eufrak twocolumn
AMSbUmsbmn AMSbUmsamn
AMSb $AMSb'162$

OPEN SESAME! SPEECH, PASSWORD OR KEY TO SECURE YOUR DOOR?

Stéphane H. Maes , Homayoon S. M. Beigi

Human Language Technologies Group,
Speech Decoding Design Department,
IBM T.J. Watson Research Center
P.O. Box 218, Route 134,
Yorktown Heights, NY 10598, USA
e-mail: smaes@watson.ibm.com

Abstract

This paper reviews the state of the art in speaker recognition. It clarifies the different technical solutions that have been explored with some success as well as the challenges and limitations of current systems. It also describes the different functions and modalities involved in speaker recognition, where the terminology is still amazingly confused: specialist often uses the same terms for different conecpts. We review the classical techniques used in speaker recognition. Finally, we introduce the revolutionary concepts of speech biometrics. By discussing the impact of these new concepts, the maturity of speaker recognition is re-focussed.

1 Introduction

Access control to locations, services and resources is one of the oldest human goals. Except for some rare societies, private, tribal, corporate, official or national properties have always existed and required means to prevent violation, damages or losses. Jean-Jacques Rousseau even attributes the origins of all wars and injustices to the inception of private property. As for any human endeavor, more and more complex techniques have been developed or imagined. They can be classified into: hiding places: treasures; strongholds: fortress, armed protection, electronic protection, remote servers; key-based access: door, vaults, encryption; knowledge-based access: login, PIN, password; biometrics.
We could present numbers of legends and stories where biometrics play or could have played major roles. Obviously "Alibaba and the 40 thieves" is the most appropriate in this context: a secret speech command is used to open a door. ¿From the beginning all the ingredients for using speech as biometrics are present..... Obviously Sesame was voice controlled, but in order to help Alibaba, the author could not equip the door with speaker verification or identification capabilities: no biometrics on Sesame! Sure, the author of the 1001 Nights would have needed a different scheme to let Alibaba enter! But what a wonderful security system it would have been! What an idea! Unfortunately, in 1992, Robert Redford showed us its modern weakness in "Sneakers"! So, is it a good or a bad concept? Will you ever equip your house with Sesame?
Hopefully, by the end of this paper, we will convince the reader of the importance and perspectives opened by "speech biometrics". By the same token we plan to radically change the meaning of these words.

2 Functions of speaker recognition

Speaker recognition is a generic term that encompasses all the activities involving matching a speech waveform to the identity of the speaker. Numbers of contradictory sub-divisions have been proposed in the literature. From paper to paper and specialist to specialist, it is very common to find the same terms used to denote diametrically opposite activities. The fact that even specialists or vendors systematically interchange the role of these categories almost from sentence to sentence does not help a broader audience to embrace the field of speaker recognition. Amazingly, this plethora of contradictions does not affect other biometrics. We conjecture that it should be attributed to a large extent to the additional confusion that exists between speech recognition and speaker recognition.
We distinguish three different functions, with some minor subdivision:

- speaker identification
- speaker verification
- speaker classification
- speaker enrollment

Our speaker recognition approach does not distinguish from a technology point of view between these different functions: our speaker recognition engines implement them all at once. Since June 1996, when the SRAPI - Speech Recognition API - committee created the SVAPI - Speaker Verification and Identification API - sub-committee, we made sure that these definitions and functions are all included in SVAPI [?]. This is especially important as the concept of speaker classification, which until recently was mostly ignored by the community. Without our effort, SVAPI would probably be limited to text-dependent/text-prompted speaker verification. There is no doubt that confusions exist even within the ranks of the proponents of speaker recognition!

2.1 Speaker identification

Speaker identification consists of identifying a speaker based on his or her voice. The speakers are already enrolled in the system. No identity claim is provided. We speak of closed-set speaker identification if we restrict the set of speakers to be identified to the enrolled speakers. If unknown speakers, not yet enrolled, must be rejectected by the system, we speak of open-set speaker identification.

