [2]. First, the speaker encoder is trained on an acoustic corpus composed of a high number of speakers. The second step - training a multi-speaker Text-to-Speech model - is not included in this study. The presented work only concerns speaker embeddings, with associated encoder. Voice cloning speaker encoders usually come from speaker classification or speaker verification tasks. Speaker classification aims to determine from which speaker a speech sample originates, within a fixed set of speakers. Speaker verification seeks to determine whether two given speech samples originate from the same speaker. X-vector model [5] is a frequently used speaker verification neural model. It takes as input variable-sized speech segments. It can be described with three blocks: frame-level layers, statistic pooling and segment-level layers. Extracted speaker embeddings, called x-vectors, correspond to one of the segment-level layers' embedding. The employed implementation is from the Kaldi ASR toolkit.

3.2. Extreme Phonetic Content Sampling

To determine the influence of phonetic content, studied samples are extracted from a female French voice (referred as Neb), containing 87 hours of speech, from SynPaFlex corpus [8]. They are constructed by randomly extracting *Sentences*, mono-syllabic words (*MSW*), phones (*Phn*), vowels (*Vowels*) or only "A" phones (*Phn-A*). Four sample durations - 1 hour, 10 minutes, 1 minute and 10 seconds - are considered to study the duration impact and to compare it with the phonetic content one. For each couple of strategy and duration, a hundred samples are used, to avoid possible margin effects. Using a very large voice is necessary to obtain several samples containing 1 hour of "A" phones only, for example.

3.3. X-Vector Analysis

Three experiments are lead with the trained x-vector model. The first one aims to ensure of the model quality, while the second and third ones are designed to determine whether sample duration and phonetic content have an impact on the produced x-vectors.

First, the goal is to assess the trained model capacity to generate a speaker-specific representation. A nearest neighbour-like classifier is implemented to classify x-vectors. Their centroid is computed as the mean of the x-vectors from the class samples. Tested x-vectors are labeled according to the nearest class centroid. A x-vector is considered as correctly classified if the labeled class is the same as the original sample class, *ie* if the correct speaker is attributed. To avoid a measure bias, when a x-vector is tested, it is temporarily

removed from its class centroid computation. A majority of correctly classified x-vectors would ensure of the model quality and reliability for further experiments, while a majority of errors could indicate training issues.

Second, the aim is to know if sample duration impacts produced x-vectors. Indeed, state of the art shows that it influences the output speech of a voice cloning system [4]. Differences between x-vectors of different duration classes could serve as a reference to study the impact of another parameter, here the phonetic content, on speaker similarity. Four sample duration are considered (*cf* 3.2). A classification method similar to the first experiment is used, with a difference: classes labels no longer correspond to speakers but to sample duration for the same speaker. Distribution of x-vectors to the centroid of their class, in terms of euclidean distances, are also studied, as well as distances between centroids.

Last, the aim is to vary the phonetic content, as described in section 3.2. The same classification method is used, and class labels correspond to content strategies. Distributions of x-vectors and distances between centroids are also compared with the duration variation measures. This should determine if the impact of sample extreme phonetic content strategies on produced x-vectors is sufficient to classify x-vectors of the same speaker by said strategies.

4. **Training and Data**

Speaker encoder training requires a high number of speakers. Yet in [2], it seems more resistant to noise than the TTS model. Thus, it can be trained with lesser quality signals. Even though the speaker encoder studied was not the x-vector model, we assume their conclusions to be extendable to other speaker verification encoders. This hypothesis serves as a basis to choose a training corpus for the x-vector model.

CommonVoice is an open-source multi-lingual corpus by Mozilla [9]. This community project allows volunteers to record speech samples via their own recording device. The corpus contains more than 12k hours of speech in around 70 languages. Due to the diversity of recording devices and background sound environments, sample quality is very variable.

