An Improved Approach to Automatic Speech Recognition – FlexSR Whitepaper

SF_01-01-24

Background

To understand the reasons why FlexSR has been developed, it is useful to broadly understand some of the history behind it.

It was as early as the 1950's that it was postulated that you could recognise linguistic features in a speech signal power spectrum and match then to the phoneme representation in words. Back then, the systems required tight adherence to a very narrow and specific rules based matching, which lacked the more flexible identification of the representation of words that the human brain adopts, where it tolerates imperfect or missing bits in the signal. This created disillusionment in these systems. Because of this, the following early systems adopted a statistical approach with decision trees and models to try and accommodate the variables, but required a lot of compute processing power. The advent of greater compute power facilitated the application more effective algorithms and use of Neural Networks in the 2000's, to improve the predicted outputs. This biased the exercise to computer science and away from the linguistic science approach, including approaches seen in more recent times, that create bigger models, requiring greater compute, to look at the entire end to end sequence of words in speech, rather that the individual sounds (phonemes) in words. Inevitably, the success of models is directly proportional to the volume of domain specific data, containing the words of interest used to train the model.

Conventional Speech Recognition requiring Acoustic Models or End to End Models

Conventionally, when a system is trained on exemplars of real speech, the system is biased to these exemplars as the standard for its trained models. In operation, the system then tries to identify which information in any given signal is relevant to its predictions on a sequence of words that have the highest probability against its trained model. But what happens when the inputs containing non-standard speech, less common word sequences, or dialect variations? It will offer a sequence of words that best match its training bias, but words may be entirely missed out, or replaced with substitutes.

The conventional resolution of the problems these variations bring, is to retrain the model, using more examples of speech for each variation, that has been manually transcribed for the training and benchmark testing. The system would then be required to apply each variation trained model, as the variations are known to occur.

The training of a required acoustic models to accommodate these variations, requires a large amount of speech, (*Nexidia have suggested this should be 250+ hours). It's never an easy task, to even collect this amount of speech, containing enough of the domain relevant variation, across a good cross section of speakers, to train or retrain a model well enough. It relies on human, manual transcription to enable the model to be trained. Unless expert in the domain and variation, human transcription can introduce its own errors. Typically, the transcription of one hour of speech into the training format will take between 4 and 10 hours depending upon complexity and ability for the

human to concentrate. This means 250 hours typically requires more than 150 man days. Training and testing time is added to this, as a number of additional man weeks.

The FlexSR alternative to Conventional Speech Recognition

The study of human speech and linguistics by the FlexSR esteemed professors, has resulted in a deeper understanding of how the human brain is so effective at recognising the sounds in words and phrases in speech, and matching them to its own mental dictionary (or Lexicon) of words and phrases. Their studies carried out their fundamental investigation of variations in speech, led to a flexible linguistic model of speech based on phonological features, the articulatory and acoustic properties of each sound that form its contrasts with others. For example, the 'voicing' feature (whether the vocal cords are vibrating or not) forms a component of the contrast between the 'p' and 'b' consonant sounds in English. The team developed the speech recognition system, now known as FlexSR, that was trained to recognise a universal set of 19 such features and could combine them to identify speech sounds, or phones. Importantly, it targets those features, essential to human understanding of speech, and ignores or tolerates those that can vary across speakers or utterances. Most importantly, the FlexSR system model does understand the linguistic dependencies and rules, that sequences of sounds must obey, but can allow for certain information to be missing or faulty in the signal, much like the human brain. This overcomes the limitations of those early linguistic approaches that put them out of favour. This is the flexible aspect of FlexSR.

FlexSR Speech Recognition applied to any chosen Language

As FlexSR is based on an acoustic description of (abstract) features, which are part of a pre-defined lexicon; there is no need to 'learn' the mapping of sounds to word forms. In fact, no training set consisting of large corpora of speech material is required at all. To implement FlexSR to a new language, we require a lexicon with standard phonetic description, and this would be simple for all languages which have published phonological analyses. No new acoustic analyses are required, but depending on what sounds exist in the language, the choice from the 15 phonological features commonly used has to be adjusted. For instance, German may require all 15 features, but Dutch could need only 14. This can be performed with little human or computer resources and can be achieved in a short time. Many world languages are published and available off the shelf in the IPA format which satisfies the phonological analysis, with a complete set of Lexicons. New Lexicons can be easily added, ad-hoc to suit required domain or situational variations. A big differentiator over the model training requirements of convention systems of Hybrid (Acoustic Model plus Language Model) or End to End Models.

