Possibilities for applied joint speech processing and computer vision systems

E S Romeu, V I Syryamkin
National Research Tomsk State University, Tomsk, Russia
E-mail: edgarsolisromeu@gmail.com

Abstract. This report addresses the current state of the Computer Vision and the Speech Processing fields. It explains some of the most used algorithms in these fields and provides an overview of the current use of these technologies in the market. The objective of this report is to find new possible industrial and commercial applications that use both Computer Vision and Speech Processing.

1. Introduction
Now a day people live with gadgets and technology that would be unthinkable some 20 years ago like tablets, Smart watches, an ubiquitous internet with ever growing capabilities and many other things, but the truth is that this is only the beginning, and there are many more developments that are being worked on to this day. Speech processing and Computer Vision are two of the fields of computer science that are having huge advances, with many of them already part of modern life, and with many more still to come.

2. Speech Processing

2.1 General Aims of the Speech Processing Field
Computer Science has many fields and one of them is Speech Processing. The aim of this discipline is to create computer systems capable of recognizing spoken words. The main idea is to detect the words spoken by a person and process that data to identify what the person is saying [1].

2.2 Brief History of the Field
This field that has been in existence since the 18th century with Wolfgang Von Kempelen’s Acoustic Mechanical Machine. A quick review of some major breakthroughs in Speech Processing would include Edison’s dictation machine in 1879, a computer system names Audrey capable of recognizing spoken digits with 90% accuracy developed in 1952 at Bell Labs and IBM’s Tangora that could be adjusted for particular voices and recognized a wide array of English words and phrases. The speech recognition software of the 1980’s required the speaker to speak slowly with space between words and without background noises. In 1997, Dragon’s Naturally Speaking Software was released and it allowed for its user to speak complete phrases, instead of only words [2].

During the first phase of speech processing software development, these systems worked by translating the sound waves into numbers and storing them, then the speech recognition would be triggered when an identical sound was introduced as input. This method is generally known as Template Matching.
Eventually Machine Learning was introduced and it revolutionized the field. The main player that used ML to its advantage was google in 2008 with its Google Search App for Iphone. Eventually that company developed an algorithm called “Hummingbird” that allowed better use of language when it is used [2].

2.3 Connectionist Temporal Classification (CTC) Algorithm, How Does It Work
A common used algorithm for Speech Processing is the Connectionist Temporal Classification (CTC). Here is a fast and simple overview of how it works:

1. Transform the analog sound waves into digital information. To do this you record a word and you sample a tiny piece of a recording by marking the height of a sound wave at a specific point in time. The idea is that we end up with an array of numbers that represent a sound wave’s amplitude every 1/16000th of a second intervals [3].

2. The information is pre-processed, one good option is to group the sampled audio into 20 millisecond long pieces, when this information is plotted the resulting graph looks like a sound wave. Sound waves are formed by the combination of many waves. It is possible to deconstruct the complex sound wave into simple sound waves by adding the amount of energy that is contained in each, this is done through Fourier Transform [3].

3. The neural network is fed the data and it tries to find out the letter that is being spoken. It is a good idea to use a recurrent neural network, since they use memory to influence their predictions, this is useful at recognizing patterns in words and guessing which letters will be next. The result is a mapping of the most likely spoken letters. Eliminating possible spaces and ignoring the repeated characters can refine this output [3].

4. Using large databases that record the likelihood of a word appearing in a language, the neural network can be trained to predict the word that is being constructed. To do this, companies like Google and Apple record their users to build great collections of thousands of recordings of words to feed them to neural networks and get even better results. That is why they install their apps freely, to gain access to users [3].

Researchers working for Microsoft used Deep Neural Networks to create a Speech Recognition system that was tested against the Switchboard Speech Recognition task which is an evaluation to find out if these systems are up to the standards of the industry. The result was a Word Error Date of 6.3%, at the time of the test in 2016, this was the best result ever obtained on NIST 2000 Switchboard which made the system the most accurate system ever developed [4].

