ENHANCEMENT OF SPEAKER IDENTIFICATION USING SID-USABLE SPEECH

Saurabh S. Khanwalkar[†], *Brett Y. Smolenski*[†], *Robert E. Yantorno*[†] *and S. J. Wenndt*[‡]

[†]Speech Processing Lab, ECE Department, Temple University
12th & Norris Streets, Philadelphia, PA 19122-6077, USA

[‡]Air Force Research Laboratory/IFEC,
32 Brooks Rd. Rome NY 13441-4514, USA

ABSTRACT

Most present day Speaker Identification (SID) systems focus on the speech features used for modeling the speakers without any concern for the speech being input to the system. Knowing how reliable the input speech information is can be very important and useful. The idea of SID-usable speech is to identify and extract those portions of corrupted input speech, which are more reliable for SID systems, thereby enhancing the speaker identification process. In this paper, usability in speech, with reference to speaker identification is presented which is called *SID-usable speech*. Here the SID system itself is used to determine those speech frames that are usable for accurate speaker identification. Two novel approaches to identify SID-usable speech frames are presented which resulted in 78% and 72% correct detection of SID-usable speech. It is also shown that SID performance can be quantified by comparing the amount of speech data required for correct identification. The amount of SID-usable speech as input was approximately 30% less than entire input data with a considerable enhancement in SID performance.

1. INTRODUCTION

Speaker identification plays an important role in electronic authentication. In an operational environment speech is degraded by many kinds of interferences. The interference can be classified broadly as stationary or non-stationary. Stationary interference is noise which can be dealt with by using de-noising and noise reduction techniques; whereas non-stationary interference could be speech from a different speaker. Such interference is a common occurrence and the corrupted speech is known as co-channel speech [1]. Traditional methods of co-channel speech processing have been to enhance the prominent speaker (target), suppress the interfering speaker speech or both. However, previous studies on co-channel speech have shown that it is desirable to process only portions of the co-channel speech which are minimally degraded [2]. Such portions of speech considered usable for speaker identification are referred to as "usable speech". Significant amount of research has been conducted in finding speech features that would yield maximum information about the identity of the speakers, thereby increasing the accuracy of the SID system. With only usable portions of speech being input to the SID system, there is an increase in the accuracy of speaker identification [3]. A number of usable speech measures for use in the usable speech extraction system, have been developed [4] [5] [6]. These measures are based on the Target-to-Interferer Ratio (TIR) of a frame of speech with a 20 dB TIR threshold to classify usable speech [7]. The usable speech concept has been incorporated for speaker recognition improvement by silence removal [8] and multi-pitch tracking algorithm.

Usable speech is application dependent, i.e. speech segments that are usable for speech recognition may not be usable for speaker identification and vice versa. In this paper, the research presented shows that the SID system itself is used to determine the usability of the speech being input to the system. This intuitive SID application dependent definition for usability in speech is termed as *SID-usable speech*. In an operational environment, it is required that there must be some way to identify SID-usable speech frames prior to being input into the SID system, i.e. have a preprocessor block of SID-usable speech extractor before the SID process. Thus, this paper presents the development of the criteria for the detection of speaker identification (SID)-usable speech segments.

2. BACKGROUND

A brief background to the speaker identification system along with SID-usable speech labeling is given in the following section. The preprocessor SID-Usable Classification systems are presented in Section 3 and the experimental evaluation of an SID system is presented in Section 4.

2.1 Vector Quantization

The SID system used in the experiments outlined below uses a vector quantization classifier to build the feature space and to perform speaker classification [10]. The LPC-Cepstral coefficients are used as features with the Euclidean distance between test utterances and the trained speaker models as the distance measure. Although the SID system used is not the latest in advancement, it serves the purpose in illustrating the concept of SID-usable speech. One would expect the same approach to work at least as well and possible better on a more advanced SID system. A vector quantizer maps k-dimensional vectors in the vector space R_k into a finite set of vectors $Y = \{y_i: i = 1, 2... N\}$ Each vector y_i is called a *codeword* and the set of all the codewords is called a *codebook.* In this system the 14^{th} order LPC-cepstral feature space is clustered into 128 centroids during the training stage which is referred as the codebook.

