

**FUSION – THE NEXT STEP IN
USABLE SPEECH DETECTION**

**Robert E. Yantorno
Speech Processing Lab
Electrical & Computer Engineering
College of Engineering
12th & Norris Streets
Philadelphia, Pa 19122-6077**

**Final Report for:
Summer Research Faculty Program**

**Sponsored by:
Research Laboratory AFRL/IF**

and

**Speech Processing Lab
Rome Labs
Rome, New York**

August 2001

ABSTRACT

Processing of co-channel speech has been a challenge to the speech processing research community for over three decades with limited success. Recently, a novel method to process co-channel speech has been proposed. Instead of enhancing the target speech, or suppressing the interfering speech, or both enhancing the target and suppressing the interferer, the proposed new method searches for “usable” speech frames to be extracted from the co-channel signal. There are presently two usable speech detection measures, Spectral Autocorrelation Peak Valley Ratio (SAPVR) and Adjacent Pitch Period Comparison (APPC). Both of these measures shows promise for being a usable speech detection measure, however, each of the measures also has a very high false alarm rate, i.e., above 20%. Therefore, to increase the effectiveness of a usable speech detection system it will be necessary to: 1.) have more than one measure (which we presently have), 2.) fuse the measures together, and 3.) ensure that the measures have complementary information.

An investigation of the complementary information, using a “quasi-correlation” method, indicates that both measures have similar types of information. It was also determined that for both measures, many of the false alarms occur in transition regions. A study of the probability density function of the SAPVR and APPC reveals that for both measures there is not a clear separation of the density functions of hits and false alarms. For example, for the SAPVR the probability density function is almost the same for both hits and false alarms, whereas for the APPC the hits and false alarms have different types of distributions, and there is some separation of the density functions of the hits and false alarms. Linear least squares fit of the semi-logarithmic probability density data have allowed us to define the probability density function with a closed form mathematical expression. This is the first step in the development of an effective fusion system. An initial qualitative study of fusion without weighting of either the Linear Opinion Pool or the Logarithmic Opinion Pool indicates that one might be able to obtain a better combined measure if only certain parts of the measure were used, in this case to use measures that fall within a certain probability range. For example, for the Linear Opinion Pool consensus (sum) rule ($C_{\text{Lin-OP}}$) the range of probability of 0.05 to 0.12 would provide good data, i.e., ratio of hits to false alarms would be larger than without fusion. For the Logarithmic Opinion Pool consensus (product) rule ($C_{\text{Log-OP}}$) the range of probability of 0 to 0.002 would provide good data.

It has been shown that Target-to-Interferer Ratio (TIR) is a good measure of the usability of speech. Therefore, it is important to determine how the usable speech measures are correlated with Target-to-Interferer Ratio (TIR). In the ideal case there should be a one-to-one correlation between the measure and the TIR. A plot of TIR versus SAPVR and TIR versus APPC indicates that there is some correlation between the measures and TIR, however, the correlation is not strong, i.e., not close to or equal to 1.

Finally, a study was performed to determine the statistical characteristics of the usable speech segments, i.e., what is the distribution of number of frames per segment. The results indicate that the probability of there being 3 or 4 frames per segment is much higher than one would expect if usable and unusable frames were distributed in a purely random fashion.

I. INTRODUCTION

Co-channel and Usable Speech

Co-channel speech processing has been a challenge to the speech processing community since the early 1970s. The usual approach is to attempt to extract the target speech from the co-channel speech. This has been attempted by either enhancing the target speech, suppressing the interferer speech, or a combination of both enhancing target and suppressing interferer. Recently it has been determined that there are portions (segments) of co-channel speech which can be considered as usable, i.e., usable as related to how the co-channel speech is to be processed. For example, if one is performing speaker identification or verification then certain segments of the co-channel speech will be either only target speaker speech or will be target speaker speech minimally degraded by the interferer's speech. The latter can happen if, for example, the target speech is voiced and the interferer speech is unvoiced. Therefore, if one can find a method for measuring the usability of co-channel speech then one has a tool which will allow one to deal with co-channel speech in a very different way than was done in the past. There are some ways in which a usability measure might be used, for example, for extracting only usable portions of speech for speaker identification or verification. Figure 1 shows the front end of a next generation speech processing system.

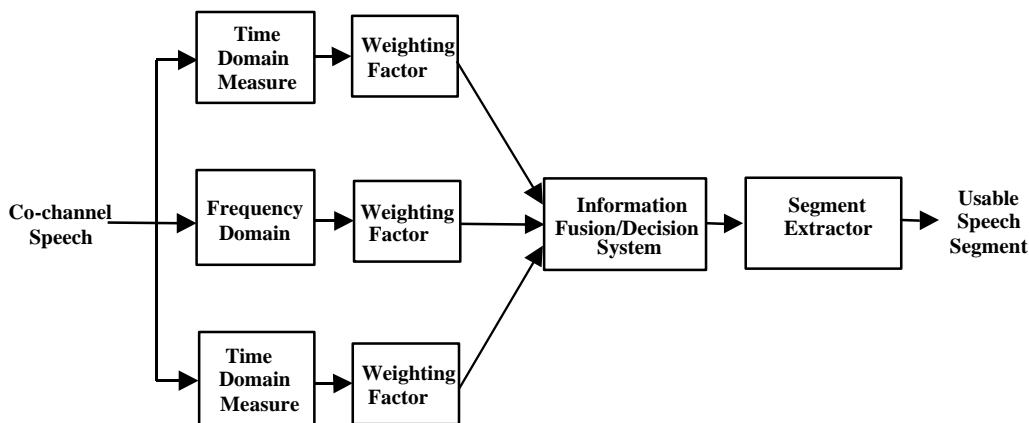


Figure1. Usable Speech Segment Detection and Extraction Sub-unit.

Only usable segments would be passed to a speaker identification system, thereby increasing the accuracy of such a system by eliminating those portions of the speech signal which would degrade its operation. It has also been proposed that the system shown in Figure 1 could also be used as a co-channel detection system (Yantorno *et al*, 2001). For this situation one could use measures and fusion/detection system as a co-channel detection system whose output could signal a speech processing system to halt its operation, until there was single speaker speech available. This operation is shown in Figure 2 below.

Once the usable segments are identified and extracted from the co-channel speech then they can be used to reconstruct the utterances of the target and/or interferer. However, the next step in the process is to identify and separate the target and interferer segments. This was the topic of research for one of the researchers from the Speech Lab at Temple University who worked on

this project here at Rome Labs this summer (Smolenski, 2001). This operation is shown in Figure 3 below. Note that speaker identification is being used in a unique way, i.e., to identify segments belonging to a specific speaker rather than identifying which speaker belongs to specific segments.

