# On the Use of Dynamic Information for Speaker Identification

Rosa M. Ortega-Mendoza, Esaú Villatoro-Tello, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez and Eduardo F. Morales

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
{rmortega, villatoroe, villasen, mmontesg, emorales}@inaoep.mx

**Abstract.** This paper describes a language independent method for speaker identification. This method is based on a novel characterization of the speech signal that captures the dynamic information contained in the cepstral coefficients. The proposed method was evaluated through several experiments on a corpus of Mexican speakers. The achieved results demonstrated the relevance of the signal characterization, reaching an identification accuracy as high as 97% under a multi-class scheme.

## 1 Introduction

The speech signal works as a vehicle for several types of information. It mainly bears a message through the language, but it also allows identifying the spoken language and establishing the emotion, the gender, the age as well as the identity of speakers.

In the task of automatic speaker recognition focuses on determining the identity of speakers through their voice. It involves two kinds of problems, namely, speaker identification and speaker verification. Speaker identification, the subject of this paper, determines which registered speaker provides a given utterance from amongst a set of known speakers. On the other hand, speaker verification accepts or rejects the identity claim of a speaker.

Speaker identification has been widely studied and it is currently performed at very high accuracy rates [1]. However, the best achieved results correspond to methods that are text dependent (users must read a predefined text) and language dependent (they depend on the availability of a phonetic recognizer) [2, 3].

In order to diminish these limitations, this paper proposes a new statistic method for speaker identification. This method is *text independent* [4], and also *language independent*. It directly works with the speech signal and avoids applying any process for phonetic recognition. In particular, the proposed method takes advantage from the dynamic information contained in the Mel Frequency Cepstral Coefficients to enhance the characterization of the speech signal.

The rest of the paper is organized as follows. Section 2 describes the proposed method. Section 3 presents the experimental results. Finally, section 4 depicts our conclusions and future work.

## 2 Proposed Method

The speaker identification task consists of two main phases: a training phase and a test phase. In the training phase an identification model is built from the recordings of every user. Subsequently, in the test phase, these models are used to determine the corresponding speaker for an input recording.

The construction of the identification models involves two processes. On the one hand, the characterization of the speech signals by the extraction of some descriptive parameters. On the other hand, the application of a stochastic procedure –over the extracted characterizations– in order to capture the distinctive regularities for each speaker. Traditionally, a cepstral analysis has been used for characterizing the signals, and Gaussian Mixture Models for inducing the identification models [4].

The contribution of the proposed method is on the application of a different speech characterization. It uses the cepstral coefficients to compute a new set of features that capture the time variations of the signals. These new features resume the dynamic behavior of the signals and thus enhance the construction of the speaker's models. In addition, our method accomplishes the statistical modeling by using some automatic classification algorithms, in particular, Naïve Bayes and Support Vector Machines.

The following sections describe in detail the proposed approach.

### 2.1 Signal Characterization

In order to construct the speech characterization the signal is represented by fixed-size segments and each segment is characterized using the Mel Frequency Cepstral Coefficients (MFCC). Basically, we consider non-overlapping segments of 30ms and calculate 16 coefficients per segment. We propose using 16 coefficients, instead of the twelve traditionally used for speech recognition, because we want to exploit all possible useful information to distinguish a speaker. Specially, the last coefficients contain information about the high frequencies that allow capturing the tone of speakers.

Using the cepstral coefficients our method constructs a more concise representation that expresses the speech information by a set of more general and time independent features. In particular, we also characterize the signals by their coefficient's variations. That is, we calculate the change of the coefficient's values between two contiguous signal segments. In order to enrich the acoustic characterization, we also compute the averages of the coefficient's variations as well as their maximum and minimum values. Several experiments were performed with the 16 MFCC and with the 64 different features to represent each signal sample.

Table 1 describes all the 64 statiscally-based features related to each one of the 16 Mel Frequency Cepstral Coefficients. In this table, $C_{ik}$ denotes the coefficient $i$ from segment $k$, $N$ indicates the number of considered segments, and $\Delta$ represents the coefficient variation between contiguous segments.

