VOICE/AUDIO INFORMATION RETRIEVAL: MINIMIZING THE NEED FOR HUMAN EARS

Mark Clements[1] and Marsal Gavaldà[2]

[1]Georgia Institute of Technology and [2]Nexidia, Inc.
Atlanta, Georgia, USA
clements@ece.gatech.edu, mgavalda@nexidia.com

ABSTRACT

This paper discusses the challenges of building information retrieval applications that operate on large amounts of voice/audio data. Various problems and issues are presented along with proposed solutions. A set of techniques based on a phonetic keyword spotting approach is presented, together with examples of concrete applications that solve real-life problems.

Index Terms— Phonetic search, speech recognition, information retrieval

1. INTRODUCTION

Information retrieval, and its closely allied disciplines of document retrieval and data mining, have long been at the leading edge of utilizing technological advancements in areas such as computation, storage, communications, databases, signal processing, pattern recognition, and machine learning. In practical terms, the goal has always been to remove the burden of analysis from a human consumer. More recently, the goal has been to add some modicum of control to an almost unmanageable overload of information. Of the major media sources —text and text-like metadata, images and video, and voice and audio— text-based information retrieval is by far the most advanced. Typical tasks involving text include creation of searchable text indices, creation of text-based natural language interfaces, document retrieval, summarization, clustering, translation, trend-spotting, language identification, sentiment and tone of a communiqué (e.g., terse, jovial, annoyed), and clandestine observations, to name a few. The data can come from internet sources, databases, emails, faxes, text messages, broadcast transcripts, legal documents, one's own computer files, and a host of other sources. We mention these foregoing tasks and data sources simply to emphasize that each of these has a correlate with voice and audio, but with much less capability for performing the desired operations.

While text and video/images can be scanned efficiently by a human consumer to verify results (e.g., a web search engine), audio cannot be rapidly displayed. The logical choice for voice information retrieval, therefore, would be to produce time-aligned transcripts for the voice signal and then bring to bear all the advanced tools of text-based processing. Manually created transcripts are sometimes available in the form of closed-captioning or from court reporters, but in general they are not. Automatically generated transcripts often include large numbers of word errors, which must be allowed for in any information extraction process. Further, a great deal of information may be missing, e.g., the identity of the person talking, his or her gender, the channel characteristics, the background conditions, the stress and intonation pattern, the accent, the speaking rate, the proficiency with the language, as well as other voice characteristics. Hence annotated text, such as diarization [1], would be desirable. But, even with these embellishments, the text form will never be as rich as the audio waveform itself.

In any information retrieval problem, be it an internet search engine, an interactive help menu, or even directory assistance, there is generally a fundamental tradeoff between finding everything you want *(recall)* and omitting things you do not want *(precision)*. Setting the proper operating point continues to be a challenge in any system.

In this presentation we discuss one approach to the voice/audio information retrieval problem related to real-life applications, along with some interesting observations and challenges.

2. DEFINING THE PROBLEM

We deviate perhaps from the classical definitions of information retrieval in that we define the overarching problem to be that of reducing the dependence on human ears for extracting information from an audio stream or record. Examples of such sources include broadcasts, web

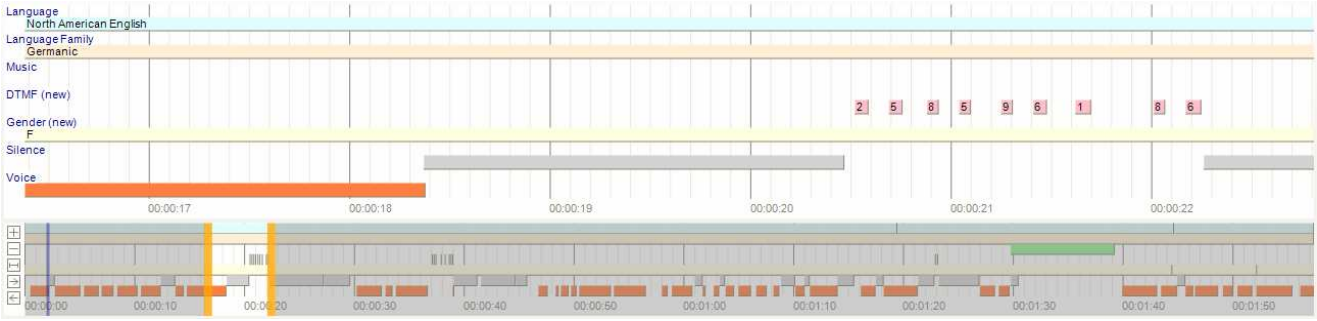

Fig. 1: Example of a media file richly annotated by detector and classifier tracks such as language, language family, music, DTMF, gender, silence, and voice activity. The interactive control consists of two panels: the lower one offers a "bird's eye view" of the entire media file (about 2 minutes long in the example), whereas the top panel shows only the zoomed-in portion, corresponding to the segment delimited by the vertical orange bars in the lower panel (about 7 seconds long in the example).

postings (such as YouTube videos), voicemails, recorded voice interactions, music, telephone recordings, and others. Desired information to be retrieved can include a wide array of descriptors that may have little do with the actual words spoken. For instance, it may be valuable to determine automatically what fraction of time in a recorded voice interaction was "hold music." Given this definition, virtually the entire universe of speech and audio signal analysis tools prove to be valuable.