2.2 Distances from Speaker Models

During the testing stage, the test utterance is divided into 'n' frames and the euclidean distance of the features of 'n' frames for 'm' trained speaker models is determined. For each speaker model, the minimum distance obtained from the codewords is considered as the distance from the model and an (m x n) classification matrix is

obtained. For simple understanding of the concept, consider SID system trained on just two speakers and tested on one of the speakers. The same concept can be then expanded to 'm' speakers. One can expect to have two histograms of the distances with significant differences between them as shown in Figure 1.

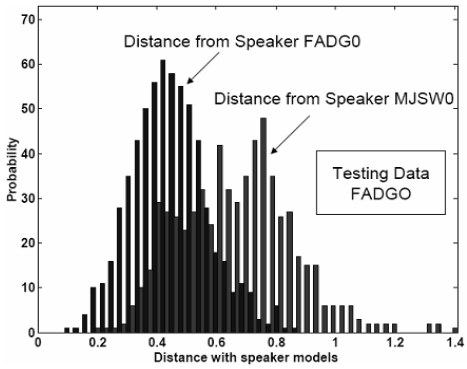

Figure 1: Histograms of Distances obtained from the Classification Matrix

The histogram with a lower mean value corresponds to the identified speaker. In this case, the left histogram corresponds to the identified speaker. It should be noted that there exists a good number of frames which have equal distances for each model. It is easy to realize that such frames contribute minimally to the speaker identification process, and might even degrade the operation with a multispeaker trained system.

2.3 SID-Usable Speech Labeling

With the knowledge of the frame distances from the speaker models, a frame of speech is defined as usable or unusable for SID system. The criterion used is the minimum of the distances from the different speaker models, and if it corresponds to the correct speaker, the frame can be termed as SID-usable. From the classification matrix, the speech frames are categorized into two classes and are labeled as '1' (usable) and '0' unusable. The labeling is done based on the following criterion:-

$$\text{Label} = \begin{cases} 1, \min(Di) = d(m,i) \\ 0, \min(Di) = d(m,i) \end{cases}$$

m = speaker index, i = frame index, d = classification matrix (m x n). In other words the criterion can be cited as: a frame of speech is considered to be usable if it yields the lowest similarity measure with the correct speaker and hence aids the speaker identification process, otherwise it is considered unusable.

3. SID-USABLE SPEECH CLASSIFICATION

Once the SID-usable speech segments are defined, it is intended to develop a method to identify usable speech segments beforehand and use them for speaker identification. Now that we have the ground truth of the real SID-usable speech (labeled SID-usable speech), one approach is to treat the identification as a classification problem and apply principles of pattern recognition and speech classification. Two classification methods to accomplish this are presented here.

3.1 Sinusoidal Model-based Classifier

Using the very well known signal analysis method of Fourier Series, the speech signal x(n) is modeled as the sum of a small number of sinusoids with time-varying amplitudes and frequencies in the presence of noise z(n).

$$x(n) = \sum_{i=1}^{p} A_i e^{j(2\pi f_i n + \phi_i)} + z(n) \quad \text{..... (1)}$$

Equation (1) above is the basic sinewave model that can be thought of as speech-independent, i.e., the model can be applied to any signal [11]. The main step in sinusoidal modeling of speech is to develop a robust procedure for extracting the amplitudes and frequencies of the component sinewaves from the speech waveform.

The ESPRIT algorithm is a signal-subspace based frequency estimation technique that is built upon the principle of eigendecomposition and it also exploits the principle of signal subspaces [12], [13]. This method is based on the decomposition of a vector space of a noisy signal by applying the eigendecomposition to the correlation matrix. The information about the harmonics can be obtained from the ESPRIT short-time spectral envelope of the speech signal. Usable speech by definition is harmonic and has a definite structure, whereas unusable speech is noise-like and lacks any kind of definite structure. The harmonic information obtained by ESPRIT algorithm is used to classify speech into usable or unusable [3]. This requires no prior knowledge of the TIR, as the real SID-usable data obtained from the SID-usable speech labeling described in the previous section is used as ground truth for SID-usable speech classification.

3.1.1. Experimental Setup and Results

Speech data from the TIMIT database was used for all the simulation experiments. Although there are more effective corpora for the evaluation of Speaker Recognition systems, the use of TIMIT database here is merely for illustration purposes. The SID-usable speech extraction concept can be applied to any currently used speech database with comparable results. The database contained ten utterances for each speaker. Forty-eight speakers were chosen spanning all the dialect regions with equal number of male and female speakers. Of the ten utterances, four utterances were used for training of the SID system. The system was tested on the remaining six utterances and the corresponding classification matrices were saved. The speech data was labeled using the classification matrix for frames of speech, 40 ms (320 samples) long to serve as ground truth for the SID-usable speech classification. The labeled data from the forty-eight speakers was used to train and test the preprocessing systems. Now, the proposed sinusoidal model-based usable speech classifier was trained for the optimum threshold using the labeled subset data of thirty-six speakers. Then, the using this threshold, the classification matrix was formed exactly the same way as the real SID-usable data classification matrix was formed with the usable (= 1) and unusable (= 0) labels. For this testing phase, the remaining twelve speakers were used. For performance evaluation of the SID-usable speech classifier, the hits and misses were calculated by comparing with the labeled SID-Usable data. The confusion matrix is constructed as given below.

$$\text{Confusion Matrix} = \begin{pmatrix} 0.78 & 0.22 \\ 0.38 & 0.62 \end{pmatrix}$$

The rows of the confusion matrix represent the actual classes of and the columns represent the identified classes. From the confusion matrix, the percentage of hits in identifying the SID- usable speech frames is 78% and false identification rate is 22%.

3.2 Support Vector Machines (SVM) Pattern Classifier

Patter recognition approach involves 2 major steps, the feature extraction and classification. For the classification of SID-usable and SID-unusable speech frames, the MFCC coefficients of each speech frame are used as features. The coefficients are obtained by 12th order Mel scale LP analysis. Classification is performed based on the support vector machines (SVM) classifier as it is the best binary classifier currently known [14]. The advantage of support vectors is the ability to incorporate non-linear relations between various parameters.

SVM relies on preprocessing the data to represent patterns in a high dimension which is typically higher than the original feature space. With an appropriate nonlinear mapping to a sufficiently high dimension, data from 2 categories can always be separated by a hyperplane. Consider a simple example of an artificial feature set of 2D points, together with the class labels (ground truth) +1 (usable) and 0 (unusable).

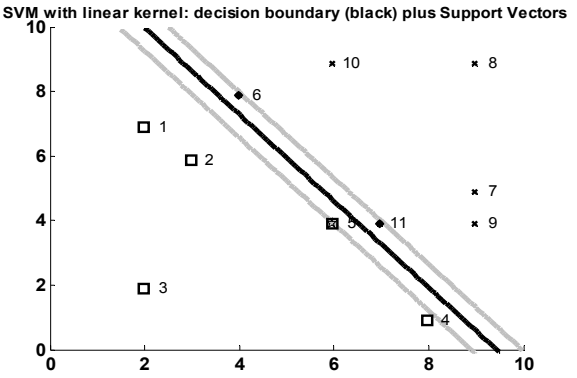

Figure 2: 2D plot of the artificial feature set along with the optimum decision boundary (hyperplane) and the support vectors.

