Separation of Non-spontaneous and Spontaneous Speech.

Owen.P. Kenny

Signals Analysis Discipline
Communications Division
Defence Science and Technology Organization
PO Box 1500, Salisbury, 5108
South Australia
Australia.

Douglas J.Nelson, John S. Bodenschatz and Heather A. McMonagle

Department of Defence
9800 Savage Road
Fort George G. Meade, Maryland 20755
USA.

Abstract

There are many situations in which it is desirable to be able to distinguish spontaneous speech and speech which is non-spontaneous. Examples of situations in which this problem may arise include forensic evidence situations, sorting voice-mail responses from voice-mail menus, and automatic segmentation of spontaneous responses from prepared questions. The later situation can occur if it is desired to create a database of spontaneous data from data which consists of spontaneous discourse responding to prepared prompts. This paper outlines and compares three methods for automatically classifying spontaneous and non-spontaneous speech and presents the experimental results comparing the performance of the methods. All three methods are based on an analysis of the probability distributions of prosodic features extracted from the speech signal. The first method uses an expansion of the of the probability distribution in terms of the statistical moments. The second method is an application of a modified Hellinger's method applied to histograms of signal amplitude and other speech features. The third method is based on a measure of the non-Gaussianity of the data.

1.0 Introduction.

There are many situations in which it is desirable to be able to distinguish spontaneous speech from non-spontaneous speech. Examples of situations in which this problem may arise include forensic evidence situations, sorting voice-mail responses from voice-mail menus, and automatic segmentation of spontaneous responses and prepared questions in a question and answer dialogue. The later situation can occur if it is desired to create a database of spontaneous data from data which consists of spontaneous discourse responding to prepared prompts.

In this paper we shall describe different means of separating non-spontaneous and spontaneous speech. We refer to non-spontaneous speech as speech which has been recorded from prepared text, or in a situation in which the speaker has thought about and prepared the content of the utterance. When people speak to intentionally record a message, or when people read from prepared text, they have knowledge of what they want to say, and it normally sounds rehearsed or non-spontaneous. Spontaneous speech results when the speaker must spontaneously create the conversation as it is spoken. Spontaneous speech tends to sound much more natural, and less "flat" than non-spontaneous speech. This is certainly a gray art since a professional actor can speak the same lines every night for years and make it sound natural, while there are many people who have difficulties communicating, with the result that their speech seems unnatural.

In order to separate these two classes of speech one has to ask what features can be used effectively. For purposes of this paper, we assume that spontaneity can be determined from the regularity of prosodic speech features. These features include pitch, speaking rate and energy dynamics, to name a few. It is reasonable to expect features such as these to be able to detect spontaneity. People reading or reciting from prepared text do not have to think of more than a few words to trigger the entire utterance. On the other hand the person who is speaking spontaneously creates the conversation as they go. Pauses in spontaneous speech occur as the speaker composes and executes the next train of thought. As a consequence the speaking rate of the two classes is different and there is naturally more/longer dead time in the spontaneous speech than non-spontaneous speech. When reading text, the tendency is to read the individual words or look ahead to the next few words while speaking. In either case, the speaking rate tends to be fairly regular and the dynamic range of the speech energy tends to be low. In spontaneous speech, the dynamic range of energy appears to be much greater. Pitch is not explicitly addressed in this paper, but one would expect the pitch dynamic range (inflection) for spontaneous speech to be much greater than for non-spontaneous speech. Many prosodic features are easily extracted using standard signal processing techniques. Energy dynamics and the speaking rate are recovered from the energy envelop of the signal. Pitch, inflection and other prosodic features require a more delicate process.

The three basic approaches which are presented here are based on modeling the probability density functions of the prosodic features of spontaneous and non-spontaneous speech to devise different distance measures for these densities. These distance measures are then used to distinguish the speech classes. In the first method, the density is expanded into its moments, which are used as coordinates of an Euclidean space. The second method is based on Hellinger's distance measure. This measure is actually an Euclidean distance on a Hilbert space whose elements are essentially probability density functions. In the third method, the density functions are modeled as symmetrical alpha stable density functions and the value of alpha is estimated using maximum likelihood parameter estimation. Since this problem is a closed set discrimination problem in which it is assumed that each data

segment consists of speech which is either spontaneous or non-spontaneous, it is sufficient to find a single decision boundary which separates the classes in the feature space. The decision boundary is estimated for each of the methods, and the methods are applied to approximately 100 utterances of hand segmented test data. Each of the discrimination methods is presented and, for each of the three approaches an error plot is given to determine its performance.

2.0 Calculation of the Speech Envelop

I n designing the speech classification algorithms, many potential features were examined. Of all of the features, it was found that by far the best were the statistics of the energy envelop of speech. The energy envelop is the time sequence which is the instantaneous power of the speech signal. The energy envelope of the speech waveform is obtained from the analytic signal by taking the natural logarithm, exponentially weigthing the real part and then low passed filtering the result. The procedure is summarized in the following set of equation.

$$Z(t) = S(t) + jH[S(t)] \tag{1}$$

where $H[x]$ denotes the Hilbert transform. Eqn (1) can be rewritten in the form,

$$Z(t) = a(t)e^{j\phi(t)}, \tag{2}$$

where $a(t))$ is the positive real envelop, and $\phi(t)$ is the signal phase. The energy envelope of the signal can then be computed from the complex log of the analytic signal

$$\ln(Z(t)) = \ln(a(t)) + j\phi \tag{3}$$

by exponentiation to produce

$$a(t) = Exp\{Re(\ln(Z(t)))\}. \tag{4}$$

An example of envelope extraction is shown in Figs (1a) & (1b) where (1a) depicts the speech waveform and (1b) its corresponding envelope. It is this envelope which was used to form the density function for the speech classes.

3.0 Features for spontaneous and non-spontaneous speech separation.

The sections which follow describe the techniques used to separate the classes of speech. In all the cases described the envelope of the waveform is obtained and used as the feature vector. Many features were considered, and results were computed using the different features, but the performance using the other features proved to be far inferior to the waveform envelop. In addition, including additional features in the feature vector did not improve the results obtained using the waveform envelop alone.

In processing the data, the major problem is the parameterization of the feature space in such a way that the density functions for the different data types can be easily parameterized and compared. Three processes were selected and the performance results of the three processes are presented here for comparison. The first method is to expand the estimated probability distributions in terms of the central moments. The principal components of this expansion are then used as a basis of comparison of the distributions. In the second method, a variation of Hellinger's method is used. In Hellingers method, probability density functions are imbedded on the unit sphere in a Hilbert space by computing the square root of the density function. Hilbert space techniques combined with principal component method may then be used to cluster the data. In the third method, the distribution is estimated from the sampled data, and the distribution is modeled as being symmetric and alpha stable. The order α of the distribution is estimated, and used to discriminated the data types.

It should be noted that in each case, care was taken to rescale the signal magnitude so that the average power of each speech data set was unity. In doing this, the data was normalized so that there was no signal bias to contribute to the discrimination of the two signal classes. In normalizing to unit power, the signals were essentially normalized to have zero mean voltage and unit variance.

3.1 Density function separation by moment expansion.

One method to classify a density function is by means of moment expansion of the density function. These moments can be used to characterize the density. In addition, the density function can be reconstructed as a power series whose coefficients are the moments $P_n = E(x^n)$, where the reconstructed density function is

$$G(t) = p_0 + p_1 t + p_2 t^2 + \ldots . \tag{5}$$

The central moments are calculated from the signal envelope directly by considering the envelope as a sampled sequence. The central moments are then computed as,

$$\mu_r = \{\sum (x_i - \bar{x})^r\}/n, \tag{6}$$

where μ_r denotes the rth moment, $\bar{x}$ is the mean of the sample points, n is the number of data points, and x_i denote the sample points.

