Unsupervised Indexing of Conversations with Short Speaker Utterances

Uchechukwu O. Ofoegbu¹, Ananth N. Iyer¹, Robert E. Yantorno¹ and Stanley J. Wenndt²
¹Speech Processing Laboratory, Temple University, Philadelphia, PA 19122-6077, USA
{uche1, aniyer, byantorn}@temple.edu
²Air Force Research Laboratory/IFEC, Rome, NY 13441-4514, USA
Stanley.Wenndt@rl.af.mil

Abstract—Two speaker indexing system for conversations are presented in this paper. The first method involves indexing two-speaker conversations. In this method, two reference models are judiciously chosen from the conversation such that they represent the two different speakers. Models are then matched to the reference speakers using distance-based comparisons. The second technique is based on first determining the number of participants in the conversation using a speaker count method termed the “Residual Ratio Algorithm” (RRA), and then indexing based on this count. The RRA involves an elimination process in which speech segments matching a chosen set of reference models are sequentially removed from the conversation and the relative amount of residual speech is observed to determine the count. The distance measures used in comparing models include the Bhattacharyya distance, the T-Square statistics and the Mahalanobis distance. Speaker comparison decisions of all three distances are combined to improve the accuracy of the system. Linear Predictive Cepstral Coefficients of voiced phonemes are used in forming speaker models. The two-speaker indexing technique was able to yield an indexing accuracy of up to 95% when evaluated using SWITCHBOARD data. The counting-indexing technique resulted in a maximum indexing accuracy of about 91% when tested on artificial conversations generated from HTIMIT data.

TABLE OF CONTENTS
1. INTRODUCTION.................................................................1
2. DISTANCE MEASURES .............................................. 2
3. SPEAKER MODEL FORMATION .....................................3
4. SPEAKER INDEXING....................................................5
5. SPEAKER COUNTING-INDEXING ..............................7
6. CONCLUSION .................................................................10
REFERENCES ..........................................................10
BIOGRAPHY ..............................................................10

1. INTRODUCTION

The art of examining multi-speaker speech data in order to determine who is speaking and when they are speaking is commonly referred to as speaker-indexing. This relatively recent aspect of speaker-recognition can be classified into two categories: supervised and unsupervised speaker indexing. The former category involves having some information about the speakers participating in the observed speech data; whereas, in the latter, this information is absent. Consequently, the problem of unsupervised indexing is more difficult than supervised speaker indexing, and has been addressed using techniques such as Neural Networks [1], Generic Speaker Modeling [2], Bayesian Information Criterion [3] and Support Vector Machines [4]. These methods, however, are complex in that they require either the detection of change points or the use of external (representative models) to compensate for the lack of speaker information. In this paper, two speaker indexing techniques designed for conversations are presented.

The first speaker indexing technique assumes that the number of speakers in the conversation is known and equal to two. Since no speaker information is available, speaker homogenous models are formed from the given conversation, and these models are assigned to different speakers using an approach termed the “Restrained-Relative Minimum Distance” (RRMD) method. This technique involves 1) computing the distance between speaker homogeneous models, formed using Linear Predictive Cepstral Coefficients (LPCCs) extracted from voiced portions of the speech data, and then 2) selecting the two models which are farthest apart as reference models to represent the two different speakers. Each of the other models is then assigned to one of the reference speakers contingent upon its passing an intra-speaker likelihood ratio test or an inter-intra speaker relative distance test. Models which fail to meet the above mentioned criteria are set aside and considered undecided/unusable for speaker indexing. This is illustrated in the block diagram of the proposed system given in Figure 1.

In telephone conversations, utterances are short because speakers change relatively frequently [5]; it is therefore
important that caution be taken when forming models. In this research, analysis is performed to determine the appropriate data size suitable for ensuring that the models formed contain speech information from only one speaker. Also, since telephone conversations are known to contain several portions of co-channel speech (two speakers speaking at the same time), models classified as unusable are assumed to fall in such category. During evaluation, however, indexing errors are classified into two forms: 1) percentage of speech that was assigned to the wrong speaker and 2) percentage of non-co-channel speech that was declared unusable by the system.

The second technique assumes that the number of speakers is unknown but can be a maximum of four. A speaker count algorithm is first implemented, and then speaker indexing is performed based on the number of speakers determined. The method involves an elimination process in which speech segments matching a chosen set of reference models are sequentially removed from the conversation. The speaker count is then obtained by observing the ratio of the length of the residual speech to the original length of the conversation. Indexing of the residual speech is performed using the minimum distance criterion.

