Leveraging Speech Models for Audio-based Lexical Retrieval in Dictionaries: the Case of the Teochew Language

Siman Chen, Ilaine Wang, Maxime Fily, Pierre Magistry
ERTIM, Inalco
Paris, France
{siman.chen, ilaine.wang, maxime.fily, pierre.magistry}@inalco.fr

Abstract

This study presents our attempt to apply Query by Example - Spoken Term Detection methodologies to a real-world, low-resource scenario: building an audio-based query functionality for the Teochew dictionary WhatTCSay. This functionality enables users to retrieve dictionary entries without prior knowledge of the writing systems in Teochew, thereby enhancing the accessibility of the dictionary and facilitating language revitalization efforts within Teochew communities. To address the retrieval task, we investigate two approaches: (i) an Automatic Speech Recognition (ASR)-based approach using text-to-text matching, and (ii) a Dynamic Time Warping (DTW)-based acoustic framework for audio-to-audio retrieval. In the first approach, we compare an automatic romanization of the spoken query against the gold romanization from the dictionary; in the second, we directly match the user's spoken query against audio recordings from the dictionary pronounced by a native speaker. Retrieval performance is evaluated using recall at rank k. Results show that text-to-text matching achieves better performance than audio-to-audio matching; however, the two approaches were not optimized under fully comparable conditions, as the ASR-based approach benefited from additional optimization, which was not equally available for the DTW method.

Keywords: Speech Models, Low-Resource Languages, Automatic Speech Recognition, Dynamic Time Warping, Teochew Language, Query-by-Example, Spoken Term Detection

1. Introduction

Dictionaries represent an invaluable linguistic resource for any language. Yet, looking up a word in a dictionary is far from a trivial task and actually requires a certain level of prior knowledge: reading and writing skills, but also familiarity with orthographic conventions, the ability to identify lemmas, and, in the case of paper dictionaries, knowledge of language-specific sorting rules. Such prerequisites are often taken for granted for languages with a well-established written tradition and institutional support, where formal school plays a central role. This assumption, however, does not hold for minority languages, and, as a matter of fact, for the vast majority of the world's languages.

Heritage languages are particularly affected in this regard, as they are typically spoken at home, transmitted orally across generations, and associated with a limited vocabulary and lower prestige.

In order to make dictionaries accessible to speakers who lack the aforementioned skills, we propose a speech-based search functionality. Our approach builds on Query-by-Example Spoken Term Detection and is applied to Teochew, a Sinitic language for which the diaspora community has already developed a dedicated dictionary.

2. Background

2.1. Teochew as a Multivariational Heritage Language

Teochew is a Sinitic language belonging to the Southern Min branch of the Sino-Tibetan family, spoken primarily in eastern Guangdong, China.

From the 18th to the 20th centuries, successive waves of migration have spread the Teochew people over the world, resulting first in large diaspora communities across Southeast Asian countries such as Thailand, Cambodia, Singapore and Malaysia, and then in Western countries, including the United States, France, and Australia (Live, 1995; McFarland, 2021).

Based on several accounts, Tan (2020) estimates the number of speakers to be between 5 and 7 million in Thailand, the largest Teochew community in Southeast Asia, and between 80,000 and 150,000 speakers in France. These figures should be treated with caution as they stem from outdated sources[1] and may not reflect the current situation, especially since population censuses do not target spoken languages but rather ethnicity, and having Teochew origins does not necessarily imply Teochew language proficiency.

[1] Estimates for Thailand were taken from sources dating from 2001 and 2004, while the lowest estimates for France date back to 1989. See Tan (2020) for more details on these questions.

Proceedings of SPEAKABLE @ LREC 2026, pages 139–149

Teochew is indeed a heritage language (HL) in all of those countries, which means that one should not presume that the Teochew language is passed down to the younger generations. A HL is often defined as a home language (Valdés, 2000), as opposed to a majority language (Montrul, 2010), and "is no longer the present dominant language where [the speaker] lives" (ElHawari, 2020). While HLs are often studied in the context of migration, in our case, Teochew is also a HL in China, spoken at home, as opposed to Mandarin.

As a result of its worldwide spread and the lack of standardization, distinct varieties have emerged between speakers in the language's homeland in southern China as well as in the global diaspora. Those varieties are shaped by long-term language contact and often exhibit phonetic interference from dominant languages as well as substantial lexical borrowing. While in China interference mostly occurs with Mandarin, in the diaspora we can observe traces of languages including Thai, Khmer, Malay, French and English.[2]

2.2. Teochew as a Language with Multiple Writing Systems

Despite being predominantly used and transmitted as an oral language, Teochew is not a non-written language. In fact, multiple writing systems are used by Teochew speakers: sinograms and the Latin-based transcriptions (or romanizations).