The data is plotted in Figure 2, where 'squares' represent one set of feature points with label Yi = +1 and 'crosses' stand for points with label Yi = 0. Now the support vector machine cassifier is trained on this feature set. Training of a support vector machine consists of finding the optimal hyperplane, that is, the one with the maximum distance from the nearest training patterns. The support vectors are those (nearest) patterns, a distance 'b' from the optimal hyperplane. The *support vectors* are the (transformed) training patterns that are close to the hyperplane. The contour plotted in black separates class +1 from class 0 (*this is the actual decision boundary*). The contour plotted in 'bold' 'squares' are the points at distance +1 from the decision boundary and the contour plotted in 'bold' 'crosses' are the points at distance -1. The main goal in training a support vector machine is to find the separating hyperplane with the largest margin; it is expected that the larger the margin, the better generalization of the classifier. Thus support vectors are the training samples that define the optimal separating hyperplane and are the most difficult patterns to classify. But, they are the patterns most informative for the classification task. The classification is performed with the real SID-usable speech (i.e. the labeled SID-usable data from the SID system itself) as ground truth target variables or class labels.

3.2.1 Experimental Setup and Results

Training and testing data are same as described in Section 3.1.1. The training stage of the MFCC-SVM classifier involves the computation of the MFCC coefficients for each frame. These are then used to form the 12 dimension feature vector for which the support vectors are computed during the training stage of the support vector machines. These support vectors are used at the testing phase to perform the SID-usable and unusable speech classification.

To analyze the classification obtained by SVM classifier, the confusion matrix was constructed and is given below. The classification results displayed here are for only a small subset of the testing dataset. Further experiments are to be performed using the RBF kernel to form the SVM classified SID-usable speech database.

$$\text{Confusion matrix} = \begin{pmatrix} 0.73 & 0.27 \\ 0.36 & 0.64 \end{pmatrix}$$

From the confusion matrix, the percentage hits in identifying the usable speech frames is 73% and false identification rate as 27%.

3.3 Comparison and discussion

The results obtained from the two classification methods are compared with the previously developed Weighted k-NN SID-usable speech classifier in the Table 1 below [9].

Table 1: Performance evaluation of the SID-Usable Speech Classifiers

	Weighted k-NN Classifier	Sinusoidal Model Classifier	SVM Pattern Classifier
Usable Hits	76%	78%	73%
Usable Misses	23%	22%	27%
Unusable Hits	68%	62%	64%
Unusable Misses	32%	38%	36%

From Table 1, it can be noted that the Sinusoidal Model SID-Usable speech measure performs better than weighted k-NN classifier for usable speech detection. Whereas, there is still need for further improvement in the performance of SVM pattern classifier.

4. SPEAKER IDENTIFICATION IMPROVEMENT

One would expect the performance of speaker identification to be higher if only the usable speech frames (SID-usable speech frames) are identified in a pre-processing system and then used for speaker identification. This suggests that the SID system should be tested with only the frames labeled as SID-usable. After performing SID-usable speech classification using the proposed classifiers, a SID-usable speech database was formed. SID-usable speech obtained from k-NN classifier was also included in the SID-usable speech database and used for SID performance evaluation [12]. There were 4 such databases formed for the entire TIMIT database: real usable data (from SID system), k-NN SID-usable data, sinusoidal model SID-usable data and SVM classified SID-usable data. For training phase, entire TIMIT database was used. The database contains ten

utterances for each speaker. Forty-eight speakers were chosen spanning all the dialect regions with equal number of male and female speakers. Of the ten utterances, four utterances were used for training the SID system.

For the testing phase, 5 testing datasets generated (4 usable speech databases and the original entire speech TIMIT database) were used. The system was tested on remaining six utterances to obtain the SID performance metric in percentage accuracy of speaker identification. The bar chart in Figure 3 below shows the results of the SID system performance for the different SID-usable speech databases.

