Accuracy Analysis for
Radio Broadcast Deployment of Captioned Radio Services

Tom Apone

Trisha O'Connell

Ruth and Carl Shapiro National Center for Accessible Media at WGBH

Introduction

This study examined different approaches to creating captions for NPR radio programs, in support of NPR/WGBH efforts to prototype and field test the technologies, service models, and operational techniques required to provide a real-time text track that can be transmitted with radio broadcasts to better serve consumers who are deaf or hard-of-hearing.

NCAM compared the performance of the two technologies currently available for this purpose — real-time stenocaptioning, which relies on a specialized court reporting professional, and a new process called "voice-writing" in which a highly skilled operator listens and repeats the audio portion of a program or lecture into specially trained speech recognition software which converts that echoed speech into text.

Stenocaptioning

Real-time captions for television broadcasts are created by court reporters retrained as stenocaptioners, using stenographic keyboards and dedicated real-time captioning software. Stenocaptioners must be able to listen to programming intently, differentiate homophones, synonyms, and unfamiliar words, at high speeds, with precision and under pressure, and instantly recall and utilize recently assigned keystrokes for program- and topic-specific vocabulary to render a 95-99% accurate real-time transcript. At the simplest

cognitive level, this involves instantaneous differentiation between homophones, which are not easily identified by software.

A stenographer uses a small black keyboard with 22 keys and depresses combinations of keys that represent phonetic sounds. In this way, the stenographer is able to type an entire word with a few keystrokes (much like depressing chords on a piano), rather then typing letter-by-letter on a traditional QWERTY keyboard. This enables typing speeds of 200 words per minute and higher for a highly skilled stenographer. On the mechanical level, stacking errors are caused by conflicting keystrokes and require strict attention to releasing keys when fingering the keyboard.

The stenographer builds a dictionary in her computer that translates these keystrokes into English. These dictionaries can grow to 100,000 entries or more over time. Caption delay is the result of the stenocaptioner's response time and the few seconds of processing time required by the software to translate steno keystrokes to English. This delay typically does not exceed two or three seconds.

A shortage of skilled stenocaptioners combined with a federal requirement to caption almost all television programming has generated interest in exploring the use of speech recognition technologies in an effort to provide another option and reduce costs.

Speech recognition

Continuous, speaker-independent speech recognition systems may one day be able to provide accurate real-time captions. Today, however, despite constant advances in the technology, speech recognition technologies cannot produce acceptable levels of accurate translations with rapid speech and minimal processing delay. Accuracy falls markedly when there are multiple

speakers or background noise. Moreover, continuous speech recognition technologies utilize language-processing capabilities to accurately interpret words based on context. This results in a significant recognition delay, as the software analyzes the appropriate word choice within context to produce a meaningful sentence. This delay can result in an entire sentence appearing as captions *after* the speakers have finished talking and the program has moved on to a different topic altogether.

There is growing interest in the use of speaker-dependent automatic speech recognition (ASR) using an approach called voice writing or voice echoing whereby a trained voice writer listens to an audio stream and repeats the words in real-time. The speech recognition software, which has been trained to the characteristics of that particular voice, then generates the corresponding text. The two most common ASR software programs in use are IBM's Via Voice and Dragon Naturally Speaking.

The process requires excellent audio processing and listening skills and the ability to concentrate and repeat audio quickly with clear diction. A good voice writer can reach reasonably good accuracy rates depending on the speed of the audio being repeated. Our voice writer for this test has several years of experience and also teaches voice writing at a local academic institution. Her system was equipped with IBM's Via Voice speech recognition software.

<u>Direct input</u>

We also tested the feasibility of feeding audio directly into a speech recognition system without a voice writer repeating the words. We attempted this with the latest version of Dragon Naturally Speaking (version 9.5) but, as expected, it proved completely unreliable and unusable. Without properly training the system to the voice it was receiving, there was no way for the system to produce

an acceptable transcript. It is possible that with some preparation, which might include having on-air personalities spend an hour training the speech recognition systems individually, reasonable accuracy rates could be attained. However, this approach would still not resolve accuracy problems with the unique voices of guests, interviewees, and callers on phone-in programs, like Car Talk.