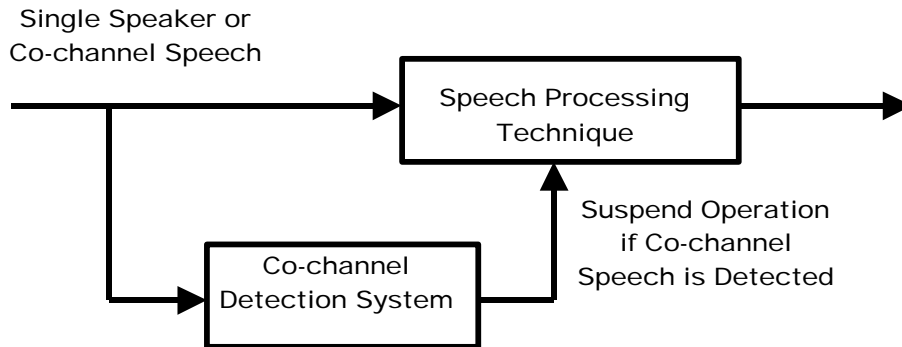


Figure 2. Co-channel Detection System

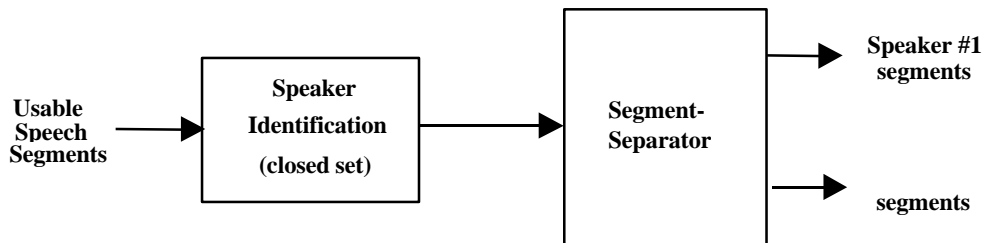


Figure 3. Segment Separator Sub-unit.

Once the segments have been separated then it is possible to reconstruct the utterances of each of the speakers, i.e., the target and the interferer. The process of reconstructing the utterances is shown in Figure 4 below. It should be noted that the sub-unit shown in Figure 4 is a suggested approach; other methods besides pitch and format extraction, and the use of abutting frames, might be used to reconstruct the utterances.

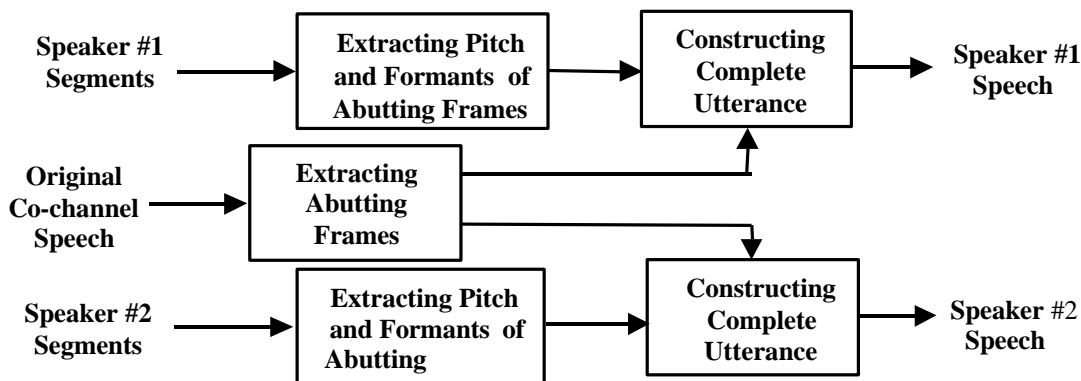


Figure 4. Target and Interferer Utterances Reconstruction Sub-unit

Therefore it is evident that possessing methods for measuring usability of co-channel speech could be very beneficial for the co-channel speech processing community. The challenge has been to find good usable speech measures. However, one would expect that a single usable speech measure would provide marginal performance due to the fact that speech is nonstationary and therefore one is confronted with many different types of speech even for a single speaker, i.e., different sounds, different energies, transition regions and even sounds which have no steady-state portion, such as glides. Due to the nonstationary nature of speech it is necessary to develop a number of usable measures, and to then fuse the measures together to provide as effective and reliable measure as possible. To date there are presently three usable measures, Spectral Autocorrelation Peak Valley Ratio (SAPRV) (Krishnamachari *et al*, 2000), kurtosis (Krishnamachari *et al*, 2001) and Adjacent Pitch Period Comparison (APPC) (Lovekin *et al*, 2001). It should be noted that the kurtosis measure is not used in the same manner as the SAPRV or APPC, i.e., identifying frames as being usable or unusable. Rather, the kurtosis is used to mark the beginning and end of usable segments.

Fusion Approaches

Given that there are two usability measures, the question then is how to use the measures to provide an effective and reliable usable speech measure. To help solve this problem one must look at the field of consensus theory, where the objective is to find consensus amongst a group of experts, i.e., make a decision based on the opinion of experts, in this case measures. Consensus theory was first applied to statistics and managerial science in the early and middle 1980s. It was then expended into the field of pattern recognition as related to multisource data sets. Related to consensus theory is statistical multisource analysis, which is a probabilistic method based on Bayesian decision theory and includes mechanisms to weigh the influence of data sources in classification. Kittler *et al* (1998) have developed a general theoretical framework for combining classifiers, i.e., two basic classifier rules, product rule and sum rule. Using these two rules the authors then show that these two rules can allow for different classifier combination strategies, such as max. rule, min. rule, median rule and majority vote. Some of these approaches represent the traditional approaches of Linear Opinion Pool (Lin-OP), Logarithmic Opinion Pool (Log-OP), and voting.

Linear Opinion Pool (Lin-OP)

The linear opinion pool is probably the most commonly used consensus rule, and the Lin-OP consensus rule (C_{Lin-OP}) can be defined mathematically in two different ways (Equations. 1 and 2) as shown below. However, Equation 1 provides more insight into how the linear opinion pool works. That is, the consensus rule is a weighted sum of the probability density functions of the various experts (in our case measures). The consensus rule is -

$$C_{Lin-OP}(p_1, p_2, \dots, p_N)(X) = \sum_{i=1}^N w_i p(x_i) \quad (1)$$

where: α_i are the weights and $p(x_i)$ are the probability density (distribution) functions of the various measures. The second formulation is -

$$C_{Lin-OP}(X = k) = \sum_{i=1}^N w_i P(X = k | D_i = j_i) \quad (2)$$

where: $P(X = k | D_i = j_i)$ is the *a posteriori* probability that the tested frame belongs to class i when the decision of the m th classifier is j_i .

Logarithmic Opinion Pool (Log-OP)

The consensus rule ($C_{\text{Log-OP}}$) of Logarithmic Opinion Pool is defined as:

$$C_{\text{Log-OP}}(X = k) = \prod_{i=1}^N P(X = k | D_i = j_i)^{w_i} \quad (3)$$

Taking the log of the products results in the following equation -

$$C_{\text{Log-OP}}(X = k) = \sum_{i=1}^N w_i \log P(X = k | D_i = j_i) \quad (4)$$

Accordingly, the linear opinion pool method can more generally be defined as a sum rule and the logarithmic opinion pool consensus rule can more generally defined as the product rule (Kittler *et al*, 1998).

Choice of Weights

The choice of weights is an important part of the consensus approach, and is usually dealt with more for Lin-OP than Log-OP, simply because any weighting method is less intuitive for the Log-OP because of the product form of the rule. However, one can use heuristics or ad hoc methods for Log-OP as well as methods outlined below for Lin-OP.

Benediktsson and Swain (1992) identify four different methods of weighting for the Linear Opinion Pool: 1.) equal weights, in this case the result of equal weighting is to take the average of the probability density functions, 2.) weights proportional to ranking, i.e., rank sources according to goodness with the best at the top, 3.) weights according to self-ranking, i.e., weights are assigned according to self-rating which means that a source's weight may vary according to the type of data being analyzed, and 4.) weights based on some comparison of previously assessed distributions with actual outcomes.

