Aalto University
School of Electrical
Engineering

ELEC-E5531 - Speech and Language Processing Seminar V D

"A Pattern Recognition Approach to Spasmodic Dysphonia and Muscle Tension Dysphonia Automatic Classification"

Mehedi Bijoy & Jack Bergkulla

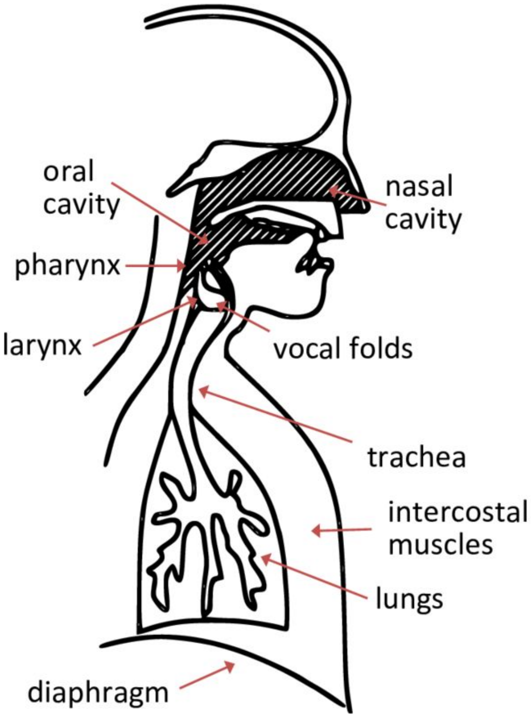
Human Voice Production

Overview of Human Voice Generation:

- Begins with air filling the lungs.
- Released air passes through the larynx, creating a sound wave.

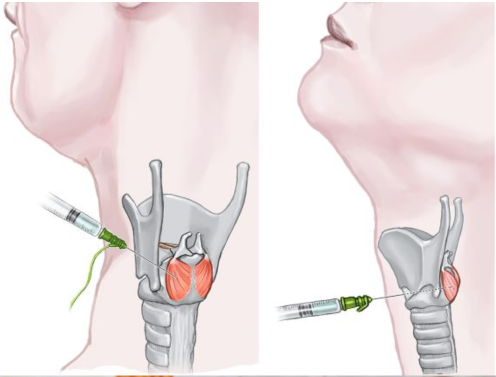
The Larynx - Voice Box:

- Larynx contains vocal cords or vocal folds.
- Also known as the voice box.



Spasmodic Dysphonia (SD) and Muscle Tension Dysphonia (MTD)

- SD and MTD are voice disorders with similar characteristics.
  - Differentiation requires experienced voice clinicians.
- Diagnosis challenging due to shared symptoms.
  - SD is a larynx focal dystonia → neurological disease, treated with surgery or botulinum toxin injections.
  - MTD is functional disorder, correctable with voice therapy.



Spasmodic Dysphonia (SD) and Muscle Tension Dysphonia (MTD)

- **Spasmodic Dysphonia (SD)**
  - Types include adductor SD (AdSD) and abductor SD (AbSD).
    - AdSD involves strong contraction causing strained voice, breaks.
    - AbSD is less common, spasms in muscles opening vocal folds.
  - Incidence is rare, affecting 30,000–50,000 in North America.
- **Muscle Tension Dysphonia (MTD)**
  - MTD involves excessive muscular tension during speech.
  - Vocal folds appear normal at rest but exhibit abnormal contraction during speech.

