Detection of a Third Speaker in Telephone Conversations

Uchechukwu O. Ofoegbu[1], Ananth N. Iyer[1], Robert E. Yantorno[1] and Stanley J. Wenndt[2]

[1]Speech Processing Laboratory, Temple University, Philadelphia, PA 19122-6077, USA
{uche1, aniyer, byantorn}@temple.edu

[2]Air Force Research Laboratory/IFEC, Rome, NY 13441-4514, USA
Stanley.Wenndt@rl.af.mil

Abstract

Differentiating speakers participating in telephone conversations is a challenging task in speech processing because only short consecutive utterances can be examined for each speaker. Research has shown that, given only brief utterances (1 second or less), humans can recognize speakers with an accuracy of about 54% on average. The task becomes even more challenging when no information about the speakers is known *a priori*. In this paper, a technique for determining whether there are two or three speakers participating in a telephone conversation is presented. This approach assumes no knowledge or information about any of the participating speakers. The technique is based on comparing short utterances within the conversation and deciding whether or not they belong to the same speaker. The applications of this research include 3-way call detection and speaker tracking, and could be extended to speaker change-point detection and indexing. The proposed method involves an elimination process in which speech segments matching two reference models are sequentially removed from the conversation. Models are formed using the mean vectors and covariance matrices of Linear Predictive Cepstral Coefficients of voiced segments in each conversation. Hotelling's T^2-Statistic is used to determine if two models belong to the same or to different speakers based on likelihood ratio testing. The relative amount of residual speech is observed after the elimination process to determine if a third speaker is present. The proposed technique yielded an equal error rate of 20% when tested on artificially simulated conversations from the HTIMIT database and 23% error rate when tested on actual telephone conversations.

Index Terms: Speaker Discrimination, Speaker Count, Telephone Conversations.

1. Introduction

Speaker recognition is a major aspect of speech processing. One common application of speaker recognition is speaker identification (SID), where speaker models are formed using a training dataset containing speech from all the speakers to be examined. An SID system is then tested using a test dataset containing speech from the same speakers [1]. Speaker recognition is also applied in speaker indexing of broadcast news data, where the utterances are labeled according to the participating speakers. This is usually accomplished by first determining speaker change points and then clustering the segments between these change points [2], [3]. Other methods of indexing speech data have also been examined [4], [5]. In the above mentioned applications, however, two important factors must be considered: 1) In SID, information about the speakers is known *a priori*, and, on many occasions, at least 5 seconds of data per speaker is available for comparison. 2) In broadcast data indexing, long speaker consecutive speaker utterances (5 to 20 seconds) are available [3], [5].

In this paper, a method for separating speakers in telephone conversations is presented. No *a priori* information about any speaker in each conversation is known. This poses a challenge in the detection problem because the system cannot be trained with information about the speakers as is the usual practice in SID systems. Moreover, unlike in broadcast news data indexing, only short consecutive utterances can be used for comparison in the case of telephone conversations [6]. The method presented here overcomes these problems by forming speaker models from short segments of the observed data.

The proposed technique is described as follows: for each telephone conversation, models are created using the mean vectors and covariance matrices of 14th order Linear Predictive Cepstral Coefficients of voiced segments. Two reference models are then selectively chosen to represent two different speakers. A sequential elimination procedure, referred to as the Residual Ratio Algorithm, is then performed in which models with relatively small T^2 distances from the reference speaker models are removed from the conversation. The presence of a third speaker is then determined by the relative amount of speech left in the conversation.

The paper is organized as follows: the use of Hotelling's T^2-Statistic in comparing the speaker models is explained in the Section 2. In Section 3, a detailed description of the Residual Ratio Algorithm is given, followed by a presentation of experimental results in Section 4. Conclusions are drawn in Section 5.