**Table 1.** Set of proposed features

| Description | Compute | Num. of Features |
|---|---|---|
| Maximum value of the coefficient's changes | $\hat{\Delta}c_i = \max_{k=2}^{N}\left(c_{ik} - c_{i(k-1)}\right)$ | 16 |
| Minimum value of the coefficient's changes | $\check{\Delta}c_i = \min_{k=2}^{N}\left(c_{ik} - c_{i(k-1)}\right)$ | 16 |
| Average value of the coefficient's changes | $\tilde{\Delta}c_i = \frac{1}{N-1}\sum_{k=2}^{N} c_{ik} - c_{i(k-1)}$ | 16 |
| Variance of the coefficient's changes | $\Delta_v c_i = \frac{1}{N-1}\sum_{k=2}^{N}\left(c_{ik} - \tilde{\Delta}c_i\right)^2$ | 16 |

### 2.2 Statistical Modeling

Once the set of features for every sample of every speaker are computed, we apply a machine learning algorithm to build the identification models. As we mentioned before, we employ two different algorithms: Naïve Bayes and Support Vector Machines. We briefly describe both algorithms.

**Naïve Bayes**. This probabilistic classifier is based on the assumption that the features are conditionally independent of each other given the target values (classes) [5]. It can be applied to learning tasks where each instance is described by a conjunction of feature values $a_1, a_2, \ldots, a_n$ and the target function $f$ can take any value from some finite set $V$. That is, given the instance $x$, the Naïve Bayes classifier assigns it the most probable target value in accordance with the following expression:

$$f(x) = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Where $P(v_j)$ represents the probability that the class $v$ has the *j-th* value and $P(a_i, v_j)$ is the conditionally probability that the feature $a$ takes the *i-th* value given that the class $v$ has the *j-th* value.

Therefore, the construction of a Naïve Bayes classifier basically involves the estimation of the probabilities $P(v_i)$ and $P(a_i \mid v_j)$ from the training data. These estimations are then used to classify new instances using *f(x)*.

**Support Vector Machines (SVM)**. This learning algorithm is specially suited to work with very high dimensional data sets. It uses geometric properties in such a way that it is capable to compute the hyperplane that best separates the training set [6]. In the case where the input space is not lineally separable, it projects the original training space to a higher dimensional feature space using a kernel in order to find an optimal hyperplane. The works by Vapnik [7] and Scholkopf and Smola [8] describe in detail the SVM algorithm.

## 4 Experiments and Results

In order to prove the proposed method we use a set of the recording samples from the DIMEx100 corpus [9]. This oral corpus for Mexican Spanish contains high quality (44 KHz) recordings from 100 different persons. For each person it includes 50 different phrases of 3.5 seconds long. In total, the corpus is about 291 seconds.

In particular, for the experimental evaluation, we randomly selected 50 persons and constructed speech samples of 3 seconds. In addition, as previously mentioned in section 3, we represented the signals by fixed-size segments of 30ms and characterized each segment using the Mel Frequency Cepstral Coefficients (MFCC).

In order to determine the relevance of the proposed method we performed the following experiments:

**Experiment 16MFCC+3**. In this case each recording sample was represented by the 16 MFCC coefficients for each segment and three additional features that captures the dynamic behavior of the signal: the minimum, the maximum and the average values for each coefficient.

**Experiment 16MFCC+4**. This experiment considered the same representation that in the previous case. It only included one additional "dynamic" feature: the variance value of each coefficient.

**Experiment 4/16**. In this experiment the recording samples were exclusively represented by the set of features expressing the dynamic information of the 16 coefficients, that is, they were represented by the minimum, maximum, average and variance values of each coefficient.

**Experiment 4/12**. For this final experiment we also only used the features that express the dynamic information, but we only considered 12 MFCC coefficients.

Table 2 shows the obtained results for the four experimental configurations. In all cases we used: (*i*) Naïve Bayes and Support Vector Machines, (*ii*) the Information Gain technique for dimensionality reduction (preserving those features having an information gain greater than zero), and (*iii*) the 10-fold-cross-validation technique for evaluation.

