SPEAKER INDEXING IN LARGE AUDIO DATABASES USING ANCHOR MODELS

D. E. Sturim[1] *D. A. Reynolds*[2]*, E. Singer*[1] *and J. P. Campbell*[3]

[1]MIT Lincoln Laboratory, Lexington, MA
[2]Nuance Communications, Menlo Park, CA
[3]Department of Defense
{sturim,dar,es}@sst.ll.mit.edu, j.campbell@ieee.org

ABSTRACT

This paper introduces the technique of anchor modeling in the applications of speaker detection and speaker indexing. The anchor modeling algorithm is refined by pruning the number of models needed. The system is applied to the speaker detection problem where its performance is shown to fall short of the state-of-the-art Gaussian Mixture Model with Universal Background Model (GMM-UBM) system. However, it is further shown that its computational efficiency lends itself to speaker indexing for searching large audio databases for desired speakers. Here, excessive computation may prohibit the use of the GMM-UBM recognition system. Finally, the paper presents a method for cascading anchor model and GMM-UBM detectors for speaker indexing. This approach benefits from the efficiency of anchor modeling and high accuracy of GMM-UBM recognition.

1. INTRODUCTION

This paper describes a method of representing and characterizing a target utterance with information gained from a set of anchor models derived from a predetermined set of speakers. Since the speakers of the target utterances are not members of the model training set, the system is capable of characterizing the target speaker with no prior knowledge of that speaker. Previous research [1, 2] suggests that the target speaker will be projected into a talker space defined by the anchor models. Since the models are created only once in the training phase, it is unnecessary to train a model for a new target speaker. Applications of the approach include speaker recognition, speaker detection, and speaker clustering for very large speaker populations where it is undesirable or infeasible to train models for every member of the target population. Another application of anchor modeling discussed in this paper is speaker indexing; that is, the use of speaker detection for the retrospective searching of large speech archives. For large archives, current state-of-the-art speaker recognition systems may be too computationally inefficient for large searches. The efficiency of the anchor system lends itself to the application of large speech archive retrieval. It is shown that although the detection performance of the anchor model system falls short of state-of-the-art Gaussian Mixture Model with Universal Background Model (GMM-UBM) speaker detection systems [3, 4], the efficiency of anchor modeling can be effectively exploited by embedding it in a two-stage cascaded system, where the role of the anchor system is to reduce the data load of the more accurate but less computationally efficient GMM-UBM.

This work was sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and not necessarily endorsed by the United States Air Force.

2. ANCHOR MODELS

The basic concept of anchor modeling is the representation of a target speech utterance with information gained from a set of models pre-trained from a defined set of talkers . In theory, the models could consist of virtually any method of speech representation. Previous work [1, 2] used speaker-dependent Hidden Markov Models (HMM) as the anchors. This study uses the GMM-UBM as the representation model for forming the anchors.

Segments of speech, s, are scored against a set of pre-trained anchor models, A_i, $i = 1, ..., N$. Each of the N anchor models yields a likelihood score and the collection of scores is used to form the N-dimensional *characterization vector*. The speech utterance is represented by this characterization vector V, where

$$V = \begin{bmatrix} p(s|A_1) \\ p(s|A_2) \\ \vdots \\ p(s|A_N) \end{bmatrix} \quad (1)$$

The characterization vector can be considered a projection of the target utterance into a speaker space defined by the anchor models. If an utterance from a single speaker projects into a unique portion of the speaker space, then the speaker representation is unique. Speaker detection is performed by considering the location of the vectors within this speaker space.

Speech segments are compared by scoring a speech segment s_u from an unknown speaker and a speech segment s_t from a target speaker against the same set of anchor models (Figure 1), thereby forming two characterization vectors, V_u and V_t, to represent the unknown and target segments of speech. A vector distance is then used to compare the speech segments.

Preliminary experiments using Euclidean, absolute value or "city block", and Kullback - Leibler distance measures showed that Euclidean distance performed best. Unit nor-

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2001	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Speaker Indexing In Large Audio Databases Using Anchor Models	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **MIT Lincoln Laboratory, Lexington, MA, USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**4**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

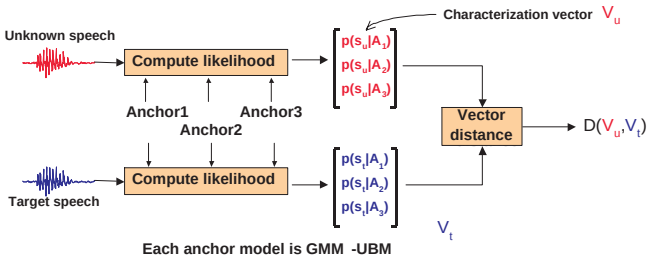

Figure 1: The anchor model system.

malizing the elements of characterization vectors in the distance calculation did not change performance.

