Speech Recognition in Machine Learning using Google Speech API and Problems related to it

Abhimanyu olla

VIT University, Vellore

Abstract- In this paper, a brief implementation of Automatic Speech Recognition (ASR) using the Googles speech API is done and further I will elaborate on some of the difficulties with Automatic Speech Recognition. I will argue that the main motivation for ASR is efficient interfaces to computers, and for the interfaces to be truly useful it should provide coverage for a large group of users.

I will discuss some of the issues that made the recognition of single speaker difficult and then extend the discussion with problems that occur when we target more than a single user.

1. INTRODUCTION

Speech Recognition is an import feature in several applications used such as home automation, artificial intelligence, etc. Far from being fad, the overwhelming success of speech-enabled products like Amazon Alexa has proven that some degree of speech support is essential aspect of household tech for foreseeable future. If you think about it, the reasons why are pretty obvious. This paper aims to provide an introduction on how to make use of Speech Recognition in Python and further point out where it lacks. Incorporating Speech Recognition in your python application offers a level of interactivity and accessibility that few technologies can match. The accessibility improvements alone are worth considering. Speech Recognition allows the elderly and the physically and visually impaired to interact with state-of-the-art products and services quickly and naturally- no GPU needed. It can be used on microcontrollers such as Raspberry Pi's with the help of an external microphone.

2. HOW IT WORKS

Before we get to the nitty-gritty of doing speech recognition in Python, let's take a moment to talk about how speech recognition works. Speech recognition has its roots in research done at Bell Labs in the early 1950s. Early systems were limited to a single speaker and had limited vocabularies of about a dozen words. Modern speech recognition systems have come a long way since their ancient counterparts. They can recognize speech from multiple speakers and have enormous vocabularies in numerous languages.

The first component of speech recognition is, of course, speech. Speech must be converted from physical sound to an electrical signal with a microphone, and then to digital data with an analog-to-digital converter. Once digitized, several models can be used to transcribe the audio to text. Most modern speech recognition systems rely on what is known as a Hidden Markov Model (HMM). This approach works on the assumption that a speech signal, when viewed on a short enough timescale (say, ten milliseconds), can be reasonably approximated as a stationary process—that is, a process in which statistical properties do not change over time. In a typical HMM, the speech signal is divided into 10-millisecond fragments.

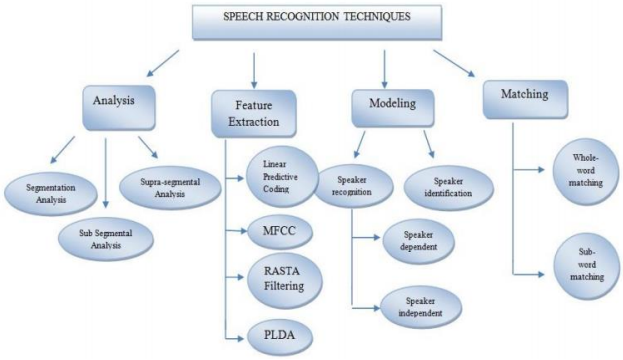
The power spectrum of each fragment, which is essentially a plot of the signal's power as a function of frequency, is mapped to a vector of real numbers known as cepstral coefficients. The dimension of this vector is usually small—sometimes as low as 10, although more accurate systems may have dimension 32 or more. The final output of the HMM is a sequence of these vectors.

To decode the speech into text, groups of vectors are matched to one or more phonemes—a fundamental unit of speech. This calculation requires training, since the sound of a phoneme varies from speaker to speaker, and even varies from one utterance to another by the same speaker. A special algorithm is then applied to determine the most likely word (or words) that produce the given sequence of phonemes.

One can imagine that this whole process may be computationally expensive. In many modern speech recognition systems, neural networks are used to simplify the speech signal using techniques for feature transformation and dimensionality reduction before HMM recognition. Voice activity detectors (VADs) are also used to reduce an audio signal to only the portions that are likely to contain speech. This prevents the recognizer from wasting time analysing unnecessary parts of the signal.

