Abstract: - This paper is intended to introduce a graduate student to the speaker verification problem. Knowledge of pioneer work, as well as real methods overview, clears the path into deeper investigation. A list of useful references and books provides further orientation. The panorama is not complete if practical considerations are not taken into account: corpora, assessment of detection performance, speech and channel variability.

Key-Words: Speaker Verification, Speaker Models, Feature Extraction, Matching score, Corpora

1 Introduction
Speaker verification is a speech processing system that accepts or rejects a person’s claimed identity. The biometric signal used is voice, a natural approach for access control to communication systems. It is also economic compared to other biometrics, and somewhat robust, depending on variables we will address later. The idea is to match a voice sample acquired in the recognition phase with a speaker model built in the training phase [1]. The decision yields to a true claimer or an impostor, but it also generates two types of errors: the false acceptance of an invalid user (FA) and the false rejection of a valid customer (FR).

The general speaker verification problem may involve a closed set of speakers or an open one. In the latter, no previous model of the speaker exists.

There is also another important classification between text-dependent (TD) and text-independent (TI) speaker verification. TD only models the speaker for a limited set of words in a known context. When the sequence of spoken words is unknown, the problem becomes more difficult and recognition rates decrease.

2 General overview
Speech production involves several processes that can be classified into four categories: semantic, linguistic, articulatory and acoustic. [B1]

Given that speech is a signal created from anatomical and learnt characteristics, there are considerable differences that make possible speaker recognition and verification.

A block diagram of the components of a speaker verification system [2] [B1] is given in Fig. 1. Verification process consists in five steps:
Data acquisition: Test speech from a microphone is transformed into digital speech.
Feature extraction: a short interval of speech is mapped into a multidimensional feature space.
Matching Method: the sequence of feature vectors is compared to speaker models by a matching method, producing a matching score
Decision making: depending on normalization and threshold level, a decision is accomplished in an hypothesis-testing problem.
Training: the acquired speech of a customer is processed to obtain a speaker model. It can be a template, a codebook or a statistical model, depending on the matching method used.

![Diagram](Fig.1)
2.1 Features

Feature extraction is the estimation of variables obtained from parameters of a speech signal. To reduce the dimension of the feature vector, a selection is made, meaning that a transformation that preserves important speaker information is performed. This new feature space enables simple measures of similarity.

One transformation widely used in ASV systems is Mel warping. It changes the frequency scale to place less emphasis on high frequencies, based on the nonlinear human perception of the spectrum. Mahalanobis distance [B2- p.241], divergence measure [28] and Bhattacharyya distance [15] are popular tools to perform feature selection.

2.2 Matching Methods

Before the matching process, speaker models have to be constructed from the features derived from the training sequences.

The classification sketched in Fig.2 introduces the subject.

- Template Models: In this case, a sequence retrieved from the training data is used to build a template. Then it is compared to another template from the test speech sequence. Templates are matched in a deterministic way, searching for an alignment of observed frames to the stored sequences attempting to minimize the distance between them.

- Stochastic Models: Pattern matching is performed measuring the conditional probability (likelihood) of the feature vector of the input sequence, for a given model. The focus is made in the concept of modeling a speaker with a conditional pdf, and to compare it against that of the claimed customer.

A short model description with advisable references follows. The reader must take into account that the list is not complete, and not all methods have the same relevance or real exploitation.

  During the production of speech, the articulators of the vocal tract change their configuration, producing variations over time. This method compensates human speech variability with an algorithm that maps time axis of both templates in order to minimize a match score. Fig. 3 shows speech variability between the observation and the training signal. After non-linear DTW, a better alignment was achieved. This algorithm was intended for TD verification, but lacked flexibility for more complex situations.

  Speech is fragmented into segments or templates. Selected features form vectors that are coded into codebook. It is designed to each speaker, clustering averaged data extracted from a specific text. Time information is not necessary, and the match score results from the minimization of the distance between the test features and that of the corresponding codebook.