The GMM-UBM anchor models described in this paper were trained using speech from 668 talkers in the NIST-1996 and NIST-1999 speech corpora.[1] The GMM-UBM algorithm used was the same as that developed for the NIST-2000 speaker recognition workshop [5, 6] but without speaker (T-NORM) and handset (H-NORM) normalizations.

2.1. Anchor Model Pruning

The full anchor model characterization vector is formed by scoring an utterance against all 668 anchor models. Methods of reducing the size of the Euclidean distance comparison were investigated in an effort to increase performance by using only those anchor models that provide good characterizing information. Reducing the size of the distance comparison reduces the dimensionality of the speaker space and increases computational efficiency.

Model pruning strategies were motivated by the observation that the vector distance between characterization vectors derived from the same talker should be small while distances between characterization vectors of different speakers should be large. Characterization vectors of two utterances from the same talker were compared and the resulting element distances, d_i, were rank ordered by magnitude, where

$$d_i = \left[(V_{t_i} - V_{u_i})^2 \right]_{i=1:N} \quad (2)$$

and V_t and V_u are two characterization vectors obtained from two target speech utterances. A percentage of the models with the lowest element distances was then chosen as the anchor model set. In a similar manner, characterization vectors of utterances from different talkers can be evaluated with Equation 2, where V_t and V_u are now characterization vectors from different talkers. With this approach, only those models with the largest element distances are chosen for the anchor model set. Using these two methods of pruning, the size of the Euclidean distance comparison was reduced by 60% while the equal error rate was improved.

3. SPEAKER DETECTION WITH ANCHOR MODELS

Results presented in this section used speech data from the NIST-2000 Speaker Recognition Workshop, sectioning the corpus into test and training sets and performing the evaluation using the protocols stipulated in [7]. Figure 2 presents the Detection Error Tradeoff (DET) curves for the NIST-2000 single speaker detection task primary condition.[2] The equal error rate for the anchor model system using the full characterization vector ($N = 668$) was 24.2% while the equal error rate of the anchor system with model pruning was 21.4%. Pruning of the models provides a relative performance increase of 11.7%.

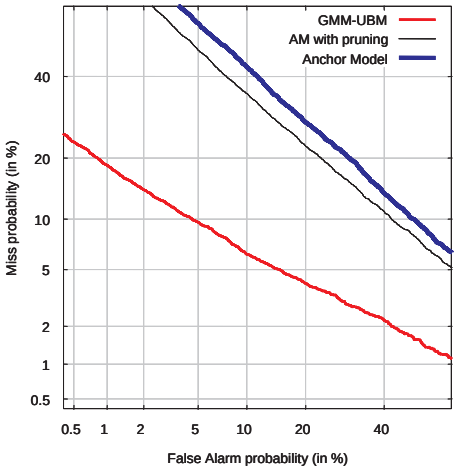

Figure 2: DET curves for the GMM-UBM and anchor model system using the primary condition of single speaker detection NIST-2000 speech corpus.

The performance of the anchor system falls well short of the 7.7% equal error rate of the GMM-UBM system. The next section discusses one application of speaker detection where the computational efficiency of the anchor modeling approach is used to advantage.

4. SPEAKER INDEXING

Speaker indexing is defined as the application of speaker detection to the retrospective search of large speech archives. Two possible uses of speaker indexing are the clustering of speech messages contained in a speech archive and the retrieval of a list of messages from an archive in response to an external query. This paper focuses on the list retrieval task. Performance in speaker detection evaluations has traditionally been reported using a (prior-independent) DET curve that describes the underlying tradeoff between misses and false alarms for a given detector and corpus. However, performance in information retrieval applications such as speaker indexing is better described using the notions of precision and recall. Detection theory and information retrieval measures are related as follows: Recall is the proportion of relevant material retrieved from the archive and so is equal to the detection probability. Precision is the proportion of retrieved material that is relevant and is given by

$$Precision = \frac{P_t(1-P_m)}{P_t(1-P_m) + (1-P_t)P_{fa}} \quad (3)$$

[1] The data used in the NIST evaluation is a subset of the Switchboard I-II data corpora.

