Robust Speaker Identification

by

Smarajit Bose

Interdisciplinary Statistical Research Unit
Indian Statistical Institute, Kolkata

Joint work with Amita Pal and Ayanendranath Basu

Overview

- Speaker Identification Problem
- Existing features/models, limitations
- Improving accuracy by
  - Combination of classifiers
  - Principal component transformation
  - Robust statistical procedures
- Results

Automatic Speaker Recognition

- The use of a machine to recognize a person from a spoken phrase
- Two different problems:
  - To identify a particular person
  - To verify a person's claimed identity

Motivation and Applications

- Security
  - Access control as a component of a biometric identification system
    - Phone banking
    - Password-less or card-less access
- Forensics
- Authentication of speech recordings
- ……

Voice as a Biometric

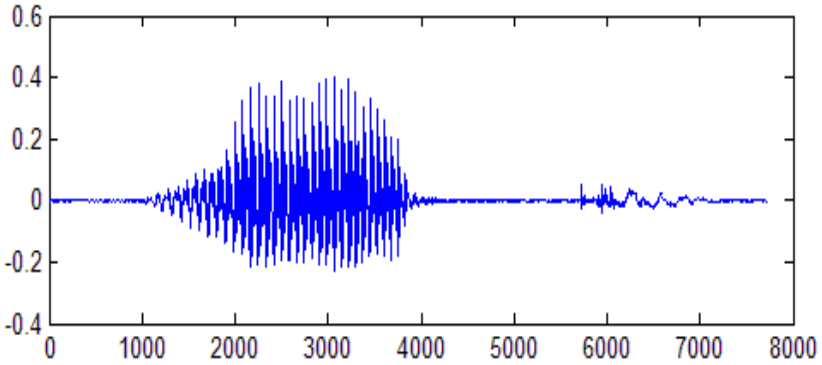
- Not as reliable as fingerprints
- Accuracy can increase if other biometrics are used in conjunction
- Collection of information easier
  - Requires relatively cheaper equipment
  - Non-invasive

Issues in Speaker identification

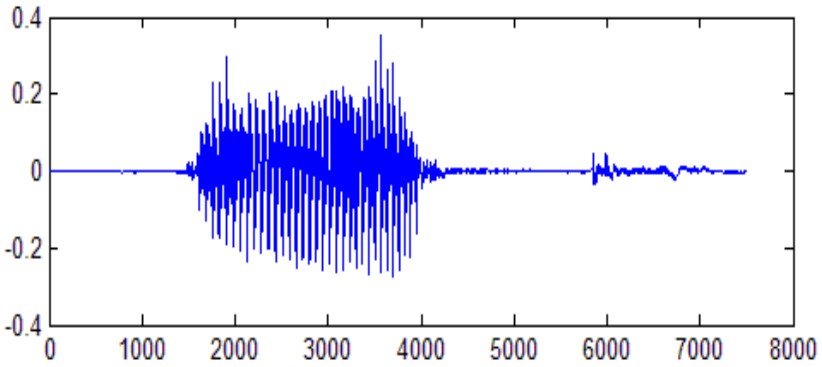
- Text-dependent vs. Text-independent
- Closed-set vs. Open-set
- Cooperative vs. Non-cooperative speakers
- Quality of speech
- Duration of Speech
- Channel of recording/transmission

Speaker Identification Example

Speaker no. 1



Speaker no. 2



Basic Components of a Speaker Recognition System

- Features
- Speaker Models
  - Specification
  - Parameter estimation
- Matching Criteria or classification rules

Features for Speaker Recognition

- Domain of Signal Processing experts
- Most successful features are essentially
  - short-duration (computed on frames of a few milliseconds duration)
  - carry spectral information, widely believed to be speaker-specific
- Commonly used features:
  - Mel Frequency Cepstral Coefficients (MFCCs)
  - Linear Prediction Cepstral Coefficients (LPCCs)

Mel-Frequency Cepstrum (MFC)

- A representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear *mel* scale of frequency.

