Key Word searching in Speech using QbE and RNN

[1]M.Mamatha
[1]*Research scholor,RU,AP*

Abstract *- The modeling of textual content queries as sequences of embeddings for conducting similarity matching headquartered search inside speech aspects has been recently shown to beef up key word search (KWS) efficiency, peculiarly for the out-of-vocabulary (OOV) phrases. This procedure uses a dynamic time warping(DTW) centered search methodology, changing the KWS problem right into a pattern search difficulty by artificially modeling the text queries as pronunciation-founded embedding sequences. This question modeling is finished via concatenating and repeating body representations for every phoneme in the keyword's pronunciation. In this letter, we advise a query model that contains temporal context information using recurrent neural networks(RNN) educated to generate practical question representations. With experiments conducted on the IARPA Babel software's Turkish and Zulu datasets, we exhibit that the proposed RNN-founded query generation yields significant upgrades over the statistical query items of prior work, and yields a comparable performance to the state-of-the-art techniques for OOV KWS.*

Index terms *— keyword search, out of vocabulary phrases, question modeling, recurrent neural networks*

I. INTRODUCTION

RETRIEVAL of spoken content is an primary mission no longer just for finding the materials of curiosity in spoken archives, but also for facilitating computerized speech mining for higher large vocabulary continuous speech recognition (LVCSR). In distinct, KWS systems aim to reap these pursuits via finding the specific ingredients of a spoken document the place a person-provided key phrase is uttered. Probably the most intuitive and handy procedure for key phrase search is to transcribe the document into text (in the type of hypotheses lattices) utilizing LVCSR programs, and then behavior a textual content-headquartered search on the LVCSR output [1]–[3]. Nonetheless, the paucity of labeled speech coaching data in low resource languages hinders the development of reliable KWS methods, resulting in error-prone KWS systems. Moreover, if a time period of curiosity comprises phrases which aren't within the coaching vocabulary of the LVCSR approach, it cannot be located within the phrase level transcriptions from that method and so cannot be within the search index. Such phrases, referred to as out-of-vocabulary (OOV) phrases, constitute some of the essential challenges of KWS in low resource languages. Retrieval of OOV terms is the essential center of attention of this letter[31]. In KWS, the hunt term, referred to as the question, is offered in text kind and KWS methods are required to find the place the query is uttered in a speech corpus. Once the speech document is indexed with a LVCSR system, the text question is easily retrieved within this index [4]–[6]. Nonetheless, when the to be had coaching data measurement is constrained, many question phrases fall outside the insurance policy of the LVCSR procedure's training lexicon. It has been estimated that simplest about 1% of the 7000 languages spoken on this planet have sufficient linguistic resources to construct trustworthy speech to text techniques [7]. Therefore, it is essential to enhance alternative strategies which might be competent to adequately retrieve OOV phrases.

One such OOV retrieval approach is the use of sub-phrase models for lattice and index iteration [8]–[11]. Yet another line of work entails extending the language mannequin and lexicon by means of automated phrase synthesis [12] and automated text crawling from internet sources [13]; these approaches attempt to transform OOV words into in-vocabulary (IV) ones guaranteeing that those words can arise on the lattice. The most generally practiced approach makes use of proxy key words; by way of using automatically generated pronunciations of the OOV keywords, an identical sounding IV words will also be observed which might be then searched for as an alternative of the exact OOV ones [14], [15]. Apart from the LVCSR related systems described above, another method has been proposed which models the key words with their phonetic indexes as point process model(PPM) and conducts the search on the report posteriorgram [16].

The utilization of query by example Spoken term Detection (QbE-STD) techniques for retrieval of OOV terms has recently been proven to furnish state-of-the art efficiency [17]. This manner includes changing the textual content queries into sequences of frames whose representations are collectively discovered with a distance metric for use within the dynamic time warping (DTW)-established search. The OOV KWS problem is as a result converted into a QbE STD mission. This approach yields superior efficiency compared to the smooth indexing and proxy key phrase-headquartered systems on OOV phrases on account that the hunt audio will not be listed to belong to phrase/sub-phrase level states, for which the choice would not be confident, due to the OOV nature of the quest. As a substitute, the document is represented in a softer type (than the soft-indexing), by using simply the frame representations and the trouble is converted to a pattern-matching task. With this approach, two predominant advantages are accomplished: (1) A similarity score for any subsequence in audio would be got ,enabling fruitful normalization strategies as a substitute than simply sum-to-one (STO) or keyword specific thresholding (KST)