3. CHALLENGES

There are three broad classes of problems that must be solved, all of which are important for the overall problem. The first set of issues relates to the *base-level signal processing*. These include not only the methodology of extracting the desired descriptors, but also such measures as accuracy, confidence, pre- and post-processing speeds, latency, and storage requirements. The second set of issues relates to the *user interface* and how the results are displayed in a useful fashion to consumers whose level of expertise may widely vary. The third set of issues relates to how to impose a higher intelligence on the extracted information to assist in actual *decision making* for custom applications.

3.1. Base-level Analysis: Phonetic Word Spotting

One set of solutions, explored by Nexidia and others [2,3], is based on a phonetic analysis of the incoming speech, along with extraction of other key descriptors. The reasons for this choice are numerous and include such factors as processing speed and latency, rapid adaptation to new languages, and completely open vocabulary. (There are both obvious advantages and disadvantages for the speech-to-text approach, but these will not currently be discussed at this time.) In phonetic word spotting, certain behaviors must be accommodated such as the confusion of homonyms in a string, the existence of false alarms (especially short audio events), and slower search speed than for text. The underlying processing comprises two phases – pre-processing and searching. The first phase pre-processes the input speech to produce a phonetic search track, which is searched during the second phase to find the query term(s). Pre-processing is performed only once for a given media segment, and stored or archived. Searching, on the other hand, is typically performed every time to locate the temporal offset(s) of the query term(s) within the audio file. The original waveforms are not involved at all during searching and could be discarded if desired in favor of compressed representations.

3.1.1 Pre-processing, acoustic model, and phonetic grammar

The pre-processing phase begins with *format conversion* of the input speech into a standard representation for subsequent handling. Then, using an *acoustic model* and *phonetic grammar,* the *pre-processing engine* scans the input speech and produces the corresponding *phonetic search track.* An acoustic model represents characteristics of both an *acoustic channel* and a *natural language*. Channel characteristics include frequency response, background noise, and reverberation.

A phonetic grammar likewise depends upon the natural language in use (particularly the set of phonemes used to represent basic sounds and meanings of the input speech). This grammar is used to identify likely end points of words in the input speech (although one should note that words *per se* are not generated during pre-processing, unlike in large vocabulary continuous speech recognition (LVCSR) systems).

Fig. 2: Example presentation of search results for a media file. The left panel shows the hits in tabular form, whereas the right panel superimposes their offset on the player's timeline (vertical red bars).

3.1.2 Phonetic search track

The end result of phonetic pre-processing of a media segment is a *phonetic search track* – a highly compressed representation of the *phonetic* content of the input speech (which cannot be reconstructed from the search track). Unlike LVCSR, whose essential purpose is to make *irreversible* (and possibly incorrect) bindings between speech sounds and specific words, phonetic pre-processing merely infers the likelihood of *potential* phonetic content of sounds (thereby deferring decisions about word bindings to the subsequent searching phase). Produced by phonetic pre-processing and required for phonetic searching, phonetic search tracks are tangible artifacts that can be treated as metadata, associated and distributed with the originating media segments, produced in one environment, stored in databases, transmitted via networks, and searched in another environment. They are distinct from the word or sub-word lattices often generated as an internal step in LVCSR.

3.1.3 Searching, keyword parsing, phonetic dictionary, spelling-to-sound, and search results

The searching phase begins with *keyword parsing* of the query term, which is specified as text containing one or more:

- words or phrases (e.g., "Osama bin Laden")
- phonetic strings (e.g., "[B IY T UW B IY]", the six phonemes representing the acronym "B2B")
- temporal operators (e.g., "brain cancer &15 cell phone," representing two phrases spoken within 15 seconds of each other)

A *phonetic dictionary* is probed for each word within the query term and it typically contains unusual words (whose pronunciations must be handled specially for the given natural language) as well as very common words (for which performance optimization is worthwhile). Any word not found in the dictionary is then processed by consulting a *spelling-to-sound* model that generates likely phonetic representations given the word's orthography.

After words, phrases, phonetic strings, and temporal operators within the query term are parsed, then actual searching commences. Multiple phonetic search tracks can be scanned at high speed during a single search for likely phonetic sequences (possibly separated by offsets specified by temporal operators) that closely match corresponding strings of phonemes in the query term. Recall that phonetic search tracks encode potential sets of phonemes, not irreversible bindings to sounds. Thus, the matching algorithm is probabilistic and returns multiple results, each as a 3-tuple:

- *Search Track* (for the media segment probably containing the query term)
- *Temporal Offset* (of the query term within the media segment, accurate to 0.01 second)
- *Confidence Level* (that the query term occurs as indicated, between 0.0 and 1.0)

Even during searching, irreversible decisions are postponed. Results are simply enumerated, sorted by confidence level, with the most likely candidates listed first. Post-processing of the results list can be automated. Example strategies include hard thresholds (e.g., ignore results below 90% confidence), occurrence counting (e.g., a media segment gets a better score for every additional instance of the query term), and natural language processing (patterns of nearby words and phrases denoting semantics).

