CONVERSATIONAL BIOMETRICS

Stéphane H. Maes
IBM T.J. Watson Research Center Yorktown Heights, NY 10598, USA
e-mail: smaes@watson.ibm.com

Abstract
In this paper, we present a new modality for speaker recognition: conversational biometrics. By combining diverse simultaneous technological approaches, high accuracy transparent speaker recognition becomes possible even in channel or environment mismatch. For speaker identification over very large populations, we combine dialogues to reduce the set of confusable speakers and text-independent speaker identification to pin-point the actual speaker. Similarly, dialogues with personal random or predefined questions are used to perform simultaneously knowledge-based and acoustic-based verifications of the user. Adequate design of the dialog allows to tailor the ROC curves to the needs of most applications.

1 Introduction

Speaker recognition encompasses all the activities involving matching a speech waveform to the identity of the speaker. Technical overviews of the field can be found in [1,9,21,12,10,11,5,20]. Speaker recognition is reaching maturity. However, multiple unknown factors still exist: unicity, cooperative speakers, robustness, etc. There is no doubt that text-constrained verification can be deployed in some specific environments or for some specific applications. Multiple commercial systems are already available. However, in most cases, these systems are only used as secondary systems or field test prototypes. On the other hand, text-independent systems are technically more challenging. Performances are lower, but they open new perspectives and application possibilities, especially since text-constrained systems are too prone to fraud by playbacks or future generations of synthesizers able to better mimic a speaker voice.

Speaker recognition has a unique advantage over other biometrics by relying on speech, the primary vector of communication. This is especially true for remote authentication. It requires only cheap sensors, microphones and handsets, that are already widely available. Furthermore, the explosion of speech recognition opens the way for voice interaction with computers and machines and therefore dramatically increases opportunities and needs for accurate speaker recognition capabilities. As part of a new user interface paradigm, un-obtrusive speaker recognition, transparent to the user and the transaction, becomes even more important.

Classical authentication relies on one of these three items: what you own, what you are and what you know. Key or card-based systems characterize what you own. PIN and password based systems rely on what you know. Voice password have been proposed: utterance verification for access control and password compliance [24,19,18]. Biometrics and in particular speaker recognition rely on what you are. In this paper, we discuss a very important new approach: speech biometrics or conversational biometrics. Text-independent speaker recognition is used to acoustically identify or verify answers from the user in dialog with the system. The questions addressed to the user can be randomly selected, follow a predefined sequence or follow a business logic. With this approach, user verification and identification rely on acoustic recognition and on the content of the answers to the questions.

The paper presents implementation architectures suitable for most application segments. We describe enrollment, identification and verification scenarios and algorithms. Additionally, we discuss the advantages of conversational biometrics over utterance verification, voice password recognition and other non-simultaneous combinations of speech recognition and speaker recognition. Eventually results are provided to illustrate the improvements achievable over conventional approaches. It constitutes a demonstration that conversational biometrics can indeed be deployed with today’s technology even as primary, and unique, security systems.

2 Functions of speaker recognition

We distinguish four different functions, with some minor subdivision:

- **Speaker identification**. It consists into 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 rejected by the system, we speak of open-set speaker identification. In terms of biometrics, speaker identification is a "many-to-one" 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. Besides classical speaker identification, some extensions exist 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 that 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 as illustrated in the audio-indexing example presented later on. 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 common components. 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.

- **Speaker verification**. Speaker verification consists of verifying the identity claim of a speaker based on his or her voice. In terms of biometrics, speaker verification is a "one-to-many" recognition task. 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.

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

- 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 unائل 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 speaker attributed to another speaker). Different sub-functions can be distinguished: speaker separation (speaker changes and regrouping segments of a same speaker) [15, 14, 4, 7]; segment clustering [4]; speaker clustering (grouping speaker based on their similarities) [3, 4, 16, 17, 25].

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.

- Speaker enrollment. In order to recognize the user based on his or her voice, 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. Voice-prints are algorithm-dependent.