In terms of biometrics, speaker identification is a "many-to-many" recognition task. The decision alternatives are equal to the size of the enrolled speakers (+ 1 in open-set case). Therefore, the accuracy of speaker identification degrades as the size of the speaker population increases. As speaker identification accuracy does not yet compare to the performances of other biometrics, we can understand why, after for more than 30 years of research, speaker identification has not yet reached maturity! Especially as speaker identification engines are not yet able to cope with uncooperative speakers! Uncooperative users are those users who disguise their voice intentionally in order to avoid being identified e.g. candidates for social benefits who cheat the system by submitting with multiple applications.

Besides classical speaker identification, some extensions exists with added functionality of providing N-best lists or confidence scores. In the former case, a speaker identification system returns a sorted list of N identities who match the best the current speaker. The latter case rather implies that the identifier will produce a confidence level for each enrolled speakers that he or she matches the current speaker. Within these frameworks, speaker identification is much closer to maturity.

The reader will note in the next session that when it comes to confidence levels and rejection as out-of-set, speaker identification and speaker verification share a common behavior. Indeed, although such strategy is computationally expensive, identification can be implemented by repeated verifications with each speaker in the enrolled population used for subsequent identity claims.

Open set speaker identification requires rejection features that can usually be directly used for verification purposes.

2.2 Speaker verification

Speaker verification consists of verifying the identity claim of a speaker based on his or her voice. The identity claim designates a speaker enrolled in the system. Otherwise, rejection is trivial. Concepts of open or closed sets are not relevant to speaker verification.

In terms of biometrics, speaker verification is a "one-to-many" recognition task. There are only two choices: accept or reject. Contrary to speaker identification, the accuracy of speaker verification is not directly dependent on the population size. However, as it is typical in biometrics, the estimate of this accuracy depends on the representation of the population samples used to evaluate the accuracy. In contrast to other biometrics, these estimators also strongly depend on the channel effects and noise corruption of the signal. In general, speaker recognition performances vary dramatically from matched conditions (same type of microphone, channel characteristics and background noise) to mismatched conditions. The difficulty to correctly accommodate the effects of these mismatches dramatically damage the performances of speaker verification engines. To a large extent, this explains the relatively limited deployment of speaker verification systems. The other major cause being due to the goat phenomena: the majority of errors committed by verification engines are concentrated over a small fraction of the population (a few percent at maximum). Unfortunately, from the point of view of an application developer, it is not acceptable to exclude from a service a portion of the population on the basis of unexplained and uncontrolled behavior of their voice! Hence, some systems were never completely deployed!

There are two types of errors: *false acceptance* (an imposter has been incorrectly granted access) and *false rejection* (the authorized user has been denied access). Depending on the application, false acceptance will often be critical. However, in the literature, the focus is more often on the *equal error rate*: the total error rate committed when the false acceptance rate is equal to the false rejection rate.

Besides classical speaker verification, we must also mention extensions where instead of hard accept or reject decisions, confidence levels are returned.

2.3 Speaker classification

Speaker classification consists of performing speaker recognition over an unknown number of unknown speakers. Usually, it means to be able to detect speaker changes, also called speaker separation, and index the resulting segments according to the identity.

This function is specifically speech related. Only portions of the concept are met in other biometrics. However, the capabilities that it offers to distinguish between different undeclared successive users of a system may also be implemented with other biometrics.

Errors are measured in terms of segmentation mistakes (segmentation points versus speaker changes, end-times in the middle of words instead of in silences), and grouping mistakes (segments of one speakers attributed to another speaker).

Different sub-functions can be distinguished: speaker separation (speaker changes and regrouping segments of a same speaker) [?, ?, ?]; segment clustering [?]; speaker clustering (grouping speaker based on their similarities) [?, ?, ?, ?, ?].

Speaker clustering in unsupervised mode involves a bottom-up clustering of the model of different speakers. On the other hand, supervised speaker clustering usually leads to classes of speakers based on their gender, age, regional accent etc.

2.4 Speaker enrollment

In order to recognize the user based on his or her voice, first we need to acquire samples of the user's voice and create a model. Such models are usually called speaker models. Often, the models used for speaker identification differ from those used for speaker verification. By analogy to fingerprints, voice-prints refer to the minimum set of characteristics of a speaker required to create the speaker models used for identification and verification. We use the same models for identification as well as verification.

Of course in text-independent mode and with the appropriate technology, any voice sample can be used to create a voice-print. However, enrollment usually involves a strict procedure that the enrolling speakers must follow step by step: e.g. repeating words digits or sentences.