- **Mel-frequency cepstral coefficients** (**MFCCs**) are coefficients that collectively make up an MFC.

- Based on a bank of filters, a set of MFCCs are computed as

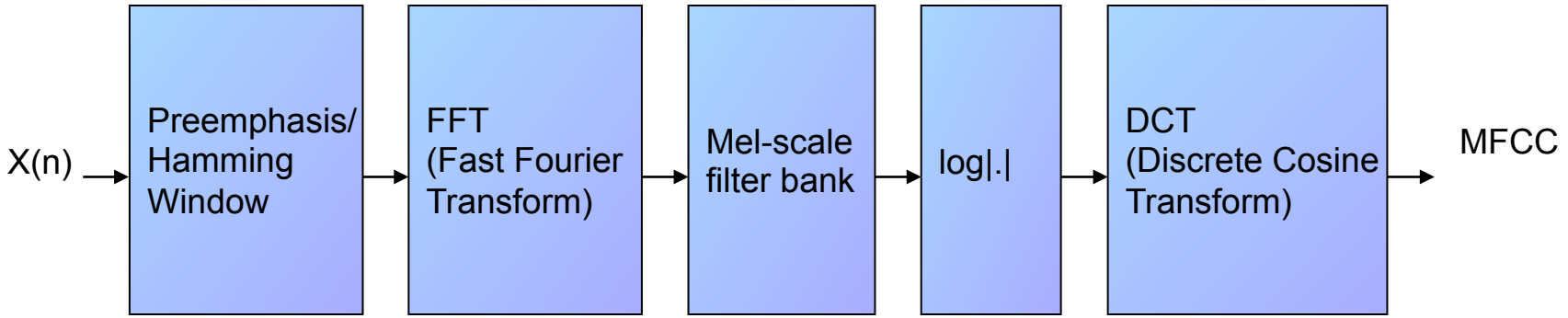
$$MFCC_i = \sum_{k=1}^{K} X_k \cos\left[i\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \quad i = 1,2,\ldots,M$$

- X_k being the log-energy output of the kth filter.

Computation of MFCCs

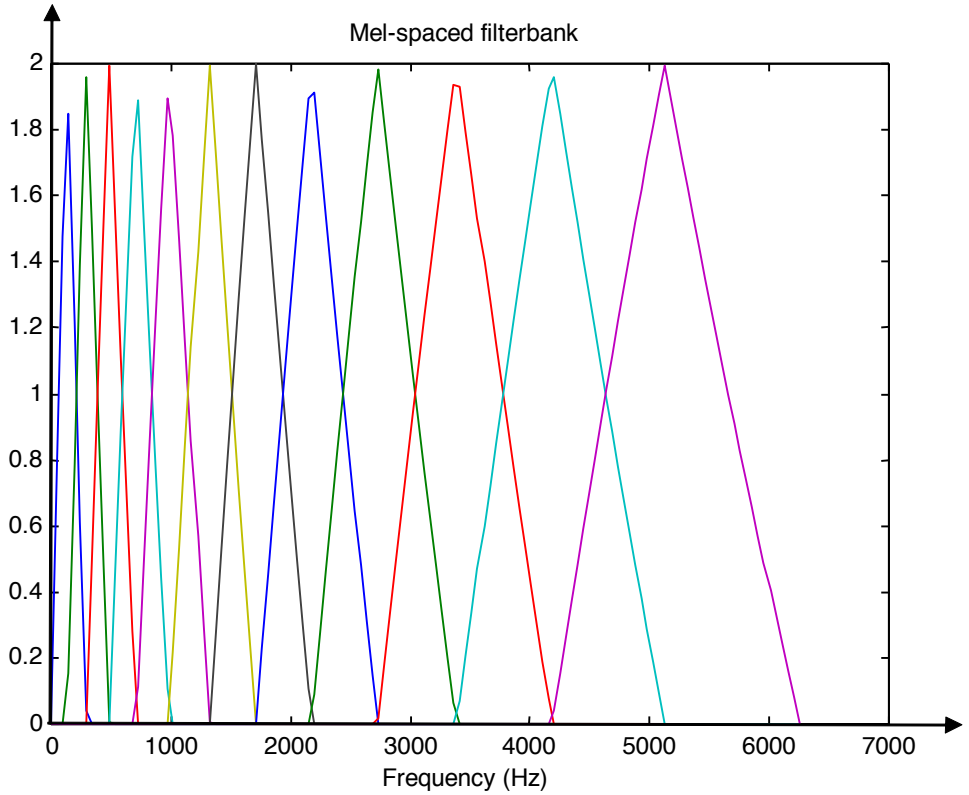
- Partition the speech signal into overlapping segments or frames
- Take the Fourier transform of signal from each frame.
- Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
- Take the logs of the powers at each of the mel frequencies.
- Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.

