Leveraging Speech LLMs for Audio-based Lexical Retrieval in Dictionaries: the Case of Audio Queries in WhatTCSay, a Dictionary app for the Teochew Language

Siman Chen, Ilaine Wang, Maxime Fily, Pierre Magistry

Leveraging Speech LLMs for Audio-based Lexical Retrieval in Dictionaries: the Case of Audio Queries in WhatTCSay, a Dictionary app for the Teochew Language

Siman Chen, Ilaine Wang, Maxime Fily & Pierre Magistry

ERTIM, Inalco

Institut national des langues et civilisations orientales

ERTIM Equipe de recherche textes, informatique, multilinguisme

Abstract

This study focuses on Query by Example - Spoken Term detection in low-resource scenarios. It reports on our experiments in building an audio-based query functionality for the diasporan Teochew dictionary WhatTCSay.
This functionality enables users to retrieve dictionary entries without prior knowledge of the writing systems in Teochew, thereby facilitating language revitalization efforts within Teochew communities.
To address the retrieval task, we explore:
(i) a supervised ASR-based approach using peng'im-to-peng'im matching and a sinogram-to-sinogram matching methods, and (ii) a non-supervised method based on Dynamic Time Warping (DTW) for audio-to-audio matching.
Retrieval performance is evaluated using recall at rank k.
Results show that peng'im-to-peng'im matching achieves the best performance, followed by audio-to-audio matching. In contrast, sinogram-to-sinogram matching performs the worst.

Context and Objectives

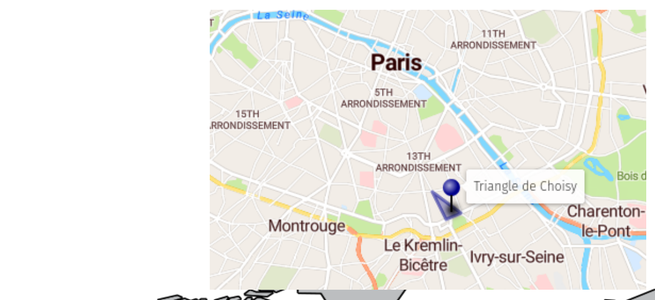
Teochew, an under-documented Sinitic language spoken both in South China and around the world in diasporan communities. Despite the renewed interest in Teochew from diasporan communities and the popularity of the dictionary app WhatTCSay, its use can be frustrating.
Heritage languages are typically acquired in the home environment with no formal instruction. Because of its dialectal status, Teochew is often devalued in comparison with mandarin. As a result, many heritage speakers do not write in their dialect, and some are unaware that it even has a written form.

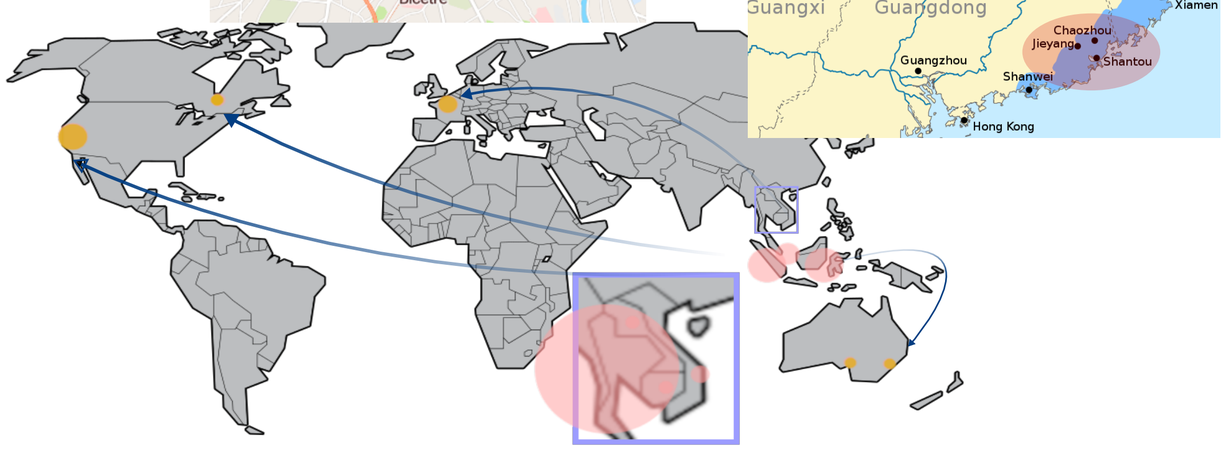
Our objectives are:
- to enable heritage Teochew speakers to search the WhatTCSay dictionary using audio queries without needing knowledge of Teochew writing systems,
- to improve the overall usability of the dictionary

