Towards to a Direct Speech to Speech for Endangered Languages in Africa

Diane Carole Tala Metalom, Jean Louis Fendji Kedieng Ebongue, Blaise Omer Yenke

Towards to a Direct Speech to Speech for Endangered Languages in Africa

Diane Carole TALA METALOM[1], Jean Louis FENDJI KEDIENG EBONGUE[*2], Blaise Omer Yenke[2]

[1]Department of Mathematics and Computer Science, Faculty of Science, University of Ngaoundere, Cameroon
[2]Department of Computer Engineering, University Institute of Technology, University of Ngaoundere, Cameroon

*E-mail : lfendji@gmail.com

Abstract

Direct speech-to-speech translation has recently been introduced with the ability to perform a textless translation of a speech in a source language to a corresponding speech in the target language. This ability allows the development of automatic speech recognition systems for under-resourced languages and even non-written languages, since transcription is no longer necessary. Because of the number of such languages in the world and in Africa in particular, the design of such systems is essential to reduce the extinction rate of these languages. Unfortunately, current work claiming to design direct speech systems for unwritten languages still conducts experiments using well-resourced languages such as French, English and Spanish. In this work, we consider the particular case of sub-Saharan African languages, specifically Fulfulde, and identify the challenges faced in developing a direct speech-to-speech system for this language. We find that a main difficulty lies in the construction of a dataset. First, syntactic and morphological differences between French and Fulfulde have been identified. A method for constructing datasets for unwritten languages, usable in direct speech-to-speech translation models, is proposed. This construction method takes into account the difficulties and peculiarities of unwritten languages. Beyond the dataset, an approach to design a corresponding direct speech-to-speech translation system is proposed.

Keywords

Direct Speech to Speech; Endangered Language ; Speech Translation

I INTRODUCTION

Direct speech to speech translation (DS2ST) has been the subject of several research projects in recent years. Kano et al in [16] claimed that the traditional cascading method of translating a source language voice to a target language voice, requiring going through automatic speech recognition and then speech synthesis, increases the error rate, due to the error in each step. This error would be minimized by going through direct translation. As a result, some researchers developed models based on sequence to sequence[6, 11], while others worked on the construction of datasets for direct speech to speech translation [13, 4, 5, 2]. Some others worked on research on the construction of direct speech to speech translation systems [17, 18]. The ultimate goal is the translation of speech from a source language into a corresponding speech in a target language. But so far, the work done in this area focuses on written languages such as Chinese-English [13], German-English [2], Spanish-English [6, 13, 11, 8]. A direct speech to speech

model has been developed for unwritten languages [12], but the authors carried out the experiments using again well-resourced languages. Therefore, there is a need to design datasets for low-resourced or unwritten languages that can be used for direct speech to speech to evaluate the performance of this approach on low-resourced or real unwritten languages.These datasets could also be used to make some applications voice-based and interactive, allowing illiterate populations working in a field such as agriculture [19] to improve their performance. The design of such datasets requires a suitable methodology. In the second section of this paper, we present a linguistic context to show the fragility of unwritten languages and the danger of a complete extinction. In the third section we briefly discuss automatic speech recognition mechanisms and the required resources. Then we show how direct speech to speech translation as been applied on "unwritten languages" by quoting authors who have developed models for these languages. We expose the problem of constructing datasets for unwritten languages in section five, before giving a dataset construction method for unwritten languages. Finally, we propose architectures for the construction of DS2ST models. The first architecture focuses on the case where the source language is written and the target language is unwritten, while the second architecture tackles the case where the source and target languages are both unwritten.

II LINGUISTIC CONTEXT

Language is the first means of communication of a population. It is also a crucial element of the identity of people. One of the best ways to promote or preserve a language is to write related documents such as dictionaries or books. Thanks to globalization and innovative methods of information and communication technologies, the development of languages could be effective. This development involves: the use of documents designed for these languages; and/or the design and deployment of applications that can help illiterate populations to master this language in their activities. The world counts 7139 languages according to The Ethnologist[1]. Africa in particular has a linguistic richness representing the cultural, socio-cultural heritage and identity of many ethnic groups. Indeed, 30% of the world's languages, i.e. 2144[2], are spoken in Africa by at least 887 million people. Among the world's languages, 42.27%[3] are endangered, and about 308 African languages are "highly threatened" to disappear [1] (14.36% of all existing languages on the continent). More severe, at least 201 African languages are known to have become extinct [7], not to mention the many otherless threatened but nevertheless vulnerable languages. This shows that the situation of languages disappearing in sub-Saharan Africa is more serious than we perceive. These languages that fade over time are mostly absent in cyberspace because they are unwritten or under-resourced. This disparity is also explained by military conquests with genocides, numerical weakness, socio-economic domination and many others. The disappearance of a language leads to the disappearance of part of the heritage of humanity in the sense that a language embodies a way of conveying knowledge. To reduce the linguistic divide, i.e. the possibility of the disappearance of certain African languages, one approach is to build automatic speech recognition systems for these languages.

