S13.11

Segregation of Speakers for Speech Recognition and Speaker Identification

Herbert Gish, Man-Hung Siu, and Robin Rohlicek
BBN Systems and Technologies
Cambridge, MA 02138

ABSTRACT

A method for segregating speech from speakers engaged in dialogs is described. The method, assuming no prior knowledge of the speakers, employs a distance measure between speech segments used in conjunction with a clustering algorithm, to perform the segregation. Properties of the distance measure are discussed and an air traffic control application is described.

1. Introduction

In many speech recognition and speaker identification applications it is often assumed that the speech from a particular individual is available for processing. When this isn't the case, and the speech from the desired speaker is intermixed with the speech from other speakers, the speech must be segregated into the speech from the individuals before the recognition or identification process can commence. The particular application we consider is the case where we have the dialog between an air traffic controller (ATC) and several pilots and we want to automatically determine the commands which the controller gave to the pilots. This application requires the separation of the controller utterances from those of the pilots before recognition can begin. Although we focus on the ATC application the methodology we present is also applicable to other situations. These include the problem of text-independent speaker identification, when there is more than one speaker.

We assume that we have no *a priori* information about any of the speakers. We further assume that the total number of speakers is also unknown. Our approach is to view the problem as one of unsupervised clustering of speakers, with each cluster representing a different speaker. The association of segments of speech with each of the clusters represents the desired segregation.

The paper focuses on the development of a distance measure between any two segments of speech, where the distance reflects whether the two segments are from the same speaker. This distance measure serves as the basis for the clustering algorithm. We discuss both the theoretical and empirical properties of the measure.

2. A Distance Between Speech Utterances

Consider that we have two segments of speech each of which is characterized by a sequence of spectral feature vectors, which we will denote by $x_n, n = 1, \ldots, N_1$, and $y_n, n = 1, \ldots, N_2$, respectively. We assume that the vectors in each of these sequences can be modeled as coming from a multivariate Gaussian distribution, and that the vectors are statistically independent. The question we pose with respect to the sequences (or original segments) is whether or not they come from the same underlying model, or, equivalently, whether the segments were uttered by the same speaker.

Formally we have the following hypothesis test:

H_o: the segments were generated by the same speaker

and,

H_1: the segments were generated by different speakers.

We test this hypothesis by using the generalized likelihood ratio test, i.e., we form the likelihood ratio of the observations with the unknown model parameters replaced by their maximum likelihood estimates. If $L(x; \mu_1, \Sigma_1)$ denotes the likelihood of the x sequence and $L(y; \mu_2, \Sigma_2)$ denote the likelihood of the y sequence then the likelihood, L_1, of the two segments being generated by different speakers is given by, $L_1 = L(x; \mu_1, \Sigma_1)L(y; \mu_2, \Sigma_2)$. Furthermore, the likelihood of the segments being generated by the same speaker is given by $L_o = L(z; \mu, \Sigma)$, where z is the union of the x and y segments.

If we let λ denote the likelihood ratio, then $\lambda = \frac{L_o}{L_1}$, giving

$$\lambda = \frac{L(z; \hat{\mu}, \hat{\Sigma})}{L(x; \hat{\mu}_1, \hat{\Sigma}_1)L(y; \hat{\mu}_2, \hat{\Sigma}_2)} \tag{1}$$

where the hat denotes the maximum likelihood estimate.

If we now use the multivariate Gaussian models in the likelihood expression we obtain that the likelihood ratio can be written as

$$\lambda = \lambda_{COV}\lambda_{MEAN} \tag{2}$$

where λ_{COV} is the likelihood ratio that tests the hypothesis that the two segments are from the same Gaussian models with the same covariance matrix, with no assumption being made about the equality of the means, and λ_{MEAN} is the likelihood ratio that tests the hypothesis that the two segments are from the same Gaussian models with the same mean, with no assumption being made about the equality of the covariances.

We have,

$$\lambda_{COV} = \left(\frac{|S_1|^{\alpha}|S_2|^{1-\alpha}}{|W|} \right)^{\frac{N}{2}} \tag{3}$$

where $\alpha = \frac{N_1}{N}$, S_1 and S_2 are the sample covariance matrices for each of the two segments, and W is their frequency weighted

CH2977-7/91/0000-0873 $1.00 © 1991 IEEE

average, viz., $W = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$ Also,

$$\lambda_{MEAN} = \left(1 + \frac{N_1 N_2}{N^2} (\overline{x_1} - \overline{x_2})^t W^{-1} (\overline{x_1} - \overline{x_2})\right)^{\frac{-N}{2}} \quad (4)$$

where $\overline{x_1}$ and $\overline{x_2}$ are the means of the segments. Derivation of the these formulas is based on well-known results in multivariate analysis (ref. [1] for example).

