2022 International Conference on Advanced Aspects of Software Engineering (ICAASE) | 978-1-6654-9202-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAASE56196.2022.9931574

Analysis of the Effect of Audio Data Augmentation Techniques on Phone Digit Recognition For Algerian Arabic Dialect

Khaled Lounnas, Mohamed Lichouri
Computational Linguistics Department, CRSTDLA
Algiers-Algeria
{k.lounnas, m.lichouri}@crstdla.dz

Mourad Abbas
High Council of Arabic Language
Algiers-Algeria
m_abbas04@yahoo.fr

***Abstract*—In this study, we describe a solution for dealing with the problem of data scarcity in Speech Processing tasks involving low-resource languages, including Automatic Speech Recognition (ASR). This method is based on a set of Data Augmentation (DA) techniques that will be applied to the small corpus that was initially used. This corpus comprises the first 100 Arabic digits uttered by two native Algerians. We used a variety of DA techniques to increase the size of this corpus, including stretching the signal without changing the pitch, simulating an environment using white noise, and finally shifting the sound. Finally, a number of experiments were carried out on two alternative configurations to assess the influence of these strategies on ASR performance. Extensive tests are carried out to verify the impact of the augmented samples in the training set or the training and testing set. Experimental results show that data augmentation plays an important role in improving the accuracy of recognition models, in which the impacts of the data augmentation methods such as Noise, Time Stretch, and rotation are slightly obvious.**

***Index Terms*—ASR, Algerian Arabic Dialect, Audio Data Augmentation, Spoken Digit**

I. Introduction

Speech recognition is a critical task for developing smart assistive devices right now, especially as the number of communication in daily life rises. This high frequency is causing a significant shift in people's lifestyles, making the development of a dependable system to assist them a significant challenge. As a first solution, mobile phones have emerged as a trendy area offering attractive platforms for speech recognition-based functions that can help in solving various issues in mobile telephony [1], automated contact centers, and consumer electronics items.

Nowadays, many applications have been implemented as intelligent telephone answering systems [2] that use standard speech modems to respond to incoming calls and recognize the call recipient and caller's name. There is also, Interactive Voice Response (IVR) systems, which can be used for mobile purchases, banking payments, services, retail orders, travel information. To host these applications, several manufacturers currently offer mobile phones with built-in voice interfaces [3], [4]. Most of these interfaces are developed to support particular languages [5], for example, English, French, and Hindi [6], whereas some many other languages and dialects aren't fully supported, especially those with low resources such as Arabic dialects.

To address this issue, the fundamental contribution of this research is the development of a system that recognizes Arabic digits spoken in Algerian dialects. In this regard, we started with the Algerian Arabic (Algiers + Blida) subset of the corpus developed in our previous work [7]. Then, to solve the issue of the relatively tiny size, we applied a variety of DA methods like temporal stretch, noise, and audio boosting. This newly produced corpus will be utilized to train acoustic and linguistic models in a number of configurations in the CMUSphinx environment[1].

This paper is organized as follows: a literature review is presented in section II. In section III, a brief description of the proposed system has been given. Section IV is devoted to experiments and results. Finally, the conclusion is presented in section VI.

II. Related Work

Almost all the work done on Arabic Spoken Digit Recognition has been concentrated on the ten first digits as in [8] for Saudi dialect, [9] for Moroccan dialect and [10] for code-switching Algerian dialect. In contrast, in [7], the authors increased the level of complexity by recording a corpus of the first 100 Algerian spoken digits (Arabic and French) and attempting to evaluate the influence of feature size on the Word Error Rate (WER). Despite their system's high performance, they have run into a problem with the amount of the recorded dataset, which prevents them from having a proper sense of the system's quality. As a result, the approach was to increase the size of the dataset artificially or by recording fresh data, which is the first contribution to this study.

In [11], Ko et al. examined three DA methods (Text-to-speech DA, Cycle-Generative Adversarial Networks DA, and Pseudo-label augmentation) as an option to solve the missing data problem in distant-talk scenarios to enhance their end-to-end ASR-based systems. Two well-known corpora, CHiME-4 and CHiME-6 Challenge, were used in the experiment. The

[1]https://cmusphinx.github.io

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:42:51 UTC from IEEE Xplore. Restrictions apply.

authors claimed that their findings revealed that combining the described methods into a single implementation improved the system's performance. Furthermore, combining all of the aforementioned techniques has given positive results, particularly when applied to the CHiME-6 Challenge dataset.

