The Role of Speaking Style and Context in the Accuracy of Automated Speech Recognition

Michael Colley

Abstract

The goal of this paper is to investigate the effect of speaking style on speech recognition accuracy. By performing an error analysis with a commercially available speech recognition system using data from both read and spontaneous speech, it is possible to estimate how much improvement in accuracy could be expected if computers were able to handle the variation that results from the difference in speaking style as easily as human listeners. While fast speech phenomenon, such as phonetic reduction and coarticulation, and parsing errors can account for over half of the errors, the study also shows the importance of context in understanding speech. A perception study used to estimate how phonetically recognizable the computer transcription errors were shows that even human listeners perform poorly when attempting to identity phonetically reduced words out of context.

Introduction

Despite the rapid advances made in speech technology over the last few decades, computer recognition of human speech still seems rather primitive compared to the visions of it portrayed in science fiction. No computer comes close to being able to transcribe speech as accurately as a human being. Robson (2008), for example, notes, “Existing [speech recognition] systems are only reliable if people talk in a steady, uniform manner. In reality, speech varies depending on where you are and what you're doing.”

The goal of this paper is to investigate the effect of speaking style on speech recognition accuracy. It also examines the role that context plays in understanding speech. By performing an error analysis with a commercially available speech recognition system using data from both read and spontaneous speech, it is possible to estimate how much improvement in accuracy could be expected if computers were able to handle the variation that results from the difference in speaking style as easily as human listeners. In recent years, researchers in speech technology

have become more interested in language variation, since this is often seen as one of the main barriers to accurate speech recognition (e.g., Strik and Cucchiarini 1999; Riley et al. 1999.) This paper aims to show that while handling language variation in it many forms (such as variation due to style, fast-speech phenomena, coarticulation, etc.) is indeed an important topic to be addressed in speech recognition, using the overall context is also crucial in understanding speech.

The speech recognition software used in this study is Dragon NaturallySpeaking, Version 9, one of the most popular and highly acclaimed speech recognition systems commercially available (Metz 2005).

Background on speech recognition

A detailed discussion of the algorithms involved in automated speech recognition is beyond the scope of this paper. (For a more detailed introduction to the topic, see Coleman 2005). Instead, I will mention some of the most important features of these algorithms that are relevant to the discussion that follows.

The first step in any speech recognition program is to prepare the speech for computer analysis. This includes sampling the speech signal at a fixed interval, say 10 or 20 ms., and creating a vector of parameters, most of which are measured from the spectra of the signal. Next, these parameters are used to estimate the probability that the particular time slice represents a particular phonetic event. The use of probabilities is important at this level because the phonetic representations depend not only on the parameters measured from the acoustic signal but also on the surrounding environment.

Most commercial speech recognizers use hidden Markov models to estimate the probability of the so-called hidden parameters given a set of observable parameters. In its simplest form, this uses Bayesian-style probabilities to calculate the most likely word or phrase given the acoustic signal. The model is called “hidden” because the parameters do not uniquely identify any particular phone or other speech event. Rather, the speech parameters are used to calculate an observation sequence, which are the likelihoods that a particular set of phones were

observed in the signal. Based on these likelihoods, the model calculates the probability of a particular sequence of phones or words occurring in a language corpus. These are referred to as N-grams, which can be bigrams (combinations of two phones/words), trigrams (combinations of three phones/words), etc. These probabilities are only as accurate as the corpus on which they are based. Since no corpus can represent the totality of human language, speech recognition systems are always biased toward a particular style of language (usually written, since most large-scale corpora, such as the 100-million word British National Corpus, consist predominately of written texts).

Previously, speech technology researchers looked to linguistics to help develop recognition systems (Coleman 2005). Thus earlier attempts at speech recognition, often referred to as knowledge-based approaches or acoustic-phonetic approaches, were more grounded in specific linguistic knowledge, such as formant frequencies and the presence or absence of voicing, similar to the way a phonetician might try to "read" a spectrogram. These systems used binary decision trees to parse the speech signal by analyzing its acoustic properties. These decisions may be of the type "is the sound voiced or voiceless?" or "is F2 high or low?" Such systems are generally not very accurate. According to Coleman (2005), one of the main problems is the nature of the decision trees; a single error in one of the decisions will affect the rest of the decision process, and it is difficult to recover from such errors. The binary nature of the decisions may also be inappropriate for human language; although binary features are common in formal phonology, actual speech parameters, such as voice onset time, formant values, and formant loci, almost always exist in continua, that is, there is usually a range of possible values for any given parameter. Another problem, as noted by Zue (1990), is the "inability to reliably extract phonetic information from the speech signal." This problem seems to be related to the inherent variability of speech, including stylistic variation, inter-speaker variation, and fast speech phenomena. The linguistic models on which knowledge-based systems were based did not take language variation into account.

Pattern-matching approaches, which are less interested in the phonetic structure of speech, rely on training the software on large corpora specifically designed for speech recognition to automatically construct a set of parameters for each vocabulary word in the system. As noted in Coleman (2005), these systems are more reliable than knowledge-based approaches, but they still have difficulty dealing with the inherent variability of speech. The usual approach to dealing with this problem is extensive training on multiple speakers, or in the case of most commercially available software, training of individual users.

More recently, there has been a resurgence in interest in knowledge-based approaches, including more attention to language variation, usually under the premise that a synthesis of pattern-recognition with phonetic knowledge could decrease error rates (e.g., Pols 1999; Strik and Cucchiarini 1999; Barry, Van Dommelen, and Koreman 2005). Although the core of most currently used systems is still based on the pattern-matching approaches, numerous proposals for adaptations have been made to handle fast-speech phenomena and other forms of linguistic variation. Some of the most relevant of these to this study are summarized here.