2. Comparing models using Hotelling's T^2-statistic

When comparing the means of two univariate random variables, a commonly used test is the t-test. However, for multivariate random variables, a generalization of the t-test is the Hotelling's T^2-Statistic. The Hotelling's T^2-Statistic is simply the square of the t-test and is suggested due to the fact that it takes all variables into consideration simultaneously [7]. This statistic has been shown to improve speaker change point detection accuracy when integrated with Bayesian Information Criterion [3]. With the T^2-Statistic, the covariance matrices of the two random variables being compared are assumed to be approximately equal but unknown. Let $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_p]$ are two multivariate random distributions of lengths n_x and n_y and number of features equal to p. Let $\boldsymbol{\mu}_x$

and $\boldsymbol{\mu}_y$ be the mean vectors of **X** and **Y** respectively and let $\mathbf{C}_x$ and $\mathbf{C}_y$ be their respective covariance matrices. Hotelling's T^2-Statistic can then be expressed as:

$$T^2 = \frac{n_x n_y}{n_x + n_y} \sum_{i=1}^{p} \sum_{k=1}^{p} (\mu_{xi} - \mu_{yi}) c^{ik} (\mu_{xi} - \mu_{yi}) \qquad (1)$$

Where μ_{xi} and μ_{yi} are i^{th} samples of the mean vectors $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ and 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:

$$C = \frac{(n_x - 1)C_x + (n_y - 1)C_y}{n_x + n_y - 2} \qquad (2)$$

Larger values of T^2 indicate more separation between the mean vectors of the two random variables being examined.

In this research, the features in question are the 14th order LPCCs. In order to validate the use of LPCCs as an appropriate feature for the speaker count procedure, some preliminary tests were performed on utterances from the HTIMIT [8] database. The LPCCs were computed on a frame-by-frame basis, with each frame being 30 milliseconds in length. The first test involved computing T^2 values for different speech utterances from the same speaker using all 384 speakers from the HTIMIT database. This was then compared with the T^2 values for speech utterances from different speakers, chosen at random from the database, using a combination of all 384 speakers. The distributions of T^2 values obtained for this initial observation are given in Figure 1 below.

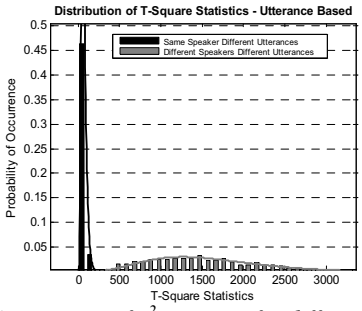

Figure 1 *Comparison of T^2 statistics for different utterances from the same speaker (black bars) and different utterances from different speakers (grey bars).The solid lines are the estimated pdfs of the T^2 statistics.*

It is clear from Figure 1 that two speakers can be effectively discriminated using the T^2-Statistics. Note that it will be impossible to compare whole utterances of speakers in the practical (conversational) application of the T^2-Statistics without prior information about speaker change points. Based on this fact, another experiment was conducted in which segments were used instead of whole utterances. In this case, speaker models were formed using consecutive voice segments from the same utterance (amounting to about 1-second) and compared with another set of five voiced segments from the same or a different speaker's utterance. The use of one or two segments was considered inappropriate, based on observations made on the SWITCHBOARD [9] conversation database that, in any conversation, each speaker's utterance would last for at least one second. This amount of data would contain five voiced segments on average [6]. Moreover, using one or two segments resulted in a sample size that was too small, causing the T^2-Statistics to increase to unreasonably large values due to inconsistent covariance estimates.

Figure 2 shows a comparison of the five-segment T^2-Statistics distributions for the same speaker and different speakers.

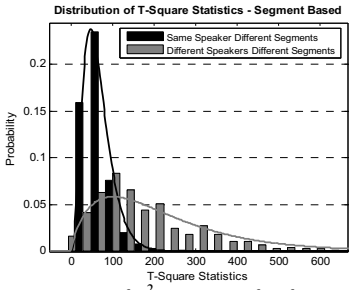

Figure 2 *Comparison of T^2 statistics for five concatenated segments from different utterances of the same speaker (black bars) and utterances from different speakers (grey bars). The solid lines are the estimated pdfs of the T^2 statistics.*

It can be noted from the above figure that the use of segments rather than utterances significantly reduces the separability of the T^2-Statistics for same speaker and different speaker speech files. One way to overcome this problem would be to use larger number of segments. This would be impractical, however, as one might end up putting segments from different speakers together.