III GENERAL INFORMATION ON AUTOMATIC SPEECH RECOGNITION SYSTEMS

Automatic speech recognition (ASR) is a technique for analyzing the human voice captured by means of a microphone to transcribe it into a text [3]. Speech synthesis is the transcription of

[1]https://www.ethnologue.com/guides/how-many-languages
[2]https://www.ethnologue.com/guides/continents-most-indigenous-languages
[3]https://www.ethnologue.com/guides/how-many-languages-endangered

a text into audio. Speech recognition and speech synthesis are speech processing techniques. These techniques make it possible to create human-machine voice interfaces where part of the interaction is done by voice. The construction of a speech recognition system involves modeling, a step that requires several resources. The diagram in Figure 1 shows the set of resources required to build an ASR system and the role of each resource in that system.

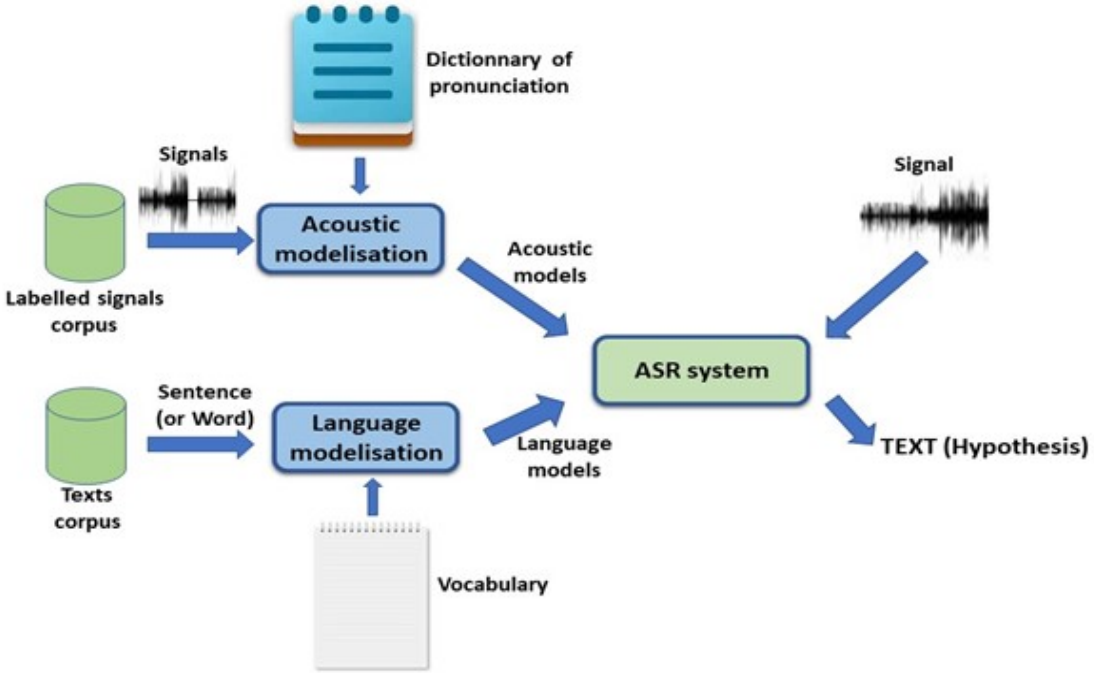

Figure 1: Construction of ASR system and required resources.

The word spoken is compared with those of the audio signal corpus, the comparison consists in subtracting the shades of gray of the pixels of the spoken word from those of the words of the corpus and repeating this for each row and column. We will be able to find the most similar signal. The voice signal is compared to the words in the pronunciation dictionary through acoustic modeling where the acoustic model gives for each sequence of words, the possible pronunciation or pronunciations with their probabilities. The recognition algorithm makes it possible to choose the most similar word, calculating the rate of similarity between the word pronounced and the various references. The language model gives the probability of each sequence of words in the target language. The construction of an ASR requires linguistic resources such as good quality pronunciation dictionaries, textual and audio corpora, as well as vocabularies [24]. However, languages that tend to disappear are unwritten and poorly endowed, which is a limit in the construction of ASRs for these languages. Recent works report on a new method of recognition: direct speech to speech recognition.

