Use of Word Pairs and Context to Achieve Better Automatic Speech Recognition Results with Foreign English Speakers

R. Schneider and D. Wilcox
Department of Engineering Technology, Kennesaw State University, Marietta, Georgia, United States

Abstract - The research and development of modern speech recognition technology has been in progress for over forty years. There are still many areas of ASR (automatic speech recognition) technology which require improvement. One of these areas is speech recognition of English spoken with an accent.

Modern ASR programs use not only acoustic properties for speech recognition, but also grammar and syntax rules. This paper will review various methods used to recognize speech and their development. Additionally, the phonetic differences between English and, respectively, Russian, Mandarin Chinese, and Hindi, will be discussed, including a review of difficult consonant and vowel sounds for each of the accents to pronounce.

This paper will also attempt to take advantage of syntax and grammar rules, specifically word pairs, to improve the probability of recognition of the speech of ESLs (second language English speakers).

Keywords: Artificial Intelligence, Speech Recognition, Accents

1 Introduction

In 1971, a study group headed by Allen Newell and funded by DARPA began a five year project with the goal of improving computer automated speech recognition. The name of the project was the Speech Understanding Research (SUR) Project [1]. One of the main objectives of the SUR Project was to add more knowledge sources to ASR programs, specifically acoustic, phonemic, lexical, parametric, semantic, and sentence knowledge sources [1].

Many different speech recognition programs were developed by the time the SUR project ended in 1976, but there are three programs that appear to stand above the rest in setting the ground work for future research: DRAGON, Hearsay, and Harpy. All three programs were developed at Carnegie Mellon University (CMU), under the leadership of Dabbala Rajagopal Reddy aka Raj Reddy, who did his graduate research at Stanford University, and later taught and researched at CMU [2]. The same group later released the SPHINX speech recognition program in the late 1980’s, widely credited as the first speaker independent ASR program.

The creator of the DRAGON Program, Dr. James Baker, later commercialized the program under the name DRAGON Naturally Speaking. Xuedong Huang joined the faculty at Carnegie Mellon University (CMU) in 1989, and directed the Sphinx-II speech recognition program under the tutelage of Raj Reddy and Kai-Fu Lee. Huang later founded the Speech Technology Group at Microsoft Research, which eventually created Windows Speech Recognition [3].

2 The Speech Recognition Model

Below is a very basic model of an early speech recognition system [4]:

![Figure 1. Early speech recognition model](image)

The acoustic front end receives the incoming voice signal and processes the signal digitally.

In the early ASR programs, once the speech pattern was converted to its frequency spectrum, the pattern was processed by comparison to both the acoustic model and the language model [4].
These days most ASRs operate on speech signals in the cepstral domain [5]. The cepstral domain of a signal is obtained by taking the Inverse Fourier Transform of the logarithm of the Fourier Transform of the signal. The cepstral domain is preferred to the frequency spectrum as it contains information not only about the dominant frequencies in the signal, but also the rate of change of the dominant frequencies. The rate of change can provide information about the pitch of an utterance, and can also help to discern voiced segments from unvoiced segments.

The cepstral domain values are in quefrequency, and are measured in seconds, although they are a combination of time and frequency values. Just as filter coefficients can be obtained for the frequency spectrum, cepstral coefficients can be obtained through a process called “liftering”. The most common type of cepstral coefficients used by modern speech recognition programs are the Mel Frequency Cepstral Coefficients [5].

The acoustic model is in essence a dictionary which defines phonemes for different frequency spectrums or cepstral domain readings. The miniaturization and dramatic cost decrease in storage devices over the past fifty years has improved this aspect of voice recognition dramatically [4].

After being compared to the acoustic model, the signal is then compared to the language model. The language model contains the statistical probability of certain words occurring based on grammar rules and other linguistic conventions. The language model may include some common phrases and can be used to choose a noun over a verb following a verb. If, for example, the remainder of the phrase “all’s well that…” was being recognized, the words “ends well” may be chosen over “tends well” or “ends swell” due to the linguistic model.

ASR Programs can have “static” acoustic and linguistic models, which are preprogrammed by the manufacturers and cannot be modified. Static models tend to have low accuracy, however, compared to “dynamic” models, which can be trained.