3. SPEECH RECOGNITION TECHNIQUES

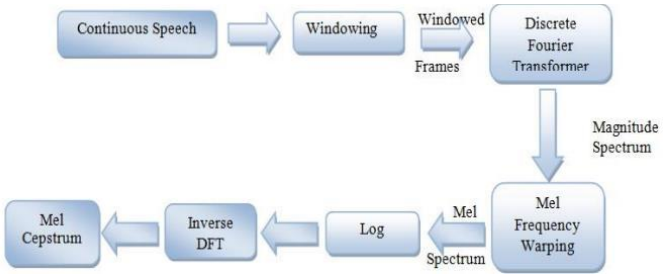
The main objective of a speech recognition system is to have capacity to listen, understand and then after act on the spoken information. A speech recognition system includes four main stages which are further classified as shown in the figure below.



3.1 Analysis: The first stage is analysis. When the speaker speaks, the speech includes different types of information that help to identify a speaker. The information is different because of the vocal tract, the source of excitation as well as the behaviour feature. The speech analysis stage can be further classified into three analyses:

a) Segmentation Analysis: In segmentation analysis, the testing to extort the information of speaker is done by utilizing the frame size as well as the shift which is in between 10 to 30 milliseconds (ms) [Range].
b) Sub-segmental Analysis: In this analysis technique, the testing to extract the information of speaker is done by utilizing the frame size as well as the shift which is in between 3 to 5 milliseconds (ms) [Range]. The features of excitation state are analysed and extracted by using this technique.
c) Supra-segmental Analysis: In Supra-segmental analysis, the analysis to extract the behaviour features of the speaker is done by utilizing the frame size as well as the shift size that ranges in between 50 to 200 milliseconds.

3.2 Feature Extraction Technique: Feature extraction is the main part of the speech recognition system. It is considered as the heart of the system. The work of this is to extract those features from the input speech (signal) that help the system in identifying the speaker. Feature extraction compresses the magnitude of the input signal (vector) without causing any harm to the power of speech signal. There are many feature extraction techniques.



The above figure is the feature extraction diagram. In this, from one side we input the continuous speech signals for the process of windowing. In the process of windowing the disruptions which are present at the start as well as at the end of the frame are minimized. After this process, the continuous speech signal is converted into windowed frames. These windowed frames are passed into the discrete Fourier transformer which converts the windowed frames into magnitude spectrum. Now in the next step, spectral analysis is done with a fixed resolution along a subjective frequency scale that is the Mel-frequency scale which produces a Mel-spectrum. This spectrum is then passed to Log and then to inverse of discrete Fourier transform which produces the final result as Mel-Cepstrum. The Mel-Cepstrum consists of the features that are required for speaker identification. A few feature extraction techniques include:

a) Linear Predictive coding: LPC is a tool which is used for speech processing. LPC is based on an assumption: In a series of speech samples, we can make a prediction of the nth sample which can be represented by summing up the target signal's previous samples (k). The production of an inverse filter should be done so that is corresponds to the formant regions of the speech samples. Thus, the application of these filters into the samples is the LPC process.
b) Mel-frequency cepstrum (MFFCs): Mel Frequency Cepstral Coefficients are based on the known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. The main purpose of the MFCC processor is to copy the behaviour of human ears.
c) RASTA filtering: RASTA is short for Relative Spectral. It is a technique which is used to enhance the speech when recorded in a noisy environment. The time trajectories of the representations of the speech signals are band pass filtered in RASTA. Initially, it was just used to lessen the impact of noise in speech signal but know it is also used to directly enhance the signal.

d) Probabilistic Linear Discriminate Analysis (PLDA): This technique is an extension for linear probabilistic analysis (LDA). Initially this technique was used for face recognition but now it is used for speech recognition.

