International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 1 pp. 292-296 Sept 2017 - March 2018 www.csjournals.com

Automatic Speech Recognition: A Review

Shubhnandan Singh Jamwal
Department of Computer Science and IT, University of Jammu, J&K
jamwalsnj@gmail.com

Abstract: The most important area of the research of the Natural Language Processing is the development of Automatic Speech recognition system. There is also a considerable interest in interdisciplinary combinations of automatic speech recognition (ASR), machine learning, natural language processing, text classification and information retrieval. In this paper the different AI models are reviewed which are used for the development of the ASR. The approaches which are used in speech recognition are Hidden Markov Model, Dynamic time warping, Gaussian Mixture Model and Artificial Neural Network.

Keywords: Automatic Speech Recognition, Hidden Markov Model, Dynamic time warping, Gaussian Mixture Model, Artificial Neural Network.

Introduction

The words in the text are main key to our understanding of the information conveyed by any document. But in the models world the text in the document is now spoken and it comes in the form of the output as speech. Therefore the systems are also accepting the input as speech and this field is the emerging field of the researchers. Franco Canavesio and Giuseppe Castagneri [1] observed that continuous improvements of algorithms for Automatic Speech Recognition (ASR), and the availability on the market of a variety of devices, will enhance the possibility of voice input applications in telecommunications. Thipe Modipa, Marelie Davel, and Febe de Wet [2] analysed the automatic speech recognition (ASR) systems which are increasingly being developed for under-resourced languages, especially for use in multilingual spoken dialogue systems. They investigated different approaches to the acoustic modelling of Sepedi affricates for ASR. They determined that it is possible to model several of these complex consonants as a sequence of much simpler sounds. This approach reduces the Sepedi phoneme inventory from 45 to 32, resulting in simpler dictionary development and transcription processes, as well as more accurate acoustic modelling. S. Stuker [3] analysed that automatic speech recognition (ASR) systems have been developed only for a very limited number of the estimated 7,000 languages in the world. In order to avoid the evolvement of a digital divide between languages for which ASR systems exist and those without one, it is necessary to be able to rapidly create ASR systems for new languages in a cost efficient way. Grapheme based systems, which eliminate the costly need for a pronunciation dictionary, have been shown to work for a variety of languages. They are thus destined for porting ASR systems to new languages. Their research paper studies the use of multilingual grapheme based models for rapidly bootstrapping acoustic models in new languages. The cross language performance of a standard, multilingual (ML) acoustic model on a new language is improved by introducing a new, modified version of polyphone decision tree specialization that improves the performance of the ML models by up to 15.5% relative.

Literature Review

Hema Raghavan and James Allan [4] highlighted the problems that arise due to variations of spellings of names that occur in text, as a result of which links between two pieces of text where the same name is spelt differently may be missed. The problem is particularly pronounced in the case of ASR text. They proposed the use of approximate string matching techniques to normalize names in order to overcome the problem and also showed that how we could achieve an improvement if we could tag names with reasonable accuracy in ASR.

Y. -C. Tam, Y. Lei, J. Zheng and W. Wang [5] analysed that detecting automatic speech recognition (ASR) errors can play an important role for effective human-computer spoken dialogue system, as recognition errors can hinder accurate system understanding of user intents. Our goal is to locate errors in an utterance so that the dialogue manager can pose appropriate clarification questions to the users. They proposed two approaches to improve ASR error detection: (1) using recurrent neural network language models to capture long-distance word context within and across previous utterances; (2) using a complementary ASR system. The intuition is that when two complementary ASR systems disagree on a region in an utterance, this region is most likely an error. We train a neural network predictor of errors using a variety of features. They also performed experiments on both English and Iraqi Arabic ASR and observed significant improvement in error detection using the proposed methods.

S. Stuker, M. Paulik, M. Kolss, C. Fugen and A. Waibel [6] described work in coupling automatic speech recognition (ASR) and machine translation (MT) in a speech translation enhanced automatic speech recognition (STE-ASR) framework for transcribing and translating European parliament speeches. They demonstrated the influence of the quality of the ASR component on the MT performance, by comparing a series of WERs with the corresponding automatic translation scores. By porting an STE-ASR framework to the task at hand, and showed how the word errors for transcribing English and Spanish speeches can be lowered by 3.0% and 4.8% relative, respectively. N. Cooke, A. Shen and M. Russell [7] highlights the use of gaze information in dynamic model-based adaptation methods for noise robust ASR. Previous use of gaze (eye movement) to improve ASR performance involves shifting language model probability mass towards the subset of the vocabulary whose words are related to a person's visual attention. Motivated to improve Automatic Speech Recognition (ASR) performance in acoustically noisy settings by using information from gaze selectively, we propose a `Selective Gaze-contingent ASR' (SGC-ASR). In modelling the relationship between gaze and speech conditioned on noise level - a `gaze-Lombard effect'-simultaneous dynamic adaptation of acoustic models and the language model is achieved. Evaluation on a matched set of gaze and speech data recorded under a varying speech babble noise condition yields WER performance improvements.