Martínez et al. (1997) examines the characteristics of slow, average, and fast speech. They describe some of the fast speech phenomena encountered in a database of Spanish sentences spoken at slow, normal, and fast rates. These include phone deletion, monophthongization, and some phonetic variants, such as the use of [r] for [s] in clusters of the type /sT/, as in *doscientos* > *dorcientos* "two hundred." The focus of the paper is on predicting how the duration of the phones varies with speaking rate. In regard to the phonetic variation, they claim that it is difficult to "accurately predict phonetic phenomena like phone elision."

Siegler and Stern (1995) discuss some possible adaptations to improve the recognition accuracy of fast speech. One of these included modification of the pronunciation dictionary, specifically eliminating schwa in certain contexts. Such modifications did not improve the accuracy of the system, however.

One of the problems with studies like these is that they attempt to improve recognition accuracy by accounting for a handful of common variants. The number of variants that can occur

in speech is far more extensive, however. (See, for example, the discussion of function words such as "the" in section 4.2) These studies seem to underestimate the amount of variation in language.

Others have attempted rule-based approaches for dealing with variation. Gravier et al. (2005) for example, proposes an algorithm to account for *liaison* in French. A word like *très* ("very") has two variants, depending on the following word: [tʁe] before a consonant, and [tʁez] before a vowel. However, this type of variation only scratches the surface of what can and does vary in natural speech, especially in the case of fast-speech phenomena (see section 4.2). Even in this case, as the authors note, it is not appropriate to assume that *only* the [tʁe] variant can occur before a consonant.

A study on speech recognition in Dutch (Wester, Kessens, and Strik 2000) found that adding frequently occurring multi-words to the dictionary, such as *ik wil* "I want" and *het is* "it is" significantly improved recognition. In addition, they used a set of phonological rules, such as /t/-deletion and /ə/-insertion, that frequently occurred in such multi-words. (For example "het is" can be [hɛtɪs], [ətɪs], or [tɪs].) The use of multi-words in combination with the phonological rules led to an even better performance, while simply adding the individual variants (such as [hət], [ət] and [t] for "het") to the lexicon led to a deterioration in performance. Furthermore, calculating the probabilities of the variants and including them in the language model also improves performance.

The most successful algorithms for dealing with variation tend to be those that are designed to automatically derive variants from a corpus. Yang and Martens (2000), for example, developed a method to derive pronunciation variants from a transcribed corpus, resulting in improved recognition accuracy. The variants were used to create transformation rules, taking into account the left and right context of the phone or phones in question. The rules are automatically generated from the corpus, thus providing an advantage over previous methods that hand-select variants of frequent words to be included in the lexicon.

Amdal, Korkmazskiy, and Surendran (2000) proposed a method called *association strength* to improve recognition for non-native speakers. Since non-native speakers often have numerous phone substitutions, such as [s] → [z], the algorithm assigns high association values between such phones. The values are derived automatically from a training corpus. Given that the non-native speakers tested in the study had different native language backgrounds, one might expect that deriving the association strengths for each speaker individually would give the best results. Nonetheless, a slightly better recognition accuracy was obtained by deriving the association strengths simultaneously for all speakers, perhaps, according to the authors, because the individual associations may suffer from scarce data, and similar variation may be present in the data despite the differences in native language.

Ravishankar and Eskenazi (1997) note that "between-word coarticulation is a major problem in the recognition of continuous, fluent speech." They developed a method to automatically derive a set of phonetic transformation rules from a corpus. However, they found that although the technique improved recognition of words containing this type of coarticulation, it also introduced other errors, suggesting that the transformations need to be applied more restrictively.

Methodology

An error analysis was performed using the Dragon NaturallySpeaking software to help determine the main reasons for the transcription errors made by the software. Due to propriety issues, I do not have access to the algorithms used by the software; thus it can only be assumed that the software is typical of other large-vocabulary speech recognition systems.

The data used for the analysis consists of twelve native speakers of North American English (seven females and five males), who were recorded in two speaking styles: a reading style and a spontaneous style. All participants were undergraduate students at a university. Background information was obtained on where the participants had lived during their lives and other languages they speak (if any). For the reading style, the participants read aloud a magazine article on global warming; for the spontaneous style, they talked about their opinions on global

warming and other issues brought up by the article. The recordings were made in a soundproof recording booth to minimize the effect of background noise on the error rates. A Microtrack M-Audio digital recorder and a lapel microphone were used. False starts and other hesitations were edited out of the recordings so that these would not bias the results. The interviewer's questions and contributions to the conversation during the spontaneous style (including any cases of overlapping speech) were also edited out of the recordings.

Once recorded, each participant also performed the training session provided by the NaturallySpeaking software. The purpose of the training session is to increase the accuracy of the transcription by adapting the software to the unique characteristics of each individual's voice. It consists of reading aloud a short passage; the program accompanies the user's speech and matches it to the written passage. To test the effect of the training session on transcription accuracy, each participant's recordings (one for reading and one for spontaneous) was transcribed by the Dragon software both before and after the training session. The program's "transcribe recording" feature was used, which allows for transcriptions of pre-recorded speech.

The word error rate was calculated for each transcription. The error rate is defined as the number of errors divided by the total number of words spoken. Any word that was mistranscribed or not transcribed at all (known as *substitutions* and *deletions*, respectively) was counted as one error. In addition, words included in the transcription that were not spoken by the participant (known as *insertions*) were counted as one error. Words that were partly mistranscribed were counted as one half of an error. These include compound words in which only a portion of the compound was transcribed correctly (eg. "motorcycle" transcribed as "motor"), mistranscribed morphology (eg. "canceled" transcribed as "cancel") and mistranscribed homophones (eg. "too" transcribed as "two").

Four error rates were calculated for each participant, one for each combination of reading vs. spontaneous speaking style and before vs. after performing the training session; these will be referred to as "reading no training," "reading training," "spontaneous no training," and "spontaneous training."