Only the French part of the corpus is used here. It contains 682 hours of speech for 12,953 speakers. This is consistent with given the corpora used in state of the art voice cloning. Moreover, using the text-independent version of the x-vector model, transcriptions are not given to the model. For reproducibility, the train, dev and test default sets are used for this study. Even though their speaker distribution is not proportionate, no speaker appears in more than one set, guaranteeing that no test speaker was seen during training. The x-vector model is trained with the train set, containing 3605 speakers for a total of 428h.

5. **Results and Discussions**

Result analysis is in two steps: a quality check for the x-vector model, then the comparison of variations depending on duration and phonetic content strategies.

5.1. **X-Vector Quality**

For *CommonVoice*'s test set, the classifier accuracy is 0.98 (15,515/15,763 x-vectors correctly labeled). Almost every x-vectors is classified as its original speaker. However, two points can mitigate these results. First, there are few - 5 at most - samples per speaker, so centroids are based on a very low number of x-vectors. Second, as *CommonVoice* test and train set originate from the same corpus, there could be a corpus bias, even if they share no sample nor speaker. To alleviate these biases, the x-vector model is tested on another corpus: a part of *MuFaSa* corpus [10] with 9 female speakers and a higher number of samples per speaker. The obtained accuracy is then 0.94 (5,204/5,550), see details in table 1a. Its

	1	2	3	4	5	6	7	8	9
1	259	2	0	0	1	12	0	0	0
2	4	325	0	0	0	6	0	0	0
3	8	7	284	14	14	16	27	7	0
4	1	0	0	24	0	0	0	0	0
5	1	0	1	24	146	6	2	1	1
6	24	5	0	9	4	415	0	10	0
7	1	0	19	32	11	0	3,275	6	25
8	4	2	0	16	2	10	0	282	0
9	0	0	0	2	2	0	0	7	194

(a) Confusion Matrix of X-Vectors for *MuFaSa* corpus.

	10s	1m	10m	1h
10s	16	22	27	35
1m	51	29	16	4
10m	1	18	35	46
1h	27	38	16	19

(b) Confusion Matrix of X-Vectors from Neb Samples of Different Duration, for Strategy *Sentences*.

Table 1. Confusion Matrices of X-Vectors for Speaker (left) and Duration (right) Classifications. Lines Are Real Labels, Columns Predicted Labels.

accuracy is still very high, which allows us to dismiss the two biases. Thus, these measures confirm the quality of the x-vector model used.

5.2. **Phonetic Content Impact in Relation to Duration Impact**

	Class: Duration			Class: Strategy	
Strategy N°i	Mean Inter-Centroid Distance	Distance of X-Vectors to their Class Centroid		Distance of Centroid N°i to Centroid N°j	
Sentences N°1	0.09	1h	0.07	2	0.41
		10min	0.16	3	0.56
		1min	0.40	4	0.67
		10s	0.76	5	1.18
MSW N°2	0.04	1h	0.04	1	0.41
		10min	0.10	3	0.44
		1min	0.30	4	0.59
		10s	0.61	5	1.17
Phn N°3	0.04	1h	0.03	1	0.56
		10min	0.07	2	0.44
		1min	0.23	4	0.56
		10s	1.02	5	1.17
Vowels N°4	0.05	1h	0.04	1	0.67
		10min	0.09	2	0.59
		1min	0.27	3	0.56
		10s	0.58	5	0.80
Phn-A N°5	0.04	1h	0.02	1	1.18
		10min	0.06	2	1.17
		1min	0.20	3	1.02
		10s	0.47	4	0.80

Table 2. X-Vector Dispersion.

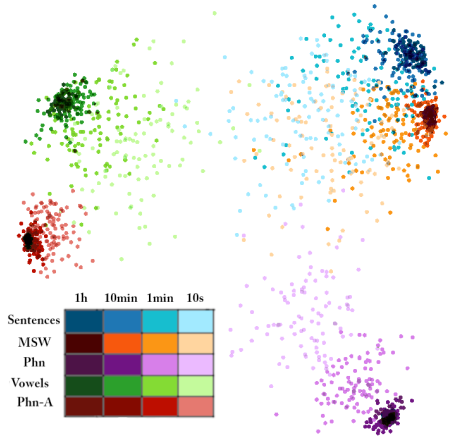


Figure 2. X-Vector PCA. Total Described Variance is 59.9%.

