ISSN:0975-9646

Multilingual Automatic Speech Recognizer

Gitanjali pawar
PG Student
Dept. of Computer Engg. Late
G.N.Sapkal COE, Nashik

Prof. N R Wankhade
Assistant Professor
Dept. of Computer Engg. Late
G.N.Sapkal COE, Nashik

***Abstract*-Automatic speech recognition (ASR) frameworks are used in every day by a large number of people worldwide to manage messages, control devices, start seeks or to encourage data input in small devices. The client involvement in these situations depends on the nature of the speech translations and on the responsiveness of the framework. For multilingual clients, a further impediment to common connection is the monolingual character of various ASR frameworks, in which clients are constrained to a single present language. In this work, presents an end-to-end multi-language ASR that permits clients to choose subjective combinations of talked language. In that system function is also performed by language identification scheme additionally with two persons communicate with different language in that one person language is converted into another person language**

Keyword: Automatic speech recognition (ASR), deep neural network (DNN), language identification (LID), multilingual.

I. INTRODUCTION

Automatic speech recognition (ASR) has become gradually relevant to date, tracking the explosive development of mobile devices. The use of voice as a natural and helpful technique for human-gadget communication is particularly relevant to hands-free things (e.g., whereas driving) and interaction with small form-factor devices (e.g., wearable's). The nature of the client involvement in these situations is essentially affected by the translation accuracy and period of time responsiveness of the ASR framework.

For multi-language clients, another obstacle to regular interaction is the basic monolingual character of ASR frameworks, in which clients can talk in just a single preset language. An incorporated end-to-end multilingual architecture that expands upon the work described in [9]. In this, monolingual speech recognizers decipher the input at constant time in every of the chosen languages, whereas the LID framework tries to work out that language is talked. A choice is then made, in view of the LID decision and on the certainty scores of the individual recognizers as to which recognition result best matches the client input. This architecture keeps away from the additional latency that an early language decision would present, what's more, advantages from the additional scores from the recognizers to better choose which result to come back to the client.

Both the ASR and LID backend depend upon DNNs. throughout late years, DNNs have achieved exceptional performance in numerous and testing machine learning applications. Those embrace acoustic modeling, visual item recognition, and various others. Compared with past acoustic modeling based on GMMs, the use of DNNs presents a few points of interest. In the first place, dissimilar to GMMs, DNNs utilize a multilevel distributed representation of the data. This makes DNNs exponentially more compact than GMMs. Second, DNNs, being simply discriminative, do not force any presuppositions about the input data distribution. Further, DNNs have demonstrated successful in exploring vast amount of data, accomplishing more strong models without passing into over fitting. In this influence this last indicate use huge amounts of training data recorded from client traffic

II. RELATED WORK

A. *LVCSR:*

In which describe a few early data sharing and model grouping experiments they did to develop the recognition of English queries made to Mandarin Voice Search, in Taiwan. [2]LVCSR-large vocabulary continuous speech recognition and it was slow in process because of to create multilingual context dependent acoustic models they evaluated different methods of parameter sharing. In this paper having semiautomatic unit selection and global phonetic decision tree as two learning methods, to address this issue via effective utilization of acoustic data from multiple languages. The key issue that the learning methods are addressing was how to balance between boosting acoustic training from multiple languages and reducing acoustic data impurity arising from language mismatch. Hence, this approach seeks to use similarities among languages and dialects, associated lend itself to an simply deployable system. However, universal models tend to be larger and better in mental confusion relative to their monolingual equivalents, resulting in probably adverse effects on transcription accuracy and decoding latency.

B. *PRLM:*

The several monolingual speech recognizers were activated and give the outcome of the LID classification and phone Recognition followed by Language Modeling (PRLM) [4]. Language identification using phoneme recognition and phonotactic language modeling follows by n-gram language models and uses PRLM. It introduce gender-dependent acoustic model. This technique was used for improving speech recognition performance. But due to gender dependent accuracy was low so our system can improve accuracy for the gender-dependent the main drawbacks of this method were the latency introduced by the LID step, and the propagation of language classification errors to the final transcription.

C. GMM based classification

SVM [4] based speaker verification using GMM Model, Gaussian mixture models with universal backgrounds (UBMs) have become the standard method for speaker recognition. A speaker model is formed by MAP version of the means of the UBM. A GMM super vector was constructed by the means of adopted mixture components. A recent research is that factor analysis of this GMM super vector was an effective method for variability compensation. They consider this GMM super vector in the of support vector machines. They construct a support vector machine using method called as GMM super vector. [6] The semi-automatic unit identification method starts from the existing phonetic inventory for multiple languages. Both learning methods, one on the use of new cross lingual speech units and another on the use of a global decision tree, were shown to produce superior speech recognition performance over the respective baseline systems. There was vast opportunity to develop new learning methods in the space of multilingual. [4] There was a parametric likelihood density function called as GMM which was represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution in a biometric system, such as vocal-tract related spectral features measured in a speaker recognition system.

