

Contents

<i>Chapter / Section</i>	<i>Page</i>
1. Introduction	1
2. Speaker Recognition Background	7
2.1 The speech wave	7
2.2 Speech signal pre-processing and feature extraction	12
2.2.1 Pre-processing	12
2.2.2 Feature extraction	14
2.3 Principles and methods of speaker recognition	23
2.3.1 Speaker recognition methods	24
3. Feature Selection for Speaker Recognition	30
3.1 Selection procedure	30
3.1.1 Exhaustive search	31
3.1.2 K-best method	31
3.1.3 Forward selection	31
3.1.4 Backward selection	32
3.1.5 The l-r algorithm	32
3.1.6 The sequential floating forward sequence	32
3.1.7 Branch-and-bound	33
3.1.8 Dynamic programming	33
3.1.9 Genetic algorithm	35
3.2 Performance criteria	36
3.2.1 F-ratio	36
3.2.2 Scatter matrices and separability criteria	37
3.2.3 Bhattacharyya distance	38
3.2.4 Bhattacharyya shape	39
3.2.5 Divergence distance	39
3.2.6 Divergence shape	39
3.3 Speaker recognition with feature selection – state of the art	40
3.4 Proposed performance criterion for speaker verification	45
3.4.1 Gaussian goodness of fit test	47
3.4.2 The Recognition Related Criterion (RRC)	50
3.4.3 Generalized performance criterion	52
4. The Proposed Speaker Verification System	54
4.1 The training stage	55
4.1.1 Front-end processing and the global feature set	56
4.1.2 The individual feature selection procedure	58
4.1.3 Impostor selection for the feature selection procedure	59
4.2 The testing stage	61

5. Experiments and Results	62
5.1 Text-dependent speaker verification	62
5.1.1 Experimental setup	62
5.1.2 The text-dependent database	63
5.1.3 Results and discussion	63
5.2 Text-dependent speaker verification in noisy speech	68
5.2.1 Training on a noisy database	68
5.2.2 Training on the clean database	74
5.3 Text-independent speaker verification	77
5.3.1 Experimental setup	77
5.3.2 The text-independent database	77
5.3.3 Results and discussion	79
6. Conclusions and Future Work	83
References	86
Appendix A. CD-HMM	91
Appendix B. GMM	96

Abbreviations

ASV	Automatic Speaker Verification
BW	Baum-Welch
CMS	Cepstral Mean Subtraction
DET	Detection Error Trade-off
DP	Dynamic Programming
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
FA	False Accept (or False Alarm)
FFT	Fast Fourier Transform
FR	False Reject
FS	Feature Selection
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IBM	Individual Background Model
LAR	Log-Area Ratio
LPC	Linear Predictive Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
PDF	Probability Density Function
RRC	Recognition Related Criterion
SFFS	Sequential Floating Forward Sequence
SNR	Signal to Noise Ratio
STD	Standard Deviation
SVM	Support Vector Machine
UBM	Universal Background Model
VAD	Voice Active Detection
VQ	Vector Quantization

Notations

x	Scalar
$\mathbf{x}, \boldsymbol{\mu}$	Column Vector
$\boldsymbol{\Sigma}, \mathbf{O}$	Matrix
$\mathbf{a}^T, \mathbf{A}^T$	Transposition of Vector/Matrix
$\boldsymbol{\Sigma}^{-1}$	Inverse of matrix
$ \boldsymbol{\Sigma} $	Determinant
$\text{tr}[\boldsymbol{\Sigma}]$	Trace
X	Set
$p(\)$	Probability density function
$\log(\)$	Logarithm (base e)
$\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$E[\]$	Expectation operation
$\text{std}[\]$	Standard-deviation operation
$\text{erf}(\)$	Error function
$x(n)$	Sampled signal at discrete time n
$x(t)$	Signal at continuous time t
$X(z)$	z-transform of $x(n)$