Having obtained the T^2 probability distribution functions for models from different speakers, as well as the same speaker, as shown in Figure 2, the problem now becomes one of deciding if two models are from the same or different speakers based on their T^2-Statistic. A direct and simple approach would be to choose a threshold by observing the mean values of both distributions, and making decisions based on this threshold. This approach could be considered sufficient if both distributions were of almost equal variances, and if they were Gaussian. However, Figure 2 shows that there is significant difference between the variances of the single-speaker and the two-speaker distributions. Furthermore, the separation between the means of the two distributions is not very significant. Finally, the pdfs are also not Gaussian, but could be described more appropriately by the more generalized Gamma Distribution. In order to overcome these problems, a T-Square Likelihood Ratio (TSLR) test is introduced and used to determine if two models belong to the same speaker or to different speakers. Let X be a random variable with mean μ and standard deviation σ. The Gamma pdf of X is expressed as:

$$\gamma = f(x \mid a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{x/b} \qquad (3)$$

where Γ(a) is the Gamma function evaluated at *a* and the parameters *a* and *b* are given by the equations:

$$a = \frac{\mu^2}{\sigma^2}; \qquad b = \frac{\sigma^2}{\mu} \qquad (4)$$

Using a standard database, representative values of a and b can be computed for both the same-speaker class and the two-speaker class of T^2 values. Let a_1 and b_1 represent the Gamma parameters for the T^2 distribution obtained when models of the same speaker are compared; and let a_2 and b_2 represent the Gamma parameters for the T^2 distribution obtained when models of the different speakers are compared. The probabilities, $f(x|a_1,b_1)$ and $f(x|a_2,b_2)$ can be computed using equation (3) above. Given the T^2-statistic, x, from any two models, one can determine if the models are from the same or different speakers simply by observing the greater of the two probabilities. In other words, the two models can be said to be from the same speaker if the single-speaker likelihood or probability, $f(x|a_1,b_1)$ is greater than the two-speaker likelihood, $f(x|a_2,b_2)$. The T-Square Likelihood Ratio is thus defined as:

$$TSLR = \frac{f(x \mid a_1, b_1)}{f(x \mid a_2, b_2)} \quad (5)$$

If the single-speaker and different-speaker cases are assumed to have equal probability, then a TSLR value above 1 will indicate that both models are from the same speaker and if the TSLR is below 1, then both models are from different speakers. Note that this test is based, not just on the mean, but also on the variance of the distributions, thereby increasing the accuracy of the T^2-statistic statistic in discriminating speakers. This TSLR test is used in the three-speaker detection algorithm described in the following section.

3. The Residual Ratio Algorithm

The three-speaker detection technique presented here is based on eliminating two speakers from a conversation and observing the relative amount of speech remaining in order to determine whether the conversation consisted of two or three speakers. This technique is referred to as the Residual Ratio Algorithm (RRA), and is described in detail as follows:

i. Speech models are formed from a given conversation by computing the mean vectors and covariance matrices of the 14th order LPCC coefficients of 5 consecutive voiced segments (representing one model).
ii. Two reference models are then chosen by obtaining all pair-wise T^2 statistics for all the models in the conversation and then choosing the two models with the highest T^2 values between them.
iii. TSLR tests are performed between one of the reference models and all other models (segments), and every model with a TSLR > 1 is considered to belong to the reference speaker and eliminated from the conversation.
iv. Step ii is repeated, using the other reference model
v. The ratio of the number of voiced segments remaining to the total number of voiced segments in the conversation is computed as the Residual Ratio for that conversation.