D. Dialect identification:

Automatic Dialect Classification (ADC) has recently gained huge amount of interest within the area of speech process. Dialects of a language ordinarily were mirrored in terms of their phoneme space, word pronunciation/selection, and speech traits. These characters were clearly visible in natural speaker-to-speaker spontaneous conversations. However, pronunciation cues in prompted/read speech were usually ignored by the community.

E. *Score normalization*:

A transformation performed [10] on the scores to improve the performance of the speaker verification system. Accepting or rejecting a applicant speaker purely based on the log likelihood scores from the UBM (background model) and the hypothesized speaker model is highly error-prone. This was due to the following reasons:

1. Limited modelling capability of the models.
2. Variations in inter-speaker scores.
3. Variations in inter-session scores.

III. PROPOSED SYSTEM

This architecture shows different languages, allowing customers to typically participate with the structure in languages. The language recognition depends on the combination of a specific DNN-based LID classifier and the translation confidences transmitted by the individual speech recognizers. In this data misused by the DNN-based LID classifier with the unusual state data related with the language model of the speech recognizer. In that system function is also performed by language identification scheme additionally with two persons communicate with different language in that one person language is converted into another person language.

This architecture sets up an instrument to execute language selection in almost continuous. This licenses customers to clear change among various languages under the existence of using a monolingual ASR. Evaluated the system to the extent both precision and response time in a huge database including certifiable development data and multiple languages. Results exhibit that the proposed architecture is outfitted for managing various languages without huge impact on exactness and latency contrasted from our monolingual speech recognizers. In system we extract acoustic feature that Mel –Frequency cepstral coefficients (MFCC). Deep Neural Networks for LID, It is a completely associated feed-forward neural network with hidden units that implemented as rectify linear units (ReLU)

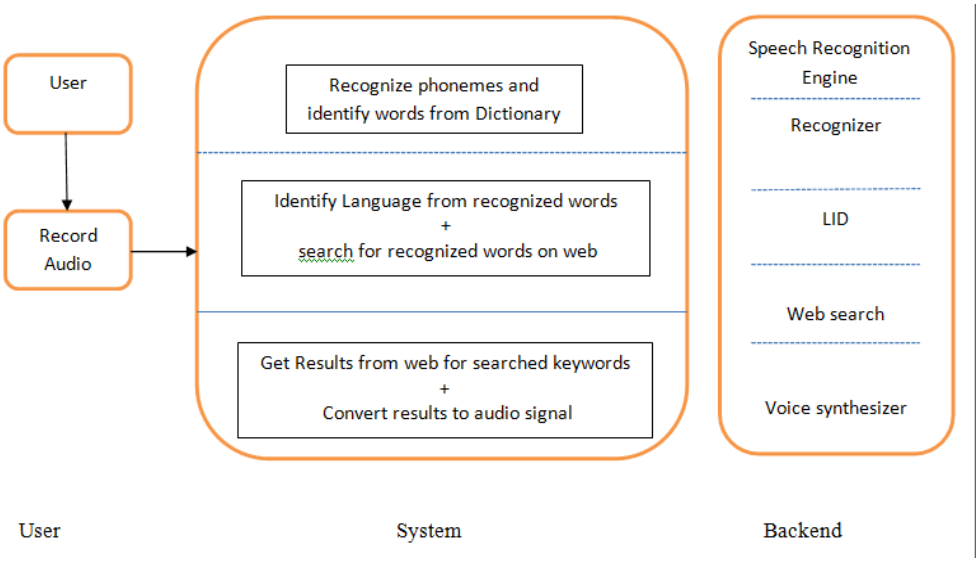

Fig -1: System Architecture

Input: audio from mobile phone, browser or similar internet connected device capable of record audio

1. Accept audio stream
2. Recognize phonemes
3. Then speech recognizer that accepts an audio stream and streams back transcription results
4. Web search server accepts Recognition transcription and retrieves search results. The results sometimes include a text summary which is intended to be read back to the user.
5. Voice synthesizer takes text as input and produces a language waveform as output. The selected language influenced using voice synthesis.

IV. Experimental Result

This system is implemented in .Net Framework 3.5 using c. For the verification of the results Indian dataset is used. In this work ASR is evaluate under the various words.

