T. Dutta, 'Text Dependent Speaker Identification Based on Spectrograms', *Proceedings of Image and Vision Computing New Zealand 2007*, pp. 238–243, Hamilton, New Zealand, December 2007.

TEXT DEPENDENT SPEAKER IDENTIFICATION BASED ON SPECTROGRAMS

Tridibesh Dutta

Indian Statistical Institute, Kolkata, India
Email: tridibesh@gmail.com

Abstract

The goal of this paper is to study a new approach to text dependent speaker identification using spectrograms. This, mainly, revolves around trapping the complex patterns of variation in frequency and amplitude with time while an individual utters a given word through spectrogram segmentation. These segmented spectrograms are used as a database to successfully identify the unknown individual from his/her voice. This is a notable drift from the usual Gaussian Mixture Models (GMM) and the Vector Quantization (VQ) techniques for text-dependent speaker identification/verification. The features used for identifying, rely on optimal spectrogram segmentation and the euclidean distance of the distributional features of the spectrograms of the unknown voice with that of a given known speaker in the database. Performance of this novel approach on a sample collected from 40 speakers show that this methodology can be effectively used to produce a desirable success rate.

Keywords: Speaker recognition, Spectrograms, Image matching, Pattern recognition, Image segmentation

1 Introduction

The process of automatically recognizing who is speaking by distinguishing qualities in a speaker's voice is called speaker recognition. For this purpose, it is important to preserve the speaker specific information in the speech signal. Human voice has lots of variations termed as intra-speaker variability. Variations in voice 'in between' speakers is called inter-speaker variation. According to the relevance to the content of speech, the speaker recognition task could be divided into 'text independent' and 'text dependent'.

Moreover, the text-dependent speaker identification can be subdivided into two further categories, closed-set and open-set problems. The closed set text-dependent speaker identification problem may be stated as follows. Out of a total population of N 'known' speakers, find the speaker whose reference pattern has closest resemblance to the sample pattern of the 'unknown' speaker who is assumed to be one of the given set of speakers. In the open set problem, a reference model for an unknown speaker may not exist. In this situation, an additional decision alternative, that the unknown does not match any of the models, is required. This speaker verification (in an open set) task is a hypothesis testing problem where the system has to accept or reject a claimed identity associated with an utterance. Since most of todays systems are based on probability calculations, two types of erroneous decisions may occur in speaker verification. A *false acceptance* is said to occur when an impostor is accepted, while a *false rejection* occurs when the system rejects a true client. There is a trade-off between these two error types. If safety is emphasized, the false rejection rate will have to increase in order to keep the false acceptance rate low. But if the system produces too many false rejections, users may find the system annoying. One common choice is to put the false acceptance and false rejection rates equal, aiming for the equal-error-rate (EER) [1].

In this paper, text-dependent speaker identification for both the closed set and open set problems have been studied with. In the proposed method, speaker identification is carried out by means of speech spectrograms. The essence of this technique lies in formulating the speaker-identification problem into pattern recognition of images and resolving it using machine learning tools.

Speaker Identification task includes the basic components: (I) feature extraction (II) speaker modeling (III) speaker matching and (IV) decision logic. The feature extraction module converts the raw speech waveform in the given sample to a spectrogram.

Distributional features of the spectrograms are then used to make representative codebooks of speak-

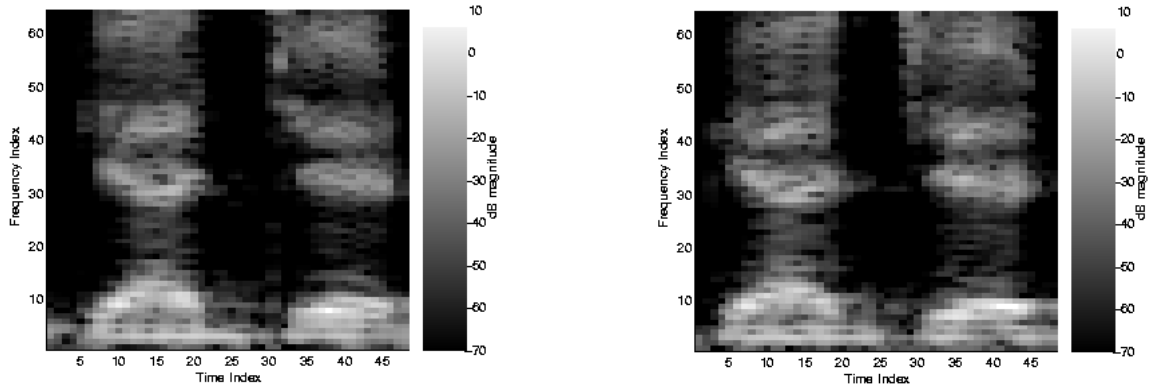

Figure 1: Depicts similarity of sample spectrograms of speech signals of 'Speaker 1' for the utterance 'gadget'.

