Phone Based Keyword Spotting for Transcribing Very Low Resource Languages

Éric Le Ferrand,[1,2] **Steven Bird,**[1] **and Laurent Besacier**[2]
[1]Northern Institute, Charles Darwin University, Australia
[2]Laboratoire Informatique de Grenoble, Université Grenobles Alpes, France

Abstract

We investigate the efficiency of two very different spoken term detection approaches for transcription when the available data is insufficient to train a robust speech recognition system. This work is grounded in a very low-resource language documentation scenario where only a few minutes of recording have been transcribed for a given language so far. Experiments on two oral languages show that a pre-trained universal phone recognizer, fine-tuned with only a few minutes of target language speech, can be used for spoken term detection through searches in phone confusion networks with a lexicon expressed as a finite state automaton. Experimental results show that a phone recognition based approach provides better overall performances than Dynamic Time Warping when working with clean data, and highlight the benefits of each methods for two types of speech corpus.

1 Introduction

Efforts are made across Australia to preserve, document and revitalize Aboriginal languages. These languages exist primarily in spoken form, and even if there often is an official orthography available, it is not widely used by local people. Making recordings of speakers has been a widespread practice for documenting traditional knowledge. However, such recordings are often not transcribed, making them hard to access.

Manual transcription is time consuming and is often described as a bottleneck (Brinckmann, 2009). While automatic speech recognition (ASR) has seen great improvements in recent years (Povey et al., 2011; Watanabe et al., 2018), it relies on a large amount of annotated data. Attempts to build ASR systems for low-resource languages end up with high word error rate or single-speaker models making them of limited use in Indigenous contexts (Gupta and Boulianne, 2020a,b).

Such methods assume that everything should be transcribed. Bird (2020) describes a sparse transcription model where we only transcribe the words we can confidently recognize, using word-spotting, while leaving the transcription of more difficult sections for later, perhaps when a speaker is available (Bird, 2020). Based on this model, Le Ferrand et al. (2020) proposed a workflow which combines spoken term detection and a human-in-the-loop to support transcription in under-resourced settings. Such a workflow avoids the use of a language model which requires too much textual data, data that we cannot find in most Aboriginal contexts, and which only needs a few spoken terms to be annotated. While they show through their simulation the capability of iterative transcription in remote communities, the precision of their method depends on the quality of the spoken queries, and the density of the resulting transcription is limited by the size of the lexicon.

Automatic phone recognition has seen progress with minimal data (Gupta and Boulianne, 2020b; Li et al., 2020). While Bird (2020) argues that phonetic transcriptions do not stand in for the speech data and cannot be segmented to generate the required higher-level word units, we can nevertheless view phone transcriptions as a speech encoding, retaining our commitment to the sparse transcription model. Such an approach has an advantage over traditional query-by-example methods in that a simple word list can be used instead of a spoken lexicon which can be challenging to collect. In this paper we show how this can be done, and compare it with dynamic time warping (DTW) (Sakoe and Chiba, 1978) commonly used for keyword spotting for Indigenous languages. We consider both methods as applied to two very low-resource languages, Kunwinjku (gup) spoken in the far north of Australia and Mboshi (mdw) spoken in Congo Brazzaville.

2 Background

Traditional ASR systems are not well suited to Aboriginal languages. The lack of data for such languages does not allow us to train an acoustic model or a language model. Additionally, the type of data usually recorded is often spontaneous and noisy which makes it difficult to transcribe, regardless of the amount of annotated data available.

Bird (2020) describes the sparse transcription model, which combines spoken term detection with a human-in-the-loop, in an iterative process. Using spoken term detection as a transcription method allows us to avoid traditional components of an ASR system, specifically the language model, to focus only on the recognition of isolated words.

Traditional Spoken Term Detection systems rely on text-based search in lattices extracted from ASR systems (Lleida et al., 2019; Saraclar and Sproat, 2004). Attempts to train ASR systems in low-resource contexts have so far provided poor results for single speaker systems (Gupta and Boulianne, 2020a,b). This makes traditional spoken term detection approaches questionable in very low-resource settings. A few papers linked to the Babel Project have explored lattice search using ASR systems trained in low-resource settings (Gales et al., 2014; Rosenberg et al., 2017). However, they work with much larger data collections than what is available in Indigenous contexts.