List of figures

<i>Figure</i>	<i>Page</i>
1.1 Basic speaker recognition system	4
1.2 A basic block diagram of the proposed speaker verification system using individual feature space	6
2.1 A schematic diagram of the human speech production mechanism [Deller et al., 1993]	7
2.2 A block diagram of human speech production [Deller et al., 1993]	8
2.3 (a) A speech signal of the word “six”; (b) blowup of the initial /s/; (c) blowup of the vowel /i/	10
2.4 Block diagram of speech signal pre-processing and feature extraction	12
2.5 The Linear Vocal Tract Model for speech production [Deller et al., 1993]	15
2.6 The Mel scale [Deller et al., 1993]	20
2.7 Mel Frequency Filter Bank [Rabiner et al., 1993]	21
2.8 VQ based method	26
2.9 HMM structure	27
2.10 Speech-recognition-based methods	29
3.1 Feature subset selection using dynamic programming	34
3.2 Estimation of verification errors from target and impostors Gaussian-like histograms	46
3.3 Gaussian fit for the histogram of target (#3) and impostors’ scores	50
4.1 The proposed speaker verification system	54
5.1 Maximum RRC criterion as a function of the feature space dimension, k , for several feature selection procedures (for speaker #3)	64
5.2 Real EER test results of the different feature selection procedures in different feature space dimension (for speaker #3)	64
5.3 Number of feature occurrences in the individual selected feature subsets (ten targets)	66

5.4	Average DET curves of speaker verification results (feature spaces: global 120 features, 24 MFCC and Del MFCC space, and 24 individual optimal space)	67
5.5	Number of feature occurrences in the individual selected feature subsets (ten targets) – using the 20dB noisy database	70
5.6	Average DET curves of speaker verification results in 20dB noisy database	71
5.7	Number of feature appearance in the individual selected feature subsets (ten targets) – using the 5dB noisy database	72
5.8	Average DET curves of speaker verification results in 5dB noisy database	73
5.9	Average EER vs. SNR of the two verification systems	73
5.10	Average DET curves of speaker verification results in 20dB noisy testing database (using clean training database)	74
5.11	Average DET curves of speaker verification results in 5dB noisy testing database (using clean training database)	75
5.12	Average EER vs. SNR of the two verification systems, using noisy testing database	76
5.13	Number of feature occurrences in the individual selected feature subsets	80
5.14	Average DET curves of text-independent speaker verification results using individual thresholds (feature spaces: 24 MFCC and Del MFCC space, and 24 individual optimal space)	81

List of tables

<i>Table</i>	<i>Page</i>
3.1 The corresponding probability falling in each subinterval, the expected number of points falling in each subinterval and the actual number of points falling in each subinterval, for one example target's scores (speaker #3) and impostors' scores from 39 speakers	48
4.1 The features and their symbols	57
5.1 Selected features for the first 5 target speakers	65
5.2 Average equal error rate of the verification results	68
5.3 Selected features for the first 5 target speakers on 20dB noisy database	69
5.4 Selected features for the first 5 target speakers (text-independent)	79

Abstract

Today's Automatic Speaker Verification (ASV) systems use a common feature space for all speakers. This common set of features is usually a set of cepstral and delta-cepstral coefficients, which are used for speech recognition tasks. This research is based on the assumption that every speaker has his own 'optimal' feature space, which optimally discriminates him from other speakers. The goal of this work is to demonstrate the significance of employing an individual feature space in modern Continuous Density Hidden Markov Model (CD-HMM) or Gaussian Mixture Model (GMM) based-verification systems.

In this research, a new criterion for feature selection was developed, which is suitable for speaker verification tasks and correlated with the recognition rate, named "Recognition Related Criterion" (RRC). Several feature selection procedures were implemented and tested along with the new feature selection criterion, such as k-best, forward selection, Sequential Floating Forward Sequence (SFFS), and Dynamic Programming (DP). We also implemented two speaker verification systems that combine an individual feature selection algorithm: the first is a text-dependent verification system, based on an HMM classifier and the second is a text-independent verification system, based on a GMM classifier.

The proposed HMM-based verification system was evaluated on a text-dependent database. A significant improvement over the "standard" Mel Frequency Cepstrum Coefficients (MFCC) space ($12 \text{ MFCC} + 12 \Delta\text{MFCC}$) in verification results was demonstrated with the selected individual feature space. An EER of 0.7% was achieved when the feature set was the MFCC space. Under the same conditions, the system based on the selected individual feature space (order of 24) yielded an EER of only 0.48%. It was found that the two best selection procedures are the DP and the SFFS. However, the SFFS

was more efficient than the DP in terms of calculation load. This system was also evaluated on a noisy text-dependent database. It was found that when using an individual feature space the verification system becomes very sensitive to changing environmental conditions, meaning, different Signal to Noise Ratios (SNR) for training and testing database.

The proposed GMM-based verification system was evaluated on a text-independent database. A significant improvement in verification results was demonstrated with the selected individual feature space. An EER of 6.14% was achieved when the feature set was the MFCC space. Under the same conditions, a system based on the selected feature space yielded an EER of only 4.15%.

Keywords:

Speaker Verification, Feature Selection, Individual Feature Space, Equal Error Rate (EER), Recognition Related Criterion (RRC), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM).