SEQUENTIAL *k*-NEAREST NEIGHBOR PATTERN RECOGNITION FOR USABLE SPEECH CLASSIFICATION

Jashmin K Shah[†]*, Brett Y Smolenski*[†]*, Robert E Yantorno*[†] *and Ananth N Iyer*[‡]

[†]Temple University, Speech Processing Lab
1947 N 12th Street, Philadelphia PA 19122-6077 USA
Email: shah, bsmolens, robert.yantorno@temple.edu
URL: http://www.temple.edu/speech_lab

[‡]Department of Electrical Engineering
The Pennsylvania State University
121 EE East Building, University Park PA 16802-2705 USA
Email: aniyer@psu.edu

ABSTRACT

The accuracy of speech processing techniques degrades when operating in a co-channel environment. Co-channel speech occurs when more than one person is talking at the same time. The idea of usable speech segmentation is to identify and extract those portions of co-channel speech that are minimally degraded but still useful for speech processing application such as speaker identification. Usable speech measures are features that are extracted from the co-channel signal to distinguish between usable and unusable speech. In this paper, a new usable speech extraction technique is presented. The new method extracts features recursively and variable length segmentation is performed by making sequential decisions on the *k*-NN pattern classifier class assignments. This new approach is able to identify 79% of available usable speech segments with 21% false alarms and it requires lesser amount of data to make accurate decisions compared to previously presented methods.

Key words: Co-channel Speech, Usable Speech, Sequential Detection, *k*-Nearest Neighbor Classifier, Speaker Identification.

1. INTRODUCTION

The performance of speaker identification system degrades under co-channel conditions, i.e., when two people are talking at the same time. Speech signal from the prominent speaker or that from a speaker of interest is termed as *target speech* and that from the interfering speaker is termed as *interferer speech*. It was shown that an energy ratio of the target and the interfering speech signals, the *Target-to-Interferer Ratio* (TIR) is a good metric to identify portions of the speech data usable for speaker identification [1].

Previous studies have revealed that about 40% of the co-channel speech has enough information about the target speaker to perform reliable speaker identification even when the overall target speech energy is equal to that of the interferer speech energy. All frames having a TIR (computed over a 10 msec frame) greater than 20dB is considered to be "usable" for the speaker identification system. It was also found that the accuracy of speaker identification system can be improved from 45% to 90% when extracted usable speech segment used as an input to the speaker identification system instead of co-channel speech [2]. Figure 1 shows the application of usable speech processing for speaker identification system where, extracted usable speech segments are used in the speaker identification system. The co-channel data considered in this research are considered to be recorded over a single microphone and hence the target speech energy and interferer speech energy cannot be computed from the co-channel data. Hence usable speech measures which have high correlation with the TIR is needed to determine the usability in co-channel speech.

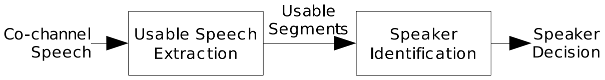

Figure 1. Application of Usable Speech Extraction System for Speaker Identification System.

Several usable speech measures have been proposed [3, 4, 5, 6, 7, 8] with mediocre performance in usable speech identification. These measures perform a frame-by-frame (of a fixed frame length) and consider periodicity or structure of the speech frame to identify usable speech segment. In this paper, we take a different approach to identify the usable speech segments. The fixed size frame analysis is eliminated by studying the speech data at a sample-by-sample level. The recursive least squares (RLS) algorithm is used to obtain the coefficients of the all-pole speech model. These coefficients are used in a *k*-Nearest Neighbor (*k*-NN) pattern classifier to obtain sample-by-sample class associations. Speech segmentation is later achieved by grouping samples belonging to the same class together by the sequential probability ratio (SPR) test. The SPR is a class ratio computed for each class and segment class associations are made comparing by the SPR with a fixed threshold. The threshold is determined based on the lenience to be given for the segmentation scheme. We will use the term RLS-*k*NN algorithm to denote the the segmentation scheme being developed.

The remainder of this paper is organized as follows. A brief background to the *k*-NN pattern classifier and concepts of sequential detection is presented in the following section. The proposed RLS-*k*NN algorithm is described in section 3. Experimental evaluation of the new classification scheme presented in section 4. Finally, conclusions are drawn in section 5.

2. BACKGROUND

The RLS-*k*NN algorithm consists of two main parts. First, the feature extraction step using the recursive least squares [9] and second, a *k*-NN classifier performing sequential classification.