As in any detection-based information extraction process, many descriptors are required to fully describe accuracy. Plots and graphs would include ROC curves, DET curves [4], etc. Although such measures are of limited value without context, a few key numbers for Nexidia's implementations are presented for comparison purposes only. The Figure-of-Merit (FOM) number is computed by measuring the probability of detection of a search term averaged over the conditions of 0 to 10 false alarms per

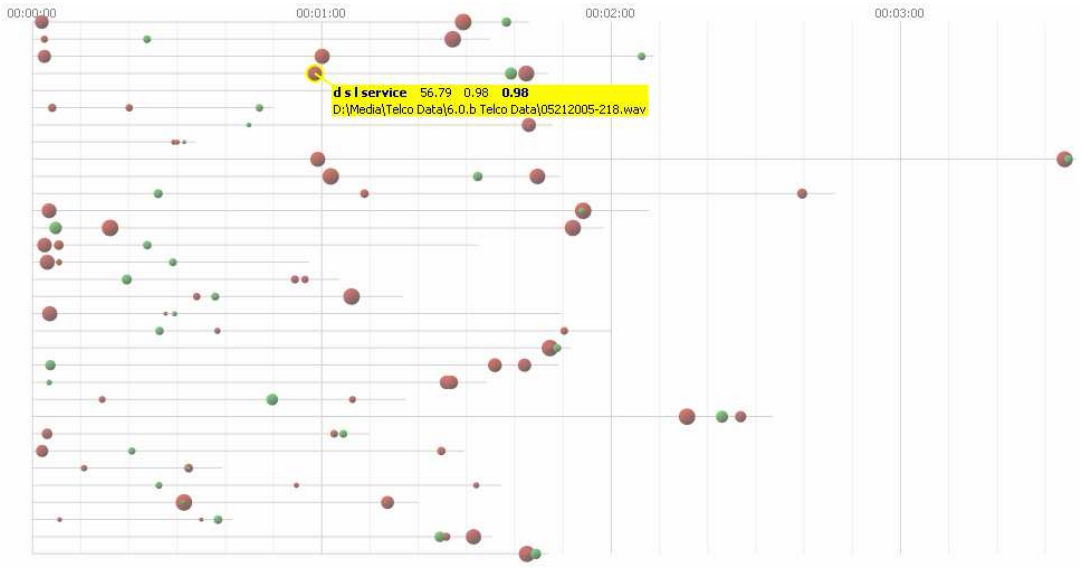

Fig. 3: Another view into a search result set. In this case, each horizontal line corresponds to a media file (thus the different lengths) and the hits are depicted by a sphere whose radius is proportional to the confidence score and whose color is determined by the query (only two in this case).

hour of audio. Here the search terms are varied from 4 to 20 phonemes in length. For broadcast news materials, this number is 87% (consistent across English, Modern Standard Arabic, etc.). For *Switchboard* landline and cellular, the numbers are 77% and 71% respectively. For mixed telephony compressed audio using G.726 at 6 kHz and 2 bits, the FOM drops to 69%, and for actual input from a production call recorder, a typical FOM score is 56%. Finally, typical FOMs for speech captured from a far field microphone during a live meeting are in the range of 45%. One should note, though, that in the latter two cases, the quality is sufficiently bad as to cause great difficulty for even a skilled human transcriptionist. Given this progression in audio quality versus accuracy, the degradation is reasonably graceful.

Another result is that the FOM shows a marked improvement as the length of the search term increases. Short terms (e.g., those consisting of only four phonemes) will produce many more false alarms than will longer terms. FOMs for terms of 20 phonemes in length give FOMs in the high 90s. These numbers, when verifiable, are very consistent across languages. A key lesson in these measurements is that real-world data often stresses the system more than do laboratory databases. But, even in the most hostile conditions, useful information can be extracted, especially if one avoids short queries.

3.1.4 Application of additional processing, annotation, and filtering

In addition to creating phonetic search tracks, the process of indexing a media file can also include other types of base-level analyses, such as:

- Voice activity detection
- Silence detection
- Music detection
- Speaker turns
- Segmenting talkers in two-way voice signals
- Language and/or language family identification
- Accent or dialect analysis
- Gender detection