Training the acoustic and linguistic models of an ASR Programs consists of speaking known words into the program, allowing the program to set acoustic baselines for future recognition. DRAGON Naturally Speaking requires training by its user, which leads to highly accurate dictation when used by the person who trained the program. Windows Speech Recognition can be trained, but training is not mandatory as with DRAGON.

Training an acoustic model can be a very time consuming task. Both DRAGON and Windows Speech Recognition, as well as most modern ASR Programs, benefit tremendously from the common use of the internet and the fact that the devices being used to run the application are most likely online. The ASR Programs use the voice input from the users to refine the acoustic and linguistic models. As a result there are millions of people essentially training the acoustic and linguistic models of the programs every day when correcting errors in dictation [6].

As ASR Programs advanced to continuous speech recognition, the language model became more complex, and comparison with the acoustic model and the language model did not necessarily take place sequentially. Recognition of an audio signal no longer has to be from left to right, and iterations can occur in the process. Different models or knowledge sources can be applied to the signal until the optimal probability of recognizing the utterance correctly is achieved.

One of the biggest challenges to ASR Programs is the issue of accents. Every language has different phonemes, and English contains some phonemes that do not exist in other languages.

This paper will examine how Chinese, Indian, and Russian accents can affect the pronunciation of certain English language phonemes.

The paper will also include a dictation test of Windows Speech Recognition and DRAGON Naturally Speaking, as dictated by 6 volunteers: 2 with a Mandarin Chinese accent, 2 with an Indian accent, and 2 with a Russian accent. As the current standard for ASR Programs is continuous speech, not conversational speech, the volunteers will be asked to enunciate and speak as clearly as possible, with a minimal yet defined break between words.

The most commonly used metric to measure an ASR Program’s accuracy is the Word Error Rate (WER) [6], which is defined as

$$\frac{S + D + I}{N}$$  \hspace{1cm} (1)

-- the sum of the number of words incorrectly substituted, deleted, or inserted, divided by the total number of words attempted. The experiment will use this metric to grade DRAGON and the Windows Speech Recognition program for the handling of each of the three accents.

3 ASR programs and recognition methodologies

3.1 The Hidden Markov Model

The DRAGON Speech Understanding System was one of the first speech recognition programs, if not the first, to make use of Hidden Markov Models [6]. The program makes use of the idea of “nesting” Hidden Markov Models, creating a hierarchy of probability calculations for each chosen path. Hence the probability of a certain “emission” or set of outputs signifying a particular word or set of words being
spoken can be obtained not only using acoustic definitions, but also grammar and linguistic rules [7].

A simplified example of this system is given in Dr. James Baker’s discussion of the first DRAGON program [7]. In this example the program being discussed is a chess program which is controlled by voice commands.

The program assumes that the first word to be spoken will be the name of a chess piece, followed by either the word “on”, the word “takes”, or the word “to”, and so on. Certain steps can be broken down further, such as the type of chess piece to be moved. The piece being moved and the current configuration of the board provide the probabilities for the Hidden Markov Model used to recognize the next word.

The Hidden Markov Model can contain models of not only words, but also of sentence structures and linguistic and grammar rules, such as a sentence having a subject and a predicate, or a verb having a high probability of following a noun [7]. Through the use of these higher level models, the Hidden Markov Model methodology can lower the perplexity of a system greatly.

3.2 Blackboard Method

The Blackboard Method or Blackboard Architecture derives its name from the analogy of comparing the mechanics of the system to a group of experts sitting around a blackboard analyzing a problem, and as each expert writes something on the blackboard in an effort to solve the problem, the other experts benefit from that additional information.

The Hearsay-I Speech Recognition System, developed by Raj Reddy, Lee D. Erman, and Richard B. Neely in Carnegie Mellon University in 1971, was the first to make use of the Blackboard Method [8]. The idea of the system was to use multiple knowledge sources on the sample of speech to be recognized. The first Hearsay System only used three knowledge sources – acoustic-phonetic, syntax, and semantics [8].

The three knowledge sources interacted with each other by each source putting forward hypotheses for the speech segment being analyzed, followed by one or both of the other sources analyzing said hypotheses, until a hypothesis could be agreed upon with a certain level of confidence by all three knowledge sources.