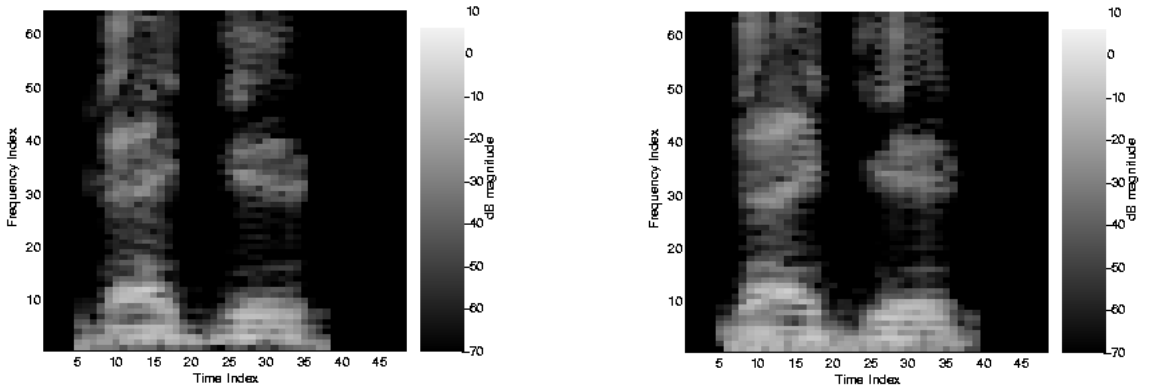

Figure 2: Depicts similarity of sample spectrograms of speech signals of 'Speaker 17' for the utterance 'gadget'.

er's voice patterns and use them to create a database. Later, when unknown samples arrive, they are used to match spectrograms from the given database. The decision logic finally makes a one-out-of-N decision, e.g. selects the speaker with maximum degree of similarity. Some of the widely used modeling techniques, such as vector quantization (VQ) [2], Gaussian mixture models (GMM) [3] and other established procedures use the same general framework for speaker identification.

A database designed for speaker identification with limited enrollment data, is used in the study. The database is collected in realistic conditions with the use of an external microphone. The database contains 40 enrolled speakers, each reciting a list of words. There are three words: 'cat', 'gadget' and 'loss'; with each enrolled speaker reciting each of the assigned words 9 times, of which 4 samples, for each word, are to be randomly chosen for training purpose and the other 5 samples for testing. The speech signals are sampled with $8 - 16kHz$. Samples from every speaker are collected in different sessions varying over time, to make our database as efficient as possible. Utterances of the speakers were recorded in a 'noise free' environment thus ensuring least loss of speaker dependent information. Also, before computation of the spectrogram, any DC offset present in the signals were removed and the signals centered around 0 vertically. The maximum amplitude of the utterances was *normalized* to $-3dB$, to ensure a fair comparison of the spectrograms.

The rest of the paper is organized as follows: in Section 2, the feature extraction and modeling is explained. The identification methods in a closed and in an open set of speakers is described in Section 3. Results are discussed in Section 4. Applications has been suggested in Section 5. Conclusions and future work has been outlined in Section 6.

2 Spectrogram Processing

It is evident from Figures 1 and 2, that the illustrated spectrograms of the utterance 'gadget' appear to be dissimilar for different speakers. Hence, an essential task of image comparison is to justify the claim. Spectrogram comparison to recognize a speaker is already an established procedure in

our text-dependent speaker identification problem. The spectrogram comparison approach for speaker identification proposed by Dutta and Basak, uses a non-parametric statistical technique namely, the Kolmogorov-Smironov test for image comparison comprising of Hollander-Wolfe statistic. In that, segmented spectrograms are compared using the cumulative distribution function of the gray-scale intensities [4,5]. Though segmentation of spectrograms has a mention in that paper [4] too, but, the optimality in the number of bands had not been dealt with.

Under assumption that images are subject to random noise, we want to test if images are the same (the speech samples are of the same speaker). We say that two images are the same if the corresponding bands of the segmented images have the same distribution properties. The choice of variable of interest to be extracted from the spectrograms is of utmost importance i.e. the variable which loses the least information about the speakers. One may choose the statistical mean (first moment), information entropy (or Shannon entropy), the second central moment, the third and so on.