Screenshot of WTCS3's interface in French when looking for Teochew dishes

What's Teochew?

潮州話 (*dio-jiu ue*)
- **low-resource** language, absent from large-scale speech datasets
- **multi-variety**, spoken in Guangdong, China and various regions in diaspora
- **multiple writing systems** : sinograms and several romanization systems (Guangdong peng'im (GDPI) vs. Gaginang peng'im (GGNPI).
- "**polyphony**" : many-to-many relations between sinograms and their pronunciations
- **tone sandhi** (8 tones)





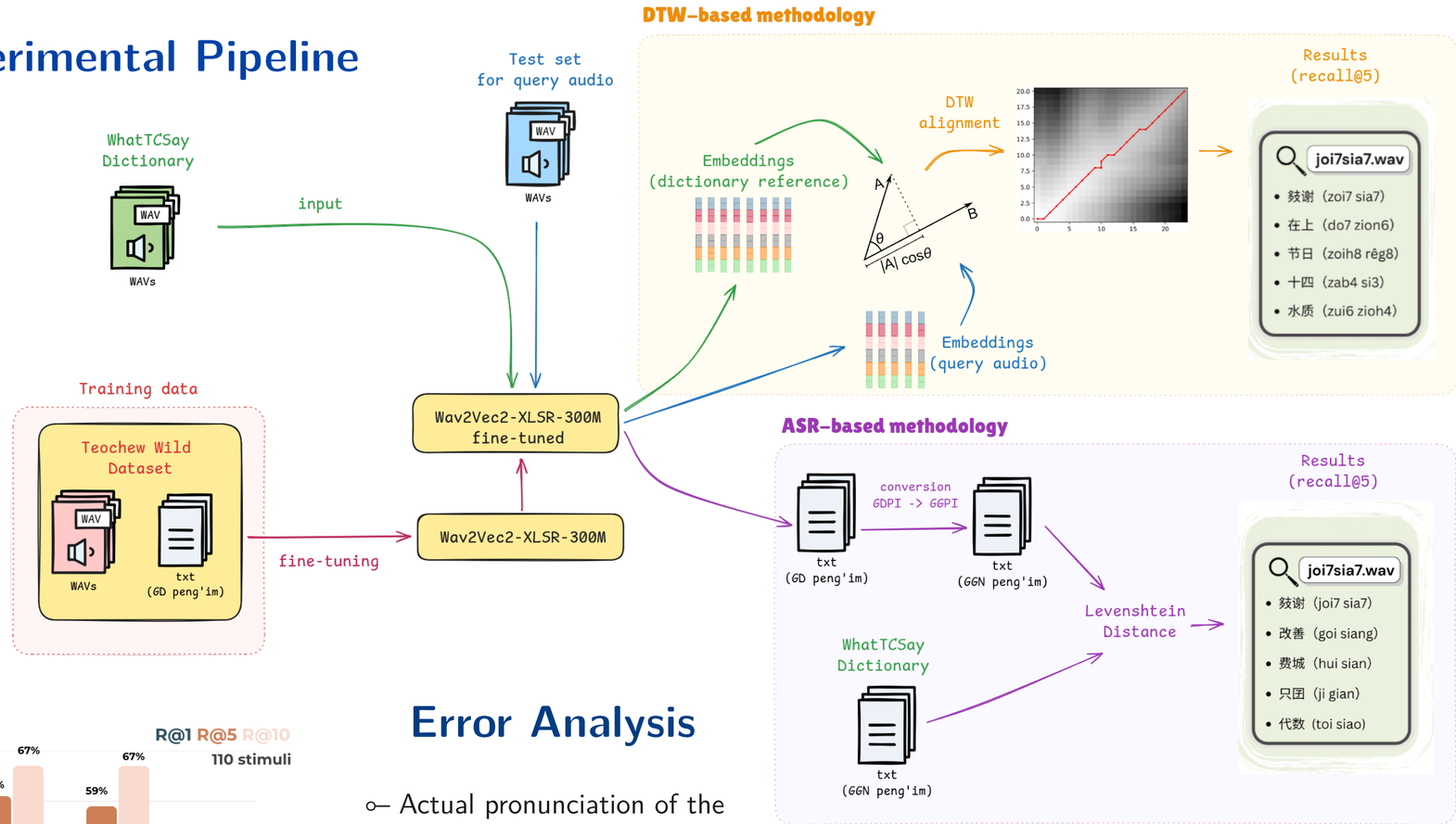
Data

1) Reference: original audio files (WTCS)
- Teochew from diaspora (USA)
- 4,600 entries, mostly isolated word (~1 s each)
- 1.28 hours of speech

2) Training data: Teochew Wild (Pan et al. 2025)
- Teochew from China,
- Phrase-level corpus: 12,500 utterances (~5.43 s per utterance)
- 18.9 hours of total speech

3) Test set for query audio
- 6 speakers, 1 Teochew from China, 5 diaspora
- 172 word-level stimuli
- Out of domain

Experimental Pipeline



Models

- Wav2Vec2 XLSR-300M fine-tuned on Peng'im (CTC loss: 0.111)
- Whisper-Medium: a publicly available fine-tuned Whisper model trained on sinograms

Model	Annotation Type	val	test
Whisper-Medium(ft)	Chinese Character	9.61	10.01
Wav2Vec2 XLSR (300M, ft)	Peng'im	13.51	12.52

Results

Peng'im-to-peng'im matching vs. sinogram-to-sinogram matching:
- Metric: Recall@K (R@1, R@5, R@10)
- Peng'im-to-peng'im outperforms sinogram-to-sinogram retrieval
- Without tone > with tone

Possible explanations:
- a large number of tonal minimal pairs (e.g., *leu5* "labeur" vs. *leu2* "toi")
- tonal sandhi (annotations in training corpus are in citation forms while tone sandhi is realized naturally in speech)
- non-standardized writing systems
