Dysphonic Voice Detection for Speakers' Biometry

Center for Biomedical Technology
Neuromorphic Speech Processing Laboratory (NEUVOX)

research group GIAPSY ©UPM

Pedro Gómez[1], Luis M. Mazaira[1], Agustín Álvarez[1], and Eugenia San Segundo[2]
[1]Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain
[2]Universidad Internacional Menéndez Pelayo (UIMP), Madrid, Spain
E-mail: pedro@fi.upm.es

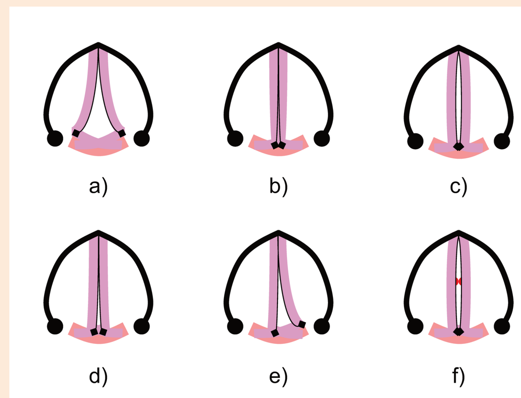
Introduction

- Phonation Distortion leaves relevant marks in a speaker's biometric profile.
- Dysphonic Voice production may be used in biometrical speaker characterization.
- In the present paper phonation features derived from the glottal source (GS) parameterization after the vocal tract inversion is proposed for dysphonic voice characterization in Speaker Verification tasks (Gómez, 2012).

Vocal Fold Images during closed phase from strobostcopic endoscopy. Left: bilateral nodules. Center: Reinke's Edema. Rigth: Polyp in right fold.

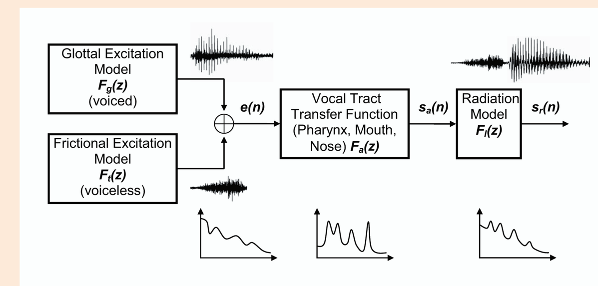
Distortion in Dysphonic Voice is mainly due to :
- Vocal Fold vibration Asymmetry
- Deficient Glottal Closure

Main causes:
- Organic pathology
- Phonation Dysfunction
- Neurological Diseases
- Emotional Perturbations

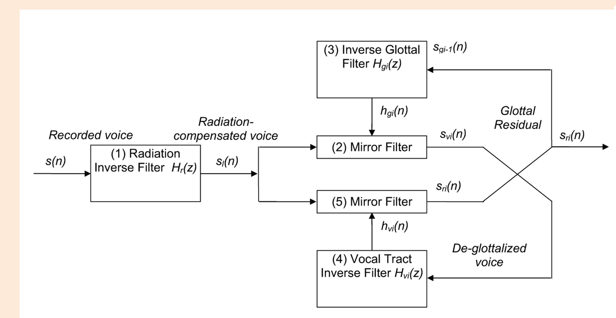


Vocal Fold symplified situations:
a) Open during breathing
b) Phonation closed phase
c) Phonation open phase
d) Deficient closure in posterior third
e) Asymmetric contact defect
f) Deficient closure in the medial third (lession)

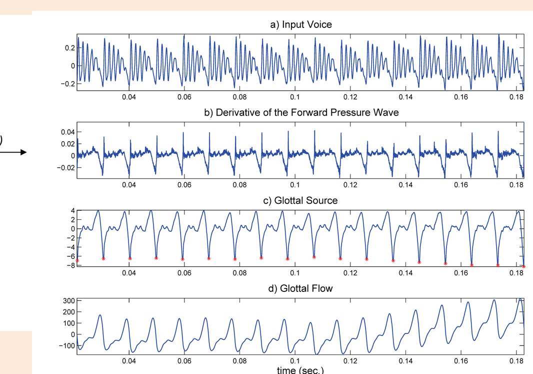
Phonation is related to Vocal Fold vibration: voiced speech segments (long vowels and fillers)
Voiced speech is inverse filtered to obtain the glottal residual, the glottal source is estimated from the residual
Different sets of parameters are estiimated from the glottal source: distortion, cepstral, spectral, biomechanical, temporal, glottal closure and tremor. For a review of algorithmic methods see (Gómez, 2009).



