International Journal of Computer and Electrical Engineering, Vol.2, No.6, December, 2010
1793-8163

# Keyword Spotting in Multilingual Environments

C. P. Santhosh Kumar and V.P. Mohandas

***Abstract*—Multilingual keyword spotting is of immense interest in the Indian context with as many as 30 languages spoken across the country, and more than one language spoken in most cities. In this paper, we present the details of a gender independent multilingual keyword spotting system developed using lattices generated by a multilingual phone decoder for two of the most widely spoken Indian languages, Hindi and Tamil. For building the multilingual phone decoder, we used phonetic as well as acoustic similarities to map phones across the two languages, and see that the approach offers promising results. A distance measure based on Kullback-Leibler divergence is used for measuring the acoustic similarity of phones. We used a hybrid hidden Markov model – neural network implementation of the phone decoder for all our experiments reported in this work.**

***Index Terms*— Keyword spotting, lattices, phone recognizer, hidden Markov model , neural network.**

## I. Introduction

Keyword spotting (KWS) systems are widely used for detection of selected words in speech utterances. Searching for various words or terms is needed in applications such as spoken document retrieval (SDR) or information retrieval or in security related applications. While large vocabulary continuous speech recognition (LVCSR) based KWS [5] is the most popular approach, it suffers from the out of vocabulary (OOV) problem. Further, LVCSR based KWS systems are not suitable for multilingual applications when words or part of sentences are borrowed across languages. This is primarily due to the difficulty in training a multilingual language model catering for more than one language at the same time. When two languages are mixed, the resulting sentence would either follow the syntax of the primary language spoken, when words alone are substituted or follow the syntax of the second language for that part of the sentence borrowed from the new language. Training the language model becomes extremely difficult when the languages are of diverse linguistic characteristics, such as Tamil and Hindi. Another approach that is widely popular uses lattices generated by a phone recognizer [5],[6]. This approach has the advantage that it can be easily ported to multilingual environments, unlike the LVCSR based approach. Reference [5] gives a comparison of different KWS approaches .

People speaking some of the languages like Tulu and Kongini are spread across the country mixing with people speaking other major languages. Thus, the accent of the people speaking these languages is influenced by many languages. Also, the number of people speaking these languages is less than a few millions, and this makes the creation of a good quality database extremely difficult. In such situations, it would be advantageous to relate the phones in the language to the phones of the major language spoken in every region and develop a multilingual solution, rather than trying to develop a phone recognizer exclusively for the language. It is also worth noting that the importance of a language related technology is not directly related to the number of people speaking that language, but could be many other reasons, or even be political at times when the security threat perception arising out of people speaking that language changes.

One plausible solution to address KWS in the multilingual environments may be to run many monolingual phone decoders in parallel to generate lattices and merge these lattices for a multilingual keyword spotting system. The main difficulty with this approach is the score normalization necessary with acoustic scores generated by different decoders due to the possible mismatch caused by potential differences in acoustic and recording conditions [9], and is often done empirically. To apply this in a general multilingual framework where the number of languages can change is not so direct, and also not reliable. A truly multilingual approach would be interesting for this kind of situations wherein we do not need to do any normalization of the scores with a single system working for multiple languages at the same time.

Phone recognition accuracy and KWS accuracy are very closely related. In [3], [4], a hybrid hidden Markov model -neural network (HMM-NN) approach was used to enhance the phone recognition accuracy using split temporal context (STC) features. This approach was reported [4] to offer state-of-the-art phone recognition performance. In this work, we use the hybrid HMM-NN phone recognizer for all our experiments. Across most of the Indian languages, there is an overlap of phones; they do share some phones across languages. The natural choice to share data across languages would be by mapping phones across languages based on phonetic information. This approach is widely used in many systems [11], [12], [13], [14], and [14] reports that phone mapping based on phonetic knowledge outperformed the automatic mapping of phones based on Bhattacharyya distance. The main disadvantage of using IPA/SAMPA for mapping phones across languages is that there may be cases where there will be more than one phone in the recognition system corresponding to the same IPA/SAMPA symbol. This is more so in the case of phone recognition systems to cater for the allophonic variations of the same phoneme, while it is not so important in speech recognition systems where the

C. P. Santhosh Kumar and V.P. Mohandas are with the Department of Electronics and Communication Engineering, School of Engineering, Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore, India ( cs_kumar@cb.amrita.edu, vp_mohandas@amrita.edu)

allophonic variations could be captured in the triphones. It is quite natural therefore to suggest that acoustic similarity also should be taken into consideration while merging phones across languages, rather than using phonetic information alone. In this work, we use acoustic similarities guided by phonetic knowledge to decide if two phones across the languages could be mapped. We use KL divergence [7] to measure the acoustic similarity of phones.

