Experiments on Speech Synthesis for Teochew, Can Taiwanese Help ?

Pierre Magistry, Ilaine Wang, Ty Eng Lim
ERTIM (Inalco), WhatTCSay
2, rue de Lille, 75007 Paris
{pierre.magistry, ilaine.wang}@inalco.fr, tlimgaginang@gmail.com

Abstract

This paper reports on our preliminary experiments in speech processing for Teochew, an under-resourced Sinitic language spoken both in China and around the world in diasporan communities. Following the recent uptick of interest in Teochew from heritage speakers of the diaspora and in order to respond to the needs of this community, we develop a Teochew Text-to-Speech system. We describe experiments to build this system and to assess the possible contribution of available resources in Taiwanese Hokkien, the closest language with a significant body of resources. The results of these experiments are not as conclusive as we expected: the Taiwanese dataset did not help our model significantly, but considering our objectives, we find it encouraging that they show that a large training dataset was not necessary for this precise task. A promising model could still be obtained with only a small dataset of Teochew. We hope that this work inspires other communities of speakers of languages in a revitalization phase.

Keywords: speech synthesis, language revitalization, Southern Min languages, Teochew, Chaozhouhua, Taiwanese, Hokkien

1. Introduction

Teochew is an under-resourced Sinitic language spoken both in China and around the world through its diaspora. In this paper, we present our preliminary experiments in speech processing for Teochew, focusing on Text-to-Speech (TTS)[1]. Our motivation is twofold.

The first goal is practical: to build a TTS system good enough to help complete the audio recordings of a Teochew/English/French dictionary app. The current version of the dictionary contains more than 6,700 entries, but only about 65% of those entries have an associated audio. Moreover, as for any dictionary of a living language, the number of entries is continuously growing, and this growth is even faster for the dictionary of a language that is still documented by its community of speakers.[2]

Secondly, we want to assess the possible contribution of Taiwanese Hokkien datasets in building a Teochew language processing application. This possibility relies on the fact that these two languages are closely related languages grouped into the Southern Min family, with Taiwanese being the most resourced language of this subgroup, as described below.

The feasibility of developing TTS tools and their relevance for language revitalization is well motivated in (Pine et al., 2022). This work is a first step in this direction for Teochew.

1.1. Teochew as a Sinitic Language

Sinitic languages is a vast group of languages both in terms of number of languages and number of speakers. Altogether, they form a major branch of the *Sino-Tibetan* language family. This group is then subdivided into dozens of branches. Going further in the divisions of language subgroups, we reach the Southern Min languages which includes both Teochew and Taiwanese Hokkien. Although they are two distinct languages with only partial mutual intelligibility, these two languages present striking similarities at every level of linguistic analysis - which is not the case when compared to Mandarin. As this work focuses on word-/phrase-level TTS, we mostly rely on phonetic similarities. Differences in romanization conventions may give a false impression of differences that only require some conversions to obtain homogeneous data. In contrast, the fact that both languages can be written with the Chinese script can also be helpful. We give more details on this issue in Section 4.

1.2. Teochew as a Heritage Language

It is noteworthy that although Hokkien and Teochew are closely related, their sociolinguistic contexts are very different. On the one hand, Taiwanese Hokkien tends to stabilize towards a standard today, thanks to the joint effort of the Taiwanese people and their government, while on the other hand, Teochew has evolved into multiple varieties among which: the variety from its original region of China,

[1] code and configuration files for our experiments are available on this git repository: https://gitlab.huma-num.fr/ertim/teochew-tts

[2] The next release of the dictionary is already planned to have more than 9,000 entries.

LREC-COLING 2024, pages 6849–6854
20-25 May, 2024.

and the variety from the international diaspora for whom Teochew is a *heritage language*. The latter can be subdivided into several other varieties, depending on both the hometown of the diaspora, the history of the family and the host country.[3]

This work has to take into account the specificity of such a situation: Teochew is spoken by different communities of heritage speakers around the world, and inevitably carries phonetic interference as well as loanwords from the local dominant language.

We also observe that while Teochew heritage speakers work towards the *revitalization* of their language, Taiwanese is at the stage of *protection* and *institutionalization* of the language.