In this research, all speaker models are created using the mean vectors and covariance matrices of 14th order Linear Predictive Cepstral Coefficients (LPCCs) of voiced segments. The distance measures used in comparing test models to reference models include the Bhattacharyya distance, the T-Square statistics and the Mahalanobis distance. Model comparison decisions are also made by a vote-based comparison of the decisions of all three distance measures.

This paper is organized as follows: an introduction to the distance measures used is given in Section 2, followed by a description of the proposed model formation technique in Section 3. In Section 4, the first speaker indexing techniques which assumes the presence of only two speakers is presented along with results obtained on actual telephone conversations. Section 5 deals with the speaker count based indexing system and its evaluation. Finally, a summary of the research is presented in Section 6.

2. DISTANCE MEASURES

The distance measures involved in this research include the Bhattacharyya distance, Hotelling’s T-Square statistics, and the Mahanalobis distance. All distances were chosen due to the fact that the feature sets used in representing speakers are multivariate random variables; therefore distance measures which take all variables into consideration are preferred since they generally utilize information not only from the mean values, but also from the covariance matrices of the feature vectors. The above mentioned distance measures have all been successfully applied in speaker recognition systems for applications such as SID [6], [7], [8], Speaker change-point detection and indexing [3], [9], and speaker count [10].

The following notations will be used. The random variables:

\[ X = [X_1, X_2, \ldots, X_p] \]  
\[ Y = [Y_1, Y_2, \ldots, Y_p] \]

denote the two multivariate random variables - of lengths \( n_x \) and \( n_y \), and number of features equal to \( p \) - to be compared. For all distances to be valid, the feature vectors are required to have pdfs \( f_r(X) \) and \( f_r(Y) \) which follow the multivariate Gaussian distribution:

\[
f_r(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}
\]

In order to be considered a valid distance metric, each distance measure, \( Q(X; Y) \), between two random variables, \( X \) and \( Y \), their pdfs or their parameters \{\( \mu_x, \Sigma_x \)\} and \{\( \mu_y, \Sigma_y \)\} satisfies the following properties:

\[
Q(X, Y) \geq 0,
\]

\[
Q(X, Y) = 0 \text{ iff } X = Y,
\]

\[
Q(X, Y) = Q(Y, X),
\]

\[
Q(X, Y) \leq Q(X, Z) + Q(Z,Y)
\]

**Bhattacharyya Distance**

The Bhattacharyya distance, belongs to a class of distance measures which compute the separation of two pdfs based on the dispersion of the likelihood ratio with respect to one of the densities [11]. The distances in this class are derived from the equation:

\[
Q(X; Y) = g \left( E_x \left\{ f \left( \frac{p_y(y)}{p_x(x)} \right) \right\} \right)
\]

where \( g \) is a function which is continually increasing on the Real Line, \( E_x \) is the expectation of the random variable \( X \), and \( f \) is a continuous convex function on the positive Real Line. The Bhattacharyya distance is derived by assigning:

\[
f(x) = -\sqrt{x}
\]

\[
g(x) = -\log(x)
\]

The general form of the Bhattacharyya distance is given by:
\[ Q_{BHA} = \log \left( \rho \left( p_x(x), p_y(y) \right) \right) \]  

(11)

where

\[ \rho \left( p_x(x), p_y(y) \right) = \sqrt{p_x(\xi)p_x(\xi)d\xi} \]  

(12)

stands for the Bhattacharyya coefficient (geometrically interpreted as the cosine of the angle between the pdfs of the two random variables being compared) which measures the amount of separation between the two feature vectors. If the random variables are assumed to be Gaussian, the Bhattacharyya distance may be expressed as:

\[ Q_{BHA} = \frac{1}{4}(\mu_y - \mu_x)^T (\Sigma_x + \Sigma_y)^{-1}(\mu_y - \mu_x) + \frac{1}{2} \log \frac{\Sigma_x + \Sigma_y}{2\sqrt{\Sigma_x \Sigma_y}} \]  

(13)

Hotelling’s T-Square Statistics

Hotelling’s T-Square Statistic is a multivariate generalization of the t-test, which is commonly used in comparing the means of two univariate random variables. Hotelling’s T-Square Statistic can be expressed as:

\[ Q_{T-SQ} (X,Y) = \frac{n_x n_y}{n_x + n_y} \sum_{i=1}^{p} \sum_{k=1}^{q} (\mu_0 - \mu_i) \Sigma^{-1} (\mu_0 - \mu_i) \]  