Flow Chart for MFCC Computation



MFCC Filter Bank

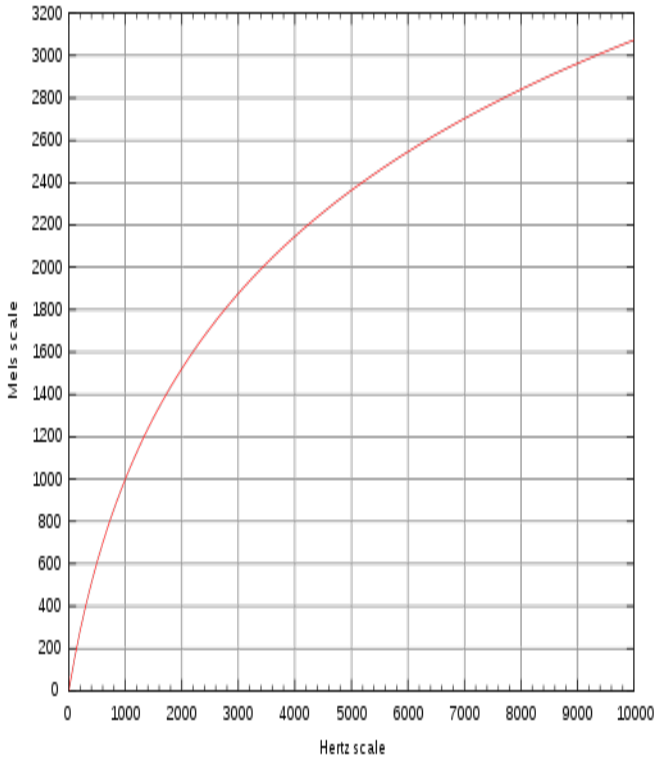
One way to simulating the spectrum is by using a filter bank, spaced uniformly on the mel scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval.



Mel Scale

- A scale of pitches judged by listeners to be equal in distance from one another
- **Mel** comes from the word **melody** to indicate this
- A popular formula to convert *f* hertz into *m* mel is:

$$m = 2595 \log(1 + f / 700)$$



Why Use the Mel Scale?

- Psychophysical studies show that human perception of the frequency contents of sounds for speech signals does not follow a linear scale.
- For each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on the so-called 'mel' scale.
- The *mel-frequency* scale is
  - a linear frequency spacing below 1000 Hz and
  - a logarithmic spacing above 1000 Hz

Successful Statistical Speaker Models

- Gaussian Mixture Models
- Vector Quantization
- Hidden Markov Models

GMMs as Speaker Models

- If $\boldsymbol{x}$ is the *D*-dimensional feature vector, then for a C-speaker problem, a given speaker is modeled as a mixture of *N* component densities

$$p(\boldsymbol{x}\,|\,\boldsymbol{\lambda}) = \sum_{i=1}^{N} p_i\, b_i(\boldsymbol{x}\,|\,\boldsymbol{\theta}_i) \qquad \sum_{i=1}^{N} p_i = 1.$$

- p_i is the prior probability for the *i*th component,
- $b_i(\cdot\,|\,\boldsymbol{\theta}_i)$ is the probability density of $\boldsymbol{x}$ in the *i*th component.
- $\boldsymbol{\lambda} = \{p_i, \boldsymbol{\theta}_i, i = 1,2,\ldots,N\}$ is the collection of unknown parameters

Speaker Recognition with GMMs built on MFCCs

• Each speech sample (training as well as test) is split into a number of overlapping segments or frames, with MFCCs computed from each segment.

