Competitive Evaluation of Commercially Available Speech Recognizers in Multiple Languages

Susanne Burger, Zachary A. Sloane and Jie Yang
interACT, Carnegie Mellon University, Pittsburgh PA, USA
sburger@cs.cmu.edu

Abstract

Recent improvements in speech recognition technology have resulted in products that can now demonstrate commercial value in a variety of applications. Many vendors are marketing products which combine ASR applications including continuous dictation, command-and-control interfaces, and transcription of recorded speech at an accuracy of 98%. In this study, we measured the accuracy of certain commercially available desktop speech recognition engines in multiple languages. Using word error rate as a benchmark, this work compares recognition accuracy across eight languages and the products of three manufacturers. Results show that two systems performed almost the same while a third system recognized at lower accuracy, although none of the systems reached the claimed accuracy. Read speech was recognized better than spontaneous speech. The systems for US-English, Japanese and Spanish showed higher accuracy than the systems for UK-English, German, French and Chinese.

1. Introduction

Evaluations of commercially available dictation systems have been continuously conducted alongside of their evolution. Starting from simple command structures and small vocabulary dictation, these systems improved to discrete dictation (pausing between words) of larger vocabulary. The current systems allow continuous dictation of different text types directly into text editors, operation of applications and desktop functions and surfing the Web using voice input, and claim they recognize speech at 98% accuracy.

Most of these evaluations looked at the recognition accuracy of read texts, usability, time and cost efficiency (e.g. Wyard, 1993; Burger, 1997; Cane, 1998; Zafar, 1999). Even for early systems, the accuracies were already relatively high -- around 98%, depending on processor power, vocabulary size and duration of training. It then became more interesting to look at the performance of such systems after only minimal training and with differing types of input speech: read and prepared text versus spontaneously uttered thoughts. Using the systems out of the box with only minimal training already decreased accuracies to below 80% (Devine, 2000), with the recognition of spontaneous speech at only 55% accuracy (Broughton, 2002). Most of the studies available were done using English systems, though Burger (1997) used German systems.

In the current evaluation, we will try to give an objective view of the state of the art in commercial recognizers and the variation in their performance across several languages, although the identity of the recognizers cannot be released. The point is to compare the different recognizers rather than to get absolute values on their performance.

The systems were chosen based on their availability in as many languages as possible. For each language, we compared different text types: read texts of different difficulty levels and spontaneously-created samples, with a given topic or situation. To balance the necessity of a sufficient number of subjects per language with the need for comparable results between subjects and languages, we decided to use the default rejection strategy of each system and to do just the minimum-required enrollment per speaker and system.

2. Data Collection

2.1 Languages

The target languages were US-English, UK-English, Iberian Spanish, French, German, Japanese, Simplified Chinese, and Traditional Chinese.

UK and US-English are considered to be two different languages for the purposes of this study because the manufacturers offer different recognition engines for each.

Similarly, the evaluated manufacturer offers two different engines for Simplified and Traditional Chinese, even if these could also be seen as two dialects of Mandarin Chinese -- Simplified Chinese developed from Traditional Chinese out of historical and political changes. At first glance the character sets for Simplified and Traditional Chinese seem to be very different from each other. However, the Simplified Chinese writing system is in fact a simplified version of the Traditional Chinese writing system. Both systems still co-exist; Simplified Chinese is used in mainland China, while Traditional Chinese is used in Taiwan and Hong Kong. Thus, we recorded speakers from mainland China for Simplified Chinese and speakers from Hong Kong and Taiwan for Traditional Chinese.

2.2 Evaluated Dictation Systems

Table 1 shows all evaluated software packages.

- UK-English: There was no extra software package for UK-English for systems C and W, but both packages offered both UK and US-English as choices.
- Spanish and French: system W was not available in a comparably recent version.
- Simplified and Traditional Chinese: There was no version of system C found for Chinese. The Chinese version of system W could only be used

in single-user mode; it was not capable of registering new users and thus it could not be evaluated in our setting.