(14)

where \( C^{ik} \) is the element in the \( i \)th row and \( k \)th column of the inverse of \( C \), the pooled estimate of the covariance matrix for both populations, expressed as:

\[ C = \frac{(n_x - 1)\Sigma_x + (n_y - 1)\Sigma_y}{n_x + n_y - 2} \]  

(15)

Note that the T-Square statistic is simply the square of the T-test, thereby taking into account the correlation of all features in the set simultaneously. It is also a scaled representation of the Mahalanobis distance; the scaling factor is derived from the sizes of the two random variables. Significantly large values of the T-Square statistic indicate more separation between the feature sets being compared.

Mahalanobis Distance

The Mahalanobis distance is simply a modified version of the Euclidean distance. Note, the Euclidean distance does not take into account the correlations of the dataset, and is sensitive to the scale of the measurements. With the Mahalanobis distance, on the other hand, each dimension is given a weight which is inversely proportional to its variance (in order words, the covariance matrices of the random variables are taken into consideration during the distance computation). The Mahalanobis distance is expressed as:

\[ Q_{MAHALANOBIS}(X,Y) = (\mu_x - \mu_y)^T \Sigma^{-1} (\mu_x - \mu_y) \]  

(16)

where \( \Sigma \) is the covariance matrix of the two random variables combined.

3. Speaker Model Formation

Taking into account that in telephone conversations, speakers change rapidly, and speaker change-points are seldom known \textit{a priori} (except if specifically determined), in this research, speakers are represented on a phoneme basis, as it is quite improbable for two speakers’ speech to be contained in the same voiced phoneme. Speaker homogeneous utterances usually consist of more than one voiced phoneme. Models are formed from speech segments created by concatenating \( N \) consecutive voiced phonemes, with \( N > 2 \), as illustrated in Figure 2 (for \( N = 3 \)). 14th order LPCCs are then computed on a frame-by-frame basis, and the mean vector and covariance matrix for each segment is used to form each speaker model.

![Figure 2: Formation of speaker homogeneous segments.](image)

In determining the number of segments used in forming the homogenous speaker segments (or ‘models’), one has to take into account the application of the system. In general, the greater the data size of the speech utterances compared, the better the speaker recognition performance. Nevertheless, in handling conversations, models must be formed with short data lengths in order to avoid overlapping speakers in the same model [9].

Figure 3 shows the intra-speaker and inter-speaker distributions for \( N = 20 \). The Mahalanobis distance was used and the LPCCs were used as features. Each homogeneous speaker model was obtained by concatenating 20 voiced phonemes from as many utterances of the same speaker as was required to yield 20 phonemes. 1000 intra-speaker and inter-speaker distances were computed.

A clear separation can be observed between the intra-speaker and inter-speaker distances when \( N = 20 \). The same trend was observed for all distance and with the LPCCs used as features (figures not shown).
It is, however, very impractical to assume 20 voiced phonemes will be available for each homogeneous speaker utterance in telephone conversations. Note that the 20 phonemes had to be extracted from 15 or more seconds of single-speaker speech data from the HTIMIT database [12]. In practical conversations, a speaker’s utterance is generally about 2 seconds in length on average [8]. It is therefore important to determine an appropriate number of phonemes, which would yield sufficient differentiation between intra- and inter-speaker distances, and also prevent grouping together of segments from two different speakers to form one model. In other words, the least number of segments with adequate separation is desired.

Figures 5 to 7 show the intra-speaker and inter-speaker distributions for N = 5 using the Bhattacharyya distance, the T-Square statistics and the Mahalanobis distance, respectively. The same data as in Figure 3 was used. Note that the separation between the intra-speaker and inter-speaker distances decreased significantly from what was obtained for N=20 (compare for the Mahalanobis distance – Figure 7). The Mahalanobis distance is shown to yield the highest separation, with the least inter class separation obtained from the Bhattacharyya distance (Figure 3). In this research, techniques will be developed to compensate for this performance degradation due to data unavailability. Additionally, information from these three distances will be fused to obtain a final comparison decision.
4. SPEAKER INDEXING

Restrainted-Relative Minimum Distance Approach

The proposed Restrainted-Relative Minimum Distance (RRMD) speaker indexing technique is described in the following steps:

Step 1: All pair-wise distances between homogeneous speaker models (formed as described in Section 3) are computed using each of the distance measures.