• GMM models for all speakers are trained by the **Expectation Maximization algorithm**, generally assuming diagonal covariance matrices.

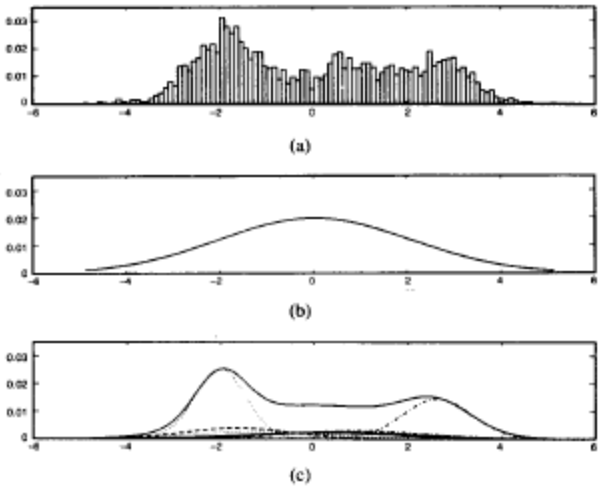
• Likelihood function for the unknown sample is computed, based on MFCC vectors obtained from all frames, assuming independence.

•The unknown sample is classified by the **Maximum Likelihood rule**.

•Spectacular performance reported by Reynolds (1995) with mixtures of 32 Gaussians as speaker models with diagonal covariance matrices.

Explanation for the Success of the GMM-MFCC approach

Individual component Gaussians represent broad acoustic classes which reflect some speaker-dependent vocal tract properties



Proposed approaches:

- **Use of the Principal Component Transformation** (PCT) on MFCC features before building individual speaker models
  - Independence of MFCC features is a questionable assumption.
  - Correlation structures vary from speaker to speaker.
- MFCC vectors computed from each frame of a test utterance are transformed by the PCT corresponding to each speaker model before matching.
- **Ensemble classification**
- **Use of robust statistical methods**

Principal Component Transformation

- A widely-used linear orthogonal transformation for converting a set of observations on possibly correlated variables into a set of observations on linearly uncorrelated variables called **Principal Components**.
- Since correlation structures differ from speaker to speaker, these transformations are also different for different speakers.

PCT (continued)

- The GMM for a particular speaker is fitted using the MFCCs transformed by the principal component transformations for that speaker.
- For testing, to determine the likelihood values with respect to a given target speaker model, the MFCCs from the test utterance are transformed by the principal component transformation corresponding to that speaker.
- Dimensionality reduction is not the primary objective of this work, and all *D* principal components have been used.

Ensemble classification: Common Techniques

- Bagging

  Classifiers built from many bootstrap samples

- Boosting

  Classifiers built by varying weights of observations

Here classifiers are built by varying different parameters of GMM-MFCC model

Ensemble Classification

- To improve classification accuracy, the outcomes of a number of competing classifiers can be combined, for example, by
  - majority voting
  - combining likelihood values from different classifiers suitably
  - ranking the speakers that are most likely from each classifier, to come up with a combined rank for making a final decision.
- In this work, likelihoods computed for a number of classifiers were combined by simple averaging.

Building Different Classifiers

With the same training data, it is possible to build distinct MFCC-GMM classifiers, for example, by

- discarding low energy frames: different threshold values yielded varying (and sometimes better) results.
- using different ranges of frequency of the cepstrum: experimentation with various frequency ranges led to varying (and sometimes better) results.