Finally the normalized central moments can be computed from the central moments by normalizing by an appropriate power of the second central moment. The normalized central moments are given by [3] as

$$\gamma_n = \mu_n/(\mu_2)^{n/2}. \tag{7}$$

A two dimensional Hilbert subspace was constructed from the data sets using the first and third normalized central moments, and the data was plotted in this space. The moments for each data set were plotted on a cartesian plane, as shown in Fig. (2) with the normalized first moment oriented in the y direction and the normalized third moment oriented in the x direction. As it can be seen the spontaneous and non-spontaneous sets of data tend to separate in this plane. The question now is to determine how well this separation method works. To obtain some measure, a threshold level was set for the normalized mean direction which was moved from 0.2 to 2.0 in interval steps of.01. The number of erroneous points were counted for each threshold position. The results of this procedure was plotted in Fig (3). It can be seen from this plot that a minimum error can be obtained for a unique threshold level. For the observation data used the minimum error was found to be around 10 percent.

3.2 Modified Hellinger's method for class separation

In Hellinger's method, the probability density functions of the desired classes are estimated from the training data. The probability density functions are estimated from each of the data sets in the test data, and the test distributions are compared to the training distributions by imbedding the space of probability distributions in a

Hilbert space by computing the square roots of the distributions[1]. To reduce the dimension of the problem, the data is projected into a principle component space, and the decisions are made within this reduced dimension space.

To begin the process, the amplitude envelope of the speech waveform is computed and the amplitude distribution is estimated by computing a histogram of the amplitudes. The distribution is normalized by scaling the observed signal to make the mean signal amplitude equal to one, making the signals invariant to receiver gain. The square root of the histogram counts are computed, and the resulting square-root distribution is then used as a feature vector. These feature vectors are the square root of the probability density function of the envelope of the signal.

To make this process more clear, we let e_1 and e_2 be the amplitude distributions of the spontaneous and non-spontaneous data, respectively. We define the Hellinger feature vectors u_1 and u_2 by

$$u_i = \sqrt{e_i} \tag{8}$$

Since probability distributions integrate to 1, the u_i defined by the condition (8) form a Hilbert space, since

$$\int_{-\infty}^{\infty} u_i(x)u_i(x)dx = 1 . \tag{9}$$

The ultimate goal is to project the observed feature vectors into vector spaces constructed from the training data for the two classes. A decision can be made as to which class the test data sets belong, using clustering techniques. The first step in achieving this goal is to perform an Eigen value decomposition on the auto-covariance matrix computed from the training data

$$VDV^H = R , \tag{10}$$

where V denotes the unitary Eigenvector matrix, D is a diagonal Eigen value matrix, H denotes the Hermitian transpose and R is the covariance matrix defined as,.

$$R = UU^H . \tag{11}$$

The U matrices are concatenated feature vectors of the training data as column vectors.

$$U = \begin{bmatrix} & & \\ u_1 & u_2 & \cdots & u_n \\ & & \end{bmatrix} . \tag{12}$$

We consider the feature vectors used to make up the matrix U to be equal up to small perturbations. The vector which represents the "average" vector of each class we call the common vector.

The task is to extract the common feature vector set from the training data. This common feature vector is extracted by taking the unity column vector in the matrix V which relates to the maximum value in the matrix D. We denote the common vector which belongs to the class of vectors u_1 as V_1 and similarly for the class of vectors u_2 we have a common vector denoted by V_2. The common feature vector for the two different classes are shown in the Fig. (4). As it can be seen, the feature vectors are different for the two different classes of speech.

We now define the projections. From these two feature vectors we define a projection vector which is defined as,

$$P = V_1 - V_2 . \tag{13}$$

The corresponding projection operator is then defined as the inner product of the observation vector and the projection vector as,

$$P[u_i] = \langle P|u_i\rangle . \tag{14}$$

This projection operator can be redefined as the common feature vector and the perturbation ξ_i, which is the difference between the common feature vector and observation vector. In this context, the projection operator is

$$P[u_i] = \langle V_1|V_i\rangle - \langle V_2|V_i - \xi_i\rangle . \tag{15}$$

As it can be seen from Eqn (15), the projection of the non-spontaneous speech set of data is projected in one direction and the spontaneous speech set is projected in the opposite direction. The result of the projection of the two sets is shown in Fig. (5).

