On the design of automatic voice condition analysis systems. Part II: review of speaker recognition techniques and study on the effects of different variability factors.

by J. A. Gómez-García, L. Moro-Velázquez &
J. I. Godino-Llorente

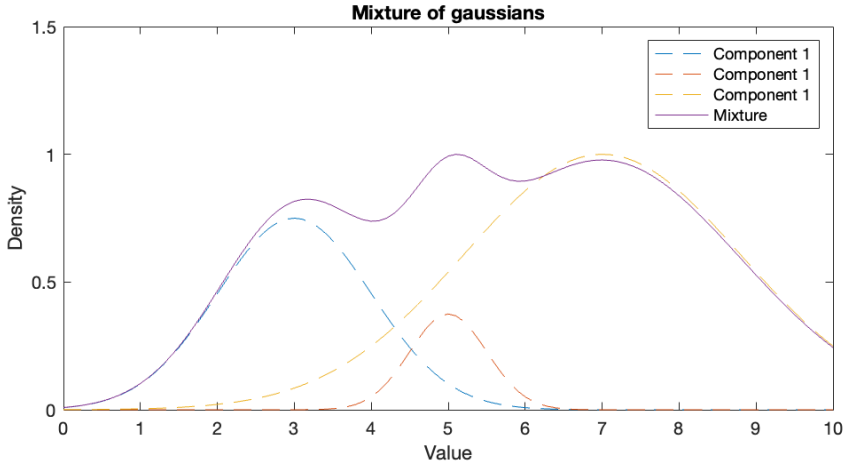
Review by Meghna Ranjit

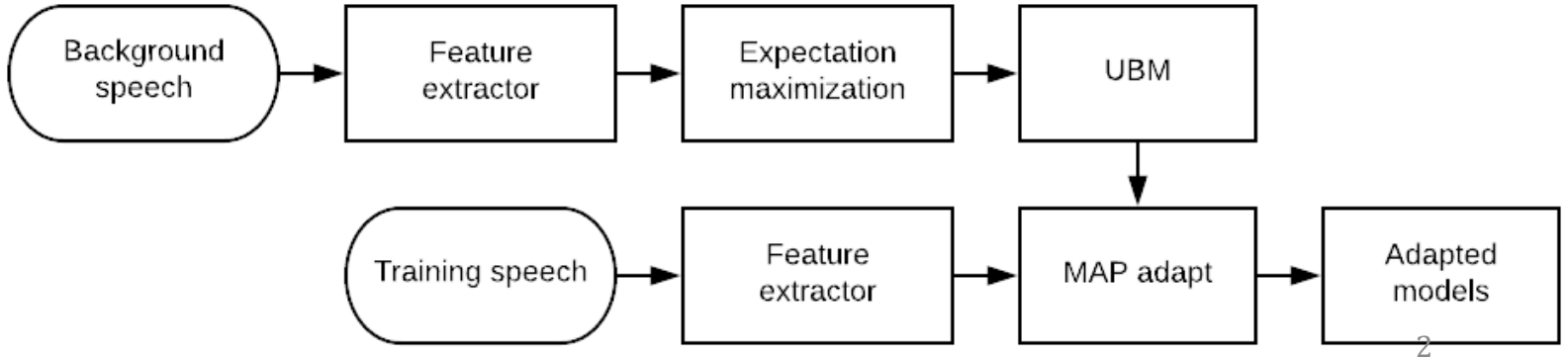
Objective of the work

- Automatic voice condition analysis provides the advantages of speed and non-invasiveness as compared to to traditional detection procedures.
- First paper: review of concepts & an insight to the state of the art
- Aim of this study is to examine variability factors affecting the robustness of systems
- Experiments are performed to test out the influence of
  - the speech task
  - extralinguistic aspects (such as sex)
  - the acoustic features
  - the classifiers

Methodologies used

- GMM represent the probability density function of a dataset by means of a linear combination of Gaussian components.
- Via GMM, a large auxiliary dataset can be modelled, termed as the Universal Background Model (UBM)
- The UBM which serves as an initialisation models, from which from which specific models are adapted using the training data