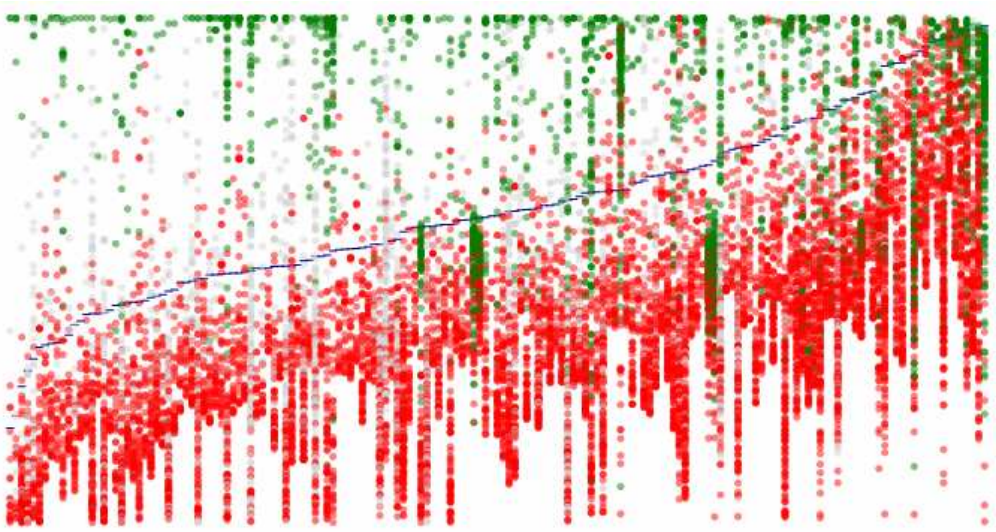

Fig. 4: A more advanced visualization of search result sets which includes "dispositioning" information, i.e., truth marks. In this case, each dot corresponds to a hit, color-coded according to its dispositoned value: red for a false alarm, green for a true positive, and gray for a phonetic partial match (as in matching "sixty" when searching for "sixteen"). Each column (X-axis) corresponds to a search term, and the Y-axis corresponds to the confidence score. In this view, the search terms are sorted by their automatic threshold (horizontal blue line), an attempt to automatically separate true hits from false alarms.

- DTMF detection and decoding
- Number string detection
- Fax/modem signal detection
- Other captured meta-data

Applying this array of detectors and classifiers results in richly annotated media files, as depicted in Fig. 1. Note that almost all of the descriptors are binary, lending themselves to detection strategies.

3.2. Presentation layer: Informative display of result sets

Once the results for a search term or query have been obtained, there are different ways of presenting them to the user. Fig. 2 shows a very common interface, where the results are presented in both tabular form and superimposed on a timeline. Fig. 3 extends the idea to a 2-D plot to display multiple media files, and Fig. 4 adds dispositioning information (i.e., whether the hits are true or false positives) to the visual representation of a large result set.

Post-processing can also be manual – particularly for short lists of results. Highly optimized and efficient human interfaces can be devised to sequence rapidly through a list, to listen briefly to each result, to determine relevance, and finally to select one or more utterances that meet specific criteria. Depending upon available time and importance of the retrieval, the list can be perused as deeply as desired.

Fig. 5 shows an interface for interactive dispositioning and convergence to the optimal sequence of phonemes to represent a query.

3.3. Decision-making: Applications solving real-world problems

The applications of information retrieval technologies for voice/audio media are varied and numerous. The most obvious one is *direct search*, i.e., using search terms and structured queries directly, in order to locate occurrences of those expressions in the audio. Typical scenarios for direct search are *forensic investigations*, e.g., combing through an archive of voice mails to find occurrences of certain phrases ("Grandma Millie," as an example of an actual coded message on one customer's site), or any situation where users perform searches to *navigate* the media, to jump to the relevant segments within the audio or video files (as when searching for, say, "taxi cab driver" to locate a particular story within a long television newscast). Additionally, if the search is conducted on a stream of audio (such as a live radio broadcast or a telephone call that is still ongoing), the application is said to perform *live monitoring*, which can

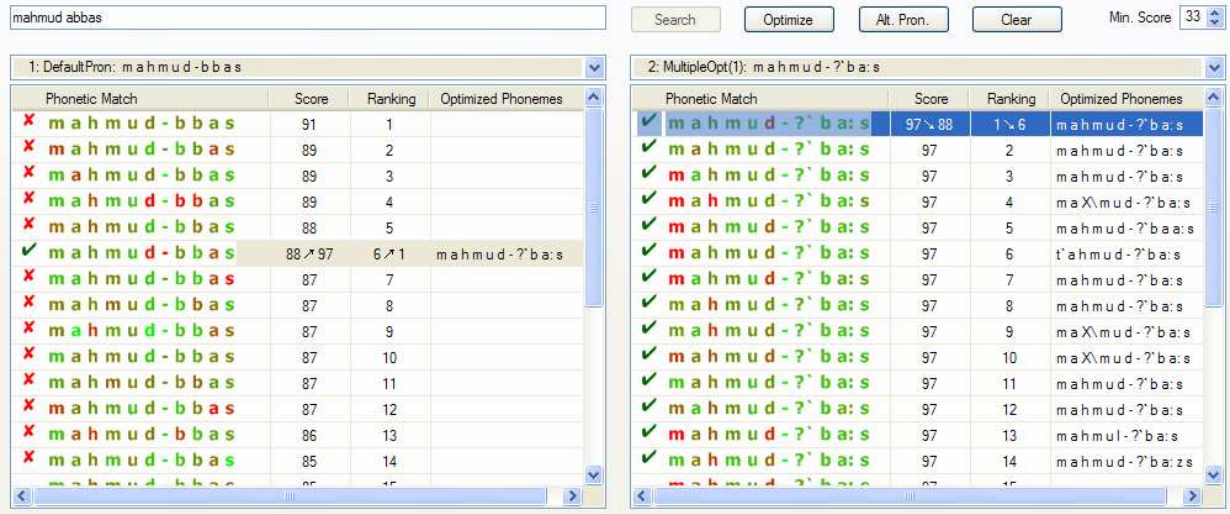

Fig. 5: Example of interactive dispositioning that leads to a more accurate sequence of phonemes to represents the query at hand. In this case, after marking the 6th hit as correct, the engine proposes a variation in the phoneme sequence that allows the user to find many more occurrences of the target search term.

trigger live alerts when the phrases of interest are spoken. Note that search expressions can be known in advance by the system (e.g., in a call center deployment, there can be a global list of search terms to be applied against all incoming calls) or be totally ad-hoc.

