Empower your Operators with Language Agnostic Speech Classifier

The Holy Grail of speech processing would be a technology that can filter, process, translate and provide verbatim English output at scale and in real time against lower quality data, which has yet to arrive... Operators need a real world tool that can assist them in filtering on Language/Dialects of interest, on individual groups of speakers, or even specific speakers themselves. If those condensed "buckets" of information could be provided quickly and accurately, language specialists would be able to focus on mining and translating just the most relevant mission specific topics.

It is possible to combine the power of audio Language ID, Speaker ID, and Group ID into one dominant tool. The approach we are discussing here offers an easy to use process that can provide analysts and linguists a better way to manage and identify language/dialects, speakers, or groups of speakers. This is all possible with a simple to use tool that can also allow users to train relevant new models in minutes on site with no network connectivity on standard Windows based laptop machines. The process would allow the end users to categorize and prioritize the information that is most relevant to their particular mission. It can also help include/exclude things based on relevance. Other approaches could take days or more, need massive processing power, and tie up crucial linguist resources for extended periods of time.

If identifying languages/dialects, speakers, and/or groups are important to your mission; this concept needs to be considered. It is language agnostic and therefore can be built for use on any language/dialect that exist today with no additional costs or technology changes. With this system, a user can define exactly the type of identifier they want to have. Whether they are looking for a particular language or dialect being spoken or want to separate out all languages that don't meet their defined criteria, this capability is built to adapt to changing mission needs.

The technology will operate as good as the training language model that has defined it. The more robust and representative the training model is the better and more specific the output will be. There are systems in broadcast environments that produce decent results against that high quality commercial traffic. This system, however, is intended to be used against massive amounts of real world noisy low quality radio and cellular traffic.

The analogy we like to use for a Language Model is a phonetic DNA for the speakers. In the case of an individual speaker, it definitely can be thought of as a one-to-one attempt to match a speaker's unique voice and pronunciation. Language Models trained on a group of speakers is similar in nature, but includes a larger phonetic alphabet and similarities in pronunciation and terminology. Using the DNA analogy, it would be how grandparents share some DNA traits with their grandchildren. During the model build process, the engine basically creates a phonetic language defined by the characteristics of the training set, i.e. similar to how cousins share DNA traits. The training data should be as representative as possible. Depending on the amount of data, the process should only take a few

minutes to complete. Any subsequent collected audio file could be compared to the newly defined "language model" and output a confidence level of the match. The DNA analogy continues as a much broader Language and Dialects Models which can be thought of as huge collection of further distant relatives that use similar terminology and pronunciation. At the end of the day, the Language Models (one or many) are used to compare the differences/similarities to the test file being processed.

This type of concept has been attempted for years. The fundamental problem with all other approaches is that they try to convert the speech into text and then do advanced text analytics and word entity extraction. The problem is further complicated because those approaches try to translate the speech into English. This is not what is being done with this approach. The only real way to address this problem today is with human force, i.e. adding more trained and knowledgeable linguists and analysts to assist with the filtering and/or base the determination of language, group, or speaker on their knowledge. Some use sensor provided physical location metadata or zip code, which is not a very good determination. Human force doesn't scale well and metadata is extremely unreliable and neither of these provides much value for Group Identification. The approach being proposed here uses a different technology and process altogether. There is no conversion to text, the audio is kept in its native format and the model being processed is a phoneme-based model.

Most systems try to generalize language so that it fits into a particular dictionary and then use that information to define the language. This tool takes a completely different approach as well as uses pronunciation variations to highlight the variances in the spoken language. Since the models are built specifically on the dialect, speaker, or group of interest, the engine is looking for differences. More training data is typically better, but the key is that you only need to train on a few minutes of representative data to start providing value. There is no dictionary to maintain. Once the model is built it is easy to continue to improve if desired. The output can include a confidence level to help determine how certain the engine is on a match (low confidence outputs can still always be evaluated by a human ear to verify).

The solution is based on years of phonetic research. The phoneme is an extremely small unit of sound in speech; it is used to distinguish one word from another. Recently a customer requested support for Igbo (a language in Nigeria). This was a language that was unfamiliar to us as well as the end user. Based on a few minutes of collected audio, they were able to quickly build an Igbo Language Model which then assisted them in filtering Igbo from all other African languages. This approach has been tested on many lower density languages and shown extremely impressive results. This approach does not require any dictionary maintenance, re-indexing, or expert assistance to work. That process to convert the audio to text and compare the text is much slower and can remove much of the uniqueness needed to properly identify the language, speaker, or group. This approach relies heavily on pronunciation specific terminology and does not require that conversion to text (that can always be done after this filtering process has narrowed the files to only those that need to be converted).

For the technology to work optimally there is some upfront effort required. Mainly that one of the users will need to have knowledge of the language and particular mission involved to vet the data to be ingested so the model is not tainted with incorrect data. After initially defining and building a few models, the models can then help separate the new files to further enhance the existing models. We will be able to provide initial training language models for many languages, but since they are not built on end user specific operational data, they should be used only for the initial filtering process. Further care should be taken when building specific speaker or dialect models because of the similarities of sounds of individual voices (particularly in noisy poor quality environment). Human ear is still the ultimate voice and language discriminator. This tool can make the human ear more productive, but not replace it. Although the initial training phase is somewhat involved, the output is intended to be used by standard operators without any linguistic or audio science background.

Although the user requires some knowledge and understanding of the language, the training process is quite simple. The user only needs to separate the training files into a directory and point the tool to the files and click 'Go'. That starts the process to build a phonetic-based Language Model that is unique to that training data. So if it is trained on one speaker, it is good on identifying that one speaker. If it is trained on one group of particular speakers, it will be good at identifying that particular group (studies have shown that a group that is together for a while starts to form their own "language"). Once the audio is separated, the actual computer time needed to build the Language Model is only a matter of minutes, i.e. building a new language model with a few hours of audio data on a standard Windows based laptop with two cores takes less than 5 minutes).

Audio processing and analytics is a difficult problem and is made even worse because of the quantity and quality of the audio, lack of language resources, and the speed at which information needs to be provided to the operators. Furthermore, if the technology is to be used in forward deployed environments in any type of force protection scenario, it needs to be something that is flexible and easy to use. Our system makes a complex problem much simpler by using this novel approach which removes much of the complexity. The technology addressed all these difficult environmental factors and provides more efficiency and productivity out of the limited language specialists. It empowers the operators, whether they are forward or stateside to focus only on filtered mission and language specific relevant output.

Field testing is ongoing, contacts available upon request.