2.1 *k*-NN Classifier

The *k*-NN classifier is a very simple non-parametric method for classification. Despite the simplicity of the algorithm, it performs very well and is an important benchmark method. The *k*-NN classifier, as described by [10, 11], requires a distance metric d, a positive integer k, and the reference templates X_m of m labeled patterns. A new input vector $\mathbf{x}$ is classified using the subset of k-feature vectors of that are closest $\mathbf{x}$ to with respect to the given distance metric d. Mathematically this can be described to compute the *a posteriori* class probabilities $P(\omega_i|\mathbf{x})$ as

$$P(\omega_i|\mathbf{x}) = \frac{k_i}{k} \cdot p(\omega_i) \tag{1}$$

where k_i represents the number of vectors belonging to class ω_i within the subset of k vectors. The main disadvantage of this classifier is distance metric computations, which increases with the increase in number of patterns in the reference templates. Pattern x is assigned to the class ω_i with the highest *a posteriori* class probability $P(\omega_i|\mathbf{x})$.

2.2 Sequential Detection

Wald [12] introduced the concept of sequential test and formulated *sequential probability ratio test* (SPRT). The test was designed to decide between two simple hypotheses sequentially. Given two constant as the upper and the lower stopping thresholds and the hypotheses H_0 and H_1, by observing the data and computing the accumulated log likelihood ratio sequentially, SPRT can make a decision on either continuing observation or stopping the testing accepting H_0 and H_1. This algorithm needs pre-determined threshold value for decision.

Sequential detection scheme over the *k*-NN class associations is made and hence automatically segmenting the speech data into two classes. Hypothesis H_0 corresponds to declaring the segment as usable speech and H_1 corresponds to declaring the segment as unusable speech. On every incoming sample of speech, the SPRT is done and one of the three possible decisions is made.

1. Decide H_0
2. Decide H_1
3. Not enough information to decide either H_0 or H_1.

If decision 1) or 2) is made, the hypothesis testing procedure stops. Otherwise, an additional observation is taken, and the test is performed again. This process continues until a decision is made either in favor of H_0 or H_1. Note that the number of observations taken to obtain a decision of H_0 or H_1 is not fixed but a random variable.

3. RLS-*k*NN ALGORITHM

The RLS-*k*NN algorithm performs segmental classification of speech data. The segmental classification is accomplished by classifying on a sample-by-sample basis. It is easy to realize that a sample-by-sample classification would need large amounts of computational power. Hence in this algorithm we extract features recursively and simultaneously perform classification.

Recursive-least squares was used to obtain the 14^{th} order auto-regressive model coefficients. A step-size of 0.4 and forgetting-factor of 0.99 was used in the computation of the weight vector $\tilde{\mathbf{w}}[n]$ for sample point n:

$$\tilde{\mathbf{w}}[n] = \tilde{\mathbf{w}}[n-1] + \kappa[n]e[n] \tag{2}$$

where $e[n]$ is the error in prediction and $\kappa[n]$ is the update factor which is computed using the step-size, forgetting factor and inverse correlation matrix. The new weight vector computed at every recursion is used to determine the *a posteriori* class probabilities $P(\omega_i|\tilde{\mathbf{w}})$ for the *k*-NN classifier using equation 1. The class probabilities are then mapped as labels $\phi[n]$: “1” representing usable speech class and “0” represents the unusable speech class:

$$\phi[n] = \begin{cases} 1, & \text{if } P(\omega_1|\tilde{\mathbf{w}}) \geq P(\omega_0|\tilde{\mathbf{w}}) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In the next step the segmental classification is done, by considering a sequence of labels obtained at each recursion. We define the SPR φ^j for segment j and classes ω_0 and ω_1 as:

$$\varphi_0 = \frac{(n-n_s^j) - \sum_{k=n_s^j}^{n} \phi[k]}{n-n_s^j} \tag{4}$$

$$\text{and } \varphi_1 = \frac{\sum_{k=n_s^j}^{n} \phi[k]}{n-n_s^j} \tag{5}$$

where n_s^j denotes the index of the beginning sample point and we will use n_e^j to denote the segment end sample point of the segment j. The class ratios φ_i^j are compared to a fixed threshold ρ. A valid range for this threshold is $0.5 \leq \rho \leq 1$. If $\varphi_i^j \geq \rho$ for any $\{i : 0,1\}$, then the segment j between indices n_s^j and $n_e^j = n$ is assigned to class ω_i and a new segment with start point index $n_s^{j+1} = n+1$