A perception test was done on a random sample of the transcription errors made by the Dragon software to help with the error analysis. This was used to determine to what extent the errors were phonetically recognizable by human listeners. Ten native speakers of North American English (six females, four males) were selected from a university. All were either faculty or graduate students from the linguistics department, or undergraduate students in a linguistics class. None had been recorded in the first part of the study. The errors were tested both in isolation and in the context in which they occurred. In both tests, listeners were able to listen to the word or phrase as often as they liked. In the isolation test, they heard a word or phrase of up to three words (depending on the error) and were instructed to type what they heard. In the in-context test, they heard a sentence or part of a sentence, starting from first pause before the error to the first pause after the error. To keep the test from becoming too long or tiring, the sentences were transcribed for them, with a blank inserted in the place of the error. They were instructed to fill in the blank with what they heard.

Each person heard half of the errors in isolation and half in context. They were counterbalanced so that no person heard any two errors twice. Thus each error was tested by five people in isolation and five people in context.

Some errors in the random sample, especially in the reading style, were not able to be tested because they were either the same error that other speakers had made, or because the context of the error overlapped with another error. To prevent people from hearing the same error twice, or hearing two errors with the same context, some errors were excluded from this test.

Error analysis and results

A multivariate repeated measures test was used to test the effects of training, speaking style, and speaker sex. Both speaking style ($p < 0.000$) and training ($p = 0.032$) were significant. To test whether the sex of the speaker had any effect on accuracy, speaker sex was added as a between-subjects factor. This factor did not interact significantly with either training ($p = 0.212$) or speaking style ($p = 0.575$). Therefore, it will not be discussed further.

In both the reading and spontaneous styles, the training lowered the average error rate; the average rate was lowered 2.0 percentage points in the case of reading, and 3.7 percentage points in the case of spontaneous speech. Paired t-tests reveal that both differences are significant, at $p = 0.04$ and $p = 0.02$ respectively. The magnitude of the difference is not large, however. In the case of reading, the average error rate decreased from 15.0% to 12.9%; in the case of spontaneous speech, the average rate decreased from 45.1% to 41.4%. Even after training, the error rates are far from the 1% claimed by Dragon Systems (stated as "up to 99% accuracy" [Nuance Communications 2006]). The magnitude of the decrease is typical of that reported in the speech technology literature for the development of various methods and algorithms that deal with variation (e.g., Riley 1999). From the viewpoint of users, however, it is unlikely that they will find this increase satisfactory. In fact, many users may not see any increase with training. Surprisingly, the lowest error rate achieved in this study (8.3%) was achieved with no training.

The largest factor on the error rates was that of speaking style. Error rates after training ranged from 8.8% to 18.6% for reading and from 23.3% to 52.9% for spontaneous speech.

To help determine the main reasons behind the errors made by the program, a random sample of errors was selected to be analyzed, using a random number generator. The randomness of the sample helps assure that the errors selected for examination are not biased toward any particular type of error. Ten errors were selected per speaking style per participant from the no-training transcription. For each error, it was noted whether the mistranscribed word was a function word and whether it was corrected with training. Function words were defined as prepositions, pronouns, auxiliary verbs, conjunctions, articles, and other grammatical particles. In addition, inflectional affixes (eg. "canceled" transcribed as "cancel") were counted as function words.

Since high-frequency function words are more likely to be phonetically reduced than content words (Jurafsky et al. 2001), it is not surprising to find that over half the errors involve function words: 65% in the reading style, and 55% in the spontaneous style. The lower

percentage in the spontaneous style is likely due to the fact that there are more errors in the spontaneous data overall. Thus it is not the case that more function words were transcribed correctly in the spontaneous speech, but rather that more words in general (both content and function) were mistranscribed.

The errors were corrected in 22.5% of cases in the reading style, and in 35% of cases in the spontaneous styles. Given the low magnitude by which the training session improved transcription accuracy, it is not surprising that in most cases the errors were not corrected with training. In addition, many words transcribed correctly without training became errors after training, thus the net gain in accuracy is even smaller than these percentages suggest.

Despite the use of a sound-proof recording room for the recordings, there are a few instances in which background noise was a problem, usually due to the lapel microphone brushing against hair or clothing. In addition, there were a few cases of clipping, due to laughter or unusually loud speech. In all, there were ten cases of background noise and/or clipping in the sample; since this is likely to be the cause of the errors, they were excluded from further analysis.

Perception test

There are many cases in the sample in which pronunciation variation appears to be a main cause of the error. Words that are phonetically reduced or contain phonetic variants that substantially differ from their canonical pronunciation can become phonetically unrecognizable without proper context. For example, when presented with a reduced form of "and" such as [n̩] in isolation, most listeners in this study were unable to recognize it as "and." By noting how many listeners were able to accurately determine the target word in isolation, the perception test was used as a measure of how phonetically recognizable the words are.

What human listeners do when listening to speech is different from what speech recognition programs do. The computer program uses the hidden Markov models to predict what phones and words are present in the speech signal. The signal is considered to be noisy or imperfect, thus not containing all of the parameters necessary to perfectly reconstruct the speech

signal. The program thus uses data from pre-transcribed corpora to predict what parameters are present in the signal.

Humans, at least in the case of speech presented in context as opposed to individual words, are constantly attempting to parse and interpret the speech, in addition to using knowledge of collocations to predict what will come next. (For evidence of this, see the TRACE model of speech perception [McClelland and Elman 1986].) In this respect, humans are much better at predicting what words should be present in a speech signal than computers are. This can be seen in the many cases in the perception test where 0/5 listeners correctly identified a word in isolation, but 5/5 identified the same word in context. For example, given a word [t^hən] in isolation, listeners interpreted it as "chin," "turn," or "ten." But given the context "Americans [t^hən] to value independence," all listeners identified it as "tend." (The software did not transcribe the word at all.) For a computer using N-grams, this would be a more difficult task, assuming that the word "tend" does not necessarily occur more frequently given the context of "Americans," "to," "value," and "independence," than other words such as "turn" and "ten."

For each error in the perception test, it was noted how many of the five participants correctly identified the word in isolation and in context. Figures 1 and 2 show the distribution of these proportions for the reading style and the spontaneous style respectively. It is not surprising that listeners perform much better when the words were presented in context; for both speaking styles, the majority of the words in context were interpreted correctly by 5/5 listeners, and rarely below 3/5 (only 2 cases out of 107). The proportions of correct identifications of the words in isolation were more evenly distributed among all possible proportions, from 0/5 to 5/5. This shows that the words vary considerably in how phonetically recognizable they are, although the two most common proportions in both cases are 0/5 and 5/5, in that order.

<fig. 1 here>

<fig. 2 here>

For the sake of analyzing the possible reasons for the errors, a word was considered phonetically recognizable if it was correctly identified by over half of the listeners in isolation,

that is by 3, 4, or 5 of the listeners. Using these criteria, over half of the words (51%) can be considered phonetically unrecognizable, meaning that there is not enough phonetic information in the words for most listeners to identify them in isolation. Thus these words are dependent on their context for comprehension. This did not appear to vary by speaking style: 53% of the errors are phonetically unrecognizable in the reading style, versus 51% in the spontaneous style. There were only two cases where the context given was still not enough for most listeners to correctly identify them: "faced with [a] problem of this magnitude" and "tuning up your car [and properly] inflating the tires." In the first case, it is not possible to tell from the context whether the article is "a" or "the"; in the second case, three people forgot to include the second word "properly" in their responses.

Sources of errors

One of the most common types of errors in automated speech recognition is parsing errors. This is due to the program's inability to segment the continuous stream of speech into individual words. Obviously, it is impossible to say with certainty whether a given error is a parsing error just by looking at the transcription. However, it is possible to make a reasonable guess. The following criteria were used to determine if the errors were likely to be parsing errors:

1. A single word is transcribed as more than one word, eg. "congressman" → "commerce and"
2. More than one word is transcribed as a single word, eg. "or at least" → "released."
3. The error is part of a long stream of errors, eg. "I think that this would have" → "it was what"
4. The error appears to be influenced by a sound in a neighboring word, eg. "uncompromising warning" → "uncompromising morning." (The presence of the velar nasal at the end of "uncompromising" caused the [w] in warning to be transcribed as "m.")

Based on these criteria, over half the errors (58%) are likely parsing errors. Parsing errors appear to be more likely in the spontaneous style, in which they occurred in 67% of the errors,

versus 50% in the reading style. This does not exclude the possibility that there are other reasons for the errors, since it is possible for an error to be both phonetically unrecognizable *and* a parsing error (this occurred in 27% of all cases).

This raises the question of whether phonetically unrecognizable words are likely to cause parsing errors. This can be tested by determining whether phonetically unrecognizable words occur more often by themselves or are involved in parsing errors. Of the words that are phonetically unrecognizable, 53% (57/107) were involved in parsing errors. A binomial test of this proportion shows that it is not significantly different from the test proportion of 50% ($p = 0.562$). Splitting the data into reading and spontaneous styles suggests that phonetically unrecognizable errors in the spontaneous style might be more likely to lead to parsing errors than in the reading style. The percentages of phonetically unrecognizable errors involved in parsing errors are 43% (23/53) in the reading style and 63% (34/54) in the spontaneous style. However, a binomial test did not show a significant difference from 50% for either style ($p = 0.410$ for reading; $p = 0.076$ for spontaneous). Although the data for the spontaneous style is approaching significance at $p = 0.076$, there is not sufficient evidence in the current sample to confirm that phonetically unrecognizable errors are more likely to be involved in parsing errors.

Given that function words are more likely to be phonetically reduced than non-function words (Jurafsky et al. 2001), it is expected that the words that are phonetically unrecognizable are more likely to be function words. Indeed, 70% (75/107) of the phonetically unrecognizable words in this sample are function words, which is significantly different from the test proportion of 50% in a binomial test ($p = 0.000$). Very common function words like “the,” “and,” and “in” pose a difficult problem for speech recognition programs because they are so dependent on context for reliable comprehension. Assimilation to the preceding and following phonetic environment is so common with these words that the phonetic realizations of several tokens of the same word may have nothing in common. For example, in the tokens of “the” in this sample, stopping and/or devoicing of the first segment is common, thus resulting in [də] or [tə]. In addition, the schwa is frequently weakened or deleted, as in [t] and [ð]. In some cases, the first

segment can assimilate to the last segment of the previous word, as in “was the” [wəzə]. Between two nasals, “the” may even become a syllabic nasal, as in “in the middle” [ɪnn̩mɪɾl̩]. However, despite the large amount of variation in a given set of tokens, some predictions can be made on what phonetic realizations are more likely. Obviously, a nasal realization of “the” is likely in the presence of other nasals. Schwa weakening or deletion appears to be more common in the presence of sonorants. Thus the variation, while not 100% predetermined, is not entirely random.

The same is generally true of non-function words as well. Most of the variation in this sample is due to predictable, phonological processes, such as deletion or weakening of unstressed segments, and assimilation of neighboring segments. Some examples from unrecognizable errors that are not function words are included in Table 1.

