Universal Phone Recognition for Language Agnostic Keyword Search

Anonymous ACL submission

Abstract

Recently, significant advanced have been made in universal phone recognition. Certain of these methods allow researchers to recognize phones in thousands of languages. In this paper, we explore the usage of such universal phone recognition for phonetic keyword search (KWS). That is, we apply these methods to search for specific sequences of phones, corresponding to keywords, in a set of audio files. We find that truly universal phone recognition might not be viable for KWS, but phone recognition systems can be fine-tuned with small amounts of data (3-5 hours of recordings) to produce useful results.

1 Introduction

Keyword spotting is a speech processing task in which spoken words or phrases are identified in one or more utterances[1]. Given a target keyword or phrase, the spotting method detects the presence, and sometimes location or timestamp, of the target keyword from raw input audio or other representations of that audio (e.g., spectrograms). Models that perform this task are used to detect wake words for digital assistants and to find (and even replace) keywords in audio files (Can and Saraçlar, 2011; Audhkhasi et al., 2017). The latter application is sometimes referred to as Keyword Search (KWS).

KWS is also beginning to be applied in industry to assist humans in the post-editing of audio content, such as podcasts[2]. In these applications, KWS techniques are used to find and remove filler words or to simply search through content for the mention of certain key terms. An editor may realize, for example, that a confidential project or person was mentioned in one or more audio files, and they could use KWS techniques to locate and redact mentions of that project or person.

This kind of editing via KWS would be extremely in the context of local language translation of healthcare and legistlative information. COVID-19 prompted a dramatically increased need for the translation of health tips and hygiene information (Hardach). This information necessarily needs to be in both audio and text formats, because of the prevalence of oral cultures and illiteracy. The quality of translations also needs to be verified before publication (Ramos, 2020; Kmiecicka, 2021; Ghobadi et al., 2017), which might require many edits. However, KWS-based augmentation of this oral translation editing process is not possible with many current KWS and Automated Speech Recognition (ASR) techniques due to the lack of transcribed or otherwise labeled speech data in local, low resourced languages (Blasi et al., 2021).

In this paper, we evaluate the efficacy of language agnostic, phonetic KWS based on recent developments in universal phone recognition. The method that we present leverages a universal phone recognizer to convert speech data into a common phonetic representation (the IPA phone inventory), regardless of language. Keyword search is then performed in this IPA phone representation based on text transliterations of keywords or phones recognized in audio recordings of keywords. To demonstrate the performance of such an approach, we apply our KWS methodology to audio recordings of English, Hindi, and Telugu utterances. We find that some fine-tuning of the universal phone recognizer, Allosaurus (Li et al., 2020) in this case, may be necessary to achieve useful KWS performance.

2 Related Work

Generally, the following approaches have been employed for keyword search and spotting: phonetic speech analytics, large vocabulary continuous speech recognition (LVCSR) based methods, end-to-end neural networks, and query-by-example (QbyE) techniques.

[1] https://paperswithcode.com/task/keyword-spotting
[2] https://bit.ly/3DbXZbw

This work is primarily inspired by phonetic KWS methods (Moyal et al., 2013; Titariy et al., 2014). In such approaches, speech data is converted into sequences of phones/phonemes. Then phoneme sequences corresponding to keywords are matched to the phonemes corresponding to the speech data using, e.g., Levenshtein distance (Navarro, 2001). The keywords need not be represented in a predefined vocabulary, but, given that similar sounds might occur in a variety of places in speech data, the approach sometimes results in false positives. Existing phonetic KWS methods methods are distinguished from the current work in that they utilize language specific phone recognition and are, thus, not language agnostic or easily adapted to new languages.

Various attempts have been made to rapidly adapt keyword search and spotting to new languages in low resource scenarios. By way of example, these include Rosenberg et al. (2017); Yusuf et al. (2019) and Liu et al. (2014). The research in this vein that is most related to the current work is that of Ferrand et al. (2021), which also attempts to utilize universal phone recognition for KWS. Ferrand et al. (2021), in contrast to the currently proposed method, relies on a lexicon of spoken words that is annotated with orthographic transcriptions, whereas our method utilizes the universal phone recognition model to obtain phonetic representations of example keywords. Further, Ferrand et al. (2021) use a mapping function to convert phonetic representations of speech data back to grapheme transcriptions before search through the reference lexicon. In contrast, we perform KWS directly using the phonetic representations.

3 Methodology

See Figure 1 for an overview of our proposed approach for KWS. The approach includes: (i) recognition of phones corresponding to speech data and phones corresponding to keyword examples; and (ii) matching the phones corresponding with keywords examples to similar phone sequences occurring in the speech data. We evaluate the efficacy of universal phone recognition in this context, and we experiment with both text and audio keyword examples. In certain cases we fine-tune the universal phone recognizer with language specific data to boost KWS performance.