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. [6]

4 Integration of speaker recognition and speech recognition

Text-constrained speaker recognition is only used for speaker verification purposes. Because the text associated to the utterance is available to the engine, more information can be extracted in the voice-prints and higher accuracy are achieved.

On the other hand, text-independent technology can advantageously implement all the modalities afore mentioned. Furthermore, because it does not constrain the content of the utterances used for enrollment or recognition, it can be used in the background of dialogs between humans or machines.

Figure 1: Integration of speech and speaker recognition engines.

Consider the system described in figure 1, which simultaneously performs speech recognition and speaker recognition on the input utterances. The audio stream is provided to the acoustic front-end as isolated utterances (command and control mode or answers to a directed dialog) or as a continuous stream. The front-end captures the audio and extracts the acoustic features (e.g. MPCC). The features can be further compressed using the algorithm described in [25]. Note that the acoustic features are shared by the speech recognition engine and any post processing system (e.g. natural language understanding module [22]). In networked applications, where the acoustic front-end can be on the client side while the other conversational functions are performed on the server side, sharable features allow one data stream at data rates as low as 4 to 5 kb/s, quite suitable for wireless modem connections. On embedded systems, only one signal processing task is performed. This reduces the CPU, memory and power requirements. The feature stream is then split up between the speech recognition engine and the text-independent speaker recognition engine.

With such a system, numerous integrated functions can be offered:

- Text-independent speaker identification can be performed recognizer to select the speaker and load speaker-dependent [15].
- Text-independent speaker classification expands the concept to load class-dependent models in the speech recognizer [13]. This increases the performances of the system without enrollment and reduces the amount required for efficient adaptation to the speaker.
- Simultaneous speech recognition and text-independent speaker identification to disambiguate commands. A typical example is a voice dailer function. Commands like "call home" or "call my wife" are ambiguous when the system is accessible to multiple users. The identity tag returned by the speaker identification system helps to complete or disambiguate the command.
Simultaneous speech recognition and text-independent speaker verification can be used for continuous access control. For example, in a command and control application or directed dialog, each command or transaction request can be executed only upon verification of the speaker. The verification can be performed on a command by command basis or on a buffered set of utterances. This also provides continuous background monitoring capabilities to certify that no speaker took place during a transaction, or after the authentication.

Obviously, such integration of the speaker recognition capability allows transparent recognition in a manner totally non-obtrusive to the user and the transaction. Also it is well known that the more data can be used for recognition, the better the performances. Therefore, it is also particularly advantageous to postpone any recognition decision later in the transaction when a decision is required.

Figure 2 illustrate another use of integrated speaker recognition and speech recognition [26]. In this example, speaker classification is implemented by a sequence of text-independent speaker segmentation, identification and verification. A speech recognizer simultaneously transcribes each resulting segment. The resulting time stamped transcripts and speaker identity tags can be stored for later search of the audio segment database (audio search) or displayed in real time with identity annotations (close captioning). Other use

5 Conversational Biometrics

Conversational biometrics, as we originally called it [20], designates the combination combine acoustic-based speaker recognition with content-based and knowledge-based recognition. It requires close integration of the text-independent engine with the entire conversational system. As illustrated in figure 3 conversational systems consist of speech recognition, speech synthesis, natural language understanding, natural language generation and dialog management [22, 8]. Indeed, the dialog management now carries a conversation with the speaker aimed at automatically identifying a cooperative user or verifying a claimant.

The dialog manager described in [22] is expanded to include an identification form and a verification form, which could be a unique form. Note that both can be combined or that the dialog can degenerate into a finite state grammar FSG machine driven Q&A.