Id	Language	Software Packages		
		System C	System W	System Y
us	US-English	x	x	x
uk	UK-English	x	x	x
ge	German	x	x	x
sp	Spanish	x	-	x
fr	French	x	-	x
jp	Japanese	x	x	x
cs	Chinese-simplified	-	-	x
ct	Chinese-traditional	-	-	x

Table 1: Evaluated software packages.

2.3 Speakers

The recognition accuracy was measured using six female and six male speakers of each of the eight languages. Each speaker was a native speaker of the tested language. They were recruited locally, primarily from Pittsburgh universities.

Speakers were allowed to exhibit slight regional variants from the understood "standard" of the language under test. None of the speakers were speech professionals. Pre-tests showed that all three systems had difficulties dealing with strong variants, dialects, and accents; these were therefore avoided when possible.

All speakers had post-secondary education, and at least twenty percent were between thirty and fifty years of age. Additional data were collected from each speaker, including age, gender, height, weight, level of education, profession, self-reported dialect or accent, place of residence during the first years of attending school, place of residence for the longest period of life, dialect/language relationship to parents, and place of origin of parents.

2.4 Recorded Samples

Each speaker recorded at least five different speech samples: a 'warm-up' text (T04) (a short piece of a well-known story of the particular language) which was not evaluated, an 'easy' text with mostly common vocabulary (T05), a 'hard' text containing more fringe vocabulary or technical jargon (T06), a short, spontaneous message left on an imaginary answering machine (T08), and spontaneous dictation of a duration of either two (T07) or four minutes (T10) for which the speaker could choose the topic. All the texts except the answering-machine message contained punctuation and formatting commands that speakers had to read as part of the text.

To avoid cultural differences, it was decided not to translate the texts for reading; for each language new texts were chosen. The texts may differ in terms of difficulty compared with the texts of the other languages, but each language was assigned an ‘easy’ and a ‘hard’ text.

2.5 Equipment and Recording Environment

The experimental environment was modeled on the target environment of commercial speech recognition systems: a Windows PC in a small office, with moderate, typical office noise and an inexpensive microphone. The experiment platform was Windows XP.

The following provides a short rundown of the physical setup:

- Dell PC with Pentium III 860MHz and 512 MB of RAM.
- Andrea Anti-Noise NC6 headset microphone.
- The computer was situated in the corner of a 4-workstation office, with dimensions of about 10’ x 16’. There were typically other people working in the office, but talking was not allowed while recording.
- The phone was unplugged and the door was locked, with a warning sign on the outside.

2.6 Recording Procedure

The experiment was designed to use pre-recorded speech from each speaker. The benefit here is that each system can transcribe the exact same speech sample. The alternative to this approach would force subjects to repeat the same texts several times; this would have led to inconsistencies due to fatigue, weariness and boredom. Furthermore, it is inconceivable to imagine that a subject could reproduce spontaneous speech exactly enough for the purposes of this study.

Each recording session was administered by an assistant fluent in the tested language, who was able to give instructions for speaking style as well as guide the participant through the recording steps. Each speaker first followed the manufacturer's standard enrollment wizard for minimum enrollment so that each system could adapt to speech and voice characteristics particular to each speaker. After completing the enrollment, speakers proceeded to read and record the experimental texts, within a simple software interface. This program provided instruction screens and sound recording functionality, and briefed speakers on how to speak to the computer (speaking pace, using spoken commands for punctuation, etc). It also included a volume meter which indicated if the speaker was within green parameters (volume acceptable) or red parameters (too loud). Speakers could re-record any of the samples or receive help from the assistant at any time. Recordings were done in 22 Kilohertz, 16 Bit. It was chosen randomly which system was trained first

2.6.1 Problems of Data Collection

Several problems were encountered during data collection:

The duration of the spontaneous recording was extended from two to four minutes after the English data had been collected. Technical problems marred the collection of the four-minute recordings in Simplified Chinese; they are consequently only two-minute recordings. The

spontaneous speech samples were extended to four minutes simply because they hold the most potential for further interesting research. Rather than having a consistent length of two minutes in all languages, we pursued the opportunity to collect larger samples in the remaining languages. Although we did not find significant changes in accuracy rates compared to the two-minutes samples, there may be effects of fatigue or speakers may behave differently because they have to produce more spontaneous thoughts for the extended duration.