Step 2: The two models with the maximum difference between them are selected to be the reference models, as this ensures that they are from two different speakers.

Step 3: Each of the other remaining models are matched to the reference models based on the following conditions

(i) The Restraining Condition: This condition is based on likelihood ratio testing of each distance in order to determine if it matches an intra-speaker or an inter-speaker distance distribution. The distances computed using the distance measures considered in this research can be approximated by either the Gaussian distribution (for the Mahalanobis and Bhattacharyya distances) or the Gamma distribution (for the T-Square statistics). Let \( \alpha_1 \) and \( \beta_1 \) be the intra-speaker parameters and \( \alpha_2 \) and \( \beta_2 \) be the inter-speaker distribution parameters – which would represent the mean(\( \alpha \)) and the variance (\( \beta \)) for the Gaussian case, and the parameters a (\( \alpha \)) and b (\( \beta \)) for the Gamma case. Representative values of \( \alpha_1 \) and \( \beta_1 \) and \( \alpha_2 \) and \( \beta_2 \) can be computed for all the distance measures using a large amount of data from a standard data base (such as the HTIMIT). Subsequently, given a distance value, \( x \), computed between two models, the two models can be said to be from different speakers if the intra-speaker likelihood or probability, \( f(x|\alpha_1,\beta_1) \) is less than the inter-speaker likelihood, \( f(x|\alpha_2,\beta_2) \), and from the same speaker if it is greater than 1. A Distance Likelihood Ratio (DLR) is thus defined as:

\[
DLR = \frac{f(x|\alpha_1,\beta_1)}{f(x|\alpha_2,\beta_2)}
\]  

If the same-speaker and different-speaker cases are assumed to have equal probability, then a DLR value above 1 will indicate that both models are from the same speaker, and if the DLR is below 1, then both models are from different speakers.

If the DLR between the observed model and the reference models to which it is closest is greater than 1, then they are considered to be of the same speaker. If this condition fails, matching is restrained and Condition (ii) below is checked.

(ii) The Relative Distance Condition: This involves computing the difference between the distances of the observed model from both reference models. In other words, let \( d_{\min} \) and \( d_{\max} \) be the distance from the observed model to its closer and farther reference models, respectively. The relatively distance parameter is defined as:

\[
D_{rel} = d_{\min} - d_{\max}
\]  

If the value of \( D_{rel} \) is greater than a threshold, suitably a value around the difference between the means of the intra- and inter-speaker distance distribution, then the restraint is lifted, and the observed model is considered to belong to the same speaker as its closest reference model. The Relative Distance condition is illustrated in Figure 8, which shows the intra- and inter-speaker T-Square distributions for \( N \) (number of voiced phonemes) = 5. The vertical dashed-lines on each distribution represent estimated means.

Figure 8: Illustration of Relative Distance Conditions for Speaker Model Matching. \( D_{rel} = 140 \).
Determining the “Best N”

Initial experiments were performed using the SWITCHBOARD telephone conversation database [13] in order to determine an appropriate value for \( N \), the number of voiced phonemes used for each speaker homogeneous model. 245 conversations were used, each with an average duration of about 400 seconds. In this case, the RRMD indexing technique was not applied. Instead, the two maximally separated models were selected as references, and the remaining were matched to the closest reference. Evaluation was performed using transcriptions provided by the Mississippi Switchboard Transcription Project [14]. The percent accuracy values were determined as follows:

- An index of 1 or 2, representing each of the two speakers in the conversation, was assigned to each speech sample based on the ground truth transcriptions.
- An index of 1 or 2 was assigned to each speech sample based on Minimum Distance (MD) matching (not RRMD) of the reference models.
- Two accuracy computations were obtained as follows: let \( L_T \) be the total number of ground truth samples let \( L_{xy} \) be the number of samples with ground truth index = \( x \) and MD index = \( y \);

\[
\text{Accuracy}_1 = \frac{(L_{11}+L_{22})}{L_T}
\]

\[
\text{Accuracy}_2 = \frac{(L_{21}+L_{12})}{L_T}
\]

The maximum of the two accuracy values was taken as the correct accuracy, expressed as:

\[
\text{Accuracy} = \text{Max(\text{Accuracy}_1, \text{Accuracy}_2)} \times 100\%
\] (20)

Figure 9 shows the percent accuracy (of all 245 conversations) of the basic MD indexing procedure with respect to \( N \), for the T-Square statistics. Note that, as expected, the accuracy increases as \( N \) increases, to a certain value and then begins to decrease, most likely due to the fact that models are being formed using phonemes of different speakers. From Figure 9, the optimum value for \( N \) is observed to be 5.