Parameters tweaked in this work

- Thresholds for frame energy (both training and testing): 0.0, 0.01, 0.014, 0.0141, 0.015
- Number of MFCCs : 20, 25, 28, 30, 35, 38, 40, 42
- Number of Filters: 24, 25, 28, 30, 35, 36, 38, 40, 42
- Minimum frequency: 0, 200,100
- Maximum frequency: 5500, 4000,6000

Matching the test utterance

- The test utterance is again broken into frames, feature vectors are generated from each frame.
- Likelihood values of each of these vectors are calculated
- Assuming independence product of these values is taken as the matching score.
- The model with the highest matching score was selected.

- This implies that if $\hat{f}_j$ is the j-th speaker GMM and f^* is the density of the new utterance from which the feature vectors $x_1,\ldots,x_N$ are produced, we maximize over j

$$\sum_{i=1}^{N} \log \hat{f}_j(x_i) \qquad \text{or equivalently,} \qquad \frac{1}{N}\sum_{i=1}^{N} \log \hat{f}_j(x_i)$$

Instead we can used some robust estimates like median or trimmed mean. In our experiments, trimmed mean worked much better.

Benchmark Data Sets

TIMIT and NTIMIT

- **TIMIT:** an speech corpus in English Owned by the Linguistic Data
- Consortium (LDC), University of Pennsylvania.
- Clean microphone recordings of 10 different read sentences.
- 630 speakers (438 males and 192 females) from eight major dialect regions of the USA.
- **NTIMIT**: the TIMIT database played through a carbon-button telephone handset.

ISIS and NISIS

- Simultaneously-recorded microphone and telephone speech
- Recorded in the Indian Statistical Institute, Kolkata
- 105 speakers (53 male + 52 female) Multiple sessions (four, with gaps
- between sessions of 1 week to 2
- months)
- Spontaneous as well as read speech
- In two languages (Bangla and English) Recorded in a typical office environment
- with moderate background noise

Results with NTIMIT

Using trimmed mean?	Number of Speakers	Data set	% Recognition score with			
			GMM	PCT-GMM	Combination of	
					GMMs	PCT-GMMs
No	100	6:4	48.0	50.0	53.0	55.3
Yes	100	6:4	50.0	52.0	56.3	61.3
No	630	6:4	34.0	40.0	40.4	47.2
Yes	630	6:4	36.0	43.0	42.0	48.8
No	100	8:2	51.0	53.0	58.0	61.5
Yes	100	8:2	59.0	60.0	63.0	67.0
No	630	8:2	41.0	48.0	48.5	56.4
Yes	630	8:2	43.0	50.0	49.8	58.3

Results with NISIS

Using trimmed mean?	Data set	% Recognition score on NISIS with			
		GMM	PCT-GMM	Combination of	
				GMMs	PCT-GMMs
No	**6:4**	**69.0**	**84.0**	**72.5**	**85.8**
Yes		**72.0**	**86.0**	**75.8**	**89.3**
No	**8:2**	**76.0**	**89.0**	**77.0**	**91.5**
Yes		**77.0**	**91.0**	**78.5**	**92.0**

Relative Performance: NTIMIT vis-a-vis NISIS

CORPUS	Using trimmed mean?	Data set	% Recognition score with			
			GMM	PCT-GMM	Combination of	
					GMMs	PCT-GMMs
NTIMIT	**No**	**6:4**	**48.0**	**50.0**	**53.0**	**55.3**
	Yes		**50.0**	**52.0**	**56.3**	**61.3**
NISIS	**No**		**69.0**	**84.0**	**72.5**	**85.8**
	Yes		**72.0**	**86.0**	**75.8**	**89.3**
NTIMIT	**No**	**8:2**	**51.0**	**53.0**	**58.0**	**61.5**
	Yes		**59.0**	**60.0**	**63.0**	**67.0**
NISIS	**No**		**76.0**	**89.0**	**77.0**	**91.5**
	Yes		**77.0**	**91.0**	**78.5**	**92.0**

