A MULTIVARIATE SPEECH ACTIVITY DETECTOR BASED ON THE SYLLABLE RATE

David C. Smith, Jeffrey Townsend, Douglas J. Nelson, Dan Richman
U.S. Department of Defense
Fort Meade, Maryland, USA 20755-6000
E-mail(Smith):dcsmit2@afterlife.ncsc.mil

ABSTRACT

Computationally efficient speech extraction algorithms have significant potential economic benefit, by automating an extremely tedious manual process. Previously, algorithms which discriminate between speech and one specific other signal type have been developed, and often fail when the specific non-speech signal is replaced by a different signal type. Moreover, several such signal specific discriminators have been combined in order to tackle the general speech vs. non-speech discrimination problem, with predictable negative results. When the number of disriminating features is large, compression methods such as Principal Components have been applied to reduce dimension, even though information may be lost in the process. In this paper, graphical tools are applied to determine a set of features which produce excellent speech vs. non-speech clustering. This cluster structure provides the basis for a general speech vs. non-speech discriminator, which significantly outperforms the TALKATIVE speech extraction algorithm.

1. INTRODUCTION

In this paper, recent work at the Department of Defense (DOD) on speech activity detection (SAD) is presented. Those who have attempted to manually locate the boundaries of speech intervals in heterogeneous acoustic signals will attest to the need for a procedure which automates this very tedious and time consuming task.

While significant work has been done on this problem, most efforts have attempted to extract speech segments from data containing one other specific signal type. Unfortunately, when such an algorithm is applied to separate speech from a different specific non-speech signal, results are generally poor. Thus, while the variance of the delta spectrum magnitude is an excellent discriminator of speech vs. music [1], it is not very effective for separating speech from certain modem signals or tones. Blind combination of such specific discriminators in order to develop a general speech vs. non-speech detector only compounds the problem. Moreover, when the number of discriminating features is large, dimension reduction is typically achieved using Principal Components [2], [3] , apparently with the hope that compressing the data according to variance considerations would unravel the confusion.

It is very easy to see that this approach can lead to disastrous consequences. Consider the hypothetical situation where a two dimensional data set defines two elongated non-overlapping ellipsoidal clusters, with parallel major axes. Principal Components designates the direction parallel to these major axes as most significant, since this direction contains most of the total variance of the data. Unfortunately, projecting the data along the major axes will not separate the groups, which overlap along this direction. The graphical method described below has no difficulty in discriminating between such groups.

The method presented in this paper for constructing speech activity detectors utilizes visual displays to select a small set of features which produce good speech vs. non-speech clusters. This cluster structure may be learned by a standard classifier, which may be applied to separate speech and non-speech. An example detector is constructed, which significantly outperforms the Canadian speech extraction algorithm, TALKATIVE [4].

2. EARLIER SAD ALGORITHMS

During recent years, several speech activity detectors have been developed and utilized with varying degrees of success.

One of the earliest was the Readability algorithm, which essentially performs an autocorrelation of an input signal segment, and classifies the segment as speech if and only if the peak of the autocorrelation is within the pitch range of speech (60-400 Hz). Unfortunately, the algorithm is easily fooled by certain signalling tones, impulsive noise, muzak, and various modems. Moreover, the algorithm is sensitive to channel gain.

An improvement over the Readability algorithm is the Nelson-Pencak algorithm [5], which sorts the power spectrum of the input signal segment, and computes the ratio of high-powered components (assumed to be signal) to low-powered components (assumed to be noise). The speech/non-speech decision is made by thresholding the resulting SNR. Pre-whitening and signal variance are used in secondary tests. The NP algorithm generally outperforms Readability and is insensitive to channel gain, although certain signals such as modulated tones and muzak produce high false alarm rates.

The TALKATIVE algorithm performed among the best in a recent evaluation of speech activity detectors conducted by DOD researchers [6]. This algorithm assumes that speech is non-stationary, and that this non-stationarity is reflected in vectors of cepstral coefficients. The Euclidean distances between nearby pairs of cepstral vectors are averaged and thresholded to give a speech/non-speech decision. TALKATIVE has difficulty with muzak and spurious tones, and degrades ungracefully in mild white noise conditions.

3. NEW APPROACH

The method proposed in this paper for constructing speech activity detectors consists of four steps: (1) assemble a pool of candidate algorithms which are surmised to have good speech vs. non-speech discrimination potential; (2) from the candidates selected in step one, utilize graphical tools such as XGobi [7] to determine a small number of features which produce good clustering of speech vs. non-speech; (3) apply a classification algorithm based on the cluster structure learned in step two; (4) apply secondary tests to remove any remaining non-speech.

4. NEW CANDIDATE ALGORITHMS

Four candidate algorithms were expected to contribute to solving the speech vs. non-speech discrimination problem. The tests were: (1) normal spectrum, (2) syllable rate detector, (3) baud rate detector, and (4) carrier detector.

The normal spectrum is computed by averaging the DFT of the time waveform over several ovelapping frames whose union is the input signal segment.

The syllable rate detector is constructed by applying a low pass filter (passband around 60 Hz) to the magnitude of the analytic version of the signal (i.e., the complex valued signal having the original signal in the real part and the Hilbert transform of the original signal in the imaginary part). The spectrum of the AM envelope of

the low-passed analytic signal is output. Since the syllable rate of speech is around 5Hz, one expects to observe a bulge in this spectrum near this value for input speech segments.

The baud rate detector determines the baud rate of a modem or the fundametal pitch frequency of speech as the peak in the cross power spectrum (the ouput of the algorithm, see [8] for details) of the magnitude of the analytic signal. Since most modems have a baud rate above 1000 Hz, while the pitch range for speech is between 60 and 400Hz, it was anticipated that this test would do most of the work in separating speech from bauded signals.

The carrier detector exploits the weak fourth order symmetry of most bauded signals by computing the crosspower spectrum of the fourth power of the magnitude of the anayltic signal. A spectral bulge is anticipated at four times the carrier frequency.

Clearly, many other prospective algorithms may be proposed, but good results were obtained starting with only these algorithms.

Each of the four tests takes a signal segment as input and produces a Fourier spectrum as ouput. The corresponding amplitude spectrum may be normalized to produce a probability vector (i.e. a real vector with non-negative components which sum to one). Features can be extracted from these probability vectors, such as statistical moments and magnitude ratios. The normalization also ensures that the tests are invariant to channel dependent gains.

For each test, we investigated three features (mean, variance and ratio of largest to average component of the probability vector) for several different signal types, computed over one second data windows. This window length is a compromise between the time thought necessary to adequately resolve the syllable rate for speech (around 5 cycles per second) and the requirement that the algorithm locate boundaries between speech and non-speech fairly accurately.

A preliminary analysis using histograms showed that the normal spectrum and the carrier detector had considerable difficulty separating speech from non-speech; consequently these candidate algorithms were disgarded. The baud rate detector and the syllable rate detector appeared to have the best potential for discrimination between speech and non-speech signals.

Fig 1. displays the the AM envelopes of a speech signal and a modem signal, while Fig 2. shows the amplitude spectra obtained in applying the syllable rate detector to these envelopes. The large spike near 5 Hz in Fig. 2 (top) represents the syllable rate.

5. FEATURE SELECTION

Starting with the six features from the remaining two candidate algorithms, we exhaustively examined the cluster structure produced by various combinations of two and three features, using the Xgobi graphics package. It was surprising to learn that the three features from the syllable rate detector alone produced the best separation of speech vs. non-speech. Fig. 3 (top) shows the clusters obtained with the three components of the baud rate detector, while Fig. 3 (bottom) illustrates the superior clustering found with the syllable rate features. Combination of various syllable rate and baud rate features scored no better than syllable rate features alone.

The fact that the AM envelope from the syllable rate detector is an excellent speech vs. non-speech discriminator is easy to understand. It is the distinctive oscillatory pattern of the envelope between relatively large amplitudes and zero which alerts one to the possible presence of speech when humans view a speech waveform. By contrast, the envelope for the bauded signal in Fig. 1 is bounded away from zero, and the oscillations are not as extreme as those in the speech envelope. Tones and several other non-speech signals also have envelopes which are bounded away from 0, with little oscillation. Indeed, we will further exploit the shape of the envelope to obtain improved speech vs. non-speech discrimination on .5 second windows, in Section 8.

6. QDA CLASSIFICATION ALGORITHM

Based on the shape of the clusters obtained with Xgobi, it was decided to apply Quadratic Discriminant Analysis [9] to learn the cluster structure and to classify new data. For different data sets, other classification methods such as CART [10] might work better, depending on the modality of the clusters. We recall that in QDA, the probability of an observation vector $\mathbf{x}$ belonging to class G_i is

$$P(G_i|x) = \frac{p_i\sqrt{|S_i|}^{-1} e^{-\rho(x,\mu_i)}}{\left(\sum_j p_j\sqrt{|S_j|}^{-1} e^{-\rho(x,\mu_j)}\right)}, \quad (1)$$

where

$$\rho(x,\mu_i) = \frac{1}{2}(x-\mu_i)^t S_i^{-1}(x-\mu_i) \quad (2)$$

is the (squared) Mahalanobis distance between $\mathbf{x}$ and ith traininig class and

$\mu_\mathbf{i}$ is the mean vector of the ith training class;
$\mathbf{S_i}$ is the covariance matrix of the ith training class;
G_i is the ith class;
p_i is the prior probability of belonging to the ith class;
$|\mathbf{S_i}|$ is the determinant of $\mathbf{S_i}$.

An observation data vector $\mathbf{x}$ is assigned to the class which has the largest $P(x|G_i)$ value.

7. EXPERIMENTS

A data base of homogeneous one-second long time waveforms containing representatives from eleven different signal types (spontaneous speech, recorded speech, music, speech noise, bauded signals, silence, clicks, tones, multi-tones, white noise, unusual noise) was constructed. Spontaneous speech and recorded speech formed one class, while the other nine signals collectively formed the non-speech category. The data base was randomly split into training and testing sets and QDA classification rates were computed for various combinations of two and three features selected from among the six candidate features. This experiment confirmed the heuristic result of Section 5 that the best separation ocurred when the three syllable rate features alone were used.

We desired to publish a speech/non-speech decision every .1 second as we processed a data file from beginning to end. To accomplish this, we marched our one second decision window through the file, sliding it by .1 second increments. In this way a fixed .1 second data segment would be part of ten different speech/non-speech decisions. A final decision for the fixed .1 second window was obtained by polling these ten decisions.

Classification rates obtained with the new algorithm (which we refer to as SRSAD) were compared with those obtained with TALKATIVE for several hand marked heterogenous acoustic files. The initial results indicated no clear winner. It was observed that most of the files where TALKATIVE outperformed SRSAD contained many short (a second or less) speech segments separated by short non-speech segments. Since TALKATIVE's decision window was .32 seconds vs. 1.0 second for SRSAD, we conjectured TALKATIVE was better able to locate speech/non-speech boundaries.

8. IMPROVEMENTS TO THE SRSAD

A shorter decision window for SRSAD seemed desirable for resolving boundaries, yet longer decision windows are better for determining the syllable rate. Indeed, experiments showed that shortening the decision window weakened the discrimination power of the algorithm.

Fortunately, a modification of the algorithm was discovered which permitted the decision window to be shortened, while simultaneously improving classification rates. Note that the shape of the graphs of the amplitude spectra in Fig. 2 are fairly similar, which makes discriminating between speech and bauded signals difficult for SRSAD. Note also that the bottom curve in Fig 2 resembles the

graph of a scaled magnitude *sinc* function, |sin(x)/x|. This is consistent with the ragged step function-like shape of the modem envelope in Fig. 1 (recall that the Fourier transform of a step function is a scaled *sinc* function). By subtracting the means of the (zero-padded) AM envelopes before taking the DFT, the spectral amplitude graphs in Fig. 4 for the speech and modem signals were obtained. The amplitude spectrum of the modem envelope no longer resembles a *sinc* function, while the amplitude spectrum of the speech signal has changed only slightly. A similar effect was observed for several other non-speech signals. When SRSAD was modified to incorporate mean subtraction, it was possible to shorten the decision window length to .5 seconds and simultaneously improve classification rates. Indeed, false alarm rates dropped by 6-15% on several files.

9. EVALUATION OF THE SRSAD

To determine how well SRSAD locates boundaries between speech and non-speech intervals, files consisting of alternating segments of speech and either tones or modems were synthetically constructed, using data not seen in training. By varying decision thresholds in TALKATIVE and speech priors in SRSAD, performance was plotted as ROC curves, with false alarm rates on the horizontal axis and detection rates on the vertical axis. Fig. 5 shows ROC curves for the SRSAD and TALKATIVE when the alternating segments were .8 seconds long. SRSAD performs about 4-5% better in detection rate at the same false alarm rate. Similar improvements were observed when the segment length was varied between .5 and 2.0 seconds.

The algorithms were also compared when uniformly distributed white noise was added to telephone quality speech. Both algorithms were run on a file consisting of several 2 to 3-second sentences uttered by various speakers, which were separated by 1.5-second intervals of silence. The ROC curves for this clean file were nearly identical for both algorithms. However, at 3dB SNR SRSAD very significantly outperformed TALKATIVE, as seen in Fig. 6. In fact, there was very little degradation in SRSAD performance even at 0dB SNR.

Finally, SRSAD and TALKATIVE were compared on 128 naturally occuring signal files containing speech and non-speech, which were obtained from diverse sources not seen in training SRSAD. These files were concatenated, and SRSAD and TALKATIVE were compared on this merged file. The resulting (transposed) ROC curve in Fig. 7 shows that TALKATIVE's false alarm rate is significantly higher than SRSAD's, at the same detection rate. This margin increases from about 2% at 82% speech detection rate to about 18% at 96% speech detection rate.

It should be emphasized that these results hold for the concatenated file, and therefore may serve as predictors of average performance only. Indeed, TALKATIVE outperformed SRSAD on a few individual files.

10. SECONDARY TESTING

While SRSAD very significantly outperforms TALKATIVE on average, certain signals are troublesome for both algorithms. Impulsive noise, very short tones and muzak often produce high false alarm rates. Although impulsive noise and short tones have envelopes which resemble step functions, mean subtraction does not significantly alter the *sinc* function shape of the amplitude spectra in these cases. This is because most of the windowed signal is zero (or nearly zero) valued, so that the mean is frequently negligible. It is anticipated that a pre-whitening filter will be effective in removing these impulsive components from the signal envelope.

Muzak is a more difficult problem. The shapes of the AM envelope of muzak and speech are very similar, and a secondary test is required to separate these two signal types. A solution that works well in low noise environments is based on the observation that the muzak envelopes rarely cross the horizontal axis, whereas speech envelopes often do. An empirically determined threshold of approximately 25% of the envelope mean was set, and the number of times the envelope crossed this threshold (excluding endpoints) was tallied. A signal segment was classified as non-speech if it never crossed the threshold; otherwise it was sent to the QDA classifier. The overall classification rates obtained in low noise environments were excellent, from just over 90% for the worst language to over 95% for English. Moreover, incorporating this test into SRSAD did not compromise classification rates obtained previously.

11. SUMMARY AND CONCLUSIONS

Starting with a suite of candidate discriminators, we were guided by visual display to select a few channel independent features which produced exellent speech vs. non-speech clustering. The best features were the mean, variance and magnitude ratio from the syllable rate detector. This heuristic conclusion was verified by performing a search over various subsets of candidate features, using QDA. A preliminary comparison of the new SRSAD algorithm with TALKATIVE revealed no clear winner. However, after modifying SRSAD by subtracting the mean of the AM envelope before transforming, classification rates improved substantially. Furthermore, it was possible to shorten the decision window from 1.0 to .5 seconds, which improved SRSAD's capacity to resolve speech/non-speech boundaries. This modified SRSAD algorithm is far superior than TALKATIVE for extracting speech segments in white noise conditions. Finally, SRSAD very significantly outperformed TALKATIVE, on average, when the algorithms were compared on a large number heterogenous acoustic signal files not seen in training.

While the improved performance of SRSAD over the TALKATIVE is significant, the importance of the technique should be emphasized. It has been demonstrated that visual displays may be employed to determine features which produce excellent clustering of speech vs. non-speech, and the cluster structure may be used as a basis for successful SAD algorithms. Use of graphical displays also avoids possible information loss which may result upon blind application of compression methods such as Principal Components.

12. ACKNOWLEDGMENT

Special thanks to David Bisant of R5 for allowing us to train on several data files which he had hand marked for another project.

13. REFERENCES

[1] Scheirer, E. and Slaney, M., "Construction of a Robust Multifeature Speech/Music Discriminator," Proceedings of IEEE, ICASSP-'97, Vol. 2, Pp. 1331-4, 1995.

[2] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, 1984, New York.

[3] Smith, D. C., "A Note on Principal Components and Improvement of the Signal-to-Noise Ratio," internal DOD Technical Report Z36-TSR/05/95, 1995.

[4] Gagnon, L. and Cyr, M., "TALKATIVE: A Computationally Efficient Algorithm to Extract Speech From Heterogenous Audio Signals," CSE E4 internal Technical Memorandum E402-7-1, 1996.

[5] Pencak, J. and Nelson, D.J., "The NP Speech Activity Detection Algorithm," Proceedings of the IEEE, ICASSP-'95, Vol. 1, Pp. 381-4, 1995.

[6] "1996 Corpus Development and Evaluation of Speech Activity Algorithms," internal DOD Technical Report, 1996.

[7] XGobi web site: http://lib.stat.cmu.edu/general/XGobi/.

[8] Nelson, D., "Special Purpose Correlation Functions for Improved Signal Detection and Parameter Estimation," Proceedings of the IEEE ICASSP-'93, Vol. IV, pp. 73-6, 1993.

[9] Higbee, K., "Variations of Linear and Quadratic Discriminant Analysis for Multivariate Classification," internal DOD Technical Report Z032-TSR/03/93, 1993.

[10] Breiman, L. et. al., *Classification and Regression Trees, Monterey,* Wadsworth and Brooks, 1984.

14. FIGURES

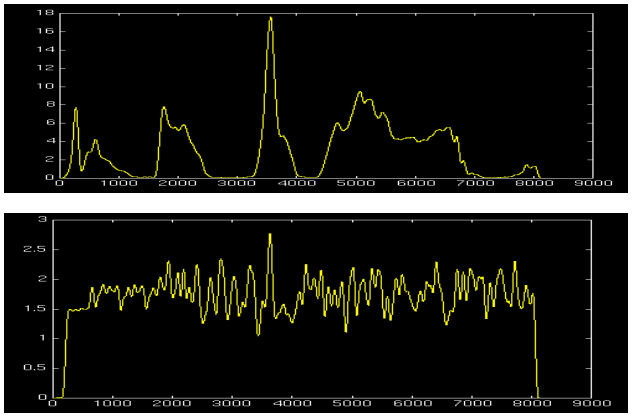

Figure 1. AM envelopes of a speech waveform (top) and a modem waveform.

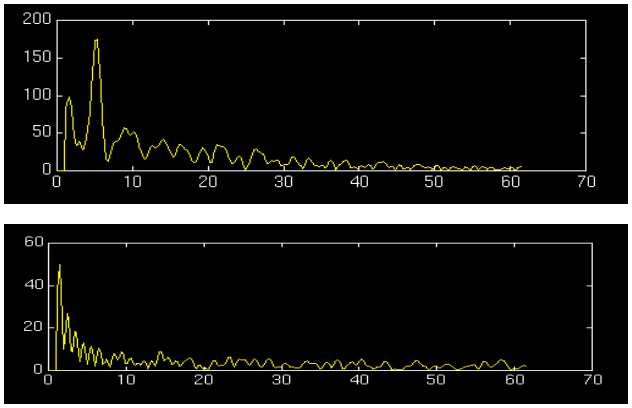

Figure 2. Amplitude Fourier spectra of the AM envelopes of a speech signal(top) and a modem signal.