- **NN**: Nearest Neighbors [15]
  A combination of both preceding methods is computed by measuring the distance between input and stored frames. The match scores result from averaging all NN distances. Although the method is more powerful, it requires large amounts of memory.

- **NTN**: Neural Tree Network [16] [17] [18] [19]
This network is a hierarchical classifier that uses a tree architecture to implement a sequential linear decision strategy. Features extracted from training data are labeled one or zero depending on the target speaker. The NTN learns to differentiate regions of feature space that belong to the claimer from those that are more likely to belong to an impostor.

There are other architectures of neural networks that offer different performances. Generally NN’s require a smaller number of parameters than VQ systems.

**PNN:** Predictive Neural Networks [25]
These kind of neural networks do not use discriminating information but genuine speaker characteristics to build a non-linear model. PNN’s predict the next frame using a back-propagation algorithm, being trained to minimize the error between the actual and the predicted frame.

**HMM:** Hidden Markov Models [B3]pp. 321
In contrast to Markov models, where each state corresponds to a deterministically observable event, in HMM’s observations are a probabilistic function of the state. In this case, the stochastic process is hidden. Fig. 4 displays a HMM of six states with \( a_{ij} \) transition probabilities and six observable states with \( b_k \) associated probabilities.

**GMM:** Gaussian Mixture Models [10]
This approach provides a probabilistic model of the underlying sounds of a person’s voice, without imposing markovian constraints among sound classes.

It has many advantages, like noise and channel robustness, efficient computation time, insensitivity to model initialization and higher identification performance that preceding methods.

As Fig. 5 shows, a GMM is a weighted (\( p_i \)) sum of M component probability density functions \( b_i \) of the random vector \( \bar{X} \). Each Gaussian pdf has a mean \( \mu_i \) and a covariance \( \sigma_i^2 \). The model is parameterized by the notation expressed in (1).

\[
\lambda = \{ p_i, \mu_i, \sigma_i^2 \} \quad i = 1, \ldots, M \quad (1)
\]

**MDD:** Mixture Decomposition Discrimination [26]
This is an alternate method to GMM, also based in HMM, which constructs a mixture profile of all HMM of the same word, spoken by the verified user. This is the basis for discriminating between users and impostors with relatively small computational complexity. Performance over a HMM TD verification system is significant when MDD is used in combination with HMM’s cohort normalization.

3 Pathfinders
The ASV field has been growing during the last three decades at an increasing pace. Based on the chronology progress found at [2] and [B1], the following incomplete summary pays tribute to some of the ASV pioneers and gives initial orientation to get involved into the subject:

- 1974 – Atal: Pattern matching of cepstral features.[3]
- 1979 – Markel and Davis: Linear Prediction of Long Term Statistics.[4]
- 1991 – Tishby: HMM (AR mix) trained with LP features.[8]
- 1995 – Reynolds: HMM (GMM) with Mel-Ceptrum coefficients. [9] [10]
The amount of research material that may found today is huge. Although some work is conducted to find new models and techniques, investigations tend to upgrade known systems in the robustness sense.

4 Corpora
Standard Speech corpora are data bases specifically created for development and evaluation in ASR and ASI/V research. Four factors have to be available in a corpus to evaluate its applicability to a speaker verification system: [27]
- Number and diversity of speakers, classified as customers and impostors.
- Number and time separation of sessions per speaker
- Type of speech (phrase, digit, read sentence, conversational speech)
- Channel, microphone and recording environment description.
A list of available corpora in different languages is listed in Table 1. Reference [27] and [28] provide further information.