Importance of Automatic Classification of SD and MTD

- Precise treatment selection.
- Early intervention and improved prognosis.
- Resource optimization → cost-effective and sustainable healthcare.
- Customized patient care and patient empowerment.
- Advancements in research and understanding.
    - More comprehensive dataset
    - A deeper understanding of the distinctive acoustic features
    - Automation in voice pathology

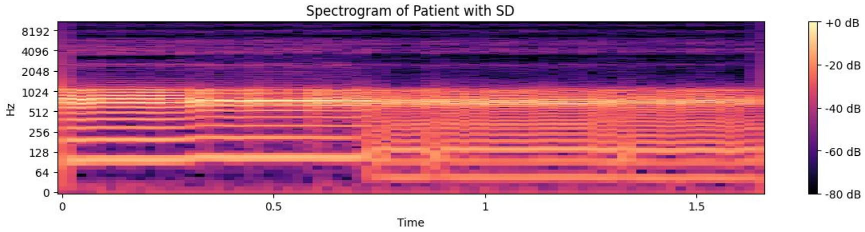
Limitations Identified by the Authors and Their Contributions in the Paper

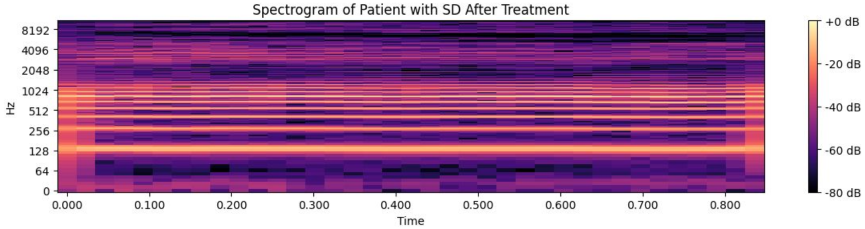
- <u>**Limitation Spotlights**</u>
  - Diagnostic challenges due to clinical expertise and absence of criteria.
    - Complex voice disorders including variability and overlap in symptoms.
  - Earlier efforts only focused on normal vs. pathological voices.
- <u>**Contributions**</u>
  - Automatic differentiation of AdSD, MTD, and normal voice.
  - Comparision between Neural Network and SVM.
  - Analysis of acoustic parameters from sustained vowels samples.

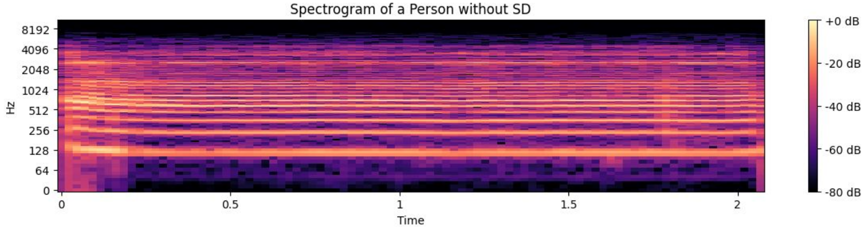
Data

- **Speaker Composition**
  - Dysphonic Speakers: 36 => 15 (MTD) + 21 (AdSD)
  - Normal Speakers: 53
- **Speech Signals**
  - Speech Type: Sustained Vowel ***/a/*** (for at least 3 seconds at a comfortable pitch and loudness)
    - Eight selected acoustic measures have been extracted.
- **Diagnostic Confirmation**
  - History, physical examination. MRI of the brain. Laryngeal EMG, Laboratory tests, Neurological evaluation, and Videostroboscopy.

Visualization







Feature Extraction

- **Degree of Voice Breaks (Unvoiceness)** → Reflects the presence of voice breaks.
- **Local F0 Ratio (Jitter)** → Indicates variations in the pitch of the voice.
- **Relative Average Perturbation (RAP)** → Provides information about pitch perturbations.
- **Five-Point Period Perturbation Quotient (PPQ5)** → Quantifies pitch perturbations.
- **Intensity (Shimmer)** → Represents variations in the amplitude of the speech signal.
- **Three-Point Amplitude Perturbation Quotient (APQ3)** → Assess amplitude perturbations in the speech signal.
- **Eleven-Point Amplitude Perturbation Quotient (APQ11)** → Offers a more detailed analysis of amplitude perturbations.
- **Harmonics-to-Noise Ratio (HNR)** → Reflects the ratio of harmonics to noise in the speech signal, with decreased values indicating certain voice disorders.