So if, for example, the acoustic-phonetic knowledge source puts forward a hypothesis of “I want to be on a sailboat”, the syntax KS (knowledge source) would then analyze said hypotheses for proper grammar and would perhaps agree with the hypothesis, at which point the semantics KS would examine the hypothesis and perhaps disagree, and modify the hypothesis to “I want to be on a sailboat”. Said modification would then trigger a review by the other two knowledge sources. The syntax KS would be likely to confirm the modification, and the acoustic-phonetic KS would also be likely to confirm it. Although the confidence of the acoustic-phonetic KS is higher in the “sailboat” hypothesis, it is reasonable to assume that the combined confidence of the acoustic-phonetic and semantics knowledge sources is higher in the “sailboat” hypothesis.

The Hearsay-II System not only added complexity by including more knowledge sources, but also by including levels of interaction, with certain knowledge sources applied to certain levels.

The main reason for the re-organization of the blackboard and the creation of the separate levels was for the sake of efficiency. If each KS tested every hypothesis put forward by the other knowledge sources as well as every modification the process would be too time consuming to be usable in applications. The different levels were created so that knowledge sources only interacted with the hypotheses of a portion of the other knowledge sources, and only under certain conditions [8].

3.3 The Harpy System

The Harpy speech recognition system, developed by Bruce T. Lowerre in 1976 under the guidance of Raj Reddy, sought to combine the advantages of the DRAGON and Hearsay Programs by using multiple knowledge sources together with a mathematically tractable model [9].

The Harpy System uses a transition network much like the Hidden Markov Models used in the DRAGON Program. There are two main differences. The first distinguishing difference is that while the Hidden Markov Models used in the DRAGON Program are pre-determined, the transition network of the Harpy System uses heuristics to modify transition probabilities dynamically based on context. The other main difference is that while the DRAGON Program searches all possible paths for a given length of speech input, the Harpy System uses a “Beam Search” to eliminate some paths based on probability.

The “Beam Search” is a modified best-first search strategy, in which the paths with the lowest transition probabilities are eliminated. While possibly eliminating the correct path, this method improved on the time response of the DRAGON System greatly [8]. Take, for example, words like “control” or “bargain”. The DRAGON System would calculate the probabilities of all the paths involving said words, whether the words were being used as a noun or a verb. Based on context, the Harpy System may be able to distinguish if the word is being used as a noun or a verb, and thus eliminate some possible paths from being investigated. This could lead to errors if, for example, the speaker was using incorrect grammar.
3.4 The Sphinx System

The Sphinx Speech Recognition System, developed by Kai-Fu Lee, Hsiao-Wuen Hon, and Raj Reddy at CMU, was the first system to demonstrate speaker independent speech recognition combined with a large vocabulary and continuous speech [10]. The first Sphinx Program was developed in 1987, at a point when computing power increased to the point of enabling newer methods to be used for ASR.

The Sphinx System made use of Linear Predictive Coding coefficients in the cepstral domain to achieve speaker independence. Linear Predictive Coding (LPC) is the process of extracting frequency spectrum or cepstral coefficients in real time to assist in speech recognition [11].

The LPC process extracts the relevant coefficients by removing the effects of formants from the speech signal and estimating the intensity, frequency, and pitch of the remaining signal. This process is called inverse filtering [11]. The derived coefficients can then be used to assist in recognition of future speech by the same speaker.

Using the LPC coefficients from the cepstral domain not only provides information about the frequency characteristics of the speaker’s voice, but also the pitch characteristics. This ability to distinguish pitch as well as frequencies enables Sphinx to recognize multiple versions of identical phonemes. The Sphinx System makes use of the fact that different versions of the same phonemes are used based on context to assist in recognition [10].

Since different speakers generate different power amplitudes in their speech signal, the Sphinx System normalizes the power reading by subtracting the maximum power value of a sentence from each power value in the sentence [10].

The basic methodologies and algorithms used by ASR programs have not changed dramatically over the past forty years. Many algorithmic improvements have been made, and the use of Deep Neural Networks has improved the heuristics of ASR, but the concepts detailed above are still in use in ASR systems today.

4 Effects of Hindu, Russian, and Chinese accents on spoken English

4.1 Phonetic variations in spoken English caused by a Hindu accent

Compared to English, Hindu has more consonant sounds (about twice as many), and about half as many vowels.

The Hindu vowel phonetics are comprised of: /a/ (about), /a/ (father), /i/ (sin), /i/ (seen), /u/ (food), /b/ (book), /e/ (made), /æ/ (sad), /o/ (soda), and /aʊ/ (now). The vowel phonemes /e/ (bet, met), /a/ (cut, love), /o/ (butter, actor), /al/ (five, eye), and /əl/ (boy, join) do not exist in Hindu, and are approximated by the closest existing phoneme [12].

The consonant phonemes /ð/ (this, mother) and /ʒ/ (pleasure, vision) do not exist in Hindu, and could cause difficulties for ASR programs when spoken by someone with a Hindu accent. Additionally, consonant clusters, especially at the beginning of words, are rare in Hindu [12].

Hindu has a predictable word stress. English speakers for whom Hindu is a first language may have trouble pronouncing words with irregular stress patterns, such as photograph and photographer [13].

4.2 Phonetic variations in spoken English caused by a Russian accent

The Russian language is very complex in terms of inflection and stress of syllables. As a rule Russian words use more syllables than English words. Additionally, sequences of three, four, or more consonants occur frequently in the Russian language [14].

Most consonant phonemes in the Russian language have a palatalized and unpalatalized pronunciation. The same phoneme can mean different things depending on if it is palatalized or unpalatalized. The English language does not differentiate between palatalized and unpalatalized phonemes, although different versions of the same phoneme are used occasionally in different words, such as key (palatalized) and coat (unpalatalized) [14].

As a result the Russian language has more consonant phonemes than English, although there are phonemes in the English language which do not exist in Russian. The Russian language only has five vowel phonemes. There are no elongated vowels or diphthongs in Russian, which may cause issues with speech recognition for people with Russian accents.

The following consonant phonemes do not exist in Russian phonology: /h/ (hit, how) – usually substituted with /x/ (butter, actor), /θ/ (this, mother), /ð/ (think, both) – usually substituted with /t/ and /d/; /w/ (wet, window) – usually substituted with /j/ (joyce, five); and /dʒ/ (just, large). Additionally, the vowel phonemes /i/ (hit, sit), /e/ (bet, met), /æ/ (cat, laugh), /a/ (cut, love), /o/ (about, sofa), /ə/ (butter, actor), /d̪/ (dad, mad), /al/ (five, eye), /aʊ/ (now, out), and /əl/ (boy, join) do not exist in Russian phonology, and will usually be substituted with the closest sounding vowel phoneme which does exist in Russian [14].

Native Russian speakers tend to produce an audible release for consonants ending a word, and are likely to transfer this tendency to English speech, creating inappropriate releases of final bursts that sound overly
careful and stilted and even causing native listeners to perceive extra unstressed syllables [14].

In general, words in Russian have more syllables, different stress patterns, and more complex consonant strings than English. As a result words with multiple consonants are unlikely to cause trouble for Russian ESLs ( speakers for whom English is a second language), except for the consonant phonemes which do not exist in Russian.

4.3 Phonetic variations in spoken English caused by a Chinese (Mandarin) accent

English can be a very difficult language for native Chinese speakers to learn. Out of the three languages studied in this paper, it is the most different from English. There are general factors that differentiate the two languages, but also specific phonemes which cause trouble for native Chinese speakers speaking English.

One of the main differences between the two languages is timing. Chinese speech timing is dictated by the number of syllables being spoken, and all syllables take virtually the same amount of time to pronounce. In English, the basis of timing is emphasis, and the emphasized syllables take more time to pronounce [15]. Native Chinese speakers may have trouble adjusting to the English speaking rhythm, and pronounce each syllable in the sentence with the same timing. Although some of the algorithms of the speech recognition program account for some discrepancies in timing of speech, it is possible that this irregularity in speech pattern, which can hinder understanding of a Chinese ESL speaker when heard by humans, could also hinder recognition by an ASR.