Fant's Speech Production Model



Algorithmic Inversion of Phonation



Perturbation Parameters
1. Absolute Pitch
2. Abs. Norm. Jitter
3. Abs. Norm. Ar. Shimmer
4. Abs. Norm. Min. Sharp.
5. Noise-Harm. Ratio (NHR)
6. Muc./AvAc. Energy (MAE)

Cepstral Parameters
7. MWC Cepstral 1
8. MWC Cepstral 2
9. MWC Cepstral 3
10. MWC Cepstral 4
11. MWC Cepstral 5
12. MWC Cepstral 6
13. MWC Cepstral 7
14. MWC Cepstral 8
15. MWC Cepstral 9
16. MWC Cepstral 10
17. MWC Cepstral 11
18. MWC Cepstral 12
19. MWC Cepstral 13
20. MWC Cepstral 14

Spectral Parameters
21. MW PSD 1st Max. ABS.
22. MW PSD 1st Min. rel.
23. MW PSD 2nd Max. rel.
24. MW PSD 2nd Min. rel.
25. MW PSD 3rd Max. rel.
26. MW PSD End Val. rel.
27. MW PSD 1st Max. Pos. ABS.
28. MW PSD 1st Min. Pos. rel.
29. MW PSD 2nd Max. Pos. rel.
30. MW PSD 2nd Min. Pos. rel.
31. MW PSD 3rd Max. Pos. rel.
32. MW PSD End Val. Pos. rel.
33. MW PSD 1st Min NSF
34. MW PSD 2nd Min NSF

Biomechanical Parameters
35. Body Mass
36. Body Losses
37. Body Stiffness
38. Body Mass Unbalance
39. Body Losses Unbalance
40. Body Stiffness Unbalance
41. Cover Mass
42. Cover Losses
43. Cover Stiffness
44. Cover Mass Unbalance
45. Cover Losses Unbalance
46. Cover Stiffness Unbalance

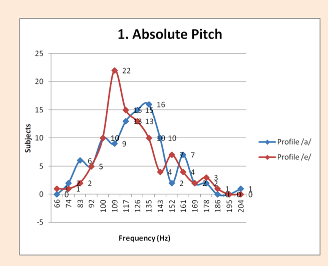
Temporal Parameters
47. Rel. Recov. 1 Time
48. Rel. Recov. 2 Time
49. Rel. Open 1 Time
50. Rel. Open 2 Time
51. Rel. Max. Ampl. Time
52. Rel. Recov. 1 Ampl.
53. Rel. Recov. 2 Ampl.
54. Rel. Open 1 Ampl.
55. Rel. Open 2 Ampl.
56. Rel. Stop Flow Time
57. Rel. Start Flow Time
58. Rel. Closing Time

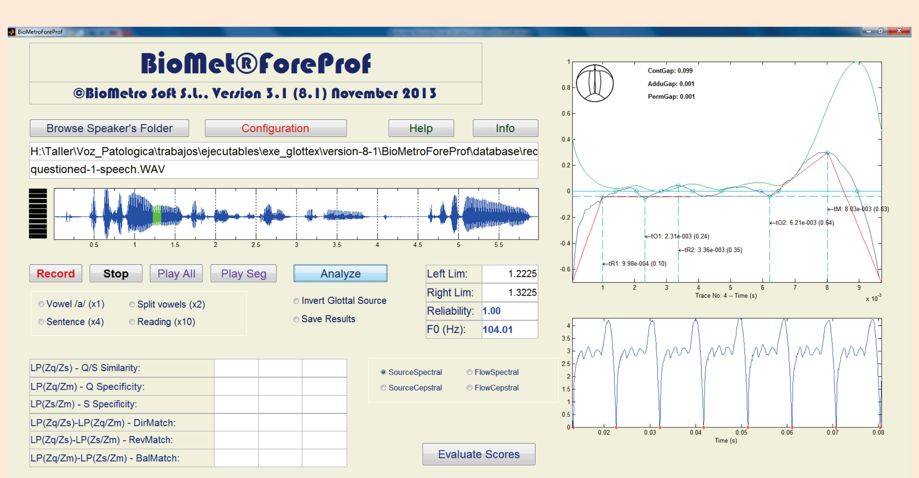
Glottal GAP Parameters
59. Val. Flow GAP
60. Val. Contact GAP
61. Val. Adduction GAP
62. Val. Permanent GAP