3.3 Modeling using SAS Densities

The speech envelope is obviously non-symmetric. However, in creating the envelop, the positive choice of the sign is somewhat arbitrary. In addition, speech is somewhat impulsive. Therefore, it is appropriate to model speech features using Symmetric Alpha Stable (SAS) densities. The SAS distributions are defined as,

$$p(x) = \exp(|\gamma x|^{\alpha}) \tag{16}$$

where the parameter $\alpha \in (0, 2]$ defines the density function of the process. When the value of $\alpha = 1$ the density function corresponds to a Cauchy distribution and when $\alpha = 2$ the density corresponds to a Gaussian density. The alpha parameter is the measurement of the impulsiveness of the signal and is invariant to scale.

To create a pseudo-symmetric distribution from unipolar data, we simply invert every other sample of the speech envelope. The distribution of the resulting sequence is symmetric and may be modeled as SAS. Alpha is estimated using the sample fractile method presented by MuCulloch[4], which is an improvement of Fama's method[2]. The result plotted in Fig. (7).

4.0 Conclusion.

In this paper we have discussed three different approaches for distinguishing between non-spontaneous and spontaneous speech. These approaches were the moment method, the modified Hellinger's method and the parameter estimation using SaS densities. For each approach the corresponding error plots were given and it was found that the minimum error in all cases was around 10 percent.

REFERECES.

[1] Donoho,D.,"Large-Sample Modulation Classification using Hellinger Representation", Center for Research on Applied Signal Processing(CRASP) Research review, 1996.

[2] Fama,E. and Roll,R.,"Parameter Estimates for Stable Distributions",Journal of the American Statistical Association, Vol. 66, Pages 331-338, 1971.

[3] Keith, Selkirk, Longman,Mathematics Handbook, the language and concepts of mathematics explained, New York Press, 1992.

[4] McCulloch,J.H.,Simple Consistent Estimators of Stable Distribution Parameters",Communications in Statistics and Simulation,Vol.15,Num.4,Pp 1109-1136, 1986.