Altincay and Demirekler (2000) have used combination schemes developed by Bloch (1996), which is two groups of operators, i.e., context dependent and context independent (which do not consider the individual performance of classifiers in combination). The context dependent operations are further decomposed into two groups; the first group is combination schemes that take into account conflict among classifiers and the second group takes into account class dependent reliability as well as global classifier reliability. The authors also provide a very nice association of the relationship between classification theory and information theory and the parallel between the channel matrix (information theory) and the confusion matrix (classification theory). Besides developing a very interesting weighting scheme, they also have developed a measure called complementariness which provides information about the complementary information provided by different classifiers. Their weighting scheme is estimated in a dynamic manner with the final weight being a multiplication of three different weights, i.e., decision dependent classifier weight, global classifier reliability, and a weighting based on the amount of conflict between classifiers.

Due to the limited time available for this research project only the Lin-OP and Log-OP methods without weight will be investigated.

Applications of Consensus Theory and Fusion to Speech

More recently, consensus theory has been applied to areas of speech processing such as speaker identification (Farrell and Mammone, 1995, Altincay and Demirekler, 2000) and speaker

verification (Farrell 1995 and Farrell *et al*, 1998) as well as speech recognition (Rogozan and Deleglise, 1998).

II. MUTUALLY EXCLUSIVE INFORMATION EXPERIMENTS

In order to make a fusion system as effective as possible one of the requirements would be that each of the measures to be fused would contain independent information, e.g., to quote Altincay and Demirekler (2000), “ The classification performance of a combined system can be better than the performance of the individual classifiers only if there exists some complementary information coming from individual classifiers. . . . if the combination is not done carefully . . . the classification performance of the combined system may even be worse than the performance of the individual classifiers.”

For some systems, determination of complementarity or mutual exclusiveness is usually straightforward. However, for the present measures comparisons of results to determine mutually exclusive information would only be possible if each measure somehow detected frames of usable data that were not detected by the other measures. However, there is other information besides detection information that may provide some insight into how mutually exclusive the information is from each measure, and that other information source would be false alarms.

A series of experiments were performed to determine the exclusivity of the information of each measure. A comparison was made of SAPVR with APPC, SAPVR with Kurtosis and APPC with Kurtosis. The data was obtained in the following way, for the SAPVR hits, a sequence was generated where a one was placed in the location where a detection of a usable speech frame was made, otherwise a zero was placed there. The resulting sequence then was a series of ones and zeros. The same thing was done for APPC and kurtosis, i.e., a sequence of ones and zeros is generated for the same speech utterance as was used for the SAPVR. Therefore, if one multiplies (in this case logically ands) the SAPVR hits sequence with the APPC hits sequence, a one will appear only at the sequence location where there is a one for both the SAPVR as well as the APPC, otherwise the results will be zero. This means that it is possible to develop a “quasi”-correlation-like process. For example, if one obtains the total number of ones in the resultant of the multiplication and divides this number by number of ones from the sequence with the largest number of ones, the results will be a number which will vary from one to zero indicating in a similar way correlation as is done using the traditional correlation coefficient, i.e., 1 being perfect correlation and zero being no correlation whatsoever. Note, sequences of ones and zeros for hits, misses and false alarms were generated and manipulated for the APPC and kurtosis measures also

The exclusivity of the measures can be more easily observed with the following sets of plots. Figure 5 (below) shows the result of analyzing the correlation of the SAPVR and kurtosis. It should be noted that the Kurtosis is being used in a manner in which it was not initially designed to be used (Krishnamarchari *et al*, 2001). The kurtosis was initially developed to be used to identify usable segments in the region close to where the kurtosis measure is above a threshold. However, in this case we are using the kurtosis measure for each frame as a detection system for usable speech. Because we are using the kurtosis in such way we expect the kurtosis measure to perform more like a detection system which would operate in a random-like fashion, i.e., if there was twice as much usable speech as there was unusable speech then one would expect the

number of hits, misses and false alarms to be about the same. Because the kurtosis measure is randomly distributed, one would expect the result of the quasi-correlation of the SAPVR and the kurtosis to produce about half the number of hits, misses and false alarms as there are for the SAPVR. This would only be true for a large sample, which is not the case here. However, there are appreciably fewer hits, misses and false alarms for the quasi-correlation as shown in the bottom panel of Figure 5 (below), this is clearly an indication of a lack of correlation between the SAPVR measure and the kurtosis measure, which was expected.

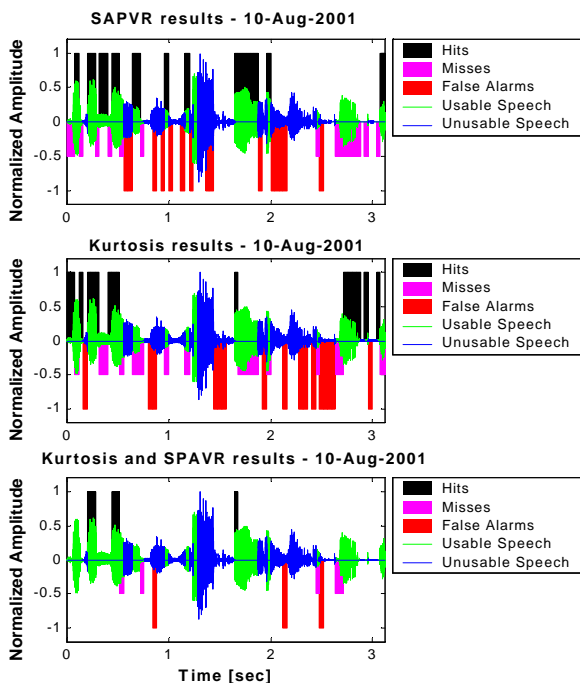


Figure 5. Correlation Analysis – Hits, Misses and False Alarms. SAPVR measure (upper panel), Kurtosis measure (middle panel) and correlation of SAPVR and Kurtosis (lower panel).

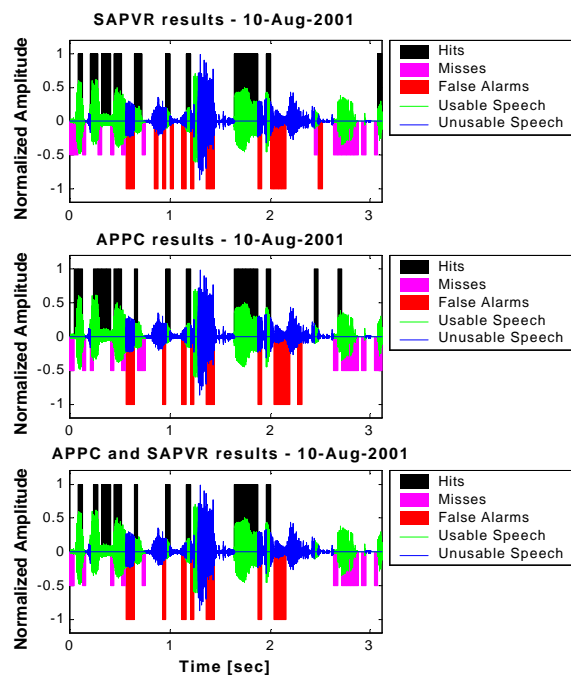


Figure 6. Correlation Analysis – Hits, Misses and False Alarms. SAPVR measure (upper panel), APPC measure (middle panel) and correlation of SAPVR and APPC (lower panel).