Sinograms. As a Sinitic language, Teochew can be written using sinograms. Yet, unlike Mandarin and Cantonese, Teochew has not benefited from a sustained effort to maintain its written form in sinograms, resulting in a heavily incomplete character set and the lack of consensus for which characters should be used for many words. Such cases include colloquial words that are not used in Mandarin, such as /ta˧pɔʊ˧/ 'boy'[3], or loanwords such as /ma˧ta˥/ 'the police', which comes from Malay.

Romanization. Multiple Latin-based transcription systems co-exist for Teochew. Historically, the *pe̍h-ūe-jī system* (also called Swatow Church Romanization) created by missionaries in the late 19th century is the first complete romanization system developed for Teochew. It was later followed by the *Guangdong Peng'im*[4] (GD), created by the Provincial Education Department of Guangdong in 1960 and based on *hànyǔ pīnyīn* used for Mandarin. On their Discord server, the international Teochew diaspora community encourages the use of *Gaginang Peng'im* (GGN), a system based on the GD but with a few modifications to better fit the pronunciation of speakers of the Southeast Asian diaspora. This version notably abandons the diacritically marked vowel <ê>, making it more accessible for typing across different keyboard layouts (see Table 1 for a full comparison of GD and GGN). As an example, the word *jap8 ghueh4* in GGN ('October') would be *zab8 ghuêh4* in GD. In both systems, the eight tones are represented using numbers.

The transcription used in the Teochew dictionary WhatTCSay is GGN, while our training set is transcribed using GD (see 4.1). In our text-based pipeline, we include a converter from GD to GGN on the transcription output.

initials		nuclei		codas	
GD	GGN	GD	GGN	GD	GGN
b	b	a	a	**b**	**p**
p	p	ai	ai	**g**	**k**
bh	bh	ao	ao	m	m
g	g	**e**	**eu**	h	h
gh	gh	**ê**	**e**	ng	ng
d	d	i	i	n	n
t	t	ia	ia		
s	s	io	io		
z	**j**	u	u		
c	**ch**	ua	ua		
r	**y**	uai	uai		
l	l	**uê**	**ue**		
m	m	ui	ui		
n	n	o	o		
ng	ng	oi	oi		
h	h	ou	ou		

Table 1: Romanization mapping between Guangdong Peng'im (GD) and Gaginang Peng'im (GGN). Non-matching graphemes are marked in bold.

Several other romanization systems exist, though they remain less widely adopted. In her dissertation on Teochew, (Tan, 2020) prefers using *cháoyǔ pīnyīn*, a transcription system created in 2010 by Chen Enquan for their dictionary. Birnie-Smith (2016) reports that Teochew Indonesians are transcribing Teochew using a system based on their usage of the Latin alphabet for Malay. Such *ad hoc* informal adaptations are widely used in the diaspora, as observed in Teochew online

[2] See McFarland (2021, 2022) for an account of studies on Southeast Asia Teochew varieties, and a thorough study of Cambodian Teochew, and Tan (2020) on the possible influence of French.

[3] See the 9 alternative forms for this word in https://en.wiktionary.org/wiki/%E4%B8%88%E5%A4%AB#Pronunciation_3 which include Hokkien-based and Teochew-based propositions.

[4] The word *peng'im* is the Teochew counterpart of *pinyin* (拼音). In this article, we use *peng'im* for any romanization.

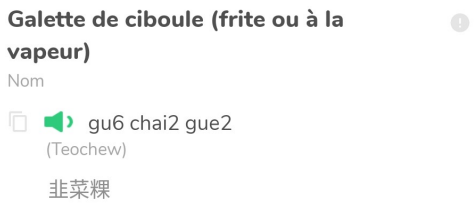

Figure 1: Screenshot of the entry for *gu6 chai2 gue2* ('chive cakes') in the French Android version of WhatTCSay3.

communities. This is one of the reasons certain communities are encouraging the use of either GD or GGN, to maximize mutual intelligibility among speakers from different regions.

2.3. WhatTCSay, a Teochew dictionary app

WhatTeochewSay (WhatTCSay) is a dictionary app for mobile phones originally crowdfunded in 2012[5] and is collaboratively developed by heritage speakers in North America and France. The current release of the app (WhatTCSay3) contains more than 6,700 entries. Each entry is composed of a word (or expression) transcribed in GGN, with a definition (either in English or French) and a part of speech. Optionally, some entries also have the corresponding sinograms, and an audio recorded by the founder. Figure 1 shows a complete entry as an example. The recording feature (indicated by the megaphone icon) is crucial for speakers who cannot read peng'im, but is only available for about 65% of the dictionary in the current version.

The next release will contain more than 10k entries and take into account accents from different regions of Guangdong as well as from different diaspora communities. To keep up with this extension, automatically generated audio will be provided for entries lacking a native speaker recording, based on the text-to-speech system developed by Magistry et al. (2024).

