Lecture Notes- Fundamentals of Speaker Recognition

Sanjeev S Koni

Assistant Professor

Speaker Recognition Principles

Depending on the application, the general area of speaker recognition can be divided into three specific tasks: identification, detection/verification, and segmentation and clustering.

The goal of the *speaker identification* task is to determine which speaker out of a group of known speakers produces the input voice sample. There are two modes of operation that are related to the set of known voices. In the closed-set mode, the system assumes that to be determined voice must come from the set of known voices. Otherwise, the system is in open-set mode. The closed-set speaker identification can be considered as a multiple-class classification problem. In open-set mode, the speakers that do not belong to the set of known voices are referred to as impostors. This task can be used for forensic applications, e.g., speech evidence can be used to recognize the perpetrator's identity among several known suspects.

In *speaker verification*, the goal is to determine whether a person is who he or she claims to be according to his/her voice sample. This task is also known as voice verification or authentication, speaker authentication, talker verification or authentication, and speaker detection. It can be considered as a true-or-false binary decision problem. It is sometimes referred to as the open-set problem, because this task requires distinguishing a claimed speaker's voice known to the system from a potentially large group of voices unknown to the system. Today verification is the basis for most speaker recognition applications and the most commercially viable task. The open-set speaker identification task can be considered as the merger of the closed-set identification and open-set verification tasks. It performs like closed-set identification for known speakers but must also be able to classify speakers unknown to the system into an "unregistered speaker" category. Speaker verification can be used for security applications, such as, to control telephone access to banking services.

Speaker segmentation and clustering techniques are used in multiple-speaker scenarios. In many speech recognition and speaker recognition applications, it is often assumed that the speech from a particular individual is available for processing. When this is not the case, and the speech from the desired speaker is intermixed with other speakers, it is desired to segregate the speech into segments from the individuals before the recognition process commences. So the goal of this task is to divide the input audio into homogeneous segments and then label them via speaker

identity. Recently, this task has received more attention due to increased inclusion of multiple-speaker audio such as recorded news show or meetings in commonly used web searches and consumer electronic devices. Speaker segmentation and clustering is one way to index audio archives so that to make the retrieval easier.

According to the constraints placed on the speech used to train and test the system, Automatic speaker recognition can be further classified into text-dependent or text-independent tasks. In text-dependent recognition, the user must speak a given phrase known to the system, which can be fixed or prompted. The knowledge of a spoken phrase can provide better recognition results. In text-independent recognition, the system does not know the phrase spoken by the user. Although this adds flexibility to an application, it can have reduced accuracy for a fixed amount of speech.

Running a speaker recognition system typically involves two phases. In the first phase, a user enrolls by providing voice samples to the system. The system extracts speaker-specific information from the voice samples to build a voice model of the enrolling speaker. In the second phase, a user provides a voice sample (also referred to as test sample) that is used by the system to measure the similarity of the user's voice to the model(s) of the previously enrolled user(s) and, subsequently, to make a decision. The speaker associated with the model that is being tested is referred to as target speaker or claimant. In a speaker identification task, the system measures the similarity of the test sample to all stored voice models. In speaker verification task, the similarity is measured only to the model of the claimed identity. The decision also differs across systems. For example, a closed-set identification task outputs the identity of the recognized user; besides the identity, an open-set identification task can also choose to reject the user in case the test sample do not belong to any of the stored voice models; a verification task chooses to accept or reject the identity claim.

Basic Structure of a Speaker Recognition System

Like most pattern recognition problems, a speaker recognition system can be partitioned into two modules: feature extraction and classification. The classification module has two components: pattern matching and decision.

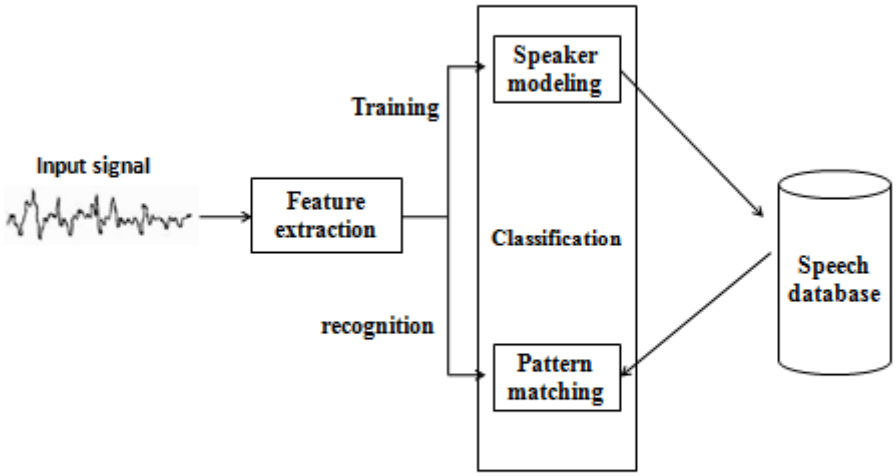

Figure 1 depicts a generic speaker recognition system.

