SPEECH TECH SYSTEMS LLC

Speech Analysis Technology for the Tactical Environment

Many new technology approaches are attempting to eliminate the human element from the media analysis process. Others are so complex that they cannot be adequately fielded or used by non-speech scientist. For defense ISR based systems, we believe these other systems are heading in the wrong direction. So how do you improve efficiency and actually expand your capabilities to keep up with the ever growing volume of data and new speech processing requirements? Many technologies and concepts continue to show promise, but when put to the test in a rough operational environment they fail.

We want to show you a way to augment your human force with audio and voice processing technology to greatly enhance their ability and make it possible to deal with these huge volumes and produce the actionable intelligence that is needed. Automated speech processing tools are making advancements, but nothing works like the human mind. So the real key is magnifying your best linguist/analyst's ability. The problem is even more pronounced when dealing with low quality audio collections such as radio or poor cellular traffic and have limited size, weight, and power (SWAP) resources.

Most organizations are familiar with the concept of Speech-To-Text (STT) type processing where voice can be converted to text and then the text can be mined for words of interested. This can be a good solution for broadcast quality with highly educated speakers. However, in a non-pristine environment against people who don't speak eloquently, the basic problem with those systems are that they make hard decisions and output on matches to their dictionary or based on grammatical rules. Their engine hears a term (rightly or wrongly) and that term is essentially written in stone in a transcript file. That transcript file is then used for every subsequent process. There are some positives here, but there are as many or more problems with this approach. These systems aren't robust to dialect variations and noisy environments. They need a really good and relevant dictionary of terms to produce a decent output. They don't easily support numerous languages as well as deal with any usual grammatical deviations (human speech does not always follow the rules). They are not set up to deal with short communication used with radio type communication. Moreover they usually require huge amounts of processing power so they are not always practical for forward deployed groups. Lastly, they have the simple problem of garbage in produces garbage out.

The fact of the matter is that dealing with voice processing, especially in this type of environment, is really difficult. Audio and voice processing problems are compounded with the numerous languages and dialects that are frequent in the areas of interest. This makes the language specialists who understand the mission and the languages very unique. Any system that is deployed must be easily adaptable, flexible, and robust enough to handle high volumes, low quality, language/dialectical variation, lightweight (with regards to SWAP) and provide almost immediate results and value to the end users.

The Speech Tech approach would include using a new Phonetic-based technology to solve voice processing requirements. There is a fundamental difference in our approach versus any other speech processing system which is that we keep the audio in its raw form and do not try to convert the audio from speech to text (STT). We create a phonetic index file that points to every phoneme (short sound) in the file. This index file can then be used to search at extremely fast speeds. We also quickly realized that having a phonetic index was beneficial with other speech processing requirements like language identification and speaker identification. No hard decisions, like writing to a transcript file, are made so a misidentification can quickly be resolved by adjusting the search parameters. We refer to it as 'indefinite depth of search', so a user has the power to immediately change the criteria until they are able to find what they are looking for. The approach is pronunciation specific so there are capabilities that can be included to assist the users in finding optimal pronunciation of terms/phrases to help them be successful. After understanding the importance of pronunciation, this led us down the path to see how individual speaker processing (like speaker ID) was possible. A large system could process millions of hours of audio and video files, but small tablet based (disconnected) systems are fast enough to handle hundreds of simultaneous feeds in real time on a windows based platform that could easily fit into a man-pack.

All the processing with this system starts with the ability to create a phonetic index of the audio. This process is so fast that it can keep up in real time and provide the ability to monitor against streaming media (multiple feeds if necessary). Searching for keywords and phrases is one of the more basic but powerful use cases, however, advanced Boolean operators can be used to build complex search queries. It is possible to include operators like ANDs, ORs, ANDNOTs, etc. to include and exclude specifics topics of interest. Accuracy can be very high, but it is completely within the control of the user who is conducting the searches. The operators are able to easily adjust the threshold on the searches so they can control exactly what is being returned back to them. This is essential when working in noisy environments or with new target speakers who language or dialect may be different than expected. Opening the results up to more false positives might be required to ensure nothing is being missed, but the penalty for doing this is only a few extra seconds of time to weed out the false positives. In harsher acoustic environments this type of flexibility is essential.

Usual search capabilities are specific to a particular language because of pronunciation and input techniques; however, it is possible to have a language agnostic solution. So in certain cases you can employ searches against any spoken language using a multilingual language pack. However, when possible, it is better to determine the language and dialect that would also provide additional information to the analyst and the system. Each language has a unique phonetic alphabet and this technology can conduct a language identification (LID) process to determine the spoken language. This approach can perform LID on any and every language and provide a percentage confidence level back that the given language is a match.

In a very similar fashion to LID, it is possible to do speaker identification (SID) with this technology. LID involves building and comparing language models that are used to compare how close a match they are to other language models. SID uses the same concept, but rather than comparing speech collected from numerous speakers to create a robust language model, it is built on a single speaker with as much examples as possible. Pronunciation, accent, educational background, environment, etc. these are all distinct influences that make a particular speaker's voice unique. It is similar to the way a biometric system uses fingerprint minutia, this system builds and weights the speech for comparison against all other speaker models. We have also had groups pursue a type of group identifier where they are looking to correlate particular groups of speakers with common speech attributes (to help identify them against all other groups). The idea is to try to filter and narrow the amount of data that the analysts and linguists need to deal with. If the intelligence is in the collected traffic, the technology provides the users many avenues to be able to find what they are looking for AND be able to find it in time where it is still relevant to the mission.

It is very possible to search and find thousands of individual terms/phrases, so for a given mission topic. Using those terms, it is very conceivable to automatically produce a translated "gist" of a conversation. This approach could provide enough value to the analysts to determine if a more thorough verbatim translation were required.

It is also possible to discern sentiment using a voice processing system, but based on our research, we believe more emphasis should be place on the context of the speech rather than based on the specific acoustic variations of the speaker. Language and culture play a crucial part and without that predetermined, it makes a difficult task almost impossible. Building proper search queries are essential to determining the sentiment of the speaker(s). With the understanding of the mission and the target language, an analyst could quickly build multiple buckets of reviewable information in ANY language to understand the threat level based on the sentiment of the voice files.

Target groups may have a unique way of communicating that basically is like a unique language of its own. This is one of the many reasons that intelligence gathering systems need to keep a human element heavily involved where it is possible for them to learn and understand this "language" over time. It will still be a while until automated tools are able to handle all the complexity of language and deal with cover and threat terms that may only be relevant to some people who understand this exclusive language. There are no system that could associate "peanut butter and jelly" spoken in a foreign language for being a code for jihad. The human mind can make that connection. The technology we are suggesting can empower the users to wield massive amounts of audio data and provide more manageable buckets for more in-depth analysis.

The technology has a flexible architecture which allows for integration into existing systems or formation of a complete stand-alone operation. There are application programming interfaces (APIs) available to make it more possible to set up to pull any media format and quickly identify the language, look for matching voice files, and then search for a list of threat or cover terms. As far as media exploitation, it is very possible to build a process to connect to or plug directly into any target system, start a routine to seek any type of media, run Language ID, Speech ID, search for numerous topics of interest in multiple languages and provide translation of critical topics to English all within a matter of seconds.

If there is a desire is for a system to assist with speech processing requirements that can support the mission, is flexible enough to work in SOF environments, and is accurate and fast enough to produce real value then we have a good solution for you. It could be plugged into your existing platforms and would not add any additional SWAP requirements. The solution would maximize operator efficiency which should reduce overall costs of any SIGINT voice processing. The ultimate goal being to multiply the value of your existing language specialists, making their life's better/easier, as well as greatly increasing their capabilities.