SPEAKER IDENTIFICATION IMPROVEMENT USING THE USABLE SPEECH CONCEPT

A. N. Iyer[†], *B. Y. Smolenski*[†], *R. E. Yantorno*[†], *J. K. Shah*[†], *E. J. Cupples*[‡] *and S. J. Wenndt*[‡]

[†]Speech Processing Lab, Temple University
12th & Norris Streets, Philadelphia, PA 19122

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

ABSTRACT

Most signal processing involves processing a signal without concern for the quality or information content of that signal. In speech processing, speech is processed on a frame-by-frame basis, usually only with concern that the frame is either speech or silence. However, knowing how reliable the information is in a frame of speech can be very important and useful. This is where usable speech detection and extraction can play a very important role. The usable speech frames can be defined as frames of speech that contain higher information content compared to unusable frames with reference to a particular application. We have been investigating a speaker identification system to define usable speech frames and then to determine a method for identifying those frames as usable using a different approach. We present a simple and intuitive definition and two methods to identify the defined usable speech frames, which have resulted 78% and 68% success rates.

1. INTRODUCTION

In an operational environment speech is degraded by many kinds of interferences. The operation of many speech processing techniques are plagued by such interferences. Usable speech extraction is a novel concept of processing degraded speech data. The idea of usable speech is to identify and extract portions of degraded speech that are considered useful for various speech processing systems. Yantorno [1] performed a study on co-channel speech and concluded that the Target-to-Interferer Ratio (TIR) was a good measure to quantify usability for speaker identification. However, the TIR is not an observable value[1] from the co-channel speech data. A number of methods termed *usable speech measures* which are indicators to the TIR have been developed and studied under co-channel conditions [2, 3, 4, 5, 6]. These measures are used as features in decision fusion systems to make an overall decision [7, 8]. On similar lines the effects of silence removal on the performance of speaker recognition were studied in [9].

In all of the above methods mentioned, usability in speech is considered to be application independent. However the concept of usable speech by definition is application dependent, i.e. usable speech for speech recognition may not be usable for speaker identification and vice versa. In this paper we present an intuitive application dependent definition to usability in speech, with reference to speaker identification and term it as *SID-usable speech*. The speaker identification system under study uses the LPC cepstal features of 14 dimensions and a vector quantizer with 128 codebooks. We also present two system which serves as a preprocessor to the speaker identification system, to identify and extract the SID-usable speech frames. A block diagram of the application of usable speech for speaker identification is illustrated in figure 1.

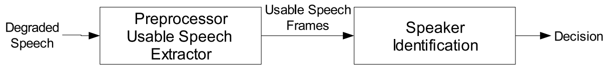


Figure 1: Block diagram of usable speech processing for speaker identification.

2. BACKGROUND

A brief background to the speaker identification system is given in the following section. The usable speech is defined in section 3. The preprocessor systems are presented in section 4 and their experimental evaluation is presented in section 5.

2.1 Vector Quantization

The speaker identification system, under study uses a vector quantization classifier to build the feature space and to perform speaker classification [10]. The LPC-Cepstrum is used as features with the Euclidean distance between test utterances and the trained speaker models as the distance measure. 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 14th order LPC-Cepstral feature space is clustered into 128 centroids during the training stage which is referred as the codebook.

2.2 Study of Distances from Speaker Models

Consider the testing stage in which the test utterance is divided into '*n*' frames and the Euclidean distance of the features of '*n*' frames with '*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. Without loss of generality, consider a system trained with two speakers and tested on one of the speakers. This two speaker system provides a simple approach to

[1] TIR is not measurable from the signal as co-channel data recorded over a single microphone are considered.

better understanding how the system functions and to be able to interpret the results due to its simplicity. One can expect to have two distributions of the distances with significant difference in the expected values as shown in figure 2. The distribution with a lower mean value corresponds to the identified speaker. It should be pointed 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 multispeaker trained system!