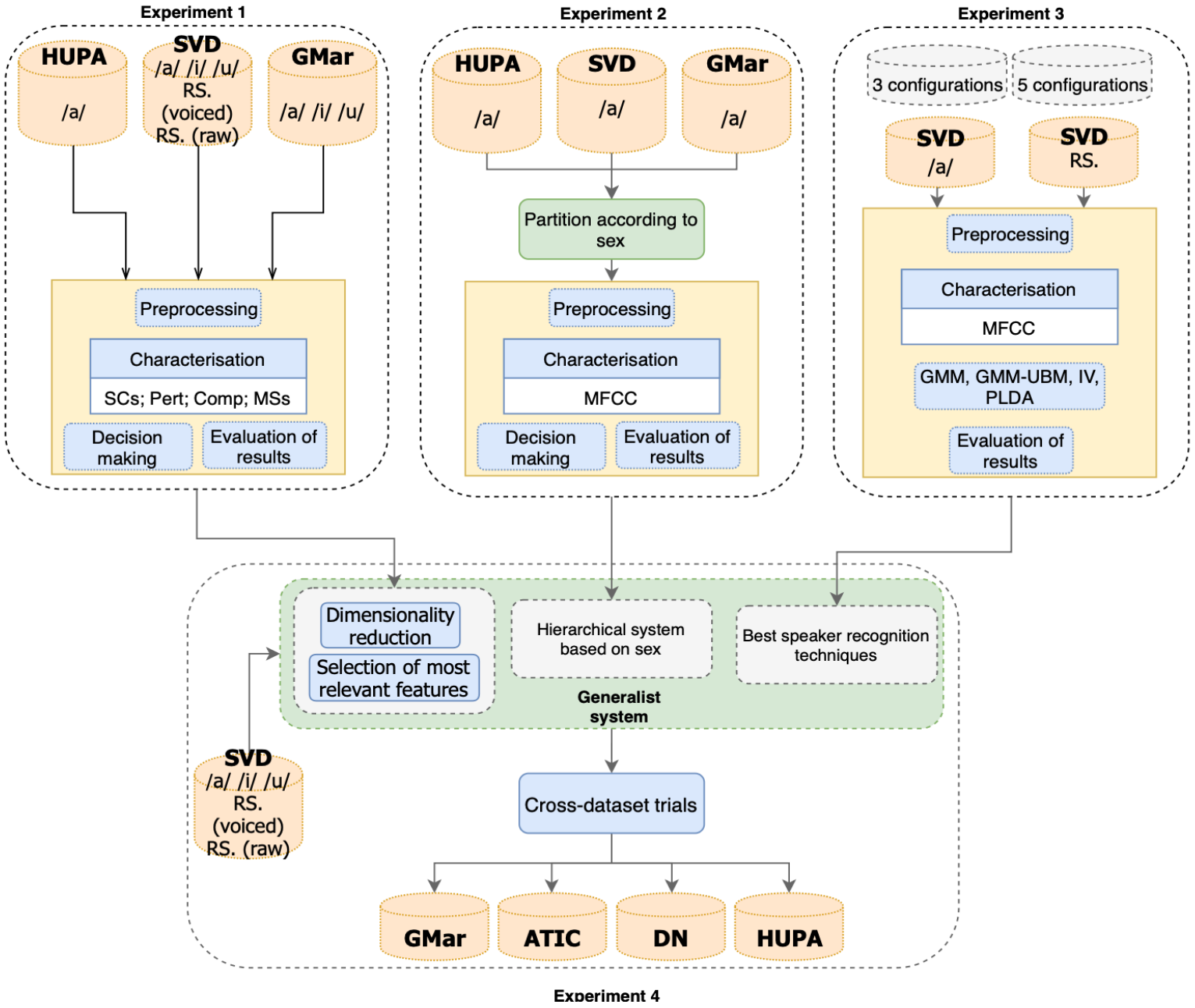


Methodologies used

- A variation to GMM-UBM is termed GMM-SVM, which is aimed at combining the discriminatory capabilities of SVM into the GMM framework.
- Likewise, a further improvement to the GMM-UBM are the i-Vectors (IV), which rely on the concept of GMM-UBM and factorial analysis for modelling the training dataset in a total variability space.
- I-Vectors are often accompanied by a Probabilistic Linear Discriminative Analysis (PLDA), which seeks to compensate for the effects of variability factors in the training data