From these likelihoods we obtain our "distances" between segments by taking the negative of their logarithms: $d_{COV} = -\log \lambda_{COV}$ and $d_{MEAN} = -\log \lambda_{MEAN}$.

Since the generalized likelihood ratio is always greater than zero and less than unity the above "distances" are always positive, although they do not satisfy the triangle inequality.

3. Theoretical Distribution of the Distances

In order to understand the statistical behavior of the results of applying the distance measures we must understand their statistical properties. Below we will discuss the theoretical basis for the distribution of the distances and then compare the theory with experiment.

The distances are parameterized by N_1, N_2, and their sum, N. When the null hypothesis is true, i.e., when the segments are from the same speaker, $2 * d$ (twice the distance), has a chi-square distribution when N is large [1]. The degrees of freedom of this chi-square is equal to the difference in the dimensionalities of the parameter spaces under the alternative and null hypotheses. If r_1 denotes the number of parameters under the null hypothesis and r_o denotes the number of parameters under the alternative, $r = r_1 - r_o$ is the number of degrees of freedom. For example, if we are utilizing spectral feature vectors with p components, and we are only considering the covariance matrices, then under H_1 there are two covariance matrices each with $p(p+1)/2$ parameters, while under H_o there is only a single covariance matrix. Thus, for this case, $r = p(p+1)/2$. In the case of the means we would have $r = p$ degrees of freedom.

We note that the asymptotic distribution does not depend on N_1,N_2, or N. We further note that this is true only under the null hypothesis. Under the alternative hypothesis the asymptotic distribution of course depends on the way the speakers differ.

4. Experimental Results

In these experiments we have only considered d_{COV}. Similar results, though not presented, were obtained with d_{MEAN}. The two distances are rather different in that they capture very different information about the speakers and the communication channels being used. For Gaussian distributions they are statistically independent. In any application one or both distances can be used depending on the particular circumstances, such as the important sources of channel variability. We note that channel, as well as speaker characteristics, play an important role in the segregation of speakers. The distance, d_{COV}, is invariant to time-invariant filter of the speech, whereas d_{MEAN} is not.

The Database

The database we will use for our experiments consists of speech collected from air traffic control dialogs between controllers and pilots. The data is collected off the air and represents a wide range of transmission quality, with the signal-to-noise ratio in the average transmission being about 15dB. The received speech is filtered to a band of 300-3300 Hz. Through out our study the spectral features we use are mel-warped cepstra, $c_1, \ldots, c_{14}$. The feature vectors do not include the cepstral coefficient c_0. The cepstral vectors are calculated every 10 ms using a 20 ms window. The cepstral vectors are screened on the basis of frame energy, and only a fraction of the highest energy frames are used. Typically, 40% of the frames will be discarded.

The stream of speech from the dialogs is separated into the speech segments from individuals primarily on the basis of energy measurements, and this is done quite reliably. Some of the segments are from the controller and the others are from the different pilots with whom the controller is engaged in dialog. In Figures 1. and 2. we have plotted histograms for the durations for the segment durations for the controllers and pilots. The controllers are seen to have a longer tailed distribution, with the median durations for controllers and pilots being 2.8 and 2.1 seconds, respectively.

The Estimated Distribution of the Distance

We take the segments from the controller and pilots and evaluate d_{COV} between all segments; these distances are collected into two sets: one set of distances from segments from the same individual, and the other composed of distances from different individuals. We have further quantified these sets of distances on the basis of the durations of each of the segments entering the distance computation. These duration categories, which are quantized, serve as index in to the collection of distances. We have 5 duration categories consisting of 0-1, 1-2, 2-3, 3-4, and 4-15 seconds, respectively. For example, we have a collection of distances that corresponds to the case where on segment has duration of 1-2 seconds and the other 3-4 seconds duration, under the conditions when the speakers are the same and when they are different. These categorizations will enable us to understand the effects of the different durations on the resulting distributions.

The asymptotic chi-square property of the null hypothesis distribution has led us to consider the gamma distribution as a means of fitting a distribution to the data. The gamma includes the chi-square as a special case and therefore is seems to be a good choice for being able to model deviations from the theoretical predictions. The gamma probability density function is given by

$$p(x) = \frac{(x/b)^{c-1}}{\Gamma(c)} e^{-(x/b)} \quad 0 \le x < \infty \quad (5)$$

The chi-square density function corresponds to having $b = 2$ and c being 1/2 the number of degrees of freedom of the chi-square.

The density function is seen to have two parameters, b and c, which are referred to as being the scale parameter and shape parameter, respectively. The method we will use to fit this model to the data is the method of matching moments. The estimators are given by $\hat{b} = \frac{s^2}{\bar{x}}$ and, $\hat{c} = \left(\frac{\bar{x}}{s}\right)^2$. where $\bar{x}$ and s^2 are the sample mean and the sample variance, respectively. See Hastings and Peacock [2] for additional details about properties of the gamma and chi-square distributions.