Query-by-Example methods have been preferred in very low-resource contexts since they only rely on acoustic comparison between spoken queries and utterances. Le Ferrand et al. (2020) explore feature representation using DTW in an iterative pipeline following the sparse transcription model (Bird, 2020), and have been able to transcribe up to 42% of a lexicon in their speech collections. This method, however, has shown limitations in terms of robustness in the face of speaker variability. Research around speech features for spoken term detection has explored the use of bottleneck features, or the hidden representation of an auto-encoder (Menon et al., 2019; Kamper et al., 2015, 2020). Such research highlights the benefits of multilingual approaches for spoken term detection when transcribed data are limited in the target language. Others have exploited neural approaches to train word classifiers from word pairs using a Siamese loss (Settle and Livescu, 2016; Settle et al., 2017), however pairs of words are required, limiting the selection to words that can be searched.

Query-by-example relies on a spoken lexicon and, by extension, a comparison between two acoustic vectors. A difference of speakers or recording channel between the query term and the speech collection has an influence on the likelihood of a given term to be retrieved. Moreover, a spoken lexicon is not simple to gather and this therefore limits the amount of terms we can retrieve. Using a lexicon made of terms recorded in isolation for spoken term detection purposes will lead to poor precision. Another solution would be to manually extract the terms of the lexicon from a speech collection which is time-consuming. Phone recognizers, like ASR systems, also need a few hours of annotated speech to provide acceptable performance (Gupta and Boulianne, 2020b; Adams et al., 2018). However, recent work has shown how multilingual phone recognizers can be fine-tuned with minimal data to work on a new language (Li et al., 2020). Raw phone transcriptions are hard to obtain as they require the skills of a trained linguist, and they cannot help directly for retrieving higher level-units (Bird, 2020). However, the orthography of most Indigenous languages is based on their phonology and there is usually a simple mapping from graphemes to phonemes can be obtained to train a phone recognizer, even with a shallow knowledge of the phonology. A spoken term detection method based on a phone recognizer could allow us to rely only on written queries following a traditional lattice-search method.

3 Methods

We begin with a lexicon of size s consisting of audio clips of spoken words, along with orthographic transcriptions, plus a speech collection in which more instances of those words may be found.

Two spoken term detection approaches, involving a multilingual component, are investigated here: (a) a baseline method based on DTW applied on multilingual BottleNeck Features (mBNF); and (b) a method based on a textual search in phone confusion networks extracted from a universal phone recognizer (P2W).

3.1 Baseline: Sparse Transcription using DTW

We first extract acoustic features from both the corpora and lexicons. Based on general performance scores reported in the literature, and in order to compare our method with another multilingual ap-

proach, we have chosen multilingual bottleneck features. These are extracted from a model trained on the Babel corpus and consist of 80 dimension acoustic vectors. They have been extracted with the Shennong library.[1] We slide each term of the lexicon along the utterances of the corpus with a step size of 30 milliseconds. We then select the best matches for each utterance-word pair based on DTW distance and retain all matches above a threshold *m* for evaluation.

3.2 Sparse Transcription using Phone Recognition (P2W)

Li et al. (2020) introduced *Allosaurus*, a universal phone recognition system which combines a language independent encoder and phone predictor, and a language dependent allophone layer with a loss function, associated with each language (Fig. 1). Allosaurus models are trained using standard phonetic transcriptions and the *allovera* database (Mortensen et al., 2020), a multilingual allophone database that can be used to map allophones to phonemes. The model first encodes speech with a standard ASR encoder which computes the universal phone distribution. Then an allophone layer is initialized with the allophone matrix and maps the universal phone distribution into the phoneme distribution for the given target language. The resulting model can be fine-tuned and applied to unseen languages.

