

FUSION OF USABLE SPEECH MEASURES USING QUADRATIC DISCRIMINANT ANALYSIS

Brett Y. Smolenski, Robert E. Yantorno

Temple University/ECEDept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA

bsmolens@astro.temple.edu, robert.yantorno@temple.edu

http://www.temple.edu/speech_lab

ABSTRACT

Speech that is corrupted by an interfering speaker, but contains segments that are still usable for applications such as speaker identification or speech recognition, is referred to as “usable” speech. The situation where there exists more than one person talking at the same time is referred to as co-channel speech. In general the above speech processing applications do not work in co-channel environments; however, they can work on the extracted usable segments. Unfortunately, currently available usable speech measures only detect about 75% of the total available usable speech with about 25% false alarms. To improve on this performance, optimal Bayesian classification is used to fuse the information of two recently proposed usable speech measures. Since these usable speech measures were gaussian distributed, quadratic discriminant functions were obtained for the optimal Bayesian classifier. Using this approach we were able to obtain a 10% improvement in the total percentage of hit over the Adjacent Pitch Period Comparison (APPC) measure, the best performing measure, and a 10% decrease in the total percentage of false alarms. This amounts to a 39% decrease in detection error.

1. INTRODUCTION

It is well known in the statistical community that one can reduce the likelihood of making type II errors, false alarms, by simply not making decisions on pieces of the data that poorly fit the assumptions of the statistical model being used [1]. The statistician simply considers these pieces of data unusable. A crude but commonly used example of this is the practice of removing *outliers* from a sample and replacing them with the mean value of the remaining samples [2].

In an operational environment the classifiers used in speech processing algorithms are plagued with all kinds of situations not accounted for in their training sets. For example, a speaker may yawn, cough, sneeze, or even laugh in the middle of an utterance. Co-channel speech is another common situation that lies outside the usual speech processing system training set [3].

The traditional approach to co-channel speech processing has been to enhance the target speech while attenuating the interfering speech [3]. Thirty years of these approaches has

produced limited results. However, recently a novel approach to co-channel speech processing has been proposed [4].

Within a co-channel utterance, where both speakers are contributing the same overall energy, there exist several segments of speech where one of the speakers is 20 dB or more above the other speaker [3]. It has been shown that when the target speaker is at least 20 dB greater than the interfering speaker, 80% reliable identification of the target speaker can be obtained [4]. Hence, these segments with a high Target-to-Interferer Ratio (TIR) may be considered usable with respect to speaker identification. Current research has shown that about 35% of a co-channel utterance is usable speech [3]. This is about the same as the average amount of silence contained in an utterance.

Recent advances in co-channel speech processing have produced several usable speech measures, which yield some indication of the TIR [5] [6] [7]. Such measures are necessary to determine usability in an operational environment, since *a priori* knowledge of the TIR would not be available. The usable speech measures used in this research were the Adjacent Pitch Period Comparison (APPC) and a Spectral Autocorrelation Peak-to-Valley Ratio (SAPVR-residual) measure [6] [8].

The APPC measure detects the structure of usable speech by comparing the Euclidean distance between adjacent pitch periods of voiced co-channel speech [6]. When there is little interference ($TIR > \pm 20$ dB) this distance is usually small and the speech is considered usable. However, when the interference is substantial ($TIR < \pm 20$ dB) this distance is usually large and the speech is considered unusable. The SAPVR-residual measure detects usable speech by examining the structure of the autocorrelation of the spectrum of the LPC residual [8]. When there is little interference a pattern of peaks and valleys occurs and the speech is considered usable. However, when the interference is substantial, such patterns fail to appear and the speech is considered unusable. Since these two measures detect usable speech in radically different ways i.e., time-domain versus frequency-domain, it is likely that each measure contains complementary information, which can be exploited in a fusion system [9].

A previously proposed technique approached the problem of fusing the above usable speech measures using independent component analysis (ICA) and non-linear estimation [10]. The idea was to use the de-correlated usable speech measures to form an estimate of the TIR. Since the response variable, TIR, is not normally distributed, non-linear estimation was required

[11]. Due to the many inherent complexities of non-linear estimation, it was thought that a better technique of fusion could be obtained by approaching it as a classification problem.

This paper provides information on the capabilities of a usable speech measure fusion system based on the optimal Bayesian classifier. The idea is to partition the space created by the usable speech measures into two classes, usable and unusable, such that the probability of getting a miss or a false alarm is minimized [12]. Thus, the usable speech measures are treated as features extracted from frames of the co-channel utterance. The resulting probability of detection error will be less than that of any of the individual measures alone, provided each measure contains complementary information [9]. Fortunately, previous research has shown that the two measures do contain complementary information that is correlated with usability [10].

2. QUADRATIC DISCRIMINANT ANALYSIS

Current research has shown that the distributions of the SAPVR-residual and the APPC measures are approximately gaussian distributed [10]. Fortunately, when the measures are approximately gaussian distributed, it can be shown that the resulting optimal partition is a quadratic curve i.e. a parabola, hyperbola, or ellipse [12]. In this situation the optimal classifier having the minimum detection error is very easy to implement.

The decision curve that minimizes the total probability of error would have to satisfy

$$P(w_{use} | X) - P(w_{un} | X) = 0$$

where X is a vector representing the values of the measures and the $w_{use/un}$ represent the two classes of usable and unusable speech. Because of the exponential form of the involved gaussian densities, it is preferable to work with the monotonically increasing logarithmic function of the two density functions [2]. Such a function is referred to as a discriminant function [2]. The following discriminant function was derived

$$g_i(x) = \ln(p(x | w_i)P(w_i)) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) + \ln P(w_i) + C_i$$

where C is a constant, equal to $-(l/2)\ln 2$, μ_i is the mean vector of the measures for the i th class and

$$p(x | w_i) = \frac{1}{2\pi^{l/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right)$$

$$\Sigma_i = E[(x - \mu_i)(x - \mu_i)^T]$$

Current research has shown that the probability of obtaining usable speech $P(w_{use})$ is about 0.35 and, hence, the probability of obtaining unusable speech is about 0.65 [3]. It should be noted that the population mean vector μ and the covariance matrix Σ are not known and must be estimated from training data [2]. Using the maximum likelihood estimates of these parameters has been shown to yield consistent estimates of the corresponding decision curves [12]. The maximum likelihood estimators for the mean vector x and the covariance matrix S for the multivariate gaussian distribution are:

$$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} \quad \text{and} \quad S = \sum_{i=1}^n \frac{(x_i - \bar{x})(x_i - \bar{x})^T}{n-1}$$

respectively [11]. The variance of these estimates decreases with sample size [11].

3. PROCEDURE

For this research all pairs of 42 single speaker utterances, 21 male and 21 female taken from the TIMIT speech corpus, were used to form a co-channel speech database of 861 co-channel utterances (42 choose 2 equals 861). These files were down sampled to 8kHz and the longer file in each pair was truncated to make both files the same length. The files were then combined at 0 dB overall TIR to form the co-channel utterance.

It should be noted that in an operational environment it is highly unlikely that two speakers would be talking over each other during the entire utterance. In addition, each utterance would not have exactly the same length or have the same energy. The reason for using this approach was to capture the worst possible scenario, with respect to both speakers, that one could expect in a co-channel environment.

Once the co-channel utterance was formed it was broken down into 40 ms frames with no overlap, since it has been demonstrated that speaker identification reliability has little dependence on overlap [4]. For each frame, the values for both measures, TIR, signal energy, and spectral flatness were computed. Signal energy and spectral flatness were necessary in order to exclude silence and unvoiced frames, since usable speech measures would not be used with these frames. Usable speech measures are designed to work with only voiced speech, since unvoiced frames provide little information useful for speaker identification [4].

To control the variability and eliminate any bias between the dialect regions only one dialect region was used (region 1 of the TIMIT data base). Research has shown that, provided the utterances are the same length, the characteristics of usable speech depend more on the speaker than on what is being said [4].

Obtaining the optimal decision boundary curve requires estimates of the mean vector and the covariance matrix of the two measures. These estimates were obtained using training data. Once these estimates were obtained it was possible to compute the optimal quadratic curve that best separated the regions of the space spanned by the two measures where the co-channel speech was usable (TIR > ±20dB) and those where the co-channel speech was unusable (TIR < ±20dB). Testing was performed by using this optimal curve on the values of the measures for the remaining co-channel utterances. Half of the 861 co-channel speech files (431) were randomly selected to train the system. The remaining half (430 co-channel files) was used for testing.

4. RESULTS

A scatter plot of the values of the two measures is shown in Figure 1 (below) for a single co-channel utterance containing 53 40ms frames. The plus signs correspond to the values of the measures when the frame of speech was usable (TIR > ±20dB) and the circles correspond to the values of the measures when the frame of speech was unusable (TIR < ±20dB). The quadratic curve is the optimal decision boundary that was computed using the training data.

The region above or below the two curves represents usable speech, while the region inside the two curves represents unusable speech. By inspecting this figure one can easily observe that for this utterance there were 7 usable frames that were missed and 4 false alarms out of a total of 53 frames of co-channel speech. Only one utterance is shown to clearly demonstrate how the system operates during the testing phase.

It should be noted that this system could easily be modified to properly weight the kind of detection errors obtained. For example, if getting a false alarm was considered twice as bad as getting a miss, a shift of coordinates would be all that was required to obtain the desired ratio of false alarms to misses [12]. For this research, the cost of getting a false alarm and the cost of getting a miss were considered equal.

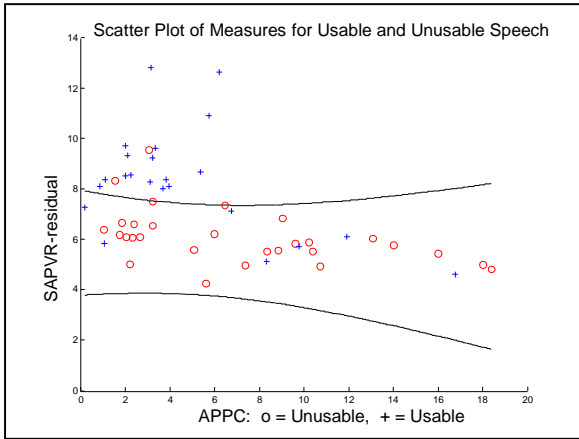


Figure 1: Scatter plot of values of SAPVR-residual and APPC for both the usable (Plus) and unusable (Circle) frames of a single co-channel utterance.

One can compare the above scatter plot to the individual histograms for each of the measures (Figure 2 below). For any given threshold used for APPC, representing any vertical line, the gray bars to the left of this line would be false alarms and any black bars to the right of this line would be misses. The opposite holds for the SAPVR-residual measure.

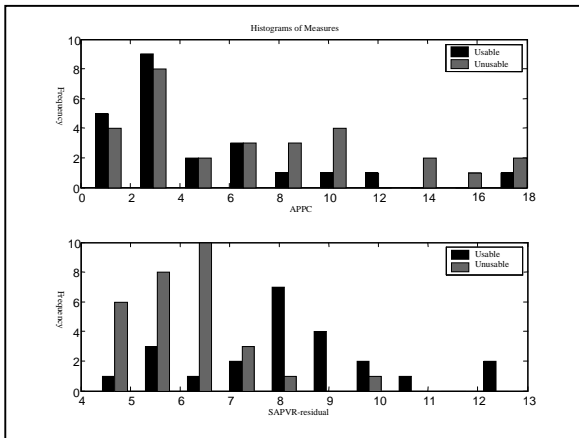


Figure 2: Histograms for the APPC (Upper part) and the SAPVR (Lower part) measures: black corresponds to usable speech frames and gray corresponds to unusable.

For any given threshold used for SAPVR-residual, again representing any vertical line, the gray bars to the right of this line would be false alarms and any black bars to the left of this line would be misses. The threshold lines in the histograms can also be plotted in the scatter plot. For the SAPVR-residual measure the threshold line would be horizontal, while a vertical threshold line would be required for the APPC measure. In comparing Figure 2 with Figure 1, it should be evident that for any pair of threshold lines chosen in Figure 2, one would produce more errors than would be obtained using the quadratic decision curve of Figure 1.

Estimates of the total probability of error were obtained using the remaining 430 co-channel files. Using the Bayesian quadratic classifier, we were able to obtain a 10% improvement in detection error over APPC, the best performing measure, and a 10% decrease in false alarms. Table 1 below shows the results of the fusion system compared to each individual measure for thresholds chosen to produce the same number of misses as false alarms. It is important to note that these results constitute a 2% improvement over the results obtained using non-linear estimation [10].

Table 1: Comparison of the individual measures and their fusion.

Results	APPC	SAPVR-residual	Fused Measures
Correct Detected	74%	71%	84%
False Alarms	26%	29%	16%

5. CONCLUSIONS

It stands to reason that if one had additional usable speech measures available to fuse, additional performance gains could be realized, assuming each additional measure contained complementary information. However, one finds that with each additional measure fused the performance gain, though still positive, would be less than that obtained from that of the previous measure [9]. This stems from the fact that the addition of more features yields less complementary information. Using principle components, preliminary research suggests that about 12 features would account for 90% of the variability in the co-channel utterance. Hence, the dimensionality of classifying usability is around 12.

A possible source of error in using this approach stems from the fact that the usable measures are only approximately gaussian. A Kolmogorov-Smirnov test for normality indicated that these usable speech measures were gaussian distributed with a 90% significance level [10]. Since usable speech measures will not in general be gaussian distributed, it will be necessary to first decompose the distributions of the measures into a sum of gaussians in order to apply this technique. One would then obtain a set of quadratic curves, one for each gaussian component. This set of curves would determine the class boundaries. Such decomposition is called a gaussian mixture model (GMM) and it can be shown that any distribution can be approximated to within arbitrary precision by a mixture of enough gaussians [2]. Also, a support vector machine (SVM) may prove to be a better classifier when the measures are not gaussian distributed.

It should also be noted that the curves obtained to determine the class boundaries are only estimates of the true class boundaries. They are consistent and unbiased estimates obtained from a finite sample [12]. However, the variability in these estimates could cause additional detection errors.

Better fusion results may be possible when one takes into account both class independent, and class dependent measure reliabilities [13]. For example, a larger Euclidean distance from the decision curve to a point in the measure space would imply a more reliable decision could be made. Further, data dependent reliability, such as knowledge that one feature performs better on frames with more energy in higher frequency components, could also be useful [14].

6. FUTURE AREAS OF RESEARCH

One may have noticed that the usability decisions obtained using this technique treat each frame of speech independently of the decisions made on previous frames. The TIR sequence has been shown to be exponentially distributed and appears to have a significant amount of correlation [10]. Hence, one should be able to exploit this inter-frame correlation using a maximum likelihood sequence estimator (MLSE), which would further reduce the probability of error [15]. Fortunately, the previous fusion technique, which used principle component analysis combined with non-linear regression, can be easily adapted to accomplish this task by incorporating previous values of the measures as addition variables [10]. The Verterbi algorithm may prove to be a more efficient way of implementing this process.

In addition, TIR has been the standard against which the effectiveness of these techniques has been evaluated. It may be more prudent to use the actual speaker identification system to delineate the classes of usable and unusable speech. One major reason for doing this would be that even for usable (single speaker) speech there may exist frames that exhibit error prone characteristics and, hence, one would want to reframe from using these frames in the speaker ID system. Such frames would be considered unusable. One could create a database of usable and unusable speech by simply separating frames that were correctly and un-correctly classified by the speaker ID system. A set of features could then be developed that one could use to distinguish the two classes. In theory the performance of any classifier could be greatly improved by having a preprocessor classify and discarded any error prone frames.

ACKNOWLEDGEMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-02-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Also, we would like to acknowledge Ananth Iyer for his useful experiments on TIR and speaker identification.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.

7. REFERENCES

- [1] D. Sheskin, "Statistical Tests and Experimental Design: A Guidebook". Gardner Press: New York, NY, 1984.
- [2] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. San Diego, CA: Academic Press, 1999.
- [3] R. E. Yantorno, "Co-Channel Speech Study". Final Report for Summer Research Faculty, Sponsored by AFRL/IF Laboratory, Rome, NY. 1999.
- [4] J. Lovekin, R. E. Yantorno, D. S. Benincasa, S. J. Wenndt, and M. Huggins, "Developing Usable Speech Criteria for Speaker Identification", ICASSP 2001, pp. 421-424, May 2001.
- [5] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions." IEEE International Symposium. Intelligent. Sig. Process. And Comm. Sys., November 2000.
- [6] J. Lovekin, K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, November 2001.
- [7] A. R. Kizhanatham, R. E. Yantorno, S. J. Wenndt, "Co-channel Speech Detection Approaches Using Cyclostationarity or Wavelet Transform". 4th IASTED International Conference Signal and Image Processing 2002.
- [8] N. Chandra, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Usable Speech Detection Using the Modified Spectral Autocorrelation Peak-to-Valley Ratio Using the LPC residual". 4th IASTED International Conference Signal and Image Processing 2002.
- [9] D. L. Hall, *Mathematical Techniques in Multisensor Data Fusion*. Boston, MA: Artech House, 1992.
- [10] B. Y. Smolenski, R. E. Yantorno, S. J. Wenndt, "Fusion of Co-channel Speech Measures Using Independent Components and Nonlinear Estimation", IEEE, ISPACS, Nov 2002.
- [11] H. Stark and J. W. Woods, *Probability, Random processes, and Estimation Theory for Engineers*. Englewood Cliffs, NJ: Prentice Hall, 1994.
- [12] B. D. Flury, *A First Course in Multivariate Statistics*. New York, NY: Springer, 1997.
- [13] H. Altincay and M. Demirekler, "An Information Theoretic Framework For Weight estimation in the Combination of Probabilistic Classifiers for Speaker Identification," *Speech Communication*. Vol. 30, pp. 255-272, 2000.
- [14] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
- [15] R. J. Freund and W. J. Wilson, *Regression Analysis: Statistical Modelling of a Response Variable*. San Diego, CA: Academic Press, 1998.