Another series of experiments were performed for the SAPVR and APPC measures and the results are shown in Figure 6 (above). There are two important observations that can be made concerning the result of the quasi-correlation of the SAPVR and the APPC as shown in the bottom panel of Figure 6 (above), and those are: 1.) that there is a high degree of correlation between the two measures as evidenced from the large number of hits, misses and false alarms, and 2.) that a large number of false alarms occur in transition regions.

III. STATISTICAL CHARACTERIZATION OF USABLE MEASURES

In order to more effectively determine how to fuse the various measures one must obtain some information about each of the measures, such as their probability distribution with respect to hits and false alarms, their reliability, etc. This has been done for the SAPVR, APPC, and kurtosis, and the results are shown and discussed below.

Experimental Conditions – Speech Data

Forty-one TIMIT database utterances were used, 21 male utterances and 20 female utterances, each about 3 to 4 seconds long. Files were filtered using a FIR filter with a corner frequency of 3.3 kHz and then down-sampled from 16 KHz to 8 kHz prior to analysis.

Results

One very important aspect of fusion of a measure is to have some idea about the distribution of hits and false alarms. In an ideal situation one would hope that the distribution of these two parameters would be very different, and more importantly, would not overlap at all in their distribution. However, this is not the case with the three measures investigated here, i.e., the SAPVR, the APPC and the kurtosis.

SAPVR

The combined distribution of both hits and false alarms for the SAPVR as can be seen, in Figure 7 below.

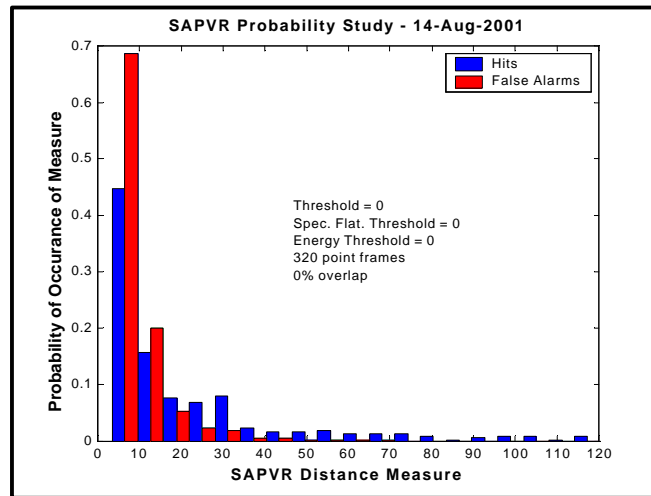


Figure 7. Probability of hits and false alarms for statistics of SAPVR from 41 speech files.

It was determined previously that an effective threshold of the SAPVR was 10. However, from Figure 7 a more appropriate threshold would be 25. It should also be noted that as the measure increases in value there are fewer false alarms. The usual approach for forming a consensus of the measures is to weigh the probability of each measure, and then to sum up those values, as outlined in Equation 2 above. However as the SAPVR measure increases the probability decreases. Therefore, because there are fewer false alarms at higher SAPVR, it would seem to be more reasonable to use a reliability number rather than a probability number, where the reliability increases as the value of the measure increases instead of decreasing, which it would do if the probability were used. The distribution of hits as well as false alarms appears to be exponential in nature. A semilog plot of the data was generated, as shown in Figures 8a (hits probability) and 8b (false alarms probability).

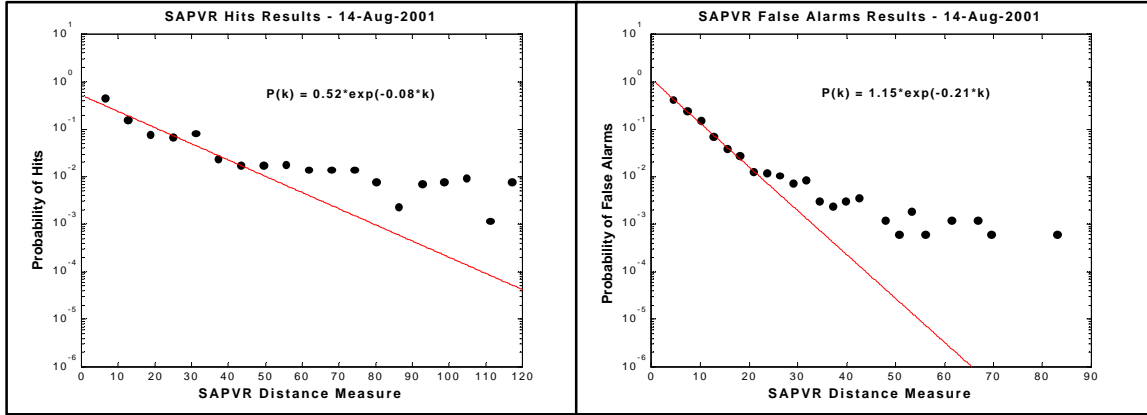


Figure 8a.

Figure 8b.

Figure 8. Semilog plot of data from Figure 7 above. Least squares fit of the data was performed to obtain the equation show on the plot. Figure 8a SAPVR hits and Figure 8b SAPVR false alarms.

A linear least squares fit of the semilog data of Figures 8a and 8b above was performed and the equations obtained were –

$$P_{SAPVR}(hits) = 0.52e^{-0.08 * SAPVR} \quad P_{SAPVR}(false) = 1.15e^{-0.21 * SAPVR} \quad (5)$$

The experiments used to obtain information about the SAPVR measure were also conducted for the APPC and kurtosis measure, and the results are shown in Figures 9 through 12.

APPC

The results of the APPC experiments are shown in Figure 9 below for the hits and false alarms probability density (distribution) of the APPC measure.

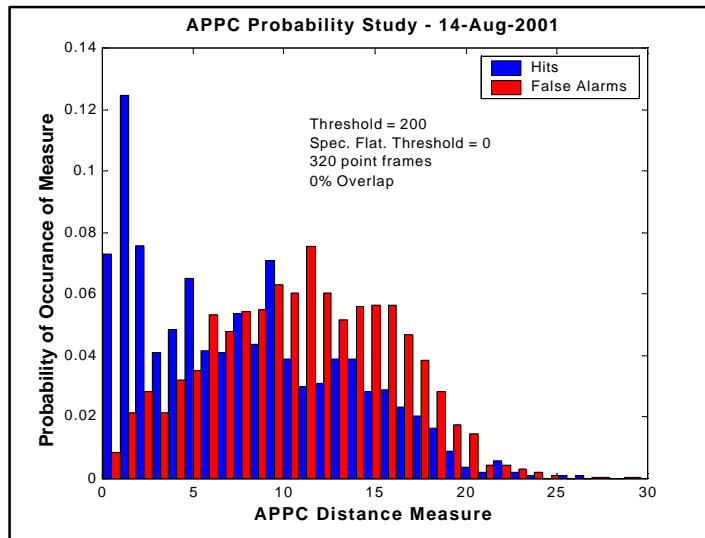


Figure 9. Probability of hits and false alarms for statistics of APPC from 41 speech files.