3.3.1 Interactive secondary search

The information retrieval experience can also be highly interactive. Indirect or secondary search refers to employing the occurrences of search expressions (i.e., the search results) as a starting point for more complex tasks, such as filtering, classification, ranking, and more elaborate statistics. *Filtering* refers to selecting a subset of media from a larger collection by presence or absence of a search expression, e.g., finding all the calls containing "cancel my account." Typically it is an iterative process, as one can, for example, refine that set of cancellation calls by searching for names of competitors within that set only, or apply set-theoretic operations (intersection, union, difference) on such sets of results. Closely related to filtering is *classification*, where results for predefined search expressions are used to automatically assign audio records to different categories (e.g., fire vs. police vs. medical in a emergency call center), either as a hard, binary decision (call X either belongs or does not belong to category Y) or as a soft one (call X belongs to category Y with probability Z). This type of soft classification or scoring can be used for *ranking*, where audio records are ordered by relevance to a particular area of interest (modeled via search expressions). For example, when intelligence analysts are faced with a pool of thousands of intercepted calls, they can specify a few search expressions relevant to their inquiry, let the system rank the calls accordingly, and thus conduct a much more targeted and efficient investigation. More generally, arbitrarily complex *statistical analysis* can be performed on search results, including establishing positive or negative *correlations* between search expressions (e.g., one may discover that the "customer satisfaction" metric is in fact negatively correlated with the "agent follows script" metric) or computing *trends* by tracking such metrics over time.

3.3.2 Pronunciation optimization

In one interactive application of phonetic technology, one can conduct searches in a language one does not speak by engaging in a pronunciation optimization dialogue with the system, where the user starts with an approximation of the target search term (e.g., grossly misspelled, or lacking diacritics), then listens to the search results and marks the few correct hits, whereupon the engine performs a correction of the phoneme sequence for the target search term based on the true hits, which leads to finding many more correct hits, etc, eventually converging onto the optimal pronunciation (typically only after one or two optimization rounds, see Fig. 5 again for an example). Fig. 6 presents an example of the phoneme graph that underlies the computation of alternative phoneme sequences.

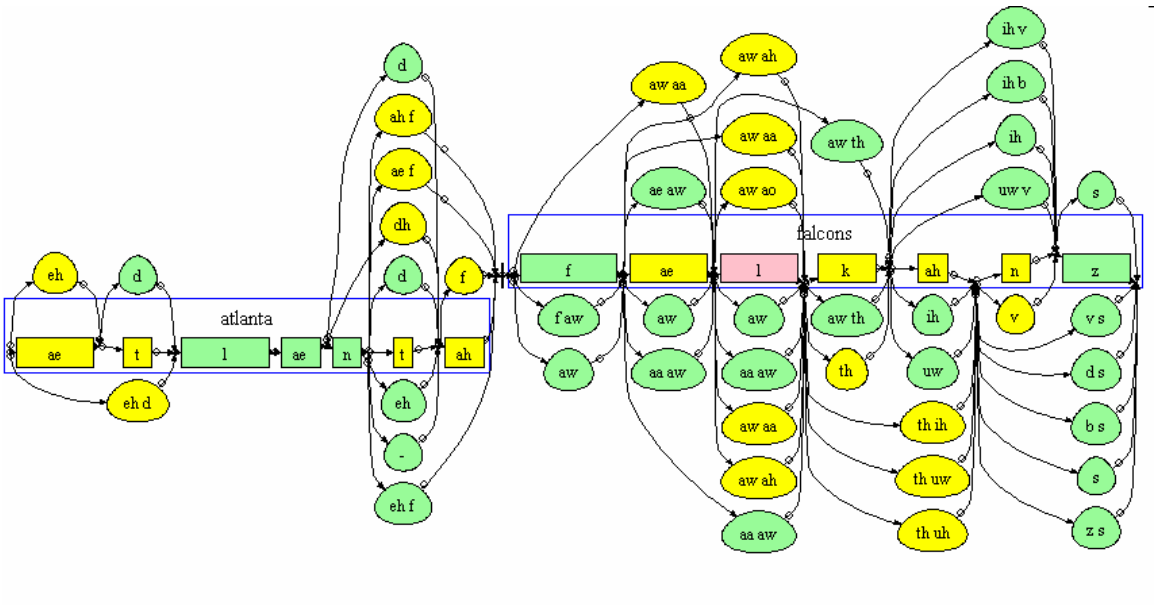

Fig. 6: Example of a phoneme lattice such as the ones that underlie the computation of alternative sequence of phonemes used for pronunciation optimization and repeated search.

3.3.3 Repeated search

In a process similar to that described above, a user can find one particular instance of a successful query and repeat the search based on a "find it again" basis. An alternative phonetic rendering of the target query is derived and substituted for the original rendering. This procedure can be illustrated by the following example. In a body of data related to foreign policy, a query was run on the phrase "nuclear weapons," and multiple correct detections were extracted. One particular correct hit had been uttered by a southern politician that had pronounced it as "nucular" (i.e., "NEW-cue-lerr" (IPA: ['n(j)u:kjə.lə(ɹ)]) rather than "NUKE-lee-ar" (IPA: ['n(j)uk.li.ə(ɹ)])). When the "find it again" process was invoked, a phonetic rendering that better matched the target utterance was derived, and the search process repeated. The results were similar to those found earlier, but with this particular talker's utterances ranked higher than others. This method is not a speaker identifier, per se, but it is highly correlated.