[18], [19] and, (2) the number of misses is decreased greatly, at the rate of quite multiplied false alarm charges, yielding a significant time period weighted worth growth; furthermore, the normalized pass entropy situated on the hits' ratings cut down. The predominant contribution of this letter may also be summarized as follows: The question modeling section of the DTW-based search is elevated by using inclusion of temporal expertise via generative RNNs. As an alternative of the statistical frame-based modeling and concatenation of special representations, we propose studying whole pseudo instance queries. In this part, this work differs significantly from the prior work. On account that the fundamental intention of this work is to handle modeling of OOV terms, we work closer to producing a sensible posteriorgram that might function a template for search in the document posteriorgram. Experiments conducted on the IARPA Babel application's [20] Turkish1 and Zulu2 datasets show that the generative modeling of query posteriorgrams using RNNs provides a 9p.C relative improvement on OOV retrieval performance when in comparison with the baseline that uses statistical body-headquartered query modeling.

Nonetheless, it must be famous that the efficiency of the QbE headquartered manner is independent of vocabulary, therefore it performs just as well on IV (In Vocabulary)terms. Although the IV efficiency of the proposed approach is modest compared to LVCSR-situated procedures, each techniques will also be further increased via combining their results.

II. METHODOLOGY

As mentioned in section I, we goal to generate extra realistic query posteriorgrams for KWS by incorporating temporal context know-how into question modeling. Posteriorgram-headquartered KWS converts the OOV KWS main issue right into a QbE-STD undertaking. A brief flowchart is furnished in Fig. 1.

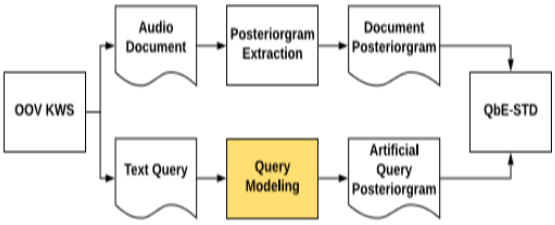

Fig. 1: Posteriorgram-based KWS flow chart

The posteriorgram representation, which is a class vs time matrix that represents the chance of a speech frame belonging to one of the finite set of courses, has been shown to provide higher results than different function representations in QbE-STD tasks, because of its speaker independence [21], [22]. In QbE-STD, given the document and question acoustic aspects (D_x $=d_{x1},d_{x2},....,d_{xT}$) and ($A_q=a_{q1},a_{q2}$ $,...,a_{qT}$) respectively, the posteriorgram representations can be expressed for the document ($X\in[0,1]$; $X=X_1...X_{xT}$ where $X(i,j)=p(s_i|a_{xj})$) and for the query ($Q\in[0,1]^{SxqT}$; $Q=q_1...q_{qT}$; $Q(i,j)=$ $p(s|a_{qj}$))where S is the number of phones. The mobile phone posteriors are bought via summing the possibilities of the context-stylish states comparable to each phoneme. Considering that in KWS, the question sequence(Q) is significantly shorter than the record sequence (X), the hunt is employed via the sub sequence DTW (sDTW) algorithm [23]. In sDTW, the boundary constraints are altered such that the short sequence is allowed to align with any sub sequence of the long sequence. The similarity rating between Q and any sub sequence of file, X(s) is calculated through the collected distance alongside the alignment course Φ

$$\text{score}(\mathcal{Q}, \mathcal{X}^{(s)}) = 1 - \frac{1}{\text{length}(\Phi)} \sum_{(r,k)\in\Phi} d(\mathrm{q}_r, \mathrm{x}_k^{(s)}) \qquad (1)$$

In KWS, nevertheless, the question just isn't spoken (Aq), but given in textual content kind. To make a QbE-STD like search feasible for KWS, the query posteriorgrams are artificially modeled making use of their (estimated) phonetic transcriptions. The unreal question modeling procedure, first proposed in [24] makes use of the grapheme to phoneme (G2P) indexes [25] of the keyword's pronunciation and concatenates body representations for each and every phoneme in the estimated pronunciation. Even as in [24], the phoneme posteriorgrams are modeled as repetitions of traditional posterior vectors for the corresponding phoneme, in [17], each mobile phone posteriorgram is a learned representation that serves as a centroid in the jointly learned distance area. Because in these methods, one body representation is learned for each phoneme, key phrase pseudo posteriorgrams are chunks of
repetitions of characteristic vectors. An illustration for the factitious query modeling is given in Fig. 2.