IV DIRECT SPEECH TO SPEECH RECOGNITION

4.1 Background

Direct speech-to-speech translation (S2ST) translates a word from a given language into a word from another language without relying on the generation of intermediate texts [8]. Several works have been done on the design of direct speech to speech translation models.

The main elements of a DS2ST model are:

- **Network LSTM (Long Short Term Memory)**: It is a recurrent neural network (RNN), which can learn long-term dependencies. It is based on the fact that long-term knowledge recall is a natural behavior of people. Many factors such as the dataset, the complexity of the model and the learning time taken by the LSTM network influence the translation process [20].
- **Encoder**: It is a component of the Encoder-Decoder architecture, it takes as input a source utterance of variable length and returns a vector of fixed length. The most common encoders used by researchers are the Variational AutoEncoder (VAE) [17, 12], to transform source speech into discrete elements, forming a sequence-to-sequence model. In papers [11, 8, 6], authors use either a 5-layer [11], or a 8-layer [6, 8] LSTM encoder to map the input spectrogram features into a fixed number of channels. The first layers represent the source content and the deep layers learn to encode the target content information.
- **Decoder**: It predicts the output sequence using the vector produced by the encoder. The LSTM decoder is used in several DS2ST works [11, 8, 6]. The 4 or 6 layer LSTM decoders give good performances [11, 20].
- **Vocoder**: This component converts the spectrogram into a waveform file that can be understood by a human. Authors in [21] describe the HiFi-GAN neural vocoder based on a generative adversarial network framework. HifiGAN is a neural vocoder based on a generative adversarial network framework. A discriminator composed of sub-discriminators is used by the model during training. Each sub-discriminator performs processing on a specific periodic part of a raw waveform. This vocoder is modified [22] to convert the units into a waveform. [23] has set up UnivNet, a neural vocoder that synthesizes high-fidelity waveforms in real time.

4.2 Existing DS2ST models

In 2019, Jia et al [6] developed a Direct speech to speech translation model and performed tests for Spanish-English translation. This model is not as efficient as the ASR (Automatic Speech Recognition) or MT (Machine Translation) or TTS translation (Text To Speech translation) models. But the authors specify that it thus opens up a new line of research. Sequence to sequence models of voice conversion were the basis for the realization of this model, the goal being to recreate a statement in the voice of another person.

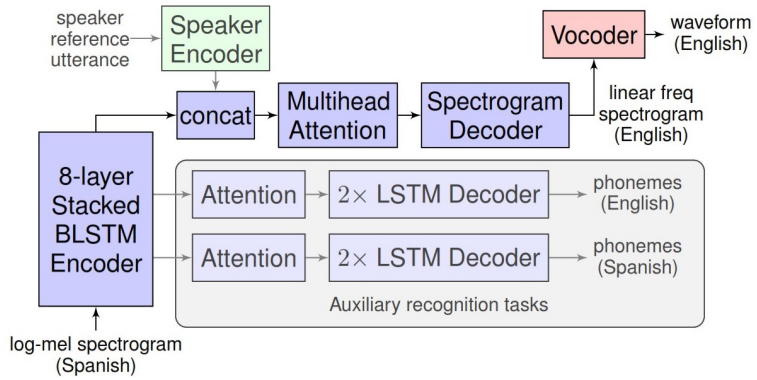

Figure 2: Model Architecture implemented in Google's Translatotron [6].

Figure 2, shows the architecture of the Direct S2ST model for translatotron described in [6]. The model aims to translate Spanish into English, it takes as input a Spanish spectogram obtained after Fast Fourrier extraction. The model provides as output an English voice (waveform). The Bi-Directional Long Short Term Memory (BLSTM) encoder receives the input file and returns

8 layers which are concatenated and the result is passed through a multihead Attention. The result of the latter is sent back to the spectogram decoder which will calculate a linear frequency spectrogram output. The Vocoder receives this output and converts it into a waveform that we can hear. To have a waveform output with a voice similar to the source file, and to maintain the voice characteristics, a complex trained network also called encoder is used. The block made up of LSTM decoders allows to perform the auxiliary tasks added to obtain better results.

Chen Zhang et al. [12] described a UWSpeech direct speech to speech translation model for unwritten languages. Figure 3 describes the drive and inference pipeline of this model. UWSpeech uses the method named XL-VAE (Cross-Lingual Variational AutoEncoder) based on the improvement of the VQ-VAE method (Vector-Quantized Variational AutoEncoder) so that it is able to recognize speech in several languages.