Estimates of mean of a random variable is a commonly used variable to study its distribution. The Shannon entropy or information entropy is a measure of uncertainty associated with a random variable. Clearly, if two images are the same, up to a small noise, they should have "close" distributional properties. The reverse though may not true. Thus, distribution analysis is helpful when images of the same content are compared.

While the mean M of a discrete random variable X dependent on observational values $\{x_1, x_2, \ldots, x_n\}$ is given by:

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

the information entropy $H(X)$ of a discrete random variable X is defined to be:

$$\begin{aligned} H(X) = E(I(X)) &= \sum_{i=1}^{n} p(x_i) log_2(1/p(x_i)) \\ &= -\sum_{i=1}^{n} p(x_i) log_2(p(x_i)) \end{aligned} \tag{2}$$

where $I(X)$ is the self-information of X, which is itself a random variable; and $p(x_i) = P(X = x_i)$ is the probability mass function of X.

The spectrograms are partitioned into several overlapping bands having 'nearly' equal bandwidths and overlaps (approximately half of band-width), for separate processing. Given a segmentation pattern, all the spectrograms in question (each of which have the same pixel matrix size), are split in a similar fashion. The number of bands a spectrogram is segmented into along any axis, depends on the band-width and overlap. It is important to note here that, as the number of bands differ, it is not always possible to segment a spectrogram into bands having a 'equal' band-width and overlap. As a remedy, the spectrograms are split into bands having a nearly equal bandwidths and overlaps. Though, the choice of the best band-width and band-overlap selection remains to be an open problem, a good success rate and speedy completion of the test may be assumed to satisfy an optimality criterion. Results on effect of segmenting the spectrograms into bands have been provided in a later section. The motivation behind decomposition of the spectrograms lie in a higher dimension comparison of the spectral features of two different images. The choice of axis (frequency/time) along which to split the spectrogram is also important. It has been shown that splitting up frequency axis into several bands produce significantly better results than time axis segmentation, primarily, due to phase shifts if working on time domain [4]. The pixel values in these frequency bands may be interpreted as the energy content in the frequency bands (which uniquely characterizes an individual) as the speech signal is swept through time. This fact lays the basis of our methodology to verify and, more importantly, identify a speaker.

The task for speaker modeling has been formulated as follows: Let μ_{ijkr} $(i = 1, \ldots, N; j = 1, \ldots, M; k = 1, \ldots, P; r = 1, \ldots, R)$ denote the mean of pixel values (a comparison of the mean with Shannon entropy has been stated later) for replicate r corresponding to the k^{th} frequency band of the spectrogram of the i^{th} speaker's utterance of the j^{th} word. Here, N denotes the number of speakers in question; M, the number of different words uttered; P, the number of frequency bands the spectrogram is split into and R, the number of replications per word used for training corresponding to each known speaker. We use these observations, to prepare our codebook corresponding to each spectrogram. A typical codebook, corresponding to the spectrogram of r^{th} replicate of the j^{th} word of the i^{th} speaker would consist of a vector with its elements representing the means of the ordered overlapping bands of the segmented spectrogram.

3 Speaker Identification

3.1 Identification in a 'closed set'

Having collected our training database of spectrograms for 40 speakers, 4 training samples for every word for every speaker is chosen randomly to be tested with. We consider a test sample comprising of an utterance of each of the 3 words of an un-

known speaker (in the closed set). An important assumption is that, the unknown speaker is in the closed set and utters the three prescribed words in a predefined order to enable identifying which sample corresponds to which word.

Let θ represent the actual identity of the unknown speaker based on the mean pixel values of the various bands. For simplicity, let the i^{th} speaker in our database be denoted by 'Speaker i' ($i = 1, \ldots, 40$). Given codebook vector C_{ijr} representing the i^{th} speaker's, r^{th} replicate ($r = 1, \ldots, 4$) of the j^{th} word ($j = 1, 2, 3$), the minimizing value i of the Euclidean distance [6,7] from the 'unknown' speaker's codebook is a plausible solution to the speaker identification problem. Mean pixel value of a particular band of the segmented spectrograms of a specific word by a speaker does not remain the same with replications due to variation in voice. In the database samples, let μ_{ijkr} (as defined in the previous section) be the mean value corresponding to the k^{th} frequency band of the spectrogram of the i^{th} speaker's r^{th} replicate of the j^{th} word. Again, let $x_{\theta jk}$ denote the mean of the unknown speaker's k^{th} frequncy band of the spectrogram corresponding to the j^{th} word.