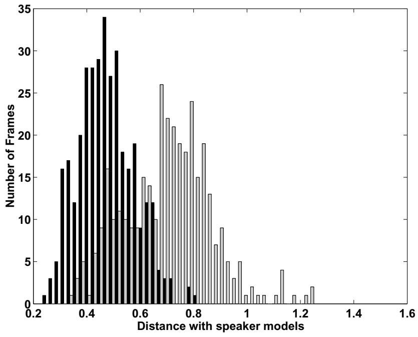


Figure 2: The histogram of the distances obtained from the classification matrix.

3. USABLE SPEECH DEFINITION

With the knowledge of the frame distances from the speaker models, a frame of speech can be defined as usable in different ways. The simplest method is to look at the minimum of the distances from different speaker models, and if it corresponds to the correct speaker, the frame can be termed as usable. From the classification matrix the speech frames are categorized into two classes and are labeled as “1” (usable) and “0”(unusable). The labelling is done based on the following criterion –

$$\phi_m(i) = \begin{cases} 1, & \min(\mathbf{D}_i) = d(m,i); \\ 0, & \min(\mathbf{D}_i) \neq d(m,i). \end{cases} \quad (1)$$

where m is the speaker index, i is the frame index, D_i is the vector consisting of distance between frame i and the trained speaker models and d is the classification matrix. In other words, the criterion can be cited as: a frame of speech is considered to be usable if it yields the smallest distance measure with the correct speaker and hence aids in the speaker identification operation, else it is considered unusable. One would expect the performance of speaker identification would be higher if only the usable speech frames are identified in a preprocessor unit and fed into the speaker identification system. Figure 3 shows the labelled speech data. The data labelled as usable is represented in gray and the unusable is represened in black. Note that it is hard to visually draw any conclusions regarding the two classes of data.

3.1 Speaker Identification Performance Metric

The speaker identified corresponds to the model which has the smaller mean value μ_c of the distances. If the next best chosen model has a mean value of μ_{c-1}, the difference between the mean values of the best two speaker models chosen by test speech data serves as a metric to quantify the speaker identification performance.

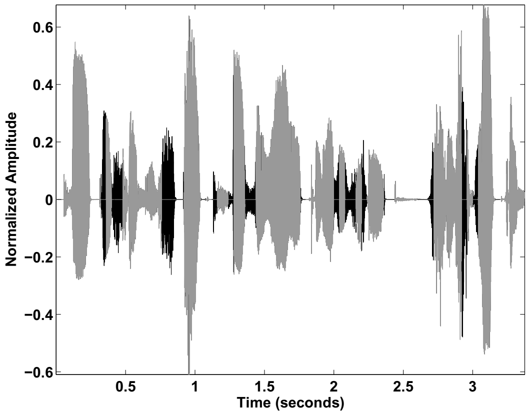


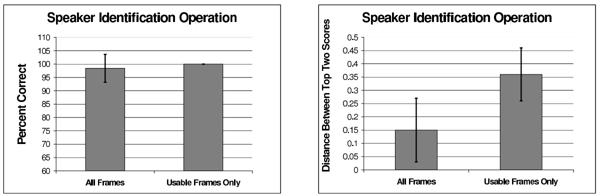
Figure 3: Labeled speech data: Usable speech is represented in gray and the unusable is represented in black.

$$\mathbf{P}_1 = \mu_c - \mu_{c-1} \quad (2)$$

It would be evident that the speaker identification performance had improved if the value of the metric is higher.

The performance of speaker identification can also be quantified by comparing the amount of speech data $\mathbf{P_2}$ (secs) required for correct identification, i.e., if less speech data is needed for good identification.

To realize these performance metrics, speaker identification experiments were performed with *a priori* knowledge of the speakers. The speaker identification system was trained on two speakers and tested on one of the speakers resulting in a collection of usable frames. The defined SID-usable data was used to test the speaker identification performance. The performance was compared for two scenarios, 1) utterances having a length equal 2 seconds and 2) usable speech segments, of average length 1.4 seconds. Data from the TIMIT database with twenty-four speakers was used for the speaker identification operation experiments and the results were analyzed and are presented in Figure 4.