2.4 The State of Speech Processing in the Industry
To learn how speech recognition can be applied in the Industry let's take the example of Apple’s Siri. It uses microphones to detect speech and then transcribes the text using Automatic Speech Recognition. The text is then transformed into parsed text and gets evaluated locally in the device, if it can’t be done properly then Siri uses Cloud web services to query the request. Once Siri has executed the command and has the results, it provides them to the user either by text or by a verbal response [5].

These companies have more ambitious plans for these products. The idea is to use speech recognition software to push for devices in houses in what is called the Internet of Things. By doing this, a customer uses a search engine by only asking a question without having to open a computer and he can also do online sales using this technology.

Customers have responded to this business model and adopted the products mentioned. For example, Tech Crunch has published a report with data captured from May 2016 to May 2017. This report found that the demographics of a “super user”, a person that spends twice the amount of time with personal assistant technologies on a monthly basis than average, is a 52-year-old woman spending 1.5 hours per month with assistant apps. This is an interesting fact that shows the possibilities for market expansion of these speech based assistants, especially because the users demographics are not dominated by young tech savvy people [6].
It is also good to point out that it is projected that by the year 2020 half of all search engine queries will be done by speech, also on 2017, 7% of all US households had smart speakers (assistants like Alexa, Siri, etc), but by 2020 the percentage is projected to raise to 75%. This has convinced major companies to invest in this technology since more people are accepting the concept of communicating with businesses through smart speakers. We have to keep in mind that these are only the possibilities of Speech Recognition through the lens of retail [6].

According to market intelligence firm Tractica, Speech Recognition software market will grow to 5.4 billion dollars by 2025 from a 136 million USD in 2016. SR is also expected to drive an expansion of sales in professional services and hardware. Tractica estimates that by 2025 the market opportunity for SR will be of 22.3 billion. That is why it is important to protect Intellectual Property related to this sphere as well as assessing the costs for related products and services [7].

Researchers working for Microsoft used Deep Neural Networks to create a Speech Recognition system that was tested against the Switchboard Speech Recognition task which is an evaluation to find out if these systems are up to the standards of the industry. The result was a Word Error Date of 6.3%, at the time of the test in 2016, this was the best result ever obtained on NIST 2000 Switchboard which made the system the most accurate system ever developed [4].

Table 1 contains some interesting startups that are using this technology and their place in the market.

**Table 1. List of some Speech Processing Start Ups [8]**

| Company | Description | Advantages | Disadvantages | Business Plans and cost |
|---------|-------------|------------|---------------|-------------------------|
| Tetra   | Software that uses speech recognition to take notes on phone calls. | Synchronization with Google Calendar. Lets you organize the conversation using keywords, tag highlights, label speakers and creates summaries. | The process is not instant; it takes 10 minutes to generate the notes. Laws regarding recording business and private phone calls, the app notifies the parties involved that they are being recorded, as a way of dealing with it. | Free release with subscription plans and business support being introduced in the future. They have obtained 1.5 million USD in seed investment. |
| Alexa   | An intelligent assistant that uses speech recognition to do online searches. | Makes a good use of speech recognition to create an easy to use interface. | Concerns from privacy advocates that have issues with the way Amazon saves the voice recordings of the users for future neural training. | 15 billion USD in projected revenue in the next three years (2019 - 2021). |
| NeuroLex [9] | A startup founded in 2016 that is developing an app that analyses the voice of a patient for signs of depression, schizophrenia | This startup points the way to using processing language in the medical profession. | The technology is still in development. More factors and tests are needed to determine if a | They have invested $210,000 USD and are looking for 1.5 million USD for |
Ummo A speech coach application that helps its user to improve their oratory skills. The app analyses the filler words used and also the pace and volume of the speech, as well as pauses, clarity and accents. The app focuses on analyzing and presenting the data of the speech, but doesn’t present solutions for the users. Only works on Apple. It costs 1.99 USD in the Apple Store. It has a monthly download rate of 9049 per month.