WTCS is a popular app among the diaspora, with almost 20k downloads in total reported by the developer in 2026. Despite encouragement to learn and use either peng'im system in the online communities, most Teochew speakers do not have proficiency in those systems, and have difficulties in reading the peng'im in the app for words that do not have an audio recording associated. It is also almost impossible for them to guess the correct spelling in the case they want to look up a word they heard. Building an audio-based lexical retrieval feature is the next step in making the app definitely accessible.

In this paper, we explore two different strategies to identify dictionary entries based on audio queries. The first approach adopts an Automatic Speech Recognition (ASR)-based pipeline in which audio queries are transcribed into peng'im and subsequently matched against candidate textual entries. The second approach relies on direct audio-to-audio matching using acoustic embeddings extracted from self-supervised speech models within a Dynamic Time Warping (DTW)-based framework.

Our main contributions are as follows: 1) a comparison of ASR-based and DTW-based methods for speech lexicon retrieval; 2) an evaluation on real-world Teochew speech data; 3) a demonstration of how self-supervised speech representations can support low-resource query-by-example retrieval.

3. Query-by-Example Spoken Term Detection

The speech and lexical retrieval task in this paper is redefined as a Query-by-Example Spoken Term Detection (QbE-STD) task. The only difference is that instead of matching all audio documents from a corpus which contain a spoken query provided by a user (Hazen et al., 2009), we match the spoken query directly with a corpus of isolated words. Existing approaches generally follow two paradigms: (1) acoustic matching, where speech features are extracted from audio signals and aligned with query representations (Naik et al., 2020; Le Ferrand et al., 2021; San et al., 2021), and (2) ASR-based pipelines, where audio is first transcribed and then matched using text-based retrieval methods (Parada et al., 2009; Lee et al., 2015; Macaire et al., 2022).

3.1. Dynamic Time Warping

The acoustic approach typically adopts a two-stage framework: extracting frame-level acoustic features from both queries and target audio, and computing a detection score for each query–target pair, typically using DTW-based template matching (San et al., 2021). DTW is an algorithm for measuring the similarity between two temporal sequences that may vary in speed or duration, which is widely adopted in speech recognition (Sakoe and Chiba, 1978).

Previous studies by Le Ferrand et al. (2020) on two endangered languages—Mboshi (Congo) and Kunwinjku (Australia) showed that classical acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual linear prediction

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

(PLP) features, can outperform neural and self-supervised representations from *Wav2Vec* models trained on those languages. This is likely due to the extremely limited amount of available training data, which constrained the effectiveness of self-supervised representation learning.

San et al. (2021) investigated spoken term detection using data from seven Australian Aboriginal languages and a regional variety of Dutch. They systematically evaluated feature extraction approaches based on *Wav2Vec 2.0* representations from both the English monolingual and multilingual *XLSR-53* models. They subsequently implemented a DTW-based detection stage. Results showed that embeddings from the 11th Transformer layer of the English *Wav2Vec 2.0* model achieved the best retrieval accuracy, outperforming MFCC and bottleneck features by 56–86%, even under mismatched speaker and recording conditions.

Their findings suggest that representations extracted from a model trained on a single language or a set of phonologically similar languages may be more beneficial for QbE-STD than a large multilingual model such as *XLSR-53* trained on a diverse set of 53 languages. This is particularly when supervised fine-tuning in the target language is not readily feasible.

3.2. Automatic Speech Recognition

Another approach to the QbE-STD task is based on ASR technique. ASR refers to the use of machines to convert human speech into corresponding text (Benzeghiba et al., 2007). With the rapid advancements in deep learning and the availability of large-scale datasets (Karpagavalli and Chandra, 2016; Ardila et al., 2020), state-of-the-art ASR systems now deliver high-precision transcription for resource-rich languages, including English, Mandarin, and others (Radford et al., 2023; Seed-ASR, 2024).

However, comparable performance in low-resource settings has only recently become achievable through large-scale and multilingual models (Pratap et al., 2024). Recent advancements in self-supervised multilingual speech models have led to a popular paradigm for solving low-resource speech recognition problems. Self-supervised learning (SSL) enables the model to learn robust acoustic representations from unlabeled data during pretraining (Conneau et al., 2021). Several studies have demonstrated that multilingual ASR architectures, such as *Wav2Vec 2.0 XLS-R* (Babu et al., 2022) and *Whisper* (Radford et al., 2023), can be effectively fine-tuned on small language-specific datasets to achieve competitive recognition accuracy. In this paper, We experiment with the multilingual *Wav2Vec2 XLS-R-300M* model (Babu et al., 2022), considering both the pre-trained base model and its fine-tuned counterpart trained on our target language Teochew.