In order to assess the fit of the gamma distribution to the data we, as an example, consider segments of 3-4 seconds duration. From these segments we compute a set of distances for the same speakers and another set of distances for different speakers. Figure 3. shows the fit of the gamma density functions for both of these cases. The quality of fit we observe in this figure is typical of what we have observed in general, which seem to be quite reasonable. In the following, we will use the gamma density as being representative of the population and not present further comparisons to the data.

Distributions as a Function of Duration

In order to be able to understand the the behavior of the distance between segments we must understand its behavior as a function of the durations of the segments. To this end, we will focus our attention on the specific case of distances between segments that were of 3-4 seconds in duration, to segments that were 0-1, 1-2, 2-3, 3-4, and 4-15 seconds in duration. On the basis of increasing total durations of the segments used in computing the distance, we refer to the sets of distances by the numbers 1-5, for both the same speaker distances and different speaker distances.

In Figure 4. we have plotted the pdf's of the same speaker distances for the different total durations. The important feature of this plot is that for increasing total duration, we have different distributions, indicating that an asymptotic distribution has not been achieved. There are a several reasons for this, one of which is that we do not have enough data to be in the asymptotic region; another reason is that the Gaussian model is only an approximation, since the data is not Gaussian. Another factor is the variability of the channel due to changing snr's.

We also observed that the c parameter of the gamma pdf's remained fairly constant at about 22.5 for the various values of total duration. This implies that the degrees of freedom in the gamma density (looking at the gamma as being a scaled chi-square) is about 45, which implies that the of the spectral information for the speaker lies in a 9 dimensional sub-space of cepstral space. An analysis of the the principal components of the the covariance matrices shows that shows that over 95% of the speech energy lies in in the first 9 principal component directions. Thus the degrees of freedom observed in the gamma distribution may well be a reflection that the true dimensionality of the parameter space is significantly less than the nominal dimensionality.

Going to Figure 5., which corresponds to the case of distances between segments from different speakers, we also see variability in the pdf's. The pdf's also having increasing mean value with increasing total segment duration. We can assess the effect of increasing the total duration, with respect to hypothesis discrimination, by considering the classification performance of a pair of segments for each of the five quantized durations. If we let $p_o(d_{COV}|\text{duration})$ and $p_1(d_{COV}|\text{duration})$ denote the pdf's of d_{COV} under the same and different speaker conditions, respectively, we can classify pairs of segments as being from the same or different speakers depending on the whether p_o/p_1 is greater or less than 1. This gives us the classification error rates as a function of duration givean in Table 1., showing improvement in classification performance with increasing duration.

Total duration having an effect on classification performance implies that d_{COV} has a different meaning depending on the duration. Figure 6. shows how the likelihood ratios, $\log(p_o/p_1)$, vary as a function of d_{COV}, for the 5 different total durations. These curves can serve as the basis for normalizing d_{COV} for the effects of total duration, e.g., equal distances correspond to equal values of likelihood ratio.

The above described experiments considered the effects of increasing total duration on the ability to distinguish the hypotheses. In a related experiment we attempted to maintain the total duration constant and examine the effect of imbalance of the durations on the pdf's of the distances. In this case we compared the pdf's generated by segment pairs both of which were in the 2-3s range against segment pairs that were in the 1-2 and 3-4s range. We did observe that there was a change in pdf's under both hypotheses, with the equal duration case having higher somewhat higher means and offering slightly greater discriminability.

Application to Clustering

The method of clustering that we employed is a conventional agglomerative technique for the construction of dendograms (see for example Everitt [3]). We form the distances between all segments and combine the closest pair of segments into a single set. This process is then repeated where the distance between two sets of segments is taken to be the maximum distance between members from each of the sets. The resulting dendogram is then split by finding the largest cluster, which is declared the cluster of controller segments, with all other segments attributed to pilots.

We have run several experiments, all giving encouraging results. Our largest experiment consisted of 423 segments, 220 of which were from the controller and 203 were from pilots. Application of d_{COV} (not normalized by duration) as the measure of the distance between the segments, in conjunction with the agglomerative algorithm noted above, resulted in a total of 6 errors. Three controller segments were misclassified as having been pilots and 3 pilot segments identified as controllers.

Note that although d_{COV} is invariant to the effects of linear time-invariant filtering, it can be aided in discriminating between pilots and controllers by aspects of channels that affect the variability of spectral information, such as additive noise.