Handling Speaker Pronunciation Variations

The problems with Non Native speaker pronunciation on Speech Recognition systems

For example, the non-native pronunciations of English result from the common linguistic phenomenon in which non-native speakers of any language tend to transfer the intonation, phonological processes and pronunciation rules of their first language into their English speech. They may also create innovative pronunciations not found in the speaker's native language. The sound pattern of the learner's first language is transferred into the second language and is likely to cause foreign accents. The mispronunciations of words by nonnative speakers reflect the influence of the

sounds, rules, stress, and intonation of their native language. The main problem that second language learners have with pronunciation has to do with their need to change a conceptual pattern appropriate for their first language that they have internalized in childhood. All of this can cause a pronunciation deviation from the ideal way a word should be pronounced in a non-native language

or by a speaker with a strong regional pronunciation that varies from the ideal. The deviation can lead to errors in word and speech recognition.

FlexSR accommodating any Non-Native or Regional Variational Pronunciation

The FlexSR research team has also used this novel system to develop a language learning application. which analyses words and sentences spoken by the user, and provides detailed feedback. It can be used so language learners can receive personalised responses to improve their pronunciation. It can also be used to profile the speaker variance from the ideal standard pronunciation and apply that to the FlexSR system to enhance the matching of the features identified in a speech signal, to the chosen Lexicon. Notably this is achieved without the need to build or train acoustic models as required by conventional systems.

Definition of Speech Recognition Accuracy

The measure generally used to evaluate speech-to-text systems is the word error rate "WER"). It consists in adding all the word errors, namely substitutions, insertions and deletions of words, and divide by the total number of words in the reference transcription: WER = (S + I + D) / N

A WER of 10% corresponds to a recognition rate of 90%. This metric is not perfect, since it gives the same importance to all the errors, while some have much more impact than others with regard to the intended application, but it has the merit of being rigorous and standard, allowing comparisons over time and between systems.

The WER is tested against a test data set that represents the language/accents/domain of interest should be kept separate from any training data that may have been used for a system using trained models.

By their nature, models introduce a bias to predict an expected sequence and content, to its most likely probability. Even when an WER regarded as close to a human at 4% against test data set, it can still be missing the most important target words or phrases, yet make the output appear highly credible and read well. For many years a published corpus of switchboard data set, based on 260 hours of US English language switchboard conversations, had been used to compare WER of systems.

So, claims of broad WER accuracy are not the metric of interest with FlexSR, where it is the specific accuracy and dependability, in finding the important target words and phrases of interest in a speech signal, that is most important.

Conclusion

FlexSR evolved from the Linguistic understanding of speech and how a human makes the sounds, to be able to describe them and identify them in a speech signal in a flexible manner, much closer to a

human brain, with minimal information. Whilst other conventional systems rely on modelling as many examples of speech, relating to the domain of interest as possible, and to predict with a level of probability, using large amount of compute to run the algorithms. Both methodologies are building out vocabularies, but a FlexSR vocabulary can be built and added to, on an easy ad-hoc basis, without the need for model training and retraining, as required in conventional systems. Conventional Automatic Speech Recognition (ASR) has been in wide use, and where its model

accurately represent the domains of interest in the languages and pronunciation required, well enough for the task, they are serving their purpose. The area of application that FlexSR addresses better, is where the ASR faces variations of speech to conventional ASR built models, in pronunciation and words or phrases that would benefit from a rapid accommodation of these variations, without training or re-training models. This is why the Professors Aditi Lahiri and Professor Henning Reetz researched the better use of linguistic theory, published a number of peer reviewed papers, built a number of proof of concepts and invented novel systems that had World-wide Patents granted between 2015 and 2018. And this is why FlexSR has a special place and role to play in the world of Automatic Speech Recognition.

Accessing FlexSR technology

FlexSR novel technology has world-wide Patents granted for; The novel inventions for Automatic Speech Recognition and a System for Automatic Speech Analysis, under the International Patent Cooperation Treaty World Intellectual Property Organisation. Is brought to market as licensable technology by FlexSR Limited, a spinout company from the University of Oxford, to OEM's, SI's and VAR's.