Ideally, all models from the first speaker should be eliminated in step ii, and all voiced segments from the second speaker should be eliminated in step iii. If the conversation consists of 2 speakers, the Residual Ratio should be zero. In practice, however, some segments from the first and second speakers may be missed in the two elimination rounds. This is illustrated in Figure 3, which shows the elimination stages of an artificial conversation simulated from three different speakers (i.e., 10 seconds each of speech data from three speakers concatenated) from the HTIMIT database. The white spaces between the colors represent the unvoiced and silence portions which were removed before processing the data.

In Figure 3, the algorithm is shown to have successfully eliminated most of the first speaker's segments in the first round (the reference model was from the first speaker) but erroneously removed few segments from Speakers 2 and 3 also. In the second round, the reference model is from Speaker 2 and all Speaker 2's segments were correctly identified and removed. However, just as the first round, some segments from Speaker 2 were also incorrectly removed. Notwithstanding these errors, it can be observed that, had there been two speakers in the conversation (only speakers 1 and 2 for instance), there would have been no speech segments remaining in the conversation after the two elimination stages. In other words, the ratio of the number of residual segments to the total number of segments is expected to be greater for three-speaker conversations than for two-speaker conversations.

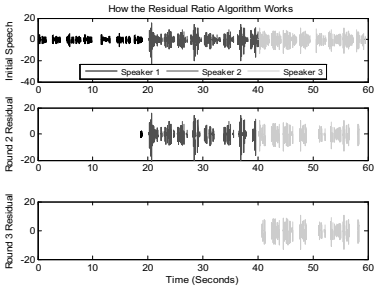

Figure 3 *Three-Speaker Illustration of the Residual Ratio Algorithm. Original Speech (top panel), speech remaining after first elimination round residual (middle panel), speech remaining after second elimination round (bottom panel).*

In order to take models other than the references into consideration during the elimination process, a reference modification step was added to the algorithm. This modification involves comparing each model, not only with the reference model, but with all prior matching models and then taking the average TSLR for all the comparisons. In other words, if N models have been matched to the reference model, then the currently observed model is said to match if

$$\frac{1}{N}\sum_{i=1}^{N} TSLR_i > 1 \quad (6)$$

Where $TSLR_i$ is the TSLR obtained (using (5)) from the comparison of the currently observed to model to the i^{th} matched model.

4. Experiments and results

The proposed technique was tested on 100 artificial two-speaker and 100 artificial three-speaker conversations simulated from the HTIMIT database. All 200 conversations were of 60 seconds in length and each speaker contributed the same amount of speech. The approach was also tested on 265 actual telephone conversations. Twenty-three (23) two-speaker conversations with lengths of about 60 seconds on average (with each speaker speaking for about 50% of the duration) were available in this database. One speaker was repeated in all 23 conversations;

therefore, by ensuring that the other speaker in each of the two combined conversations was different, three-speaker data could be created. Only 242 such data were possible. The total length of each three-speaker telephone data was about 2 minutes, and the amount of contribution per speaker in each conversation varied from conversation to conversation. One speaker spoke about half of the time, while each of the other two speakers spoke for about 25% of the time on average. It must be noted that no standard three-way telephone conversation database is currently available; hence the creation of such data from an existing two-way telephone database (Database 2), and a standard speaker identification database recorded over the telephone (HTIMIT).

The parameters a_1 and b_1 were determined by computing T^2-statistics between two models from the same speaker for all 384 files in the HTIMIT database. The parameters a_2 and b_2 were also computed for models of different speakers by comparing each speaker in the database with a different speaker chosen at random from the database. These values were used in the TSLR tests for all experiments presented.

Using the RRA as described in the previous section, Residual Ratios were computed for the test conversations described above. The problem was treated as a three-speaker detection problem, and a third speaker was considered present in a conversation if the Residual Ratio was more than a chosen threshold (which could be chosen from Figure 4). Using all the Residual Ratios obtained for both classes of conversations as thresholds, percent detection hits and false alarms were computed for each of the classes as follows:

i. If the algorithm detected a third speaker in a three-speaker conversation, a hit was said to have occurred.
ii. If the algorithm detected a third speaker in a two-speaker conversation, a false alarm was said to have occurred.
iii. If the algorithm failed to detect a third speaker in a three-speaker conversation, a miss was said to have occurred.