4. MODELING TECHNIQUES

The goal of the modelling techniques is to produce speaker models by making use of the features extracted (feature vector). the modelling techniques are further categorized into speaker recognition & identification. Speaker recognition can be further classified into speaker dependent and speaker independent. Speaker identification is a process in which the system is able to identify who the speaker is on the basis of the extracted information from the speech signal. In speech recognition process we can use the following modelling approaches:

a) **Acoustic-Phonetic approach:** The basic principle that this approach follows is identifying the speech signals and then providing these speech signals with apt labels to these signals. Thus, the acoustic phonetic approach postulates that there exists finite number of phonemes of a language which can be commonly described by acoustic properties.
b) **Pattern recognition approach:** It involves two steps: Pattern Comparison and Pattern Training. It is further classified into Template Based and Stochastic approach. This approach makes use of robust mathematical formulas and develops speech pattern representations.
c) **Dynamic Time Warping (DTW):** DTW is an algorithm which measures whether two of the sequences are similar that vary in time or even in speed. A good ASR system should be able to handle the different speeds of different speakers and the DTW algorithm helps with that. It helps in finding similarities in two given data keeping in mind the various constraints involved.
d) **Artificial Intelligence Approach (AI):** In this approach, the procedure of recognition is developed in the same way as a person thinks, evaluates (or analyses) and thereafter makes a decision on the basis of uniform acoustic features. This approach is the combination of acoustic phonetic approach and pattern approach.

5. MATCHING TECHNIQUES

The word that has been detected is used by the engine of speech recognizer to a word that is already known by making use of one of the following techniques:

a) **Sub word matching:** Phonemes are looked up by the search engine on which the system later performs pattern recognition. These phonemes are the sub words thus the name sub word matching. The storage that is required by this technique is in the range 5 to 20 bytes per word which is much less in comparison to whole word matching but it takes a large amount of processing.
b) **Whole word matching:** In this matching technique there exists a pre-recorded template of a particular word according to which the search engine matches the input signal. The processing that this technique takes is less in comparison to sub word matching. A disadvantage that this technique has is that we need to record each and every word that is to be recognized beforehand in order for the system to recognize it and thus it can only be used when we know the vocabulary of recognition beforehand. Also, these templates need storage that ranges from 50 bytes to 512 bytes per word which very large as compared to sub word matching technique.

6. DIFFICULTIES

6.1 Human comprehensive compared to ASR

Humans use more than their ears when listening, they use the knowledge they have about the speaker and the subject. Words are not arbitrarily sequenced together, there is a grammatical structure and redundancy that humans use to predict words not yet spoken. Furthermore, idioms and how we 'usually' say things makes prediction even easier. In ASR we only have the speech signal. We can of course construct a model for the grammatical structure and use some kind of statistical model to improve prediction, but there is still the problem of how to model world knowledge, the knowledge of the speaker and encyclopaedic knowledge. We can, of course, not model world knowledge exhaustively, but an interesting question is how much we actually need in the ASR to measure up to human comprehension.

6.2 Noise

Speech is uttered in an environment of sounds, a clock ticking, a computer humming, a radio playing somewhere down the corridor, another human speaker in the background etc. This is usually called noise, i.e., unwanted information in the speech signal. In ASR we have to identify and filter out these noises from the speech signal.

Another kind of noise is the echo effect, which is the speech signal bounced on some surrounding object, and that arrives in the microphone a few milliseconds later. If the place in which the speech signal has been produced is strongly echoing, then this may give raise to a phenomenon called reverberation, which may last even as long as seconds.

6.3 Spoken Language is not the same as written language

Spoken language has for many years been viewed just as a less complicated version of written language, with the main difference that spoken language is grammatically less complex and that humans make more *performance errors* while speaking. However, it has become clear in the last few years that spoken language is essentially different from written language. In ASR, we have to identify and address these differences.

Written communication is usually a *one-way communication*, but speech is dialogue-oriented. In a dialogue, we give feed-back to signal that we understand, we negotiate about the meaning of words, we adapt to the receiver etc.

Another important issue is *disfluences* in speech, e.g. normal speech is filled with hesitations, repetitions, changes of subject in the middle of an utterance, slips of the tongue etc. A human listener does usually not even notice the disfluences, and this kind of behaviour has to be modelled by the ASR system.

Another issue that has to be identified, is that the grammaticality of spoken language is quite different to written language at many different levels. Some differences being:

- In spoken language, there is often a radical reduction of morphemes and words in pronunciation.
- The frequencies of words, collocations and grammatical constructions are highly different between spoken and written language.
- The grammar and semantics of spoken language is also significantly different from that of written language; 30-40% of all utterances consist of short utterances of 1-2-3 words with no predicative verb.