While focusing on the labeling DA process and dealing with the data sparsity issue, the authors of [12] used two DA approaches based on Vocal Tract Length Perturbation (VTLP) and Stochastic Feature Mapping (SFM) because of their effect on increasing speaker and speech variations of the limited training data. Furthermore, a layered architecture was used to implement a combination protocol of the aforementioned techniques. Experiments on Assamese and Haitian Creole, two IARPA Babel development languages, revealed improved performance in automated speech recognition (ASR) and keyword search, according to the authors (KWS).

Based on past work and as our second contribution, we choose to apply three data augmentation approaches that can help us increase the size of the corpus without impacting the audibility and comprehensibility of the audio recordings.

III. Proposed System

We'd like in this part to go deeper over each component of our proposed system as follow.

A. Architecture

This section describes our experience using CMU Tools to build and develop an Algerian Arabic vernacular speech recognition system. The core components of a typical ASR system are depicted in Figure 1.

The primary purpose of our study is to develop an Algerian Arabic Darija automated speech recognition system that employs Mel frequency spectral coefficients in the feature extraction phase and GMM-HMM scheme combination approaches in the training phase. We built our Darija acoustic model using SphinxTrain tools based on the dictionary, language model, and voice data, and we utilized the Pocketsphinx decoder [14] in the recognition phase. Our adopted dictionary file was made up of the first hundred uttered digits, followed by their transcriptions. Table I shows a selection of corpora and their transcriptions.

B. Linguistic Material

Building a typical corpus is a required phase in the development of any engineering system, hence building a large corpus is critical to the success of any machine learning-based model. As a result, and in order to address the key issue raised in our earlier paper, we attempted to enlarge the initial corpus that was developed in the first place in [7].Three DA algorithms are used to extend the audio recording during the extension process. This will enable us to create a new corpus that is three times as large as the original. It will also allow us to test our system in a variety of situations that might alter the voice signal. The table below provides a quick summary of the new corpora's background information.

TABLE I
The First Hundred Digits, with their Syllabus and Transcription in English and Algerian ARABIC.

Digits	English Transcription	Algerian Dialect Transcription
00	CIFER CIFER	صفر صفر
01	CIFER WAHED	صفر واحد
02	CIFER ZOUJ	صفر زوج
03	CIFER TLATHA	صفر تلاتة
04	CIFER REBAA	صفر ربعة
05	CIFER KHEMSSA	صفر خمسة
06	CIFER SETTA	صفر ستة
07	CIFER SEBAA	صفر سبعة
08	CIFER THMENYA	صفر تمنية
09	CIFER TESAA	صفر تسعة
...	...	...
98	THMENYA OU TESAAINE	ثمنية وتسعين
99	TESAA OU TESAAINE	تسعة وتسعين

TABLE II
Details on the New corpus. AAD stands for Algerian Arabic Dialect, BAD: Blida Arabic Dialect, respectively.

Features	Value
Sampling rate	16 KHz
Number of bits	16 bits
Number of Channels	1, Mono
Audio data file format	.wav
# Speakers	2
#Speakers per dialect	1
# Dialect	2
# Language	1
# Tokens per speaker	1500
# speaker's gender	Male
# Data augmentation algorithms	4
# Total number of tokens	12000
#Number of digits	100 digits (AAD) 100 digits (BAD)
# Repetitions per word	15
Condition of noise	normal life
Preemphased	$1 - 0.97z^{-}1$
Window type Hamming	25.6 ms
Frames overlap	10 ms

C. Data Augmentation (DA)

Data augmentation is the process of generating fresh synthetic training samples from our initial training set by making minor adjustments. The objective is to make our model less vulnerable to these disturbances while also improving its generalizability. In our example, we'll add noise, stretching, and rolling [2]. We used the code from the Alibugra GitHub repository [3] to alter our audio files. In this code, we used three functions: the first will help us to add white noise, which is based on a normal distribution with a mean equal to 0, whereas the variance is equal to 1 and a length equal to the size of the

[2]https://www.kaggle.com/CVxTz/audio-data-augmentation
[3]https://github.com/alibugra/audio-data-augmentation/

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:42:51 UTC from IEEE Xplore. Restrictions apply.