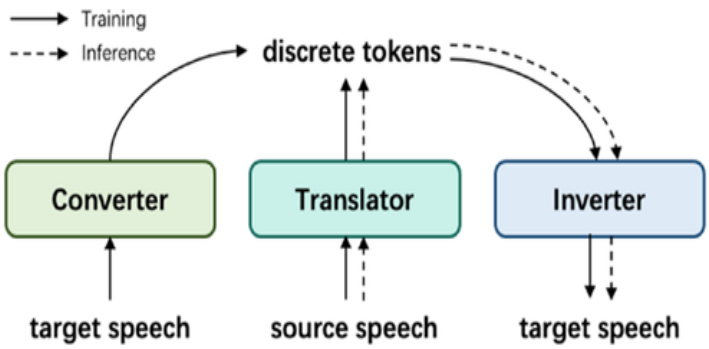

Figure 3: The training and inference pipeline of UWSpeech [12].

UWSpeech consists of a converter that transforms unwritten target speech into discrete tokens. The translator in this model translates the speech of the source language into discrete tokens of the target language. Finally, the inverter converts the discrete tokens translated into unwritten target speech. For good accuracy, XLVAE improves the discretization (transformation of speech into a discrete token, using the converter) and reconstruction (synthesis of discrete token speech, using the inverter) capabilities of the VQ-VAE method. The tests in UWSpeech are done with written languages,namely Spanish and English. This work does not really integrate unwritten languages. Consequently, it would be interesting to show how to design datasets of unwritten languages for the construction of direct speech to speech models in order to build models and perform tests on unwritten languages.

Table 1: Summary of works on Direct Speech to Speech Translation.

Papers	Methodology	Experimentals data	Results	Advantages	Disadvantages
[16]	Training of an attention-based encoder-decoder component. Use of transcoders to connect the resulting components. One transcoder that transfers attentional context information from hidden acoustic representations to hidden linguistic representations of the target language, and a second that do the reverse work. Using a source language speech encoder and three decoders that predict source and target language text transcripts.	Basic Travel Expression Corpus (BTEC), authors choose English-to-Spanish, Japanese-to-Korean, English-to-Japanese and Japanese-to-English.	Test done for specific languages shows that BLEU and METEOR score of transcoder are better than Google Transformer, Google RNN, Cascade RNN and cascade transformer.	Tests were done for the translation of several languages and by using two metric scores we can see that the model is better.	There is still a need to go through transcripts.
[12]	Conversion and translation of the source language into discrete units of the target language using the translator, then synthesis of the target speech from the discrete units in an inventer.	Fisher Spanish-English dataset.	UWSpeech achieves 17.33 BLEU points about 10 points higher than VQ-VAE.	Good performance.	Tests are still done using written languages.
[6]	The model consists of the attention, a vocoder, a decoder and an encoder stack. The attention-based sequence generates the spectrograms, the vocoder converts the target spectrograms into waves. The encoder stack maps the input features of the 80-channel log-mel spectrogram into hidden states that pass through an attention-based alignment mechanism to condition a decoder that predicts 1025-dimensional log-mel spectrogram frames corresponding to the translated speech.	Large scale conversational corpus of parallel text and read speech pairs from [25], and the Spanish Fisher corpus of telephone conversations and corresponding English translation.	The use of the Griffin-Lim and WaveRNN (to evaluate naturalness) vocoder on the conversational corpus gives respectively a MOS (Mean Opinion Score) of 3.20 and 4.08, and respectively 3.05 and 3.69 for the Fisher corpus.	The transcribed model focuses on the naturalness of the speech.	The cascade model has a higher BLEU score than this system.

Papers	Methodology	Experimentals data	Results	Avantages	Disadvantages
[8]	Predicting self-supervised discrete representations of target speech in the case where source and/or target transcripts may not be available. For written languages, the model presents a framework that jointly generates speech and text by combining the S2ST and S2T tasks via a shared decoder and a partially shared decoder. For the case of unwritten target languages, the authors extend the use of discrete units to text-to-speech translation. With multitask learning using both discrete representations for source and target speech. Authors show that it is possible to train a DS2ST system without using transcripts.	Fisher Spanish-English corpus, dataset composed of 139000 sentences.	In the case where the source and target languages are written, the BLEU score of this model is 37.2 against 39.5 for a cascade model. When the source language is written and the target language is unwritten, the BLEU score is 33.8 versus 41.0 for the cascade model. For unwritten source and target languages, the BLEU score is 27.1 versus 9.4 for UWSpeech and 0.6 for translatotron.	This model deals with the cases where the source and target languages are written, the cases where the source and target languages are unwritten and the case where the source language is written and the target language is unwritten.	Test done with a corpus of written languages considering that the target language is unwritten.
[26]	Connection by a single attention module of a source speech encoder, a target phoneme decoder and a target mel-spectrogram synthesizer. Trained of model with speech-to-speech translation and speech-to-phoneme translation objective.	Two spanish-english datasets and a multilingual-english dataset.	For translation of spanish to english, BLEU score gives 57.3 and MOS gives 3.24. Translatotron 2 achieved translation quality comparable to the baseline ST model, translatotron 2 is also highly effective for multilingual S2ST.	This model predicted speech naturalness.	All languages used for the tests are well-ressourced.