(a) (b)

Figure 4: Speaker identification performance comparison with speech data and extracted usable frames. a) percentage accuracy in speaker identification and b) difference in distance ($\mathbf{P}_1$) between the best two speakers selected. Note - black vertical lines are standard error bars.

The system was succesively trained with two utterances accounting all combinations of male / female speakers and tested on a total of 384 utterances. The values represented in the chart are the average values over all the test utterances.

Observing Figure. 4 it can be noted that by using only usable speech segments, the speaker identification system has higher performance with respect to both the metrics based on five different pieces of information. First. the average difference between the best two scores is higher with usable speech case. Second, the amount of usable speech was approximately 30% less than the all frames data without the systems performance being compromised. Third, the standard deviation of the usable speech difference scores were smaller, indicating a higher confidence level in the identified speaker. Fourth, for the usable speech case the percent correct was 100% versus 94% for the all frames case. Fifth, the standard error for the percent correct is zero as compared with for all frames condition. Therefore, it can be concluded that using only usable speech improves the speaker identification performance significantly.

4. USABLE SPEECH IDENTIFICATION

In an operational environment it will be essential that there must be some way to identify SID-usable speech frames prior to being input into the speaker identification process. Two methods to accomplish this are presented here. The weighted *k*-NN is used as a blind system performing classification. The use of speech features with decision tree algorithms is motivated by the fact that certain classes of speech contain more information compared to the others.

4.1 Weighted *k*-NN Pattern Classifier

The *k*-Nearest Neighbor rule [11] is a very intuitive method that classifies unlabelled samples based on their similarity with samples in the training set. The *a posteriori* class probabilities $P(\omega_i|\mathbf{x})$ of test vector $\mathbf{x}$ for the usable and unusable classes $\{\omega_i; i = 1,2\}$ is determined by

$$P(\omega_i|\mathbf{x}) = \frac{1}{d_i}.\frac{k_i}{k}.p(\omega_i) \tag{3}$$

That is, the estimate of the *a posteriori* probability that $\mathbf{x}$ belongs to class ω_i is merely the fraction k_i of the samples within the *k*-nearest neighbors, that are labelled ω_i and weighed inverse proportionally to the average similarity measure d_i with each class samples. Further it is weighed with respect to the class probabilities $p(\omega_i)$. Usually for an even class problem, k is chosen to be odd to avoid a clash. The *k*-NN rule relies on the proximity measure and the Euclidean distance is between the 14th order LPC-Cepstrum coefficients of the test pattern and the training templates was considered. The value of k was chosen as 9, as it resulted in reasonable classification results.

4.2 Decision Trees

Prior studies [12] have shown unvoiced frames of speech do not contribute significantly to speaker identification. This study is to determine if there exists a relationship between speech classes and their contribution to speaker identification. For example, some classes of speech might not help the speaker identification process such as nasals which have zeros and hence would not give satisfactory results in speaker identification, because the features used by the SID are based on the autoregressive. The problem addressed in the next section can be summarized as follows Identify speech classes from speech data and study the relation between speech classes and their contribution to speaker identification.