The enrollment texts offered by the systems are very similar. To have the same enrollment condition for all three systems, and to have the possibility to enroll systems repeatedly providing the exact same condition, the desired experiment design would have used one pre-recorded enrollment text read by each speaker for all three systems. This would have shortened the enrollment period for the benefit of more time to record speech samples. Unfortunately, not all of the systems could be enrolled by pre-recorded sound files. We decided to have all users actively enroll in each system and save the enrollment profiles for re-usage.

Audio for the enrollments was recorded in parallel by a program developed for the experiment. This background recording procedure worked only for German, Japanese, French and Spanish.

The experiment setting required that the systems transcribe pre-recorded wav files to ensure that all of them received the same input. The manufacturer of one of the three systems provided us with an additional function of transcribing pre-recorded samples because this function was originally not included in the software package. The manufacturer now plans to provide this function in future versions, though.

Pre-tests showed that the systems worked optimally at different volume levels. Therefore, we decided to deviate from pure black-box testing to manipulate the volume of some recordings in order to have a comparable and fair condition for all three systems.

Also, many of the Chinese recordings were lower in volume than the other recordings -- there were even volume differences between recordings of the same speaker. We boosted the volume of those recordings.

2.7 Manual Transcription

Human transcribers transcribed the speakers' recordings to produce a reference transcription. The spontaneous speech recordings were manually transcribed at the word level, and punctuation and formatting commands were specially labeled as such. Disfluencies in speech were additionally marked. The read speech was validated and adapted to the original speech sample, including punctuation commands, disfluencies and misreading.

The transcribers were asked to give a subjective grading of how sloppy or well articulated each speaker spoke; grade 1 was best, grade 6 worst.

2.8 Corpus

The entire recorded corpus consists of over 25 hours of recorded speech produced by 96 speakers in the eight tested languages, with about three hours of speech per language.

2.9 Automatic Transcription

Assistants later fed the speech samples into the speech recognition systems' transcription functions using the recorded speaker profiles to produce the recognized-text outputs.

All dictations systems were used in their default accuracy setting (systems W, Y: medium, system C: low) for automatic recognition.

3. Evaluation

Using the SCLITE evaluation tool provided by the National Institute of Standards and Technology, the output transcribed by the systems and the hand-transcribed outputs were compared to calculate word error rate (WER).

Accuracy for English, French, German and Spanish was calculated as a function of words; Chinese and Japanese as a function of characters. The Japanese evaluation was done manually, because of the mixed usage of hiragana, katakana and kanji characters; in this mixed system there are multiple 'correct' ways to transcribe the same speech.

3.1 Analysis

For the measurement of accuracy rates we compared the transcription of the recorded audio file and the output which each system produced for this audio file. In the case of audio files with multiple versions (i.e. those with boosted or cut volume), the version with the best results went into the analysis.

Reference transcriptions and output transcriptions needed to be preprocessed and normalized to be used by SCLITE: In the case of the reference transcriptions that meant:

- removing everything aside from the actual transcription of the spoken words (no disfluency labels, comments, etc)
- converting everything to lower case
- normalizing spelling inconsistencies (e.g. proper names, different spelling possibilities)

In the case of the automatically-produced transcriptions, produced by the systems:

- all executed commands were converted to the command tag version of the reference (to see which commands were correctly executed, which were transcribed as words and which words were executed as commands)
- everything was made lower case
- format differences were normalized
- in the case of system C, digits had to be written out
- different spellings of vocabulary were normalized (hyphens, proper names, etc)

SCLITE outputs several detailed reports per file, such as significance tests and lists of confused vocabulary. The current analysis looks at the following files:

- Ensemble.es: the results as a percentage (containing WER, insertions, deletions, substitutions)
- Ensemble.prn: a lined comparison of the reference text and all outputs.

Results give the percentage of recognition read from the Ensemble.es files.
SCLITE has special options to handle character error rate and Chinese GB coding. All Chinese transcriptions, references and outputs were preprocessed, all disfluency markers and other ASCII annotations were removed and spaces were added between all characters. All Chinese transcriptions were converted into GB3213 encoding. Several Traditional Chinese reference transcriptions and output files contained one or two characters which created problems in SCLITE, and which subsequently had to be found and converted manually.

