# Abstract

Voice conversion (VC) is a method to transform the voice of one speaker (source) so that it is perceived as spoken by another specified speaker (target). This technology has applications in text-to-speech customization, voice dubbing, education, speaking aids, entertainment, and a possible malicious application to deceive the speaker identification (SID) and speaker verification (SV) systems that use speech biometric. This thesis concentrates on the development of a new voice conversion system, use of voice conversion technology to generate synthetic speech attack on speech-based biometric systems, and finally, proposing an efficient solution to successfully discriminate natural speech from synthetic speech.

In the first contribution of this thesis, we propose a new VC method using i-vectors which consider low-dimensional representation of speech utterances. An attempt is made to restrict the i-vector variability in the intermediate computation of total variability ($\mathbf{T}$) matrix by using a novel approach that uses modified-prior distribution of the intermediate i-vectors. This $\mathbf{T}$-modification improves the speaker individuality conversion. For further improvement of conversion score and to keep a better balance between similarity and quality, band-wise spectrogram fusion between conventional joint density Gaussian mixture model (JDGMM) and i-vector based converted spectrograms is employed. The fused spectrogram retains more spectral details and leverages the complementary merits of each subsystem. The results show that the proposed technique can produce a better trade-off between similarity and quality score than other state-of-the-art baseline VC methods. Furthermore, it works better than JDGMM in limited VC training data. The proposed VC performs moderately better (both objective and subjective) than mixture of factor analyzer based baseline VC. In addition, the proposed VC provides better quality converted speech as compared to maximum likelihood-GMM VC with dynamic feature constraint.

The second work in this thesis presents an experimental study to evaluate the robustness of speech-based biometric systems (Gaussian mixture model (GMM) based SID systems, GMM with universal background model (GMM-UBM) and GMM supervector with support vector machine (GMM-SVM) based SV systems) against voice conversion disguise. Voice conversion is conducted by using GMM, weighted frequency warping (WFW) and variation of WFW ($\text{WFW}^{-}$), where energy correction is disabled. Experimental results show that the GMM-SVM SV systems are more resilient against voice conversion spoofing attacks than GMM-UBM SV systems, and all SID and SV systems are most vulnerable towards GMM-based conversion than WFW and $\text{WFW}^{-}$ based conversion. From the results, it can also be said that, in general terms, all SID and SV systems are slightly more robust to voices converted through cross-gender conversion than intra-gender conversion. This work extended the study to find out the relationship between VC objective score and SV system performance. The results of this experiment show an approach on quantifying objective score of voice conversion that can be related to the ability to spoof an SV system.

In the third contribution of this thesis, we propose a new approach to detect synthetic speech using score-level fusion of front-end features namely, constant Q cepstral coefficients (CQCCs), all-pole group delay function (APGDF) and fundamental frequency variation (FFV). CQCC and APGDF were individually used earlier for spoofing detection task, and yielded the best performance among magnitude and phase spectrum related features, respectively. The novel use of FFV feature to extract pitch variation at frame-level, provides complementary information to CQCC and APGDF. Experimental results show that an overall equal error rate (EER) of 0.05% with a relative performance improvement of 76% over the next best-reported results is obtained using the proposed method. In addition to outperforming all existing baseline features for both known and unknown attacks, the proposed feature combination yields superior performance for ASV system (GMM with universal background model/i-vector) integrated with countermeasure framework. Further, the proposed method is found to have relatively better generalization ability when either one or both of copy-synthesized data and limited spoofing data are available a priori in the training pool.

***Keywords:*** All-pole group delay function (APGDF), anti-spoofing, constant Q cepstral coefficient (CQCC), fundamental frequency variation (FFV), identity vector (i-vector), joint density Gaussian mixture model (JDGMM), modified-prior, score-level fusion, spectrogram fusion, spoofing attack, voice conversion.