TRIAGE FRAMEWORK FOR RESOURCE CONSERVATION IN A SPEAKER IDENTIFICATION SYSTEM[•]

A. Jairam, E. Singer, D. A. Reynolds

Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, MA 02420
{arvind, es, dar}@ll.mit.edu

ABSTRACT

We present a novel framework for *triaging* (prioritizing and discarding) data to conserve resources for a speaker identification (SID) system. Our work is motivated by applications that require a SID system to process an overwhelming volume of audio data. We design a triage filter whose goal is to conserve recognizer resources while preserving relevant content. We propose triage methods that use signal quality assessment tools, a scaled-down version of the main recognizer itself, and a fusion of these measures. We define a new precision-based measure of effectiveness for our triage framework. Our experimental results with the 35-speaker tactical SID corpus bear out the validity of our approach.

Index Terms— Speaker recognition, speech intelligibility, signal detection, speech processing, acoustic signal processing

1. INTRODUCTION

With the existence of ever expanding archives of unlabeled audio data, the need for techniques to search and sort these archives based on audio content is gaining importance. One such technique is speaker indexing, where a large audio archive is processed with an automatic speaker identification (SID) system to retrieve data likely spoken by a set of desired speakers and a sorted list of data likely spoken by these speakers is presented the user for further processing, analogous to a sorted list of results from a search engine query. We assume that the user will not evaluate the entire output set from the recognizer, but rather a small fraction (e.g., the highest scoring 1%). As the archive volume increases, the straight-forward approach of simply scoring all data with the SID system will not scale well for a fixed compute allocation. Thus, a method is sought to discard a large fraction of the input data prior to SID processing while maximizing the precision of the retained data, so as to increase the usefulness of the system from the user's perspective.

We address this problem by introducing a triage framework wherein the archive data is first processed to winnow out segments that are likely not to be scored well by the SID system (and so not appear in the top ranked data viewed by the user) saving SID computations for more "relevant" data. We consider a naïve triage method as a baseline and introduce more sophisticated approaches that yield better performance. To evaluate performance of the triage system relative to the user's experience, we define a new measure of effectiveness based on precision to compare precision of various smart triage approaches to the naïve triage baseline.

2. TRIAGING METHODS

The most straightforward method for reducing the computation burden of a recognizer is to simply decimate the input data (or equivalently, to discard a fixed proportion of data at random). This method, which we call naïve or random triage, will not, on the average, change the proportion of data that will be correctly classified. Random triage serves as a useful baseline when evaluating other triage methods, since any type of "smart" triage should refine the retained data, i.e., increase the proportion of data that will be correctly classified, in addition to reducing the computational load by discarding data.

One approach for smart triage involves making signal quality measurements on the input data. It is well known that the performance of SID systems tends to degrade in the presence of distortion; therefore, eliminating inputs in which these distortions are detected should simultaneously reduce the computation load on the recognizer and increase the volume of correct material presented to a user. For this purpose, we seek to identify signal features that correlate well with SID performance. We considered using various candidate features, including signal-to-noise ratio (SNR), the presence of silences or tones, pitch, and other spectral and cepstral statistics. A survey of available methods determined that two existing tools would suffice to conduct a proof-of-concept study of the triage framework.

The first tool chosen was an SNR estimator adapted from the version available from NIST [1]. We reasoned that an SNR estimate would require a simple, fast calculation and would correlate with SID performance since speech samples that are degraded by noise would likely be prone to more recognition errors. The second was an objective speech quality assessment tool available as ITU specification P.563 [2]. As shown in Figure 1, this tool calculates a range of characteristic speech parameters and computes a score between 1 and 5 (5=best) that is indicative of subjective speech quality. The P.563 tool is designed to assess speech quality without a reference signal by identifying 46 signal-related parameters, including basic speech descriptors such as pitch and speech level, indicators of unnatural speech based on vocal tract parameters, noise analysis parameters, and parameters related to interruptions or mutes. P.563 was designed to correlate well with human perception, and has a measured correlation coefficient of p=0.9.

[•] This work is sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

We used the signal quality tools out of the box, with the understanding that implementation in a real system would require optimizing the tools to run much faster than any SID system, which we posit could reasonably be done. The scope of the present work is to demonstrate the feasibility of a triage framework rather than to optimize the measurement tools.