On this letter, we endorse coaching a recurrent neural network that may 'generate' the factitious query, given its (estimated) pronunciation. The RNN is expert with the Viterbi alignment labels of the educational audio as enter, and the educational posteriorgram as the output. The recurrent structure of the neural community makes use of the temporal knowledge in producing the posteriorgram frames. The go entropy cost is minimized comparing the actual posteriorgrams with the generated ones. With confusions are modeled by non-recurrent weights. Considering that labels to the generator network are output activations of the decoder that is used to obtain the document posteriorgram, extraordinary ground fact posterior vectors are provided for the same enter phoneme.

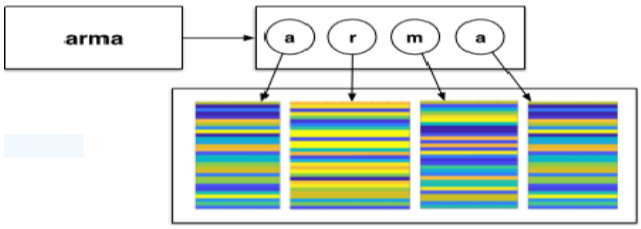

Fig. 2: Pseudo question modeling for template matching-based KWS

This technique, we anticipate to get the following contributions to the posteriorgram-founded KWS scheme: (1) The inter phoneme for this reason an averaged activation is anticipated for each phoneme in the language. (2) The intra-phoneme version as well as the confusion induced by using context is modeled with the aid of the recurrent layers. The generator community takes the indexes of the pronunciation for the text keyword as input, and outputs a posteriorgram that will emulate an actual posteriorgram for an utterance of that key phrase.

The pronunciation of the question term is obtained by means of a G2P system and represented as sequences of indexes. Network inputs are concatenations of 1-hot vectors every with a one on the index of the corresponding phoneme in the pronunciation. To facilitate a practical temporal flow and model the period of the keyword, each and every phoneme frame is repeated as generally as its normal duration estimated from the training alignment.

The document and training posteriorgrams are obtained from DNN acoustic items (with perceptual linear prediction (PLP) function as input) knowledgeable with the Kaldi speech attention toolkit [28]. The RNNs for query modeling are carried out with the Keras Toolkit [29] using the Theano again-finish. Sooner or later, the ratings for each key phrase are normalized with the aid of for the reason that their distributions over the set of one of a kind lower back hit hypotheses, following the recipe explained in [30].

III. EXPERIMENTS

In our experiments, each and every language has two scan speech documents each with its set of keywords. The first set is referred to as the dev-set and comprises a ten-hour spoken record. The Turkish and Zulu dev-sets contain 86 and 809 OOV terms for Turkish and Zulu respectively. We use the term weighted value (TWV) because the important KWS analysis metric. TWV is a linear mixture of the precision and consider at a pre defined world threshold. For progress experiments, were port two TWV scores: the maximum TWV (MTWV) and the top-rated TWV (OTWV). MTWV is defined because the TWV bought at the pleasant world threshold, whereas OTWV is the MTWV acquired with a separate, ultimate, threshold values for every keyword.

From these experiments, we opt for the quality performing techniques, now not by using direct observation of coaching or validation loss, however through staring at the precise KWS efficiency, evaluated by using the MTWV.
1) RNN memory: The first fundamental parameter to come to a decision in the generator progress is the size of the memory for use in posteriorgram new release, that is the quantity of prior frames that the error is again propagated to. In the Turkish devset experiments we evaluate 5 reminiscence settings : 10 frames (take into account the phone), 40 frames (the syllable), 100 frames (time period), one hundred fifty and 200 frames (longer terms or phrases). It must be noted that each and every body is 10 ms and the quantity of frames are chosen to have the corresponding meanings in the paren the seson natural. Even though the phrases in the key phrase record are by no means straight visible within the generator coaching, we observe that inclusion of extra reminiscence upto 100frames(one term length)accelerated the KWS performance, the place longer reminiscence prompted excessive fluctuations in efficiency and a reduce averaged performance. The outcome of these experiments are shown in Fig. 3.
2) Structure resolution: When we pick the first-class memory parameter (100 frames), we then come to a decision the structure on which to continue the experiments. The development experiments are carried out on thirteen specific models on Turkish dev-set.

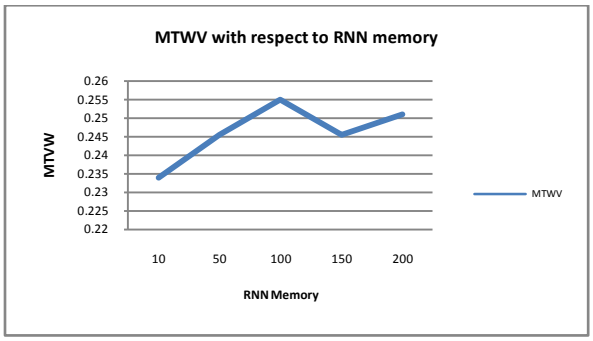

Fig 3:Line graph of the MTWV values obtained by various memory settings (τ).

There are a number of observations on the preliminary generator experiments on the Turkish dev-set. Gated RNNs perform higher than vanilla RNNs as they're less susceptible to the vanishing gradient trouble for lengthy temporal reminiscence (up to a second). Making the networks wider or deeper does no longer look aid with the MTWV performance; we speculate that given the confined amount of coaching information, as the quantity of parameters increase, generalization to previously unseen speech materials (OOV phrases) decreases. It will have to even be famous that the proper early stopping standards are utilized for every of the items.

Procedure evaluation : After the development and process set-up experiments on the Turkish dev-set, we pick the top five first-rate performing architectures to scan on Zulu. We experiment on extraordinary units since the two languages inevitably possess structural variations which would influence generator experiments. For Zulu, GRU-1-S used to be the high-quality performing mannequin, which isn't unlike LSTM-1-S in that it also has rectangular weight matrices and a gated constitution [32]. The KWS performance evaluation is employed on the following programs:

- Proxy: The work in [14], retrieving acoustically equivalent proxy key terms for OOV key words.
- Statistical body-founded query Modeling (SFQM): The work in [24], posteriorgram-based keyword search with frame degree question modeling making use of averages of posteriors.
- Generative query Modeling (GQM): The proposed, posteriorgram-established key phrase search with RNN-generated query items. The dev-set efficiency

metrics can also be obvious in table II, which can be in comparison with the equivalent work of [17].

The SFQM numbers on this letter are bigger than the fashioned work mentioned in[24]considering the fact that we use a distribution-founded normalization process (b2-norm) introduced in [30] whereas [24] uses STO normalization. The dev-set experiments function a approach resolution and tuning section. Having realized the KWS parameters comparable to premier selection threshold values, network architecture, ranking pruning threshold and so on, the genuine evaluation is done on a separate dataset, which is called the evalpart1, additionally supplied by the Babel application. The evaluation audio for each and every language is set 5 hours lengthy and the keyword units have 1216 and 1112 OOV terms for Turkish and Zulu, respectively. For the eval-set experiments we additionally document the specific TWV (ATWV) that is the process efficiency outcomes for a suite international threshold, discovered within the dev-set experiments.
The dev-set performance metrics can be visible in Table I, which is also when put next with the an identical work of [17].

IV. CONCLUSION AND DISCCUSION

In this work, we proposed a approach of modeling queries for QbE-situated OOV time period retrieval. By using modeling queries as pseudo posteriorgrams, we're competent to reframe the keyword search project as a QbE one. Previously, the usage of phoneme normal vectors as units has been proven to out participate in competing OOV retrieval ways. Utilizing that as a baseline, we advocate utilizing RNNs to generate pseudo-posteriorgrams which account for phonemic context. The ensuing question units are for this reason dynamic, incorporating context-elegant decoder confusions, not like the baseline which represent every question phoneme with an unchanging vector for the duration of its evolution despite context. We exhibit the superiority of the RNN-generated question mannequin to the normal model baseline. Extra work may just contain setting up a unified procedure that eschews the DTW completely, as well as normalization approaches that close the gap between the ATWV/MTWV and the OTWV.

REFERENCES

[1] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata: Application to spoken utterance retrieval," in Proc. Workshop Interdiscip. Approaches Speech Indexing Retrieval HLT-NAACL, 2004, pp. 33–40.

[2] D.Can and M. Sarac¸lar, "Lattice indexing for spoken term detection," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 8, pp. 2338– 2347, Nov. 2011.

[3] M. Sarac¸lar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in Proc. HLT-NAACL Main Proc., 2004, vol. 51, pp. 129–136.