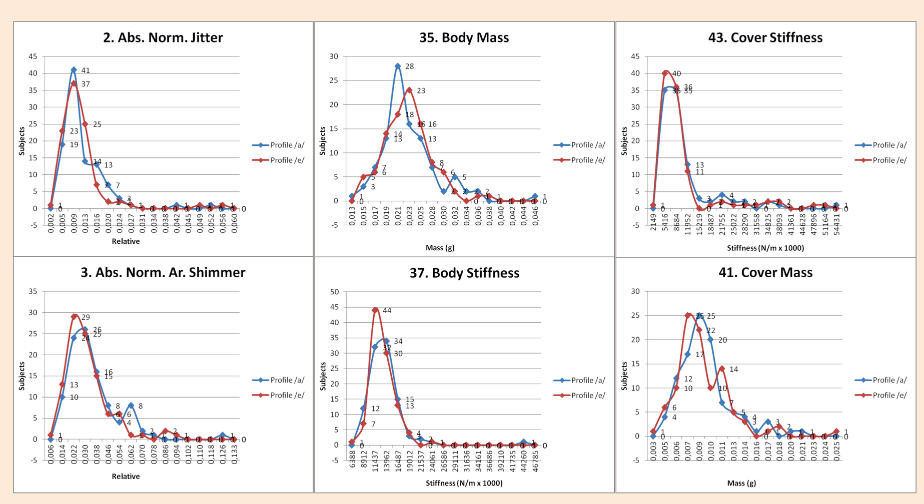
Tremor Parameters
63. 1st. Order Cyc. Coeff.
64. 2nd. Order Cyc. Coeff.
65. 3rd. Order Cyc. Coeff.
66. Tremor Frequency
67. Estimation Reliabilty
68. Tremor rMS Amplitude

Materials and Methods

- Fillers and log vowels (lasting more than 100 ms)
- Segmented from a telephonic (GSM) database of 100 male speakers
- Corresponding to phonations of vowels between /a/ and /e/
- 68 parameters were obtained from each 100 ms long segment
- A comparative paired test showed equivalence between /a/ and /e/ phonations







- Sample Matching Paradigm:
  Zq: Questioned sample set (test: blue)
  Γs: Gaussian Mixture Model from Suspect's sample set (control: red)
  Γm: Gaussian Mixture Model from Line-Up sample set (model: green)
- Prosecurot's vs Defender's Hyphothesis is evaluated as a Log-Likelihood Ratio:

$$\lambda_{pd} = \log\{\Pr(\mathbf{Z}_t|\Gamma_c)\} - \log\{\Pr(\mathbf{Z}_t|\Gamma_m)\}$$

Where conditional probabilities are given by:

$$\Pr(\mathbf{Z}_t \mid \Gamma_c) = \frac{1}{(2\pi)^{M/2}|\mathbf{C}_c|^M} e^{-1/2(\boldsymbol{\mu}_t-\boldsymbol{\mu}_c)^T \mathbf{C}_c^{-1}(\boldsymbol{\mu}_t-\boldsymbol{\mu}_c)}$$

$$\Pr(\mathbf{Z}_t \mid \Gamma_m) = \frac{1}{(2\pi)^{M/2}|\mathbf{C}_m|^M} e^{-1/2(\boldsymbol{\mu}_t-\boldsymbol{\mu}_m)^T \mathbf{C}_s^{-1}(\boldsymbol{\mu}_t-\boldsymbol{\mu}_m)}$$