Ongoing Work and Future Directions

- Recall that for identification of the speaker of a test utterance
  - The utterance is split into frames, and feature vectors are generated from each frame.
  - Likelihood values of each of these vectors are calculated.
  - Assuming independence, the product of these values is taken as the matching score.
  - The speaker model with the highest matching score is selected.
- If $\hat{f}_j$ is the GMM for the *j*-th speaker and f^* is the pdf from which the feature vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ computed from the test utterance arise, then for speaker recognition,

$$\sum_{i=1}^{N} \log \hat{f}_j(x_i)$$

is maximized.

- This implies that if $\hat{f}_j$ is the j-th speaker GMM and f^* is the density of the new utterance from which the feature vectors $x_1,\ldots,x_N$ are produced, we maximize over j $\sum_{i=1}^{N} \log \hat{f}_j(x_i)$ which is equivalent to maximizing

$$\frac{1}{N}\sum_{i=1}^{N} \log \hat{f}_j(x_i) = \int \log \hat{f}_j \, dF_n$$ where F_n is the empirical cdf

This is an estimate of $\int f^* \log \hat{f}_j$, maximizing which is equivalent to minimizing $\int f^* \log \frac{f^*}{\hat{f}}$, the KL divergence

This is not robust, particularly in noisy situations. We can perhaps do better with Hellinger Distance.

References

- **D. Reynolds and R. Rose**, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models,” *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83, 1995.
- **D. Reynolds, “**Large Population Speaker Identification Using Clean and Telephone Speech”, *IEEE Signal Processing Letters*, Vol. 2, No. 3,1995.
- **N. Z. Tishby**, “On the application of mixture AR hidden Markov models to text independent speaker recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 39, no. 3, pp. 563–570, 1991.
- **F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B.-H. Juang**, “A vector quantization approach to speaker recognition,” *AT&T Technical Journal*, vol. 66, no. 2, pp. 14–26, 1987.

Thank you!

Building Different Classifiers

With the same training data, it is possible to build distinct GMM-MFCC classifiers, for example, by

- changing methods of calculating frame energy
- discarding low energy frames
  - Different threshold values yield varying (and sometimes better) results.
- using different ranges of frequency
  - Experimentation with various frequency ranges led to varying (and sometimes better) results.

Combining Classifiers

The results of implementing the different classifiers so constructed can be combined by

- majority voting
- likelihood values from different classifiers can be combined in a suitable way
- ranking the speakers that are most likely from each classifier, to come up with a combined rank for making a final decision.

Parameters tweaked in our work

- Thresholds for frame energy(both training and testing): {0.0, 0.01, 0.014, 0.0141, 0.015}
- Number of MFCCs :{20, 25, 28, 30, 35, 38, 40, 42}
- Number of Filters:{24, 25, 28, 30, 35, 36, 38, 40, 42}
- Minimum frequency:{0, 200,100}
- Maximum frequency:{5500, 4000,6000}
- Percentage of reduction of frames (for test):{12, 13, 13.5, 14,14.5, 15, 16, 18}

Significance of parameters tweaked

- Frame energy: For each speech frame, the energy at time n for the l-th mel-scale filter is

$$E_{mel}(n,l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l} |V_l(\omega_k)\,Y(n,\omega_k)|^2$$

  - where V_l is the response of the lth mel-scale filter, and

$$Y(n,\omega_k) = x[m]w[n-m]e^{-j\frac{2\pi}{k}m} \qquad A_l = \sum_{k=L_l}^{U_l} |V_l(\omega_k)|^2$$

  - $x[m]$, $w[m]$ are respectively the speech signal and a window function. At time point m.
- Frequency range: $\min\{L_l, l = 1,2,\ldots L\}$ to $\max\{U_l, l = 1,2,\ldots L\}$
- Frame reduction: discarding non-informative frames

