A Survey: The Capabilities of Automatic Speech Recognition (ASR) Systems Facilitating Speech User Interfaces

Nicholas C. Defranco

Software Development and Data Science, Seneca College

BTH 645: Multimedia Elements for User Interfaces

Sunny Shi

March 12th, 2023

Abstract

Automatic Speech Recognition (ASR) is a process of transcribing speech similar to the job of human transcribers. ASR is a multi-purpose technology and, as such, is a very dense topic. Due to its multitude of utilities, ASR caught the attention of technical experts with a diverse set of backgrounds. Computer scientists, medical specialists, education planners etc., are all actively researching with many simultaneous ongoing projects to improve the usability of multimedia user interfaces. Specifically, the primary motivation behind all these research initiatives is maximizing accessibility for disabled users. This survey delves into moderate detail about some of the incentives in numerous fields.

1. Introduction

In the past several years, research on advancements in Automatic Speech Recognition (ASR) technology often involves deep learning approaches. Such approaches attempt to mimic human's ability to process the two (2) types of speech: audio, i.e., sound waves produced from human vocal cords, and visual, i.e., mechanical movements of specific facial organs from which humans can exclusively use to discern language (Alharbi et al., 2021). Systems leveraging ASR provide an alternate method to interact with multimedia user interfaces. One of the most prevalent of such systems is voice assistants, which enable users to use countless "smart" features in a hand-free fashion. Another use case is in speech identification systems.

This topic is relatively new in multimedia projects. Unsurprisingly, researchers encountered many difficulties in many facets of this uncharted technology. With the increasing number of discovered problems, researchers conducting studies silo their work and focus on one problem at a time. Because the field of ASR is so expansive, the article cannot cover the entire scope of ASR. Instead, it will cover a select amount of research to provide a sufficient basis. The core topic of this survey is to explore advancements in multimedia user interfaces concerning ASR. Thus, this survey centres on improvements in human-to-machine interactions (and not human-to-human interactions).

The order in which the content appears in this survey is as follows: In part 2, a discussion on past feats and discoveries, followed by part 3, what issues currently plague researchers, and finally, part 3, which discusses what the future has in store.

2. Mature Features of ASR that provide adequate performance

Understanding speech involves comprehending two components of language: phonetics (pronunciation) and syntax (grammar). Similarly, engineers design ASR systems to perceive language based on the same two (2) components. Engineers represent the components with trained acoustic and language models for phonetics and syntax, respectively. In contemporary research, researchers construct these models using various deep-learning approaches. With these approaches, researchers have achieved Word Error Rates (WERs), an ASR accuracy metric, below ten percent (10%) (Passricha & Aggarwal, 2018).

2.1 Acoustic Model Architecture

The most common approach to creating Acoustic Models (AMs) is using a combinative model of a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models. A CNN processes raw speech data in two (2) main steps: (1) extracting features from the speech wave and (2) classifying them into discrete categories (Passricha & Aggarwal, 2018).

First, an ASR system must perform spectral analysis, which distills information from the speech signal (Li et al., 2019). Feature extraction is particularly applicable in speaker verification systems, which must collect various characteristics by challenging the speaker's identity (Hossan et al., 2010). The most common way to extract features is to use some implementation of Mel Frequency Cepstral Coefficients (MFCC). Generally, an MFCC is a spectrum of a pre-determined number of coefficients, features of interest or formants within an audio sample. Formants are part of the spectral envelope, a graphical representation of phonemes, syllables, and words. The spectrum indicates the prevalence of these features at a given time. Simultaneously, the MFCC omits

information in the speech signal. Most notably, the pitch is usually of dubious value in this field (Hossan et al., 2010). Section 3.2 discusses an application of this property of MFCCs in the problem of child speech recognition.

Classification of specific sounds is contingent on past sounds. Consider intonation variances when asking a question versus when making a claim. To this end, the second step uses LSTM, which mimics the recollection abilities of humans when they decipher speech from an external source. The design of classical LSTM includes a network, i.e., a Recurrent Neural Network (RNN), of memory blocks. Each of these store speech fragments (the features) within memory cells (Sak et al., 2014). These features of speech data, known as frames in this context, are translated into a finite set of symbols, which constitute special features of a language, i.e., the formants.

There is a variant of LSTM known as Bidirectional LSTM (BLSTM), which improves upon its unidirectional counterpart. The design of a BLSTM involves two distinct Recurrent Neural Networks (RNNs), one for processing the past and the other for processing the future (Graves & Schmidhuber, 2005).