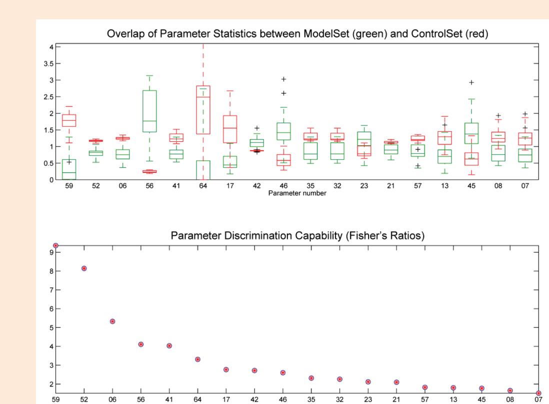
Details are to be found in (Gomez 2012)

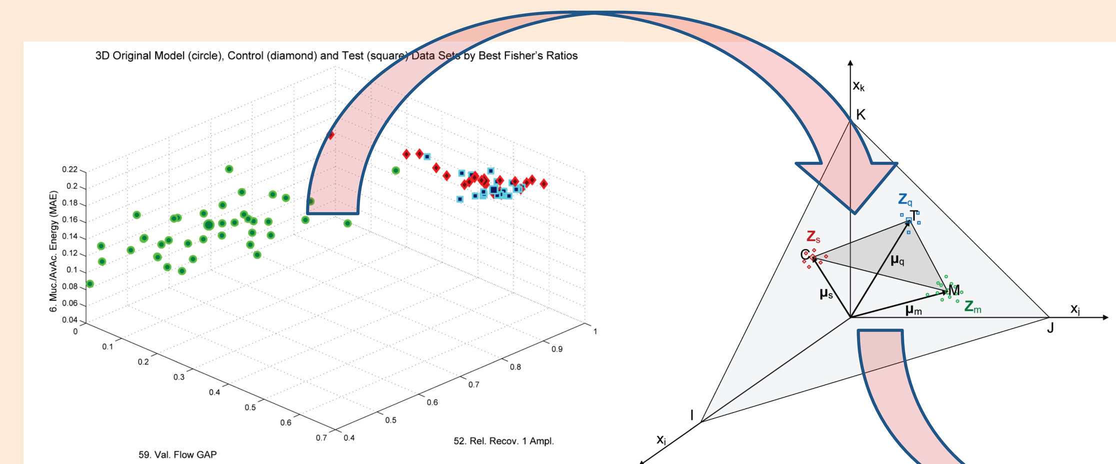
Sample Matching is based on comparing Phonation Descriptors from an unknown speaker (test set: blue) against a suspect speaker (control set: red) referred to a Line-Up speaker set (model set: green) to evaluate the Prosecutor's Hypothesis against the Defender's Hypothesis (Taroni, 2006):

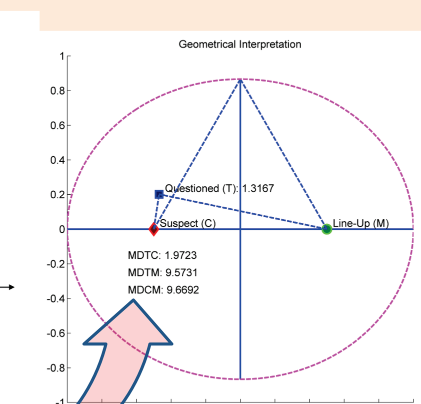
- Parameter selection from Fisher's Discriminant Ratios

$$C_{Fi} = \frac{\mu_{im} - \mu_{ic}}{\sqrt{\frac{s_{im}^2}{n_m} + \frac{s_{ic}^2}{n_c}}}$$

μ_{im}, μ_{ic}: model and control sample averages for parameter i
s_{im}, s_{ic}: model and control sample standard errors for parameter i
n_m, n_c: model and control set sample sizes