It is evident that the distribution of hits and false alarms is different. From Figure 9 it is also evident that the optimal value for the threshold would be 5 which is not that different from what was determined as an optimal threshold of 6 (Lovekin, *et al*, 2001). A semilog plot of the data of Figure 9 was performed and is shown in Figures 10a (hits probability) and 10b (false alarms probability) of the APPC measure,

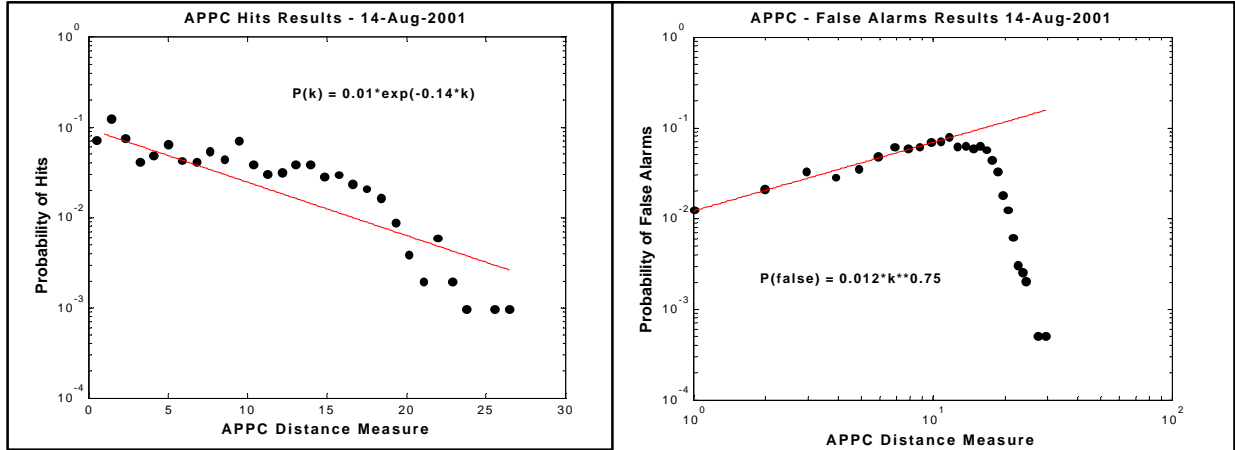


Figure 10a.

Figure 10b.

Figure 10. Semilog plot (Fig. 10a) and log-log plot (Fig. 10b) of data from Figure 9 above. Least squares fit of the data was performed to obtain the equation shown on the plot. Figure 10a APPC hits and Figure 10b APPC false alarms.

The equations for the probability of APPC hits and false alarms is –

$$P_{APPC}(hits) = 0.01e^{-0.14*APPC} \quad P_{APPC}(false) = 0.012 * APPC^{0.75} \quad (6)$$

Note, because the distribution for the false alarms is not exponential, a log-log plot was constructed, a linear least squares fit of the data was performed and results are shown above in the right-hand equation. Note that the data of the APPC distance measure of Figure 10b is over the range 0 to 10 and is well within the range used for hits detection, i.e., 0 to 6.

KURTOSIS

The results of the kurtosis experiments are shown in Figures 11 and 12 below.

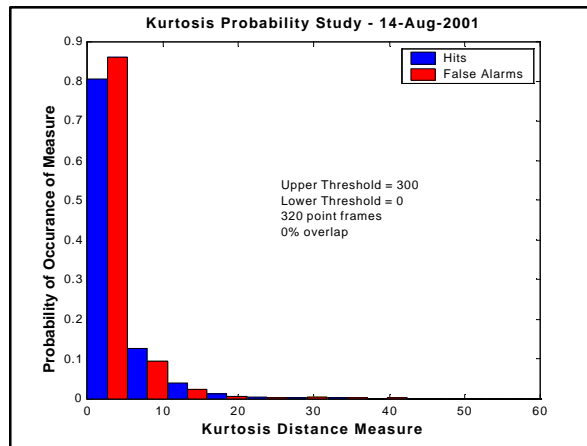


Figure 11. Probability of hits and false alarms for statistics of kurtosis from 41 speech files.

As noted earlier, the kurtosis is not being used in the manner for which it was initially developed, i.e., as a marker for the beginning or ending of usable segments. Therefore, one should not expect that the kurtosis would provide a separation of hits and false alarms but rather that both the hits and false alarms should have the same type of distribution and both hits and false alarms should track each other, which is the case as seen in Figure 11 above. Semilog plot of the data from Figure 11 is shown in Figures 12a (hits probability) and 12b (false alarms probability) of the kurtosis measure.

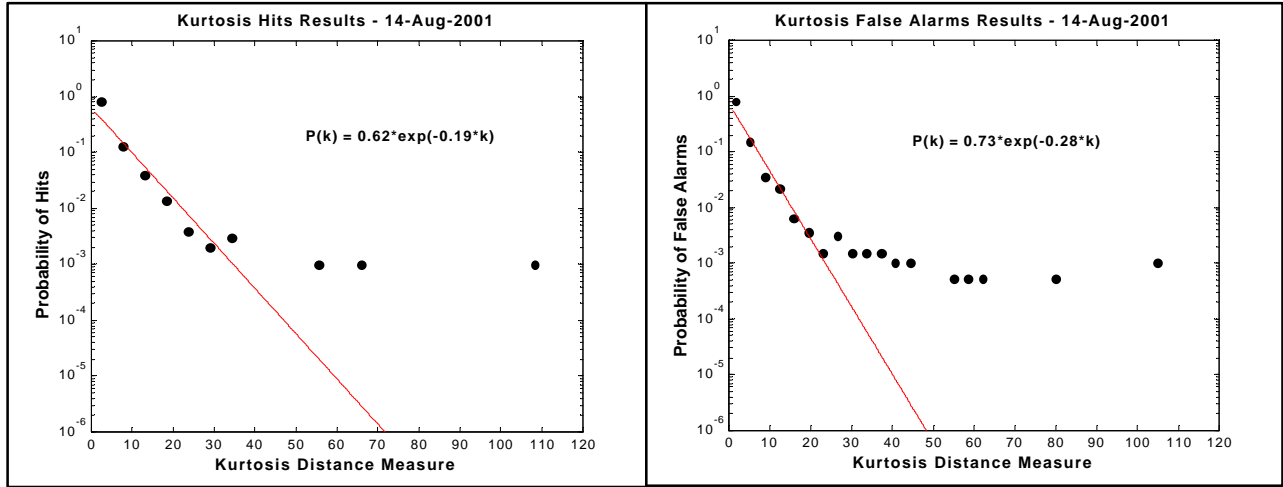


Figure 12a.

Figure 12b.

Figure 12. Semilog plot of data from Figure 11 above. Least squares fit of the data was performed to obtain the equation shown on the plot. Figure 12a kurtosis hits and Figure 12b kurtosis false alarms.

The Equations relating the kurtosis hits and false alarms are:

$$P_{kurtosis}(hits) = 0.62e^{-0.19 * kurtosis} \quad P_{kurtosis}(false) = 0.73e^{-0.28 * kurtosis} \quad (7)$$

IV. FUSION RESULTS

The goal of this summer's research was to investigate using fusion as a way to enhance the detection of usable speech. Earlier evidence of the probability density information for one of the major measures, i.e., SAPVR, indicates that separation of false alarms from hits is not possible, which presents a problem. It has also been observed that the SAPVR and APPC are very similar types of measures as indicated by experiments performed to determine how much mutually exclusive information there was in each measure. Therefore, one would expect to be as successful as one might wish to be using fusion. However, it will be helpful to perform some preliminary fusion experiments in order to gain knowledge, which will be useful when more measures are developed.

Experimental Conditions

The same set of data and experimental conditions used for the experiments outlined below are the same as those used for Section III – Statistical Characterization of Usable Measures outlined above.