Evaluation of the RRMD Approach

The intra-speaker parameters \( \alpha_1 \) and \( \beta_1 \), and the inter-speaker parameters, \( \alpha_2 \) and \( \beta_2 \), used in the DLR tests were obtained from 1000 (each) model-based intra- and inter-speaker distances computed using the HTMIT database (for each distance measure). Models were formed for \( N = 5 \), based on the accuracy plot shown in Figure 9. Speaker indexing was then performed on 100 five-minute conversations from SWITCHBOARD. The RRMD technique was used, with the \( D_{\alpha} \) threshold (which controls the amount of undecided speech samples) varying between different ranges for different distance measures depending on the means of their intra- and inter-speaker distributions.

Two types of error were determined, one being the indexing error, computed as \( I_{err} = 100 - \text{Accuracy} \), determined using equation (20). The other error was determined based on the proportion of undecided speech samples that were not labeled as co-channel speech from the ground truth transcription, as compared to the total number of ground truth speaker samples, \( L_T \). In other words, let \( N_u \) be the number of samples declared undecided/unusable by the proposed technique, and let \( N_c \) be the number of samples amongst them that were labeled as co-channel data from the ground truth transcriptions. The ‘undecided error’ was computed as:

\[
U_{Err} = \frac{N_u - N_c}{L_T} \times 100\%
\] (5.4)

Figures 10 to 12 show the average percent error with respect to varying \( D_{\alpha} \) thresholds for both types of errors using the Bhattacharyya distance, the T-Square statistics and the Mahalanobis distance, respectively. The trade-off of higher indexing accuracy for loss of ‘usable’ (non-co-channel) data is evident in all three figures. For the maximum observed threshold, a minimum indexing error is obtained, corresponding to a maximum ‘undecided error’.

![Classification Error with Respect to Mean-Difference Threshold Bhattacharyya](image)

Figure 10: Average percent indexing accuracy with respect to varying \( D_{\alpha} \) thresholds. Indexing was performed using the Bhattacharyya distance with the RRMD Technique.
In the case of the Bhattacharyya distance, the maximum indexing error is observed to be about 7.4% with an undecided error of 3.5%. The minimum indexing error is close to 6.5%, which is also the ‘equal error rate’. Note that, with this distance, increasing the threshold beyond the equal error rate does not result in an increase in the indexing accuracy, even though more data is lost. The T-Square statistics exhibits slightly better performance than the Bhattacharyya distance. In this case, the maximum indexing error is 10% with no data lost. The error can be reduced to close to 5% with about 8% of the data lost. The minimum error rate is about 6%. The performance of the Mahalanobis distance is similar to that of the T-Square statistics, only slightly better in terms of equal error rate and minimum indexing error, but slightly worse in terms of maximum error.

A speaker indexing system for two speaker conversations has been presented. In the next section, a system that assumes no knowledge about the number of participants in the conversations is presented.

5. SPEAKER COUNTING-INDEXING

In this section, a speaker count technique for counting up to K speakers in a conversation is presented and evaluated. The conversation is then indexed depending on the number of speakers determined.

Speaker Count Algorithm

A technique referred to as the Residual Ratio Algorithm (RRA), whereby a speaker count of up to K speakers can be determined, is presented. A detailed description of this technique is given below:

i. Speech models are formed as described in Section 3, with N (number of phonemes used to form one model) = 5.

ii. A reference model is chosen at random, and DLR tests are performed between this model and all others. Every model with a DLR > 1 is considered to belong to the reference speaker, and eliminated from the conversation along with the reference model itself, and the Residual Ratio – the ratio of the size of residual speech to the original size of the conversation - is determined. This completes the first elimination round.

iii. Step ii is repeated for the second round; however, the following procedure is used in order to ensure that the new reference model is valid, i.e., not one that belongs to the first reference speaker, but were erroneously mismatched in the first round. The ratio of size of the speech that was matched to the second reference to the total amount of speech is observed, and the process is repeated until this ratio is greater than a chosen threshold determined \textit{a priori}. Once this condition is satisfied, the Residual Ratio for the second round is determined.

iv. Step iii is repeated until the \((K-1)\)th round.