<table 1 here>

All of these errors contain examples of common phonological processes of English (and other languages), including “t” epenthesis between a nasal and a fricative (“hence retaining” [-ntʂ-], assimilation to a retroflex place of articulation (“hence retaining”, “industry” [-ʂɹ-]), deletion of unstressed vowels (second vowel in “industry”), deletion of homoorganic stops after nasals “winter” [-n-], and glottalization of intervocalic stops (“wiggle” [-ʔ-]). There is also one possible regional variant present, that of the lowering of the final [i] in “industry” to [eɪ], a common feature of Southern English, spoken by a native of Houston, TX.

In fact, only a handful of cases can be found in the sample where regional variation seems to have been a factor in the errors, all of them involving vowels. Table 2 lists these.

<table 2 here>

With only 13 out of a total of 240 errors, these types of errors represent only 5% of the errors in the sample. One reason for this may be that most of the regional variation in American English involves vowels, and as a result, most speech recognition programs give priority to consonants, which tend to be less variable. In addition, the sample of speakers was not designed to be representative of the many varieties of English that exist; a different sample of speakers may result in many more cases of errors due to this type of variation.

There are many examples in the sample of transcriptions that are not plausible interpretations of the phonetic input, given what is known about what is likely to vary in speech. Examples of this are shown in Table 3.

<table 3 here>

Unlike the previous examples, these do not contain examples of common phonological processes. In this case, the program seems to have made the following unlikely substitutions: [b] for [l], [f] for [b], [n] for [b], and [ð] for [m]. Although some of these may be plausible under the right phonetic environment, (such as [ð] becoming [m] due to a preceding nasal), none of these occur in such environments. All of these examples were correctly identified in isolation by more than half the participants in the perception study, and thus cannot be considered phonetically unrecognizable. These errors, then, fall under the category of "no known reason," which makes up 18% of the total errors. Of course, there are numerous other reasons for errors not addressed in this paper, including changes in speaking rate, loudness, or pitch, unexpected phonation types (eg. breathy, creaky), inaccurate probabilities in the language model, inaccurate analysis of the acoustic parameters, etc.

Comparing the errors made by the computer software with those made by the human listeners when hearing the words in isolation reveals some interesting information on the strategies used by the computer vs. humans when faced attempting to recognize words with insufficient phonetic input. In some cases, the human errors agree or partially agree with the computer errors, as shown in Table 4.

<table 4 here>

This type of agreement, however, is relatively infrequent, occurring in only 12% of the phonetically unrecognizable errors. More common are errors like those in tables 5 and 6. Table 5 lists those in which most participants agree in their responses, but the computer transcription is different. For those in Table 6, the participant responses are not in agreement, but there is a pattern in the responses that is not present in the computer transcription. In all of these cases, the

participant responses are closer to the original text than the computer transcription. Together these make up 44% of the phonetically unrecognizable errors.

<table 5 here>

<table 6 here>

Although in most of the rest of the errors the participant guesses were as inaccurate as the computer transcription, these examples show that in many cases, the human errors were better guesses given the phonetic input than the computer transcription. For the word "said," for example, all the participant guesses correctly began with an "s," and two of the three had the correct vowel [ɛ], while the computer transcription "that" had neither of these.

Discussion

The results of the error analysis help reveal some of the main reasons why it is difficult for computers to accurately transcribe speech. In terms of the difference between the reading style and spontaneous speech, it appears that the computer transcription of spontaneous speech is more likely to contain parsing errors. It is common in the spontaneous speech style that long sequences of words are mistranscribed. For example:

"…I'd definitely wanna preserve more energy and…", transcribed as "that there is or Martin Sheen"

Because there are so many mistranscribed words, it is hard to tell in these cases what caused the errors. One possibility is that there may be more coarticulation between words in spontaneous speech than in the reading style, and thus it becomes more difficult to parse. It does not appear to be the case that the errors in spontaneous speech are more phonetically unrecognizable than in the reading style. The main difference is that there are simply more of them, reflected in the higher word error rate in the spontaneous style.

Algorithms that specifically deal with language variation, as described in Yang and Martens (2000), Amdal, Korkmazskiy, and Surendran (2000), Ravishankar and Eskenazi (1997), and others should, in theory, be able to eliminate some of the errors analyzed in this paper. In cases where most human listeners are able to accurately identify a word in isolation, there

appears to be enough phonetic information in the signal to identify the word. Since I do not have access to the algorithms used in NaturallySpeaking, I am unable to evaluate how effective they might be in eliminating errors due to linguistic variation. However, it is possible that the problem is not with the algorithms themselves, but with the data on which they are trained. An algorithm that automatically derives phonological rules and probabilities from a corpus is only as good as the corpus that is used.

Some of the most common corpora used to train such algorithms are summarized below.

TIMIT, consisting of "630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences" (Linguistic Data Consortium 2008)

WSJ1 (Wall Street Journal), consisting of dictations made by journalists

DARPA, read and prepared speech from broadcast news

None of these corpora would prepare a speech recognition system for the kinds of variation encountered in this study, nor the kind produced by many potential consumers of the NaturallySpeaking software. The TIMIT corpus includes readings of phonetically rich sentences like "She had your dark suit in greasy wash water all year" (Linguistic Data Consortium 2008). The reading data used in this study, on the other hand, consisted of excerpts from a magazine article, with a total of 1253 words. Furthermore, readers were told that they would be discussing the article afterwards, thus their attention was focused less on the phonetic content of the words and more on the meaning of what they were reading. Isolated and "phonetically rich" sentences like those used in the TIMIT corpus focus readers' attention more on the phonetic content of the sentence. As numerous sociolinguistic studies have shown (e.g. Labov 2006), the amount of attention paid to speech can affect the frequency with which a particular linguistic variant is used. Thus, isolated sentences are likely to contain much fewer instances of a variant than a read text, which in turn will contain fewer variants than a spontaneous conversation. Thus the probabilities calculated for these variants from a corpus like TIMIT would not be accurate for the data used in this study.