Results

Only the Linear Opinion Pool (Lin-OP) and the Logarithmic Opinion Pool (Log-OP) were tested. Also, due to time constraints it was not possible to determine any effective method of weighting for either method.

Because the kurtosis measure was not being used in the manner for which it was designed to be used only the SAPVR, APPC, Lin-OP and Log-OP were investigated, and the results are shown in Figures 13a (sum approach - Lin-OP) and Figure 13b (product approach - Log-OP) below.

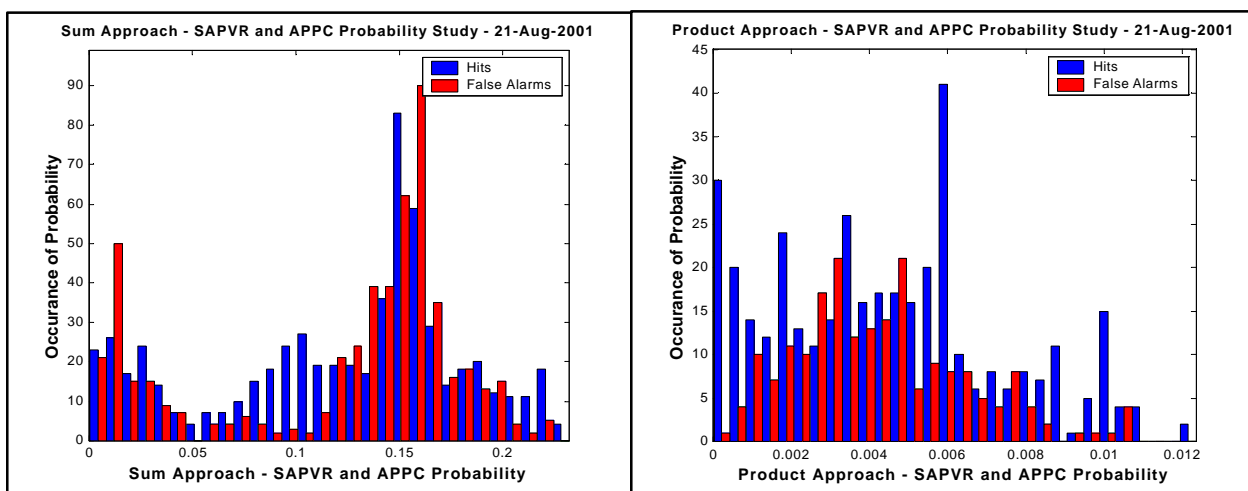


Figure 13a.

Figure 13b.

Figure 13. Histogram of the probability of hits and false alarms for statistics of sum approach (Figure 13a) and product approach (Figure 13b).

Fusion without weighting of either the Linear Opinion Pool or the Logarithmic Opinion Pool indicates that one might be able to obtain a better combined measure if only certain parts of the measure were used, in this case if measures that fall within a certain probability were used. For example, for the Linear Opinion Pool consensus (sum) rule (C_{Lin-OP}) it would appear that the range of probability of 0.05 to 0.12 would provide good data, i.e., ratio of hits to false alarms would be larger than without fusion. For the Logarithmic Opinion pool consensus (product) rule (C_{Log-OP}) the range of probability of 0 to 0.002 would provide good data.

V. CHARACTERIZATION OF MEASURES AND USABLE SEGMENTS

Although not directly related to fusion, the following sets of experiments were conducted to better understand the measures presently available, as well as to develop criteria to be used to evaluate future measures. To better understand each of the measures in terms of their weaknesses and strengths a series of experiments were performed, to see how each of the measure correlates with TIR. Also, a series of experiments were performed to determine the distribution of usable segments. It is theorized that the usable segments should not be randomly distributed. Although the choice of whether a segment is usable or unusable is similar to the choice of a fair coin toss, i.e., equally probable, the relationship between frames is not statistically independent, i.e., a frame following a usable frame may be more likely to be usable than unusable, and therefore this could result in a higher probability of multi-frame segments.

Experimental Conditions

The set of data and experimental conditions used for the experiments outlined below are the same as those used for Section III – Statistical Characterization of Usable Measures outlined above.

Results

SAPVR

A plot of TIR versus SAPVR distance measure was constructed and is shown in Figure 14 below.

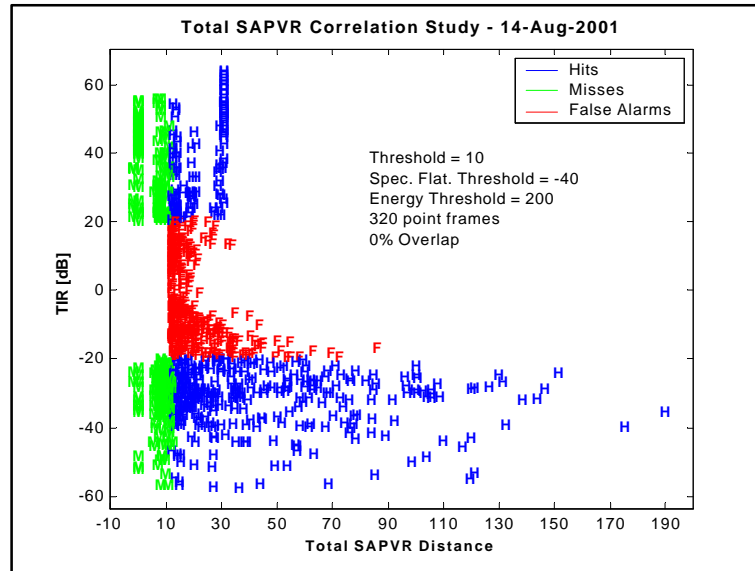


Figure 14. Target-to-Interfere Ratio (TIR) versus SAPVR distance measure for 41 speech files.

It should be noted that when performing the experiments the same male speech file was used as the target speech and then the interferer speech was one of 40 other male and female speech files. Because the target speech was not changed for 40 experiments the TIR versus SAPVR distance measure for values of TIR above 20 dB is very structured, as can be observed in Figure 14 as well as Figures 15 and 16 below.

APPC

A plot of TIR versus APPC distance measure was constructed and is shown in Figure 15 below. It should be noted the APPC distance measure is calculated using variable frame length, i.e., frames whose length are equivalent to the pitch period. Because the SAPVR measure uses a fixed frame length it was necessary to construct an APPC distance measure that was also based on a fixed frame length so that we could manipulate all of the measures together. Therefore, it was necessary to change the variable frame length results into fixed frame length results. This can be accomplished two different ways. The first is to simply take the average of the distance measure for 320 points (the frame length used in the experiments with the fixed frame measures), and if the average distance is below the threshold, then that frame would be considered as usable.