4. Results

The results we report here are the results for word/character accuracy:
System C and system Y recognized at comparable accuracy, although the margin of difference varies from language to language.
Considering only the languages where systems C and Y were both available -- that is, excluding Chinese, C showed 76% accuracy averaged over all speakers, all languages and all recorded samples, and Y showed an accuracy of 75%. The average for system W was at 66%. These results fell far short of the recognition accuracies advertised by the manufacturers, though one must acknowledge that minimal user training and variations in input style (read and spontaneous) contributed to this. The systems' accuracy also varied between the languages tested.
The following sections look at this in more depth.

4.1 Best System

Figure 1 shows which system "won" the evaluation per language. The bars in figure 1 show accuracy rates over all speakers and texts per language in percentage. The little dots denote the highest and lowest observed recognition accuracy of all speakers and texts.

- Systems C and Y worked best in their Japanese, US-English and Spanish versions, with average recognition accuracies between 77% and 90%.
- UK-English, German and French showed lower results, between 63% and 71% accuracy.
- System W was always in third position for the languages where it was evaluated.
- Simplified and Traditional Chinese could only be evaluated by using system Y, which produced low results (60% and 54%).

The systems had difficulties recognizing some of the speakers while working very well for others. Standard deviation between speakers was lowest for the Japanese systems (3.5) and for the Spanish systems (4.1). The highest deviation between speakers was observed for the Chinese speakers (13.6)

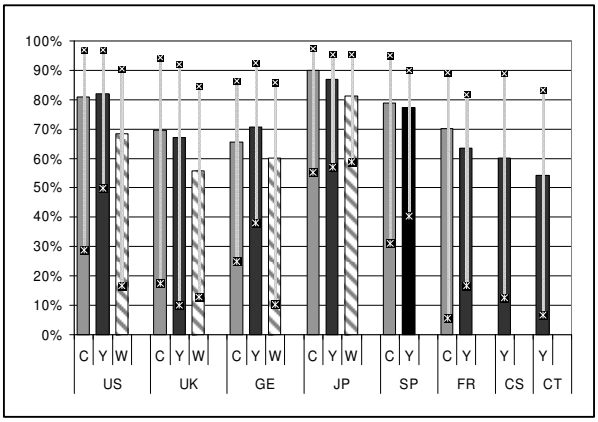

Figure 1: Accuracy rates (%), average over all recordings and all speakers per language and system. Dots show maximum and minimum accuracies.

The highest accuracy rates reached by single speakers vary between 82% and 98%, but recognition could be as low as only 6%.
A female Japanese speaker produced the best results: System C recognized her four-minutes of spontaneous speech (T010) at 98%. The same sample was recognized at 93% by system Y and at 84% by system W. This speaker was also the favorite speaker of system C for her other recoded samples, recognized on average at 91%, while her samples were recognized by system Y at 89% and by system W at 84%.
Two US-English speakers also reached recognition rates of 97% with their two-minute spontaneous speech samples (T07) -- one speaker with system C and the other with system Y.

4.2 Best Recognized Text Types

Figure 2 shows the difference in percentage of accuracy rates of read speech versus the results for spontaneous speech. All of the very high recognition rates mentioned above occurred in the spontaneously-produced speech samples. However, looking at the average values of accuracy, all non-Japanese language systems were most successful in recognizing the read samples. In Japanese, the best-recognized recording was the four-minute spontaneous dictation T10 and the differences between read speech and spontaneous speech (including T08) were very low (under 10%). System Y for US-English, Japanese and Spanish and system C for Japanese had the lowest differences in accuracy between recognitions of spontaneous and read speech.