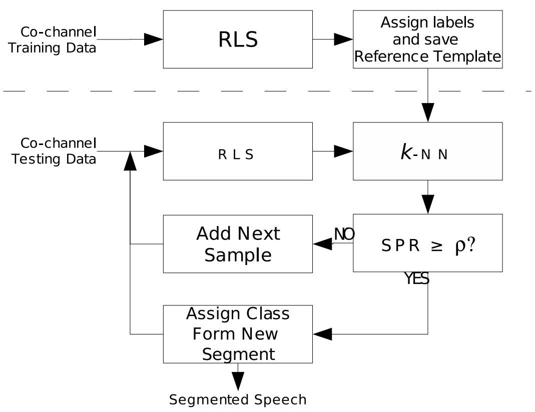

Figure 2. Usable Speech Segment Classification Using Sequential *k*-NN Classifier.

The RLS-kNN algorithm performing classification of co-channel data is illustrated with the block diagram shown in figure 2. Co-channel speech is the input to the system and the output is speech segmented into usable and unusable classes. the speech segment with usable or unusable labels. The steps shown above the dashed line represents the training process and the steps below the dashed line represents the testing process.

4. EXPERIMENTAL SETUP AND RESULTS

To evaluate the proposed segmentation algorithm, two separate schemes for training and testing were designed. The training scheme requires *a priori* knowledge of the class associations of the training data.

4.1 Training

The training process involves the RLS coefficient computations and assigning labels based on the TIR values and the TIR threshold of 20dB. Speech data was taken from TIMIT database was used for all the experiments. A subset of 42 files from speakers spanning the entire dialect regions were chosen. The original speech was sampled at 16 kHz and re-sampled to 8 kHz after low-pass filtering to 3 kHz. Two utterances were read at a time (hence making a total number 861 co-channel utterances) and their amplitudes were scaled such that the energy of the two utterances are equal over the entire utterance. The scaled speech data was added to simulate co-channel data recorded over a single microphone. The TIR values were computed over fixed frame sizes of 10 m sec. The RLS coefficients $\tilde{\mathbf{w}}[nT]$ were extracted to make the coefficients synchronous with the TIR values. T is the number of samples corresponding to the frame size of 10 m sec.

4.2 Testing

Of the 861 co-channel utterances created, 431 utterances were used for the training process and the remaining 430 utterances were used for testing and evaluation. The first step in the testing stage is to perform RLS and obtain the coefficients. At every recursion the k-NN classifier performs classification and assigns class labels as described in section 3.

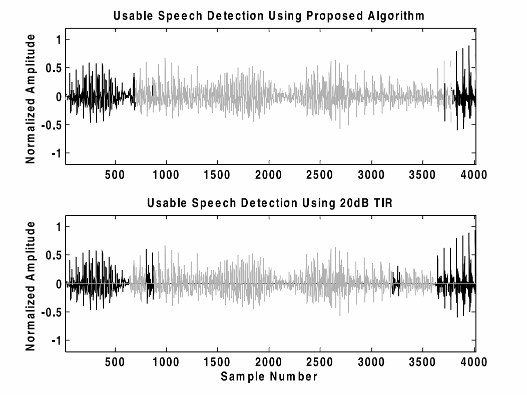

Figure 3. Comparison of Detection of Usable and Unusable Speech Segment: Upper panel using sequential k-NN pattern classifier and bottom panel using 20dB TIR, Black color shows usable speech segment and gray color shows unusable speech segments.

In this experiment, the value of k was chosen as 9, and SPR of 0.65 was chosen. These numbers gave the best performance in the experiments. A Correct detection (hit) is said to occur when both the RLS-kNN classifier and TIR identifies a segment of speech belonging to the same class. False alarms occur when RLS-kNN classifier and TIR declares a speech segment to belong in different classes. Figure 3 shows the comparison of detection of usable and unusable speech segment between proposed algorithm and using TIR with 20dB threshold.

The classification performance of the sequential k-NN pattern classifier is evaluated based on the confusion matrix computed and given below.

$$\text{Confusion matrix} = \begin{bmatrix} 0.79 & 0.21 \\ 0.23 & 0.77 \end{bmatrix}$$

The rows of the confusion matrix represent the performance actual classes and the columns represent the identified classes. The first row represents performance of classifying usable speech and the second row represents the performance of classifying unusable speech. The percentage of correct identifying usable speech is 79% and unusable is 77%. The false alarms are 21% and 23% respectively. This gives the overall identification rate of 78%.