The other approach is to see if at least half the points of the fixed frame are below the threshold then the frame is considered as usable. The latter approach yields results much closer to the results obtained using the variable frame approach versus those obtained using the average value

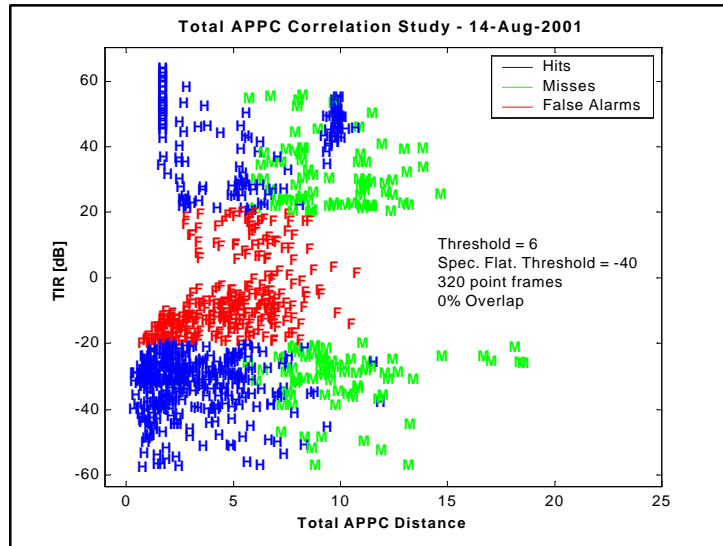


Figure 15. Target-to-Interfere Ratio (TIR) versus APPC distance measure for 41 speech files.

of the APPC distance measure.

KURTOSIS

A plot of TIR versus kurtosis distance measure was constructed and is shown in Figure 16 below.

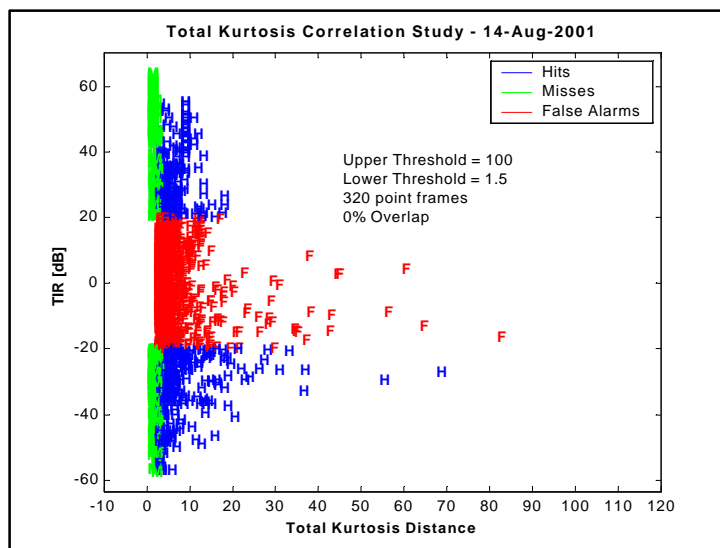


Figure 16. Target-to-Interfere Ratio (TIR) versus kurtosis distance measure for 41 speech files.

The plots of TIR versus SAPVR (Figure 14) and TIR versus APPC (Figure 15) indicates that there is some correlation between the measures and TIR, however, the correlation is not strong, i.e., close to or equal to 1. We would not expect any correlation between the kurtosis measure and TIR and this is evident in Figure 16.

Usable Segment Statistics Study

When performing a decision on whether a frame being analyzed is either usable or unusable, the decision is binary and the choice of the frame being usable or unusable is equally probable. Therefore, one might expect that one could obtain at least 50% correct decisions under those conditions. However, when dealing with segments one must consider many other possibilities besides single frame decisions, i.e., what is the probability of two usable frames being adjacent to each other, or three frames, etc. One can easily imagine that the probabilities of those sets of conditions would be less than 50%, and that the probability would decrease with increasing number of adjacent frames. To test this hypothesis two different conditions were investigated: 1.) a purely random set of conditions of the occurrence of different length frames, and 2.) the situation of observing different length segments of co-channel speech.

Experimental Conditions – Random Sequence

A 10,000 point sequence was generated in Matlab using the gaussian random function. If the random value generated was greater than zero a 0 was assigned to that point, if the value of the random number generated was less than or equal to zero a 1 was assigned. Therefore, a zero represents a usable frame and a one represents an unusable frame.

Results

A 10,000 point sequence was generated, and the data was analyzed for various numbers of frames per segments; the results are shown in Figures 17a and 17b and 18.

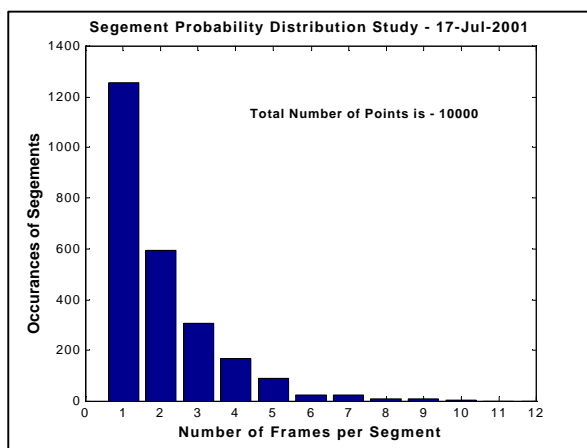


Figure 17a.

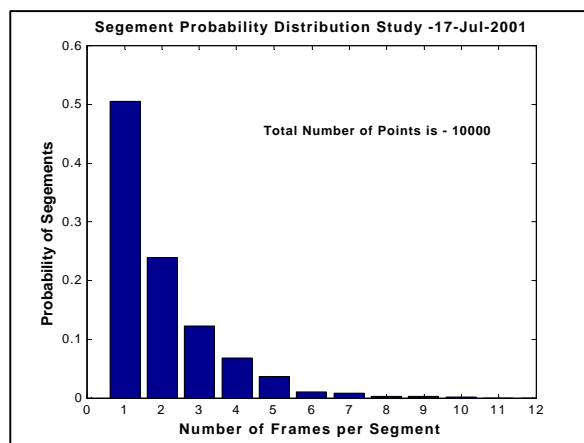


Figure 17b.

Figure 17. Segmental probability study where occurrences of frames is random with equal probability of a frame being either usable or unusable. Figure 17a is histogram of data for a 10,000 point sequence, where each point is considered a frame. Figure 17b is the probability distribution of the data of Figure 17a.

As can be observed with Figure 17b above, the probability of occurrence of different length segments decreases by about half for each increase in the number of frames in a segment.

Performing a linear least squares fit of the data shown in figure 18 below resulted in the following equation:

$$P(k) = 0.867e^{-0.657*k} \quad (8)$$

where: P(k) is the probability of the occurrence of k frames in a segment, and k is the number of frames per segment.

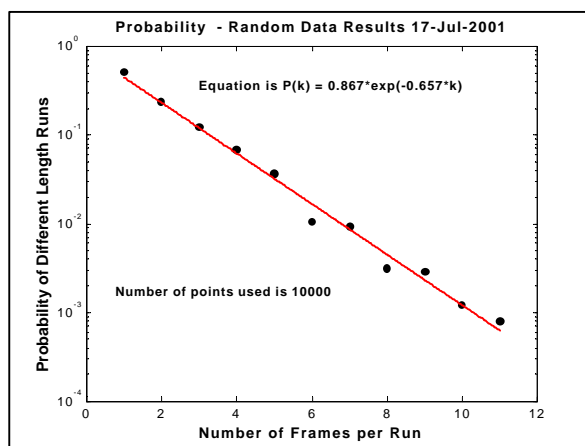


Figure 18. Semilog plot of data from Probability as shown in Figure 17b. above. Equation obtained from linear least squares fit of the data. 10,000 points was used.

The results of the segment size study of speech was interesting but not surprising. For various runs it was observed that there were an unusually large number of segments of size larger than a single frame. This is evident in Figure 19a., where it is equally probable to observe either 2 or 3 frame segments and that these are about half as probable of occurring as a single frame.