[2] The NIST "primary condition" uses 2 minute training segments and 15-45 second test segments collected with an electret microphone.

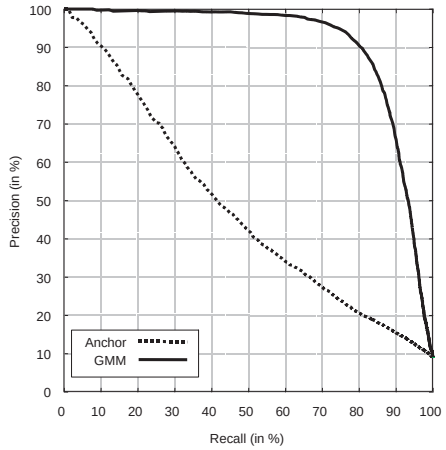

Figure 3: Precision versus recall plot for the GMM-UBM, and anchor model, with $P_t = 9\%$.

where P_t is the target probability (richness) of the archive, P_m is the probability of a miss, and P_{fa} is the probability of false alarm. These relationships can then be used to derive speaker indexing performance (in terms of precision vs. recall) from a DET plot for any given target probability P_t.

4.1. Evaluation of the GMM and Anchor models for Speaker Indexing

Figure 3 shows the precision versus recall tradeoff for the GMM-UBM and anchor model speaker detectors using the DET plots of Figure 2 (NIST-2000 speech corpus) and an archive richness $P_t = 9\%$ (the richness of the NIST-2000 corpus). As expected, the GMM-UBM method outperforms the anchor model. It is worth noting that the curves tend to move toward the upper right with increasing P_t and toward the lower left with decreasing P_t.

Another measure of a speaker detector's value for speaker indexing applications is its computational efficiency. Here it is assumed that each item in the archive is represented by a model (trained off-line) against which a query is scored. For the GMM-UBM, each 10ms frame of the query is first scored against the 2048-component universal background model and then against 5 components of each of the archive models [5]. For anchor model based speaker indexing, the query is first converted to a characterization vector by scoring it against the 668 anchor GMMs. The resulting characterization vector is then compared to each archive characterization vector (trained off-line) using a 668-element Euclidean distance. Figure 4 plots the number of 38-dimensional Gaussian computations (or equivalent) required for a 1 minute query. (It is assumed that the computation time for one 38-element Gaussian and 38 Euclidean distances are equal.) The plot for the anchor model system stays flat to about 10^6 because the computation is dominated by the conversion of the query to a characterization vector. Note that this is true for the pruned anchor system as well. It is apparent that the anchor model speaker indexing system has significant computational advantages for archives containing more than about 1000 items. It should be noted that methods exist for speeding up the computation required for the GMM that

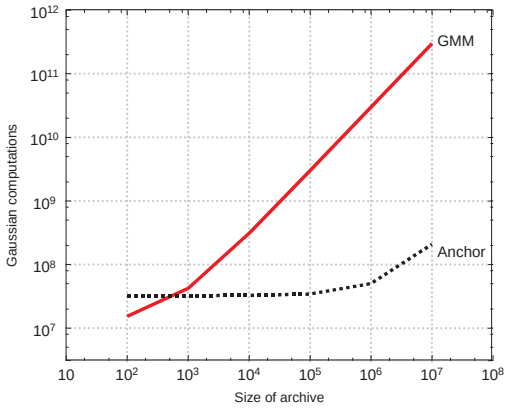

Figure 4: Plot of computational efficiency for the GMM-UBM and anchor model speaker detectors.

Figure 5: Cascaded speaker detection system.

would improve the efficiency of both the GMM-UBM and anchor model systems.

4.2. Cascading

Figures 3 and 4 show the tradeoff of computational efficiency versus accuracy for speaker indexing. The GMM-UBM has superior detection performance while the anchor system provides the computational efficiency that is essential when searching large archives. In an effort to gain a better tradeoff between computational performance and accuracy, the anchor and GMM-UBM speaker detection systems were combined in a cascade as shown in Figure 5. The objective of cascading is to construct a system containing the positive aspects of both algorithms. The anchor model is employed in the first stage to reduce the amount of computational loading for the GMM-UBM speaker detection system. The GMM-UBM is then used to provide maximum recognition performance.