The performance of speech recognition system can be evaluated in terms of Word Error Rate (WER) lower error rate shows superior accuracy in speech recognition.

Table 1: WER of ASR system for different languages

WER Calculation			
	S + D + I	N	WER = (S+D+I)/N
Marathi	3	36	0.083333333
English	4	29	0.137931034
Hindi	7	39	0.179487179

The table 1 shows WER for ASR system in which S is the number of substitutions, D is the number of deletions, I is insertions, N is number of words in reference

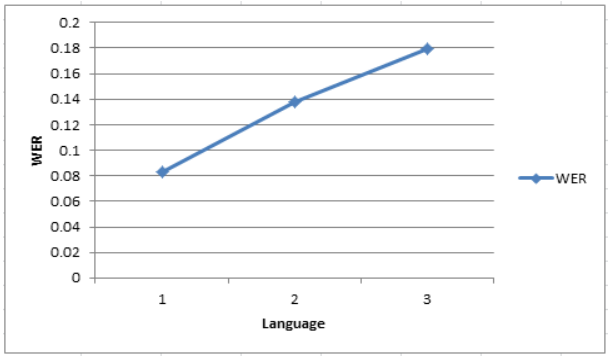

Fig 2.WER for word set on different languages

In figure 2 shows graph for three languages in which lower WER is more accurate .

V. Conclusion

The propose architecture supports different languages, allowing users to naturally relate with the system in several languages. The language detection relies on the combination of a specific LID classifier and the transcription confidences emitted by the individual speech recognizer. The audio information oppressed by the DNN-based LID classifier is high-level information related to the language model of the speech recognizer which complementing. Unlike other methodologies, this architecture creates a mechanism to implement language selection in almost real time.

The appearance of using a monolingual ASR allows users to transparent switch among changed languages under. We calculate the system in terms of both accuracy and response time in a huge database including real traffic data and languages. This will show that the propose architecture is capable of managing various languages without major impact on accuracy compared to our monolingual speech recognizers.

VI. Acknowlegement

The success and the final outcome of this project required a lot of advice and help from many people and I am extremely fortunate to have got this throughout the end of my project work. I am grateful to my guide, Prof. N.R.Wankhade sir, for his guidance and support. Thanks to all staff members of my college.

References

[1] Javier Gonzalez-Dominguez,David Eustis, Ignacio Lopez-Moreno,Andrew Senior,Franoise Beaufays,and Pedro J. Moreno, A Real-Time End-to-End Multilingual Speech Recognition ArchitectureIEEE 2015.

[2] T. Schultz and A. Waibel, Language independent and language adaptive large vocabulary speech recognition, inProc. ICSLP, 1998, vol.1998, pp. 18191822.

[3] H. Lin, L. Deng, J. Droppo, D. Yu, and A. Acero, Learning methods in multilingual speech recognition, in Proc. NIPS, Vancouver, BC, Canada, 2008.

[4] H.-A. Chang, Y. H. Sung, B. Strope, and F. Beaufays, Recognizing English queries in Mandarin voice search, inProc. IEEE Int. Acoust., Speech, Signal Process. (ICASSP), Conf.,May2011,pp. 50165019.

[5] H. Lin, J. T. Huang, F. Beaufays, B. Strope, and Y. H. Sung, Recognition of multilingual speech in mobile applications, inProc. IEEE Int. Conf. Acoust.,Speech, Signal Process. (ICASSP), Mar. 2012, pp. 48814884.

[6] G. Heigold, V. Vanhoucke, A. W. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, Multilingual acoustic models using distributed deep neural networks, inProc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 86198623.

[7] A. Mohamed, G. Dahl, and G. Hinton, Acoustic modeling using deep belief networks,IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 1422,Jan. 2012.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet,Front-end factor analysis for speaker veri_cation, IEEE Trans.Audio, Speech, Lang. Process., vol. 19, no. 4, pp. 788798, Feb.2011.

[9] G. Liu, Y. Lei, and J. H. Hansen, "Dialect identification: Impact of difference between read versus spontaneous speech," in EUSIPCO '10,2003–2006.

[10] R. Auckenthaler, M. Carey, andH. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems,"Digital Signal Process., vol. 10, no. 1/2/3, pp. 42–54, 2000.

[11] G. Tucker and A. Tucker, "A global perspective on bilingualism and bilingual education, ser. ERIC (Collection)," ERIC Clearinghouse on Languages and Linguistics, 1999 [Online]. Available:http://books.google.com/books?id=sEB5tgAACAAJ

[12] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior,V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,"IEEE Signal Process. Mag.,vol.29,no. 6, pp. 82–97, Nov. 2012.