Maria Shugrina[8] addressed the problem of formatting the output of an automatic speech recognition (ASR) system for readability, while preserving word-level timing information of the transcript. Our system enriches the ASR transcript with punctuation, capitalization and properly written dates, times and other numeric entities, and our approach can be applied to other formatting tasks. The method we describe combines hand-crafted grammars with a class-based language model trained on written text and relies on Weighted Finite State Transducers (WFSTs) for the preservation of start and end time of each word.

Fernando de-la-Calle-Silos, Francisco J. Valverde-Albacete, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno [9] presented the advances in the modeling of the masking behavior of the human auditory system (HAS) to enhance the robustness of the feature extraction stage in automatic speech recognition (ASR). The solution adopted is based on a nonlinear filtering of a spectro-temporal representation applied simultaneously to both frequency and time domains---as if it were an image---using mathematical morphology operations. A particularly important component of this architecture is the so-called structuring element (SE) that in the present contribution is designed as a single three-dimensional pattern using physiological facts, in such a way that closely resembles the masking phenomena taking place in the cochlea. A proper choice of spectro-temporal representation lends validity to the model throughout the whole frequency spectrum and intensity spans assuming the variability of the masking properties of the HAS in these two domains. The best results were achieved with the representation introduced as part of the power normalized cepstral coefficients (PNCC) together with a spectral subtraction step. This method has been tested on Aurora 2, Wall Street Journal and ISOLET databases including both classical hidden Markov model (HMM) and hybrid artificial neural networks (ANN)-HMM back-ends. In these, the proposed front-end analysis provides substantial and significant improvements compared to baseline techniques: up to 39.5% relative improvement compared to MFCC, and 18.7% compared to PNCC in the Aurora 2 database.

Yun-Nung Chen, Kai-Min Chang, and Jack Mostow [10] reported on a pilot experiment to improve the performance of an automatic speech recognizer (ASR) by using a single-channel EEG signal to classify the speaker's mental state as reading easy or hard text. We use a previously published method (Mostow et al., 2011) to train the EEG classifier. They used its probabilistic output to control weighted interpolation of separate language models for easy and difficult reading. The EEG-adapted ASR achieves higher accuracy than two baselines. We analyze how its performance depends on EEG classification accuracy. This pilot result is a step towards improving ASR more generally by using EEG to distinguish mental states. Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church [11] processed spoken documents with few resources. Moreover, connecting black boxes in series tends to multiply errors, especially when the key terms are out-of-vocabulary (OOV). The proposed alternative applies text processing directly to the speech without a dependency on ASR. The method finds long (~ 1 sec) repetitions in speech, and clusters them into pseudo-terms (roughly phrases). Document clustering and classification work surprisingly well on pseudo-terms; performance on a Switchboard task approaches a baseline using gold standard manual transcriptions.

Hari Krishna Maganti and Daniel Gatica-Perez [12] observed that accurate speaker location is essential for optimal performance of distant speech acquisition systems using microphone array techniques. However, to the best of their knowledge, no comprehensive studies on the degradation of automatic speech recognition (ASR) as a function of speaker location accuracy in a multi-party scenario exist. In this paper, they described a framework for evaluation of the effects of speaker location errors on a microphone array-based ASR system, in the context of meetings in multi-sensor rooms comprising multiple cameras and microphones. Speakers are manually annotated in videos in different camera views, and triangulation is used to determine an accurate speaker location. Errors in the speaker location are then induced in a systematic manner to observe their influence on speech recognition performance. The system is evaluated on real overlapping speech data collected with simultaneous speakers in a meeting room. The results are compared with those obtained from close-talking headset microphones, lapel microphones, and speaker location based on audio-only and audio-visual information approaches.