2.2 Language Model Architecture

A Language Model (LM) predicts the likelihood of forthcoming words in speech. The first step in training an LM is to give it an adequate basis for the language in question to make intelligent predictions. Unequivocally, larger speech training data sets bolster throughput in ASR systems (Bulyko, 2007). Therefore, teaching the ASR system a breath of vocabulary is imperative for promising results. Researchers from the Linguistic Data Consortium (LDC) accumulated a voluminous amount of speech data, including fictitious news stories from human subjects posing as anchorpeople (Graff, 2002) and telephone conversations between strangers (Kumar et al., 2020). These, and many more, speech data sources are freely available on the World Wide Web (WWW).

The most popular LM is the n-gram model, which studies the likelihood of word sequences of length n. An n-gram is a set of n successive words within a sentence (Jurafsky & Martin, 2023, p. 3). This model predicts how likely the last word in an n-gram would appear based on how many occurrences the final word in the n-gram follows the rest of the set in the training data (Xiong et al., 2018). A reasonable assumption is that a bigram, an alias for two-gram, i.e., a model containing sequences of length two (2), approximately offers the same efficacy as any other length, simplifying the computation process (Xiong et al., 2018). This approximation is known as the Markov assumption (Jurafsky & Martin, 2023, p. 4).

3. Encountered Challenges in ASR

3.1 Speech Overlapping

Researchers collectively achieved minimal Word Error Rates (WERs) in ideal circumstances. However, conditions are sometimes suboptimal. Some are susceptible to various distortions. One kind of distortion is a discerning speech from a single target speaker in a recording with multiple speakers talking simultaneously. This would require the ASR system to be selectively attentive, effectively mimicking the widely-known human phenomenon, the cocktail party (Isik et al., 2016).

Researchers devised a technique called Deep Clustering (DPCL) to rectify this issue. DPCL entails different properties of mixed speech throughout the recording, known as speaker embeddings, and constructing a neural network of feature vectors. The neural network intelligently predicts portions of the speech recording that originate from the target speaker.

Unfortunately, as Menne et al. (2019) points out, the sole use of DPCL does not provide optimal results for more realistic data samples. However, the same researchers also state that when DPCL works in tandem with DNN-HMM, it produces more promising results (Menne et al., 2019). An explanatory description of DNN-HMM, a hybrid acoustic model, goes beyond the scope of this survey.

3.2 Child Speech Recognition

The architectures discussed in the previous section are exclusively optimized for adult speech, neglecting any use cases for children. The reasoning behind the lack of focus on children is that the task of child speech recognition is far more nuanced for several reasons. Firstly, since children's vocal organs are still developing, phonetic sounds may drastically differ from child to child. Secondly, children have less interpersonal communication experience than adults, resulting in more unintentional speech errors. Finally, only a relatively small amount of children's sample speech data is publicly available (Kumar et al., 2020). This data differs from the sample adult data availability discussed in section 2.2, which, in contrast, is plentiful. With all these hurdles in mind, researchers understand that the design of their acoustic models demands more attention.

Studiously, researchers opted to reuse their existing trained acoustic models to recognize adult speech to simplify the problem. Empirical evidence from performance evaluations shows that a mixed-data model achieves better results than a model using only one type of speech data (Kumar et al., 2020). The model had to coerce the spectrum of possible frequencies in the dataset such that differences in vocal organs do not adversely impact the prediction. Researchers achieve Mixed-data acoustic models by combining speech data using Vocal Tract Length Normalization (VLTN) (Serizel & Giuliani, 2015). VLTN is a procedure that functions as a pre-processing stage before the model receives the data as input. The design of VLTN works in conjunction with MFCC (discussed in section 2.1). Using the VLTN procedure, data models can ignore specific properties of speakers' voices to generalize them for better, more consistent results. With the help of VLTN, ASR systems' performance enhanced, reaching 16 percent (16%) WER (Serizel & Giuliani, 2015).

4. Future Research Directions in ASR

4.1 Integrating ASR in the Second Language Learning Classes

ASR empowers inclined language learners to self-learn. Theoretically, one could learn an entire language in this way. Realistically, however, earnest learners prefer to learn in a classroom

where they can interact with others. ASR can still be advantageous for these learners too. One merit of ASR is that it can positively change the learning dynamic in second language classes, serving as an invaluable tool to teachers and students alike. Traditionally, these classes spend a disproportionate amount of time on written exercises. The usual culprit for this shortcoming is time, setting, and motor-skill (ability to perform basic movements) constraints. The negative consequences of this curriculum include lacking self-efficacy and self-confidence in one's abilities; exhibiting apparent deficiencies when communicating with other speakers (Liu et al., 2022). Consequently, skill evaluation tests may suggest their abilities are insufficient for real-world usage.

Liu et al.'s (2022) study analyzed students' attitudes toward ASR as the primary learning method in the classroom. The students were actively learning English as a Foreign Language (EFL). According to the results, the students spoke highly of the potential of ASR in the classroom overall. The inherent flexibility, addressing the shortcomings of traditional classroom learning, primarily attributed to these results. In addition, the students attested that the ASR systems' objective performance appraisals gave them a deeper understanding of their strengths and weaknesses. In response, the ASR system helpfully tailors subsequent practice exercises to focus on the student's weak points. As a result, the learning experience motivated students to work to the best of their abilities (Liu et al., 2022). Moreover, their findings also suggest that while students appreciate the benefits, they believe it should only be a supplementary tool as teachers still provide an irreplaceable instructional experience. Most notably, teachers contribute to academic success by solidifying a core knowledge base, which would help guide students in their future studies (Liu et al., 2022).

Ultimately, schools must be willing to adapt their learning methodologies to leverage ASR. Unfortunately, schools will not likely adapt in the foreseeable future. Despite the previously mentioned merits, the reluctance of education planners could pose an impediment.

4.2 Spotting Dysfunctions without Human Intervention

4.2.1 Developmental Language Disorder in Bilingual Children

Developmental Language Disorder (DLD) is an inconspicuous language learning dysfunction. In the field, a Speech-Language Pathologist (SLP) is the designation who diagnoses DLD. DLD in bilingual children can be difficult to identify, especially when one of the two languages is foreign to the delegated SLP (Albudoor & Peña, 2022). SLPs must holistically examine their patients' language-learning abilities by testing both known languages individually. If the SLP is unfamiliar with one of the two languages, they must rely on a heuristic to obtain conclusive insights. Unfortunately, with the unavailability of resources such as appraisal guides and reliable human resources like interpreters; and time limitations, SLPs may inadvertently declare false positives on bilinguals while working (Albudoor & Peña, 2022). ASR can potentially help fill this knowledge gap.

Albudoor & Peña's (2022) study assessed the feasibility of using a custom ASR system to grade children's performance on grammar and pronunciation tests. They benchmarked the system against human graders. The results from the ASR system were comparable to that of the manual markers. Although the test may have had some bias that weighed in favour of the ASR system, the test indicates ASR systems could still be a viable tool for SLPs. The main incentive for SLPs to embrace ASR in the field is that it enables them to tackle language barrier issues (Albuddor & Peña, 2022).

4.2.1 Identifying Parkison's Disease in Patients

Parkinson's Disease is an untreatable malady affecting the nervous system's ability to produce dopamine, a natural hormone influencing aspects of mood and controlling mobility. Unfortunately, once infected, the progression of the disease is irreversible. Currently, it is only possible to mitigate the worsening of symptoms with prescribed medication. Medical professionals recognize that prompt diagnosis is key for effective mediation treatment. Terriza et al. proposed a specialized ASR system to help reach that goal (Terriza et al., 2022).

Terriza et al. (2022) suggested using laughter to determine dopamine levels. Deliberately, they trained an ASR system to acquire information solely from laughter rather than regular speech. They based the design on the grounds of contemporary human anatomy research, which shows that laughter and speech are alike in that their production processes are similar (Provine & Emmorey, 2006). However, unlike speech, laughter candidly conveys mood, streamlining the system's predictive techniques (Terriza et al., 2022). Such simplifications bear improvements to ASR system efficacy.

4.3 Recognizing Visual Speech

Most research in the field delves into acoustic modelling, giving little attention to visual modelling concepts. Given circumstances when a visual-capturing device is present, like using a computing device equipped with a reasonably-performing camera, ASR systems can draw out additional worthwhile speech information. Effectively, supplementary information may be used as a fallback when the quality of information from audio extraction is poor. Visual modelling is powerful as it can circumvent noise when inspecting distorted speech samples. Remarkably, utilizing this approach can advance research on solving the challenging problem of deciphering mixed-speech samples (see section 3.1) (Debnath et al., 2022). AV-ASR, Audio-Visual Automatic Speech Recognition, probes and perceives data from audio and optical stimuli, eliciting more comprehensive predictions.

Fundamentally, including visuals as a constituent part of ASR requires computer vision knowledge. According to Debnath et al., visual speech is dual-featured: motion and guise or look of mouths (Debnath et al., 2022). Studying the appearance of visual speech entails examining a still image at a given time. In other words, it is static. On the other hand, motion detection looks for appearance changes as time goes on, i.e., motion is dynamic.

Here, features refer to visual speech's *valuable* properties, excluding properties such as luminescence changes and the colour of skin tone because, as with many computer vision

problems, these do not provide any meaningful information. Hence, algorithms typically start by converting true-colour input images into a grayscale representation disregarding the other dubious properties. From there, the algorithm derives an edge map. Edge maps make up the basis of feature detection. When examined, edge maps can uncover the desired (static) features. Changes in edge maps as time moves indicate motion. Thus, detecting motion requires a sequence of appearance samples (that is, frames). This procedure is called the Local Binary Pattern (LBP) (Debnath et al., 2022).

References

[1] Albudoor, N., & Peña, E. D. (2022). Identifying Language Disorder in Bilingual Children Using Automatic Speech Recognition. *Journal of Speech, Language, and Hearing Research, 65(7),* 2648–2661. https://doi.org/10.1044/2022_JSLHR-21-00667

[2] Aldarmaki, H., Ullah, A., Ram, S., & Zaki, N. (2022). Unsupervised Automatic Speech Recognition: A review. *Speech Communication, 139*, 76–91. https://doi.org/10.1016/j.specom.2022.02.005

[3] Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access, 9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

[4] Bulyko, I., Ostendorf, M., Siu, M., Ng, T., Stolcke, A., & Çetin, Ö. (2007). Web resources for language modelling in conversational speech recognition. *ACM Transactions on Speech and Language Processing, 5(1)*, 1–25. https://doi.org/10.1145/1322391.1322392

[5] Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: N-gram Language Models* (3). Pearson Education. https://web.stanford.edu/~jurafsky/slp3/3.pdf

[6] Debnath, S., Roy, P., Namasudra, S., & Crespo, R. G. (2022). Audio-Visual Automatic Speech Recognition Towards Education for Disabilities. *Journal of Autism and Developmental Disorders.* https://doi.org/10.1007/s10803-022-05654-4

[7] Debnath, S., & Roy, P. (2021). Appearance and shape-based hybrid visual feature extraction: toward audio-visual automatic speech recognition. Signal, Image and Video Processing, 15(1), 25–32. https://doi.org/10.1007/s11760-020-01717-0

[8] Graff, D. (2002). An overview of Broadcast News corpora. *Speech Communication, 37(1)*, 15–26. https://doi.org/10.1016/S0167-6393(01)00057-7

[9] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005, 4,* 2047–2052 vol. 4. https://doi.org/10.1109/IJCNN.2005.1556215

[10] Hossan, M. A., Memon, S., & Gregory, M. A. (2010). A novel approach for MFCC feature extraction. *4th International Conference on Signal Processing and Communication Systems*, 1–5. https://doi.org/10.1109/ICSPCS.2010.5709752

[11] Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-Channel Multi-Speaker Separation using Deep Clustering. *arXiv e-prints,* arXiv:1607.02173 https://doi.org/10.48550/arxiv.1607.02173

[12] Kumar, M., Kim, S. H., Lord, C., Lyon, T. D., & Narayanan, S. (2020). Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children. *Computer Speech & Language, 63*, 101101. https://doi.org/10.1016/j.csl.2020.101101

[13] Li, W., Zhang, P., & Yan, Y. (2019). TEnet: target speaker extraction network with accumulated speaker embedding for automatic speech recognition. *Electronics Letters, 55(14)*, 816–819. https://doi.org/10.1049/el.2019.1228

[14] Liu, J., Liu, X., & Yang, C. (2022). A study of college students' perceptions of utilizing automatic speech recognition technology to assist English oral proficiency. *Frontiers in Psychology, 13,* 1049139–1049139. https://doi.org/10.3389/fpsyg.2022.1049139

[15] Menne, T., Sklyar, I., Schlüter, R., & Ney, H. (2019). Analysis of Deep Clustering as Preprocessing for Automatic Speech Recognition of Sparsely Overlapping Speech. *Proceedings of INTERSPEECH 2019,* 2638-2642. https://doi.org/10.21437/Interspeech.2019-1728

[16] Passricha, V., & Kumar Aggarwal, R. (2018). Convolutional Neural Networks for Raw Speech Recognition. *From Natural to Artificial Intelligence - Algorithms and Applications.* https://doi.org/10.5772/intechopen.80026

[17] Provine, R. R., & Emmorey, K. (2006). Laughter Among Deaf Signers. *Journal of Deaf Studies and Deaf Education, 11(4)*, 403–409. https://doi.org/10.1093/deafed/enl008

[18] Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. *ArXiv*, 338-342. https://doi.org/10.48550/arXiv.1402.1128

[19] Serizel, R., & Giuliani, D. (2015). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 135-140. https://doi.org/10.1109/SLT.2014.7078563

[20] Terriza, M., Navarro, J., Retuerta, I., Alfageme, N., San-Segundo, R., Kontaxakis, G., Garcia-Martin, E., Marijuan, P. C., & Panetsos, F. (2022). Use of Laughter for the Detection of Parkinson's Disease: Feasibility Study for Clinical Decision Support Systems, Based on Speech Recognition and Automatic Classification Techniques. *International Journal of Environmental Research and Public Health, 19(17),* 10884. https://doi.org/10.3390/ijerph191710884

[21] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 Conversational Speech Recognition System. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5934–5938. https://doi.org/10.1109/ICASSP.2018.8461870