3.3.4 Clip spotting

There is often a need to find audio clips in a long stream (weeks or even months) for compiling statistics, verifying compliance, collecting royalties, etc. Such clips can include music, voice, sound effects, or both, and may have noise or other distortion added to the data. A text rendering of the audio stream would be of limited utility for this task, and a search on the original audio would in general be prohibitively slow. A demonstrated solution is to perform the search on the *phonetic search track.* Although little if any direct meaning can be assigned to the phonetic tracks of non-speech sounds, they are sufficiently consistent that target clips of two or more seconds in length can be found with almost perfect accuracy.

3.3.5 Voice user interface

An added layer of difficulty occurs when the query is input using voice rather than a keyboard (we ignore the scenario of uttering the letters themselves). This is desirable if one wishes to retrieve items from voicemail archives or as an assistive aid to the blind (described below). One approach is to use the N-best automatically derived phonetic renderings of the input and search on all of these, with suitable weighting. Depending on he application or the user preferences, the operating point can be set in such a way as to virtually assure the user that the desired query terms will be found. Testing shows that decreased accuracy always results from such input, under even the best of conditions. When the inputs were simulated by clipping snippets of words directly from the *Switchboard* audio, FOMs decreased by 40% absolute. Under live conditions with speaker adaptation, however, drops on the order of only 10% were observed.

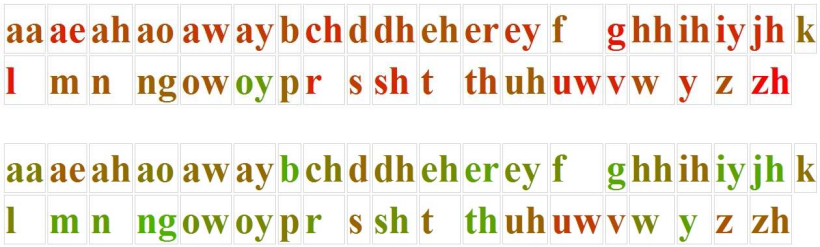

Fig. 7: Sample output from the phonetic scoring portion of a language assessment application. The tables show the average score for each phoneme across all the words in the script, color-coded in a continuous gradient from pure red to pure green. The top table corresponds to a low-scoring speaker, the bottom table to a high-scoring speaker.

3.3.6 Query building

In higher-order information retrieval, one must examine more than individual words and phrases in order to classify audio data. The best query method would be by natural language [5] which is making rapid advances. More simply, classifying text files can often be accomplished by using Boolean search that involves using nested ORs, ANDs, and NOTs along with word separation proximities. It would be attractive to be able to use this vast body of knowledge directly on the audio itself, but without transcripts for training, other methods must be employed. If the underlying processing results in keyword detection along with time delimiters and confidence scores, the Boolean search can use time separation and can include the confidence levels of the various terms (as opposed to text where each term is either present or absent). Further, the training set might consist of a set of audio records that are classified, but not transcribed, putting the query writer at a severe disadvantage in forming any statistical basis for the search. Here, useful queries can be derived interactively, where rough approximations of possible terms and structures are posited for a training set, and then automatically modified for better classification performance. Note that this can be done in individual environments where the audio quality is of a particular nature and the dialog context has various levels of constraints. The major advantage of this strategy is to diminish the required expertise for a query writer. Typical end products of this process are queries that optimally balance terms' detectabilities versus their importance for classification.

3.3.3 Latent semantic indexing

Latent semantic indexing (LSI) is also a popular method for information retrieval of text [6]. Text representations of the training sets are required to train such models, and as such, a transcription of each record must exist. For analysis, an automatically transcribed version is then useful for classification or clustering. In practice, imperfect transcripts still work, but only if the error rates are within reasonable limits [7]. Keyword spotting can also be used in the analysis (assuming the training was performed with correct text), but the important modification that confidence levels must be used in the final weighting.

3.3.4 Language proficiency assessment

Another application is the automatic assessment of a speaker's proficiency in a particular language. Although as a first impression this task might appear not to fall within the audio information retrieval umbrella, by our earlier definition of removing human listening from a process, it does. This current application of phonetic technology that solves this real-life problem is a system that automatically ranks speakers according to their pronunciation and fluency, which is useful for call center companies that need to screen thousands of applicants for agent positions. A script is prepared that covers all the phonemes of the language, and the agent applicants are recorded reading it, at which point the "language assessor" system analyzes it to compute the pronunciation and fluency scores. The pronunciation score models how "native" the speaker sounds, and can be computed as a function of the average/median scores for each phoneme across the script. The fluency is typically computed as the ratio of the sum of the durations for each word in the script over the total duration, thus penalizing false starts and other speech disfluencies. Absolute speed (as in phonemes per second) can be added into the final, global score as well. Fig. 7 presents phoneme scores for two speakers of different ability. Worth noting is the fact that a necessary component for this system is transcript synchronization, whereby each applicant's speech can be time aligned with the presented materials. This process is described below.