**Table 2.** Identification accuracies

| Experiment | Naïve Bayes | SVM |
|---|---|---|
| 16MFCC+3 | 85.59% | 85.94% |
| 16MFCC+4 | 90.47% | 89.55% |
| 4/16 | 94.56% | **97.56%** |
| 4/12 | 89.59% | 94.28% |

From table 2 we can derive the following conclusions. First, the dynamic information contained in the Mel Frequency Cepstral Coefficients is very useful for speaker identification. In particular we can observe that using only this information it was possible to achieve an accuracy of 97.56%. Second, using 16 coefficients produced better results than only considering the traditional twelve. This fact indicates that the highest coefficients (the highest frequencies) are relevant for distinguishing among different speakers. Finally, given that both learning algorithms produced similar results, we presume that the proposed signal characterization is

pertinent for the task. In other words, the achieved results were not a direct consequence of the applied classifier.

## 5 Conclusions and Future Work

This paper described a new method for speaker identification. This method has two main characteristics. On the one hand, it is *text independent* since it does not force users to read a predefined text. On the other hand, the method is *language independent* because it directly works over the speech signal and does not depend on any phonetic segmentation process.

The proposed method is mainly based on a new, simple and compact, signal characterization. This characterization is obtained from the Mel Frequency Cepstral Coefficients of the speech signals and only considers 64 features that resume the *dynamic behavior of the signals*. With this new signal characterization we are able to improve the construction of the speaker's models. Particularly, the presented results indicated an identification accuracy as high as 97%.

Finally, it is important to mention that our results are still preliminary and therefore more experiments are necessary to conclude about the real pertinence of the approach. In order to satisfy this condition we plan to participate in some recognized evaluation forums. Specifically we plan to evaluate our method in the speaker identification task of the NIST forum [10].

**Acknowledgments:** This work was done under the partial support of CONACYT (project grant 43990). We also thank SNI-Mexico and INAOE for their assistance.

## References

1. Reynolds D.A.: An Overview of Automatic Speaker Recognition Technology. In *Proceedings of the IEEE ICASSP, 2002*. Orlando, FL, 2002.
2. García-Perera P., Mex-Perera C. and Nolazco-Flores J. SVM Applied to the Generation of Biometric Speech Key. In *Proceedings of the 9th Iberoamerican Congress on Pattern Recognition (CIARP 2004)*, pages 637-644. Puebla, Mexico, 2004.
3. Campbell W. M., Campbell J. P., Reynolds D. A., Jones D. A. and Leek T. R., Phonetic Speaker Recognition with Support Vector Machines. In *Proceedings of the Neural Information Processing Systems Conference 2003*, pages 1377-1384. Vancouver, BC, Canada, 2003.
4. Bimbot F., Bonastre J-F., Fredouille C., Gravier G., Magrin-Chanolleau I., Meignier S., Merlin T., Ortega-García J., Petrovska-Delacrétaz D. and Reynolds D. A Tutorial on Text-Independent Speaker Verification. In *EURASIP Journal on Applied Signal Processing*, volume 2004, issue 4, pages 430-451, 2004.
5. Mitchell T. *Machine Learning*. McGraw Hill, 1997
6. Stitson M. O., Wetson J. A. E., Gammerman A., Vovk V. and Vapnik V.. 1996. Theory of support vector machines. *Technical Report CSD-TR-96-17*. Royal Holloway University of London, England, 1996.
7. Vapnik V. *The Nature of Statistical Learning Theory*. Number ISBN 0-38794559- Berlin: Springer-Verlag, 1995.
8. Scholkopf B. and Smola A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

9. Pineda L., Villaseñor-Pineda L., Cuétara J., Castellanos H. and López I. DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. In *Proceedings of the IX Ibero-American Conference on Artificial Intelligence (IBERAMIA 2004)*, pages 974-983. Puebla, Mexico, 2004.
10. NIST speaker recognition evaluations. *The NIST year 2006 Speaker Recognition Evaluation Plan*. http://www.nist.gov/speech/tests/spk/index.htm