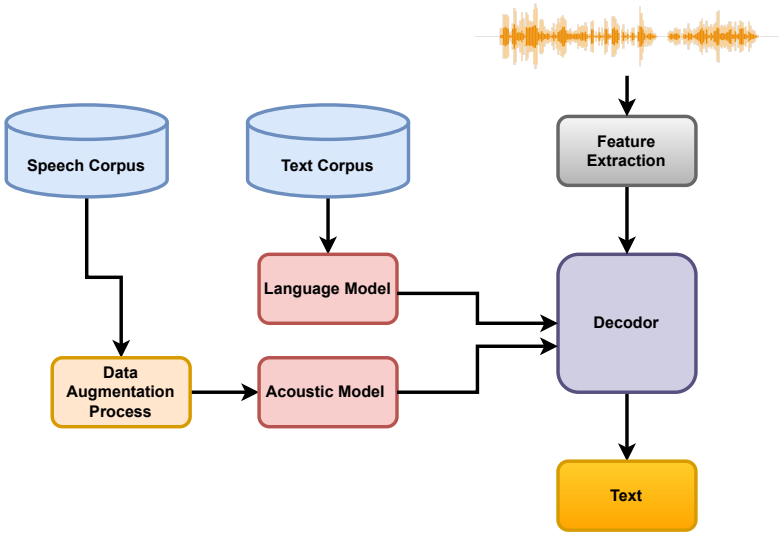

Fig. 1. Integration of Audio Data Augmentation process with a basic ASR System.

audio file[4]. The noise's amplitude was limited, so we could still hear the word despite the noise. We then call a method that rotates an audio file around a defined axis. The audio file's content gets moved as a result. A segment is returned to the first place after rolling from the first to the final position. Finally, raw data is subjected to a time stretching process [5], which involves altering the speed or length of an audio signal without modifying its pitch.

D. Feature Extraction

The feature extraction method turns the speech waveform into a sequence of feature vectors that only include the data required to identify a given utterance. It has a big influence on how well speech recognition systems work. One of the most often used feature vector extraction approaches is Mel-Frequency Cepstral Coefficients (MFCC), which are used to recreate the human ear.

E. Acoustic model

One of the most difficult components of the technology is improving the accuracy of the ASR system. The acoustic model is crucial in its enhancement. In order to improve accuracy, the acoustic model is essential. The acoustic model's primary goal is to compute the likelihood of observed feature vectors given linguistic units using a statistical method called the Hidden Markov Model (HMM) with a mixed density Gaussian distribution (phones, words, and subparts of phones). The Gaussian Mixture Model (GMM) is used to calculate the likelihood of a given feature vector $P(O|Q)$ for each HMM state **Q**, which corresponds to a **O** phone or sub-phone. For identifying limited amounts of words, using an HMM state to represent a phone is adequate. Three HMM states are the most frequent way to represent a phone; instead of one plus two non-emitting states on each ends, each phone contains three emitting HMM states.

F. Language model

The language model (LM) guides the search for proper word sequences in the speech recognition system. There are many various types of models that may be used to characterize a language during the recognition phase, such as grammar, phonetics, and statistical language models. In this paper, we develop our system's language model using the CMU-Cambridge statistical language modeling tools.

IV. Experiments and Results

We ran several tests to evaluate the performance of the proposed spoken digit recognition system, for the Algerian Arabic dialects (BAD, AAD). It is worth noting that we used CMU-sphinx in the system design [6]. According to our previous work in [7], the best size of the acoustic features (MFCCs) is 13 coefficients, and the best number of GMMs is 32 to recognize Algerian dialects. This is why we will continue with this setup for the ongoing experience.

To recognize the Algerian Arabic dialects, we will conduct multiple experiments of ASR on the clean dataset as well as the transformed one (Noise, Roll, and TimeStretch) as follows:

- In Experiment 1, we will conduct a comparison of the impact of clean or transformed dataset on Spoken Digits Recognition WER.
- In Experiment 2, we will compare the impact of the fusion of the clean and one of the transformed datasets on Spoken Digits Recognition WER.
- In Experiment 3, we will see how merging our clean dataset together with two of the modified datasets affects the WER of spoken digit recognition.

[4]https://numpy.org/doc/stable/reference/random/

[5]https://librosa.org/doc/main/generated/librosa.effects.time stretch.html

[6]https://cmusphinx.github.io/

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:42:51 UTC from IEEE Xplore. Restrictions apply.

- In Experiment 4 , we will investigate the effect of grouping the clean dataset with all of the transformed datasets mentioned above on the WER of spoken digit recognition.