The same is true for other corpora like WSJ1 and DARPA. Newscasters and journalists often have training or extensive practice with dictation and these professions generally require speakers to pay close attention to how they speak. Average speakers, however, have little or no experience in dictation and, except perhaps in the first few minutes of speech when being recorded, pay much less attention to how they speak. The use of corpora such as these will likely result in an underestimation of the amount of variation in the speech of many potential users.

However, as the results of the perception test show, even the best algorithms for handling linguistic variation cannot be expected to raise the recognition accuracy to the level of a human listener. In over half of the errors tested from this sample, human listeners required hearing the context in which the error occurred in order to be able to correctly identify the word(s). Theoretically, the hidden Markov models enable the computer to use the context together with the acoustic signal to predict the most likely word. But human listeners are far better at this than computers, since they not only use frequent collocations to predict what will come next, they also make predictions based on syntax, pragmatics, etc. In fact, as shown in several examples in this study, human listeners are able to "hear" words that are either not present in the speech signal, or are so reduced phonetically that they bear no resemblance to their canonical pronunciations. An example of the former case is in the phrase "or at least" [ɹlist], in which the word "at" is not pronounced. In the perception tests, 0/5 participants were able to correctly identify the phrase in isolation (all five guessed "released," which is also the transcription made by the software), while 5/5 were able to identify the phrase in context ("cost of going green *or at least* greener").

As in the case of using corpora to automatically train a computer program to handle variation, the accuracy of the hidden Markov models also depends on having corpora that are representative of the types of language that the computer may attempt to transcribe. Since most large corpora are composed of written material (reflecting the difficulty in transcribing large amounts of speech versus collecting written material that is already in electronic form), the probabilities used to predict the most likely word from the context are primarily based on written

language. This may be another reason why most speech recognition systems are better at transcribing read documents than spontaneous speech. For the developers of dictation software, like Dragon NaturallySpeaking, this may be seen as less of a problem; programs like this one are designed to replace the keyboard interface of the computer with a dictation interface. Thus, speakers are expected to adapt their speech to a style that is easier for the program to recognize.

However, the idea that speakers must adapt to computers rather than the other way around raises issues of prescriptivism. An older version of the Dragon software offered the following tip to users:

> Speak naturally and continuously, but pronounce each word clearly. …if you say 'Didja eat,' a person will probably understand that you're asking, 'did you eat?' But Dragon NaturallySpeaking has trouble interpreting mumbled or slurred speech. (Dragon Systems, Inc. 1999)

The equation of common phonological assimilation processes as illustrated in "didja eat" with "mumbled or slurred speech" seems to link normal conversation (as opposed to 'proper' dictation) with an inferior way of speaking. Dismissing such processes as "mumbled" and "slurred" puts the blame of mistranscription on the user, when in fact it may be that the program isn't based on an accurate model of the users' speech. Expecting users to use *less* coarticulation than they otherwise would is the opposite of speaking "naturally." It also means that the software is not likely to have many applications beyond dictation, such as providing closed captioning for movies and television shows and transcribing courtroom testimonies.

Conclusions

It seems unlikely that a single breakthrough in speech recognition technology will suddenly allow computers to understand speech as well as a human being does. Progress in this area has proceeded in small steps, and will likely continue to do so. As the analysis of errors in this paper has shown, there is no one reason why computers have trouble transcribing speech as accurately as a human. This paper looked at two possible reasons: errors due to linguistic variation and parsing errors. There are undoubtedly many other sources of error that were not

addressed. In many cases, it is impossible to examine any one error and determine its cause; it may be due to a number of different factors.

This paper has shown that many aspects of language must be addressed in order to obtain better accuracy in speech recognition systems. Algorithms have been developed for dealing with phonological and phonetic variation, but the algorithms are only as good as the corpora used to train them. Thus a useful model of language variation would only be possible with a large, spoken corpus, consisting of a variety of speaking styles, and many speakers from a variety of social and regional backgrounds. Improvements in how language variation is modeled in computer software have the potential to lower error rates, but only to a certain point. Even human listeners are highly dependent on context for understanding speech due to the large amount of variation present in the speech signal. Linguistic variation would not be possible without the predictive abilities that listeners have; otherwise it would jeopardize our ability to comprehend speech. This suggests that methods based on N-grams alone are not enough to accurately predict and transcribe the content of human speech.

Tables

TABLE 1

Examples of unrecognizable errors with non-function words

Original text	*IPA*[a]	*Computer transcription*	*Participant translations*[b]
hence retaining	[hɛntʂɹɪtʰeɪnɪŋ]	entertaining	entertaining, hentritainin [sic], intertaining [sic]
industry	[ʔn̩dʂɹeɪʰ]	straight	dentistry, prince tree, demonstrate, pidinstry [sic], didn't stray
wiggle	[wɪʔl̩]	little	will the, willow (2), will or, well a
winter	[wənɚ]	one are	one or (2), opener, winner, wondering

a. Impressionistic transcription made by the author

b. The numbers in parentheses indicate the number of responses in cases where multiple participants responded with identical transcriptions.