2. Datasets

The starting point of this study is the availability of the recordings made for *What Teochew Say* (WhatTCSay), a dictionary app for mobile phones. The app is currently available for Android and iOS and is collaboratively co-developed by heritage speakers in North America and France. It was originally crowdfunded in 2012.[4] This dataset was shared with us under a CC-BY-NC 4.0 License.

As a result of the commitment of the Taiwanese people and with the support of the Taiwanese government, corpora and tools have been built for Taiwanese in recent years, making this language a medium-resourced language. More importantly, some of these resources are readily available thanks to the open-source policy supported by the community. These resources compiled for a related language are likely to be helpful to our work. Among those, we should mention the early Sui-siann dataset (Ì-Thuân, 2018) and the large *Taiwanese Across Taiwan* corpora (Liao et al., 2020). However, since we are aiming at producing audio for a dictionary, in this study we chose to work with the data shared by the Ministry of Education with the 教育部臺灣閩南語常用詞辭典, i.e., MOE's *Dictionary of Frequently-Used Taiwan Southern Min* (Taiwan Ministry of Education, 2012 – 2023), more likely to correspond to our first application.

2.1. WhatTCSay Recordings

Our main training data in Teochew is a set of recordings made by the founder of WhatTCSay (WTCS) and consists of 4388 mp3 files used for training, and 19 kept for testing, totaling 80 hours of speech and 9146 syllables. As expected for recordings aimed at illustrating the entries of an dictionary, most of this data are single words pronounced out of context. However, the dictionary also contains a phrasebook section with helpful sentences divided into themes, and some of these sentences were also recorded.

2.2. MOE Recordings

The *Dictionary of Frequently-Used Taiwan Southern Min* was released in 2008 and updated since then. Its content is available under a Creative Common License since 2012.

Beside lexicographic data, audio recordings for both word entries and example sentences are also provided, for a total of 20,744 wav files summing up to more than 500 hours and 32474 syllables.

3. TTS models and Toolkits

With the advances in neural speech processing, numerous TTS models have been proposed in recent years and it is now relatively easy to run experiments with those models on our dataset thanks to the availability of different toolkits. We sought for a state-of-the-art model available in the toolkits with two properties that are crucial for our study: *multi-speaker*, in order to be able to cope with our heterogeneous data, and *end-to-end*, to process romanized data simply (see below). We chose to use VITS (Kim et al., 2021), and rely on the CoquiTTS toolkit for the implementation (Eren and Team, 2023). Unless stated otherwise, we kept most parameters at their default value (as of version 0.17.5).

It is worth mentioning that recently Meta released a large set of TTS models for 1,100 languages, also using VITS (Pratap et al., 2023). Among those, one model is labeled with the iso code `nan` but it is unclear which language it actually corresponds to. The *nan* code actually covers the whole Southern Min family without distinguishing between languages. Digging into the original training data, we could only find misleading or erroneous metadata, with inconsistent mentions of Swatow, Taiwan and Penang, three locations where different Southern Min languages are spoken (respectively Teochew, Taiwanese and Penang Hokkien). We conclude that the data corresponds to some variety of Hokkien (for which better training data could be found, as we mentioned in Section 2). This makes Meta's model irrelevant for our goal which is to synthetise Teochew.

[3] Large communities of speakers were established first in South East-Asia (Thailand, Cambodia, Laos, Vietnam, Malaysia, Singapore, Indonesia) and in Western countries (the United States, France, and Australia) after a second wave of migration (Live, 1995; Tan, 2020).

[4] https://www.theteochewstore.org/blogs/latest/123903619-whattcsay-teochew-language-learning-app-now-available-for-free-the-story-behind

4. Phonologies and Romanizations

Teochew and Hokkien romanizations can all be considered shallow orthographies. The reliable grapheme to phoneme correspondence with today's end-to-end neural architectures enables us to avoid an additional step of phonetic conversion: we can go directly from the romanized script to the speech signal.