Conversational identification consists of a dialog that reduces the set of confusable speakers handled by the speaker identification engine. It supposes cooperative users. For example, an IVR (Interactive Voice Response) system interrogates the speakers as follows:

- "What is your name"
- I am John Doe
- "What city are you calling from?"
- I am in Manhattan
... By now, out of the pool of millions of users enrolled in the system, the dialog has reduced the set of candidates to a subset for example smaller than say twenty to fifty speakers. If the sub-set is still bigger the dialog can continue. Text-independent speaker identification can now be performed to extract the best candidate or a N-best list. We denote $p(X)$ as the probability to correctly identify speaker $X$ by conversational biometrics, $p(X|\mathcal{S})$ as the probability to correctly identify (and/or reject) speaker $X$ out of a set of speakers $\mathcal{S}$ using the text-independent identification engine and $p(\mathcal{S} \ni X|\mathcal{P})$ as the probability to correctly extract a subset $\mathcal{S}$ that contains speaker $X$ by carrying the dialog $\mathcal{P}$. We have,

$$p(X) = p(x|\mathcal{S}) \cdot p(\mathcal{S} \ni X|\mathcal{P})$$

Over the telephone, a dialog as described in [22], or a FSG dialog, has an accuracy of 93% to 97%. With an acoustic identification average accuracy of 90% over sub-sets of 20 to 50 speakers, implemented with a telephony variation of [3, 6, 2], the overall accuracy over a large population of 200,000 speakers is now between 83.7% and 87.3%. Richer free-flow dialogues, with mixed initiatives and possibility for the user to correct or complement his or hers input further, increase the accuracy by increasing $p(x|\mathcal{S})$ and $p(\mathcal{S} \ni X|\mathcal{P})$. Any state of the art acoustic-only text-independent speaker identification system would do very poorly on populations of this size. Indeed, remember that the probability of incorrect identification increases proportionally to the size of the population. Furthermore, the computational complexity is also much smaller than what is required by acoustic-only identification. Eventually, acoustic hierarchical methods as described in [2] present higher error probability as each intermediate decision has an additive contribution to the total probability of error. Therefore the probability of error is also significantly increased with the size of the population.

Conversational verification consists of a dialog to perform a knowledge-based verification of the speaker in parallel with the acoustic-based verification. It is a powerful mechanism to combat the limitations still inherent to speaker verification systems. Consider a automated phone banking application driven by an IVR. The following dialog takes place:
- "What is your mother’s maiden name?"
- My mother’s name was Doe
- "What is your favorite color?"
- I like red
- ...

The questions can be randomly generated out of information collected during enrollment or they can be dynamically generated based on past history. With well designed dialogs,

\[ p(FA|X) \propto p_{FA}(X) \left( \frac{1}{N} \right)^k \]

where \( p(FA|X) \) is the probability of false acceptance by conversational biometrics verification, \( p_{FA}(X) \) is the probability of acoustic false acceptance, \( N \) is the underlying average perplexity for each answer and \( k \) is the average number of questions per dialog \( D \). Similarly, for the false rejection (FR):

\[ p(FR|X) = p_{FR}(X) + p_{err}(D) - p_{FR}(X)p_{err}(D) \]

where \( p(FR|X) \) denotes the probability of false rejection by conversational biometrics verification, \( p_{FR}(X) \) is the probability of acoustic false rejection probability and \( p_{err}(D) \) is the probability of making an error in the dialog. In the telephony example described previously, where the acoustic equal error rate is between 3% and \( p_{err}(D) = 3\% \) with \( N = 20000 \) and \( k = 3 \), we have an operating point at \( p(FA|X) \approx 3 \times 10^{-15} \) and \( p(FR|X) = 0.039 \). Additionally, recovery mechanism embedded in the dialog \( D \) allows to reduce the false rejection to any arbitrary level. Also, note that the performances are not affected by non-cooperative users. Acoustic verification can also be performed in continuous duration of the rest of the transaction.

Initial enrollment is implemented with a dialog that authenticates the user based on extraneous information and stores the answers to an initial set of questions. Acoustic baseform generation is used to store and represent out-of-vocabulary input. Later, additional questions can be generated based on past history of transactions or queries done by the user. Although it is beyond the scope of this paper, transcription, indexing, topic detection, speaker classification and non-acoustic information can be used to further improve the performance of verification and identification, even for non-cooperative users.

6 Conclusion

We have demonstrated the advantages of integrating speaker recognition and conversational systems to implement conversational biometrics. Appropriate design of the application allows to perform simultaneous content/knowledge-based recognition with high accuracy even in challenging conditions or over very large populations.

7. References