As for speech recognition, the principle is that there is no better enrollment data than more data! The more data that is available for a speaker the more accurate the voice-prints will be. Especially if this data can be collected over multiple mismatched conditions representative of the actual mismatches experienced during recognition.

3 Modalities of speaker recognition

There are multiple modalities in speaker recognition, i.e. different types of constraints imposed on the utterances used for enrolment or recognition. We distinguish between:

- text-dependent speaker recognition
- text-prompted speaker recognition
- text selected by user speaker recognition
- text-independent speaker recognition

3.1 Content-constrained speaker recognition

The first three categories may be defined as text-constrained speaker recognition.

3.1.1 Text-dependent speaker recognition

The content of the testing utterance matches the content of the enrolment utterance. In other words, the engine knows explicitly what the user is saying. The text can be different from user to user and it is possible that multiple texts are associated with each speaker.

3.1.2 Text-prompted speaker recognition

As in the previous case, the speaker recognition engine knows what the user is saying or supposed to say. The system asks the user to repeat a text, usually obtained by combining elementary keys or units; typically sequences of digits. During enrolment, the user is also prompted to repeat combinations of tokens.

3.1.3 Text selected by user speaker recognition

The users select a password or some key sentences to repeat at recognition. During enrollment, the user is asked to repeat their selected utterance. At recognition he or she repeats the same sentence. Furthermore, this sentence can contain the actual identity claim of the user. For example, it can be achieved by implementing speaker verification with open set speaker identification.

3.1.4 Usage

In practice, text constrained speaker recognition is only appropriate for verification tasks where a separate process devoted to text-constrained speaker recognition is acceptable. Speaker identification and classification can rarely accommodate such constraints: their applications usually require free speech capabilities.
Fraud remains easy when it comes to text-dependent speaker verification and even to a lesser extent text-selected by user speaker verification: play-backs or synthesis. Text-prompted speaker verification seems to solve these issues unfortunately, it is a major burden for the user. It also significantly slows down transactions.

3.2 Text-independent speaker recognition

3.2.1 Definition

This category designates speaker recognition on free speech utterances. The content of the speech utterance is completely unknown to the engine and does not have to be related to the enrollment utterances. In true text- independent speaker recognition, the recognition can be done in a language different from the language used during enrollment.

3.2.2 Usage

Contrary to text-constrained speaker recognition, text-independent speaker recognition presents the unique advantageof being applicable to any speech utterance. Speaker recognition can therefore be performed in the background of a regular conversation, request or transaction. Enrollment can be simplified: any speech from a speaker can be used to enroll this speaker.
By allowing the recognition to happen in the background of a transaction, with an IVR or an operator, the scenario of the transaction should forbid playbacks or syntheses. Indeed with the current technology it is not possible to generate the dialogs in real time. Also, synthesis effects can be reliably detected from the signal (e.g. pitch discontinuities).
Text-independent modality is mandatory for identification or classification.

3.3 Availability

Most of the current technology providers in speaker verification have opted for text-constrained speaker verification rather than text-independent speaker verification which is technically still more challenging. However, the non-obstrusive character of text-independent speaker verification presents multiple advantages including the capability to continuously process speech until a decision can be made.
Multiple research organizations have engaged in efforts to develop and improve text-independent speaker recognition. We are pursuing such research efforts and our speaker recognition engines are text-independent with a strong emphasis on its integration with our speech recognition engines.

4 Technology

The limited space available for this review paper forces us to limit the technical review of speaker recognition. We suggest that the reader consults the following papers for a more detailed discussion: [?, ?, ?, ?, ?, ?].
A speaker recognition system typically encompasses two elements: an acoustic feature extractor and a feature classifier.

4.1 Acoustic features

Acoustic features characterize the vocal tract characteristics of a speaker of one hand and the source properties (suprasegmental features, pitch etc) on the other hand. In contrast to speech recognition that in the vast majority of the systems uses Mel Frequency Cepstral Coefficients (MFCC), a whole variety of exotic or proprietary acoustic features are used for speaker recognition. The features usually considered to be the most appropriate are LPC cepstral coefficients. MFCCs are considered less efficient; however, systems developed by speech recognition providers often use them.