This list can be made even longer. The important point is that we cannot view speech as the written language turned into a speech signal, it is fundamentally different, and must be treated as such.

6.4 Continuous Speech

Speech has no natural pauses between the word boundaries, the pauses mainly appear on a syntactic level, such as after a phrase or a sentence. This introduces a difficult problem for speech recognition — how should we translate a waveform into a sequence of words? After a first stage of recognition into phones and phone categories, we have to group them into words. Even if we disregard word boundary ambiguity, this is still a difficult problem. One way to simplify this process is to give clear pauses between the words. This works for short command-like communication, but as the possible length of utterances increases, clear pauses get cumbersome and inefficient.

6.5 Channel variability

One aspect of variability is the context were the acoustic wave is uttered. Here we have the problem with noise that changes over time, and different kinds of microphones and everything else that effects the content of the acoustic wave from the speaker to the discrete representation in a computer. This phenomenon is called *channel variability*.

6.6 Speaker variability

All speakers have their special voices, due to their unique physical body and personality. The voice is not only different between speakers, there are also wide variations within one specific speaker.

Some of these variations are listed below in subsections.

6.6.1 Realization

If the same words were pronounced over and over again, the resulting speech signal would never look exactly the same. Even if the speaker tries to sound exactly the same, there will always be some small differences in the acoustic wave you produce. The realization of speech changes over time.

6.6.2 Speaking style

All humans speak differently, it is a way of expressing their personality. Not only do they use a personal vocabulary, they have a unique way to pronounce and emphasize. The speaking style also varies in different situations, we do not speak in the same way in the bank, as with our parents, or with our friends. Humans also communicate their emotions via speech. We speak differently when we are happy, sad, frustrated, stressed, disappointed, defensive etc. If we are sad, we may drop our voice and speak more slowly, and if we are frustrated, we may speak with a more strained voice.

6.6.3 The sex of the speaker

Men and women have different voices, and the main reason to this is that women have in general shorter vocal tract than men. The fundamental tone of women's voices is roughly two times higher than men because of this difference.

6.6.4 Anatomy of vocal tract

Every speaker has his/her unique physical attributes, and this affects his/her speech. The shape and length of the vocal cords, the formation of the cavities, the size of the lungs etc. These attributes change over time, e.g. depending on the health or the age of the speaker.

6.6.5 Speed of speech

We speak in different modes of speed, at different times. If we are stressed, we tend to speak faster, and if we are tired, the speed tends to decrease. We also speak in different speeds if we talk about something known or something unknown.

6.6.6 Regional and social dialects

Dialects are group related variation within a language.

- *Regional dialect* involves features of pronunciation, vocabulary and grammar which differ according to the geographical area the speaker come from.
- *Social dialect* is distinguished by features of pronunciation, vocabulary and grammar according to the social group of the speaker.

In many cases, we may be forced to consider dialects as 'another language' in ASR, due to the large differences between two dialects.

6.7 Amount of data and search space

Communication with a computer via a microphone induces a large amount of speech data every second. This has to be matched to group of phones (monophones/diphones/triphones), the sounds, the words and the sentences. Groups of groups of phones that build up words and words build up sentences. The number of possible sentences is enormous. The quality of the input, and thereby the amount of input data, can be regulated by the number of samples of the input signal, but the quality of the speech signal will, of course, decrease with a lower sampling rate, resulting in incorrect analysis. We can also minimize our lexicon, i.e. set of words. This introduces another problem, which is called out-of-vocabulary, which means that the intended word is not in the lexicon. An ASR system has to handle *out-ofvocabulary* in a robust way.

6.8 Ambiguity

Natural language has an inherent ambiguity, i.e. we cannot always decide which of a set of words is actually intended. This is, of course, a problem in every computer-related language application, but we will here discuss what kind of ambiguity that typically arises within speech recognition.

There are two ambiguities that are particular to ASR, homophones and word boundary ambiguity.