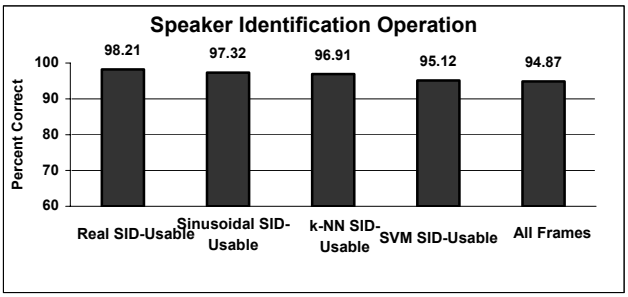

Figure 3: SID performance comparison with the different generated SID-usable speech data.

From Figure 3, by using only SID-usable speech, the SID system has a better performance. The amount of real SID-usable speech was approximately 30% less than all frames data without the SID system performance being compromised. Moreover, the performance of the SID system is better for the input SID-usable data obtained from the SID-usable speech classifiers than the all frames data.

5. CONCLUSION

In this paper, usability in speech, with reference to speaker identification, which is called *SID-usable speech,* was presented. Here the SID system was used to determine those speech segments that are usable for accurate speaker identification. Two novel approaches to identify SID-usable speech frames were presented which resulted in 78% and 72% correct detection of SID-usable speech. We have shown that SID performance can be quantified by comparing the amount of speech data required for correct identification. The amount of SID-usable speech was approximately 30% less than entire input data without the SID system performance being compromised. Therefore, it can be concluded that using only SID-usable speech improves the speaker identification performance.

6. ACKNOWLEDGEMENTS

The Air Force Research Laboratory, Air Force Material Command, and USAF sponsored this effort, 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.

7. DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of 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. Lovekin, R. E. Yantorno, S. Benincasa, S. Wenndt and M. Huggins, "Developing usable speech criteria for speaker identification," *Proc. ICASSP* 2001, pp. 421-424, 2001

[3] S. Khanwalkar, B. Y. Smolenski, R. E. Yantorno, "Speaker Identification Enhancement under Co-Channel Conditions using Sinusoidal Model based Usable Speech Detection", IEEE international Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2004).

[4] K. Krishnamachari & R. E. Yantorno, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," IEEE Inter. Symposium on Intelligent Signal Processing & Comm. Systems, pp. 710–713, Nov 2000.

[5] J. M. Lovekin, K. R. Krishnamachari, and R. E. Yantorno, "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, pp. 139–142, Nov 2001.

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

[7] R. E. Yantorno, "Co-channel speech and speaker identification study," Tech. Rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome labs, New York, 1998.

[8] J. K. Kim, D. S. Shin, and M. J. Bae, "A study on the improvement of speaker recognition system by voiced detection," *45th Midwest Symposium on Circuits and Systems, MWSCAS*, vol. III, pp. 324–327, 2002

[9] A. N. Iyer, B.Y Smolenski, R. E. Yantorno, J. Cupples, S. Wenndt, "Speaker Identification Improvement Using The Usable Speech Concept," *European Signal Processing Conference (EUSIPCO 2004)*

[10] F. K. Soong, A. E. Rosenberg and B. H. Juang, "Report: A vector quantization approach to speaker recognition," *AT&T Technical Journal*, vol. 66, pp. 14–26, 1987.

[11] R. J. McAulay, T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP–34, No. 4, pp. 744-754, August 1986.

[12] R. Roy, A. Paulraj and T. Kailath (1986). "ESPRIT: A Subspace Rotation Approach to Estimation of Parameters of Sinusoids in Noise," IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-34, pp. 1340–1342, October

[13] G. D. Manolakis, K. V. Ingle and M. S. Kogan, *Statistical and Adaptive Signal Processing: Spectral Estimation, McGraw-Hill Science/Engineering/Math (December 1999)*

[14] V.N.Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, New York, ISBN 0-387-94559-8, 1995.