Databases utilized

- Three datasets containing normophonic and pathological recordings are used as training corpora:
  - Hospital Universitario Prıncipe de Asturias (HUPA): only vowel vowel /a/
  - Gregorio Marañón Hospital (GMar) : /a/, /i/ and /u/
  - Saarbrücken (SVD) voice disorders corpora: /a/, /i/ and /u/, running speech
- Similarly, four datasets are also utilised for the construction of the UBM.
  - EUROM: multiple languages, read speech
  - PhoneDat-I corpora, German corpus, read speech
  - Massachussets Ear and Eye Infirmary (MEEI), English corpus, read speech
  - The Albayzin dataset, Spanish corpus, read speech
- Two extra corpora are used for cross-dataset trials:
  - The Hospital Doctor Negrin dataset (DN), vowel /a/
  - The Aplicaciones de las Tecnologías de la información y la comunicación corpora.


Experiment 1
HUPA
/a/
SVD
/a/ /i/ /u/
RS.
(voiced)
RS. (raw)
GMar
/a/ /i/ /u/
Preprocessing
Characterisation
SCs; Pert; Comp; MSs
Decision making
Evaluation of results
Experiment 2
HUPA
/a/
SVD
/a/
GMar
/a/
Partition according to sex
Preprocessing
Characterisation
MFCC
Decision making
Evaluation of results
Experiment 3
3 configurations
5 configurations
SVD
/a/
SVD
RS.
Preprocessing
Characterisation
MFCC
GMM, GMM-UBM, IV, PLDA
Evaluation of results
Dimensionality reduction
Selection of most relevant features
Hierarchical system based on sex
Best speaker recognition techniques
Generalist system
SVD
/a/ /i/ /u/
RS.
(voiced)
RS. (raw)
Cross-dataset trials
GMar
ATIC
DN
HUPA
Experiment 4

Experiment 1: variability due to the acoustic material and the feature set

- A total of 9 trials are carried out using
  - the HUPA dataset (sustained phonation of the vowel /a/);
  - 3 trials with GMar (sustained phonations of vowels /a/, /i/, /u/);
  - 5 with SVD (sustained phonations of vowels /a/, /i/, /u/, raw running speech, and voiced segments of the running speech task).
- Pre-processing steps
  - Down-sampled to 20 kHz
  - Max-normalised
  - Framed and Hamming windowed
- GMM classifier used
- Metrics: AUC, Accuracy, sensitivity, specifity

Features

- Perturbation features (Pert set): measure the presence of additive noise resulting from an incomplete glottal closure of the vocal folds, and the presence of modulation noise which is the result of irregularities in the movements of the vocal folds.
  - Normalised Noise Entropy (NNE), Cepstral Harmonics-to-Noise Ratio (CHNR) and Glottal-to-Noise Excitation Ratio (GNE).
- Spectral and cepstral features (SCs set): measure the harmonic components of the voice.
  - Perceptual Linear Prediction coefficients (PLP) Mel-Frequency Cepstral Coefficients (MFCC), Smoothed Cepstral Peak Prominence (CPPS) and Low-to-High Frequency Spectral Energy Ratio (LHr).

Features

- Features based on modulation spectrum (MSs set): characterises the modulation and acoustic frequencies of input voices
  - Modulation Spectrum Homogeneity (MSH), Cumulative Intersection Point (CIL), Rate of Points above Linear Average (RALA) and
  - Modulation Spectrum Percentile $(MSP)_m$, where the sub-index is referred to the percentile that is used, i.e. MSP_{25}, MSP_{75} and MSP_{95}
- Complexity (Comp set): characterises the dynamics of the system and its structure
  - Dynamic invariants (Dyn subset): eg. Recurrence Period Density Entropy (RPDE)
  - Features which measure long-range correlations (LR subset): such as Detrended Fluctuation Analysis (DFA)
  - Regularity estimators (Reg subset): such as Approximate Entropy (ApEn)
  - Entropy estimators (Ent subset): Permutation Entropy (PE)

Results

HUPA database:

Set	Subset	ACC	SP	SP	AUC
Pert	-	76.61 ± 4.30	0.77	0.77	0.85
MSs	-	71.77 ± 4.57	0.74	0.70	0.79
SCs	CPPS+LHr	62.10 ± 4.93	0.60	0.64	0.69
	MFCC	69.62 ± 4.67	0.66	0.74	0.79
	PLP	66.94 ± 4.78	0.58	0.77	0.80
Comp	LR	56.18 ± 5.04	0.51	0.62	0.60
	Dyn	65.59 ± 4.83	0.63	0.68	0.75
	Reg	69.89 ± 4.66	0.68	0.72	0.78
	Ent	75.00 ± 4.40	0.75	0.75	0.83

GMar database:

		Vowel /a/				Vowel /i/				Vowel /u/			
Set	Subset	ACC	SP	SE	AUC	ACC	SP	SE	AUC	ACC	SP	SE	AUC
Pert	-	65.35 ± 6.56	0.65	0.65	0.77	66.32 ± 6.72	0.65	0.68	0.73	61.36 ± 7.19	0.64	0.59	0.70
MSs	-	67.82 ± 6.44	0.65	0.70	0.76	67.82 ± 6.44	0.65	0.70	0.76	63.64 ± 7.11	0.63	0.64	0.72
SCs	CPPS+LHr	66.83 ± 6.49	0.67	0.66	0.73	60.53 ± 6.95	0.62	0.59	0.67	59.09 ± 7.26	0.58	0.60	0.68
	MFCC	69.31 ± 6.36	0.69	0.69	0.77	60.53 ± 6.95	0.61	0.60	0.65	62.50 ± 7.15	0.62	0.63	0.68
	PLP	68.81 ± 6.39	0.67	0.70	0.76	58.42 ± 7.01	0.60	0.57	0.65	67.05 ± 6.94	0.70	0.64	0.73
Comp	LR	56.18 ± 5.04	0.51	0.62	0.60	61.05 ± 6.93	0.59	0.64	0.65	61.36 ± 7.19	0.64	0.59	0.74
	Dyn	65.59 ± 4.83	0.63	0.68	0.75	64.74 ± 6.79	0.63	0.67	0.70	58.52 ± 7.28	0.52	0.64	0.70
	Reg	69.89 ± 4.66	0.68	0.72	0.78	61.58 ± 6.92	0.48	0.75	0.73	63.07 ± 7.13	0.60	0.66	0.68
	Ent	75.00 ± 4.40	0.75	0.75	0.83	60.53 ± 6.95	0.61	0.60	0.68	63.07 ± 7.13	0.67	0.59	0.73

Results

SVD database:

		Running speech - raw speech			
Set	Length	ACC	SP	SE	AUC
CPPS+LHr	25 ms	62.26 ± 2.46	0.60	0.64	0.67
MFCC (18)	20 ms	80.32 ± 2.02	0.74	0.84	0.86
PLP (18)	20 ms	77.90 ± 2.11	0.65	0.85	0.85

	Running speech - extracted vowels			
Set	ACC	SP	SE	AUC
Pert	66.69 ± 2.39	0.67	0.67	0.74
CPPS+LHr	56.15 ± 2.52	0.52	0.58	0.58
MFCC (16)	76.96 ± 2.14	0.70	0.81	0.84
PLP (18)	75.55 ± 2.18	0.67	0.80	0.82