4.2.1 Speech Feature Detectors

Acoustic feature detection is the search for different (acoustic) features. Examples of acoustic features include voicing, nasality and sonorance. While acoustic features are used to differentiate between various segment categories, for example, nasality may indicate the presence of nasal, or it may indicate the presence of nasalized vowel. Eight feature detectors were used in this research, which includes sonorant, vowel, nasal, semivowel, voice-bar, voiced fricative, voiced stop and unvoiced stop. Together with the feature detectors, spectral flatness value was also considered which gives a voiced/unvoiced decision. The computation of most feature detectors is based on a volume function. The volume function represents the quantity analogous to loudness, or acoustic volume of the signal at the output of a hypothetical band-pass filter. The volume function can be computed using the following equation [13].

$$\mathbf{VF}(i) = \frac{1}{\mathbf{N}_i}\sqrt{\sum_{m=\mathbf{A}}^{\mathbf{B}} \left|\mathbf{H}_i(e^{j\pi\frac{m}{256}})\right|^2} \tag{4}$$

where i is the current frame index, N_i is the number of samples, $\mathbf{A}$ is the index of low cutoff frequency and $\mathbf{B}$ is the high cutoff frequency. Each feature detection algorithm computes a feature value, which is a ratio of volume functions computed in two frequency bands. The feature values are converted into a decision based on fixed thresholds to indicate the presence of the corresponding feature in a given frame of speech [13].

With the feature decisions, the class can be classified through a sequence of questions, in which the next question asked depends on the answer to the current question. This approach is particularly useful for such non-metric data, since all of the questions can be asked in a “true/false” and does not require any notion of a distance measure. Such algorithms build a decision tree based on the entropy or the information content of each feature. The traditional C4.5 algorithm [14] was used for this work.

5. EXPERIMENTS AND RESULTS

A subset of speech data from the TIMIT database was used for all the experiments. The experiments were designed to use all the speech files for each speaker. 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 speaker identification system. Then the system was tested on the remaining six utterances and the corresponding classification matrices were saved. The speech data were labeled using the classification matrix and equation given in section 3 for frames of speech, 40ms long.

The labeled data from the forty-eight speakers was used to train and test the preprocessing systems. A subset of thirty-six speakers were used to train the *k*-NN pattern classifier and the decision tree algorithms. The data from the rest twelve speakers were used for testing and performance evaluation of the preprocessing systems. The performance of the systems are tabulated in table

Table 1: Performace evaluation of the preprocessing systems in identification of usable speech frames.

	Weighted k-NN	Decision Trees
Usable hits	77.64%	67.99%
Usable miss	22.36%	32.01%
Unusable hits	67.99%	56.59%
Unusable miss	32.01%	43.41%

In table 1, a hit is defined as the number of usable frames identified as correctly by the method and a miss is defined as the number of usable frames declared as unusable.

5.1 Speaker Identification Improvement

The next step in using the usable speech concept for speaker identification is to evaluate the speaker identification performance with the preprocessor unit. The training and testing data used for this purpose are the same as described in section 5. However, the *a priori* knowledge of the speakers identity is ignored and the usable speech frames are extracted using the schemes described in sections 3.1 and 3.2. The speaker identification system was successively trained using four training utterances and tested with utterances from one of the speakers. The result of correct identification of speakers with the weighted k-NN scheme was 97% and with the decision tree scheme was 95%. These results can be compared to 94% correct identification without the preprocessor system.

6. CONCLUSIONS

A method to label frames of speech as SID-usable or SID-unusable is defined. Two methods to identify the defined SID-usable speech segments are also developed, from the areas of pattern recognition and data mining. We have shown an 50% reduction in speaker identification errors by using the usable speech concept. As a next step, various other classification schemes are being developed and a general definition for all applications is derived.

7. 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. Further we wish to thank Rajani Smitha for performing the speaker identification experiments.

8. 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. 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] B. Y. Smolenski and R. E. Yantorno, "Fusion of usable speech measures using quadratic discriminant analysis.," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2003*, 2003.

[8] J. K. Shah, B. Y. Smolenski, and R. E. Yantorno, "Decision level fusion of usable speech measures using consensus theory," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2003*, 2003.

[9] 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.

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

[12] 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.

[13] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, Wiley, New York, 1999.

[14] R. Quinlan, "Discovering rules from large collections of examples: a case study," *Expert Systems in the Micro-electronic Age, Edinburgh University Press, Edinburgh*, pp. 168–201, 1979.