- polyphony: many-to-many relations between sinograms and their readings

Error Analysis

- Actual pronunciation of the speaker exhibit variation than the canonical reference annotation in the dictionary.
- ASR model prefers the actual variety of pronunciation.
- The evaluation shows that the models can capture inter-dialectal variation in a non-standardized, multidialectal context.

Peng'im-to-peng'im matching vs. audio-to-audio matching:
ASR outperforms DTW (R@5: 68% vs. 55%)
- ASR: various transcriptions -> Levenshtein distance can recover the correct entry by partial overlap
- DTW: speaker-related variability sensitivity, misalignments can be sensitive to local acoustic distortions and temporal

stimuli (ghao5)	s1	rank	s2	rank	s3	rank	s4	rank	s5	rank	s6	rank
ASR	giao	-1	ghao	1	gho	3	bhao	4	ghoo	2	ghaog	1
DTW	giao6	-1	N/A	1	deu5	-1	gou3	-1	N/A	1	pang1	-1

Conclusion & Future Work

Conclusions
- **ASR-based retrieval**: offers interpretability and the possibility of downstream text-based applications, but requires a consistent transcription method across the pipeline.
- **DTW-based retrieval**: relies solely on speech representations. The method bypasses orthographic inconsistencies and spelling variation, but is more speaker-dependent.

Further investigation...
- for ASR: tone removal in ASR transcriptions before applying Levenshtein distance already increased recall by 4%.
- for DTW: denoising/silence removal improved rank for 8 out of 10 queries.
- retrain/fine-tune the model with additional speech data from the diaspora
- examine how ASR models encode tonal features found in Sinitic languages (eg. tone sandhi, contour tones).

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 2025-AD011014016R1 made by GENCI.