V DIRECT SPEECH TO SPEECH FOR THE LANGUAGES OF SUB-SAHARAN AFRICA

5.1 Case of the Fulfulde language

Fulfulde is one of the most common African languages in sub-Saharan Africa. There are about fifteen countries where the language is spoken, most often by a very strong national community. Despite the fact that Fulfulde is an under-rejuvenated language, researchers in language processing have observed and noted differences with the French language on the phonetic, morphological, phonological and prosodic levels.

5.1.1 Phonetics, phonology and transcription

The phonological system of Fulfulde has 28 consonant phonemes b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, z, y,ɓ,ɗ, γ, η, nb, nd, ng, nj, ny and 5 vowels a, i, o, e and u [9]. Generally, the letters a, b, d, f, i, k, l, m, n, o, p, r, t, w are written and pronounced in the same way as in French [10]. Fulfulde and French have eight letters that are written the same way, but are not always pronounced in the same way. These are: c, e, g, h, j, s, u, y. The letters ɓ,ɗ,γ and η are specific to fulfulde. Double letters are pronounced hard when it comes to consonants and long for vowels. In Fulfulde, the demonstrative adjective precedes the common noun it determines and agrees in nominal class with it. Its phonetic form may vary depending on the situation of the reference.

5.1.2 Morphology

Fulfulde has specific morphological rules different from those of French and English. For instance, the proper names of people, animals and places do not bear specific morphological marks. They are often used in the sentence without class marker (determinant). Simple common nouns are morphologically composed of a radical and a class suffix, called a nominal class index or marker. This suffix marker is used to form the determinant of the name. It also indicates, by its nature, whether the noun is singular or plural, whether it expresses the dimensional class of the diminutive or augmentative. The demonstrative adjective precedes the common noun it determines and agrees in nominal class with it.

5.1.3 Prosody

In Fulfulde, a sound usually corresponds to a letter. Any letter that is written shall be pronounced separately and distinctly. There are no fixed rules for the transition from a written language such as French or English to Fulfulde, which makes translation and the construction of datasets difficult.

Like Fulfulde, most of the languages of Sub-Saharan Africa have specificities that do not allow them to be easily understood, hence the particularity of the dataset construction program for these languages .

5.2 Constructing datasets of unwritten languages for direct speech to speech

The construction of the dataset for speech to speech translation consists of recordings of pairs of audio segments, the first audio corresponding to the source audio (here a written language) and the second audio is the target audio (unwritten language) [13]. There are two ways to create datasets: speech translation where data is generated based on complete audio recordings or their transcripts and simultaneous translation where data is transcribed in real time by human interpretation. For the construction of datasets using speech translation, we need the source

audio and their transcriptions into text. Simultaneous translation would be the best option for building direct speech to speech datasets of unwritten languages (source and target), If the source language is a written language, one can combine the two methods. The steps that we can remember for the construction of the datasets of direct speech to speech translation are as follows. It is noted here that the source language is written and the target language is unwritten. Figure 4 gives ataset construction steps for direct speech to speech for unwritten languages.

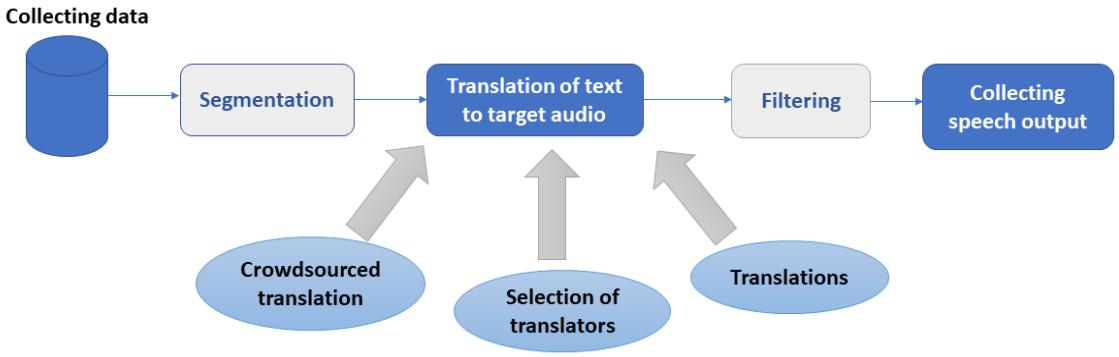

Figure 4: dataset construction steps for direct speech to speech for unwritten languages .