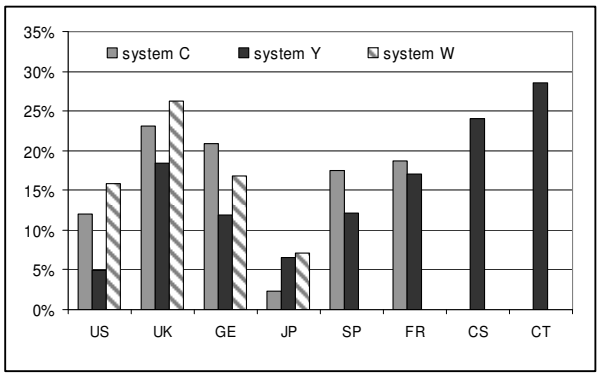

Figure 2: Difference of recognition accuracy (%) between read and spontaneous samples.

4.3 Influence of Articulation, Speed and Disfluencies

The grades labelers gave to a speaker's articulation and speaking style, as well as the number of labeled disfluencies and the number of words a speaker spoke per minute gave some idea of the correlation of accuracy rates and speaking style.

A preliminary analysis shows that for speakers of English, Japanese and Chinese high grades for articulation correlated with high recognition rates. Articulation grades were weakly correlated with recognition for German speakers and not at all for Spanish and French speakers.

German speakers had the highest correlation between a high number of disfluencies and low recognition rates. English speakers only exhibited this correlation with system Y.

French and Spanish speakers had a significant correlation between speed and recognition rates: faster speaking led to lower recognition rates. None of the other languages had any correlation between speed and recognition rates.

4.4 Differences between the Languages

The evaluated systems produced slightly different results depending on which language package was used: system C recognized UK-English, German, Spanish and French best, while system Y had better results with US-English and Japanese.

In contrast to the non-Japanese participants, Japanese speakers spoke slow and articulated well when they spoke spontaneously.

Certain languages have posed significant problems to the systems.

The UK-English speakers clearly had stronger and more varied dialects than the US-English speakers, especially when they produced spontaneous speech.

All systems had problems with German composites, especially if these were plural or declined. Also the frequent reduction of the final “e” in German verbs (e.g. “finde” becomes “find”) was not handled very well. The same was observed for the assimilation of the final “e” followed by “es” (eng.: it): e.g. “sage es” becomes “sag’s”.

Both French and Spanish posed problems, since they show a lot of assimilation during spontaneous and faster speech. The systems also had difficulties to differentiate between French endings if they sound similar but are written different: e.g. –ieu, -eaux, -ant, -ont, -et, -é. Often plural forms were not recognized.

The recognition of Chinese showed low accuracy, though it is unclear whether a repetition of the Chinese recordings without technical problems would lead to better results for both Simplified and Traditional Chinese. There was confusion observed between Chinese “s” and “sh”. We had the impression that female Chinese speakers in particular spoke more softly when they produced their spontaneous samples.

5. Discussion

5.1 Why does System W have Lower Accuracy?

System W has significantly lower results than the other two systems. Older evaluations, however, reported accuracies for system W in the same range or even higher than other systems. To our initial surprise, it was not possible to purchase a recent version of System W in Spanish and French. The last version available was from 2003. This matches the time when the manufacturer of system W gave the distribution rights to another organization. We assume that further improvements of this system were stopped at that time.

5.2 Is it the Languages Themselves or the Attention they get?

The results indicate that speech recognition in the commercial sector currently gives differing amounts of success for different languages. US-English, Japanese and Spanish outperformed UK-English, French, German and Chinese in our evaluation. In 4.4 we reported the differences we found within the languages themselves.

Another reason for the differing results could be the attention the languages get by the manufacturers. Because of market demands, manufacturers may have put more effort into the training and development of systems for certain languages. English and Spanish are certainly commercially more interesting than German because more people speak English and Spanish and the market for speech recognition in these languages is big. The demand of toys, games, and technical gadgets equipped with speech technology is very high in the Japanese market. For Japanese, there are more established commercial speech recognition systems available than for the other languages, indeed. We were considering a fourth very established system from Japan, but it did not support multi-user mode. These emphases of the Japanese market may explain a higher demand for well-developed speech recognition systems and the very good results for Japanese.

5.3 What is "Dictation"?

The difference between read texts and freely dictated texts shows that desktop dictation is still far from the type of

dictation where users develop texts on the fly and speak spontaneously. Desktop dictation works best for reading prepared texts to produce automatic transcriptions. The accuracy of spontaneous speech found in our experiments is on average over all languages 13% lower than for read speech.