6.8.1 Homophones

The concept homophones refer to words that sound the same, but have different orthography. They are two unrelated words that just happened to sound the same. In the table below, we give some examples of homophones:

One analysis	Alternative analysis
The **tail** of the dog	The **tale** of the dog
The **sail** of boat	The **sale** of boat

It's impossible on the word level in ASR to distinguish between homophones, we need a larger context to decided which is intended. However, as is demonstrated in the example, even within a larger context, it is not certain that we can choose the right word

6.8.2 Word boundary ambiguity

When a sequence of groups of phones are put into a sequence of words, we sometimes encounter word boundary ambiguity.

Word boundary ambiguity occurs when there are multiple ways of grouping phones into words. An example, to illustrate the difficulty:

It's not easy to wreck a nice beach.

It's not easy to recognize speech.

It's not easy to wreck an ice beach.

This example has been artificially constructed, but there are other examples that occurs naturally in the world. This can be viewed as a specific case of handling the continuous speech, where even humans can have problems with finding the word boundaries.

7. DISCUSSION

In this paper, I have addressed some of the difficulties of speech recognition, but not all of them. But one thing is certain, ASR is a challenging task. The most problematic issues being the large search space and the strong variability. We think that the problems are especially serious, because of our low tolerance to errors in the speech recognition process. Think how long you would try to communicate verbally with a computer, if it understood you wrong a couple of times in a row. You would probably say something nasty, and start looking for the keyboard of the computer. So, there are many problems, but does this mean that it is too hard, that we actually should stop trying? Of course not, there have been significant improvements within ASR, and ASR will continue to improve. It seems quite unlikely that we will ever succeed to do perfect ASR, but will surely do good enough. One thing that should be investigated further, is if humans speak differently to computers. Maybe it isn't natural for a human to communicate in the same way to a computer as to a human. A human may strive to be unambiguous and speak in a hyper-correct style to get the computer to understand him/her. Under the assumption that the training data also is given in this hyper-correct style, this would simplify the ASR. However, if not, it may be the case that hyper-correct speech even makes the ASR harder. And if this is not the case, we may investigate how we as human speaker can adapt to the computer to increase the quality of the speech recognition. As pointed out before, our goal is not a 'natural' verbal communication, we want efficient user interfaces.

8. REFFERENCES

[1] N. M. Ben Gold. Speech and Audio Signal Processing, processing and perception of speech and music. John Wiley & Sons, Inc., 2000.

[2] J. H. M. Daniel Jurafsky. Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Upper Saddle River, New Jersey 07458, 2000.

[3] J. Holmes. An introduction to sociolinguistics. Longman Group UK Limited, 1992.

[4] E. A. Jens Allwood. Corpus-based research on spoken language. 2001.

[5] K. E. Mats Blomberg. Automatisk igenk¨anning av tal. 1997.

[6] B. Schneiderman. The limits of speech recognition. Communications of the ACM, 43:63–65, 2000.

[7] Santosh K.Gaikwad and Pravin Yannawar, A Review, International Journal of Computer Applications A Review on Speech Recognition Technique Volume 10– No.3, November 2010

[8] Rybach, D.; C. Gollan; G. Heigold; B. Hoffmeister; J. Lööf; R. Schlüter; H. Ney (September 2009). "The RWTH Aachen University Open Source Speech Recognition System". Interspeech-2009: 2111–2114.

[9] Sanjivani S. Bhabad Gajanan K. Kharate International Journal of Advanced Research in Computer Science and Software Engineering, An Overview of Technical Progress in Speech Recognition Volume 3, Issue 3, March 2013

[10] Wiqas Ghai and Navdeep Singh International Journal of Computer Applications (0975 – 8887) a Literature Review on Automatic Speech Recognition, Volume 41– No.8, March 2012.

[11] Melanie Pinola (2011-11-02). "Speech Recognition Through the Decades: How We Ended Up With Siri". www.techhive.com.

[12] Anil K.Jain, et.al., Statistical Pattern Recognition: A Review , IEEE Transactions on Pattern Analysis and Machine Intelligence ,Vol.22, No.1, January 2000.

[13] Celso Auguiar, in CCRMA - Center for Computer Research in Music and Acoustics. Stanford University on Modelling the Excitation Function to Improve Quality in LPC's Resynthesis.