		Vowel /a/				Vowel /i/				Vowel /u/			
Set	Subset	ACC	SP	SE	AUC	ACC	SP	SE	AUC	ACC	SP	SE	AUC
Pert	-	68.79 ± 2.32	0.68	0.69	0.78	64.37 ± 2.39	0.64	0.65	0.72	59.88 ± 2.45	0.59	0.60	0.66
MSs	-	66.67 ± 2.35	0.60	0.70	0.73	63.20 ± 2.41	0.64	0.63	0.70	61.05 ± 2.44	0.53	0.66	0.67
SCs	CPPS+LHr	61.12 ± 2.44	0.61	0.61	0.68	57.87 ± 2.47	0.55	0.59	0.62	58.13 ± 2.47	0.55	0.60	0.63
	MFCC	70.48 ± 2.28	0.67	0.73	0.77	68.73 ± 2.32	0.67	0.69	0.76	65.41 ± 2.38	0.64	0.66	0.71
	PLP	71.07 ± 2.27	0.70	0.72	0.77	68.21 ± 2.33	0.67	0.69	0.75	64.69 ± 2.39	0.62	0.66	0.71
Comp	LR	61.31 ± 2.43	0.59	0.63	0.68	59.04 ± 2.46	0.57	0.60	0.61	58.06 ± 2.47	0.57	0.59	0.64
	Dyn	63.78 ± 2.40	0.62	0.65	0.73	62.29 ± 2.42	0.62	0.63	0.68	61.51 ± 2.43	0.58	0.64	0.68
	Reg	67.75 ± 2.34	0.55	0.75	0.74	63.46 ± 2.41	0.62	0.65	0.69	63.78 ± 2.40	0.50	0.72	0.67
	Ent	68.08 ± 2.33	0.68	0.68	0.75	61.25 ± 2.43	0.60	0.62	0.65	59.95 ± 2.45	0.57	0.62	0.66

Experiment 2: design of a hierarchical system based on the sex of the speaker

- To analyse the impact of the sex of the speakers in detection tasks, a sex-independent and a sex-dependent system are designed.

HUPA database:

	Subtype	ACC	SP	SE	AUC
	Fe.+Ma.	73.66 ± 4.47	0.72	0.70	0.81
S.D.	*Fe.*: MFCC(20)	73.45 ± 5.76	0.71	0.75	0.81
	Ma.: MFCC(20)	73.97 ± 7.12	0.69	0.80	0.85
S.I.	MFCC(12)	69.62 ± 4.67	0.71	0.77	0.79

Gmar database:

Type	Subtype	ACC	SP	SE	AUC
	Fe. + Ma.	70.74 ± 2.27	0.68	0.73	0.78
S.D.	*Fe.*: MFCC(18)	72.17 ± 2.91	0.70	0.73	0.79
	Ma.: MFCC(16)	68.68 ± 3.62	0.63	0.71	0.77
S.I.	MFCC(18)	70.48 ± 2.28	0.67	0.73	0.77

SVD database:

Type	Subtype	ACC	SP	SE	AUC
	Fe.+Ma.	70.79 ± 6.27	0.72	0.70	0.78
S.D.	*Fe.*: MFCC(16)	70.45 ± 7.78	0.69	0.72	0.77
	Ma.: MFCC(10)	71.43 ± 10.58	0.76	0.66	0.82
S.I.	MFCC(20)	69.31 ± 6.36	0.69	0.69	0.77

Experiment 3: testing out the performance of classification techniques used in speaker recognition

- Ancillary datasets are employed for training the UBM and compensation models:

	Configuration	Datasets	Speech tasks	Content
Sustained phonation	C_1	HUPA	/a/	No.
	C_2	HUPA, MEEI, GMar	/a/	No. + Dy.
	C_3	HUPA, MEEI, GMar, EUROM	/a/+/i/+/u/+RSv.	No.
Running speech	C_1	HUPA, GMar, EUROM, Albayzin	/a/+/i/+/u/+RSv.	No. + Dy.
	C_2	HUPA, GMar, EUROM, Albayzin	/a/+/i/+/u/+RSv.	No.
	C_3	MEEI, EUROM, Albayzin	RSr.	No.
	C_4	EUROM	RSr.	No.
	C_5	EUROM, PhoneDat-I	RSr.	No.

- GMM, GMM-UBM, IV, PLDA and GMM-SVM classifiers are employed.
- The number of Gaussian components is varied in such a manner that $\{2^i\}$: $i \in Z;\ 1 \leq i \leq 9$.