Macaire et al. (2022) propose an ASR-driven QbE framework in which speech segments are first transcribed, and subsequently matched with corpus transcriptions using Smith-Waterman algorithm (Lecouteux et al., 2012). Their findings indicate that for french-related creole languages such as Gwadloupéyen and Morisien, a French monolingual model fine-tuned with extremely limited annotated data (as little as 10 minutes) can achieve usable performance, highlighting the potential of self-supervised ASR for low-resource linguistic documentation.

Motivated by recent progress in ASR and speech representation learning, this work investigates two complementary strategies for identifying dictionary entries from audio queries. The first approach adopts an ASR-based pipeline in which spoken queries are transcribed into peng'im and matched against candidate textual entries. The second approach performs direct audio-to-audio matching using acoustic embeddings within a DTW-based framework.

4. Methodology

4.1. Datasets

In this study, we use three different datasets.

Train. For model training, we employ the Teochew Wild corpus[6] (Pan et al., 2025), the first publicly released Teochew speech dataset with sinograms and Guangdong Peng'im (GD), a romanization system presented in Section 2.2. The corpus consists of 18.9 hours of speech collected from online media, including recordings from 20 Teochew speakers from China (11 male and 9 female), with a total of approximately 12,500 sentence-level utterances.

Test. Our test data set for assessing the retrieval results is a set of recordings made by the founder of WhatTCSay (WTCS), a speaker from the diaspora community. The dictionary consists of 4,603 mp3 files, totaling 1.28 hours of speech and 9,423 syllables (Lim et al., 2024). As expected for recordings for a dictionary, most of this data consist of single words pronounced in isolation (see Section 2.3 for more information on the dictionary).

Evaluation. To evaluate the retrieval performance of our system, we constructed an evalua-

[6] https://huggingface.co/datasets/panlr/teochew_wild

tion set consisting of 172 audio stimuli corresponding to 31 different words present in the dictionary.

We originally presented a list of 40 words (definitions in French or in English, transcription in GGN) to Teochew heritage speakers. Words were selected to cover the Teochew's full phonemic inventory at least three times. Speakers were asked to pronounce as many words as possible, depending on their knowledge. We then filtered out noisy samples and samples for which the words were pronounced by less than four speakers, so that we can have comparable data, while ensuring that the resulting 31 words still covered the Teochew phonemic inventory.

These recordings were collected by six Teochew speakers using their own mobile phones, including five from the diaspora and one from China (for a total of three male and three female). The 172 resulting stimuli are distributed as follows: 35 monosyllabic stimuli, 101 disyllabic stimuli and 36 stimuli with three or more syllables. None of them were included in the training data. This reflects realistic usage conditions, and forms an out-of-domain evaluation set.

4.2. Experimental Pipeline

4.2.1. Fine-Tuning

Fine-tuning a pre-trained model is a common approach for low-resource settings. Rather than training a model from scratch, fine-tuning adapts the parameters of a large pre-trained model to the target task using a relatively small amount of labeled data. This approach is particularly effective when the source and target domains are related, allowing the model to transfer previously learned representations while specializing for the new linguistic and acoustic characteristics (Baevski and Mohamed, 2020).

For the current work, we choose to fine-tune *Wav2Vec2-XLS-R-300M* for both ASR transcription and feature extraction. As a multilingual extension of the original *Wav2Vec2* framework, *XLS-R* scales pre-training to 436K hours of speech across 128 languages, enabling the model to learn robust cross-lingual representations and improve generalization in multilingual and low-resource settings (Babu et al., 2022).

All Teochew recordings were loaded and uniformly resampled to 16 kHz to ensure consistent input features. The corresponding transcripts were pre-processed by removing punctuation marks that did not contribute to phonetic realization and could not be represented acoustically. We further normalized the text to retain only linguistic symbols relevant to the Teochew peng'im romanization scheme. The hyperparameters are given in Table 2 and the results are shown in Table 3.

Parameter	Value
pretrained_model	Wav2Vec2-XLS-R-300M
attention_dropout	0.1
hidden_dropout	0.1
feat_proj_dropout	0.1
mask_time_prob	0.05
layerdrop	0.1
ctc_loss_reduction	mean
train_batch_size	16
num_train_epochs	50
fp16	True
learning_rate	3e-4

Table 2: Values of the hyperparameters used to fine-tune the Wav2Vec2-XLS-R-300M model on Teochew.

ASR Backbone	Val WER	Test WER
Wav2Vec2-XLS-R-300M (fine-tuned)	13.51	12.52

Table 3: ASR performance of the fine-tuned Wav2Vec2-XLS-R-300M model on peng'im transcriptions (WER, %).

4.2.2. ASR-Based Method

We employed the fine-tuned *XLS-R* model trained on Guangdong Peng'im (GD) to transcribe our test set, which consists of 172 stimuli produced by six speakers. It has 35 monosyllabic words, 101 disyllabic words, 36 words of three or more syllables.