<table>
<thead>
<tr>
<th>Corpus Name</th>
<th>Origin Language</th>
<th>Applications &amp; Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>LDC English</td>
<td>ASR</td>
</tr>
<tr>
<td>SIVA</td>
<td>ELRA Italian</td>
<td>ASV/I Tel. handsets</td>
</tr>
<tr>
<td>PolyVar</td>
<td>ELRA French</td>
<td>ASV/I Tel. handsets</td>
</tr>
<tr>
<td>POLYCOST</td>
<td>ELRA English</td>
<td>ASV/I Tel. handsets</td>
</tr>
<tr>
<td>YOHO</td>
<td>LDC English</td>
<td>TD ASV High quality</td>
</tr>
<tr>
<td>GAUDI / AHUMADA</td>
<td>CICYT Spanish</td>
<td>ASV/I Tel / In situ</td>
</tr>
<tr>
<td>SwitchBoard</td>
<td>LDC US English</td>
<td>ASV/I NIST Eval.</td>
</tr>
</tbody>
</table>

Table 1: Speech Corpora

5 Speech variability
Taking into account the way corpora have been collected, actual speaker verification is contaminated by factors of speech variability. In forensic context, for example, results may change considerably. [29]
There are two main categories:

5.1 Peculiar intra-speaker variability
Characteristics to be taken into account are: manner of speaking, age, gender, inter-session variability, dialectal variations, emotional conditions, health condition.

5.2 Forced intra-speaker variability
These effects may change speaker-dependent features: Lombard effect, external-influenced stress, cocktail-party effects.
To perform verification tasks in variable conditions, matching methods must be careful to include some type of score normalization.

6 Channel variability
Condition between recording session of training data and actual test speech signal may vary in many uncontrolled situations.
Differing microphones, transmission link effects, acoustic environment are possible causes of changes in test signal features, a pose a hard challenge to matching methods.
One way to deal with a part of the problem is to “clean” the input data during training. Compensation techniques widely used include cepstral mean substration, RASTA filtering, and spectral substration. Newer methods are speaker model synthesis and feature mapping [30].
On the output side, score domain compensation tries to act on score scales and shifts due to channel variations.

7 Actual Applications
Examples of applications for ASV systems [20] [21] are mostly in the low security area. Voice activated door locks, computer and website access controls, telephone banking systems for transaction authentication that does not require a PIN, law enforcement in home-parole and prison call monitoring, as well as voice samples for forensic analysis.[22]

8 Seeking robustness
The performance of an automatic speaker recognition system is often expressed in an equal error rate (EER), which measures the performance level where false acceptance and false rejection values become equal.
[23] But it is not always the adequate parameter to acknowledge the performance sought. A banking transaction must avoid FR error, in order to preserve customers, letting the bank pay for the capital loss. On the other hand, in the forensic context, FA error is unacceptable, because it incriminates innocent people. In practice, there is always a trade-off between FA and FR errors, thus selecting the models is a criteria that depends on the application. A means of representing performance on detection tasks is the DET curve (Detection Error trade-off), which plots FR error (log) as a function of FR error (log) [24].

In figure 6, an example of a DET curve is presented. Different systems of ASR are evaluated. The diagonal represent the EER.

As it was point out in previous topics, robustness problem also concerns the way speech is acquired. When the test samples come from a wide variety of handsets, channels and acoustic environments, model adaptation and other improved compensation techniques have to be developed.

9 Discussion
Choosing features for a standard ASI/V system depends on many aspects. Long-term averages of spectral parameters were first used, but their loss of speaker-dependent information and the interval size required precluded their application. Viewing the problem as a matter of separation of different pdf’s in a M-dimensional space, [B2] a careful selection of each feature gives better results and uses lower computation. Therefore, few features that are selves independent but have small intra-speaker and large inter-speaker variances result in better separated clusters for a specific individual. Inherent characteristics, determined by the anatomy of the vocal tract, are better represented by stochastic models, and therefore suitable for TI recognition.

10 Conclusions
Speaker verification is a research subject that involves large amounts of complex material. By introducing the basic system, describing its components, giving vast references of well-known papers and over viewing others aspects such as speech and channel variability, the reader is prepared to begin studying any of the preceding topics.

Bibliography


References