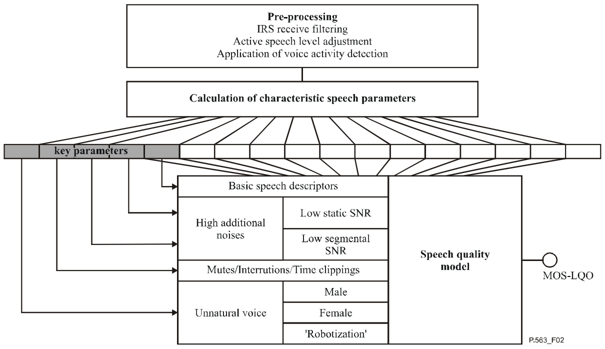

Figure 1: *Details of ITU P.563 method for objective speech quality assessment.*

A second approach to triage is to design a fast, scaled-down version of the full SID recognizer and use its output score to pass data selectively to the full recognizer. The rationale for this method is that a scaled-down recognition system, while not maintaining the accuracy of a full scale system, would still provide a potential means of rejecting likely misclassified inputs.

Finally, we consider a fusion triage system which uses the SNR, P.563 and fast recognizer scores as elements of a feature vector. We construct decision regions in the three-dimensional space and discard files accordingly. For the current study, only decision regions corresponding to hypercubes derived from the single feature regions have been considered for the sake of simplicity, but arbitrary decision regions may be employed.

3. TRIAGE ANALYSIS

Performance of speaker recognition systems is traditionally reported using (prior independent) detection error tradeoff (DET) curves that describe the relationship between missed detections and false alarms [3]. Although useful as a measure of core performance, DET curves are not well suited to evaluating the utility of triage for an end user. Rather, we seek to evaluate the impact of combining audio triage and speaker recognition on the quality of the information presented to the user. For this purpose, we choose to use precision as our metric and we propose a measure of effectiveness to evaluate the impact of audio triage on a speaker recognition system.

3.1. Precision Metric

Precision measures the relevance of information in a sorted queue (the output queue provided to the user for further evaluation, in our case). At any given point in the queue, the precision is the cumulative proportion of correct recognitions, i.e., matches between the hypothesized speaker and the true speaker. We used the precision formulation described in [4] where, for a queue of length N sorted in descending order and for a given richness (*a priori* target probability) r, the precision at any level of the sorted queue is given by

$$\text{Precision}(i) = \frac{P_d(i) * r}{P_d(i) * r + P_f(i) * (1-r)}$$

where $P_d(i)$ and $P_f(i)$ are the cumulative proportions of detections and false alarms, respectively, and i=1,2,…,N.

3.2. Triage Measure of Effectiveness

Since random triage can always be used to achieve a desired reduction in computational load, it serves as a useful baseline for comparison with more sophisticated triage approaches. Rather than presenting separate curves comparing precision for smart and random triage at all levels of a queue, we instead propose a summary statistic to characterize triage measure of effectiveness (MOE): the precision gain of feature-based triage relative to random triage (discarding the same number of files) at 1% of the sorted queue. Thus a precision gain of 0 would indicate that the triage scheme under consideration is no better than random triage, while a positive MOE indicates an improvement.

An example of the proposed triage MOE is illustrated in Figure 2. The plots show precision vs. proportion of queue for random (lower curve) and smart (upper curve) triage. Details regarding the corpus and experimental setup are presented in Section 4. For this example, the precision at 1% of the queue increases from 0.42 with random triage to 0.71 with feature-based triage, resulting in an MOE of 69%.

