Choosing the Right Technology for your Speech Analytics Project

by Marie Meteer, Ph.D.

Introduction

Speech Recognition technology is an important consideration for any successful speech analytics project. The core operation--finding exactly where a term was spoken in an audio file--has had many characterizations over the history of speech recognition, including wordspotting, spoken term detection and audio mining - but a full speech analytics system builds on that to provide much richer detail about the content, so it can be useful for business intelligence applications. There has been a long-standing debate between the merits of two very different approaches to solving the speech recognition challenge - phonetic search vs. transcription based (often called LVCSR for Large Vocabulary Continuous Speech Recognition[1]). In this paper, we look at the history, research and tradeoffs between these approaches to help you determine the best approach for your speech analytics project. We also look at new techniques coming out of the labs, and whether they will have a significant effect.

The Two Approaches – Phonetics and LVCSR

Phonetic search preprocesses the audio into the possible sequences of sounds or "phonemes" and encodes the result in a lattice of possibilities. Then the search terms are translated into a sequence of phonemes and the search determines whether this sequence is somewhere in the lattice. There are two advantages of this approach. First, the initial processing time is very fast, since the "vocabulary" is just the set of sounds in the language. However, the search is much slower since it cannot be efficiently indexed the way words can. The second advantage is that even if the search term is totally new, such as a name that has just newly been introduced into the spoken language (like the new drug term "cialis"), the term may still be found if that sequence of phonemes exists (" S IY AH L IH S "). The disadvantage is that since there are many possible sequences in the lattice, the term may be found in many places where it was never said (e.g., if the actual words were "see a list").

About Marie Meteer
Dr. Marie Meteer is an independent consultant and Adjunct Professor at Brandeis University with over 25 years experience in speech recognition and natural language processing. She is the founder and Executive Director of the Speech Technology Consortium, a nonprofit organization focused on IP and patent related issues. She was a co-founder and Vice President of Speech of EveryZing, a digital media merchandising platform which uses speech and natural language processing to drive search engine optimization. Previously, Dr. Meteer was at BBN technologies for 20 years in positions ranging from research scientist to commercial business development and management. Dr. Meteer received her Ph.D. in Computer and Information Sciences from the University of Massachusetts.

[1] Also Large Vocabulary Conversational Speech Recognition, this is the term used by the research community, particularly in the context of DARPA funded research.

Transcription-based approaches (LVCSR) transcribe the audio into a sequence of words and then use standard text-based search methods to find the search terms. Since the transcription based approach uses a dictionary of generally 50,000 - 100,000 words and statistical methods to confirm the likelihood of different word sequences (like "the side effect of cialis include" or "cialis pills"), the accuracy is much higher than just the single word lookup of a phonetic approach, so it is more likely that if the word is found, it was really spoken. However, the disadvantage is that the words in the search terms need to be in the dictionary in advance of processing the audio in order to be output by the transcription engine. The initial processing of the audio takes longer than with a phonetic approach because of the large vocabulary, however, searching is instantaneous.

LVCSR – Different than Phonetics?   Or really just MORE than Phonetics?

Before going further, it's important to point out that the comparison of "phonetics" and "LVCSR" is not comparing one technique that uses phonetics and one that doesn't. All LVCSR acoustic models are phonetically based. In both approaches, every phoneme is modeled in its phonetic context (e.g. the "t" between "s" and "r" in "street" is different acoustically than the "t" between the start of a word and "ah", as in "top"). While phonetic approaches do take into account what the possible sequences of sounds are and their frequency (for example, groups of consonants, such as "stldr" never occur in English), they do not take into account any higher level knowledge of the language.

In contrast, in an LVCSR approach, the larger context that the sounds occur in are taken into account as well, so in building a lattice of alternatives, only sequences that occur in real words in the language are included and the likelihood of one word to follow another is taken into account in how likely each alternative is. This compensates for the fact that the sounds are very ambiguous and tend to merge with neighboring sounds (e.g. "dish soap") and the same sequence of sounds can be different word sequences: "let us pray" vs. "lettuce spray". The LVCSR approach algorithmically determines which alternatives are likely in the language, letting the computer mathematically sift through the results quickly and efficiently, rather than requiring a significant manual effort. The knowledge of what words exist and are likely in the target language is a major factor in higher accuracy. The down side is that words that are not in the dictionary are not recognized. Of course, with dictionaries between 50 and 100,000 words, new words are fairly rare.