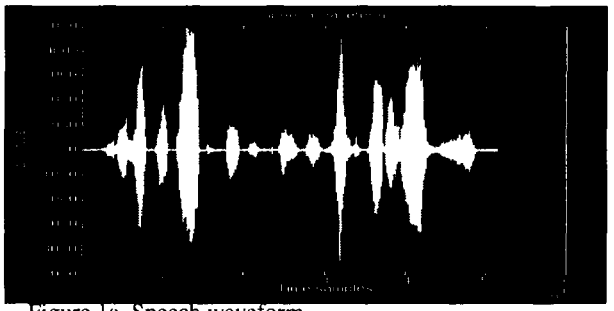

Figure 1a. Speech waveform

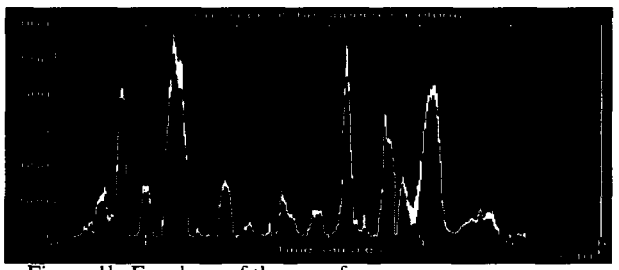

Figure 1b. Envelope of the waveform

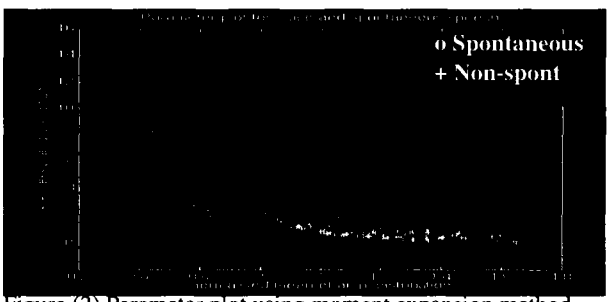

Figure (2) Parameter plot using moment expansion method

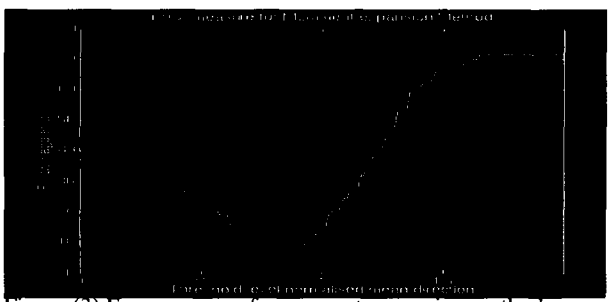

Figure (3) Error measure for moment expansion method.

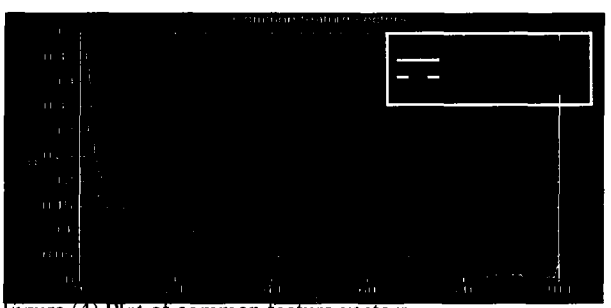

Figure (4) Plot of common feature vectors.

Figure (5) Projections using modified Hellinger's Method

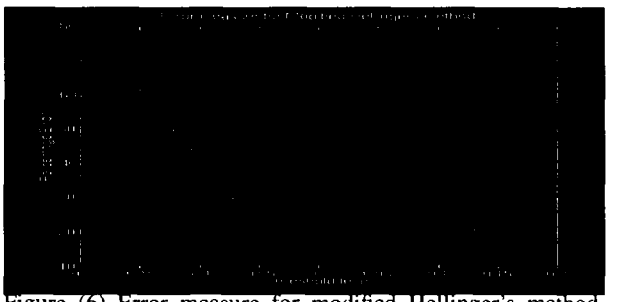

Figure (6) Error measure for modified Hellinger's method, showing the minimum error to be around 10 percent.

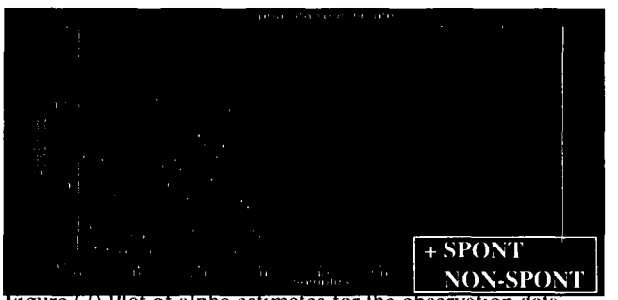

Figure (7) Plot of alpha estimates for the observation data. The corresponding error plot is shown in Fig. (8)

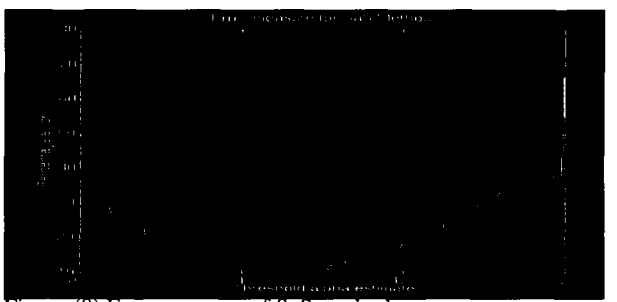

Figure (8) Error measure of SaS method.