Commercial dictation systems are obviously not practical for what is typically understood as the creative process of developing a document to have it then transcribed by another person. The form of dictation these systems can handle best is merely reading prepared text to a system. This is conceivably useful for such things as medical reports or law texts which consist often of pre-formulated text modules.

Shneiderman (2000), points out that even if the recognition of dictation input is increasingly accurate, the adoption outside the disabled-user community has been slow compared to visual interfaces. Some reasons for this may be the fatigue people feel from speaking continuously or the disruption in an office filled with people speaking. Humans also type and think better than speak and think because of the way our brain functions. Therefore, spoken commands and prepared text may work well, but creating and elaborating at the same time is inconvenient.

6. Conclusion and Future Work

We evaluated accuracy rates of three commercially available desktop speech recognition systems across eight languages. The results show that system C and system Y perform at almost the same accuracy, depending on the tested language. System W always showed accuracies significantly lower than the other two systems. The systems recognized read speech better than spontaneous speech. Japanese, US-English, Spanish were better recognized than UK-English, French, German and Chinese.

We already started to evaluate the lists of alternatives the systems offer in the case of misrecognized words. Often, the correct word is already on top of such lists. Sometimes similar sounding words or differing version of the correct words are offered. Also, the correlations of speaking style and accuracy which we already had preliminarily looked at promise to be worthy a deeper analysis.

A problem we would like to focus on in our further work with this corpus is how the actual error correction process works using the methods proposed by the dictation systems. A study done by Halverson (1999) shows that correcting recognition errors by using spoken commands for corrections often results in a cascade of new errors, where the correction of the error was repeatedly not recognized and had to be corrected before the actual error cold finally be corrected.

7. Acknowledgements

We are very grateful to Maria Kernecker, Aparna Nayak-Guercio, Kazumi Maniwa and Lu Huan who helped collecting the data. We thank Spike Katora, Nolan Pflug, Jiazi Ou, Jen Anderson, Li Jiang, Paul Hsu, Sveket Duran, Shahid Durrani, Shiun-Zu Kuo and Joseph P. Friday who solved lots of technical challenges and helped conducting the evaluation.

We thank Microsoft Inc. for the generous gift which made this evaluation possible.

8. References

Wyard, P. (1993). The Relative Importance of the Factors Affecting Recognizer Performance with Telephone Speech. In Proceedings of EUROSPEECH 1993, (pp. 1805-1808). Berlin 1993

Burger, S. & Tillmann, H.G. (1997). Comparison of Commercial Dictation Systems for Personal Computers. In Forschungsberichte des Instituts fuer Phonetik und Sprachliche Kommunikation der Universitaet Muenchen (FIPKM), volume 35 (pp. 107 – 114). Muenchen 1997

Cane, J., (1998). Comparing Dragon Naturally Speaking and IBM ViaVoice Gold, ENW International, http://www.enw-ltd.com/vv-dragonComparison.htm

Zafar, A., Overhage, J.M., McDonald, C.J., (1999). Continuous Speech Recognition for Clinicians. The Journal of the American Medical Informatics Association. 1999, volume 6, (pp. 195 – 204).

Devine, E.G., Gaehde, S.A., Curtis, A.C. (2000). Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. The Journal of American Medical Informatics Association, 2000 Sep-Oct 7(5) (pp. 462 - 468).

Broughton, M. (2002). Measuring the Accuracy of Commercial Automated Speech Recognition Systems during Conversational Speech. Workshop on "Virtual Conversational Characters: Applications, Methods, and Research Challenges". 2002, Melbourne, Australia.

Shneiderman, B. (2000). Limits of Speech Recognition. Communications of the ACM, volume 43, Number 9 (2000), (pp. 63-65).

Halverson, Ch., Horn, D.B., Karat, C., Karat, J. (1999). The Beauty of Errors: Patterns of Error Correction in Desktop Speech Systems. Paper presented at INTERACT 99, the International Federation for Information Processing conference on Human-Computer Interaction. Edinburgh, Scotland, 1999.