In choosing the best technology for your needs, the first step is to understand what the tradeoffs are.

Trade Offs

In weighing the options, there are several factors to take into account. The first is precision vs. recall. Precision is a measure of how accurate the items returned are. If the search returns 10 items and 8 match the search term, the precision rate is 80%. Recall is a measure of how many of the items that were actually in the documents being searched were found. So if there were

10 instances of a search term in the documents and the search returned only 6 of them, but didn't find the other 4, it would be said to have a "60% recall rate". Unfortunately, reports of high "accuracy" are often misleading, as a 90% "recall rate" can be achieved, but it might come at the cost of a 10% "precision rate" – meaning that for every 9 of 10 correct results, there were 10x times that many false positives that have to be manually filtered out of the result set. The best systems find a balance between precision and recall that allows for the most optimized result set. Phonetic approaches generally have higher recall, since there are many phoneme sequences to match against, but low precision, whereas transcription based approaches have much higher precision since they are more likely to contain the words that were actually said, but lower recall due to unusual words or recognition errors.

If document (or call) retrieval is more important than finding every place each word is spoken, then it is more important to have higher precision than higher recall. An example of this is finding "calls of interest", where you want to see how often some topic is coming up in a set of calls. If the topic is occurring frequently enough to be interesting, then missing a few calls is not problematic, however, if there are a large number of false positives, the cost of filtering through calls that are not on topic will be much greater. Furthermore high speech recognition accuracy isn't critical either, since an important word in that topic may occur multiple times in a call, however in order to retrieve the call, the target term only needs to be recognized correctly once.

Another element to consider is how frequently new words or names occur in your domain as search terms, and whether or not those search terms are critical to your stated business goal. For example, a legal forensics expert may wish to know where the exact name of the defendant appeared, so as to present this evidence in court, so a phonetic approach which can handle unusual names and has higher recall would be a better approach. However, a business analyst may not care about specific proper names, instead concentrating on just whether a customer was asked their name as part of a credit card validation, or that a call transfer occurred. Typically Out-of-Vocabulary (OOV) terms are rare (by definition, if a term is well-known and often spoken it would already be part of the LVCSR dictionary, or could be added before processing.) In these situations, a transcript-based approach would be better suited.

Another component is how much audio needs to be processed vs. how frequently that audio is searched. For example, in some applications, the goal is simply to have the audio ready to be searched, but only on rare occasions is it actually queried, such as the large volumes of audio that may someday need to be used in a legal audit or forensics search. In those cases, a phonetic approach that can construct a lattice very quickly and stored for possible later search is preferred as the cost of ingesting the audio is low and the high cost of search is less of a factor since it is relatively rare. However, if the audio is being actively searched multiple times as part of competitive intelligence or process improvement, then an LVCSR or transcript approach is preferred, as the indexed text search algorithms are orders of magnitude faster than the linear search of the phonetic string sequences. Indexing phonetics is not useful, as there are only 36-40 unique phonemes, whereas as with 50-100 thousand words, the text index is quite fast and efficient. Phonetics also requires a larger footprint for storage, since a word has an average of 4 phonemes and many variations of each phoneme sequence is stored in the lattice, which may also present an issue for large scale projects or enterprise use.

In the next section, we examine the history of speech recognition development, and survey the results of several formal industry research projects to see where each approach has fared better or worse, and where these two competing technologies are going in the near future (and as important, where the funding dollars and resources will continue to be spent to improve the next generation of speech).

Results from the Research Community

Speech recognition as we know it today began in the mid 1970's with the DARPA funded "Speech Understanding Project" [i], a formal concentration of the leading vendors in the space, which included the likes of research giants BBN, CMU, IBM, MIT, and SRI, all of which are still among the recognized leaders in speech recognition research.