Materials used

We obtained digital recordings from four NPR programs in January 2008. WGBH Radio division provided high-quality mp3/wav versions from the archives to simulate as nearly as possible the quality of HD Radio. The programs represent a random selection of segments from popular programs and present a variety of styles and challenges. The four programs are: Morning Edition, Fresh Air, All Things Considered and Car Talk. These programs were chosen to represent a range of program types, each of which presents unique challenges with pacing, topics, the number of guests and different voices, unfamiliar names and terms, etc.

Morning Edition	Fresh Air	All Things Considered	Car Talk
Mostly Scripted	Interview	Scripted w/Discussion	Listener Call-in
News	Known guests/topics	News / Known experts	Unknown guests/topics
Medium/fast pace	Slow pace	Medium pace	Fast pace
Multiple Speakers	Multiple Speakers	Multiple Speakers	Multiple Speakers

The segments varied in length from 9 to 18 minutes. The stenocaptioner and the voice writer were provided with these four audio files, along with a list of proper names and terms that appeared in each of the segments. This is a standard practice in the captioning industry, because to get the best translation rates and the best accuracy, new entries and unfamiliar terms must be added to dictionaries every day.

The tests were conducted independently. After the transcripts were completed, each segment's transcript was compared to a verbatim master transcript of the same segment using a simple tool from AJC Software called AJC Diff. This utility compares two text documents side by side and highlights the differences.

General observations

The steno text was of generally higher quality and had long passages of near perfect text with little missing. This was produced by an experienced stenocaptioner who works with news programs daily and is familiar with many of the topics covered in these programs.

The voice writer dropped more words, as is evident just from the raw word count in those transcripts. Car Talk was the most difficult program because of the speed of the conversation, the unfamiliar terms and the tendency of the hosts to "talk over" each other. The verbatim file contained 3335 words; the stenocaptioner's text had 40 fewer words (3295) while the voice writer's text had

almost 453 fewer words (2882). That said, much of the voice writer's text is readable and acceptable. Morning Edition had the lowest accuracy in both cases, largely because of unfamiliar terms and proper names.

Error Counting

Accuracy rates were based on the National Institute of Standards and Technology (NIST) approach. This system counts three types of errors:

Substitutions -- where one word is substituted for another. This may be a homonym, a misspelling or a completely incorrect word.
Additions -- where an unneeded word is added to the transcript. This is a word that was not spoken in the original audio stream and should not be in the text.
Deletions (Omissions) -- where words that should appear in the transcript are missing or dropped.

While this has become a standardized approach to counting errors, it differs slightly from the traditional method of measuring accuracy for stenocaptioners and it does not reflect comprehensibility or readability. In many cases, a dropped word or phrase such as "you know," does not reduce the comprehensibility of the transcript, but it does count as two errors in this system. Finally, simple differences in punctuation were not counted as errors, nor were differences in contractions – i.e., "I am" was considered equivalent to "I'm." In some cases, we made judgment calls as to whether something was an error or what category of error it was. For example, partially spoken words or words that were difficult to hear, due to interruption or "talkover" by another voice, were usually omitted. This was a significant issue in Car Talk.

NIST has hosted speech recognition competitions regularly over the past decade. As noted, their error counting approach has become the standard by which

speech recognition tools are measured but it is a more strident approach than the captioning industry has traditionally taken. The example above – dropping "you know" from the text – would not be considered a significant error. The viewer would not miss these two words; in fact, the text is probably easier to read without the aside. Similarly, some of the errors noted below, which count as two or three errors, would only be considered a single error when assessing caption accuracy.

The software tool we used, AJC Diff, allows the user to open two documents side by side and highlights the differences between them. By taking each transcript and placing one word per line, AJC Diff lines up the text well and color-codes each of the three types of errors. Here is an example from All Things Considered comparing the master transcript on the left to the voice writer's transcript on the right.