Given an unknown speaker with identity θ, for the i^{th} speaker and j^{th} word, let $p_{\theta|(i,j)}$ denote:

$$min_{1 \leq r \leq 4} \left\{ \left| \sum_{k=1}^{10} (x_{\theta jk} - \mu_{ijkr})^2 \right|^{0.5} \right\}$$

It is important to note here, that each spectrogram was split into 10 overlapping bands along the frequency axis. The unknown person θ is classified as the m^{th} person if:

$$\frac{1}{3} \sum_{j=1}^{3} p_{\theta|(i,j)} \tag{3}$$

achieves minimum for $i = m$, i.e. the Euclidean distance between the 'unknown' speaker's samples from the database samples of 'Speaker m' averaged over the three words is minimum. Here, the different words uttered serve as statistical blocking factors, enhancing recognition rate.

3.2 Identification in an 'open set'

In this case, the objective is slightly different and more difficult. The problem is to successfully identify a speaker who is in the set of 40 speakers and reject those who are not. Given a word, let two samples belong to the same cluster (i.e. the same speaker as in our case), if the Euclidean distance is less than some threshold distance d_0 [6]. It is immediately obvious that the choice of d_0 is very important. Large values of d_0 will result in *false acceptance*. If d_0 is small, it'll lead to *false rejection*. Hence, the choice of the threshold 'd_0' has to be such that it is greater than 'average within speaker euclidean distances', but, less than the 'between speaker euclidean distances'. Here, the modified codebook for each replicate of the database speakers would contain the contents as in the closed set case, as well as, the threshold value (the average within speaker distance) of the Euclidean distance for the corresponding word of the database speaker.

Therefore, an 'unknown' speaker is said to be the m^{th} speaker in the database if and only if for each word his Euclidean distance value (as is stated in the preceding section) is less than the threshold value d_0 for each word corresponding to the m^{th} speaker. Experimental results have been provided in the following section by randomly eliminating from the database, a set of 5 speakers, and then choosing a speaker from the original 40 speakers to test for identification.

4 Results

Successful identification in a closed set of speakers by choosing the statistical mean of the pixel values of each band and a comparison with entropy as the dependent variable has been depicted in Table 1. Corresponding 'success rates' when identifying speaker by spectrogram distances (Hollander-Wolfe Statistic) [4] and popularly used VQ has also been reported in Table 2.

As is evident from the Tables, the CPU run time registered, to identify a speaker is lesser in the proposed technique than the conventional VQ technique or the Hollander Wolfe measure.

Though, successful identification of a speaker from just a word by calculating the minimum 'Euclidean distance' (based on 4 training samples), may be as low as $75 - 80\%$ (using mean), combining results from the 3 words, computing the aggregate distance (as stated in Speaker Identification) and choosing an appropiate database size for every speaker (4 speech samples for every word corresponding to each speaker as in the case study), one can obtain as good as 100% success rate in identification in a closed set problem. Results are stated in Table 1. While, identification rate when the spectrogram is not segmented is as low as 27.5% (when using mean of the pixel values of the entire spectrogram), segmenting the spectrogram along frequency axis and working with the mean values of the frequency bands yielded better results which is as shown in Figure 3. Results shown in Table 1, also concludes that an appropriate choice of the study variable has a boosting effect on the success

Table 1: Success Rates in closed set identification (Spectrograms of pixel matrix size 253×271 segmented into 10 bands with common bandwidth 46 and band overlap 23 along frequency axis).

Distributional property used in our proposed technique	**Success Rate**	**CPU run-time to identify a speaker**
Mean	100%	0.98 sec.
Entropy	97.5%	1.1 sec.

Table 2: Success Rates in closed set identification (Spectrograms of pixel matrix size 253×271 segmented into 10 bands with common bandwidth 46 and band overlap 23 along frequency axis).

Techinque used	**Success Rate**	**CPU run-time to identify a speaker**
Vector Quantization (using MFCC)	100%	1.7 sec.
Hollander-Wolfe Statistic	85%	1.4 sec.

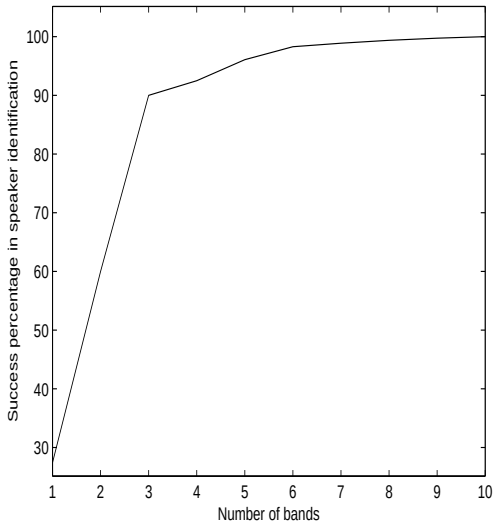

Figure 3: 'Successful identification Rate' Vs. 'Segmented spectrogram' when comparing the 'unknown speaker' (Speaker ID:20) with the known database.