Results with 100 NTIMIT Speakers (6:4)

Experiment	Threshold for frame energy	No. of MFCCs	No. of filters	Minimum frequency	Maximum frequency	No. of correctly classified test utterances	Accuracy %
Benchmark	0	20	36	200	4000	144	36
1.1	0.015	20	36	200	4000	150	37.5
1.2	0	25	24	0	5500	126	31.5
1.3	0.015	25	24	0	5500	136	34
1.1-1.3 Combined						168	42

Results with 100 NTIMIT Speakers (8:2)

Experiment	Threshold for frame energy	No. of MFCCs	No. of filters	Minimum frequency	Maximum frequency	No. of correctly classified test utterances	Accuracy %
Benchmark	0	25	24	0	5500	77	38.5
2.1	0.015	25	24	0	5500	92	46
2.2	0	20	36	200	4000	82	41
2.3	0.015	20	36	200	4000	86	43
2.1-2.3 combined						94	47

Results with 630 NTIMIT Speakers (6:4)

Experiment	Threshold for frame energy	No. of MFCCs	No. of filters	Minimum frequency	Maximum frequency	No. of correctly classified test utterances	Accuracy %
Benchmark	0	20	36	200	4000	611	24.25
1.1	0.015	20	36	200	4000	744	29.52
1.2	0	25	24	0	5500	651	25.83
1.3	0.015	25	24	0	5500	780	30.95
1.1-1.3 Combined						848	33.65

Results with 630 NTIMIT Speakers (8:2)

Experiment	Threshold for frame energy	No. of MFCCs	No. of filters	Minimum frequency	Maximum frequency	No. of correctly classified test utterances	Accuracy %
Benchmark	0	25	24	0	5500	379	30.08
2.1	0.01	25	24	0	5500	477	37.86
2.2	0.015	25	24	0	5500	491	38.97
2.3	0.01	25	24	200	4000	457	36.27
2.1-2.3 Combined						519	41.19

Results with 630 NTIMIT Speakers (8:2)

Experiment	Threshold for frame energy	No. of MFCCs	No. of filters	Minimum frequency	Maximum frequency	No. of correctly classified test utterances	Accuracy %
Benchmark	0	25	24	0	5500	379	30.08
3.1	0.01	25	24	0	5500	477	37.86
3.2	0.015	25	24	0	5500	491	38.97
3.3	0.01	25	24	200	4000	457	36.27
3.4	0.015	20	36	200	4000	443	35.16
3.1-3.4 Combined						537	42.62

Improved Speaker Recognition on NTIMIT

By Principal Component Analysis

Independence of MFCCs

- Questionable assumption
- Correlation structures different for speakers

Idea:

- Use principal component transformation before building individual speaker models
- Transform test utterance by corresponding PC before matching with a speaker model

Results with 100 NTIMIT Speakers (8:2)

Experiment	Threshold for frame energy	No. of MFCCs	No. of Eigen Vectors Taken	Minimum frequency	Maximum frequency	Percentage of reduction of frames (for test)	No. of correctly classified test utterances	Accuracy %
1.1	0.0	38	37	0	5500	13.5	101	50.50
1.2	0.0	38	37	0	5500	15	103	51.50
1.3	0.0	38	37	200	4000	13.5	110	55.00
1.4	0.0	38	37	200	4000	15	108	54.00
1.5	0.01	38	37	0	5500	13.5	107	53.50
1.6	0.01	38	37	0	5500	15	107	53.50

Results with 100 NTIMIT Speakers (8:2) (contd.)

Experiment	Threshold for frame energy	No. of MFCCs	No. of Eigen Vectors Taken	Minimum frequency	Maximum frequency	Percentage of reduction of frames (for test)	No. of correctly classified test utterances	Accuracy %
1.7	0.01	38	37	200	4000	13.5	115	57.50
1.8	0.01	38	37	200	4000	15	117	58.50
1.9	0.0141	38	37	0	5500	13.5	108	54.00
1.10	0.0141	38	37	0	5500	15	107	53.50
1.11	0.0141	38	37	200	4000	13.5	111	55.50
1.12	0.0141	38	37	200	4000	15	112	56.00

Results with 100 NTIMIT Speakers (8:2) (contd.)