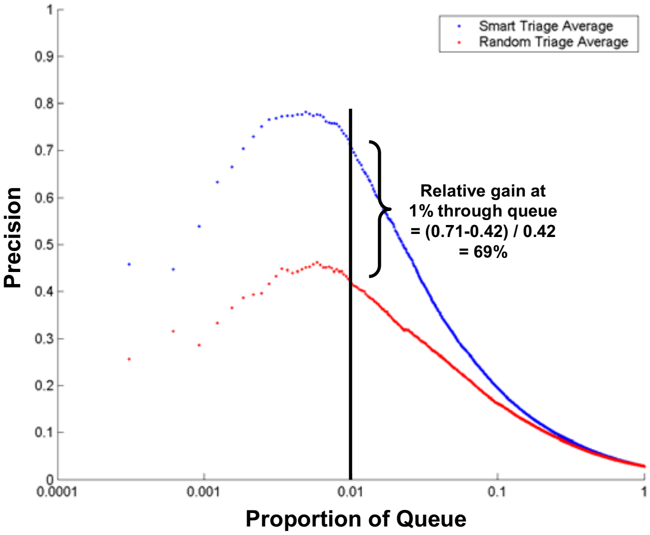

Figure 2: *Illustration of use of proposed triage measure of effectiveness (MOE). Red line (lower plot): Precision vs. queue using random triage. Blue line (upper plot): Precision vs. queue using methods detailed in Section 4.*

4. EXPERIMENTAL FRAMEWORK

In our experimental setup, we passed test input data in parallel through two distinct pathways: 1) through a SID recognizer trained on data separate from the test data, and 2) through a triage filter. The triage filter used 1, 2, or 3 triage features (SNR, P.563, and fast recognizer score), as well as a threshold for those features, to generate a mask. We then applied this mask to the SID output scores to emulate the process of discarding data via triage, measured the precision of the masked output queue, and computed the triage measure of effectiveness.

4.1. Speaker Identification System

The SID system used for this study is an adapted Gaussian Mixture Model recognizer developed at Lincoln Laboratory (see [5] for details). Target models (one for each speaker) are adapted from a 64-mixture universal background model trained using an aggregation of data from all speakers. For each input file, a log likelihood ratio score is computed for each speaker model and the highest score and speaker label are reported as the SID output. As the main focus of this work was on triage approaches, emphasis was not placed on optimization of the SID system for this data set.

4.2. Corpus

We selected the Tactical Speaker Identification (TSID) speech corpus for our study [6]. The TSID corpus contains 26 distinct English sentences spoken by each of 35 speakers and using up to 4 military transmitters and 7 military receivers. One of the receivers was a handset-mounted wideband microphone that provided the reference signal. (The corpus actually contains additional tasks from each speaker as well, including utterances of digits and giving impromptu directions based on a map, but we did not use the former because the utterances are too short, and we did not use the latter because of the non-uniformities in the utterances.) The receivers used for the data collection were dispersed geographically along a two-mile stretch of terrain, and the transmitters were collocated, as shown in Figure 3. Data was not available from all combinations of speaker, receiver, and transmitter. The resulting 17,370 sentences used for this study have a mean duration of 4 seconds, which is generally shorter than ideal for effective speaker identification. Of these files, 1630 (all the reference wideband files from transmitters 1 and 2) were used to train the target speaker models for the GMM recognizer.

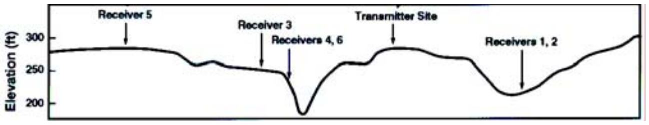

Figure 3: Relative positioning of transmitters and receivers for TSID data collection.

What renders the TSID corpus especially attractive for the present study is the presence of real distortions for some receiver/transmitter pairs. Although TSID has a relatively small number of speakers, we believe that using these real distortions rather than artificially adding distortions to existing larger corpora was more relevant for examining our triage techniques. The distortions are a result of the topography and intrinsic properties of the transmitters and receivers, and they manifest themselves as a variety of audible phenomena, including high and low frequency tones, clipping and robotization, and additive noise. We expected that some of these distortions would degrade SID performance, making this an ideal corpus for a triage study. In addition, the presence of data samples from a reference wideband source allowed us to label these utterances as "good." We expected that these utterances would yield the best recognizer performance.

4.3. Decision Regions

We analyzed the distribution of the TSID files in the SNR-P.563 space. The 17,370 TSID files were manually classified based on a subjective listening assessment. For each combination of speaker/receiver/transmitter, we listened to 1 sentence (out of 26 total) and assigned all 26 sentences from that triple to one of five categories based on subjective perception: 1) reference wideband; 2) noisy or unintelligible; 3) robotization or clipping present; 4) low or high frequency background tone present; and 5) other distortion present. Of these, only membership to Category 1 is objective, since we know which data was recorded in the reference wideband condition.