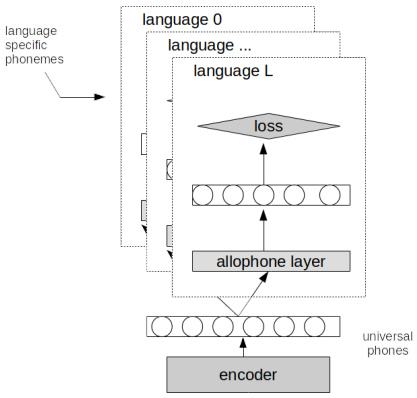

Figure 1: Allosaurus model (Li et al., 2020)

In the current context, since we only have an orthographic transcription for Kunwinjku, we transliterate it into IPA with the mapping shown in Table 1. The transcription contains some English words which will be mapped as if they were Kunwinjku words (e.g., school is written /s?kool/ instead of /skʊl/). For Mboshi, the orthographic transcription already mostly matches the corresponding phonetic transcription.[2]

graphs	a	b	d	h	e	i	ch	y	o	k	dj	s	r	rr
phones	ɑ	b	d	ʔ	ɛ	i	ʃ	j	ɔ	k	ɟ	s	ɻ	r
graphs	ng	rd	rl	nj	rn	u	f	l	m	n	w	p	t	
phones	ŋ	ɖ	ɭ	ɲ	ɳ	u	f	l	m	n	w	p	t	

Table 1: Grapheme to phoneme mapping for Kunwinjku

We fine-tuned the original pretrained model with the training and validation subsets described in Section 4 following the mapping described above, resulting in one new phone recognition model per language. We used the resulting models to automatically extract phones in confusion networks from the validation and test sets of the two languages (Mboshi and Kunwinjku) (Fig. 3).

The graph extracted is a confusion network (confnet) and consists of a size *s* sequence of phones and the top *k* likely alternatives for each phone (see Fig. 3). For each phone in the graph a probability score between 0 and 1 is assigned. We also map the lexicons into phones and convert them into a finite state automaton (FSA) in which each final state corresponds to the end of a given word (Fig.2). We explore, in the confusion networks related to our collection, every path which corresponds to a valid transition in the FSA and has a probability strictly greater than zero. If a path reaches a terminal state in the FSA, we extract the word and a score corresponding to the mean of the accumulated likelihood scores. Like the baseline with DTW, we then select the best match for each pair utterance/word pairs based on the likelihood score and keep for evaluation the matches above a threshold *n*. For both systems, we do not keep for evaluation the pairs which correspond to the query instances used to build the lexicons.

4 Data

We are using a corpus of spontaneous speech in Kunwinjku built from several sources. The training, validation and test set are described in Table 2. The training and validation sets are built from transcribed recordings made for language descrip-

[1] https://docs.cognitive-ml.fr/shennong/

[2] The tones are marked in the orthographic transcription but this feature is not taken into account in the Allosaurus model. We thus decided to treat the orthographic transcription as a phonetic transcription so the accentuated vowels are considered as new phones.

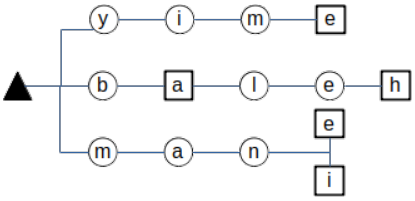

Figure 2: Example of lexicon converted into a FSA

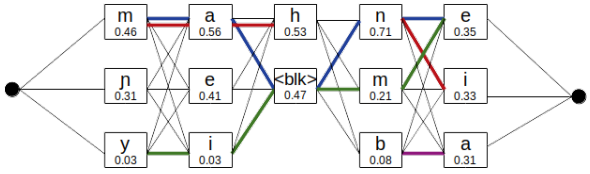

Figure 3: Example of search in a graph confusion network

tion purposes around language and emotion. They also contain some recording of guided tours of an Aboriginal town. The test set contains exclusively guided tour recordings. The orthographic transcription has been force-aligned using the MAUS forced aligner (Kisler et al., 2017). The train and validation sets contain the same 5 speakers and the test set has a non-overlapping set of 5 speakers.

We are also using a corpus of Mboshi speech which consists of 4.5 hours of speech elicited from text with orthographic transcription and a forced alignment at the word level (Godard et al., 2017). Training, validation and test sets have been extracted from the corpus and are described in Table 2. The same three speakers are represented among the three partitions.

The lexicon queries (for spoken term detection) are made of 100 words for Mboshi and 60 words for Kunwinjku. We randomly selected in the test set words which occur at least 3 times in the corresponding corpus. For each word, we manually selected examples clearly pronounced, respecting the speaker distribution of the test set (Table 3 and 4), and clipped them out.

Partitions	train	valid	test
Kunwinjku	35min45	7min39	19min43
Mboshi	21min10	10min03	3h56min

Table 2: Partition duration

Speaker	RB	TG	GN	SG	MM
Distribution	10%	25%	15%	38%	12%

Table 3: Speaker distribution across Kunwinjku lexicon

Speaker	AB	KO	MA
Distribution	63%	33%	4%

Table 4: Speaker distribution across Mboshi lexicon

5 Results

5.1 Phone Error Rate (PER)

We first evaluate the PER for both languages on the validation set. For Kunwinjku the PER started at 55.45%, and we obtained 38.82% after the system early stopped at the 24th epoch. For Mboshi the PER started at 59% and reached 38.72% at the 29th epoch. Although the PER is low considering the small amount of data used for fine-tuning Allosaurus, we would expect a bigger difference between Kunwinjku and Mboshi considering that Mboshi is read speech without foreign words and Kunwinjku is spontaneous speech containing English words. To estimate the performances for each language, we computed the PER on the test set between the top 1 phones generated by Allosaurus and the gold standard. For Kunwinjku the PER is at 39% and for Mboshi at 44%.

5.2 System performances

We evaluate the proposed methods using precision, recall and F-score.

We provide for each language the scores based on a threshold that is optimized on the respective validation sets. For the P2W method, the optimized threshold is set at 0.77 for Kunwinjku and 0.631 for Mboshi. For the DTW baseline, it is set at 0.217 for Kunwinjku and 0.174 for Mboshi. The results are detailed in Table 5. In Mboshi, the method outperforms the baseline with DTW with recall and precision. In Kunwinjku, the method does not outperform the baseline in terms of F-scores. We can see that while the baseline brings more candidates than P2W, our method is more precise. While it is clear that a phone recognition based method provides better overall performance on clean speech, the gap between the F-scores of each method in Kunwinjku is small which can make them both beneficial.

The Kunwinjku corpus contains spontaneous speech. We can observe elision phenomenon and fast speech which are not well supported by an approach based on recognition of canonical, lexical phone sequences. Figures 4 and 5 show that, while our approach seems to be more consistent across

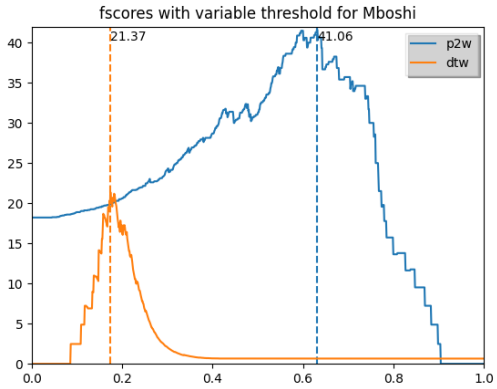

Figure 4: F-scores for Mboshi with variable thresholds on validation set

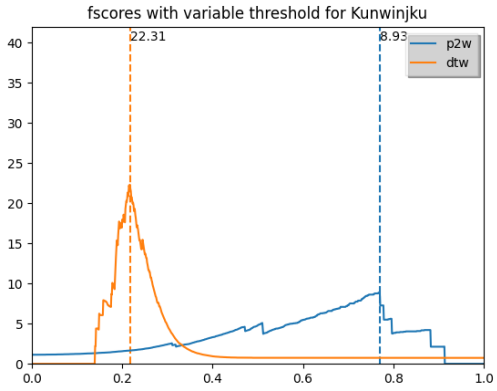

Figure 5: F-scores for Kunwinjku with variable thresholds on the validation set

	recall	precision	F-score
DTW_mb	14.55%	20.46%	17.01%
P2W_mb	22.61%	45.97%	30.31%
DTW_kun	42.09%	22.81%	29.59%
P2W_kun	17.41%	62.50%	27.23%

Table 5: Performance of spoken term detection on the test set with the optimized threshold

thresholds, it is less efficient than DTW for noisy and spontaneous speech corpora.