Results

- Running speech (right)
- Sustained phonations

Configuration	Classifier	ACC	SP	SE	AUC
–	GMM	70.48 ± 2.28	0.67	0.73	0.77
C_1	GMM-UBM	69.27 ± 2.37	0.68	0.70	0.76
	IV	67.15 ± 2.41	0.66	0.68	0.75
	PLDA	71.66 ± 2.31	0.70	0.73	0.79
	GMM-SVM	71.18 ± 2.32	0.69	0.73	0.77
C_2	GMM-UBM	69.20 ± 2.37	0.69	0.69	0.76
	IV	68.38 ± 2.38	0.66	0.70	0.75
	PLDA	72.76 ± 2.28	0.72	0.73	0.79
	GMM-SVM	72.01 ± 2.30	0.69	0.74	0.77
C_3	GMM-UBM	68.24 ± 2.39	0.67	0.69	0.75
	IV	70.77 ± 2.33	0.71	0.71	0.78
	PLDA	71.32 ± 2.32	0.69	0.73	0.80
	GMM-SVM	71.73 ± 2.31	0.70	0.73	0.78

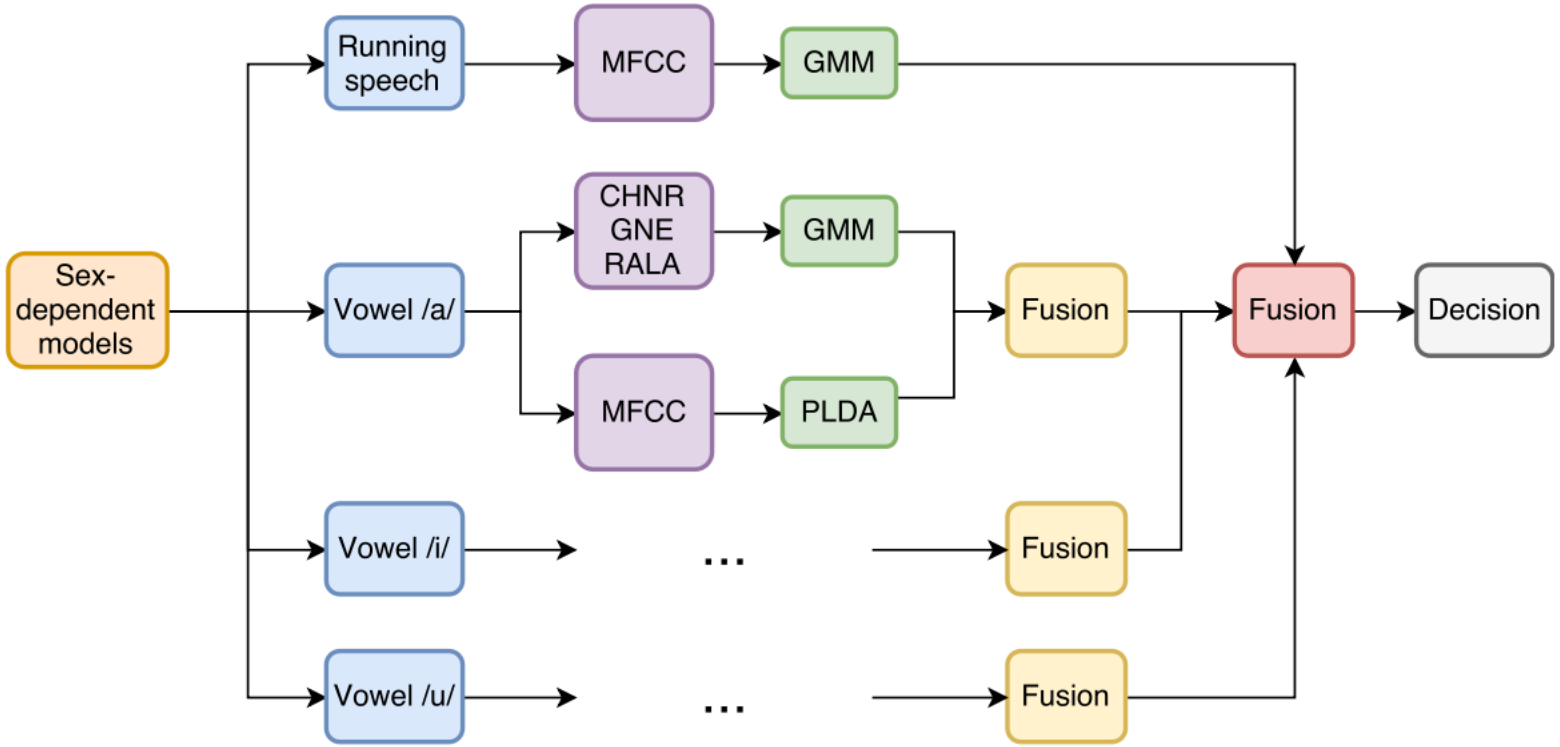
Configuration	Classifier	ACC	SP	SE	AUC
–	GMM	80.32 ± 2.02	0.74	0.84	0.86
C_1	GMM-UBM	76.70 ± 2.15	0.75	0.78	0.84
	IV	71.79 ± 2.29	0.70	0.73	0.79
	PLDA	74.01 ± 2.23	0.74	0.74	0.82
	GMM-SVM	77.03 ± 2.14	0.75	0.78	0.85
C_2	GMM-UBM	78.44 ± 2.09	0.72	0.82	0.86
	IV	73.00 ± 2.25	0.71	0.74	0.81
	PLDA	75.62 ± 2.18	0.75	0.76	0.85
	GMM-SVM	78.17 ± 2.10	0.76	0.80	0.85
C_3	GMM-UBM	78.44 ± 2.09	0.72	0.82	0.86
	IV	73.00 ± 2.25	0.71	0.74	0.81
	PLDA	75.62 ± 2.18	0.75	0.76	0.85
	GMM-SVM	78.71 ± 2.08	0.76	0.80	0.86
C_4	GMM-UBM	76.96 ± 2.14	0.74	0.79	0.84
	IV	72.73 ± 2.26	0.72	0.73	0.81
	PLDA	75.62 ± 2.18	0.74	0.76	0.84
	GMM-SVM	77.77 ± 2.11	0.71	0.81	0.86
C_5	GMM-UBM	78.64 ± 2.08	0.73	0.82	0.85
	IV	70.65 ± 2.31	0.69	0.71	0.79
	PLDA	76.09 ± 2.17	0.74	0.77	0.83
	GMM-SVM	79.25 ± 2.06	0.76	0.81	0.86