Models of ASR

There is considerable interest in interdisciplinary combinations of automatic speech recognition (ASR), machine learning, natural language processing, text classification and information retrieval. Many of these boxes, especially ASR, are often based on considerable linguistic resources [11]. The different AI models which are used in speech recognition are Hidden Markov Model, Dynamic time warping, Gaussian Mixture Model and Artificial Neural Network. Dynamic time wrapping measures the differences and similarity between two time or frequency-varying temporal sequences. Most of the researchers of the Indian languages are using the dynamic time wrapping methods. The researchers are also using the combination of HMM based DTW techniques for recognizing speech.

Hidden Markov model (HMM) is commonly employed for acoustic modelling in speech. HMM is very common model among researchers for the development of the ASR's but sometimes in big vocabularies the use of the HMM is not recommended because of its performances. The HMM are replaced by Artificial Neural networks (ANN) which are considered as powerful tools in speech recognition. These models are commonly used for the speech recognition systems. ANN's can handle a large volume of data effectively provided the computing power of the system is not limited. Gaussian Mixture Model is also used by the researchers on Indian languages. The performance of the ASR of Indian Languages using Gaussian mixtures proved very much beneficial.

Conclusions

Automatic Speech Recognition is the technology that allows human beings to use their voices with a computer. The interface is modelled using various machine learning models in such a manner that sophisticated variations are also recognized. Hidden Markov Model, Dynamic time warping, Gaussian Mixture Model and Artificial Neural Network are commonly used for the development of the ASR's. The developments of the ASR's are considered as a part of Natural Language Processing. Directed Dialogue conversations and Natural Language Conversations are commonly studied by the researchers for the development of the ASR.

References

[1] Franco Canavesio and Giuseppe Castagneri. 1986. FEEDBACK FORMAT AND USER TRAINING IN ASR OF DIGITS. SIGCHI Bull. 18, 2 (Oct. 1986), 71–73. https://doi.org/10.1145/15683.1044101

[2] Thipe Modipa, Marelie Davel, and Febe de Wet. 2010. Acoustic modelling of Sepedi affricates for ASR. In Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT '10). Association for Computing Machinery, New York, NY, USA, 394–398. https://doi.org/10.1145/1899503.1899552

[3] S. Stuker, "Modified polyphone decision tree specialization for porting multilingual Grapheme based ASR systems to new languages," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 2008, pp. 4249-4252, doi: 10.1109/ICASSP.2008.4518593.

[4] Hema Raghavan and James Allan. 2004. Using Soundex codes for indexing names in ASR documents. In Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004 (SpeechIR '04). Association for Computational Linguistics, USA, 22–27.

[5] Y. -C. Tam, Y. Lei, J. Zheng and W. Wang, "ASR error detection using recurrent neural network language model and complementary ASR," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 2312-2316, doi: 10.1109/ICASSP.2014.6854012.

[6] S. Stuker, M. Paulik, M. Kolss, C. Fugen and A. Waibel, "Speech Translation Enhanced ASR for European Parliament Speeches - On the Influence of ASR Performance on Speech Translation," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, USA, 2007, pp. IV-1293-IV-1296, doi: 10.1109/ICASSP.2007.367314.

[7] N. Cooke, A. Shen and M. Russell, "Exploiting a 'gaze-Lombard effect' to improve ASR performance in acoustically noisy settings," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 1754-1758, doi: 10.1109/ICASSP.2014.6853899.

[8] Maria Shugrina. 2010. Formatting time-aligned ASR transcripts for readability. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10). Association for Computational Linguistics, USA, 198–206.

[9] Fernando de-la-Calle-Silos, Francisco J. Valverde-Albacete, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno. 2015. Morphologically filtered power-normalized cochleograms as robust, biologically inspired features for ASR. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 23, 11 (November 2015), 2070–2080. https://doi.org/10.1109/TASLP.2015.2464691

[10] Yun-Nung Chen, Kai-Min Chang, and Jack Mostow. 2012. Towards using EEG to improve ASR accuracy. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12). Association for Computational Linguistics, USA, 382–385.

[11] Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church. 2010. NLP on spoken documents without ASR. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10). Association for Computational Linguistics, USA, 460–470.

[12] Hari Krishna Maganti and Daniel Gatica-Perez. 2006. Speaker localization for microphone array-based ASR: the effects of accuracy on overlapping speech. In Proceedings of the 8th international conference on Multimodal interfaces (ICMI '06). Association for Computing Machinery, New York, NY, USA, 35–38. https://doi.org/10.1145/1180995.1181004