Chapter 1.8
Speaker Recognition

Shung-Yung Lung
National University of Taiwan, Taiwan

ABSTRACT

This chapter presents some of the main contributions in several topics of speaker recognition since 1978. Representative books and surveys on speaker recognition published during this period are listed. Theoretical models for automatic speaker recognition are contrasted with practical design methodology. Research contributions to measure process, feature extraction, and classification are selectively discussed, including contributions to measure analysis, feature selection, and the experimental design of speaker classifiers. The chapter concludes with a representative set of applications of speaker recognition technology.

INTRODUCTION

Before addressing the various speaker recognition technologies and systems, it is appropriate to review the acoustical bases for speaker recognition. Research on speaker recognition, including identification and verification, has been an active area for several decades. The distinction between identification and verification is simple: the speaker identification task is to classify an unlabeled voice token as belonging to (having been spoken by) one of a set of N reference speakers (N possible outcomes), whereas the speaker verification task is to decide whether or not an unlabeled voice token belongs to a specific reference speaker (two possible outcomes: the token is either accepted as belonging to the reference speaker or is rejected as belonging to an impostor). Note that the information in bits, denoted I, to be gained from the identification task is in general greater than that to be gained from the verification task.

Speaker identification as just defined is also sometimes called "closed-set" identification, which contrasts it from "open-set" identification. In open-set identification the possibility exists that the unknown voice token does not belong to any of the reference speakers. The number of possible decisions is then $N + 1$, which includes the option to declare that the unknown token belongs to none of the reference speakers. Thus

open-set identification is a combination of the identification and verification tasks that combines the worst of both—performance is degraded by the complexity of the identification task, and the rejection option requires good characterization of speech feature statistics.

During the past few years, text-independent (or "free-text") speaker recognition has become an increasingly popular area of research, with a broad spectrum of potential applications. The free-text speaker recognition task definition is highly variable, from an acoustically clean and prescribed task description to environments where not only is the speech linguistically unconstrained, but also the acoustic environment is extremely adverse. Possible applications include forensic use, automatic sorting and classification of intelligence data, and passive security applications through monitoring of voice circuits. In general, applications for free-text speaker recognition have limited control of the conditions that influence system performance. Indeed, the definition of the task as "free-text" connotes a lack of complete control. (It may be assumed that a fixed text would be used if feasible, because better performance is possible if the text is known and calibrated beforehand.) This lack of control leads to corruption of the speech signal and consequently to degraded recognition performance. Corruption of the speech signal occurs in a number of ways, including distortions in the communication channel, additive acoustical noise, and probably most importantly through increased variability in the speech signal itself. (The speech signal may be expected to vary greatly under operational conditions in which the speaker may be absorbed in a task or involved in an emotionally charged situation.) Thus the free-text recognition task typically confers upon the researcher multiple problems—namely, that the input speech is unconstrained, that the speaker is uncooperative, and that the environmental parameters are uncontrolled.

What is it about the speech signal that conveys information about the speaker's identity? There are, of course, many different sources of speaker identifying information, including high-level information such as dialect, subject matter or context, and style of speech (including lexical and syntactical patterns of usage). This high-level information is certainly valuable as an aid to recognition of speakers by human listeners, but it has not been used in automatic recognition systems because of practical difficulties in acquiring and using such information. Rather, automatic techniques focus on "low-level" acoustical features. These low-level features include such characteristics of the speech signal as spectral amplitudes, voice pitch frequency, formant frequencies and bandwidths, and characteristic voicing aperiodicities. These variables may be measured as a function of time, or the statistics of long-term averages may be used as recognition variables. But the real question, the essence of the problem, is this: How stable are these speaker discriminating features? Given a speech signal, is the identity of the speaker uniquely decodable?

The fact is, however, that different individuals typically exhibit speech signal characteristics that are quite strikingly individualistic. We know that people sound different from each other, but the differences become visually apparent when comparing spectrograms from different individuals. The spectrogram is by far the most popular and generally informative tool available for phonetic analysis of speech signals. The spectrogram is a running display of the spectral amplitude of a short-time spectrum as a function of frequency and time. The amplitude is only rather crudely plotted as the level of darkness, but the resonant frequencies of the vocal tract are usually clearly represented in the spectrogram. Note the differences between the individual renditions of this linguistic message. Segment durations, formant frequencies, and formant frequency transitions, pitch and pitch dynamics, formant amplitudes, all exhibit gross differences from speaker to speaker. Thus these speakers would be very easy to discriminate by visual inspection of their spec-