Methodology

- Neural Network
    - The feature dimensionality is reduced using PCA.
    - Utilized multilayer perceptron (MLP)
- Support Vector Machine
- Leave-One-Out Method (LOO)
    - Applied LOO method due to the small dataset.

Principal Component Analysis

1. Take 8 dimensional dataset
2. Compute mean of each dimension
3. Calculate covariance matrix:

$$\mathrm{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})$$

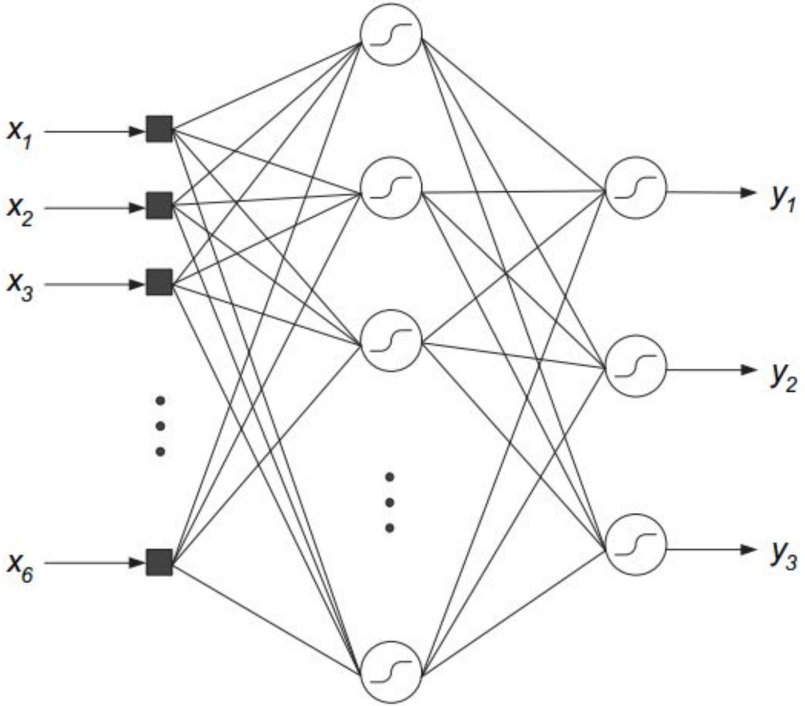
4. Find eigenvectors and eigenvalues:

$$\det(A - \lambda I) = 0$$

5. Sort the eigenvectors by decreasing eigenvalues.
6. Transform samples to new subspace.

Method - MLP

- Input dimension is reduced from 8 to 6 using PCA.
- Number of neurons in hidden layer varied from 8 to 34.
- Output: Adductor SD, MTD, and Normal.
- Activation function: TanH
- Epoch: 100
- Evaluation: Accuracy using Leave-One-Out method