To enable comparison with the WhatTCSay dictionary, we used the Gaginang Peng'im (GGN) system for consistency. Model outputs in GD were further converted to GGN using a rule-based converter that supports conversion between different Teochew romanization systems[7]. The workflow of this method is illustrated in Figure 2.

We then used the Levenshtein distance (Levenshtein, 1966) to measure the difference between two peng'im sequences. Initial weights were set to their default values, with substitution, insertion, and deletion costs equal to 1.

To better reflect Teochew writing system, we incorporated digraphs adapted to GGN that correspond to single phonemes (e.g., the consonant clusters <bh>, <gh>, <ng>, <ch>, which respectively correspond to the phonemes /b/, /g/, /ŋ/ and /ʃ/, and the vowel unit <eu>, which is not a diphthong but corresponds to /ɯ/).

We further analyzed the ASR outputs generated for the 4,603 audio recordings in the dictionary by comparing them with their reference peng'im transcriptions following the syllabic structure (onset, nucleus, coda). The edit-distance weights were then adjusted to account for frequent phonetic confusions that share articulatory features. We placed

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

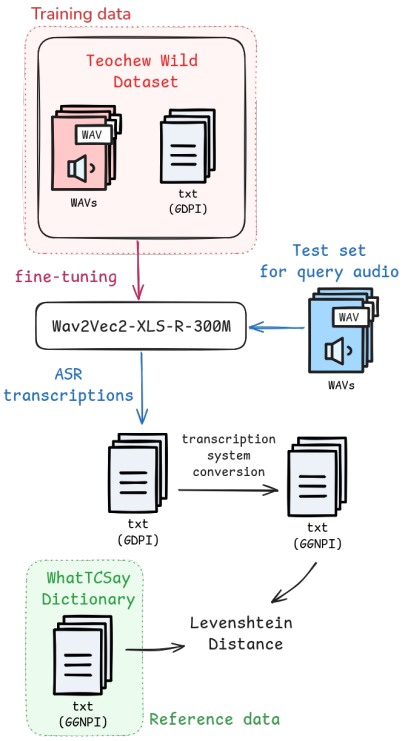

Figure 2: Workflow for the ASR-based methodology

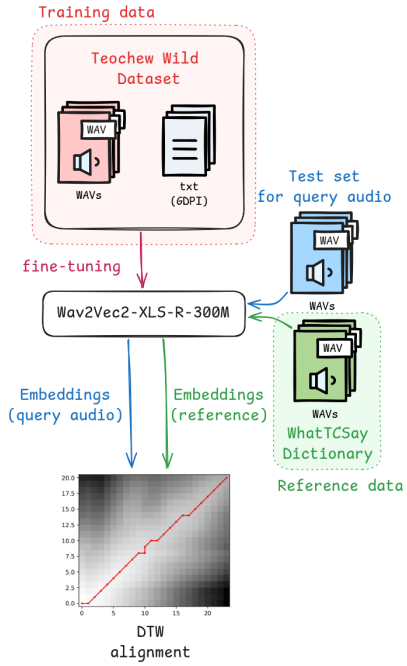

Figure 3: Workflow for the DTW-based methodology

a stronger emphasis on vowel variation, as vowels are known to be highly unstable across pronunciations (Dolgopolsky, 1964). To mitigate the risk of overfitting, six-fold cross-validation was conducted. Based on the resulting performance, the vowel substitution cost was reduced to 0.7. The four most frequent vowel confusion observed in the transcription output (i↔e, a↔o, eu↔o, eu↔a)[8] were assigned further lower weights (0.25). These adjustments also reflect speaker perception, as those contrasts are often considered equivalent by speakers in different dialectal contexts. For example, the word *new* is pronounced /siŋ/ by some speakers and /sɛŋ/ by others.

4.2.3. DTW-Based Method

This method (Figure 3) aims to retrieve dictionary audio entries by direct spoken audio matching without relying on textual representations as an intermediate step. We compare the original *Wav2Vec2-XLS-R* model and the fine-tuned version for feature extraction.

In our setup, no temporal pooling was applied (`pooling = None`) in order to preserve the full sequential structure of the speech representations. This allows DTW to perform frame-level alignment on fine-grained acoustic sequences. Dictionary candidates were subsequently ranked according to their DTW distance, where lower scores indicate higher similarity.

[8]Where <e> is /ɛ/, <o> is /ɔ/ and <eu> is /ɯ/.

4.3. Evaluation Metrics

For evaluation, we compute Recall at rank k (R@1, R@5, R@10), which measures the proportion of queries for which the correct dictionary entry is present among the top-k retrieved results (Bruno et al., 2002). This metric closely reflects real-world dictionary search scenarios, where the correct entry is expected to appear among the top-ranked results.

5. Results

5.1. ASR Performance

We used the fine-tuned model to transcribe the 172 stimuli and performed top ten retrieval under different query settings. In 74% of cases, the target entry does appear among the top ten results (see Table 4). However, a preliminary analysis revealed that the eight-tone system seems to introduce notable ASR errors. Consequently, we performed a tone-level evaluation by comparing the ASR outputs with the reference peng'im transcriptions. Across 9,343 aligned syllables, the system exhibited a tone error rate of 51.42%. This can be partly attributed to tonal variation across accents, as well as inconsistencies between the training corpus and the dictionary. Indeed, tone sandhi occurs in Teochew and the training corpus is anno-

tated with lexical tones (e.g., *jap**8** ghueh4*), while the dictionary shows tones after sandhi rules were applied (e.g., *jap**4** ghueh4*). Such tonal discrepancies introduce additional noise when computing Levenshtein distance. To mitigate this effect, we removed tone information from the queries and further evaluated a weighted Levenshtein condition. The results for three different query settings are presented in Table 4, where we can see that we do indeed manage to meaningfully improve recall, reaching 82% at recall@10.

ASR Query Setting	R@1	R@5	R@10
Query with tones	0.50	0.67	0.74
Query removing tones	0.55	0.67	0.76
Query removing tones + weighted Levenshtein	**0.57**	**0.77**	**0.82**

Table 4: ASR query retrieval performance under different query normalization settings.

5.2. DTW Performance

We leveraged speech representations extracted from both the pre-trained *XLS-R* model and its fine-tuned counterpart trained on Teochew data. And we applied a dynamic time warping (DTW)–based retrieval approach. Evaluation was conducted on an unseen, out-of-domain teochew test set comprising six speakers not included in the training data. Figure 4 reports the retrieval performance for both models using features extracted from all 24 transformer layers of each model. The fine-tuned model achieves nearly 50% relative improvement compared to the pre-trained baseline.

The base model achieves its best recall scores at intermediate layers, particularly around layer 13. In contrast, the fine-tuned model shows an improvement toward deeper layers, with the final layer yielding the highest recall scores, indicating that task-specific fine-tuning reshapes the representational hierarchy (see Table 5). Interestingly, we also observe that intermediate layers of the fine-tuned model (layers 12–17) perform comparably well in retrieving the correct entry within the top-10 results. Overall, model differences become most pronounced in the final layers where the effects of the learning objective are strong.

Method	R@1	R@5	R@10
XLS-R-300M (layer 13)	0.20	0.33	0.36
XLS-R-300M-ft (layer 24)	**0.42**	**0.59**	**0.63**

Table 5: Comparison between the two speech models from layer 1-24.

When comparing the two methods, the retrieval results show a consistent performance advantage

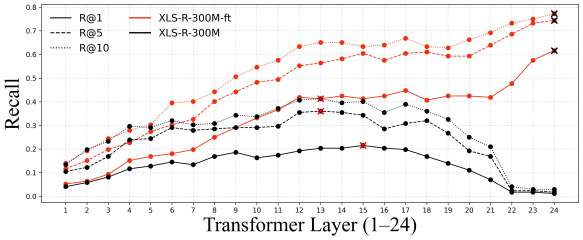

Figure 4: Retrieval results from both models' transformer layer 1-24 using DTW; the crosses indicate highest score achieved on dataset.

Method	R@1	R@5	R@10
ASR-based retrieval	**0.57**	**0.77**	**0.82**
DTW-based retrieval	0.42	0.59	0.63

Table 6: Comparison between ASR-based and DTW-based retrieval on the six-speaker evaluation set.

for the ASR-based method. It yields relative improvements of 35.7%, 30.5%, and 30.2% for R@1, R@5, and R@10 respectively.

Method	1-syll	2-syll	3+-syll
ASR-based retrieval	**0.50**	**0.75**	**0.86**
DTW-based retrieval	0.35	0.58	0.68

Table 7: Average recall (mean of R@1, R@5 and R@10) across syllable-length groups for both methods.

Table 7 illustrates how syllable length affects performance in the DTW-based method. The results show that across all syllable categories, the ASR-based method consistently outperforms the DTW-based method, and both show higher recall as word length increases. Overall, this advantage aligns with our expectations: even when ASR transcription introduces substitutions or deletions, Levenshtein distance can still recover the correct entry by matching partially shared segments. In contrast, DTW is more sensitive to local acoustic distortions and temporal misalignments (Permanasari et al., 2019).

6. Discussion

The ASR-based pipeline produces textual transcriptions that enable Levenshtein-based retrieval, even under partial overlap. This makes the approach more robust to accent variation, and provides more effective retrieval cues than purely acoustic similarity. This is particularly important for Teochew, a multivariational language characterized by pronunciation variation across speakers. However, the ASR-based pipeline requires

converting transcription outputs from one peng'im system to another before applying Levenshtein distance, and removing tones, which introduces additional process. In contrast, the DTW-based approach operates directly on speech representations, which has the advantage of bypassing orthographic or writing conventions. However, this method can be highly sensitive to factors such as background noise, volume variation, and pronunciation differences (Permanasari et al., 2019). We further assess the impact of acoustic artifacts on retrieval performance, we selected a subset of stimuli that contained noticeable silence, background noise, or both. The results shows that 8 out of 10 queries showed rank improvement when silence or noise is removed, which indicates that preprocessing has a positive effect.

Real-Time Retrieval. When deploying the system in real-life applications, apart from the evaluation metrics on recall scores, other factors like real-time retrieval and user experience also play an important role. During a workshop organized with the local Teochew community in Paris to test our system, we observed that users were generally satisfied as long as the spoken word appeared in the results page.

In addition, mismatches sometimes arose when speakers pronounced multi-word expressions as a single unit. For instance, *jiahbeung* ('to eat (rice)'), does exists in this version of the dictionary, but as separate entries: *jiah* ('to eat') and *beung* ('rice'). In such cases, prompting users to pronounce words separately after an unsuccessful search could improve usability.

Another possible strategy to reduce frustration is to offer two options after each query: repeating the same word or pronouncing a different one. Repeated attempts could potentially be leveraged to refine the retrieval process by incorporating multiple trials of the same spoken query.

Layer-Wise Analysis of SSL Representations. Our findings further suggest that the DTW-based retrieval method relying on speech representations is not as fully unsupervised as initially assumed. Using representations extracted from the fine-tuned model trained on Teochew consistently outperform those from the base model. This suggests that the ASR fine-tuning objective may contribute positively to the audio lexicon retrieval task, particularly when the fine-tuning language aligns with the target language.

Our DTW-based method shows that layers contain useful abstractions and generalizations of acoustic information. Performance varies across layers, indicating that the choice of representation level has an impact on retrieval quality.

For both the pre-trained *XLS-R* model and its fine-tuned counterpart, intermediate layers perform well in retrieving the correct entry within the top-10 results (see Figure 4). This is consistent with previous layer-wise analyses of self-supervised speech models, which suggest that middle layers often encode strong phonetic information Pasad et al. (2021) and may serve as competitive representations for speech-related tasks, and that the last layer might not always be the optimal choice (Bartelds et al., 2022; Cho et al., 2023). For example, Hao et al. (2024) report that the optimal layer for acoustic-to-articulatory inversion (AAI) task is typically located around two-thirds of the model depth.

However, the optimal layer to use for downstream evaluation may vary depending on factors such as optimization, data, and downstream task (Bordes et al., 2023). Further investigation is required to better understand these layer effects, particularly in low-resource settings.

We also observed that retrieval performance differs dramatically between models in the final layers. This behavior may be linked to ASR fine-tuning, during which the final layers become tailored to the ASR learning objective that learn to differentiate between different phonemes. Such task-oriented representations may positively transfer to the query-by-example retrieval setting, where distinguishing fine-grained phonetic differences is essential.

7. Conclusion and Perspectives

This study addresses the Teochew language, an under-resourced Sinitic language within the NLP community. We explore different strategies to identify dictionary entries based on audio queries.

The first approach uses an ASR-based pipeline to transcribe audio queries into peng'im for lexical matching, while the second performs direct audio-to-audio matching using DTW over self-supervised speech embeddings. Overall, the ASR-based method achieves superior performance.

Despite requiring additional linguistic normalization, the ASR-based method offers higher interpretability and enables integration with downstream text-based applications. In future work, it would be interesting to incorporate a language model during ASR fine-tuning to further constrain and refine the transcription outputs.

Based on ASR error patterns, we further adapted the Levenshtein distance to better account for vowel variability, which also reflects phonetic variation observed across dialectal pronunciations. As a result, this pipeline becomes less out-of-domain and more guided by supervised linguistic knowledge compared to the DTW-based

approach. These observations also raise questions regarding how to incorporate similar vowel-weighting strategies into DTW-based matching.

The DTW-based method, however, has the advantage of bypassing orthographic inconsistencies and spelling variation, which are common challenges for under-resourced languages lacking standardized writing systems. By operating directly on speech representations, our results show that embeddings extracted from the fine-tuned *XLS-R* model significantly outperform those from the pre-trained base multilingual model, which suggests that ASR fine-tuning may facilitate positive transfer to query-by-example retrieval tasks.

These findings also suggest that DTW-based retrieval is not as fully unsupervised as initially assumed, as performance remains strongly influenced by supervised ASR fine-tuning. A layer-wise analysis further reveals that the pre-trained model reaches peak retrieval performance around layer 13, whereas the fine-tuned model achieves peak performance at deeper layers (around layers 15–17 and the final layer). These observations motivate future work exploring detailed error analysis and visualization of phonetic features encoded in these layers, as well as strategies such as truncating or reinitializing higher transformer layers prior to fine-tuning.

8. Limitations

This work does not explicitly investigate how tonal information is represented or modeled. Future research should examine how ASR systems capture tonal features in Sinitic languages, such as contour tones and tone sandhi. This can provide insights into 'unretrievable' queries.

9. 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 2025-AD011014016R1 made by GENCI.

We are also grateful to the "Les Jeunes Teochew de France" association in Paris, for allowing us to recreate the "salle mauve" on their Discord for our recording needs. We would like to thank especially the members who generously contributed to the construction of the evaluation set used in the experiments.

10. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of Interspeech 2022*, Incheon, South Korea.

Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of Self-supervised Pre-Training for ASR. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7694–7698. IEEE.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural Representations for Modeling Variation in Speech. *Journal of Phonetics*, 92:101137.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. 2007. Automatic Speech Recognition and Speech Variability: A Review. *Speech communication*, 49(10-11):763–786.

Jessica Rae Birnie-Smith. 2016. Ethnic Identity and Language Choice across Online Forums. *International Journal of Multilingualism*, 13(2):165–183.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. 2023. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*.

Nicolas Bruno, Surajit Chaudhuri, and Luis Gravano. 2002. Top-k Selection Queries over Relational Databases: Mapping Strategies and Performance Evaluation. *ACM Transactions on Database Systems (TODS)*, 27(2):153–187.

Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. 2023. Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proceedings of Interspeech 2021*, pages 2426–2430.

Aron B. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej severnoj evrazii s verojatnostej točky zrenija [a probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia]. *Voprosy jazykoznanija*, 2:53–63.

Rasha ElHawari. 2020. *Teaching Arabic as a heritage language*. Routledge.

Yun Hao, Reihaneh Amooie, Wietse de Vries, Thomas Tienkamp, Rik van Noord, and Martijn Wieling. 2024. Exploring Self-Supervised Speech Representations for Cross-lingual Acoustic-to-Articulatory Inversion. In *Interspeech 2024*, pages 4603–4607. ISCA.

Timothy J. Hazen, Wade Shen, and Christopher White. 2009. Query-by-Example Spoken Term Detection using Phonetic Posteriorgram Templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 421–426. IEEE.

Shunmugam Karpagavalli and Edy Chandra. 2016. A Review on Automatic Speech Recognition Architecture and Approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4):393–404.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Phone Based Keyword Spotting for Transcribing Very Low Resource Languages. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 79–86, Online. Australasian Language Technology Association.

Benjamin Lecouteux, Georges Linares, and Stanislas Oger. 2012. Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. *Computer Speech & Language*, 26(2):67–89.

Lin-Shan Lee, James Glass, Hung-Yi Lee, and Chun-An Chan. 2015. Spoken Content Retrieval-Beyond Cascading Speech Recognition with Text Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

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.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. *Findings*.

Pierre Magistry, Ilaine Wang, and Ty Eng Lim. 2024. Experiments on Speech Synthesis for Teochew, Can Taiwanese Help? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6849–6854.

Joanna Rose McFarland. 2021. Language contact and lexical changes in Khmer and Teochew in Cambodia and Beyond. In Tom Hoogervorst and Caroline Chia, editors, *Sinophone Southeast Asia: Sinitic Voices across the Southern Seas*. Brill.

Joanna Rose McFarland. 2022. *Aspects of Cambodian Teochew Grammar: A Radical Construction Grammar Account*. Ph.D. thesis, Nanyang Technological University.

Silvina Montrul. 2010. Current Issues in Heritage Language Acquisition. *Annual review of applied linguistics*, 30:3–23.

Prajyot Naik, Manisha Naik Gaonkar, Veena Thenkanidiyoor, and Aroor Dinesh Dileep. 2020. Kernel based Matching and a Novel training approach for CNN-based QbE-STD. In *2020 international conference on signal processing and communications (SPCOM)*, pages 1–5. IEEE.

Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran. 2009. Query-by-Example Spoken Term Detection for OOV terms. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 404–409. IEEE.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Yurika Permanasari, Erwin H. Harahap, and Erwin Prayoga Ali. 2019. Speech Recognition using Dynamic Time Warping (DTW). *Journal of physics: Conference series*, 1366(1):012091.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1094–1101. IEEE.

Seed-ASR. 2024. Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition. Technical report, ByteDance.

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.

Guadalupe Valdés. 2000. Spanish for Native Speakers: AATSP Professional Development Series Handbook for Teachers K-16 (Vol. 1). *New York*.

11. Language Resource References

Lim, Ty Eng and Wang, Ilaine and Magistry, Pierre. 2024. *Audio files from WhatTCSay 3*. Zenodo.

Linrong Pan, Chenglong Jiang, Gaoze Hou, and Ying Gao. 2025. Teochew-wild: The first in-the-wild teochew dataset with orthographic annotations. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.