3. Computer Vision

3.1 What Is the Aim of Computer Vision?
Computer Vision tries to create computer systems capable of reconstructing properties of images fed into the system mainly shapes, illuminations and color distributions [10].

3.2 Computer Vision Algorithms
There are a multitude of Computer Vision related algorithms and their effectiveness depends on the type of image that is being subjected to analysis. Histograms of Oriented Gradients (HOG) are a method to transform an image into data that can be more easily interpreted by a computer, by getting rid of information that is not important.

1. Turn the image being detected black and white.
2. Look at every pixel and focus on how dark it is compared to surrounding pixels and mark the gradient (draw an arrow on the direction the pixel is getting darker, gradients show the flow of from dark to light).
3. Break the image into small 16×16 pixel squares.
4. In each pixel count how many gradients point to a major direction.
5. Replace the squares in the images with an arrow showing the strongest arrow directions.
6. The result is a HOG image pattern.

HOG images are compared to other patterns that were acquired while training the neural network with other faces [11]. Figure 1 shows the result of transforming a color image into gradients [12].

![Input image](image1.png) ![Histogram of Oriented Gradients](image2.png)

**Figure 1.** Histograms of Oriented Gradients.
HOG can be used with other algorithms like Face Landmark Estimation to generate programs that recognize faces.

Face Landmark Estimation finds 68 specific landmarks that can be found in any face (eyes, chin, eyebrows) and maps them. The results of this algorithm are rotated, scaled and sheared, so as to present the landmarks as a 2d face viewed from the front. This is useful since many times people are turning their heads when photographed and for a computer it’s difficult to recognize the same face viewed from different angles [11]. An example of this is presented on Figure 2.

![Figure 2](image)

**Figure 2.** This image shows the specific landmarks detected by the machine [13].

Deep Convolutional Neural Networks are used to train computers to find useful 128 measurements to recognize each face, this is known as embedding. The process is done by loading two different pictures of a same person and then a picture of a different face, the algorithm looks at the measures it is currently creating for the images and tweaks the neural network so the measurements for the pictures of the same person are closer, than the ones of the other face [11].

### 3.3 Computer Vision Related Start Up

Table 2 shows a Computer Vision Start Ups.

| Company   | Description                                                                 | Advantages                                                                 | Disadvantages                                                                 | Costs                                          |
|-----------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------|-------------------------------------------------------------------------------|------------------------------------------------|
| Omnify    | Analyzes outdoor audiences and creates behavioral insights about the audience. | Applies computer vision for marketing and retail purposes.                | It is a new company, hasn’t even been around for a year so it yet has to prove itself on the market. | Founded on march 2018, it has gained 2 million USD on founding. |
| Blippar   | European company founded in 2011 that specializes in Augmented Reality      | They focus on applying AR to consumer experience with virtual try-ons, loyalty schemes, gamification, etc. | Technology hasn’t been adopted in a massive scale.                           | Has had 4 founding rounds the latest on august 2018. It has obtained 131.7 |
6

III International Conference "Cognitive Robotics"
IOP Conf. Series: Materials Science and Engineering 516 (2019) 012044
doi:10.1088/1757-899X/516/1/012044

6 million USD in investment.

Apprentice AR applied for laboratories and assembly lines. Founded in 2014 in New York City. Uses smart glasses to create an AR environment. Workers can follow instructions that are displayed to them on the glasses. The idea can be adapted to more industries. Has obtained 10.5 million USD in two rounds of investment.

Magicleap Founded in USA in 2011, offers wearables (special glasses) to create an AR platform. Developers create software programs that users access through the Magic Leap One specs. The software offers focuses on games, but it also has interior design software, an AR based internet browser, and a social network. It has yet to gain massive exposure and adoption. Has obtained 2.3 Billion USD in founding in 5 investment rounds.