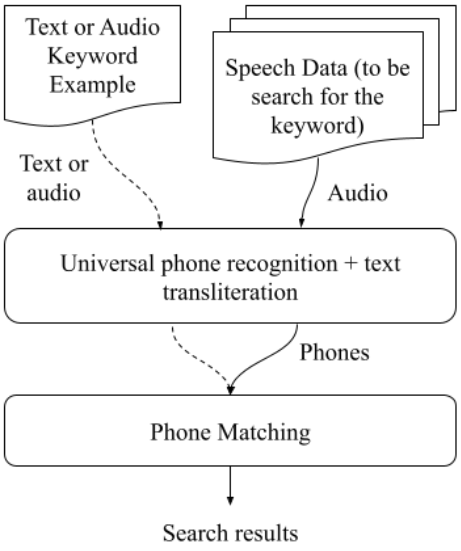

Figure 1: An overview of our method for phonetic keyword search (KWS)

3.1 Phone recognition

We convert speech data, audio keyword examples, and text keyword examples into corresponding sounds represented by phonetic symbols from the International Phonetic Alphabet (IPA). For speech/audio data, phone recognition is performed with the Allosaurus[3] universal phone recognizer, which can be applied without additional, language specific training (Li et al., 2020). To transliterate text data to phones we use the Epitran[4] grapheme-to-phoneme converter (Mortensen et al., 2018), which is specifically designed to provide precise phonetic transliterations in low-resource scenarios.

In order to ensure consistency between Allosaurus and Epitran, we took advantage of Allosaurus's inventory customization feature, giving it the phone inventories specified by the relevant language in Epitran. When this inventory is supplied as input, Allosaurus will only output symbols from the inventory.

Allosaurus is not compatible with mp3 format, so we also used the PyDub[5] library to convert any mp3 files to wav format.

3.2 Matching phones to keywords

We perform KWS with both text and audio examples of keywords. Both formats are converted into the same phonetic representation (IPA), and, thus, the same matching methodology is

[3]https://github.com/xinjli/allosaurus
[4]https://github.com/dmort27/epitran
[5]https://github.com/jiaaro/pydub

used for either format. Specifically, we use the "find_near_matches" Levenshtein distance based search within the fuzzysearch[6] Python library. This search implementation allows one to find near substring matches within a larger string. In all of our experiments, we set the max allowed distance parameter for determining a match to the length of the given example keyword (in terms of the number of IPA phone characters) divided by two.

3.3 Fine-tuned phone recognition

As mentioned, we use a universal pre-trained model from Allosaurus in order to recognize phones in speech data and audio keywords. However, we also investigate the boost in performance that can be acheived by fine-tuning the universal model with additional language specific data. This sort of fine-tuning has been shown to reduce Phone Error Rates (PERs) by 40%+ with even small amounts of language specific data (Siminyu et al., 2021). In our case, we use transcribed audio files to fine-tune Allosaurus after transliterating the transcriptions to IPA using Epitran.

The fine-tuning of Allosaurus followed used the instructions for and implementation of fine-tuning in the Allosaurus GitHub repository[7]. This implementation utilizes early stopping to avoid overfitting, where training stops if the validation PER is worse than previous PERs.

4 Experiments

To demonstrate the performance of our language agnostic (and fine-tuned) KWS methods, we search for a number of keywords in English [eng], Hindi [hin], and Telugu [tel] speech data. For English and Hindi, we utilize data from Common Voice[8][9]. For Telugu, we use data from the Microsoft Speech Corpus (Indian languages)[10].

These audio datasets contain two parts: (1) audio files; and (2) transcriptions corresponding to each audio file. The Hindi and Telugu datasets were used to fine-tune language specific Allosaurus models. However, following the work of Siminyu et al. (2021), we need only a fraction of the available data to fine-tune Allosaurus. For Hindi, we randomly selected 3000 files, or around 3.5 hours of recordings. For Telugu, we filtered out 3600 audio files by file size to get around 3GB of "medium" sized files (2000 files above 1MB in size and 1600 files between 500Kb-1MB), because the Telugu data had more variance in file size as compared to the Common Voice data. The Hindi and Telugu datasets were split into 80% for fine-tuning Allosaurus and 20% for evaluating the phonetic KWS method. A pre-trained Allosaurus model for English was already available, and, thus, we did not fine-tune a model for English.

The remaining 20% of the filtered datasets and 213 randomly selected files from the English Common Voice dataset were used to evaluate our phonetic KWS methods (described in Section 3). Certain keywords (around 20 for each languages) were chosen based on their occurrences in this data, where each keyword occurs in 3-5% of the files. To get audio examples of these keywords, a single native speaker (for each language) was recorded speaking the keywords with an iPhone 12.

5 Results

Table 1 shows the accuracy and recall of phonetic KWS search using both the universal and language specific (i.e., fine-tuned) phone recognition models. Generally, the fine-tuned, language specific phone recognition models boost KWS performance. The difference in performance between the universal and language specific models can be up to 20%+. This suggests that some fine-tuning is required for acceptable performance in phonetic KWS. However, we fine-tuned these models using existing ASR transcripts corresponding to 3-5 hours of recordings. This data is still quite small compared to datasets for modern ASR, which might include 20,000 hours of audio or more.