TABLE 2

Errors likely due to regional variation

Type of Variation	*No. of cases*	*Example*	*Machine transcription of example*
/ai/ monophthongization	4	"I'm" [æm]	"on"
/oʊ/ fronting	3	"know" [nəʊ]	"now"
/i/ lowering	2	"reasons" [rezəns]	not transcribed
raising before nasals	2	"and" [ɛn]	"in"
/u/ fronting	1	"new" [niʊ]	not transcribed
high-front pre-nasal merger	1	"India" [ɛndijə]	"and he"

TABLE 3

Examples of computer transcriptions that are not plausible interpretations of the phonetic input

Original text	*IPA*	*Computer transcription*
lot	[lɑɾ]	but
but	[bət]	for
but	[bʌtˀ]	not
maybe	[meɪbɪ]	they been

TABLE 4

Examples of agreement or partial agreement between computer transcription errors and human perception errors

Original text	*IPA*[a]	*Computer transcription*	*Participant translations*[b]
and properly	[ənpɹɑpə˞li]	improperly	improperly (3)
just	[ɪs]	this	this (2), let's
the	[θʌ::]	it though	though (3), thought

a. Impressionistic transcription made by the author

b. The numbers in parentheses indicate the number of responses

TABLE 5

Examples of agreement among human perception errors but not computer transcription error

Original text	*IPA*	*Computer transcription*	*Participant translations*
the	[də]	and	that (5)
rise	[ɹaɪʰs]	write	rice (4), riotous
foul	[faʊɬ]	fell	fouled (4), felt
and	[ɪn]	*not transcribed*	in (5)

TABLE 6

Examples in which a pattern is present in the human perception errors but not computer transcription error

Original text	*IPA*	*Computer transcription*	*Participant translations*
said	[sɛd^{n}]	that	send, sound, sent
wiggle	[wɪʔl̩]	little	will the, willow (2), will or, well a
don’t	[dð]	will	gone, not, Don, ga [sic], knock
long anyway	[lɔŋɪniweɪ]	longingly	you won't anyway, belong anyway (2), along anyway

Figures

Figure 1

Graph showing the distribution of the proportions of participants that correctly identified each word in isolation vs. in context (data from the reading style)

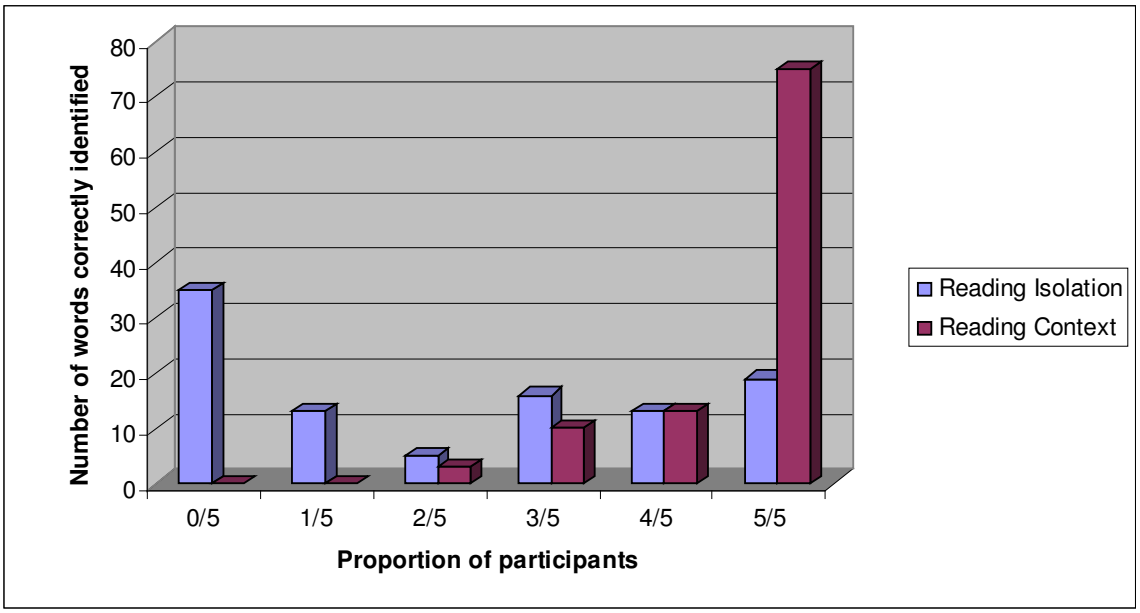
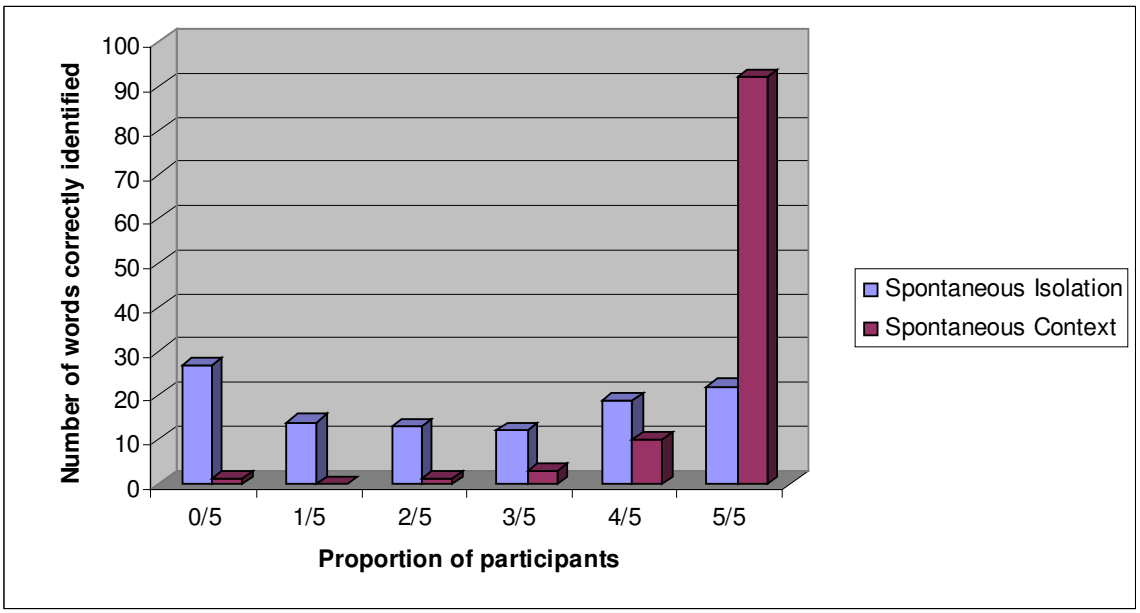


Figure 2

Graph showing the distribution of the proportions of participants that correctly identified each word in isolation vs. in context (data from the spontaneous style)



References

Amdal, Ingunn, Filipp Korkmazskiy, Arun C. Surendran. 2000. "Data-Driven Pronunciation Modeling for Non-Native Speakers Using Association Strength Between Phones." *Proceedings of the ISCA Workshop on Automatic Speech Recognition—Challenges for the New Millenium*: 85-90.