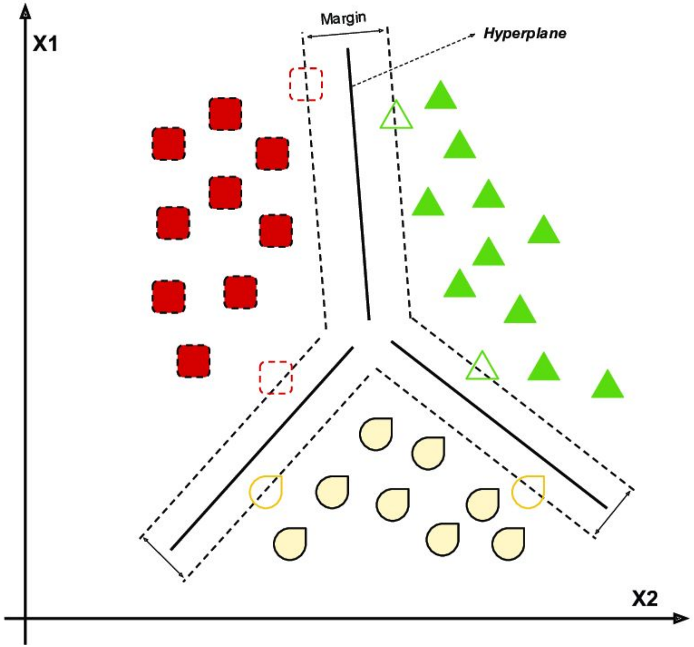
Method - SVM

- Finds a hyperplane that best separates the data.
- Mathematical formulation: $y(x) = \sum_{i=1}^{N} g_i K(x, x^{(i)}) + b$

  where, x is the input feature vector, K(x, x^i) is the kernel function and b is the bias.
- Kernel tricks:

  - Polynomial: $(\mathbf{x}^T\mathbf{y}+1)^p$

  - Gaussian Radial Basis Function: $\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2\right)$
- Unlike MLP, SVM does not depend on weight initialization.



Results - Neural Network

TABLE 1.
MLP NNs: Results of Tukey's Multiple Comparison Test

Group (Number of Hidden Units)	Mean Error ± Standard Deviation (%)	Groups With Means Not Significantly Different
32	11.01 ± 1.52	32 28 34 24 26 22 30 20 18 16 14
28	11.13 ± 1.57	32 28 34 24 26 22 30 20 18 16 14
34	11.21 ± 1.61	32 28 34 24 26 22 30 20 18 16 14
24	11.25 ± 1.62	32 28 34 24 26 22 30 20 18 16 14
26	11.33 ± 1.63	32 28 34 24 26 22 30 20 18 16 14
22	11.36 ± 1.56	32 28 34 24 26 22 30 20 18 16 14
30	11.38 ± 1.59	32 28 34 24 26 22 30 20 18 16 14
20	11.42 ± 1.67	32 28 34 24 26 22 30 20 18 16 14
18	11.52 ± 1.75	32 28 34 24 26 22 30 20 18 16 14
16	11.73 ± 1.92	32 28 34 24 26 22 30 20 18 16 14
14	11.80 ± 1.58	32 28 34 24 26 22 30 20 18 16 14
12	12.64 ± 1.98	12 10 8
10	12.83 ± 1.96	12 10 8
8	13.43 ± 2.37	12 10 8

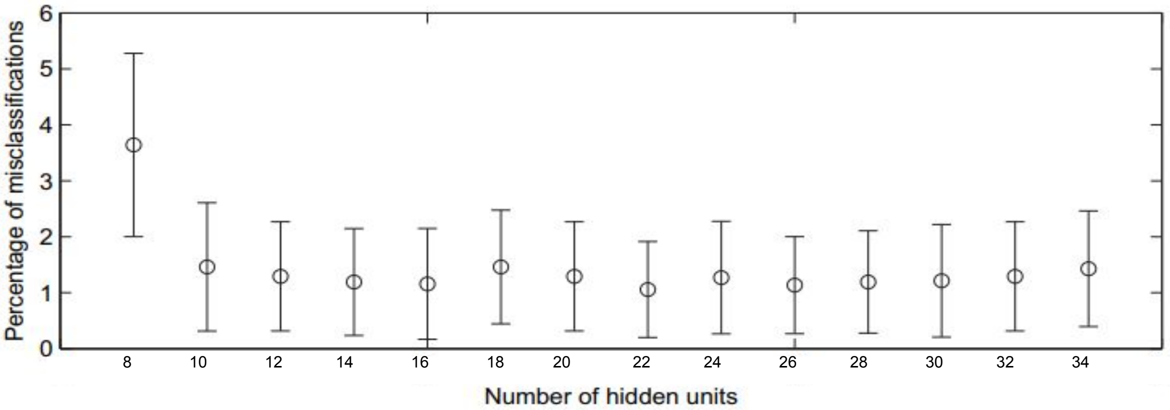


TABLE 6.
MLP NNs: Classification in Two Categories

Actual Class	Predicted Class: Pathological	Predicted Class: Normal	Correct Classifications (%)
Pathological	35.13	0.87	97.58
Normal	0.07	52.93	99.87
Total			98.94

Average of the 100 confusion matrices for 8 hidden units.