However, there is not a unique romanization system for all the Southern Min languages but multiple competing systems for every single language. This creates very misleading mismatches likely to confuse our system. To be able to benefit from a combination of different language resources from Teochew and Taiwanese, we created a mapping between the romanizations used in our dataset.

A short overview of the different writing systems is provided in Table 1. Created by Western missionaries, the two POJ are often called "Church Romanization". Guangdong Peng'im is promoted by the Guangdong province government and shows some similarities with the Hanyu Pinyin for Mandarin. Our work focuses on the last two romanizations as they are used in our dataset (respectively MOE and WTCS): (1) Tâi-lô (TLPA) promoted by Taiwan (ROC)'s Ministry of Education and clearly inspired by POJ, and (2) the Gaginang Peng'im (GGN), used in the Teochew diaspora, and derived from Guangdong Peng'im, with some modifications to make it easier to type in with different keyboards.

As a result, for the experiment in this paper, we need to provide a mapping between GGN and TLPA. The mapping is summarized in Table 2.

Most of the non-matching cases in romanizations are cases where a similar sound is spelled differently (bh and b for /b/). These are easy to convert but there are more complex cases where a sound is present in one language and not the other (for example, the tones and the coda -n in Taiwanese). In such cases, we have two options: either finding the closest match (what we did with the tones) or looking at more systematic correspondences based on cognates and possible reading of sinograms (Chinese characters). For the -n coda in Taiwanese, we noticed that sinograms read with a -n coda in Taiwanese tend to have a -ng coda in Teochew.

We implemented our conversion rules by modifying ParseTC[5] to add TLPA as input and get GGN as output.

Another important issue is *tone sandhi*. When used in context, tones of syllable are likely to change (regularly) and differ from so-called *lexical tone* (the tone that would be used when the syllable is pronounced isolated. The WhatTCSay resource writes the tones as they are realized *after* applying sandhi rules, while the MOE resource writes the lexical tones, *before* applying sandhi rules. As a result, before applying the tone conversions presented in Table 2, we had to apply tone sandhi on the Taiwanese transcripts. We assume sandhi applies everywhere but the last syllable. this is sound for short items like dictionary entries, but it may lead to a few errors in longer phrases containing sandhi boundaries which are difficult to guess.

[5] https://github.com/learn-teochew/parsetc

System	language	date
Peh-oe-ji (POJ)	Hokkien	1820
Peh-ue-ji (POJ)	Teochew	1875
Guangdong Peng'im	Teochew	1960
Tâi-lô (TLPA)	Hokkien (TW)	2006
Gaginang Peng'im (GGN)	Teochew	2012

Table 1: Different romanization systems with their date of introduction.

initials		nuclei		coda	
GGN	TLPA	GGN	TLPA	GGN	TLPA
bh	**b**	ai	ai	p	p
p	**ph**	**ao**	**au**	k	k
b	**p**	ia	ia	h	h
m	m	**iao**	**iau**	t	t
ng	ng	ieu	ieu	ng	ng
n	n	iou	iou	m	m
gh	**g**	oi	oi	**ng**	**n**
k	**kh**	ou	ou		
g	**k**	a	a	tones	
d	**t**	**eu**	**o**	1	7
t	**th**	e	e	2	6=2
j	**ts**	i	i	3	3
ch	**tsh**	**o**	**oo**	4	4
s	s	u	u	5	1
h	h	ng	ng	6	5
y	**j**	m	m	7	7
l	l			8	8

Table 2: Romanization mapping between GGN and TLPA. Non-matching graphemes are put in bold letters. This table does not include the nasalization noted -n in GGN and -nn in TLPA and a mapping between the tones.

5. Experiments

To conduct an evaluation as close as possible to our first application (the synthesis of dictionary entries), we relied on the local community of speakers. We were involved in a cultural event co-organized by the non-profit organization *Les Jeunes Teochew de France*. Teochew speakers attending the event were invited to judge our system's outputs to enable us to compute a Mean Opinion Score.

In total, 20 respondents contributed 380 scores.

5.1. Design

The samples used for the evaluation were selected from the dictionary according to the following principles: (1) the subset must contain all of the phonemes of Teochew, (2) each entry has to be common or easily understandable by most heritage speakers, and (3) the length of entries from the subset must be somewhat representative of the whole dataset. The resulting subset therefore comprises 19 entries, ranging from 2 to 9 syllables. These entries were removed from the training data.

