Abstract

Extracting speaker-specific information from speech is of great interest to both researchers and developers alike, since speaker recognition technology finds application in a wide range of areas, primary among them being forensics and biometric security systems.

Several models and techniques have been employed to extract speaker information from the speech signal. Speech production is generally modeled as an excitation source followed by a filter. Physiologically, the source corresponds to the vocal fold vibrations and the filter corresponds to the spectrum-shaping vocal tract. Vocal tract-based features like the mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients have been shown to contain speaker information. However, high speed videos of the larynx show that the vocal folds of different individuals vibrate differently. Voice source (VS)-based features have also been shown to perform well in speaker recognition tasks, thereby revealing that the VS does contain speaker information. Moreover, a combination of the vocal tract and VS-based features has been shown to give an improved performance, showing that the latter contains supplementary speaker information.

In this study, the focus is on extracting speaker information from the VS. The existing techniques for the same are reviewed, and it is observed that the features which are obtained by fitting a time-domain model on the VS perform poorly than those obtained by simple transformations of the VS. Here, an attempt is made to propose an alternate way of characterizing the VS to extract speaker information, and to study the merits and shortcomings of the proposed speaker-specific features.

The VS cannot be measured directly. Thus, to characterize the VS, we first need an estimate of the VS, and the integrated linear prediction residual (ILPR) extracted from the speech signal is used as the VS estimate in this study. The voice source linear prediction model, which was proposed in an earlier study to obtain the ILPR, is used in this work.

It is hypothesized here that a speaker's voice may be characterized by the relative proportions of the harmonics present in the VS. The pitch synchronous discrete cosine transform (DCT) is shown to capture these, and the gross shape of the ILPR in a few coefficients. The ILPR and hence its DCT coefficients are visually observed to distinguish between speakers. However, it is also observed that they do have intra-speaker variability, and thus it is hypothesized that the distribution of the DCT coefficients may capture speaker information, and this distribution is modeled by a Gaussian mixture model (GMM).

The DCT coefficients of the ILPR (termed the DCTILPR) are directly used as a feature vector in speaker identification (SID) tasks. Issues related to the GMM, like the type of covariance matrix, are studied, and it is found that diagonal covariance matrices perform better than full covariance matrices. Thus, mixtures of Gaussians having diagonal covariances are used as speaker models, and by conducting SID experiments on three standard databases, it is found that the proposed DCTILPR features fare comparably with the existing VS-based features. It is also found that the gross shape of the VS contains most of the speaker information, and the very fine structure of the VS does not help in distinguishing speakers, and instead leads to more confusion between speakers. The major drawbacks of the DCTILPR are the session and handset variability, but they are also present in existing state-of-the-art speaker-specific VS-based features and the MFCCs, and hence seem to be common problems. There are techniques to compensate these variabilities, which need to be used when the systems using these features are deployed in an actual application.

The DCTILPR is found to improve the SID accuracy of a system trained with MFCC features by 12%, indicating that the DCTILPR features capture speaker information which is missed by the MFCCs. It is also found that a combination of MFCC and DCTILPR features on a speaker verification task gives significant performance improvement in the case of short test utterances. Thus, on the whole, this study proposes an alternate way of extracting speaker information from the VS, and adds to the evidence for speaker information present in the VS.