We present in Table 6 the top 5 false positives across methods and languages. We could only report the top 4 for P2W in Kunwinjku since most of the errors were isolated cases. We can see for P2W that the errors are made between very similar words. For Mboshi, the top 5 only includes tonal differences between the query and the hit. For Kunwinjku, the errors are made between similar words, some of which are morphologically related (balanda (man), balandaken (of the man); karrire (we-INCL go), ngarrire (we-EXCL go)). For DTW, the errors are not as consistent and the hits seem to only match subparts of the query terms (wa, **wá**re; marn**bolh**, bonj).

5.3 Speaker analysis

Le Ferrand et al. (2020) pointed out the limitation of their method in terms of cross speaker spoken term detection. To compare the two approaches on this aspect, we analyze each true positive that is output by each system: we check if the word matched is pronounced by a same or different speaker that the query term. Even if we only use the written forms of the queries for P2W, we also make the same analysis.

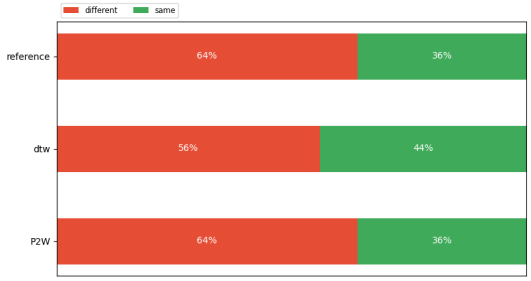

Figure 6: Proportion of same-speaker/different-speaker retrieval in Kunwinjku

Figure 7: Proportion of same-speaker/different-speaker retrieval in Mboshi

Figures 6 and 7 present the proportion of spoken terms retrieved from same-speaker or different-speaker For a fair comparison, we also compute the distribution of same/different speaker between the lexicons and all the words to be retrieved in the corpora (reference). We can see that P2W method follows the general distribution in the corpora while the baseline DTW retrieves mostly terms pronounced by the same speaker.

Mboshi P2W		Kunwinjku P2W		Mboshi DTW		Kunwinjku DTW	
Query	Hit	Query	Hit	Query	Hit	Query	Hit
ádzá	ádza	balanda	balandaken	abvúa	wa	munguyh	bonj
ádzá	adzá	birrimarnbom	birrimanbun	mwána	wa	kahdi	konhda
ngala	ngalá	mani	yiman	mvúá	wa	kunak	konhda
ngaa	ngáa	karrire	ngarrire	wáre	wa	kunred	konhda
okándá	ωkándá			ngaa	ngá	marnbolh	bonj

Table 6: Top 5 false positives

6 Combining the methods

We mentioned in Section 2 that DTW and P2W each have their own strengths. As we know, DTW will cope more easily with spontaneous speech and co-articulation effects such as assimilation and elision. Phone recognition allows us to avoid gathering spoken queries and retrieving terms with exact matching between written forms. To highlight the complementarity of the methods, we analyse the intersection of their true positives in Figure 8. We show that across both corpora the intersection of the true positives is small, and so combining the two methods can help us increase the coverage of the transcription to reach up to 49.99% for Kunwinjku and 32.16% for Mboshi.

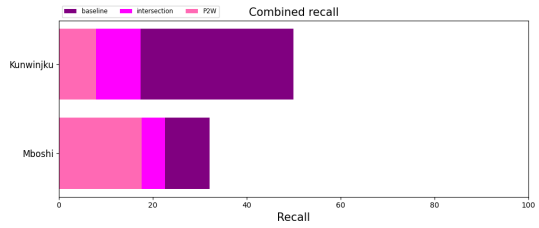

Figure 8: Relative coverage of the combined methods

We analysed the most common terms retrieved by DTW which have been ignored by P2W. For Kunwinjku, the glottal stop and doubled consonant are the phones the least properly recognized (*wanjh* written *wanj kunwardde* written *kunwarde* for example). More generally, since the data used in Kunwinjku is spontaneous speech, most of the missed hits by P2W are due to highly mistaken phone transcriptions by allosaurus. For Mboshi, beyond the main easily-confusable phones (*o* / *ω*, *e* / *ϵ* for instance) the main missed hits are due to tones or long vowels not being correctly recognized.