Chinese is a tonal language, meaning that certain phonemes can have different meanings based on tone. The phoneme “fēi”, for example, can mean “coffee”, “not”, “fly”, or “luxurious” based on the tone. There are 5 distinguishing tones in Mandarin Chinese which can exist for every syllable: high, high rising, low falling rising, high falling, and neutral [16].

English speakers use tone to indicate emotion or the level of certainty of an utterance, differentiating it between a question, statement, or suggestion. It is likely that the speech recognition programs use the tone of a spoken speech segment for context, and that improper intonation may lead to difficulty in recognition [16].

There are also differences in the way utterances are formed by Mandarin Chinese and English speakers. There are no voiced stops, affricates, or fricatives in the Chinese language. Chinese stops, affricates and fricatives are distinguished by being either aspirated or unaspirated. As a result, all of the phonemes associated with voiced stops, affricates and fricatives in English do not have parallels in Chinese [16].

Additionally, Chinese and English speakers form phonemes which use the front of the tongue and the hard palate in combination in different ways. As a result, the phonemes formed by the front of the tongue and the hard palate are not shared at all by the two languages, except for the “y” phoneme, although the “y” phoneme may be a problem for native Chinese speakers when followed by a high “i” or “I” sound, such as in “east” or “yeast” [16].

Since there are no voiced stops in Chinese, the voiced stops “b”, “d”, and “g” may be incorrectly substituted with a “p”, “t”, or “k”, respectively (“pill” instead of “bill”, “tu” instead of “do”, and “ket” instead of “get”) [16].

The Chinese “r” phoneme is pronounced quite differently from the English “r” phoneme, and as a result the “l” phoneme bears more similarity to the English “r” for most Chinese speakers, leading to a substitution of “l” for “r” phonemes (“lice” instead of “rice”) [16].

A major problem is with the common final consonant in English. This feature is much less frequent in Chinese, and the only consonant Chinese words end with is the /n/ or /ŋ/. This results in native Chinese speakers either failing to produce the consonant or adding an extra vowel at the end of the word. For example, “hill” may be pronounced as if without the double “ll” but with a drawn out “l”, or as rhyming with “killer” [17].

There are also many differences in how vowel sounds are created in the two languages. Chinese ESLs have trouble distinguishing the difference between /æ/ (cat, black), /a:/ (arm, father), and /ɔː/ (call, four), as well as the difference between /ɛ/ (bait, made), /ɛ/ (bet, head), /ʌ/ (about, sofa), and /o/ (go, hope).

Finally, diphthongs such as in “weigh”, “now”, or “deer”, which do not exist in Chinese, are often shortened to a single sound [17].

5 Setup of experiment

The experiment was comprised of 6 speakers, 1 male and 1 female for each language. The speakers with the Russian accent are both in their sixties. The speakers with the Hindu accent are both in their twenties. The male speaker with the Chinese accent is also in his twenties. The female speaker with the Chinese accent is in her seventies.

The female Chinese speaker also works at a school which teaches English to Chinese people. It is likely that her profession aided in overcoming difficulty in the pronunciations of challenging words.

The 6 speakers were recorded in a quiet environment, but not a sound-proof one. They were recorded using an Audio-Technica AT2020USB Microphone on the Audacity Audio Program running on a Lenovo 510s Touch Laptop.
using Windows 8. The voices were recorded at a sampling rate of 16 kHz.

In order to ensure consistency in the sample used for DRAGON and Windows Speech Recognition, the recordings were used for recognition, not the original speech. Once recorded, the voice sample was converted to a WAV file, and the WAV file was played on an ASUS N53S Laptop using Windows 7 running the Audacity Audio Program. The WAV file was played through an FIIOE10 amplifier and a Pioneer VSX520 Receiver, feeding only one PSB Image 4T Speaker. Once played through the speaker, the same Audio-Technica Microphone was used to input the WAV file into the Lenovo Laptop, at first running WSR, and then DRAGON.

The DRAGON Program started with two advantages over WSR. In order to begin using a new profile, a multi paragraph speech segment must be spoken to train DRAGON to adapt to the new profile’s voice. All speakers read said paragraph, which was played through the PSB Speaker to train the DRAGON Program.

WSR also requires training with every new profile created, but the training is only one sentence, and appears to only be used to set the audio levels of the microphone.