Evaluators are provided with the entry definition in French to ensure that they expect a specific word or phrase. They are invited to give a score to three different sound files for the same word/phrase. One of them (`m3`) is the original recording by a human, while the two others are the outputs of two different models: `m1`, trained only on the WTCS dataset, and `m2`, trained on a combination of WTCS (Teochew) and MOE (Hokkien) recordings. We used Label-Studio (Tkachenko et al., 2020-2022) to collect the scores.

In Experiment 1, we evaluate the models which are considered the best ones by the training algorithm (using the default settings for VITS). In particular, the number of iterations on the training data is automatically chosen. However, after noticing that the actual best model may not be the one selected, we build a small validation set of phrases to allow a speaker of Teochew to evaluate the model at different stages of the training process and select the best *checkpoint* according to her own preferences. We conduct a second experiment using those manually selected models.

5.2. Results and Discussion

Mean Opinion Scores (MOS) are reported in Table 3. The first experiment shows a significant improvement when augmenting the Teochew data with Taiwanese. However, this effect disappears in the second setting. At this stage, it is difficult to draw any definitive conclusion on the transfer between the two languages. It may be the case that TTS benefits more from homogeneous (single speaker) data, or that the synthesis of simple phrases does not require too much data. Our script conversion and transfer strategy are likely to be more relevant for other tasks such as speech-to-text.

By looking at per-item MOS displayed on Figure 1 with the peng'im transcription used to generate the speech and listening to the corresponding audio files, and discussing with the participants, we can make the following observations: the model trained on Taiwanese data tends to be worse on shorter words, for which augmenting the training data may not be needed. The worst rated output sentence is "*gian7 heu2 go3 ain2 yioh4 gu2*" (*How long does it take to go there?*). Although the voice sounds natural, the audio file for this sentence contains clearly wrong phonemes (/b/ for /h/, and a low tone on /gu2/). On the other hand, the audio generated for "*dui2 m7 ki2*" (*sorry*) corresponds to the correct phonemes but the signal is unclear on the acoustic level. In both cases, better audio could be obtained by re-running the same model multiple times on the same phrases. This non-deterministic nature of the model is difficult to handle, but may lead to improvement if we can find a way to select a better output between multiple samples.

Experiment 1		
model	MOS	stdev
m1 (Teochew data)	2.60	1.17
m2 (mixed data)	2.90	1.16
m3 (human)	4.46	0.833
Experiment 2		
model	MOS	stdev
m1 (Teochew data)	3.66	1.03
m2 (mixed data)	3.37	1.01
m3 (human)	4.57	0.730

Table 3: Mean Opinion Score and standard deviations, on a scale of 1 to 5.

A single opinion score is also difficult to interpret. In some cases, regional accent or lexical preferences may interfere with the judgment tasks. For instance, we used the word "ja1 bhou6 peng7 iu2" (*girlfriend*), but the word "*neung6 peng7 iu2*" is actually more common, which confused some of our evaluators as they did not expect to hear the former and were even unsure of the correctness of the word for some of them.

6. Applications and Future Work

This work present preliminary experiments in building a TTS for Teochew. Our promising results correspond to a TTS already close to be useful for simple tasks such a isolated words generation to complete the recordings of the dictionary. However, our experiments did not provide a definitive answer about the contribution of Taiwanese data and call for more investigations, including a more fine-grained evaluation on specific phonetic features on which the two languages differ.

One of our next steps will also be to make the models available through a *discord* bot to encourage and assist heritage speakers in practicing writing in Teochew, with features such as peng'im dictation and quizzes. The bot can also provide convenient ways to obtain feedback from the community, which is very active on dedicated Discord servers. We are also aiming at supporting the diversity of Teochew accents with a multi-speaker model.