We use lattice based keyword spotting approach [5], [6] using an HMM-NN decoder. Our experiments are evaluated using *Figure-of-Merit* (FOM) [10], which is the average of correct detections per 1, 2, . . . 10 false alarms per hour.

## II. Hybrid HMM-NN System

In the hybrid HMM-NN system, critical band energies are obtained in the conventional way [1], [3]. Speech signal is divided into 25 ms long frames with 10 ms shift. The Mel filterbank is emulated by triangular weighting of FFT-derived short-term spectrum to obtain short-term critical-band logarithmic spectral densities. Temporal Patterns (TRAP) feature vector describes a segment of temporal evolution of critical band spectral densities within a single critical band. The central point is the current frame and 15 frames from the past make a left context (LC) feature and similarly 15 frames from the future make the right context (RC) feature vector.

Subsequently, these feature vectors are processed for dimensionality reduction. We used Discrete Cosine Transform (DCT) for its simplicity to reduce the 16 dimensional LC and RC feature sizes to 11 [1], [2], [3], [4].

To further enhance the accuracy of the system, we concatenated 15 critical band features to generate input to two separate LC and RC neural networks. Outputs of these classifiers are subsequently merged together using another neural net. Outputs of all neural networks represent phone state posterior probabilities, and phone models have three states each. Details of the implementation can be found in [3], [4].

## III. Distance Measure for Phones

KL-divergence [7] between two HMM models is defined as [8]:

$$D(\lambda_1,\lambda_2)=\frac{1}{T}[\ln \mu(O_T \mid \lambda_1)-\ln \mu(O_T \mid \lambda_2)] \quad (1)$$

with:

$$\mu(O_T \mid \lambda_1)=\sum_{all\ S_T}\prod_{t=1}^{T} a_{s_{t-1}s_t}\, p_{s_t}(o_t) \quad (2)$$

where $O_T=(o_1,o_2,...,o_T)$ is the observed sequence, $S_T=(s_0,\ s_1,\ \ldots,\ s_T)$ is the corresponding unobserved state sequence. $a_{s_{t-1}s_t}$ and $p_{s_t}(o_t)$ are the state transition probability and probability of observation $o_t$ being produced at state $s_{t_k}$ and $T$ is the number of samples in the example of $\lambda$. In our work, we use continuous density HMMs where the probabilities are modeled using neural networks, and the probability was not measured in the likelihood sense across all possible state sequences as in (1), but for the best state sequence $S_{opt}$. Further, we need to account for the many examples used for evaluating the distance. Thus, eqn. (2) may be modified for the $k^{th}$ example as:

$$\mu(O_{T_k} \mid \lambda)=\prod_{t=1}^{T_k} a_{s_{t-1}s_t}\, p_{s_t}(o_{t_k}) \quad (3)$$

It may be noted that the optimum state sequence $S_{opt}$ is obtained by the force alignment of the observations against the correct transcription. The target phone being considered to be mapped to is assumed to follow the same state sequence. Also, in [8], it is shown that distance between two HMM models is more sensitive to the observation probability and is less sensitive to the state transition probability. Our experiments suggest that the recognition accuracy of the HMM based systems did not change with using tied state transition matrices across all models, so we used a tied state transition matrix system, where all the models share the same state transition matrix. The effect of the state transition matrix on $\mu(.)$ may be set to a constant multiplication factor, since $\mu(.)$ for all the models share the same state sequence, and thus (3) may be simplified as:

$$\mu(O_{T_k} \mid \lambda)=\prod_{t=1}^{T_k} p_{s_t}(o_{t_k}) \quad (4)$$

and the distance $D(\lambda_1,\lambda_2)$ may thus be modified as:

$$D(\lambda_1,\lambda_2)=\frac{1}{T}\sum_{k=1}^{N}\sum_{t_k=1}^{T_k}\{\ln(p(o_{t_k} \mid \lambda_1,s_{t_k}) - \ln(p(o_{t_k} \mid \lambda_2,s_{t_k}))\} \quad (5)$$

where $T$ is the total number of frames from all examples, N is the number of examples of the model $\lambda_1$ and $s_{t_k}$ is the state at time $t_k$ obtained as a result of the forced alignment. It may be noted that (5) reflects the frame recognition accuracy, and to make it reflect the phone recognition accuracy as desired, we modify (5) to:

$$D(\lambda_1,\lambda_2)=\frac{1}{N}\sum_{k=1}^{N}\frac{1}{T_k}\sum_{t_k=1}^{T_k}\{\ln(p(o_{t_k} \mid \lambda_1,s_{t_k}) - \ln(p(o_{t_k} \mid \lambda_2,s_{t_k}))\} \quad (6)$$