The P.563 vs. SNR scores were then plotted for the sentences, as shown in Figure 4. We observe that "good" files, i.e., files in Category 1, tend to have high SNR and high P.563 scores. (We note that SNR estimates greater than 50 dB are probably erroneous, as those speech samples do not sound particularly clear.) Also, "bad" files, in categories 2, 3 or 4, tend to have low SNR or low P.563 scores. Consequently, with the expectation that files in Category 1 would yield good SID performance and files in categories 2, 3 and 4 would yield poorer performance, we segmented the (SNR, P.563) feature space into two classes, as shown in Figure 4. Our triage decision rule is then: discard utterances with low SNR (≤25 dB) or low P.563 score (≤3). We remark that since these decision thresholds (SNR=25dB and P.563=3) were chosen without any particular optimization, one would expect other multi-feature decision regions to produce better performance.

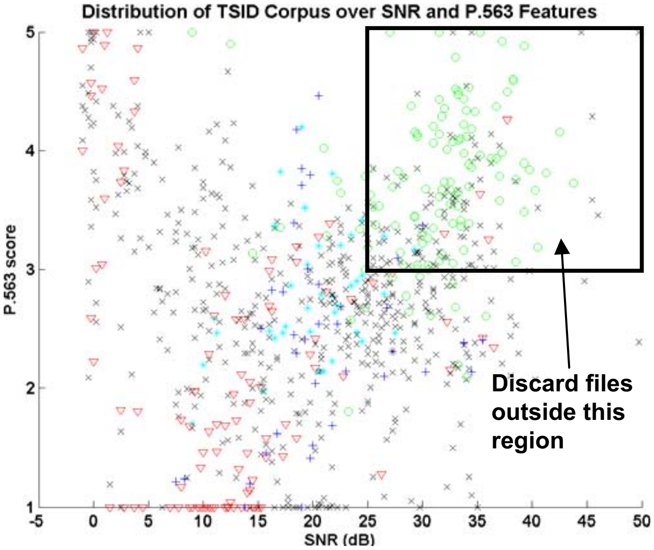

Figure 4: SNR and P.563 scores for TSID files, according to subjective perceptual categories: Green circle=1, Red triangle=2, Blue plus sign=3, Cyan star=4, Black X=5. See text.

Similarly, a decision region using the fast recognizer score was obtained using the data plotted in Figure 5, which shows per-category histograms of the fast recognizer scores. For this study, the fast recognizer was a GMM SID system that used 2 mixtures per speaker model and 1-out-of-10 frame decimation. We observe that there is considerable overlap between scores for Category 1 ("good") files and those of other categories. For this study, we chose a score of 0 as a decision threshold for this feature. During development testing, we swept out fast recognizer feature scores with 4, 8, 16, and 32 mixtures and thresholds varying from -2 to 2, but found that using 2 mixtures and a threshold of 0 yielded nearly the greatest precision gain, so we show results only for that combination.

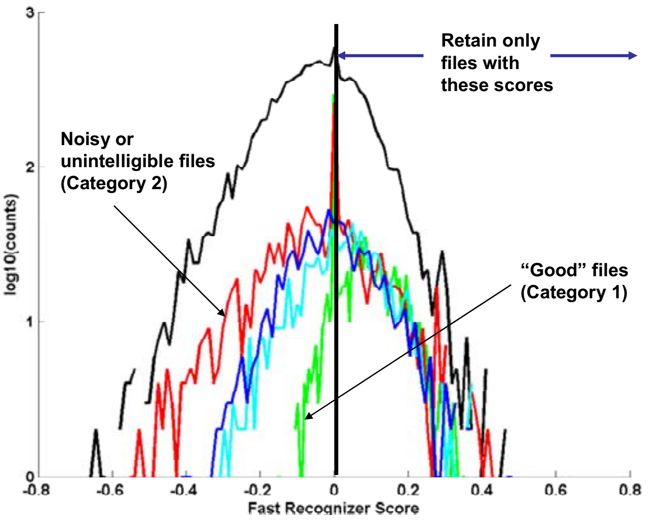

Figure 5: *Fast recognizer scores for TSID corpus (colors indicate subjective perceptual categories: Green=1, Red=2, Blue=3, Cyan=4, Black=5; see text).*

4.4. Results

Table 1 shows the results of using the various triage methods discussed in this paper. For triage using a single feature, at an approximately constant discard rate of 60%, SNR is the most useful feature (MOE=53%), followed by P.563 and then the fast recognizer. Thus, it is possible to reduce the computation load on the recognizer by 60% while increasing the relative precision (with respect to random triage) of the top 1% of the sorted queue by 53%. At an approximately constant discard rate of 82%, the most effective pair of triage features is (SNR, P.563), followed by (SNR, fast recognizer) and (P.563, fast recognizer). Furthermore, use of all three features results in 90% of files discarded and a 117% gain in precision relative to random triage at the 1% level of the queue.

Discard:					**% discarded**	**MOE, in %**
SNR ≤ 25 dB	**P.563 ≤ 3**	**Fast Recognizer Score ≤ 0**	**SNR ≤ 32 dB**	**SNR ≤ 35 dB**		
X					61	53
	X				60	26
		X			59	19
X	X				82	89
	X	X			83	57
X		X			82	74
			X		82	64
X	X	X			90	117
				X	90	47

Table 1: *Triage performance summary showing percent of files discarded and measure of effectiveness for all combinations of triage features.*

To compare single-feature triage with multiple-feature triage, we selected the single feature with the best performance (SNR) and determined the thresholds necessary to achieve the discard rates shown in the two-feature and three-feature cases. Thresholds of 32dB and 35dB yielded discard rates of 82% and 90%, respectively, with MOEs of 64% and 47%, respectively. We observe that at a discard rate of 82%, SNR triage alone outperforms (P.563, fast recognizer) triage, but does not perform as well as the other two-feature combinations. Hence, fusing multiple features may be better than using single features. This is only an initial attempt at fusing features, as we have only used multidimensional decision regions whose boundaries are derived from intersections of single feature regions.

5. CONCLUSIONS AND FUTURE WORK

We presented a framework for triage of data motivated by an audio archive search scenario in which large volumes of data were processed by a SID system and the sorted results presented to a user. To reduce the computational burden on the recognizer and boost its effectiveness to the user, a set of triage methods and evaluation metrics were proposed. The TSID corpus was selected for this study because it contained multiple speakers and many instances of real degraded speech. Three feature estimation tools were considered: SNR, P.563, and a fast GMM SID. Triage using the SNR tool led to the greatest improvement in relative precision, followed by P.563 and the fast GMM SID. Using multiple features for triage produced additional improvement despite the fact that the fusion methods used in this study were simplistic.

We note that no effort was made to either optimize the code used to compute the triage features or to determine the computational complexity of the tools employed in this study. Rather, we have simply asserted that these measurements could be made much more efficiently than running a SID system. The additional work required to verify these engineering judgments was beyond the scope of this study. In addition, the selection of feature thresholds was performed in a fairly simplistic manner and more sophisticated methods, such as using supervised learning, may lead to additional gains in performance. Further investigation of triage feature selection, computational tradeoffs, and supervised learning are prime candidates for future work in triage.

6. REFERENCES

[1] "The NIST Speech SNR Measurement," http://www.nist.gov/smartspace/snr.html

[2] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," http://www.itu.int/rec/T-REC-P.563/en

[3] A. Martin, G. Doddington, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proceedings of EuroSpeech 1997*, volume 4, pp. 1895-1898, 1997.

[4] D. E. Sturim, D.A. Reynolds, E. Singer, and J.P. Campbell, "Speaker Indexing in Large Audio Databases Using Anchor Models," *Proceedings of Acoustics, Speech, and Signal Processing*, Salt Lake City, pp. 429-432, May 2001.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[6] D. Graff, D. A. Reynolds, and G. C. O'Leary, *Tactical Speaker Identification Speech Corpus (TSID)*, Linguistic Data Consortium, http://www.ldc.upenn.edu