Additionally, the DRAGON program allowed the setting of “English with an Indian Accent” for the 2 speakers with the Hindu accent, and allowed setting of “other accent” for the Russian accent speakers and the speakers with the Chinese accent, as there were no “Russian Accent” or “Chinese Accent” choices.

Aside from the training segments for DRAGON and WSR, the participants read the following twelve sentences:

Group 1:
--The towel is wet
--The year started in January
--She put the soy on the window
--The swing is for leisure
--The car was scrapped
--Jen was not in the zone

Group 2:
--The beach towel was soaking wet
--The New Year started in the month of January
--She put the soy beans on the window sill
--The swing set is for leisure time
--The automobile was in the scrap yard
--Jennifer was zoned out

Each of the first 6 sentences contains 2 difficult words, and at least 4 of the words in Group 1 are difficult for each accent. The goal of the second group of sentences is to use word pairs, context, and word choice to help the ASR with recognition, as “window sill” may be easier to recognize than “window”, “January” may be easier to recognize with “month of” in front of it, “automobile” may be easier to recognize than “car”, etc.

6 Results

The additional training appears to have paid off, as the results from the DRAGON Program were far better than WSR. It appears perhaps more training is necessary in order to be able to use WSR to recognize accented speech.

Table 1 below shows the results for the 6 speakers. It appears that the results from WSR perhaps could be run again with more training data, as the Word Error Rate appears to be too high to be meaningful.

Table 1. Key word error rate and total word error rate

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAGON 1st Key Words WER</td>
<td>50%</td>
<td>58.3%</td>
<td>25%</td>
<td>33.3%</td>
<td>58.3%</td>
<td>25%</td>
</tr>
<tr>
<td>DRAGON 2nd Key Words WER</td>
<td>50%</td>
<td>25%</td>
<td>16.7%</td>
<td>25%</td>
<td>33.3%</td>
<td>7.3%</td>
</tr>
<tr>
<td>DRAGON Total WER</td>
<td>43.8%</td>
<td>34.2%</td>
<td>9.6%</td>
<td>16.4%</td>
<td>39.7%</td>
<td>16.4%</td>
</tr>
<tr>
<td>WSR 1st Key Words WER</td>
<td>75%</td>
<td>93.7%</td>
<td>33.3%</td>
<td>38.3%</td>
<td>92.7%</td>
<td>50%</td>
</tr>
<tr>
<td>WSR 2nd Key Words WER</td>
<td>58.3%</td>
<td>75%</td>
<td>58.3%</td>
<td>58.3%</td>
<td>92.7%</td>
<td>50%</td>
</tr>
<tr>
<td>WSR Total WER</td>
<td>63.0%</td>
<td>89.0%</td>
<td>42.5%</td>
<td>58.9%</td>
<td>72.6%</td>
<td>41.1%</td>
</tr>
</tbody>
</table>

7 Conclusions

Out of the 6 speakers, the accents of the female Chinese speaker and the 2 speakers with the Hindu accent were the least pronounced. It appears that the less pronounced accents of the Hindu speakers, and the fact that the accent of the new user in DRAGON can be set to “English with an Indian Accent”, helped the DRAGON recognition of the Hindu speakers immensely.

Out of the 6 DRAGON trials, improvement was seen in 5 out of 6 speakers in the recognition of key words in Group 2, with the last speaker getting 50% recognition on key words in both Groups 1 and 2. Out of the 5 WSR trials, results were mixed, with 2 speakers showing improvement in Group 2 in key word recognition, 3 of the speakers obtaining identical recognition rates for key words, and 1 speaker getting a better recognition rate on key words in Group 1 than in Group 2.
Although some of the results do show improvement in the recognition of difficult words in Group 2, it appears that better choice of complementary words may lead to better results. The word “soaking”, while possessing strong contextual attachment of the word “wet”, is difficult for many accents to pronounce because of the “oa” sound. “Automobile” was chosen as a more acoustically distinctive word then “car”, but was pronounced incorrectly by 2 speakers.

Additionally, any further test run on WSR may need to include more training data, as the recognition rate is extremely low, and may not provide useful information in regards to the effectiveness of using context to improve recognition.

8 References