Experiment	Threshold for frame energy	No. of MFCCs	No. of Eigen Vectors Taken	Minimum frequency	Maximum frequency	Percentage of reduction of frames (for test)	No. of correctly classified test utterances	Accuracy %
1.13	0.015	38	37	0	5500	13.5	111	55.50
1.14	0.015	38	37	0	5500	15	110	55.00
1.15	0.015	38	37	200	4000	13.5	115	57.50
1.16	0.015	38	37	200	4000	15	113	56.50
1.17	0.011	38	37	200	4000	13.5	***118***	***59.00***
1.18	0.011	38	37	200	4000	15	***118***	***59.00***

Ensemble Classification after PCA

Combined Experiment	Number of Speakers	No. of correctly classified test utterances	Accuracy %
1.6, 1.18	100	123	61.50
1.3, 1.12, 1.14	100	124	62.00
1.5, 1.9, 1.17, 1.18	100	***125***	***62.50***
1.1, 1.3, 1.10, 1.11, 1.13	100	***125***	***62.50***

Results with 630 NTIMIT Speakers (8:2)

Experiment	Threshold for frame energy	No. of MFCCs	No. of Eigen Vectors Taken	Minimum frequency	Maximum frequency	Percentage of reduction of frames (for test)	No. of correctly classified test utterances	Accuracy %
2.1	0.0	38	37	0	5500	13.5	565	44.84
2.2	0.0	38	37	0	5500	15	566	44.92
2.3	0.0	38	37	200	4000	13.5	615	48.81
2.4	0.0	38	37	200	4000	15	618	49.05
2.5	0.01	38	37	0	5500	13.5	603	47.86
2.6	0.01	38	37	0	5500	15	590	46.83
2.7	0.01	38	37	200	4000	13.5	649	51.51
2.8	0.01	38	37	200	4000	15	642	50.95

Results with 630 NTIMIT Speakers (8:2) (contd.)

Experiment	Threshold for frame energy	No. of MFCCs	No. of Eigen Vectors Taken	Minimum frequency	Maximum frequency	Percentage of reduction of frames (for test)	No. of correctly classified test utterances	Accuracy %
2.9	0.0141	38	37	0	5500	13.5	595	47.22
2.10	0.0141	38	37	0	5500	15	595	47.22
2.11	0.0141	38	37	200	4000	13.5	***659***	***52.30***
2.12	0.0141	38	37	200	4000	15	957	52.14
2.13	0.015	38	37	0	5500	13.5	578	45.87
2.14	0.015	38	37	0	5500	15	587	46.59
2.15	0.015	38	37	200	4000	13.5	652	51.75
2.16	0.015	38	37	200	4000	15	644	51.11

Ensemble Classification after PCA

Combined Experiment	Number of Speakers	No. of correctly classified test utterances	Accuracy %
2.5, 2.11	630	687	54.52
2.3, 2.10, 2.12	630	700	55.56
2.3, 2.7, 2.9, 2.16	630	***701***	***55.63***
2.3, 2.4, 2.8, 2.14, 2.15	630	***701***	***55.63***

Summary of Results

Number of Speakers → / Data Set ↓	100						630					
	Previous	Present		Previous Combined	Present Combined		Previous Combined	Present		Previous Combined	Present Combined	
		GMM	PCA-GMM		GMM	PCA-GMM		GMM	PCA-GMM		GMM	PCA-GMM
8:2	46.00	56.00	59.00	47.00	61.50	62.50	38.97	42.14	52.30	42.62	49.05	55.63
6:4	37.50	48.25	50.75	42.00	53.00	55.25	30.95	34.96	42.26	33.65	40.36	45.99