Classification error curves were obtained by plotting the percent misses and false alarms against all possible thresholds (all Residual Ratios). Figure 4 shows classification curves for the HTIMIT (black) conversations and for Database 2 conversations (grey). Misses and false alarms are plotted in solid and dotted lines respectively. It is observed from Figure 4 that equal classification error rates of 20% for the HTIMIT and about 23% for Database 2 are obtained. The lengths of the conversations were varied between 30 seconds and two minutes, but no significant difference in results was observed in spite of these differences in length. From Figure 4, it is clearly observed that not having equal contribution from all speakers does not adversely affect the performance of the technique.

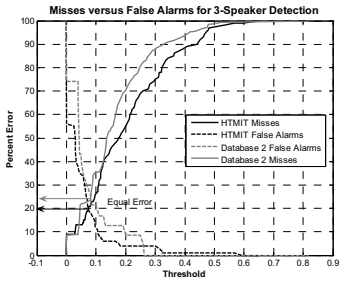

Figure 4 *Classification Error Curves for three-speaker detection using 200 artificial conversations from the HTMIT database (grey) and 265 conversations from the Database 2 (black).*

5. Conclusion

Comparing speakers participating in a telephone conversation is usually a difficult task because, even when speaker change points are known, only short utterances per speaker can be obtained for comparison. Studies have shown that humans are only able to differentiate between speakers using brief (one second or less) utterances with an accuracy of about 54% [10]. Moreover, attempts to distinguish between speakers using short utterance lengths in conversations have reported up to 41% detection error [11]. The proposed technique has been able to detect the presence of a third speaker in telephone data with accuracy comparable to human performance, with no *a priori* knowledge of speaker-change points or any of the participating speakers. Results also indicate that the proposed technique is reasonably independent of data. Further enhancements of this research could include a more discriminative approach to updating reference models and also the formation of models based on pre-determined speaker change points. Applications of this research can be extended to automatic segmentation and indexing of telephone conversations.

6. Acknowledgements

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number FA8750-04-1-0146.

7. References

[1] Reynolds, D. A. and Rose, R. C., "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," IEEE Trans. Speech and Audio Process., pp. 72–83, 1995.

[2] Chen, S., Gopalakrishnan, P., "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion". Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[3] Zhou, B. W. and Hansen, J. H., "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion", Proceedings of ICSLP, 2000.

[4] Kwon, S. and Narayanan, S., "Unsupervised speaker Indexing using Generic Models", IEEE Trans. on Speech and Audio Processing, vol. 13 (5), pp.1004-1013, 2004.

[5] Nishida, M. and Ariki, Y., Real Time Speaker Indexing Based on Subspace Method - Application To TV News Articles and Debate, Proceedings of ICSLP, 1998.

[6] Iyer, A. N., Ofoegbu, U. O., Yantorno, R. E., Wenndt, S. J., "Speaker Modeling in Conversational Speech with Application to Speaker-Count", ICSLP, 2006 (submitted).

[7] Manly, B., "Multivariate Statistical Methods -A Primer", 2nd edition, Chapman & Hall. 1994

[8] Reynolds, D., "HTIMIT and LLHDB: Speech corpora for the Study of Handset Transducer Effects", ICASSP, 1997.

[9] Godfrey, J. J., Holliman, E. C., and McDaniel, J., "SWITCHBOARD: Telephone Speech Corpus for Research and Development", Proceedings of ICASSP, 1992.

[10] O'Shaughnessy, D., "Speech Communications: Human and Machine", 2nd edition, Wiley-IEEE Press. 1999.

[11] Delacourt, P., Kryze, D. and Wellekens, C. J., "Speaker-based Segmentation for Audio Data Indexing", Proceedings of the ESCA ETRW workshop, UK, 1999.