IT	IT
WAS	WAS
HELD	HELD
IN	IN
THE	THE
ATRIUM	INTERIM
OF	OF
A	THE
BUILDING	BUILDING
AT	AT
STETSON	STETSON
UNIVERSITY	UNIVERSITY
AND	WITH
ABOUT	ABOUT
300	300
PEOPLE	PEOPLE
WERE	
PACKED	PACKED
IN	IN
PRETTY	
TIGHTLY.	TIGHTLY.
THE	THE
CANDIDATE	CABINET
MADE	MADE
IT	IT
CLEAR	CLEAR
RIGHT	
AT	AT
THE	THE
TOP	TOP
OF	OF
HIS	HIS
SPEECH	SPEECH
THAT	THAT
A	A
LOT	LOT
IS	IS
AT	AT
STAKE	STAKE
HERE.	HERE.

Seven errors are highlighted in this sentence: four substitutions (in yellow) and the three deletions (in red), but there are really only two significant errors in the voice writer's text: "interim" substituted for "atrium" and "cabinet" substituted for "candidate." If those two words had translated correctly, the sentence would be readable and accurate, though not verbatim.

One can also see the phonetic similarities at work here. As speech recognition software continues to improve and incorporates more "cognitive" aspects to analyze context, an error like "interim" will be less likely to occur – that is, "the interim of the building" makes no sense and eventually the computers and software will be powerful enough to recognize that and correct it in context. However, the substitution of "cabinet" for "candidate" may not be flagged by language processing tools as an inappropriate word choice yet it is an error that changes the meaning of the sentence.

Both stenocaptioners and voice writers are trained to produce a coherent, comprehensible output as their top priority and are encouraged to simplify, rephrase and drop words, if necessary, to enhance readability. Normal conversation is approximately 180 words per minute and when speeds exceed 200 words per minute, it is likely that either method will fall behind and be forced to drop some material in order to keep up. Likewise, a typical caption watcher will struggle to read text that is displaying at speeds in excess of 200 words per minute.

The four segments totaled about an hour of audio with a total of 9530 words. Overall, the stenocaptioner's text is superior with accuracy rates ranging from 96.6% for Car Talk to 97.9% for Fresh Air. The stenocaptioner's average across all

segments was 97.2%. The voice writer's accuracy ranged from 63% for Morning Edition to 80.6% for Fresh Air. The voice writer's average was 74.1%.

It is worth noting that each performed best on their Fresh Air segment. This was a long discussion about the Civil War but it was fairly slow-paced, kept to a single topic and had only one guest. The voice writer had more problems than the stenocaptioner with the news topics on Morning Edition and All Things Considered. Though the field is relatively new, there is evidence to suggest that stamina may be a problem for voice writers; it is difficult for a voice writer to perform the process for a long period of time without fatigue and deteriorating quality.

Below is a summary of errors and rates for the four segments.

	Morning Edition	Fresh Air	All Things Considered	Car Talk
Length (TRT)	9:02	16:04	12:32	18:30
Topics	Kenya/ Hurricanes	Civil War Deaths	Primary Coverage	Autos/ Misc.
Verbatim Word Count	1281	2576	2338	3335
Steno				
Word Count	1279	2571	2327	3295
Substitutions	18	31	26	48
Additions	9	7	7	10
Deletions (drops)	10	16	27	55
Total Errors	37	54	60	113
Error Rate	2.9%	2.1%	2.6%	3.4%
Accuracy Rate	**97.1%**	**97.9%**	**97.4%**	**96.6%**
Voice Writer				
Word Count	1207	2429	2027	2882
Substitutions	205	232	258	343
Additions	78	56	70	98
Deletions	191	212	366	363
Total Errors	474	500	694	804
Error Rate	37%	19.4%	29.7%	24.1%
Accuracy Rate	**63%**	**80.6%**	**70.3%**	**75.9%**

Accuracy and readability

Many passages in these transcripts have a significant number of errors, but are still quite readable. (As noted, both stenocaptioners and voice writers are trained to occasionally drop phrases and shorten or rephrase sentences in an effort to produce a comprehensible text.) Some examples follow.