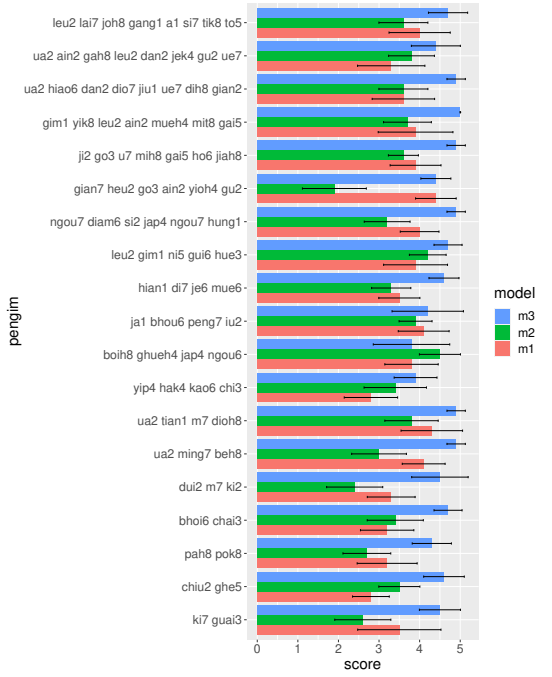

Figure 1: Mean Scores per item for each model on Experiment 2, with t-test Confidence intervals.

From a more technical point of view, we need to address the issue of selecting the “best" model without relying on the algorithm or on a single speaker. Crowdsourced voting on Discord or through the WTCS application could be good strategies.[6]

7. Ethics Statement

This research is purposely *community-in-the-loop*, our motivation comes primarily from the community of heritage speakers of Teochew, particularly active nowadays. We try to define our objective according to the community needs. WTCS is a popular app among them[7] and as stated before, the rapid growth in the number of entries for WTCS results from the fact that the community of speakers is still currently working to document Teochew. This makes the use of a TTS system crucial in order to match this speed. This coverage is literally vital for a community struggling with the promotion and the revitalization of its language: both in the diaspora and in China, the younger generations are less likely to speak or even understand Teochew. Without audio recordings, the dictionary is in fact hardly accessible for them, although they are the primary target.

[6] In the time between the submission of this paper and the conference, we managed to mitigate some of these issues by also training and ASR system to select the best sample from multiple outputs. We also obtained better results with two stages of training the model, one with all the data and a second one with only Teochew data to specialize the model. We want to thank the reviewers for the encouraging and insightful comments on our work.

[7] The developer reports more than 10k downloads in total.

Acknowledgments

This work is funded by the DiLSi ANR project ANR-23-CE38-0004-01. It was granted access to the HPC/AI resources of [CINES/IDRIS/TGCC] under the allocation 20XX-AD011014016R1 made by GENCI.

We are also grateful to the “Les Jeunes Teochew de France" association for welcoming and promoting our research during their events.

8. Bibliographical References

Gölge Eren and The Coqui TTS Team. 2023. Coqui TTS.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Yu-Sion Live. 1995. Les Chinois de Paris: groupes, quartiers et réseaux. In Antoine Marès and Pierre Milza, editors, *Le Paris des étrangers depuis 1945*, pages 343–357. Éditions de la Sorbonne.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359, Dublin, Ireland. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint*.

My Dung Adeline Tan. 2020. *L’expression du déplacement en chaozhou : les formes introduisant un groupe nominal locatif et l’encodage de la trajectoire*. Ph.D., Institut National des Langues et Civilisations Orientales, Paris.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling soft-

ware. Open source software available from https://github.com/heartexlabs/label-studio.

9. Language Resource References

Yuan-Fu Liao, Chia-Yu Chang, Hak-Khiam Tiun, Huang-Lan Su, Hui-Lu Khoo, Jane S. Tsay, Le-Kun Tan, Peter Kang, Tsun-guan Thiann, Un-Gian Iunn, Jyh-Her Yang, and Chih-Neng Liang. 2020. Formosa speech recognition challenge 2020 and taiwanese across taiwan corpus. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 65–70.

Taiwan Ministry of Education. 2012 – 2023. *MOE's Dictionary of Frequently- Used Taiwan Southern Min* 教育部臺灣閩南語常用詞辭典.

Ì-Thuân. 2018. *SuíSiann Dataset*.