論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Robust Background Models for Speaker Verification
著者(和文)	ﾋﾞｽﾜｽｻﾝｷﾞｰﾀ
Author(English)	SANGEETA BISWAS
出典(和文)	学位:博士（学術）, 学位授与機関:東京工業大学, 報告番号:甲第10048号, 授与年月日:2016年1月31日, 学位の種別:課程博士, 審査員:篠田　浩一,亀井　宏行,徳永　健伸,藤井　敦,石田　貴士
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第10048号, Conferred date:2016/1/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（論文博士）

論　文　要　旨（英文）　（800語程度）

（Summary）

報告番号	乙 第　　号	氏　名	BISWAS Sangeeta

（要　旨）

This thesis deals with short utterances and inter-session variability which considered as two important obstacles in text-independent speaker verification. It consists of six chapters.

Chapter 1 introduces automatic speaker verification system (ASVS) as a biometric system and describes the proposed framework for dealing with short utterances and inter-session variability in order to improve ASVS' performance.

Chapter 2 gives a technical overview of an ASVS and describes the speech databases, protocols and evaluation metrics used in our work.

Chapter 3 summarizes previous related works. Many known factors such as noise, recording devices, transmission channels, speakers' physiological/emotional state, age, short duration of speech as well as unknown variability of speech signal, make reliable discrimination of speakers a complicated and challenging task. By using a background model, the mismatch between the training and authentication sessions or the shortcomings of short speech for reliable parameter estimation of speaker model was greatly reduced. In order to train a good background model, two conditions need to be fulfilled: training data of the background model should be plentiful, and should have similar properties as the evaluation data, i.e., the target speakers and authentication data. Since in very few applications, the domain of authentication data can be pre-determined, researchers mainly try to match the domain of background data and the domain of target speakers. There is a trade-off between these two conditions. Using gender-dependent clusters is one good compromise for this trade-off. Obviously, speakers' acoustic properties depend not only on gender but also on the physical characteristics of the vocal tract, dialect, age etc. In addition, channel factors or background noise is known to greatly affect the acoustic properties of a recording. Considering that all known and unknown factors have similar effect on all speakers, in this research it is accepted the existence of sub-clusters in the set of target speakers and proposed to choose background model or background data for sub-clusters rather than only for a larger cluster like gender.

Chapter 4 describes our proposed acoustic forest and the motivation for using it. When training data of target speaker is very short (e.g., 10 seconds), it is not possible to estimate the model parameters of a Gaussian

mixture model (GMM) reliably by using the popular maximum-a-posteriori (MAP) adaptation technique. Structural modeling of human voice characteristics using structural MAP (SMAP) adaptation is a good solution of this case. However, in SMAP adaptation, only a single tree structure is used to model the acoustic space of all the target speakers. Since different target speakers have different acoustic spaces depending on factors such as their language, accents or pronunciations, it is reasonable to think that the optimal tree structure differs from speaker to speaker. In this chapter, it is proposed to grow an acoustic forest with different tree structures for SMAP adaption. In order to combine the decision of several SMAP adapted GMM-SVM systems, three types of score fusion techniques are proposed to use. Experimental setup and results regarding acoustic forest are also described in this chapter.

Chapter 5 presents our proposed method for removing irrelevant and noisy training data from the background model, probabilistic linear discriminant analysis (PLDA) model. When training data of target speakers is reasonably long (e.g., more than 2 minutes), channel variability compensation is considered as the main challenge. Recently, systems combining i-vector and probabilistic linear discriminant analysis (PLDA) have become one of the state-of-the-art methods in text-independent speaker verification using long utterances. General trend is to train gender-dependent PLDA models using all available data. There is no work focused on the data selection part of this background model. Selecting k nearest neighbors it is shown that we can improve the system performance by choosing a subset of the available training data of the PLDA model. In order to avoid the difficulty of optimizing k on a development set, a robust way of selecting k, named flexible k-NN (fk-NN), which uses a local distance-based outlier factor (LDOF), is presented here. It is also discussed how to use the enrolment speaker dependent k instead of using the same k for all enrolment speakers in fk-NN. The effect of i-vector selection on known and unknown non-target trials is also a topic of this chapter. By conducting experiments on male and female trials of several telephone conditions of the NIST Speaker Recognition Evaluations (SRE), it is shown that significant performance improvements can be achieved by using our proposed data selection methods.

Chapter 6 concludes the thesis summarizing the main results of proposed methods and outlining future research lines.

In summary, this thesis shows that it is beneficial to consider the existence of sub clusters in the set of target speakers and improve verification performance both for short and long training data of target speakers by focusing on background models.