## IV. Experiments and Results

In all our experiments reported in this work, we used 2.0 hours of Hindi and 1.2 hours of Tamil telephone quality speech recorded at 8 kHz. The number of phones used for

TABLE I – PHONE RECOGNITION ACCURACY ( IN PER CENT ) OF THE MONOLINGUAL SYSTEMS

| Hindi | Tamil |
|---|---|
| 43.31 | 56.88 |

TABLE II – KWS ACCURACY (IN FOM PER CENT) OF THE MONOLINGUAL SYSTEMS

| Hindi | Tamil |
|---|---|
| 49.38 | 64.54 |

TABLE III – EXAMPLES OF PHONE CLUSTERS USED FOR BUILDING THE MULTILINGUAL SYSTEM

| Cluster name | Phones from Hindi | Phone from Tamil |
|---|---|---|
| Cluster_a | ax aa | ax axn aa aan |
| Cluster_pb | p P b f | P f b bh pd bd |

Hindi is 57, 43 for Tamil. We used the most frequently occurring 80 and 60 words for Hindi and Tamil respectively for KWS evaluation. All systems used in this experiment use 300 neurons to model the probability distributions and all phone models have three states each. The size of the network was chosen to match the limited data available.

### A. *Monolingual systems*

We first built monolingual systems for Hindi and Tamil. The phone recognition accuracy and KWS performance of the two systems are listed in Table I and Table II respectively.

### B. *Multilingual system for Hindi and Tamil*

Any similarity measure for a phone should be indicative of how likely is the source phone to be confused with the target phone to which it should be mapped. We used KL divergence in (6) for mapping the phones across languages. We noticed that (5) and (6) did not lead to any difference in the results, though (6) is more accurate from the phone recognition accuracy considerations while (5) is indicative of frame recognition accuracy. The size of the multilingual phone set was varied by choosing different thresholds for the phone distance. For the development of the multilingual system, we attempt to map the phones of Tamil to an acoustically and phonetically similar phone in Hindi. We refer to Tamil as the source language and Hindi as the target language. Phones are grouped into phonetically motivated clusters for the source and the target languages, with names of the clusters matching if they represent the same group. Subsequently, the distance of every phone in the source language, Tamil, to every phone in the target language, Hindi, in the same phonetic cluster to which the source language phone belongs is calculated. The closest phone in the target language is chosen for mapping if the distance is below a chosen threshold value. This threshold can be effectively used to control the number of phones in the multilingual decoder. Table III shows some examples of the phone clusters used for developing the multilingual system. If more than one source language phone is mapped to a target language phone, we let only the closest source language phone to be mapped and the rest of the phones will have language specific acoustic models. This approach helps maintain the acoustic resolution of both the languages. Table IV and V respectively show how consonants and vowels in Tamil are mapped to phones in Hindi in the multilingual system with 70 phones, that was found to be the optimum configuration for phone mapping. Phone recognition and KWS accuracy are evaluated across language specific phones