At that time searching for words in audio used pattern matching approaches with "whole word models." A reference example of the word being searched for was extracted from the audio and then turned into a template that was compared directly against the audio with a sliding window until a match was found. This "acoustic match" process was very slow, and only worked well in targeted well-controlled acoustic environments. In addition to the obvious challenges with creating a general enough template to account for different voices and acoustic conditions, the search terms had to be known in advance.

In the 1980s, several breakthroughs pushed speech recognition forward: Hidden Markov Models (HMMs), context dependent phonemes, and more efficient search algorithms (not to mention faster computers). This allowed researchers to expand the dictionary from 100's of words to thousands, and at the same time allow for differing accents, pronunciations, and acoustics. Throughout the 1990s and into the 2000's, significant DARPA funding pushed the LVCSR technology forward quickly, such as on the DARPA projects EARS and GALE[2], with many different sites and annual evaluations to compare progress across sites. Key to the performance improvement was not just better acoustic models, but more constraints in the form of grammars and statistical ngram language models to constrain the search.

As performance improved, keyword spotting followed suit, first using HMM models of keywords and modeling the rest with "fillers" [ii]. However, as the vocabulary size of the recognizer went from hundreds of words to tens of thousands of words by the late 1990's, most audio search moved to simply using the output of the LVCSR systems as input to text-based search algorithms.

Different phonetic approaches were attempted for audio search in the late 1990's, but the phoneme accuracy was very poor without the constraints provided by the dictionary and

[2] EARS: Effective Affordable Reusable Speech to Text (http://projects.ldc.upenn.edu/EARS/); GALE: Global Autonomous Language Exploitation (http://projects.ldc.upenn.edu/gale/index.html)

language models. Ng and Zue [iii] (1999) officially reported a 35% phoneme error rate in their trials, and even more recently, research focused on phoneme accuracy using the TIMIT read speech corpus [iv] was only able to reduce the error rate to 24%. With roughly one in every 4 phonemes wrong, this equates to a statistical 100% word error rate (4 phonemes per word). Phonetic search today tries to address accuracy by searching a lattice of phoneme alternatives, not a single phoneme string, but it still means that at every point there are multiple possibilities and no guarantee that the correct sequence is actually in the lattice. (More “hits,” but also a lot more “False positives.”) Sifting out the good results from the false positives is then a manual time-consuming process afterwards.

NIST Evaluations in Spoken Document Retrieval

The first DARPA sponsored evaluation of audio search was in 1999 as a part of the Text Retrieval Conference (TREC) that was well established in evaluating document retrieval systems. As in the evaluations that pushed speech recognition forward in the early 1990’s, these evaluations were open to multiple sites and run by NIST (National Institute of Standards and Technology), who set down guidelines, specified the test data and scored the results. The evaluation was intended to bring together the speech recognition research community and the information retrieval community to address audio search. While some of the participants had their own speech recognizers and created a unified system, others used the output from an LVCSR system provided by NIST.

Despite the fact that the recognizers at that time had word error rates of upwards of 50% (words, not phonemes), the systems performed quite well, with retrieval rates close to 80% for the top system. One of the reasons for this is that speech recognizers tend to get little words wrong, like “in” and “an”, which are not important for search and retrieval, and get the content words right. Another reason is that important words tend to occur multiple times in the document and it only needs to be correctly recognized once for it to bring back the document.

Overall the evaluation was “declared a success” [v]. While in this evaluation all of the systems were based on LVCSR transcription, there was work going on at the time on using a phonetic approach, for example [vi]. While they showed that phonetic search could theoretically outperform LVCSR, they did not use a state-of-the-art LVCSR recognizer, but rather a desktop dictation engine (IBM ViaVoice) that would not handle the differing acoustical environment and speakers well when compared to more robust research engines such as IBM’s Attila, BBN’s Byblos, and Nuance’s Dragon Recognizer, which are designed for LVCSR and transcription applications.

In the 2006 NIST Evaluation on Spoken Term Detection[3] [vii], which was on telephony speech, the competition was dominated by LVCSR approaches, which scored highest in the evaluation. Eleven sites participated, including sites which had participated before (BBN, IBM, SRI) as well as others, for example, Queensland University of Technology from Australia and Brno University from the Czech Republic.