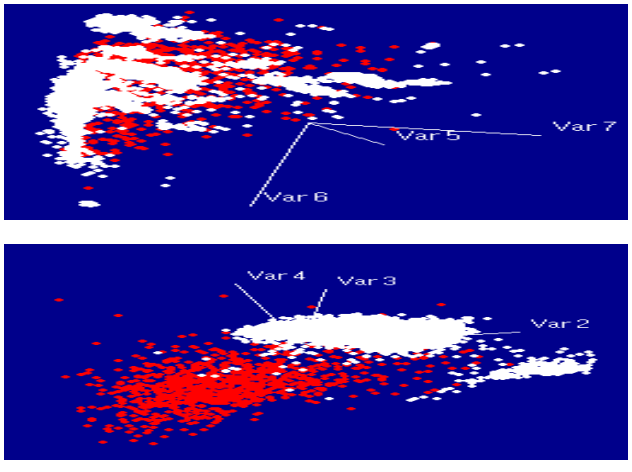

Figure 3. Xgobi scatter plots of speech and non-speech using three features derived from the baud rate detector (top) and from the syllable rate detector (non-speech in white).

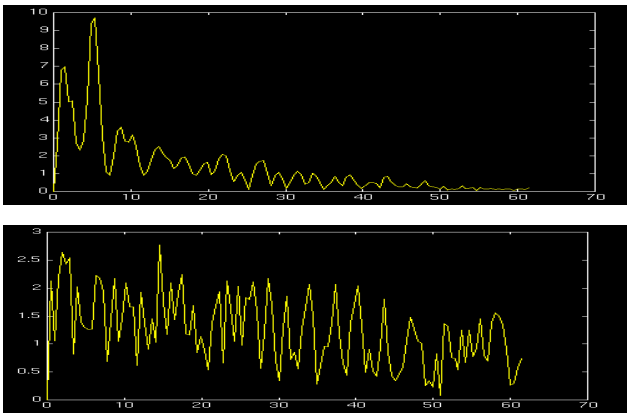

Figure 4. Amplitude Fourier spectra of the AM envelopes with means removed of a speech signal (top) and a modem signal.

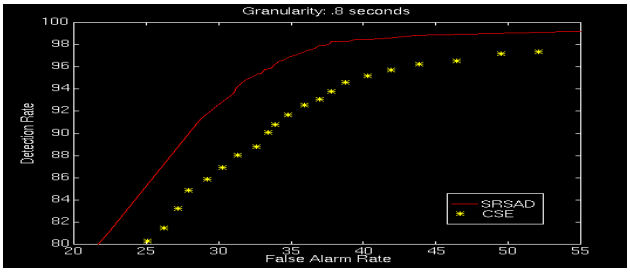

Figure 5. ROC curves of SRSAD and TALKATIVE with alternating .8 second segments of speech and non-speech.

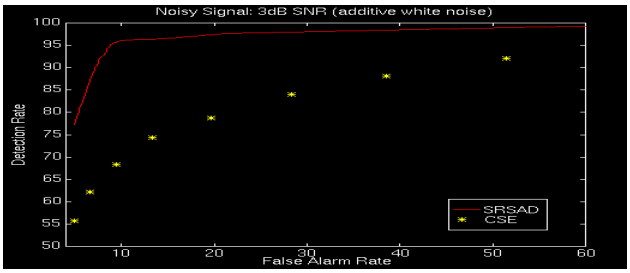

Figure 6. ROC curves of SRSAD and TALKATIVE for noisy (3 dB) speech segments separated by silence.

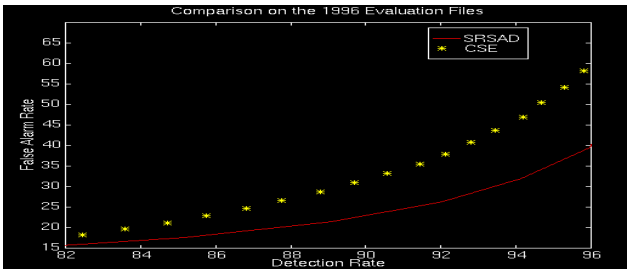

Fig. 7. Comparison of the SRSAD and TALKATIVE on a large number of heterogeonous data files (transposed ROC curves).