Ideally, all reference models should belong to different speakers, and all models from the \(k\)th \((k = 1, 2, ..., K-1)\) reference speaker should be eliminated in the \(k\)th round, and if there are \(k\) speakers, the Residual Ratio after the \(k\)th round should be ideally zero. In practice, however, some models may be mismatched in the elimination rounds, with some models belonging to the references being missed, and some models being wrongfully eliminated as illustrated in Figure 13, which shows the elimination stages of an artificial conversation simulated by concatenating speech signals from two different speakers from the HTIMIT database. In other words, the first half of the conversation consisted of speech from one speaker while the second half consisted of speech from another speaker.
and all Speaker 3’s segments were correctly matched and removed. However, some segments from Speaker 2 were also incorrectly removed. Notwithstanding these errors, the ratio of the number of residual segments to the total number of segments was still greater for the three-speaker conversation than for the two-speaker conversation in the experiments shown in Figures 12 and 13.

In determining the speaker count based on the Residual Ratios computed, a tree-type classification procedure where the Residual Ratio is observed for each round, and if it is below a certain threshold (obtained by observing residual ratios for a total of 4,000 artificially generated 1-4 speaker conversations from the HTIMIT database) for the $k^{th}$ round, the speaker count is considered equal to $k$.

**Speaker Indexing**

After the number of speakers in the conversation has been determined, the conversation is indexed as follows:

- Models that initially matched the valid reference models are given the same index (i.e., considered to be of the same speaker) as the reference models.

- Let the speaker count be $C$, if $C < K$, then each unmatched (residual) model is assigned to the model amongst the first $C$ reference models from which it has minimum distance.

- If $C = K$, then Step iv of the RRA is repeated for the $K^{th}$ round. The models which matched the $K^{th}$ reference are assigned the same index as that reference, and the unmatched models are assigned to the model amongst the first $C−1$ reference models from which it has minimum distance.

**Decision-Based Fusion of Distance Measures**

The speaker counting-indexing technique described above was implemented using the Bhattacharyya distance, T-Square statistics and the Mahalanobis distance separately. In order to take information from all three distance measures into consideration, a simple decision-level fusion approach was applied in comparing the models in the conversation to the reference models chosen. This method is a simple voting approach which involves computing DLR tests between the test model and the reference model using all three distances, and then deciding if there’s a match between the models by taking the decision made by the most distance measures. For instance, if GLR values from two distance measures indicate that the test and reference models are from the same speaker, and only one measure indicates that they are not, the two models are considered to match. In indexing, the residual models are assigned to the reference models from which they have the minimum distance, as indicated by the most number of measures.
Experiments and Results

All experiments were performed using artificial conversations from the HTIMIT database, since databases with telephone conversations consisting of more than two speakers are currently unavailable. The intra-speaker parameters $\mu_1$ and $\sigma_1$, and the inter-speaker parameters $\mu_2$ and $\sigma_2$, used in the DLR tests were obtained from 1000 (each) intra- and inter-speaker Mahalanobis Distances as described in Section II. A maximum count of $K = 4$ speakers was considered in this research. Each conversation was about 60 seconds in length, and each speaker contributed an approximately equal amount of speech.

Speaker Count—The proposed RRA technique was tested on 4000 testing conversations having the same statistics as the training data. The speaker count accuracy was determined in three different ways as described below:

1. One or more speakers: an accurate count was obtained if there was one speaker and proposed system yielded a speaker count of 1, or if there were two, three or four speakers, and the proposed system yielded a speaker count greater than 1.

2. One, two or more speakers: an accurate count was obtained if there was one or two speakers and the proposed system yielded a count of one or two, respectively, or if there were two or three speakers and the proposed system yielded a count greater than 2.

3. One, two, three or four speakers: an accurate count was considered if the proposed system yielded the correct number of speakers.

The accuracy rate of the system was obtained as the ratio of the number of correct speaker counts to the total number of conversations. Figure 15 below shows the accuracy rates for each of the distance measures, as well as the results obtained from combining comparison decisions of all three distances.

From Figure 15, it can be observed that the performance of the proposed technique diminishes with increase in the complexity of the task (as a measure of the number of speakers to be determined). Note that the event of a four-speaker telephone conversation is relatively unlikely, compensating for the relatively low accuracy obtained in the one, two, three or four speaker count case. The Mahalanobis distance yields the best performance whilst the Bhattacharyya distance yields the least. Note that combining the decisions results in a slight improvement in the overall performance of the counting system.