To evaluate the performance of the cascade, it is first necessary to identify the operating point of the anchor system. Define q to be the fraction of the archive processed by the second system of the cascade (i.e., the probability that the first system declares a target). Note that q is the denominator of Equation 3:

$$q = P_t(1 - P_m) + (1 - P_t)P_{fa} \quad (4)$$

where $(1 - P_m)$ is the probability of detection and P_{fa} is the probability of false alarm for the anchor model speaker detector. Given that the richness of the archive (P_t) is defined by the application, choosing a unique value for q identifies a (P_{fa}, P_m) pair from the DET curve (Figure 2) and represents the chosen operating point for the anchor system.

The precision versus recall curve for the cascaded system can be calculated in the same manner as in Section 4.1. Figure 6 presents precision versus recall for the cascaded

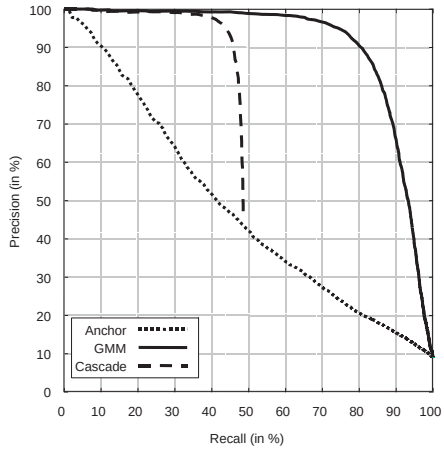

Figure 6: Precision versus recall plot for the GMM-UBM, anchor model, and cascaded system with $q = 10\%$ and $P_t = 9\%$.

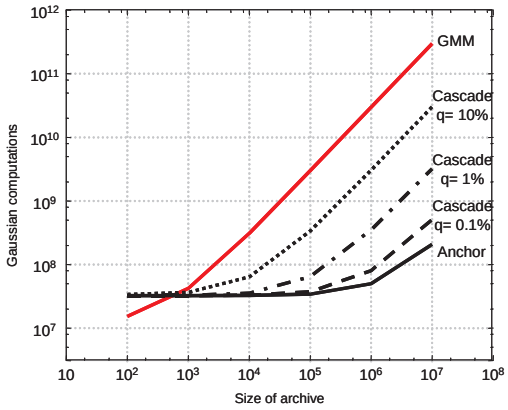

Figure 7: Estimated number of Gaussian (or equivalent) computations, 1 minute query.

system with $q = 10\%$ and an archive richness of $P_t = 9\%$. The effect of the cascade is to slightly reduce the performance in operating regions of low recall and to drastically reduce performance in regions of mid-to-high recall, relative to the GMM system.

Figure 7 displays a plot of the estimated computational efficiency for the GMM-UBM, anchor model, and cascaded speaker indexing systems. As the amount of reduction in archive size increases (smaller q), the computational efficiency of the cascaded system also increases.

5. SUMMARY

This paper presented a method of characterizing a segment of a talker's speech with information gained from a set of pre-trained anchor models. The anchor models were derived from a set of predetermined speakers. Characterization vectors were then formed by scoring the target speech segment against the set of anchor models. A method for refining the anchor modeling system was presented increased recognition performance.

Anchor modeling was then applied to the speaker detection problem. Detection error tradeoff performance showed that the anchor modeling system fell short of a state-of-the-art GMM-UBM system. It was further shown that its computational efficiency was superior to that of the GMM-UBM. Comparison of the anchor model and GMM-UBM systems for speaker indexing showed a similar tradeoff between precision versus recall performance and computational efficiency.

A cascaded speaker indexing system was proposed that utilized the anchor model system as the first stage and the GMM-UBM as the second stage. In this configuration, the anchor model reduced the data loading on the GMM-UBM while slightly reducing performance in operating regions of low recall. The effect of the cascaded system was to combine the advantages of both systems at the expense of some loss in both computational performance and detection accuracy. For large archives, the recognition performance of the anchor system and the lack of computational efficiency of the GMM-UBM system could preclude their application to speaker indexing. The cascaded system may offer a viable solution to the speaker indexing application.

6. REFERENCES

[1] Douglas E. Sturim, *Tracking and Characterization of Talkers Using a Speech Processing System with a Microphone Array as Input*, Ph.D. thesis, Brown University, 1999.

[2] Teva Merlin, Jean-François Bonastre, and Corinne Fredouille, "Non directly acoustic process for costless speaker recognition and indexation," *International Workshop on Intelligent Communication Technologies and Applications*, 1999.

[3] Douglas Reynolds, Thomas Quatieri, and Robert Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[4] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[5] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997.

[6] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[7] NIST, *The 2000 NIST Speaker Recognition Evaluation Plan*, Linthicum, MD, June 2000, http://www.nist.gov/speech/tests.