Results - Neural Network

TABLE 2.
MLP NNs: Best Confusion Matrix for 14 Hidden Units

Actual Class	Predicted Class AdSD	MTD	Normal	Correct Classifications (%)
AdSD	20	1	0	95.24
MTD	5	10	0	66.67
Normal	0	0	53	100.00
Total				93.26

TABLE 3.
MLP NNs: Best Confusion Matrix for 16 Hidden Units

Actual Class	Predicted Class AdSD	MTD	Normal	Correct Classifications (%)
AdSD	17	4	0	80.95
MTD	2	13	0	86.67
Normal	0	0	53	100.00
Total				93.26

TABLE 4.
MLP NNs: Best Confusion Matrix for 22 Hidden Units

Actual Class	Predicted Class AdSD	MTD	Normal	Correct Classifications (%)
AdSD	18	3	0	85.71
MTD	3	12	0	80.00
Normal	0	0	53	100.00
Total				93.26

TABLE 5.
MLP NNs: Average of the 100 Confusion Matrices for 32 Hidden Units

Actual Class	Predicted Class AdSD	MTD	Normal	Correct Classifications (%)
AdSD	15.78	4.48	0.74	75.14
MTD	4.45	10.44	0.11	69.60
Normal	0.01	0.01	52.98	99.96
Total				88.99

Results - SVM

TABLE 9.
Confusion Matrix for SVMs. Classification in Two Classes with Polynomial Kernel ($p = 2$)

Actual Class	Predicted Class: Pathological	Predicted Class: Normal	Correct Classifications (%)
Pathological	35	1	97.22
Normal	1	52	98.11
Total			97.75

TABLE 10.
Confusion Matrix for SVMs. Classification in Two Classes with RBF Kernel, $\sigma = 0.5$

Actual Class	Predicted Class: Pathological	Predicted Class: Normal	Correct Classifications (%)
Pathological	34	2	94.44
Normal	3	50	94.34
Total			94.38

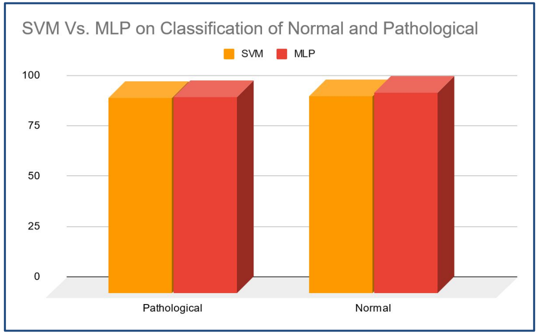
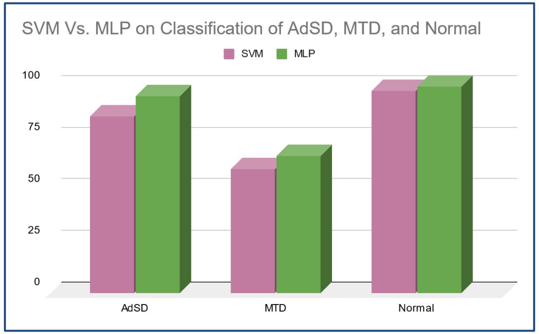


TABLE 7.
Confusion Matrix for SVMs. Classification in Three Classes with Polynomial Kernel ($p = 2$)

Actual Class	Predicted Class: AdSD	Predicted Class: MTD	Predicted Class: Normal	Correct Classifications (%)
AdSD	18	3	0	85.71
MTD	5	9	1	60.00
Normal	0	1	52	98.11
Total				88.76