The baseline provides a match for every utterance/query pair if no threshold is applied. However, since P2W is restricted by the phones output by the phone recognizer, we have a limited amount of candidates regardless of the threshold. As mentioned before, this has the advantage of being more precise, but can easily miss a match if the phone lattices contain many mistakes. In view of this, we combine the two methods as follows. For each utterance/query pair brought by P2W, we first keep for evaluation the candidates which have a score greater than the P2W threshold. Then we send to evaluation every pair having a distance less than the DTW threshold. We provide in Table 7 the results for the same optimized thresholds mentioned before.

	recall	precision	F-score
comb_mb	24.89%	45.54%	32.19%
P2W_mb	22.61%	45.97%	30.31%
comb_kun	35.76%	31.48%	33.48%
P2W_kun	17.41%	62.50%	27.23%

Table 7: Performance of the combined methods

The described way of combining the methods outperforms both P2W and DTW approaches in terms of F-score. For Mboshi, we can observe a small increase of the recall with a precision barely affected. For Kunwinjku, the results are less clear. While the F-score outperforms both the baseline and P2W, combining the methods double the recall but decreases by half the precision.

7 Conclusion

This paper compares two methods of spoken term detection, one based on DTW with bottleneck features, and one based on on phone recognition. Both methods have been applied on two very low-resource languages, namely, a corpus in Mboshi recorded in a controlled environment, and a corpus of spontaneous speech in Kunwinjku recorded in remote communities. Experimental results shown that a few minutes of transcribed speech can be

used to fine-tune a universal phone recognizer. Then searching terms in a confusion network with a lexicon expressed as a FSA outperforms the baseline for Mboshi but not for Kunwinjku.

A text-based approach has the advantage over traditional Query-by-example that a set of written queries is easier to gather than spoken queries. Further analysis has shown that the proposed phone recognition approach is more robust to speaker variability and tends to be more accurate than DTW overall. However, the baseline seems to have a better coverage over the corpora and to be more suitable with noisy data.

One method relies on canonical orthography while the other relies on acoustic comparison. Both methods have their own benefits depending on the type of data they are applied to. Experimental results have shown that it is possible to take advantage of both methods to increase the overall recall while maintaining precision at an acceptable rate.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46:713–744.

Caren Brinckmann. 2009. Transcription bottleneck of speech corpus exploitation. *Proceedings of the 2nd Colloquium on Lesser Used Languages and Computer Linguistics*, pages 165 – 179.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, et al. 2017. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2521–27.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5818–22. IEEE.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. 2020. Multilingual acoustic word embedding models for processing zero-resource languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6414–6418. IEEE.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech and Language*, 45:326–347.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous communitỳ. In *COLING 2020*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 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.

Eduardo Lleida, Alfonso Ortega, Antonio Miguel, Virginia Bazán-Gil, Carmen Pérez, Manuel Gómez, and Alberto De Prada. 2019. Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media. *Applied Sciences*, 9(24):5412.

Raghav Menon, Herman Kamper, Ewald van der Westhuizen, John Quinn, and Thomas Niesler. 2019. Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. *Proceedings of Interspeech 2019*, pages 3475–3479.

David Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan Black, Florian Metze, and Graham Neubig. 2020. Allovera: a multilingual allophone database. In *LREC 2020: 12th Language Resources and Evaluation Conference*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit.

In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

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

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49.

Murat Saraclar and Richard Sproat. 2004. Lattice-based search for spoken utterance retrieval. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 129–136.

Shane Settle, Keith Levin, Herman Kamper, and Karen Livescu. 2017. Query-by-example search with discriminative neural acoustic word embeddings. *Proc. Interspeech 2017*, pages 2874–2878.

Shane Settle and Karen Livescu. 2016. Discriminative acoustic word embeddings: Tecurrent neural network-based approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 503–510. IEEE.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *Proc. Interspeech 2018*, pages 2207–2211.