TABLE IV – PHONE MAPPING FOR CONSONANTS IN THE MULTILINGUAL SYSTEM

| Phone (ASCII) | Alphabet – Example Transcription | | Phone (ASCII) | Alphabet – Example Transcription | |
|---|---|---|---|---|---|
| | Tamil | Hindi | | Tamil | Hindi |
| k | க் - கல்வி /k ax l v ih/ | क - कलम | dh | த் - பந்தம் | द - देर /dh ae r/ |
| kd | | क - बतक /b ax tx ax kd/ | dhh | | ध - धीरे /dhh iy r ey/ |
| kh | | ख - खत /kh ax th/ | | | / |
| g | க் - பங்கு /p ax ng g U/ | ग - गरज | nd | ந் - பந்து /p ax nd dh U/ | |
| gh | | घ - घर /ghh ax r/ | p | ப் - பாடம் /p aa d ax m/ | प - फल |
| ng | ங் - தங்கம் /tx ax ng g ax m/ | ङ - गङ गा /g ax ng g aa/ | pd | | प - आप /aa pd/ |
| ch | ச் - காட்சி /k aa t ch ih/ | च - चल | f | | फ - फिर /f ih r/ |
| chh | | छ - छल /chh ax l/ | b | ப் - அம்பு /ax m b U/ | ब – बाल /b aa l/ |
| jh | ஜ் - ஜாதி /j aa tx ih/ | ज - जल | bd | | ब - रिशब /r ih sh ax bd/ |
| z | | ज - जरा /z ax r aa/ | bh | | भ - भाग /bh aa g/ |
| jhh | | झ - झगड /jh ax g ax dd/ | p | ப் - லாபம் /l aa P ax m/ | |
| ny | ஞ் - ஞானம் /ny aa n ax m/ | | m | ம் - கரும்பு /k ax r uh m b U/ | म माल |
| t | ட் - அட்டை /ax t ey/ | ट टमाटर /t ax m ax t ax r/ | y | ய் - யமுனை /y ax m uh n ax y/ | य - यदि |
| td | | ट - लौट /l aw td/ | r | ர் - மரம் /m ax r ax m/ | र - रेल /r ae l/ |
| txh | | ठ - ठण्डा /txh nn d ax/ | tr | ற் - ஒற்றன் /ao t tr ax n/ | |
| d | ட் - கடன் /k ax d ax n/ | ड - डाक | dr | ற் - கன்று /k ax n dr U/ | |
| dd | | ड - पकड /p ax k ax dd/ | rr | ற் - அறம் /ax rr ax m/ | |
| dxh | | ढ - ढाल /dxh aa l/ | l | ல் - மாலை /m aa l ax y/ | ल - लाल |
| nn | மணி /m ax nn ih/ | गणित /g ax nn ih th/ | ll | ள் - உள்ளம் /uh ll h ll ax m/ | |
| ddn | ண் - கண் /k ax ddn/ | ण – रावण /r aa v ah ddn/ | lzh | ழ் - பழம் /p ax lzh ax m/ | |
| th | த் - தட்டு /th ax t h t U/ | त - तट | v | வ் - வயது /v ax y ax dh U/ | व - वन |
| Th | த் - மதி /m ax Th ih/ | | sh | ஷ் - விஷம் /v ih sh ax m/ | श - शरीर |
| thh | | थ - छथ /chh ax thh/ | s | ச் - சங்கு /s ax ng g U/ | स - सब |
| h | பட்டம் /p ah t h t ah m/ | फूलना /p ax n h n ax g/ | hh | க் - பகல் /p ax hh ax l/ | ह – हल /hh ah l/ |
| | | | n | ன் - நான் /nd aa n/ | न - नहीं /n ah hh ihn/ |

TABLE V – PHONE MAPPING FOR VOWELS IN THE MULTILINGUAL SYSTEM

| Phone | Alphabet – Example Transcription | | Phone | Alphabet – Example Transcription | |
|---|---|---|---|---|---|
| | Tamil | Hindi | | Tamil | Hindi |
| ax | அ - அம்மா /**ax** m h m aa/ | अ - अलग | ay | எ - எட்டு /**ae** t h t U/ | ऐ - ऐनक |
| axn | | अं - अंगर /**axn** g uw r/ | ayn | | ऐं - बैंक /b **aen** k/ |
| aa | ஆ - ஆடு /**aa** d U/ | आ - आज | **ey** | ஏ - ஏணி /**ey** nn ih/ | ए – एक /**ey** k/ |
| aan | | आं - सांझ /s **aan** jhh/ | eyn | | एं - में /m **eyn**/ |
| **ih** | இ - இலை / **ih** l ax y/ | इ - इतना /**ih** th n aa/ | ao | ஒ - ஒன்று /**ao** n dr U/ | |
| iy | ஈ - ஈகை /**iy** hh ax y/ | ई - ईशवर | ow | ஓ - ஓம் /**ow** m/ | ओ - ओर |
| iyn | | इं - ईंठ /**iyn** thh/ | own | | ओं - आँखों /aan kh **own**/ |
| uh | உ - உணவு /**uh** nn ax v U/ | उ - उन | aw | ஔ - ஔவை /**aw** v ax y/ | औ - औरत |
| uhn | | उं - उंगली /**uhn** g ah l iy/ | awn | | औं - सौंदर य /s **awn** dh ax r y/ |
| U | உ - தும்பு /tx ax p h p **U**/ | | | | |
| uw | ஊ - ஊர் /**uw** r/ | ऊ - बहू | | | |
| uwn | | ऊं - कहूं /k ax hh **uwn**/ | | | |

TABLE VI – PHONE RECOGNITION ACCURACY ( IN PER CENT ) OF THE MULTILINGUAL SYSTEM

| Hindi | Tamil |
|---|---|
| 41.82 | 50.24 |

TABLE VII – KWS ACCURACY (IN FOM PER CENT) OF THE MULTILINGUAL SYSTEM

| Hindi | Tamil |
|---|---|
| 44.31 | 58.64 |

of the multilingual phone decoder for a meaningful comparison with their monolingual counterparts. Table VI and VII list the phone recognition and KWS accuracy respectively.

