

SPEAKER IDENTIFICATION IMPROVEMENT USING THE USABLE SPEECH CONCEPT

A. N. Iyer, B. Y. Smolenski, R. E. Yantorno

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

J. Cupples, S. Wenndt

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 identify usable speech frames and then to determine a method for identifying those frames as usable using a different approach. A 100% accuracy can be achieved in speaker identification by using only the extracted usable speech segments.

1. INTRODUCTION

Usable speech by definition is application dependent, i.e. usable for speech recognition may not be usable for speaker identification and vice versa. A number of usable speech measures independent of any application for use in the usable speech extraction system, have been developed [1] [2] [3] [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 [9]. What is presented here is a paradigm shift related to the determination of usable speech. In this paper we present a study of the speaker identification system and the development of criteria for the determination of speaker identification (SID)-usable speech segments. In an operational environment the knowledge of which frames of speech are usable will not be known and hence an usable speech identification system is presented to identify SID-usable speech. This system serves as a preprocessor to the speaker identification process. A brief background to the speaker identification system and the usability criteria is elaborated in the next section.

2. USABLE SPEECH FOR SPEAKER IDENTIFICATION

2.1. Vector Quantization

The speaker identification 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-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 $Y = \{y_i: i = 1, 2, \dots, N\}$. 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. The system was trained with two speakers and tested on one of the speakers. This two speaker system provides a simple approach to better understanding how the system functions and to be able to interpret the results without any oversights or limitations 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 1. The left distribution 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!

2.3. Usable Speech Labelling

Once the distances are obtained, a frame of speech can be defined as usable in different ways. The simplest method

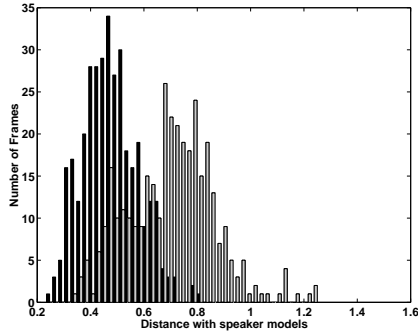


Fig. 1. The histogram of the distances obtained from the classification matrix.

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 front-end unit and fed into the speaker identification system. A set of experiments were performed on the speaker identification system with only the frames labeled as usable and hence validate the above statement.

2.4. Speaker Identification Performance Metric

The difference between the distances of the best two speaker models chosen by test speech data serves as a metric to quantify the speaker identification performance. 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 required for correct identification, i.e., if less speech data is needed for good identification. The speaker identification system was trained on two speakers and tested on one of the speakers resulting in a collection of usable frames. The identified 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 seg-

ments, 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 2.

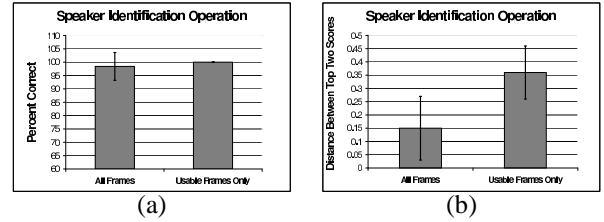


Fig. 2. Speaker identification performance comparison with speech data and extracted usable frames. a) percentage accuracy in speaker identification and b) difference in distance between the best two speakers selected. Note - black vertical lines are standard error bars.

The system was trained with all combinations of male / female speakers and a total of 384 testing utterances were utilized. The values represented in the chart are the average values over all the test utterances.

Observing Figure. 2 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 96% 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.

3. USABLE SPEECH IDENTIFICATION

Once the usable speech segments are defined it is intended to identify usable speech segments prior to the speaker identification process. Two methods to accomplish this are presented here.

3.1. 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

$$P(\omega_i|\mathbf{x}) = \frac{1}{d_i} \cdot \frac{k_i}{k} \cdot p(\omega_i) \quad (2)$$

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.

3.1.1. Experimental Setup and Results

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 2.3 for frames of speech, 40ms long. The labeled data from the forty-eight speakers was used to train and test the preprocessing systems. The training stage of k-NN pattern classifier involved computation of LPC-Cepstrum and these instances were saved and were used to determine the nearest neighbors during the testing phase. The classifier performance was computed from the confusion matrix constructed and is given below.

$$\text{Confusion matrix} = \begin{bmatrix} 0.7764 & 0.2236 \\ 0.3201 & 0.6799 \end{bmatrix}$$

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. 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.

3.2.1. 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 are 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 is 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].

$$\mathbf{VF}(i) = \frac{1}{N_i} \sqrt{\sum_{m=\mathbf{A}}^{\mathbf{B}} |\mathbf{H}_i(e^{j\pi \frac{m}{256}})|^2} \quad (3)$$

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 [14].

3.2.2. Experimental Setup and Results

Train and test data are described in Section 3.1.1. Nine speech features are computed for each frame of speech and the corresponding feature scores are computed. The training data is used in the inductive learning procedure to create the decision tree. The classification performance of the decision tree created is evaluated based on the confusion matrix computed and presented below.

$$\text{Confusion Matrix} = \begin{bmatrix} 0.6799 & 0.3201 \\ 0.4341 & 0.5659 \end{bmatrix}$$

The percentage of hits in identifying the usable speech frames is 68% and false identification rate is 32%.

4. DISCUSSION

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. The decision tree approach is speaker independent as the features used are speech dependent and not speaker dependent. Next step in this direction is to study the speaker identification system with different train and test conditions using the SID-usable speech frames. Various other classifiers such as Support Vector Machines are also being investigated for performing the classification task.

5. 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.

6. 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.

7. REFERENCES

- [1] K. R. Krishnamachari and R. E. Yantorno, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions.," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 710–713, Nov 2000.
- [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 ratio using the lpc residual," *4th IASTED International Conference Signal and Image Processing*, pp. 146–150, 2002.
- [4] 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.
- [5] 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.
- [6] 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.
- [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] Y. Shao and D-L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 205–208, 2003.
- [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 Acoustics 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.