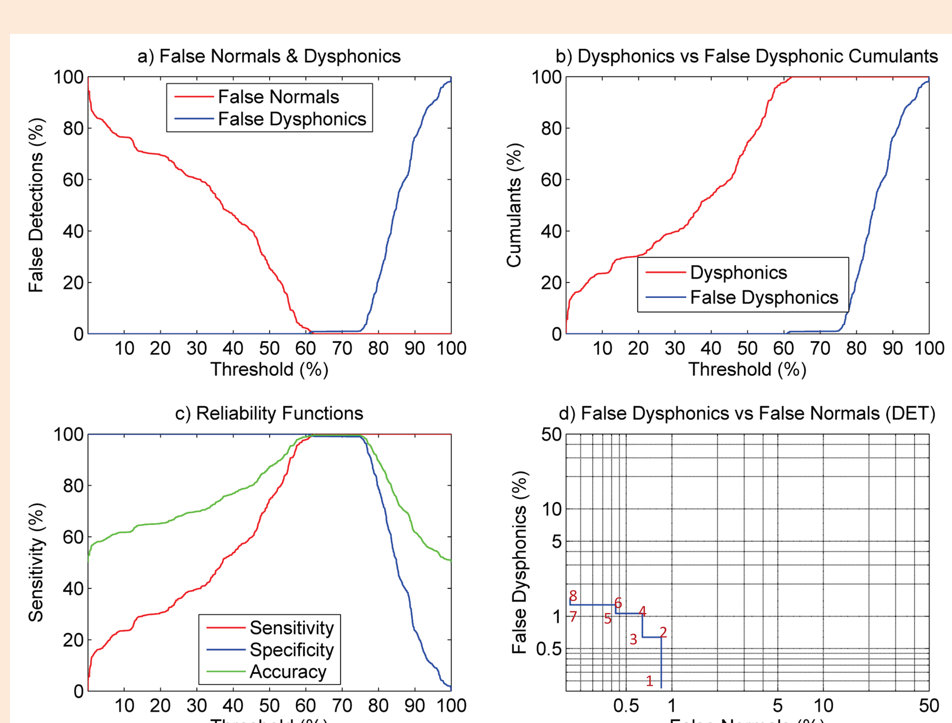


Validation & Sample Matching Results

First Experiment:
- Splitting the 100 male speakers in two equal-size subsets by their normophonic condition.
- Using a normative database validated by Hospital Gregorio Marañon of Madrid with samples of /a/ (50 male speakers).
- LLR's estimate the conditional probability of a given sample being normophonic or dysphonic (10-fold cross-validation).

Objectives:
- Estimate the discrimination accuracy of the methodology and the best parameters.
- Produce two reference subsets from GSM quality /e/ of use in Spanish.

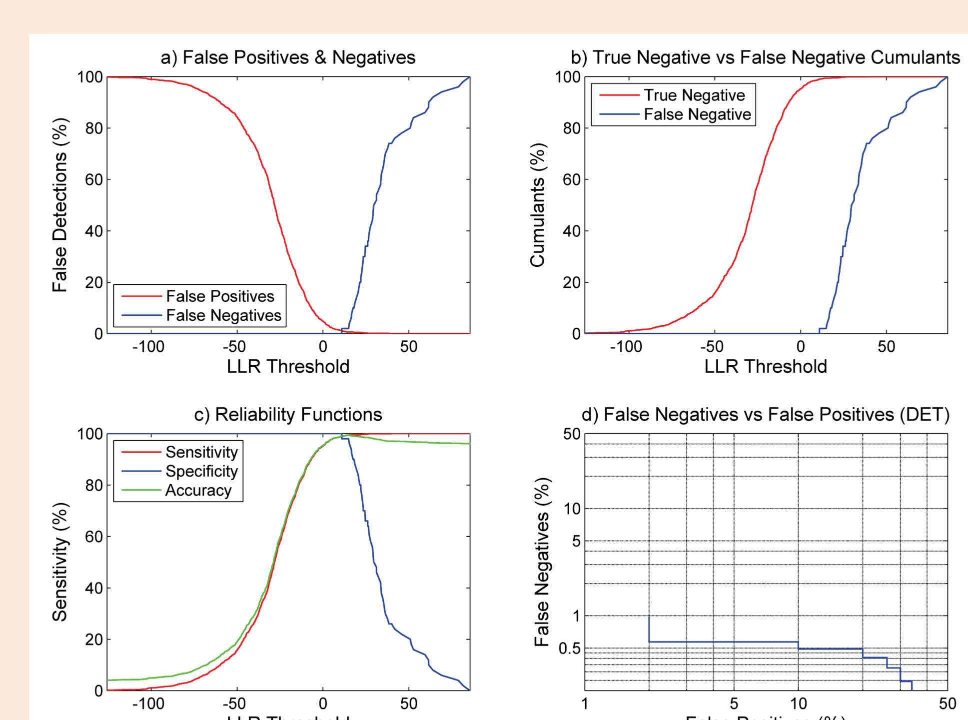
Results:
- Normophonic vs dysphonic cumulants, sensitivity, specificity and accuracy, and Detection-Error Trade-off plots.