5. Discussion

We have developed and approach for segregating speech from different speakers that requires no prior information about the speakers. The method is uses "distance" between speech segments that is based on the likehood ratio of speech segments being generated from a common model, using multivariate Gaussian assumptions.

We have explored the statistical behavior of the distance function and have compared its empirical distributions to those based on theory. A better understanding of these distributions depend on measurable quantities, such as N, N_1, and N_2, will enable us to improve our performance by better accounting for the affects of these quantities. This will be important to applying our methodology to more difficult problems, e.g., separating the individual pilots from each other.

In our application obtaining speech segments that consisted of speech from individuals was not difficult. In other applications forming of segments can be challenging and result in more complex algorithms.

6. References

1. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, J. Wiley & Sons, New York, 1984, pp. 404-450.

2. N. A. J. Hastings and J. B. Peacock, *Statistical Distributions*, John Wiley & Sons, New York, 1974, pp. 68-73.

3. B. Everitt, *Cluster Analysis*, Halsted Press, New York, 1980, pp. 24-35.

Acknowledgment

This work was supported by Defense Advanced Projects Agency and Rome Laboratories under Rome Laboratories contract number F30602-89-C-0170.

segment durations 3-4 (sec) &	0-1	1-2	2-3	3-4	4-15
classification error (%)	39	21	14	12	10

Table 1: Error as a function of duration

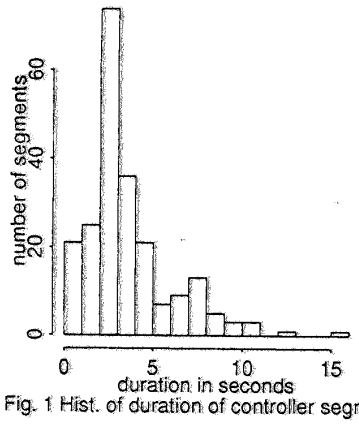

Fig. 1 Hist. of duration of controller segments

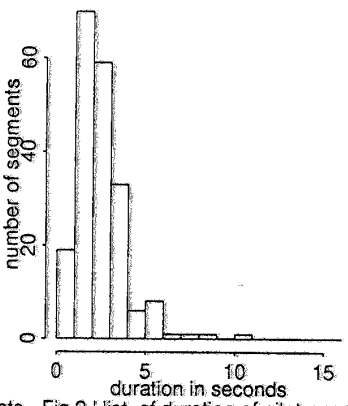

Fig 2 Hist. of duration of pilot segs.

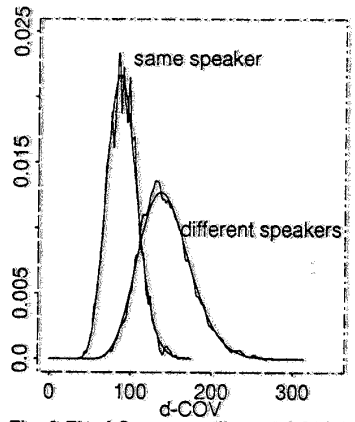

Fig. 3 Fit of Gamma pdf's to d-COV data

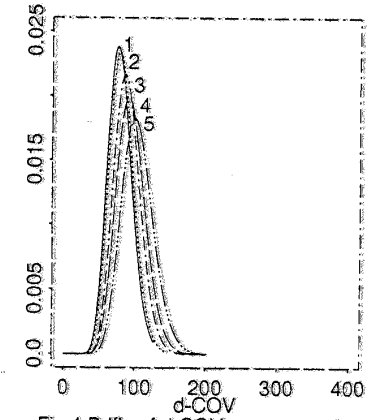

Fig 4 Pdf's of d-COV: same speakers

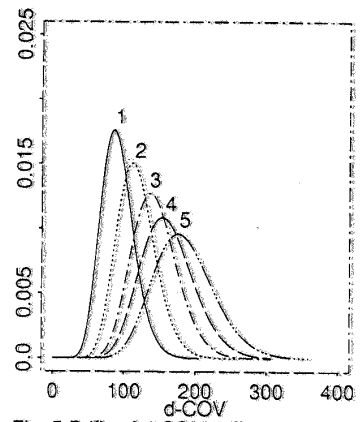

Fig. 5 Pdf's of d-COV: different speakers

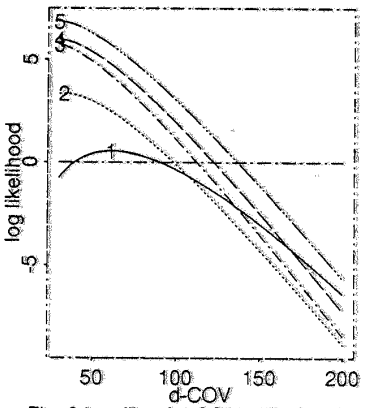

Fig. 6 Log lik. of d-COV: diff. durations