Barry, William J., Wim A. Van Dommelen, Jacques Koreman. 2005. "Phonetic Knowledge in Speech Technology—and Phonetic Knowledge from Speech Technology?" *The Integration of Phonetic Knowledge in Speech Technology*, ed. William J. Barry and Wim A. van Dommelen, 1-12. Dordrecht, The Netherlands: Springer.

Biber, Douglas. 1994. "An Analytical Framework for Register Studies." *Sociolinguistic Perspectives on Register*. Oxford: Oxford Univ. Press.

Byrd, Dani; Cheng Cheng Tan. 1996. "Saying Consonant Clusters Quickly." *Journal of Phonetics* 24: 263-282.

Coleman, John. 2005. "Introduction to Speech Recognition Techniques." *Introducing Speech and Language Processing*, 157-183. Cambridge Univ. Press.

Dragon Systems, Inc. 1999. *Dragon NaturallySpeaking User's Guide.* Newton, MA: Dragon Systems.

Gravier, Guillaume, Francois Yvon, Bruno Jacob, Frédéric Bimbot. 2005. "Introducing Contextual Transcription Rules In Large Vocabulary Speech Recognition." *The Integration of Phonetic Knowledge in Speech Technology*, ed. William J. Barry and Wim A. van Dommelen, 87-106. Dordrecht, The Netherlands: Springer.

Jurafsky, Daniel; Alan Bell, Michelle Gregory, William D. Raymond. 2001. "Probabilistic Relations Between Words: Evidence from Reduction in Lexical Production." *Frequency and the Emergence of Linguistic Structure*, ed. J. Bybee and P. Hopper. Amsterdam: Benjamins.

Keating, P.A. 1998. "Word-level Phonetic Variation in Large Speech Corpora." Paper presented at *The Word as a Phonetic Unit* in October 1997, ZAS Papers in Linguistics 11, ed. A. Alexiadou et al., 35-50.

Kirchner, Robert Martin. 1998. *An Effort-Based Approach to Consonant Lenition*. Ph.D. Dissertation, Univ. of California, Los Angeles.

Koopmans-Van Beinum, F.J. 1980. *Vowel contrast reduction: An Acoustic and Perceptual Study of Dutch Vowels in Various Speech Conditions.* Ph.D. Thesis, University of Amsterdam.

Labov, William. 2006. *The Social Stratification of English in New York City*. 2nd Ed. Cambridge Univ. Press.

Linguistic Data Consortium. 2008. "TIMIT Acoustic-Phonetic Continuous Speech Corpus". http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1. Accessed Oct. 28, 2008.

Martínez, F., D. Tapias, J. Álvarez, P. León. 1997. "Characteristics of Slow, Average and Fast Speech and Their Effects in Large Vocabulary Continuous Speech Recognition." *Proceedings of Eurospeech*, Rhodes, Greece: 469-472.

McClelland, J.L., J.L. Elman. 1996. "Interactive Processes in Speech Perception: The TRACE model." *Parallel Distributed Processing Volume 2: Psychological and Biological Models*, ed. J.L. McClelland, D.E. Rumelhart, and the PDP Research Group. Cambridge: MIT Press.

Metz, Cade. 2005. "Speech recognition gets better." *PC Magazine*, 24(11): 52.

Nuance Communications, Inc. 2006. "Dragon NaturallySpeaking 9: Preferred Edition" Burlington MA.

Pols, Louis C.W. 1999. "Flexible, Robust, and Efficient Human Speech Processing Versus Present-Day Speech Technology." Proceedings of the International Congresses of Phonetic Sciences (ICPhS 99), San Francisco: 9-16.

Ravishankar, M.; M. Eskenazi. 1997. "Automatic Generation of Context-Dependent Pronunciations." *Proceedings from Eurospeech*, Rhodes, Greece: 2467-2470.

Riley, M.; W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, G. Zavaliagkos. 1999. "Stochastic Pronunciation Modelling from Hand-Labelled Phonetic Corpora." *Speech Communication* 29: 209-224.

Robson, David. 2008. "Artificial Tongue Mimics Human Speech." *New Scientist* 2666: 26.

Siegler, M.A.; R.M. Stern. 1995. "On the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition." *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Detroit: 612-615.

Strik, Helmer; Catia Cucciarini. 1999. "Modeling Pronunciation Variation for ASR: A Survey of the Literature." *Speech Communication* 29: 225-246.

Van Son, R.J.J.H.; Louis C.W. Pols. 1999. "An Acoustic Description of Consonant Reduction." *Speech Communication* 28: 125-140.

Wester, Mirjam, Judith M. Kessens, Helmer Strik. 2000. "Modeling Pronunciation Variation for a Dutch CSR: Testing Three Methods." *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, Beijing, China: 417-420.

Yang, Qian; Jean-Pierre Martens. 2000. "Data-Driven Lexical Modeling of Pronunciation Variations for ASR." *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, Beijing, China: 417-420.

Zue, Victor W. 1990. "The Use of Speech Knowledge in Automatic Speech Recognition." *Readings in Speech Recognition*, ed. Alex Waibel and Kai-Fu Lee, 200-213. San Mateo, CA: Morgan Kaufmann Publishers.