Morning Edition samples

Verbatim	**Steno**	**Speech**
HUNDREDS OF EXPECTANT DISPLACED KENYANS ARE JOSTLING ONE ANOTHER. THEY'RE CAMPED OUTSIDE MOI AIR FORCE BASE, ACROSS THE STREET FROM ONE OF NAIROBI'S BIGGEST SLUMS, MATHARE.	HUNDREDS OF PECK TONIGHT DISPLACED KENYANS ARE JOSTLING ONE ANOTHER, THEY'RE CAMPED OUT SIDE MOI AIR FORCE BASE, ACROSS THE STREET FROM ONE OF NAIROBI'S BIGGEST SLUMS, MATHARE.	HUNDREDS OF THE EXPECTED DISPLACED TENANTS ARE CAMPED OUTSIDE WHILE THE AIR FORCE BASE ACROSS THE STREET FROM NAIROBI'S BIGGEST SLUM.
26 words	2 subs, 2 additions	3 subs, 1 add, 7 drops

Both the steno and the voice writer mistranslated "expectant" and in fairness, it was hard to understand the reporter in that instance. Beyond that, the steno text is quite accurate and readable, with a slight difference in punctuation and splitting the word "outside" into two words. The voice writer text is comprehensible but substitutes "tenants" for Kenyans." She also drops the phrases "jostling one another" and "one of" along with the proper names of the Air Force base (Moi , substituting "there") and the slum (Mathare).

For the stenocaptioner, "peck tonight" (for "expectant") probably represents the worst error in the entire segment. Two other significant errors later in the

segment are “demand ago” in place of “demanding a” and “mag shetties” for “machetes.” The first represents a “stacking error,” as it is known in the court reporting profession and the second is a “fingering error,” or simply a “mis-stroke.” The steno either added an extra key or hit keys so rapidly in combination that keystrokes were combined in an unintended way. Beyond these three errors, the remaining issues are minor substitutions, such as “out side” (two words) for “outside,” and dropping a few non-critical words and phrases, such as “you know.”

For the voice writer, there were more problems. Considerably more text has been dropped and more substitutions made, but even in these cases, much of the meaning is retained. Here is a comparison from later in the Morning Edition segment dealing with hurricanes.

Verbatim	**Steno**	**Speech**
>> ALL THIS DEBATE WE'RE HAVING UP HERE IS A LITTLE IRRELEVANT FOR PRACTICAL, YOU KNOW, COASTAL CONCERNS UNLESS YOU'RE WORRIED ABOUT 200 YEARS FROM NOW, MAYBE. WHAT PEOPLE WANT TO KNOW IS WHETHER THEY'RE GOING TO GET CLOBBERED.	>> ALL THIS DEBATE WE'RE HAVING UP HERE IS A LITTLE IRRELEVANT FOR PRACTICAL COASTAL CONCERNS UNLESS YOU'RE WORRIED ABOUT 200 YEARS FROM NOW, MAYBE, WHAT PEOPLE WANT TO KNOW IS WHETHER THEY'RE GOING TO GET CLOBBERED.	>> ALL THIS DEBATE WE ARE HAVING HERE IS A LITTLE IRRELEVANT FOR PRACTICAL, YOU KNOW, CLOSE TO CONCERNS ON THE SHORE WORD ABOUT 20 YEARS FROM NOW, MAYBE. WHAT PEOPLE WANT TO KNOW IS IF THEY WILL GET CLOBBERED.
38 Words	2 drops	8 subs, 2 adds, 2 drops

In this passage, the only stenocaptioner error is the dropping of “you know.” The voice writer has considerably more errors, yet the passage is still readable.

"Close to" instead of "coastal" is a significant error as is "on the shore word" in place of "unless you're worried" but the most critical error may be the substitution of "20" in place of "200" years. These three errors in one sentence make it very difficult, if not impossible, to follow and you get the wrong time frame from the comments, even if you can interpret the sentence.

Conclusions and Recommendations

Currently, stenocaptioning remains the more accurate system for producing real-time text. With the higher costs and limited availability of good stenocaptioners, a hybrid approach may be worth considering. It may be possible to utilize already existing scripts, some direct voice recognition of on-air talent and some voice echoing to tackle different parts of the problem and create a functional, cost-effective system.