All the 4 experiments will be conducted in two setups [15]:

1) Applying Data Augmentation on both train and test set.
2) Applying Data Augmentation on train set only.

A. Experiment 1: Clean vs Transformed

The reported results presented in table III will help us to infer that for Arabic, when the data set is not altered in any way (clean environment), our system will perform at its best, reaching a WER of 7.4% in the case of the Algerian dialect. Otherwise, the system's dropped dramatically by 4.1%, 6.5%, 10.3%. This permits us to note that there's a lot of emphasis on the listed modifications.

TABLE III
THE BEST PERFORMANCE OF THE ASR SYSTEM FOR ALGERIAN ARABIC (AAD AND BAD): EXPERIMENT1

Language	Corpus	WER (%)
Arabic	Clean (Cl)	7.36
	Noise (Ns)	11.49
	Roll (Rl)	13.89
	Time Stretch (Ts)	17.71

B. Experiment 2

The performance of the speech recognition system in front of DA is shown in table IV, where we opted to merge DA methods two by two with the raw data to highlight what impact each method has on the system. We contemplated testing our system twice: once with raw data (setup 2) and the other with raw data along with one of the selected transformations (setup 1).

It should be noted from the finding presented in the table above that when we test our system using DA, the error gradually increases with regard to the raw data, and the system that involves merging raw data with noisy data is the least expensive in terms of error. The reason is that the information is preserved in the case of the application of noise compared to other methods that modify the content of the audio signal.

TABLE IV
THE BEST PERFORMANCE OF THE ASR SYSTEM FOR ALGERIAN ARABIC (AAD AND BAD): EXPERIMENT2

Language	Training Corpus	Testing Corpus	WER (%)
Arabic	Cl + Ns	Clean	9.64
		Cl + Ns	11.85
	Cl + Rl	Clean	9.92
		Cl + Rl	14.8
	Cl + Ts	Clean	8.46
		Cl + Ts	17.18

C. Experiment 3

Table V compares different multi-DA-based corpora in which the clean environment coexists with two other DA methods. It is obviously observed that adding time stretch to the remaining DA techniques affects the recognition rate of our system. However, the worst configuration to avoid is the fusion of clean, roll, and time stretch since it affects the system performance gradually with an error of 16.62 percent.

The best finding was achieved at 14.77% when removing Time Stretch from the corpus configuration. Secondly, it should be noted that the best performance was obtained when learning our system with the combination of "clean," "noise," and "time stretch," resulting in a WER of 9.52% compared to 15.15% (an improvement of 5.63 %) when testing with the DA methods, reflecting our system's good behavior in raw conditions.

TABLE V
THE BEST PERFORMANCE OF THE ASR SYSTEM FOR ALGERIAN ARABIC (AAD AND BAD): EXPERIMENT3

Language	Training Corpus	Testing Corpus	WER (%)
Arabic	Cl + Ns + Rl	Clean	9.92
		Cl + Ns + Rl	14.77
	Cl + Ns + St	Clean	9.52
		Cl + Ns + St	15.15
	Cl + Rl + St	Clean	10.43
		Cl + Rl + St	16.62

D. Experiment 4

Table VI depicts the construction of a global system in which all of the discussed transformation are applied and merged to form a single system. According to the results, there is a decrease in WER due to the increased data size. As a result, there is no need to use this transformation to augment the size of the data because it has a negative impact on system performance.

TABLE VI
THE BEST PERFORMANCE OF THE ASR SYSTEM FOR ALGERIAN ARABIC (AAD AND BAD): EXPERIMENT4

Language	Training Corpus	Testing Corpus	WER (%)
Arabic	Cl + Ns + Rl + St	Clean	10.35
		Cl + Ns + Rl + St	16.76

Overall, we found that using DA on the train set is more important and rational than on the test set. To put it another way, if we want to expand the corpus, we must extend the training set only, not the test, otherwise it will be incoherent.

V. DISCUSSION

Based on all of the aforementioned analyses, we can reveal that nothing beats the raw environment for our system's best performance, outperforming the other DA techniques. For second rank, we have white noise, which has a lower impact on our system than the other DA, it can create a favorable mix with the raw environment indicated by the lowest recognition rate. It should be mentioned that joining some DA sets (like St and Rl) is not recommended due to the deformation of the data supplied by the latter, resulting in a significant error rate.

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:42:51 UTC from IEEE Xplore. Restrictions apply.