Experiment 4: combination of the best systems

- This experiment is built around the lessons learnt during the first three experiments
- Two trials are considered in the current experiment:
  - One which provides a single decision about the condition of patients by combining the results of the systems based on the vowels /a/, /i/, /u/ and running speech
  - The other designed to test the capabilities of the system in a cross-dataset scenario.
- Three dimensionality reduction methods are employed to rank the features from the most to the least relevant
  - Maximal Information Maximisation (MIM), Minimal Redundancy Maximal Relevance (mRMR) and Joint Mutual Information (JMI).
- Fusion of results: logistic regression is employed to fuse the system using the most consistent features and a GMM classifier, and the system based on MFCC and classification based on PLDA.

Results

- As per experiment 1, CHNR, GNE and RALA are the best features for sustained phonations
- As per experiment 2, sex-separated models perform proficiently



Database	Vowel	ACC	SP	SE	AUC
GMar	/a/	72.00 ± 7.19	0.60	0.83	0.82
	/i/	62.67 ± 7.74	0.47	0.77	0.75
	/u/	72.67 ± 7.13	0.71	0.74	0.79
	All	74.00 ± 7.02	0.64	0.83	0.82
HUPA	/a/	74.46 ± 4.43	0.65	0.85	0.87
ATIC	/a/	78.21 ± 9.16	0.90	0.74	0.93
DN	/a/	82.87 ± 5.49	0.76	0.89	0.94

References

[1] Gómez-García, J. A., Laureano Moro-Velázquez, and Juan Ignacio Godino-Llorente. "On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors." *Biomedical Signal Processing and Control* 48 (2019): 128-143.

[2] Moro-Velázquez, Laureano, et al. "Modulation spectra morphological parameters: a new method to assess voice pathologies according to the GRBAS scale." *BioMed research international* 2015 (2015).

[3] Dehak, Najim, and Stephen Shum. "Low-dimensional speech representation based on factor analysis and its applications." *Johns Hopkins CLSP Lecture* (2011).

Questions

1. In addition to the sex of the speaker, what other extralinguistic traits can speakers be separated into for potentially better classification performance?
2. For what reasons would one employ a GMM-UBM classifier instead of an ordinary GMM classifier? How does a GMM-UBM setup work?
3. How does feature extraction differ for the running speech as compared to the sustained phonations?