4. Some Patents Examples
Table 3 shows some examples of international patents that make use of these technologies. Note that patents number WO 2014/107076 and number WO 2015/176950 use both technologies in their invention.

| Patent Holder | Description | Date of Publication | Number of Patent |
|---------------|-------------|---------------------|------------------|
| Vorwerk & Co. Interholding GMBH (Germany) | A speech recognition device for kitchen appliances, so that it receives commands. It also includes a device for optical detection. | 05.05.2015 | WO 2015/176950 |
| Ford Global Technologies, LLC. (United States) | Portable speech recognition device for vehicles. | 04.05.2017 | WO 2017/074328 |
| Samsung Electronics Co Ltd (South Korea) | Speech recognition system that recognizes user’s voice and commands and looks in its database if the commands match others saved in memory and if they do then follows the instructions related to the pre-established command. | 10.07.2014 | WO 2014/107076 |
| Samsung Electronics Co Ltd (South Korea) | A device that detects the gaze of its users and once detected launches an interaction interface where the user can speak commands to the system. | 24.04.2014 | WO 2014/061916 |
| Xiaomi Inc. (China) | Device that recognize user’s face and that prompts an internet fueled search for the user’s information. | 20.11.2015 | WO 2017/000486 |

5. Results
What type of products can be developed now? There are fields that seem to be ready for a product that uses both CV and SP. One is the entertainment field where products like videogames can be developed. An example could be a game where the user gives commands to the system and the game could detect the user’s face and access his data without the need of a password. That data could be stored in cloud services making easier user mobility between systems.

Other useful way these technologies could be implemented is in the field of marketing. Agencies could implement interactive interfaces that promote brands. Earlier Blippar was mentioned as a company already working in the VR realm, so its possible that they could implement speech recognition to their products.

Finally another obvious choice would be the field of Security. Devices that use voice and image patterns are being developed for the purpose of user access to systems as diverse as cars, computers, bank accounts and smartphones without the need of passwords or keys. As the technology develops these security solutions will be used more extensively and will become the most basic way of enforcing security for a device, building, archive, etc.

6. Summary
Speech Processing (SP) and Computer vision (CV) are two disciplines that have seen great advances in the last decade, specially thanks to the constant development of Machine Learning methods like the Neural Networks. Now a day these technologies are being adopted massively by companies and the public and are becoming ubiquitous in daily life.

This situation will only grow in the future as adoption grows and newer enterprises arise seeking to capitalize in the state of the technologies. New methods of Computer Recognition for both Audio and Image will be developed as Artificial Intelligence capabilities advance and more complex products become feasible.

Acknowledgments
The paper was supported by “The Tomsk State University competitiveness improvement programme” under grant (No 8.2.24.2018) and the Russian Foundation for Basic Research (grant No 16-29-04388).

The authors are grateful to Professor Stanislav V. Shidlovskiy and Tatiana B. Rumyantseva from Tomsk State University for their assistance in preparing the paper.

References
[1] Shikhman M., Shidlovskiy S. 2018 IOP Conference Series: Materials Science and Engineering 363 (1) doi: 10.1088/1757-899X/363/1/012019
[2] Abhang P.A., Gawali B.W. 2016 British Journal of Applied Science & Technology 10(5) 1–13 doi: 10.9734/BJAST/2015/19000
[3] Lawson S 2018 The Past, Present and Future of Speech Recognition Technology ClickZ URL: https://www.clickz.com/report-past-present-future-speech-recognition-technology/209060/ af
[4] Machine learning is Fun Part 6: How to do Speech Recognition with Deep Learning Medium URL: https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a
[5] Luxton D 2015 Artificial Intelligence in Behavioral and Mental Health Care (Academic Press)
[6] Szeliski R 2011 Computer Vision: Algorithms and Applications, Texts in Computer Science (Springer-Verlag London Limited) doi: 10.1007/978-1-84882-935-0_14
[7] Dalal N., Triggs B. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) doi: 10.1109/CVPR.2005.177
[8] Shidlovskiy S. 2016 Boolean Differentiation Equations Applicable in Reconfigurable Computational Medium MATEC Web of Conferences 79 doi:10.1051/matecconf/20167901014
[9] Kazemi V., Sullivan J. 2014 IEEE Conference on Computer Vision and Pattern Recognition doi:10.1109/CVPR.2014.241