3.3.5 Transcript synchronization

Although synchronization of text with audio is a well-explored application in the ASR domain, where *forced*

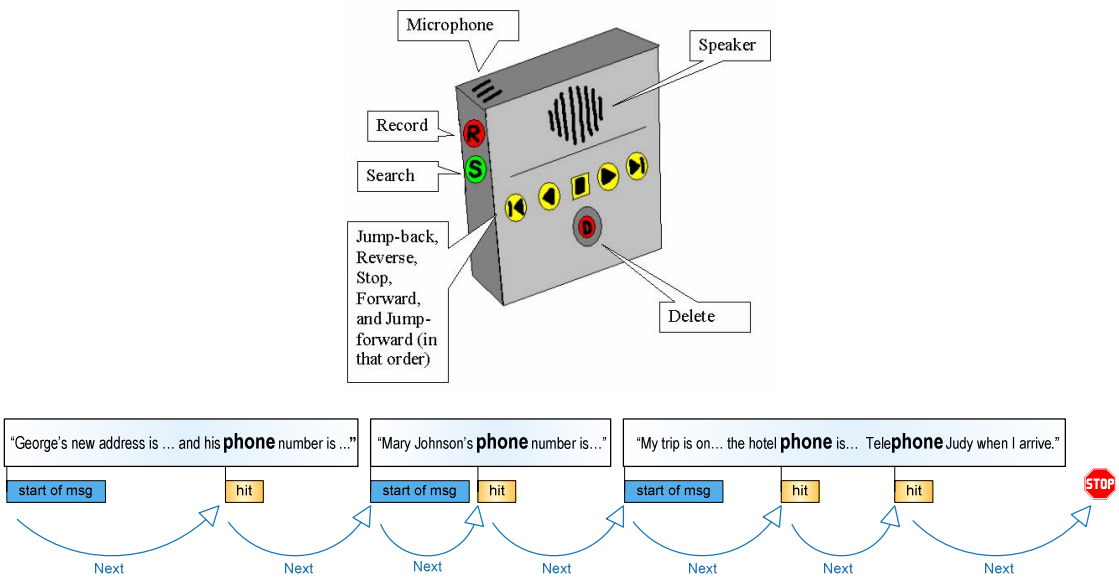

Fig. 8: Sketch of the device for note retrieval for the blind (top), and example of the navigational steps to find the occurrences of "phone" (bottom).

recognition and Viterbi alignment is common [8], the keyword spotting approach is much more robust to transcription errors, disfluencies, and omitted content. This is particularly important for talkers that may be experiencing difficulty in a test condition, as in language assessment. Interestingly, this robust alignment process has found application in film production where synchronization to the movie script is used to locate and align all the different shots (angles, takes) of the same scene. It is also used in the legal community, where video testimony is required by law to be time aligned with non-verbatim court transcripts.

3.3.6 Other information retrieval tasks

When confronted with customers who wish to further minimize the need for human supervision, a few unexpected requests emerge. One example is that of assessing how long originators of calls are put on hold. An obvious solution would be to electronically relay that the hold button was pushed, or if this is not available, to process the audio to detect the hold music. But as the service representatives learned that hold times were being monitored, they simply placed the phone on their desks. Hence, a new piece of information was necessary to detect. As a matter of practicality, it would be unreasonable to require a complete reprocessing of an audio archive to update a condition for search. Therefore, examples as the one described above should be detectable using the existing search track.

3.4 Usability Study: Note Retrieval for the Blind

Let us now describe an interesting proof of concept that was designed for personal information retrieval where the only available medium was audio. Individuals with severe visual impairments often have difficulty taking and retrieving notes. Currently available assistive devices are little more than portable voice recorders with a file structure. To retrieve stored information, the user must remember which note contains the desired information. Focus groups had earlier determined that this community desired a better mechanism so that more notes could effectively be managed and retrieved over a longer period of time of time. The solution, whose interface is sketched in Fig. 8, integrated voice storage along with keyword spotting to enable content addressable notes. The interface was designed to parallel the functionality of existing voice recorder devices, with which the users were already familiar. Although the actual tested device was a mock-up of the desired interface controller connected to a laptop computer, this fact was transparent to the users during the study. When users input recorded messages, a phonetic track was stored in parallel for later access. There was no keyboard interface and queries for retrieval were entered by voice. Nine visually impaired (VI) subjects were trained on the 8-button interface which also included distinctive speech and non-speech audio icons for navigation. They were each given a set of 50 information bearing messages that they were asked to store in their

device. The messages included such information as recipes, meeting times and places, phone numbers of individuals, etc. In the evaluation, the subjects were recalled 4 to 6 weeks later and asked detailed questions regarding information within these messages. Given that much of the detailed information was not easily remembered, the subjects were forced to rely on the recorded notes. When timed, the VI subjects were able to retrieve the information over twice as fast with the voice-enabled search as they were with only the recordings. The response was overwhelming in their desire to have such a device to use on a daily basis. In cases such as this, it is apparent that even in a setting with imperfect precision and/or recall, such technology can be very useful. One can easily see how the number of stored messages could essentially be unlimited were such a retrieval tool be available.

3.5 Major Challenges

Unquestionably, there continue to be major challenges in audio retrieval. Although many of the issues are in common with all ASR type approaches, systems based on phonetic search have some clear differences with speech-to-text systems. As such, there are important research directions that are still fertile.