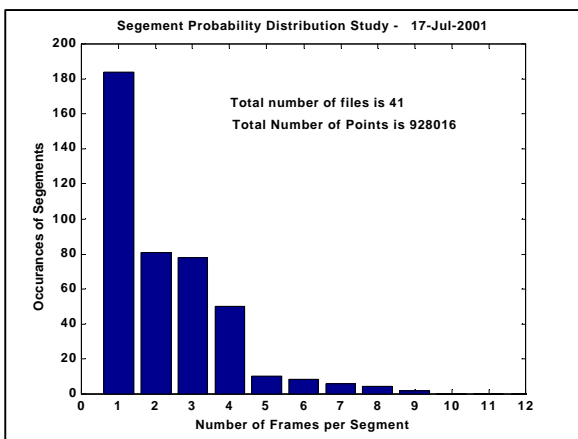


Figure 19a.

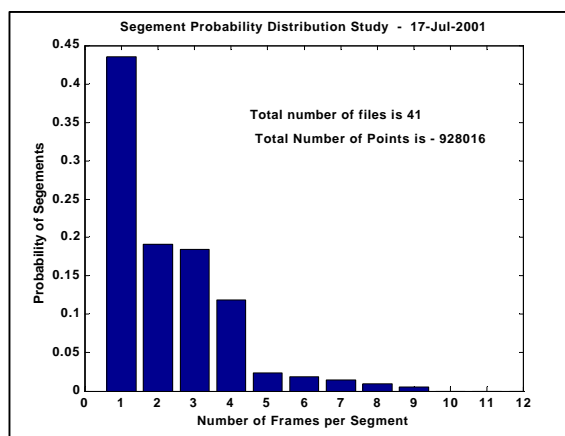


Figure 19b.

Figure 19. Segment distribution of speech using a TIR of 20 dB. Figure 19a is histogram of data from 41 files. Figure 19b is the probability distribution of segments of speech with a TIR greater than 20 dB.

A least squares fit of the data of Figure 19b above was performed and the results are shown in Figure 20 below. The equation for the probability is

$$P(k) = 0.738e^{-0.568*k} \quad (9)$$

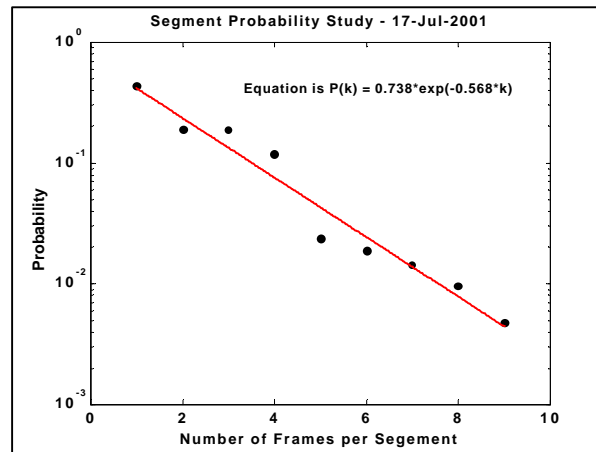


Figure 20. Semilog plot of data from Probability as shown in Figure 19b above. Equation obtained from linear least squares fit of the data of Figure 19b above.

Conclusions

As expected there is a definite difference in the occurrence of different length speech segments (Figure 19b) as compared with frames occurring in a random way (Figure 17b). A comparison of Figure 17b with Figure 19b illustrates that there is a much greater probability of segments of 3 or 4 frames occurring than one would expect if usable and unusable frames occurred in a purely random fashion (as shown in Figure 17b).

VI. CONCLUSIONS

Two areas of study were conducted, i.e., fusion of two usable speech measures and characterization of the measures and usable speech segments. The fusion approach was only marginally successful in that it has provided us with an idea about what is necessary for our measures in order to provide for an effective and efficient fusion system, i.e., that the measures must have mutually exclusive information if they are to be useful in a fusion system. Also, it is hoped that other usable speech measures will also have distributions of hits and false alarms that will not have very similar distribution characteristics as is the case with the SAPVR and will be more like the APPC. Also, that we can improve the measures by excluding transition regions or by handling them in a different way, i.e., possibly developing a usable speech measure exclusively for transition regions.

The characterization of the measures has provided us with another tool to better understand how our measures work, in this case to see how the usable speech measures correlate with the TIR. In the ideal case the correlation should be 1 but we do not expect this to happen. However, we do hope that they will show more correlation than what is observed for the SAPVR and APPC.

Finally, it will be important to investigate the usefulness of using frames adjacent to useable speech segments. Presently a hard decision is made as to whether a frame is usable or not, but frames abutting usable speech segments could very well be helpful for such things as speaker segments separation as well as speaker identification.

VII. REFERENCES

1. Altincay, H. and Demirkler, M., An Information Theoretic Framework for Weight Estimation in the Combination of Probabilistic Classifiers for Speaker Identification, *Speech Comm.*, Vol. 30, pp:255-272, 2000.
2. Benediktsson, J. A. and Swain, P. H., Consensus Theoretic Classification Methods, *IEEE Trans. Sys., Man, and Cybernetics*, Vol. 22, No. 4, pp:688-704, Jul./Aug. 1992.
3. Bloch, I., Information Combination Operators for Data Fusion: A Comparative Review with Classification. *IEEE Trans. Sys., Man and Cybernetics – Part A: Sys. And Humans*, Vol. 26, No. 1, pp:52-67, Jan. 1996.
4. Farrell, K. R. and Mammone R. J., Data Fusion Techniques for Speaker Recognition, “Modern Methods of Speech Processing”, R. Ramachandra and R.J. Mammone editors, pp:279-297, 1995.
5. Farrell, K. R., Text-Dependent Speaker Verification Using Data Fusion. *ICASSP*, pp:349-352, 1995.
6. Farrell, K. R., Ramachandra, R. P., and Mammone, R. J., An Analysis of Data Fusion Methods for Speaker Verification, *ICASSP*, pp:1129-1132, 1998.
7. Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J., “On Combining Classifiers”, *IEEE Trans. Pattern Anal. and Mach. Intell.*, Vol. 20, No. 3, pp:226-239, Mar. 1998.
8. Krishnamachari K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J., Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions. *IEEE International Symposium on Intelligent Signal Processing and Communication Systems 2000*, November. 2000.
9. Krishnamachari, K. R., Yantorno, R. E., Lovekin J. M., Benincasa, D. S., and Wenndt, S. J., “Use of Local Kurtosis Measure for Spotting Usable Speech Segments in Co-channel Speech.” *ICASSP 2001*.
10. Lovekin, J. L., Krishnamachari K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J., Adjacent Pitch Period Comparison as a Usability Measure of Speech Segments Under Co-channel conditions. *IEEE International Symposium on Intelligent Signal Processing and Communication Systems 2001* (submitted).

11. Rogoza, A., and Delegise, P., Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Comm.*, Vol. 26, pp:149-161, 1998.
12. Smolenski, B., Co-Channel Usable Speaker Segment Separation, Final Report for Graduate Student Summer Research Program, Sponsored by AFRL/IF Laboratory, Rome, NY. 2001
13. Yantorno, R. E., Krishnamachari, K. R., Lovekin, J. M., Benincasa D. S., and Wenndt, S. J., "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Employed as a Co-channel Detection System.", *IEEE Workshop on Intelligent Signal Processing*, Hungary, 2001.