## V. Conclusion

We presented the results of a multilingual gender independent keyword spotting system using a multilingual phone recognizer for languages Hindi and Tamil. The approach is particularly of interest for languages that are acoustically similar, but linguistically (lexicon, grammar, etc.) different. The approach offers the capability to bootstrap an existing system to a new language with minimum amount of training data, and is very relevant in the Indian context where words/phrases across languages, are shared during normal conversations in the native language.

## References

[1] H. Hermansky, and S. Sharma, "TRAPS- classifiers of temporal patterns", 5th International Conference on Spoken Language Processing, Sydney, Nov. 1998.

[2] P. Jain and H. Hermansky, Beyond a single critical-band in TRAP based ASR, in Proc. Eurospeech2003, Geneva, Switzerland. Sept. 2003.

[3] P. Schwarz, P. Matejka, and J. Cernocky: Hierarchical structures of neural networks for phoneme recognition, in *Proc. ICASSP 2006*, Toulouse, pp. 325-328, 2006.

[4] P. Schwarz, Phoneme recognition using long temporal context, Ph.D thesis, Brno University of Technology, 2008.

[5] I. Szoke, P Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso and J. Cernocky, Comparison of Keyword Spotting Approaches for Informal Continuous Speech, Interspeech 2005 - Eurospeech, Lisabon, Portugal, Sep. 2005, pp. 633–636.

[6] M. Saraclar and R. Sproat, Lattice-Based Search for Spoken Utterance Retrieval, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics , Boston, Massachusetts, USA, May, 2004.

[7] Kullback, S., Information Theory and Statistics. Wiley, New York, 1958.

[8] Juang, B., Rabiner, L., A probabilistic distance measure for hidden markov models. In: AT&T Technical Journal, Vol. 64, No. 2. pp. 391–408, 1985.

[9] B. Ma, C Guan, H Li and C.H. Lee, "Multilingual speech recognition with language identification", In ICSLP-2002, 505-508, 2002.

[10] J.R. Rohlcek, W. Russell, S. Roukos, and H. Gish, Continuous hidden Markov modeling for speaker independent word spotting, International Conference on Acoustics, Speech and Signal Processing, ICASSP-89, Glasgow, UK, May 1989.

[11] Schultz, T., Kirchhoff, K., Multilingual Speech Processing. Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5, 2006.

[12] Schultz, T., Waibel, A., Language independent and language adaptive acoustic modeling for speech recognition. In: Speech Communication, Vol. 35. pp. 31–51, 2001.

[13] Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, Chin-Hui Lee, A study on multilingual acoustical modeling for large vocabulary ASR, Proc. ICASSP 2009.

[14] Nieuwoudt, C., Botha, E., Cross-language use of acoustic information for automatic speech recognition. In: Speech Communication, Vol. 38. pp. 101–113, 2002.

**C.P. Santhosh Kumar** was born in Kerala, India. He got his B.Sc(Engg.) degree in Electrical Engineering from Govt. College of Engineering, Trivandrum, Kerala, India in 1984 and M.Tech degree from Indian Institute of Technology, Kharagpur, India in 1986. Since then, he has worked with IPC Corporation Ltd., Singapore, Institute of Systems Science Singapore, Kent Ridge Digital Labs., Singapore, Lernout&Hauspie Asia Pacific, Singapore. He is at present with Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore, India, working towards his Ph.D degree.

Santhosh was a visiting researcher at the Institute for Infocomm research, Singapore in 2005, Faculty of Information Technology, Brno University of Technology, Czech Republic in 2007, University of Auckland, Auckland, New Zealand and University of New South Wales, Sydney, Australia in 2009.

Santhosh is at present working on developing technologies for multilingual speech recognition and keyword spotting. His areas of research interests are speech and language processing with application to the Indian context.

**V.P. Mohandas** was born in Kerala, India. He did his B.Tech from Regional Engineering College, Calicut, M.Tech from College of Engineering, Trivandrum, Kerala, and Ph.D from Indian Institute of Technology, Bombay. He is at present the Chairman, Department of Electrical Engineering, School of Engineering, Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore, India. His areas of research interests are Financial Engineering, and Soft Computing.