3.5.1 Cross language search

One worthy goal would be to define all utterances for all words in all languages using an IPA-like set of units. This process would allow any phonetic string to be detected without the need for individual models for each language. Further, searching for foreign words or expressions in a given language would be greatly facilitated. Our experience is that as the number of unique languages in one's repertoire expands, the effects of context dependent phones causes the number of models to expand much faster than linear. In addition, not all phonemes as categorized by the IPA are similar enough acoustically to use a shared model. Preprocessing time and phonetic search track size also expand beyond manageable levels. For now, one language at a time, along with automatic language ID as a front-end filter is the practical solution.

3.5.2 Cross acoustical search

One problem that enters detection-based systems relates to accurate estimation of the confidence levels of the detected items. For a given acoustic condition known in advance, scores can be normalized based on such things as noise levels, bandwidth, codec used, and query length and content to give a more accurate estimate of the true probability of correct. When mixed data occurs (a typical example might be a broadcast news segment that includes high-quality anchor speech interspersed with distorted telephone feeds from a field reporter), setting thresholds and normalizing levels to compensate for differing conditions becomes very hard. Promising approaches include automatic classification of acoustic conditions as well as adjusting the thresholds to produce a constant false alarm rate.

3.5.3 Search speed

Even though some embodiments of phonetic search have been shown to operate at up to 2 million times faster than real time, this direct search is still too slow for brute-force searching of large data sets. Four thousand hours of audio is about the limit if one wishes a response time of 7 seconds. Further, detectors operating at one false alarm per hour would still generate four thousand false alarms. Clearly a different strategy would be required. Extensive use of any metadata could effectively filter the set of records to search, speeding up search and reducing false alarms. Another strategy for direct search is to set the operating point at an extremely low false alarm rate. DET (detection/error tradeoff) curves [4] useful for such analysis. Two points from such a curve recently evaluated on a 4,000-hour telephony corpus place the probability of detection for a 12 phoneme query at 35% and for a 20 phoneme query at 85% with a false alarm rate of 0.01 per hour. Though imperfect, this rate is still useful for retrieval and may be the desired operating point.

As exhibited by internet search engines, it is possible to access an almost unlimited amount of text data in rapid fashion based on indexing and pre-searching. Without such indexing, neither text nor audio can be effectively searched. The main difference with audio is that the search for initial indexing is much slower than for text. Nevertheless, preliminary studies suggest that most of the familiar strategies of crawling, caching, reverse indexing, etc translate directly to the audio task. A mixture of indexing followed by direct search for refinement is not unlike the search engine strategy of presenting putative hits and letting the user scan the results for refinement.

3.5.4 Integration of Speech-to-text with Phonetic Search

If a speech-to-text engine is combined with a phonetic search engine naively, one clearly gets the worst of both worlds – slow processing time and the existence of false alarms. Clearly, a strategy should exist for obtaining the best of both worlds. Combining best-of-breed systems in an optimal manner is yet to be done.

4. SUMMARY

Audio information retrieval is a multi-faceted problem which parallels other IR tasks in many ways. It also has some unique properties that require special accommodations and creativity. A host of practical problems can be solved today using a variety of approaches including systems built on an underlying phonetic search strategy.

5. REFERENCES

[1] Reynolds, D.A., and Torres-Carrasquillo, P, "Approaches and applications of audio diarization," in *Proc. of ICASSP 2005*, vol. 5, pp. 953-956.

[2] Cardillo, P. Clements, M., Miller, M., "Phonetic Searching vs Large Vocabulary Continuous Speech Recognition," in *International Journal of Speech Technology* , pp. 9-22, January 2002.

[3] Foote, J.T., Jones, G. J. F., Sparck Jones, K., & Young, S. J., "Talker-Independent keyword spotting for information retrieval," in *Proc. Eurospeech 1995*, vol. 3, pp. 2145–2149.

[4] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech 1997*, vol. 4, pp. 1895-1898. (Also available at www.nist.gov/speech/publications/papersrc/det.doc.)

[5] Zhong, Y. & Gilbert, J.E. "A Context-Aware Language Model for Spoken Query Retrieval," in *International Journal of Speech Technology,* 2005, vol. 8, no. 2, Springer, pp. 203-219.

[6] Kurimo, M, "Fast latent semantic indexing of spoken documents by using self-organizing maps," in *Proceedings of ICASSP 2000*, vol. 6, pp. 2425-2428.

[7] Gauvain, Jean-Luc and Lamel, Lori, "Structuring Broadcast Audio for Information Access," in *EURASIP Journal on Applied Signal Processing Volume,* 2003, issue 2, pp. 140-150.

[8] Wightman, C. and Talkin, D., "The aligner: Text-to- speech alignment using Markov models," in Van Santen, J., Sproat R., Olive J., and Hirschberg J.., eds*., Progress in Speech Synthesis*, pp. 313–323, Springer-Verlag, 1996.

[9] Hansen, J.H.L., Rongqing Huang, Bowen Zhou, M. Seadle, J.R. Deller, Jr., A.R. Gurijala, M. Kurimo, and P. Angkititrakul. SpeechFind, "Advances in spoken document retrieval for a National Gallery of the Spoken Word," in *IEEE TSAP Sep. 2005,* pp. 712-730.