The feature extraction module estimates a set of features from the speech signal that represent some speaker-specific information. The speaker-specific information is the result of complex transformations occurring at different levels of the speech production: semantic, phonologic, phonetic, and acoustic. The semantic level deals with transformation caused on the speech signal according to the communicative intent and dialog interaction of the speaker. For example, the vocabulary choice and the sentence formulation can be used to identify the socioeconomic status and/or education background of the speaker. The phonological level deals with the phonetic representation of the communicative intent. For example, duration and selection of phonemes, intonation of the sentence can be used to identify the native language and regional information. The phonetic level deals with the realization of the phonetic representation by the vibration of the vocal cords and the movements of articulators (lips, jaw, tongue, and velum) of the vocal tract. For example, speaker can use a different set of articulator movements to produce the same phoneme. The acoustic level deals with the spectral properties of the speech signal. For example, the dimensions of the vocal tract, or length and mass of vocal folds will define in some sense the fundamental and resonant frequencies, respectively. Despite the variety of speaker-specific information, the set of features should have the following characteristics:

• occur naturally and frequently in normal speech

• be easily measurable

• have high variability between speakers

• be consistent for each speaker

• not change over time or be affected by the speaker's health

• not be affected by reasonable background noise nor depend on specific transmission characteristics

• show resistance to disguise or mimicry

In practice, not all of these criteria can be applied to the parameters used by the current systems.

The pattern matching module is responsible for comparing the estimated features to the speaker models. There are many types of pattern matching methods and corresponding models used in speaker recognition. Some of the methods include Hidden Markov Models (HMM), and vector quantization (VQ). In open-set applications (speaker verification and open-set speaker identification), the estimated features can also be compared to a model that represents the unknown speakers. In a verification task, this module outputs a similarity score between the test sample and the claimed identity. In an identification task, it outputs similarity scores for all stored voice models. The decision module analyzes the similarity score(s) (statistical or deterministic) to make a decision. The decision process depends on the system task. For closed-set identification task, the decision can just select the identity associated with the model that is the most similar to the test sample. In open-set applications, the systems can also require a threshold to verify whether the similarity is valid. Since open-set application can also reject speakers, the cost of making an error needs to be considered in the decision process. For example, it is more costly for a bank to allow an impostor to withdraw money, than to reject a true bank customer.

The effectiveness of a speaker recognition system is measured differently for different tasks. Since the output of a closed-set speaker identification system is a speaker identity from a set of known speakers, the identification accuracy is used to measure the performance. For the speaker detection/verification systems, there are two types of error: false acceptance of an impostor and false rejection of a target speaker. The performance measure can also incorporate the cost associated with each error, which depends on the application. For example, in a telephone credit card purchase system, a false acceptance is very costly; in a toll fraud prevention system, false rejection can alienate customers.

Robust Speaker Recognition Applications

Speaker recognition technologies have wide application areas. Here we list some example applications of speaker recognition technologies.

• ***Security***: speaker recognition technologies can provide transaction authentication, facility or computer access control, monitoring, telephone voice authentication for long distance calling or banking access etc.

• ***Personalisation***: with speaker recognition technologies, we can implement intelligent answering machines with personalized caller greetings; we can build personalized dialog systems: a dialog system can recognize the user, greet to the user directly, and direct the user through the system to destination successfully via shorter path according to the user's profile.

• ***Information Retrieval***: speaker recognition can provide a way to manage and access the multimedia databases, which is to retrieve information according to interested speakers.

• ***Speaker tracking***: it is desired to know who is speaking in a tele-conference especially when there are many attendants in the tele-conference and the attendants are not very familiar with each other.

All these applications require robust speaker recognition techniques. For example in the telephone-aided services, users may call in under all kinds of acoustic conditions (in the office, on the street etc.) and use different telephone networks (land-line or cellular). In the meeting scenarios, participants may talk while moving around facing the microphone in different directions and different distances. Mismatched conditions may be encountered at any time in these cases. Therefore robustness is one of the critical factors that decide the success of speaker recognition in these applications.