Duration classification within a phonetic content strategy has an accuracy of 0.41 for *Sentences*, 0.33 for *MSW*, 0.45 for *Phn*, 0.37 for *Vowels* and 0.46 for *Phn-A*. As an example, the confusion matrix for *Sentences* is available table 1b. For a given strategy, x-vectors are not easily separable by sample duration. On the contrary, phonetic content classification is perfect for all durations, except for 10s with an accuracy of 0.96. It shows that phonetic content strategies render perceptibly different x-vector classes.

Moreover, for a given strategy, euclidean distances presented in table 2 show that duration classes centroids are particularly close. Given classification results, duration classes could thus be be distributed with nearly concentric shapes, with different radii. Hence, more measures are required to conclude on the relative importance of phonetic content in regards to duration. To this end, x-vector mean distances to their centroid within duration classes are considered, *ie* x-vector dispersion according to duration, for a given content strategy. They are mostly inferior to the distances between strategy centroids, with the punctual

exception of the 10s class. Therefore, class separation and distance both are superior for phonetic content than for duration. This result is corroborated by the Primary Component Analysis (PCA) illustrated figure 2. This allows to conclude that studied extreme phonetic contents have a higher impact on x-vectors than duration for a target voice characterisation.

6. Conclusion and Future Work

With the improvement of voice cloning systems, it becomes conceivable to apply them to phonetically constrained voices, and more particularly to pathological voices. By reproducing the first steps of a state of the art voice cloning system, objective measures can determine the impact of phonetic content on speaker embeddings. Linked to its medical application, this study focuses on extreme content. Obtained results show an impact on the produced x-vectors. Observed variations are larger than for duration, which in state of the art influences speaker proximity of the produced speech. These results are an incentive for thorough studies, with automatic and perceptive tests, on the links between phonetic content and cloned voice quality. This complementary study is in progress, with available cloned samples. Last, as extreme contents seem to implicate acoustic variations, further studies could be led with actual samples from patients suffering from organic dysphonia, including tests with remediation strategies to overcome the impact of phonetic content.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011011870R1 made by GENCI.

References

[1] F. Le Huche and A. Allali. *La voix*. 2e édition. Collection Phoniatrie. Elsevier Masson, 2010.

[2] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Moreno, and Y. Wu. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis". In: *Proc. of the Neural Information Processing Systems Conf.* 32. 2018.

[3] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. "Neural Voice Cloning with a Few Samples". In: *Advances in Neural Information Processing Systems* (2018), pp. 10019–10029.

[4] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, C. Gulcehre, A. v. d. Oord, O. Vinyals, and N. de Freitas. "Sample Efficient Adaptive Text-to-Speech". en. In: *Proc. of the Int. Conf. on Learning Representations*. 2019.

[5] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. "Deep Neural Network Embeddings for Text-Independent Speaker Verification". In: *Proc. of Interspeech*. 2017.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". en. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2018.

[7] C. E. Steuer, M. El-Deiry, J. R. Parks, K. A. Higgins, and N. F. Saba. "An update on larynx cancer". In: *CA: A Cancer Journal for Clinicians* 67.1 (2017), pp. 31–50.

[8] A. Sini, D. Lolive, G. Vidal, M. Tahon, and E. Delais-Roussarie. "SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis". In: *Proc. of the 11th Int. Conf. on Language Resources and Evaluation (LREC)*. Miyazaki, Japan, 2018.

[9] Mozilla. *CommonVoice*. Consulted in December 2020. URL: commonvoice.mozilla.org.

[10] A. Sini. "Characterisation and generation of expressivity in function of speaking styles for audiobook synthesis". Theses. Université Rennes 1, 2020.