The proposed algorithm was compared with the best performing usable speech measure: Adjacent Pitch Period Comparison (APPC) under the same experimental conditions and the results are presented in figure 4. It should be noted that the sequential k-NN pattern classifier was able to increase correct detection of usable speech by 6% and reduce false alarms by 18% relatively with respect to the performance of APPC.

It was observed that the proposed technique of sequential k-NN required only 48 sample points on an average to make a class decision, i.e., atleast one of the SPR exceeding 0.65. Other usable speech measures use a fixed 320 samples (40 ms) frame of speech for usable speech detection. Hence, we can also conclude that the new method requires less data to make a decision in the statistical sense.

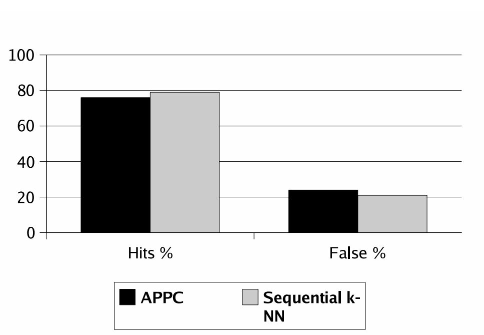

Figure 4. Comparison of Results of Detection of Usable Speech: Gray bars represent performance using APPC measure and black bars using sequential k-NN classifier.

5. CONCLUSIONS

The purpose of this paper was to develop a sequential *k*-NN classifier and evaluate it to classify the usable and unusable portions of co-channel speech in the context of speaker identification. It was found that by using 14-RLS coefficients as a feature and sequential *k*-NN classifier; we were able to achieve identification rate of 78%. It was noticed that one can obtain desired performance rate by changing SPR. Also, it was observed that the proposed algorithm requires less data to decide on the class memberships.

In the proposed algorithm, we have used only one set of features (RLS); however due to the fact that speech is nonlinear in nature and single a feature can not model entire system. This leads to poor classification performance. Therefore to improve the classification rate, one can think of using the usable speech measures itself as a feature for the sequential *k*-NN classifier. In the current research, the SPR was chosen based on heuristics, it is of our next interest to look at the receiver operating characteristic (ROC) curves and decide on the best value for SPR.

6. ACKNOWLEDGMENT

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 thereon.

7. DISCLAIMER

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

REFERENCES

[1] R. E. Yantorno, “Co-channel speech study, final report for summer research faculty program,” Tech. Rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1999.

[2] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D.B. Benincasa, and S. J. Wenndt, “Developing usable speech criteria for speaker identification,” *IEEE, International Conference on Acousitcs and Signal Processing*, pp. 424–427, May 2001.

[3] J. M. Lovekin, K. R. Krishnamachari, and R. E. Yantorno, “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*, pp. 139–142, Nov 2001.

[4] N. Chandra and R. E. Yantorno, “Usable speech detection using modified spectral autocorrelation peak to valley ration using the lpc residual,” *4th IASTED International Conference Signal and Image Processing*, pp. 146–150, 2002.

[5] A. R. Kizhanatham, R. E. Yantorno, and B. Y. Smolenski, “Peak difference autocorrelation of wavelet transform (pdawt) algorithm based usable speech measure.,” *IIIS Systemics, Cybernetics and Informatics*, Aug 2003.

[6] N. Sundaram, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, “Usable speech detection using linear predictive analysis - a model-based approach,” *IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS*, 2003.

[7] A. N. Iyer, M. Gleiter, B. Y. Smolenski, and R. E. Yantorno, “Structural usable speech measure using lpc residual,” *IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS*, 2003.

[8] Y. Shao and D-L. Wang, “Co-channel speaker identification using usable speech extraction based on multipitch tracking,” *IEEE International Conference on Acoustics, Speech, and Signal Processing,*, vol. 2, pp. 205–208, 2003.

[9] S. Haykin, *Adaptive Filter Theory*, Pearson Education, 4 edition, September 2001.

[10] E. Fix and J. L. Hodges, “Discriminatory analysis - nonparametric discrimination: Consistency properties,” Tech. Rep. Project 21-49-004, Report No.4, USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, New York, 2nd edition edition, 2001.

[12] A. Wald, *Sequential Analysis*, Wiley, New York, 1947.

[13] C. W. Baum and V. V. Veeravalli, “A sequential procedure for multihypothesis testing,” *IEEE Transaction on Information Theory*, vol. 40, 6, no. I, pp. 1994–2007, 1994.

[14] V. Matta S. Marano and P. Willett, “Sequential detection of almost-harmonic signals,” *IEEE Transaction on Signal Processing*, vol. 51, 2, pp. 395–406, February 2003.