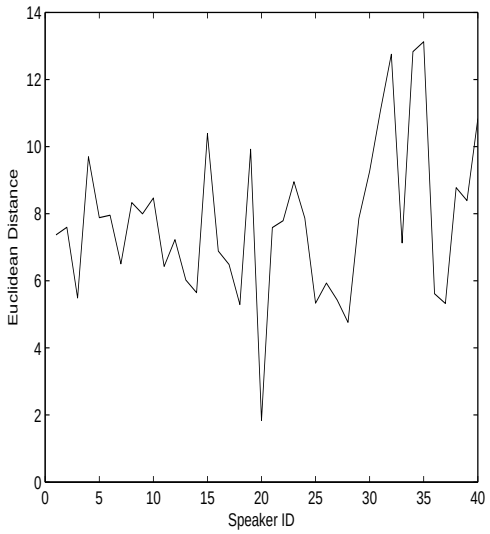

Figure 4: Euclidean distance Vs. Speaker when comparing an 'unknown speaker' ('Speaker ID:20') with the known database samples.

rate. Conducting 200 tests (each test comprising of 3 test spectrograms corresponding to the three words uttered by a speaker amongst the closed set of speakers) for each mentioned procedure, Success Rates, when it is known that the unknown speaker is from the closed set, have been computed which is as shown in Table 1. Figure 4 gives a plot of the comparative Euclidean distance an 'unknown speaker' (Speaker ID:20) has with the database samples of the speakers $1, \ldots, 40$.

In the open set classification, given a word, using the average 'within speaker distance' d_0, as the threshold value, for each word (corresponding to every speaker), the false rejection or false acceptance rates in identification when a 'unknown' speaker may or may not be in the closed set of speakers, was determined. This method of computation of d_0 satisfied the equal-error-rate criterion (EER) (stated in the *'Introduction'* section), which was computed to be 0.175. On increasing the value of d_0, as expected, the rate of *false acceptance* increases, while the value of *false rejection* falls,

which is certainly not desirable.

5 Applications

This methodology can be used to identify speakers in password protected zones where a database of voices of speakers can be used as passwords. This model, if required, can be made more dynamic by adding the "most recent successful voice acceptance" of a particular speaker into his/her database of samples, discarding his/her spectrogram corresponding to earliest voice sample in the database. This dynamic model, takes into consideration the change in voice of a particular speaker over time.

6 Conclusion and Future work

This paper presents a method for text-dependent speaker identification based on extracting unique speaker effects on the pronunciation of a word. It can be viewed as a clustering problem in which the each spectrogram band represent a cluster and its mean pixel value, the centroid of cluster. Hence, given an unknown speaker's utterance of a known word, we would be looking for the database sample of that particular word with ordered cluster centroids having having the closest euclidean distance with those of the unknown speaker.

Future work will focus on more robust nearest neighbor classifiers, better selection of words, optimality of bandwidth selection, implementation of this technique on a large-scale and in text-independent case. Also, it would be important subsequently, to reduce its computational complexity and computation time even further.

References

[1] J. Olsson. "Text Dependent Speaker Verification with a Hybrid HMM/ANN System," Thesis Project, downloadable at http://www.speech.kth.se/prod/publications/files/1630.pdf

[2] F.K. Soong, A.E. Rosenberg A.E., B.H. Juang, and L.R. Rabiner. "A vector quantization approach to speaker recognition," AT & T Technical Journal, 66:14-26, pp. 1987.

[3] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Commun. 17 (1995), pp. 91-108.

[4] Tridibesh Dutta and Gopal K. Basak. "Text dependent speaker identification using similar patterns in spectrograms," PRIP'2007 Proceedings, Volume 1, pp. 87-92, Minsk, 2007.

[5] E. Demidenko. "Kolmogorov-Smirnov image comparison," Lecture Notes Comp Sci 3056: 933-938, 2004.

[6] R. O. Duda, P. E. Hart, D. G. Stork. "Pattern Classification," John Wiley and Sons, 2006.

[7] Trevor Hastie, Robert Tibshirani, Jerome Friedman. "The Elements of Statistical Learning: Data Mining, Inference and Prediction," Springer, 2001.