[3] NIST workshop: http://www.itl.nist.gov/iad/mig/tests/std/

A few of the participants did in fact submit a phonetic recognizer, but admitted the limitations, as in this quote from Wallace [viii] et al. "Further performance improvements … are required to compete with systems that incorporate an LVCSR engine….it [a phonetic system] is prone to high levels of false alarms." Szoke [ix], et al (2005a) did a direct comparison of keyword, phonetic and LVCSR systems and found that "The best accuracy is provided by the system searching in LVCSR word lattices" mainly because the language model significantly improves accuracy and using a word lattice of alternatives compensates for errors. Both point out that the phonetic approach works better for words that are not in the vocabulary, however, these are not frequent enough to give better overall performance.

Mapping Features to Functionality

Evaluations in the research community focus on word or retrieval accuracy, however, a successful Speech Analytics project has other important considerations. From a practical perspective of the cost, both in equipment and human time, there can be significant considerations. There is also the cost and benefit in other capabilities beyond search, such as topic identification, discovery, and clustering. In this section we look at how the various features of these approaches map to the functionality that you need to build a successful system.

Processing speed and search speed

When addressing the speed of a system, there are two distinct operations to be considered. The first is the initial processing of the audio, which for phonetics produces a lattice of phonemes and for transcription produces a sequence of words or a lattice of word alternates. In both cases, this process is only done once for the audio. The second operation is the actual search over this data structure. In most speech analytic applications this will be done quite frequently, and therefore being able to apply indexed and text-based search algorithms is key. For these types of applications, transcription based approaches are more efficient overall. There are some use cases for search-once design, but they are generally limited to legal/forensic search, or as an added feature to an audio archive, and for these a phonetic approach is more efficient.

The un-bearable weight of QA

Regardless of the approach, most applications require that someone check to determine if what is returned is accurate. In a phonetic approach, a user must listen to the selection to determine its relevance. Unfortunately, there is no way to easily skim the acoustic signal without actually listening to the audio. And without artificially speeding up playback, the fastest this can be done is at a rate of 1x times the duration of the audio (30 seconds of audio requires 30 seconds of listening).

An LVCSR approach provides a transcript of the words around the key term allowing users to visually skim the "snippet" where the search term was found and make a determination, similar to how people “skim” the page of Google results to find the web page they are really interested in. Then they can choose to listen for additional information or confirmation only when necessary. Since the standard reading rate is quite high (250-300 wpm) compared to speaking (120-150 wpm), you can manually QA the results of a transcription engine at about twice the rate of a phonetic engine. Paired with the fact that the phonetic engine is more prone to producing more “possible hits” (hits plus false positives), it is a double penalty in time and effort to QA the same data as compared to an LVCSR engine (which is using the “context” of the language model to eliminate as many false positives as possible before the manual QA step begins).

Going Beyond Search

Search accuracy (precision and recall) is not the only reason for deploying a speech analytics solution. Often the desired output is business intelligence and analysis, similar to data mining. Because LVCSR outputs a text string, it can often be fed into these other business intelligence applications directly, enabling the use of NLP algorithms and techniques for topic identification and discovery. These techniques go beyond the ability to search for something that you know in advance and enable new trends and unexpected topics to emerge. Unfortunately, Phonetic strings are not well understood in both text-analytics and data-mining application suites, and are a poor fit for this type of solution, limiting them to search-only applications.

Also, enhanced search, such as query expansion and “more like this”, also requires the use of full text, not just the point in the audio where a term was found. Having words allows the user to more quickly jump to the context of the audio, rather than having to go back and listen to the audio to get context.

Finding New Words

The greatest advantage of a phonetic approach is that words that are not in a predefined vocabulary can still be found. Of course, as mentioned earlier, as recall goes up, precision goes down and many similar sounding words will be found as well. This is particularly a problem for short words (like “toe,” which would get confused with “tow,” ”tow boat,” “tomato,” “potato.” Only if the queries are very distinctive, such as long fixed phrases with no variation, can the precision of a phonetics solution be higher.

On the flip side of the accuracy coin, the degree to which out of vocabulary words affects performance depends on the application. According to a study by Logan [x], in news sampled from November 1999 to June 2000, the OOV rate was roughly 1.3%. While this is very low, when they looked at user queries, they found that the OOV rate on the queries was 12.6%. This difference is understandable in broadcast and newspaper media, since in general what is most interesting in the news will be names of people and places, and many of those will be new to the general population. However, In the NIST 2006 evaluations, the domain was generic telephony conversations and the search was on conversational and topical words, most of which were in the vocabulary so the OOV rate was around 0.5%.

The size of the vocabulary of the recognizer is an important consideration. Woodland [xi] compares OOV rates using different vocabulary sizes on a 10 hour subset of broadcast news, ranging from 1.6% OOV with a 55K word vocabulary up to 22.3% on a 3K word vocabulary. Most speech systems provide tools to add words to the vocabulary, so one solution is to simply add the domain specific words to the dictionary. This will allow that sequence of phonemes to be recognized since the underlying models of all of the phonemes in the language have already been trained. On the other hand, if these new words are frequent and important in a specific domain, for example if they are the primary product names of a client, then transcribing some text and training the language model (which captures word sequences) is also easily done. Again, the underlying phonetic models don't need to be changed. In news media, the words that are most important will recur over multiple days and weeks of news. One approach is to use text based news to flag new words, which can then be quickly added to the dictionary, keeping the overall OOV rate low.

Another approach is to map the query term to a word that is in the dictionary. For in the SRI/OGI Spoken Term Detection System [xii], they used an approach called "term mapping" that first translates the word to its phonetic spelling and then finds a match in the dictionary allowing for certain substitutions of phonemes and both partial and combinations of words. For example if the search term is "inkjet" then just separating the terms into "ink jet" will solve the problem. When "Verizon" was formed, searching for the word "horizon" would frequently find it, and in our earlier OOV example of "Cialis", the phrase "See Alice" would correctly return similar results.

Research Directions

Current research in speech recognition overall is focused on bringing down the word error rate, which directly impacts spoken term detection since more words are recognized accurately, and on making LVCSR systems faster, which improves scalability by reducing the initial processing time for the audio. Both of these directions make transcription approaches even more attractive for speech analytics.

Research directly on spoken term retrieval has focused on the problem of identifying words that are not in the dictionary. All begin with large vocabulary systems since all evaluations have shown that when the search term is in the vocabulary, this is the most effective approach. While some have looked at using subword units that are bigger than single phonemes (e.g. syllables or other word parts that can be combined later as phrases, like the "un' and "ly" syllables of "unfortunately" → "un fort you gnat ly") the biggest gains have been in hybrid systems that mix a large vocabulary recognizer with subword recognizer.

SRI [xiii] used this hybrid approach, combining the most frequent words in the vocabulary with "graphones" [xiv], which are automatically created subword units from the text which are then given pronunciations in the dictionary. These are then combined with the most frequent words into a new language model. They found that the errors from OOV were cut in half by using the hybrid system.

An alternative hybrid approach is to create multiple models, a word model and a subword model, then select which to search on depending on whether the word is OOV, which can easily be determined by looking it up in the dictionary. Work done at Cambridge Research Systems [xv] explored a variety of combinations of transcription types:

- Word recognition
- Subword "particles" which were automatically determined from a corpus
- Phonemes

Then depending on whether the search query terms were "in" or "out" of vocabulary, results could be combined from each model, with the in-vocabulary terms sent to the known dictionary model, and the OOV terms processed by the sub-word and phonetic expansion model. As research continues in this direction, we can expect to see these techniques introduced into commercial systems. Hybrid systems in particular can be built with little change to the architecture of a modern LVCSR system.

Both phonetic and LVCSR systems make use of "fuzzy matching" or query expansion, which allows for confusable phonemes to be substituted. Cambridge University included expansion of the query based on a confusion matrix, but only when the word was not in vocabulary since it lowered accuracy on words that were in vocabulary.

Conclusion: Making Choices

As with any technology choice, the final decision comes down to looking not at what the "best technology" is, but rather what is "best fit" for your business problem in terms of cost, value, and manual effort. In general, phonetics is an appropriate fit for search-seldom applications, like large archives and legal forensics. For larger enterprise application like speech analytics and business intelligence mining, the LVCSR approach is better suited.

The following chart provides some quick comparisons that are useful for determining what your true requirement may be.

Phonetic	Transcription/LVCSR
Looking for rare events or words that may only happen once.	Searching the audio for things that are relatively common in your domain.
Processing very large amounts of audio quickly, but only if infrequently searching	Searching many times over the same audio.
Cost of missing an instance of the term being searched is high, like "bomb."	Cost of sifting through many false positives is high, like "Obama," "bombardier," "that movie was da bomb"
Cost of false positives is low.	Cost of missing some instances is low.
Main activity is limited to search, that is finding	Main activities include search, but also broader types of

instances of specific terms.	analytics and discovery
Audio-searching only required (phonemes are not words)	Audio and text-related searching and analysis required (text, chat, email, surveys, etc.)

Best Uses for Phonetics

- Large Archives
- Legal or Forensic Search

Best Uses for Transcription/LVCSR

- Large-scale Speech Analytics Implementations
- Marketing and Business Intelligence
- Multi-channel Data-mining and Search
- Topic Identification and Discovery

References

[i] D.H. Klatt, “Overview of the ARPA speech understanding project”. In Lea, W. ed. *Trends in Speech Recognition,* Englewood Cliffs, NJ: Prentice-Hall. 1980.

[ii] J.R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, M. Siu, "Phonetic training and language modeling for word spotting," ICASSP, 1993.

[iii] K. Ng, and V. Zue, “Phonetic Recognition for Spoken Document Retrieval”, In Proceedings of ICASSP 98, pp. 325-328. 1998.

[iv] P. Schwarz, P. Matjka, and J. Cernock. Towards lower error rates in phoneme recognition. In *Proc. of 7th Intl. Conf. on Text, Speech and Dialogue*, number ISBN 3-540-23049-1 in Springer, page 8, Brno, 2004.

[v] J. Garafolo, C. Auzanne, and E. Vorhees, “The TREC Spoken Document Retrieval Task” A Success Story”, in NIST Special Publication, number 240 pp.83-92, 1998.

[vi] Peter S. Cardillo, M, Clements and M.S. Miller. 2002. “Phonetic Searching vs. LVCSR: How to Find What You Really Want in Audio Archives”, INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY, Volume 5, Number 1, 9-22. 2002.

[vii] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," in Proc. 2007 SIGIR Workshop on Searching Spontaneous Conversational Speech, Amsterdam, July 2007.

[viii] R. Wallace, R. Vogt, and S. Sridharan, “A phonetic search approach to the 2006 NIST spoken term detection evaluation,” in Proc. Interspeech’07, Antwerp, Belgium, 2007.

[ix] I. Szoke, P. Schwarz, P. Matejka , L. Burget, M. Fapso, M. Karafiat, J. Cernock, “Comparison of keyword spotting approaches for informal continuous speech”, Proc. Interspeech. Lisbon, Portugal, September 2005

[x] B. Logan, P. Moreno, J. Thong, E. Whittaker, Ed (2000): "An experimental study of an audio indexing system for the web", In *ICSLP-2000*, vol.2, 676-679.

[xi] P.C. Woodland, S.E. Johnson, P. Jourlin, and K Sparck-Jones, “Effects of out of vocabulary words in spoken document retrieval”, In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000.

[xii] D. Vergyri, I. Shafran, A. Stolcke, R. Gadde, M. Akbacak, B. Roark, and W. Wang, “The SRI/OGI 2006 Spoken Term Detection System,” in Proc. Interspeech, 2007, pp. 2393–2396.

[xiii] M. Akbacak, D. Vergyri A. Stolcke A. “Open-vocabulary spoken term detection using graphone-based hybrid recognition systems” Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.

[xiv] M. Bisani and H. Ney, “Open vocabulary speech recognition with flat hybrid models”, *in Proc. of Interspeech*, pp. 725–728, Istanbul, Turkey, 2005.

[xv] B. Logan, J. V. Thong, and P. Moreno, “Approaches to reduce the effects of OOV queries on indexed spoken audio,” IEEE Trans. on Multimedia, vol. 7, no. 5, pp. 899–906, October 2005.