[4] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in Proc.

[22] X. Anguera, L. J. Rodriguez-Fuentes, I. Szoke, A. Buzo, and F. Metze, "Query-by-example spoken term detection evaluation on low-resource languages," in *Proc. Int. Workshop Spoken Lang. Technol.Underresourced Lang.*, 2014, vol. 24, pp. 24–31.

[23] M. Mu¨ller, Information Retrieval for Music and Motion. Berlin, Germany: Springer-Verlag, 2007.

IEEE Int. Conf. Acoust., Speech, Signal Process., 2008, pp. 4969–4972.

[5] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using min- imal speech resources," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2006, vol. 1, pp. 949–952.

[6] S.-w. Lee, K. Tanaka, and Y. Itoh, "Combining multiple subword repre- sentations for open-vocabulary spoken document retrieval," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2005, vol. 1, pp. 505–508.

[7] Y. Zhang, "Unsupervised speech processing with applications to query- by-example spoken term detection," Ph.D. dissertation, Dept. Elect. Eng. Comp. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2013.

[8] Y. He et al., "Using pronunciation-based morphological subword units to improve OOV handling in keyword search," IEEE/ACM Trans. Audio, Speech, Lang. Process, vol. 24, no. 1, pp. 79–92, Jan. 2016.

[9] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 615–622.

[10] D. Karakos and R. Schwartz, "Subword and phonetic search for detecting out-of-vocabulary keywords," in Proc. Interspeech, 2014, pp. 2469–2473.

[11] L. Burget, "Hybrid word-subword decoding for spoken term detection," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 42–48.

[12] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in lowresource languages," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 8560–8564.

[13] A. Gandhe, L. Qin, F. Metze, A. Rudnicky, I. Lane, and M. Eck, "Using web text to improve keyword spotting in speech," in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2013, pp. 428–433.

[14] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2013, pp. 416–421.

[15] M. Sarac¸lar et al., "An empirical study of confusion modeling in key- word search for low resource languages," in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2013, pp. 464–469.

[16] C. Liu, A. Jansen, G. Chen, K. Kintzley, J. Trmal, and S. Khudanpur, "Low- resource open vocabulary keyword search using point process models," in Proc. Interspeech, 2014, pp. 2789–2793.

[17] B. Gundog˘du, B. Yusuf, and M. Sarac¸lar, "Joint learning of distance metric and query model for posteriorgram-based keyword search," IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1318–1328, Dec. 2017.

[18] D. R. Miller et al., "Rapid and accurate spoken term detection," in Proc. Interspeech, 2007, pp. 314–317.

[19] Y. Wang and F. Metze, "An in-depth comparison of keyword specific thresholding and sum-to-one score normalization," in Proc. Interspeech, 2014, pp. 2474–2478.

[20] M. Harper, "IARPA Babel program," Accessed: Dec. 2017, 2014. [Online]. Available: https://www.iarpa.gov/index.php/research- programs/babel

[21] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term de- tection using phonetic posteriorgram templates," in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2009, pp. 421–426.

[24] L. Sarı, B. Gu¨ndog˘du, and M. Sarac¸lar, "Fusion of LVCSR and poste- riorgram based keyword search," in Proc. Interspeech, 2015, pp. 824– 828.

[25] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," Speech Commun., vol. 50, no. 5, pp. 434–451, 2008.

[26] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recur- rent neural network architectures for large scale acoustic modeling," in Proc. 15th Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 338– 342.

[27] A. Graves, A.-rahman Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," Acoustics, speech signal process. (icassp), IEEE Int. Confe. pp. 6645–6649, 2013.

[28] J. Trmal et al., "A keyword search system using open source software," IEEE Spoken Lang. Techn. Workshop (SLT), pp. 530–535, Dec. 2014, doi: 10.1109/SLT.2014.7078630

[29] F. Chollet et al., "Keras," 2015. [Online]. Available: https://github.com/ fchollet/keras

[30] B. Gu¨ndogdu, "Keyword search for low resource languages," Ph.D. dis- sertation, Electrical Engineering Dept., Bogazic¸i Univ., Istanbul, Turkey, 2017.

[31] Batuhan Gundogdu , Bolaji Yusuf and Murat Saraclar "Generative RNNs for OOV Keyword Search ", in IEEE SIGNAL PROCESSING LETTERS, VOL. 26, NO. 1, JANUARY 2019