Speaker Indexing—The speaker indexing method was evaluated only on conversations for which the speaker count was correct. This amounted to 65% of the conversations when counting up to 4 speakers – i.e. 2600 artificially generated conversations. Figure 16 below shows the indexing accuracy obtained using the three distance measures individually, and by combining their model comparison decisions.

From Figure 16, it is observed that the combined distance method yields the highest accuracy.

6. CONCLUSION

Two speaker indexing techniques have been presented. The first is a simple method with a side-product of being able to detect and remove ‘unusable’ data or data which are liable to cause errors. The second is based on first determining the number of speaker in the conversation and then indexing it. Both systems have been able to yield indexing accuracies of above 90% as show in Sections 4 and 5. Moreover, a speaker count technique, with reasonable counting accuracy was developed in the process. The performances of the proposed systems are comparable to state of the art speaker recognition systems which do not face the challenges of limited data availability and lack of a priori information about the speakers.
This research can be applied to speech data that involves short utterance conversations such as monitoring inmate telephone conversations, detection of unidentified speakers on pilot radio channels, and identification of terrorist by their voices on recorded telephone or radio channels.

A possible further research endeavor could be an attempt to evaluate the indexing accuracy of the system including conversations for which the speaker count determined was incorrect. This would involve the development of a probability metric to determine the likelihood of accurate indexing in spite of a wrong speaker count.

REFERENCES


BIOGRAPHY

Uchechukwu O. Ofogebu received the B.S. and M.S. degrees in electrical engineering from Temple University, Philadelphia, in 2003 and 2005 respectively. She is currently pursuing a Ph.D. degree in engineering at Temple University. Her research interests include speaker discrimination for criminal activity detection applications and enhancing pre-college engineering education. Ms. Ofogebu is a member of the American Society of Engineering Education, the Society of Women Engineers and Eta Kappa Nu.

Ananth N. Iyer received the B.E. degree in electrical and electronic engineering from University of Mysore, India, in 1997 and two M.S. degrees in electrical engineering from Temple University, Philadelphia and the Pennsylvania State University, University Park, PA. He is currently pursuing a Ph.D. degree in engineering at Temple University. His research interests are in modeling for speaker identification and developing numerical methods for signal processing. Mr. Iyer is a member of the IEEE Signal Processing Society, International Speech Communication Association and the American Acoustical Society.

Robert E. Yantorno received the B.S.E.S. degree in 1970 from the University of Rhode Island, Kingston, Rhode Island, and the Ph.D. in Bioengineering in 1978 for the University of Pennsylvania. Dr.
Yantorno joined the faculty of Electrical and Computer Engineering at Temple University, in 1982, and is presently a Professor. He was also past chair of the Philadelphia chapter of the IEEE Signal Processing Society. As chair he organized a number of workshops and a short course, as well as an all-day speech processing educational seminar. He has also been on the organizing committee of a number of national and international conferences. Dr. Yantorno is presently the Director of the Speech Processing Laboratory, and his research interests include; intelligent signal and speech processing, speaker identification, speaker indexing, low-bit-rate speech coding, speech excitation, articulation modeling, and speech quality. He has been a consultant for AT&T Bell Labs in Murray Hill, New Jersey, and also for the Speech Processing Lab of Air Force Research Labs in Rome, New York. Dr. Yantorno has served as a reviewer for the IEEE Transactions on Speech Processing and the Journal of the Acoustical Society of America. Dr. Yantorno received the university-wide Lindback Award for Distinguished Teaching from Temple University, and the “Long-time Continuous Contributions to the Delaware Valley and the Philadelphia Section in Education, Research and Service” from the Philadelphia Section of the IEEE. Dr. Yantorno is a member of the ASEE, IEEE, Tau Beta Pi, Phi Kappa Phi, and Sigma Xi.

Stanley J. Wenndt received his B.S. degree in Electrical Engineering from Iowa State University in 1987. He received his M.S. and Ph.D. in Electrical Engineering from Colorado State University in 1991 and 1997 respectively. Since 1994, he has been employed by the Air Force Research Laboratory in Rome, New York where his research interests are in channel estimation, speaker model mitigation, confidence scores, open-set modeling, dialect identification, and noise removal.