TABLE 8.
Confusion Matrix for SVMs. Classification in Three Classes with RBF Kernel, $\sigma = 0.5$

Actual Class	Predicted Class: AdSD	Predicted Class: MTD	Predicted Class: Normal	Correct Classifications (%)
AdSD	16	2	3	76.19
MTD	4	9	2	60.00
Normal	0	0	53	100.00
Total				87.64



Comparison with Concurrent Papers

Paper	Dataset Split: P/N	Features	Classification method	Result
[1]	Total: 124 Split: 50/50%	Mean energy of speech	Random Forest	93.5 % accuracy with a population of 100 trees
[2]	Total: 2000+ Split: 33/66 %	Frequency, intensity, Harmonic to Noise Ratio	K-Nearest Neighbor, SVM, Decision Tree (DT)	DT algorithm yielded best classification accuracy of roughly 86.66 %.
[3]	Total: 120 Split: 50/50 %	Number of features between 42-60.	Naïve-Bayes (NB), MLP, SVM, Random Forest	100% accuracy using NB. NB > RF > MLP > SVM Vowel a and e best results
This Paper	Total: 89 Split: 60/40 %	Unvoiceness, Jitter, shimmer, RAP, PPQ5, APQ3, APQ11, HNR	MLP, SVM	93.26$ accuracy using MLP MLP > SVM

Summary:

- A larger dataset does not guarantee better performance. (ref. [2])
- Number of extracted features shows a linear relationship with model performance. (ref. [3])

[1] A Novel Algorithm for Detecting Spasmodic Dysphonia Voice Pathology using Random Forest Frame Work
[2] Spasmodic Dysphonia Detection Using Machine Learning Classifiers
[3] Vocal Test Analysis for the Assessment of Adductor-type Spasmodic Dysphonia

Findings and Contributions

- Automatic classification of SD and MTD using MLP and SVM based on acoustic features extracted from sustained vowel ***/a/*** samples.
- In the case of MLP, they experimented with various sizes of hidden units.
    - Stabilized error rate observed after 10 hidden units.
    - Best results:
        - 14 hidden units for AdSD and 16 hidden units for MTD.
- In the case of SVM, polynomial kernel outperformed Gaussian radial basis function kernel for both two-class (normal and pathological) and three-class classification (MTD, AdSD, and normal).

Limitations of the Paper

- **Small Dataset** → may restrict the generalizability of the findings.
- **Single Speech Sample** → may not fully capture the characteristics of disease.
- **Limited Acoustic Features** → only from sustained vowel /a./ and avoiding others.
- **Insufficient External Validation** → limits the robustness.
- **Lack of Interpretability** → limits the applicability.
- **Absence of Comprehensive Evaluation** → precision, recall, and F1 score.
- **Issues Associated with the Proposed MLP:**
  - Missing information to reproduce the model: Loss, optimizer.
  - Vanishing gradient: TanH
- **Lack of Comprehensive Analysis with Other Concurrent Methods.**

Future Research Direction

- Creation of a more diverse dataset.
- Exploration of alternative acoustic feature selection strategies.
- More advanced machine learning and deep learning models.
- Integration of multimodal method: Incorporate facial expression along with voice.
- Domain Adaptation for similar diseases.
    - Transfer learning.
    - Meta learning.
- Development of explainable automatic voice disorder classifier.

Aalto University
School of Electrical
Engineering

Thank You!

Any Questions?

Mehedi Bijoy & Jack Bergkulla

Assignment

- What are the major types of spasmodic dysphonia? Briefly explain their characteristics with appropriate examples.
- Record yourself pronouncing the sustained vowel /a/ for 2 seconds. Then, create a spectrogram of your recording. In your answer, include the spectrogram you created with “Spectrogram of the Audios.png”, and compare your spectrogram to the spectrograms of a healthy person (Audio #1) and a person with Spasmodic Dysphonia (Audio #2).
    - Audio Files and their Spectrograms: **link**