Further, we find that the usage of text vs. audio keywords in the KWS produces a mixed bag of results. For Telugu, we see that searches using audio keyword examples outperform text based searches by almost 30% when using the universal phone recognizer. However, we see almost no difference in the Hindi results. This is likely due to variations in the quality, variety, and pre-processing of the audio samples. We only gathered audio keyword examples from one native speaker per language, and we expect that the results might show a more consistent trend if we expanded this data using more speakers and/or a wider variety of recording devices.

[6] https://github.com/taleinat/fuzzysearch
[7] https://github.com/xinjli/allosaurusfine-tuning
[8] https://commonvoice.mozilla.org/en/datasets
[9] https://commonvoice.mozilla.org/en/datasets
[10] https://msropendata.com/datasets/7230b4b1-912d-400e-be58-f84e0512985e

Universal			Keywords	AvgOcc	Accuracy	Recall
	Text	English	20	5%	87.80%	29.60%
		Hindi	21	5%	88.50%	28.70%
		Telugu	23	3.64%	93.10%	17.30%
	Audio	English	20	5%	59.57%	48.39%
		Hindi	21	5%	78.27%	28.90%
		Telugu	23	3.64%	70%	48.40%
Language specific			Keywords	AvgOcc	Accuracy	Recall
	Text	English	20	5%	78.50%	70.50%
		Hindi	21	5%	88.90%	30.90%
		Telugu	23	3.64%	80.20%	46.70%
	Audio	English	20	5%	53.76%	52.53%
		Hindi	21	5%	78.13%	30.56%
		Telugu	23	3.64%	79.30%	36.50%

Table 1: Phonetic keyword search (KWS) performance using both universal and language specific phone recognition models and using both text keyword examples and audio keyword examples. The number of keywords per language is shown along with the average percentage in which those keywords appear in the data (AvgOcc).

6 Conclusions and Future Work

Using universal phone recognizers, we demonstrate language agnostic, phonetic keyword search (KWS) functionality. We find that fine-tuning phone recognizers with a small amount of language specific data (3-5 hours of recordings) significantly improves the performance of KWS. This sort of fine-tuning is likely needed if one wants to apply the methodology in practice. In the future, we would like to further investigate the performance of input text keywords vs. input audio keywords, and we would also like to scale this audio search up to larger datasets with more recording and more languages.

References

Kartik Audhkhasi, Andrew Rosenberg, Abhinav Sethy, Bhuvana Ramabhadran, and Brian Kingsbury. 2017. End-to-end asr-free keyword search from speech. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4840–4844.

Damián E. Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *ArXiv*, abs/2110.06733.

Dogan Can and Murat Saraçlar. 2011. Lattice indexing for spoken term detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:2338–2347.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Spoken term detection methods for sparse transcription in very low-resource settings. *ArXiv*, abs/2106.06160.

Mehdi Ghobadi, Golnaz Madadi, and Bahareh Najafian. 2017. A study of the effects of time pressure on translation quantity and quality. *International Journal of Comparative Literature and Translation Studies*, 5:7–13.

Sophie Hardach. The languages that defy auto-translate. *BBC*.

Eliza Kmiecicka. 2021. Mistakes in specialist translations and their possible consequences in the legal communication.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Chunxi Liu, Aren Jansen, Guoguo Chen, Keith Kintzley, Jan Trmal, and Sanjeev Khudanpur. 2014. Low-resource open vocabulary keyword search using point process models. In *INTERSPEECH*.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Ami Moyal, Vered Aharonson, Ella Tetariy, and Michal Gishri. 2013. Phonetic search methods for large speech databases. In *Springer Briefs in Electrical and Computer Engineering*.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Fernando Prieto Ramos. 2020. Facing translation errors at international organizations: What corrigenda reveal about correction processes and their implications for translation quality. *Comparative Legilinguistics*, 41:133 – 97.

Andrew Rosenberg, Kartik Audhkhasi, Abhinav Sethy, Bhuvana Ramabhadran, and Michael Picheny. 2017. End-to-end speech recognition and keyword search on low-resource languages. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5280–5284.

Kathleen Siminyu, Xinjian Li, Antonios Anastasopoulos, David Mortensen, Michael R Marlo, and Graham Neubig. 2021. Phoneme recognition through fine tuning of phonetic representations: a case study on luhya language varieties. *arXiv preprint arXiv:2104.01624*.

Ella Titariy, N. Lotner, Michal Gishri, and A. Moyal. 2014. A hybrid keyword spotting approach for combining lvcsr and phonetic search.

Bolaji Yusuf, Batuhan Gundogdu, and Murat Saraçlar. 2019. Low resource keyword search with synthesized crosslingual exemplars. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27:1126–1135.