The findings reported are indeed vulnerable to different limitations. Despite that we have shown the robustness of our system experimentally, we have overlooked the process of searching for ideal parameters because we constructed our system based on the core parameters of our prior work. It is vital to subject our system to the same expert circumstances in order to make it more generalist. Another feature that helps us to improve our model is dialectal and data augmentation variability, which means that the more dialects and data augmentation techniques our system learns, the more universal it becomes and the better it generalizes.

VI. Conclusion

We sought to re-validate the findings of a prior work that dealt with digit identification in the Algerian Arabic dialect but was not properly examined due to the tiny corpus at the time. As a result, the goal of this work was to use a range of DA methods to enhance the corpus size. We've started applying noise, roll, and temporal stretch-based approaches to increase our initial corpus.

Following that, a method based on assessing our system's behavior was presented in order to perform a comparative study with the clean one in order to acquire a basic grasp of the most reliable design in a variety of DA-related scenarios.

The obtained findings demonstrated the relevance of Noise transformation over Time Stretching and Roll transformations for data enhancement and overall correctness. We've also mentioned that DA approaches should be used only on the training set, not the test set. More testing is required to determine the ideal transformation settings, such as the range of noise levels that the Noise transformation accepts.

References

[1] Varga, I., Aalburg, S., Andrassy, B., Astrov, S., Bauer, J. G., Beaugeant, C., ... & Hoge, H. (2002). ASR in mobile phones-an industrial approach. IEEE transactions on speech and audio processing, 10(8), 562-569..

[2] Lobanov, B. M., Brickle, S. V., Kubashin, A. V., & Levkovskaja, T. V. (1997). An intelligent telephone answering system using speech recognition. In Fifth European Conference on Speech Communication and Technology.

[3] Wu, S. L., Kingsbury, B., Morgan, N., & Greenberg, S. (1998, December). Performance improvements through combining phone-and syllable-scale information in automatic speech recognition. In ICSLP (Vol. 1, pp. 160-163).

[4] Tabani, H., Arnau, J. M., Tubella, J., & González, A. (2017). Performance analysis and optimization of automatic speech recognition. IEEE Transactions on Multi-Scale Computing Systems, 4(4), 847-860.

[5] Salimbajevs, A. (2018, May). Creating Lithuanian and Latvian speech corpora from inaccurately annotated web data. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[6] Deka, B., Chakraborty, J., Dey, A., Nath, S., Sarmah, P., Nirmala, S. R., & Vijaya, S. (2018, May). Speech corpora of under resourced languages of north-east india. In 2018 Oriental COCOSDA-International Conference on Speech Database and Assessments (pp. 72-77). IEEE.

[7] Lounnas, K., Abbas, M., & Lichouri, M. (2021). Towards Phone Number Recognition For Code Switched Algerian Dialect. In Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021) (pp. 290-294).

[8] Alotaibi, Y. A., Alghamdi, M., & Alotaiby, F. (2008). Using a telephony Saudi accented Arabic corpus in automatic recognition of spoken Arabic digits. In Proceedings of 4th International Symposium on Image/Video Communications over Fixed and Mobile Networks (pp. 43-60).

[9] Satori, H., Hiyassat, H., Haiti, M., & Chenfour, N. (2009). Investigation Arabic Speech Recognition Using CMU Sphinx System. International Arab Journal of Information Technology (IAJIT), 6(2).

[10] Lounnas, K., Satori, H., Hamidi, M., Teffahi, H., Abbas, M., & Lichouri, M. (2020, April). CLIASR: a combined automatic speech recognition and language identification system. In 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) (pp. 1-5). IEEE.

[11] Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In Sixteenth annual conference of the international speech communication association.

[12] Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(9), 1469-1477.

[13] Yılmaz, E., Heuvel, H. V. D., & van Leeuwen, D. A. (2018). Acoustic and textual data augmentation for improved asr of code-switching speech. arXiv preprint arXiv:1807.10945.

[14] Ezzine, A., Satori, H., Hamidi, M., & Satori, K. (2020, June). Moroccan Dialect Speech Recognition System Based on CMU SphinxTools. In 2020 International Conference on Intelligent Systems and Computer Vision (ISCV) (pp. 1-5). IEEE.

[15] Song, C., Xu, W., Wang, Z., Yu, S., Zeng, P., & Ju, Z. (2020). Analysis on the impact of data augmentation on target recognition for UAV-based transmission line inspection. Complexity, 2020.

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:42:51 UTC from IEEE Xplore. Restrictions apply.