Second Experiment:
- Matching each normative speaker's sample set against each other.
- Using as model the normative subset produced in the first experiment.
- Using as control the non-normative subset produced in the first experiment. 50 target vs 1225 non-target speakers.

Objective:
- Estimate the discrimination accuracy of the sample matching methodology in target vs non-target detection.

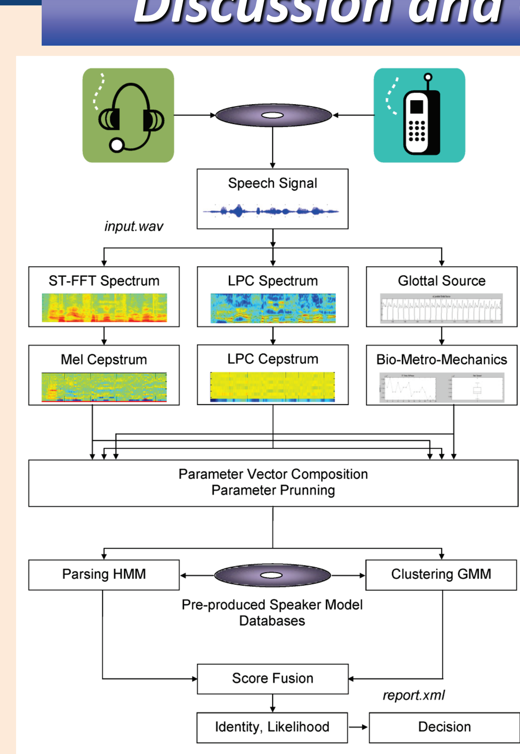
Results:
- False target vs false non-target detection cumulants, sensitivity, specificity and accuracy, and Detection-Error Trade-off plots.



Experiment	Samples	No. Tests	Accuracy (%)	LLR	EER	p-values
First	50 N + 50 D	90 Samples vs Model x 10 times cross-val. = 900	99.57	NA	0.638 (3)	0.00638, 0.00638
Second	50 N + 50 D	50 Samples vs each: 51*50/2 = 1275 (50 target + 1225 non-target)	99.29	10.88	NA	0.02, 0.0057

Discussion and Conclusions

- Sensitivity to nomophonic vs dysphonic phonation is large enough to allow forensic matchnig of voiced speech segments.
- Glottal source parameterizations from /a/ and /e/ are interchangeable, and can be used in cross-matching with no signiticant differences.
- Accuracy=(FP+FN)/(TP+TN) of target vs not-target matches grants applicability to real forensic cases.
- Margin of optimum LLR values grants strength of evidence to be over 4 in Lucy's Scale (Lucy, 2005).
- Distinction between normophonic and dyspohonic phonation seems to be feasible from parameterizations of glottal source.
- Questioned vs Suspect's Sample matching in reference to Line-Ups may be taken to meaningful 2D plots.
- Hybrid matching combining standard MFCC's and glottal source derived parameters may attain rather low equal error rates with telephone-quality speech (Khoury, 2013).



References

Gómez, P., et al. (2009). Glottal Source Biometrical Signature for Voice Pathology Detection. *Speech Comm.*, vol. 51, pp. 759-781.
Gómez, P., et al. (2012). Distance Metric in Forensic Voice Evidence Evaluation using Dysphonia-relevant Features. *Proc. of the VI Meeting of Biometric Recognition of Persons*, Ed. Universidad de Las Palmas de Gran Canaria, pp. 169-178.
Khoury, E., Mazaira, L.M., et al. (2013). The 2013 Speaker Recognition Evaluation in Mobile environments. *Proc. of the 6th IAPR International Conference on Biometrics*, Madrid, Spain.
Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists*. Wiley.
Taroni, F., Aitken, C., Garbolino, P., Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. Wiley.

Acknowledgments

This work is being funded by grant TEC2012-38630-C04-04 from Plan Nacional de I+D+i, Ministry of Economy and Competitivity of Spain.