1. Search and download data: The first step is to download audio or text data of source language, dictionaries, or HTML files that contain manual transcriptions of the audio data. During this step, audio recordings are produced for the target language and also for the source language when the latter is also unwritten. When the source language is written, steps 2 and 3 are performed.
2. Segmentation: Cutting transcribed texts into sentences and utterances.
3. Text to target audio:This step consist to obtain translations. The texts in the source language from the segmentation phase are transcribed into audio of the target language. This transcription is done by experienced people who master the source and target languages. To achieve this goal, we must go through [14]
   - Crowdsourced translations: collection of crowd-sourced translations of the collected data.
   - Selection of preferred translators: Select successful translators based on the criteria described by Zaidan in [15] For this step, a specialist in the target language will be called in to listen to the various translators' results and then select the best translators.
   - Complete translations: The selected translators will translate the collected data .
4. Filtering: The next step is to create a YAML file containing the recording time of sentences, words and utterances. During this phase, we eliminate noisy segments, long silences, words that do not match.
5. Collecting speech output: After obtaining the collection of translations, a split of the data must be made into training, development and test sets. This strategy will allow to verify the efficiency of the currently developed model.

5.3 Construction of the DS2ST model for weakly resourced languages

The construction of the model can be done according to whether the source language is written or not.

5.3.1 Case where the source language is written

The source language being written, it is possible to get transcripts. Automatic Speech Recognition (ASR) and Text to Speech Translation (T2ST) will be used. The ASR is used for the recognition of the source speech, while the T2ST is used for the translation of the text from the written language to the unwritten target speech. For T2ST, the model will be described so that as input, we have the source text of variable length which passes through an LSTM encoder as developed in [11]. However, an encoder of 6 layers is considered to improve the performance. The encoder transforms the text taken as input by a source state of fixed form. The state passes through a decoder to produce a spectogram of the speech in the target language. Finally the spectogram passes through the vocoder to produce the target speech. Figure 5 shows the architecture of this model.

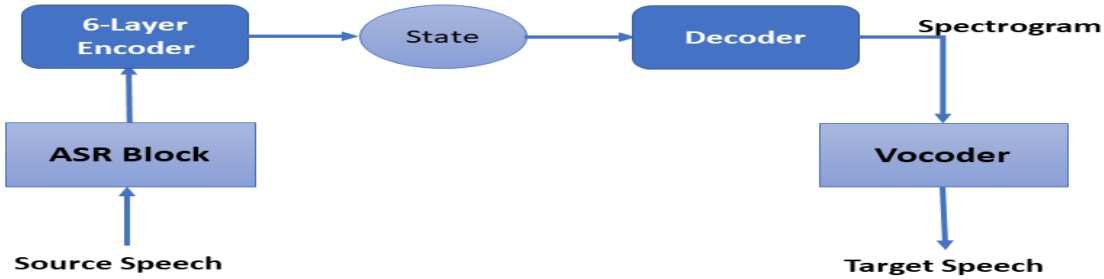

Figure 5: Architecture of model for translation Written-Unwritten language.

5.3.2 Case where the source language is unwritten

In the case where the source and target speech are unwritten, there will be no intermediate texts. For this, we will use a VQ-VAE encoder that takes as input the spectogram of the source speech. This encoder will generate a discrete representation given the log-mel filter input obtained by the source speech passed through the feature extraction. We will then train a sequence to sequence (Seq2Seq) model with two 6-layer LSTM decoders. The first decoder focus on the recognition of the source speech into its corresponding discrete tokens. This decoder is part of the auxiliary task as proposed in [8]. Adding auxiliary tasks to a translation model regularizes the encoder and helps the model learn attention that significantly improve performance. The other decoder for the translation of the source speech into discrete tokens of the target language. The architecture for this model is given in figure 6

VI CONCLUSION AND OUTLOOK

6.1 Conclusion

This work aimed to shed light of on the possibility of using Direct Speech to Speech to help reducing the rate of extinction of endangered languages, particularly in Africa. For this aim, there is an urgent need to develop datasets for unwritten languages and to test the models developed for these languages, since authors who developed direct speech to speech model for unwritten languages only ran tests using written languages. Although the paper focused on Fulfulde, it paved the way for other similar languages.

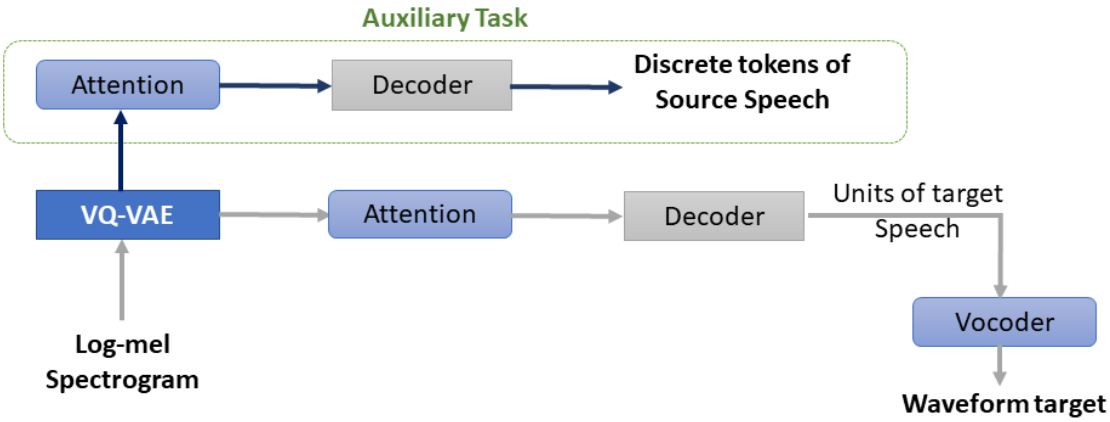

Figure 6: Architecture of model for translation Unwritten-Unwritten language .

6.2 Outlook

In future works, a dataset of French words for the source language and Fulfulde words for the target language, to perform tests on the Direct Speech to Speech system for unwritten languages developed in [12]. Additionally, a DS2ST system based on the discrete unit for translation from French source text to Fulfulde speech will be built. Finally relevant particularities of Direct Speech to Speech translation for unwritten language will be exposed.

REFERENCES

[1] Batibo, Herman M. 2005. "Language Decline and Death in Africa: Causes, Consequences and Challenges". Language Decline and Death in Africa. Multilingual Matters. https://doi.org/10.21832/9781853598104.

[2] Beilharz, Benjamin, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. "LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition." ArXiv:1910.07924 [Cs], March. http://arxiv.org/abs/1910.07924.

[3] Besacier, Laurent. 2014. "Automatic Speech Recognition for Under-Resourced Languages: A Survey." Speech Communication, 16.

[4] Cattoni, Roldano, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. "MuST-C: A Multilingual Corpus for End-to-End Speech Translation." Computer Speech and Language 66 (March): 101155. https://doi.org/10.1016/j.csl.2020.101155.

[5] Godard, P., G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, et al. 2018. "A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments". ArXiv:1710.03501 [Cs], February. http://arxiv.org/abs/1710.03501.

[6] Jia, Ye, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model." In Interspeech 2019, 1123-27. ISCA. https://doi.org/10.21437/Interspeech.2019-1951.

[7] Kandybowicz, Jason, and Harold Torrence, eds. 2017. Africa's Endangered Languages: Documentary and Theoretical Approaches. New York, NY: Oxford University Press.

[8] Lee, Ann, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, et al. 2021. "Direct Speech-to-Speech Translation with Discrete Units". ArXiv:2107.05604 [Cs, Eess], July. http://arxiv.org/abs/2107.05604.

[9] Tourneux, Henry, and Yaya Daïrou. 1998. Peuhl Dictionary of Agriculture and Nature (Diamaré, Cameroon): followed by a French-Fulfulde index. Dictionaries and languages. Paris: Wageningen: Montpellier: Editions Karthala; CTA; CIRAD Editions.

[10] Tourneux, Henry, and Giuseppe Parietti. 2018. Fulfulde-French/French-Fulfulde Dictionary (Fulani Dialect[of] Diamaré, Cameroon) Illustrations by Christian Seignobos. Mimep-Docete. https://halshs.archives-ouvertes.fr/halshs-02103976.

[11] Weiss, Ron J., Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. "Sequence-to-Sequence Models Can Directly Translate Foreign Speech." ArXiv:1703.08581 [Cs, Stat], June. http://arxiv.org/abs/1703.08581.

[12] Zhang, Chen, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2020. "UWSpeech: Speech to Speech Translation for Unwritten Languages." ArXiv:2006.07926 [Cs, Eess], December. http://arxiv.org/abs/2006.07926.

[13] Zhang, Ruiqing, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. "BSTC: A Large-Scale Chinese-English Speech Translation Dataset." ArXiv:2104.03575 [Cs], April. http://arxiv.org/abs/2104.03575.

[14] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. "Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus." In Proceedings of the 10th International Workshop on Spoken Language Translation: Papers, Heidelberg, Germany.

[15] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in Proceedings of ACL, 2011.

[16] T. Kano, S. Sakti and S. Nakamura, "Transformer-Based Direct Speech-To-Speech Translation with Transcoder," 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 958-965, doi: 10.1109/SLT48900.2021.9383496

[17] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. "Speech-to-speech translation between untranscribed unknown languages". In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 593–600. IEEE

[18] Lee, Ann, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, et Wei-Ning Hsu. 2021. « Textless Speech-to-Speech Translation on Real Data ». ArXiv:2112.08352 [Cs, Eess], décembre. http://arxiv.org/abs/2112.08352.

[19] Fountsop, Arnauld N., Jean L. Ebongue Kedieng Fendji, and Marcellin Atemkeng. 2020. "Deep Learning Models Compression for Agricultural Plants". Applied Sciences 10, no. 19: 6866. https://doi.org/10.3390/app10196866

[20] Recep Sinan and Necaattin. 2019. "Development of Output Correction Methodology for Long Short Term Memory-Based Speech Recognition". Sustainability 2019, 11, 4250; doi:10.3390/su11154250

[21] Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan,1997. “JANUS-III: Speech-to-speech translation in multiple languages,” in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, vol. 1, pp. 99–102.

[22] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, 2020. “HiFiGAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” Advances in Neural Information Processing Systems, vol. 33.

[23] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, Juntae Kim. 2021. "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation". ArXiv:2106.07889v1 [eess.AS] 15 Jun 2021

[24] Fendji, J. L. K., Tala, D. M., Yenke, B. O., Atemkeng, M. 2021. "Automatic Speech Recognition using limited vocabulary: A survey". arXiv preprint arXiv:2108.10254.

[25] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, Yonghui Wu. 2019. "Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation". arXiv:1811.02050 [eess.AS] 2022-06-03 04:33:20

[26] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, Roi Pomerantz. 2022 . "Translatotron 2: Robust direct speech-to-speech translation". https://openreview.net/forum?id=HTfUrAxjPkR

A BIOGRAPHY

Diane C. M. Tala was born in Bandjoun, Cameroon in 1987. She received her B.S. degree in fundamental Computer Science from University of Dschang, Cameroon in 2011 and the M.S. degree in Systems and Software in Distributed Environments from the University of Ngaoundere in 2015. She is currently pursuing a Ph.D. degree in Computer Sciences at the Faculty of Sciences of the University of Ngaoundere in Cameroon. Since 2018, she is a Computer Science Teacher for secondary school. Her research interests include Speech Recognition, Natural Language Processing, Network Cost Modeling, Simulation, and Neural Networks.

Jean Louis K. E. Fendji (Member, IEEE) was born in Douala, Cameroon in 1986. He received the B.S. and M.S. degrees in Computer Science from the University of Ngaoundere, Cameroon, in 2010 and the Ph.D. degree from the University of Bremen, Germany, in 2015. From 2011 to 2013, he was a Research Assistant with the BMBF Project between the University of Bremen and the University of Ngaoundere. Since 2015, he

has been a Senior Lecturer with the Computer Engineering Department, University Technology Institute, At the University of Ngaoundere. He is the author of two book's chapters and more than 20 publications. He has been working for more than 10 years on Network modeling and optimization. His current research interests include Machine Learning, Deep Learning, NLP, Speech recognition, Optimization, AI in Agriculture and Education.

Blaise O. Yenke (Member, IEEE) received the Ph.D. degree in international joint supervision from the University of Yaounde 1, Cameroon, and the University of Grenoble, France, in 2010. He is an Associate Professor and a Researcher in computer engineering. He is also the Head of the Department of Computer Engineering, University Institute of Technology of Ngaoundere, Cameroon. His current research interests include distributed systems, high-performance computing, fault tolerance, network modeling, simulation, sensor networks design, and sensor's architecture.