Authors: Amlu Anna Joshy, Rajeev Ranjan (2022)

Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques

Presentation by: Priyanshi Pal, Anusha Porwal

Aalto University
School of Electrical
Engineering

Table of contents

01 Introduction

02 Feature and Experiment Design

03 Databases Description

04 Classifier Design

05 Results

06 Discussion & Conclusion

01 INTRODUCTION

Manifestations:

- Imprecise articulation (Weak facial reflexes)
- low audibility
- atypical prosody
- variable speech rate
- Hyper nasality
- Harsh voice quality
- Increased fatigue

Dysarthria

How is severity typically assessed?

- Objective (acoustic & physiological measures)
- Subjective (perceptual: SLP)

Motivation for automatic severity classification:

- Expert evaluation varies and is expensive
- Keep track of client during rehabilitation
- Improving ASR for dysarthric patients

02 Feature Selection

Which features are selected for this study and why?

Basic Speech Features
MFCC, CQCCs

Speech Disorder Specific Features
Prosody, Glottal , Phonetic and Articulation Based

I-vector subspace modelling
iMFCC, iCQCC

Basic Speech Features

Features which mimic human auditory system. (Automate perceptual assessment)

- Mel-frequency cepstral coefficients (MFCC)
  - modelling pathological speech
  - Better than log filterbanks, comparable to i-vectors
  - irregular vocal folds movements

- Constant Q Cepstral coefficients (CQCC)
  - Excellent for speaker recognition in the recent years
  - Result of coupling between CQT and traditional cepstral analysis [more closely related to human perception]

Focus: Investigate Performance of various Deep learning models to see if improvements can be made over machine learning classifies

Speech Disorder Specific Features

- Prosodic features (103): duration, fundamental frequency, pitch and energy contours
- Phonetic features (28): Variation in phonation quality
- Glottal features (36): glottal inverse filtering/adaptive inverse filtering
- Articulatory features (488): retardation in Lip, tongue, jaw

NOTE: These features are extracted per utterance

Motivated by past works, authors suggest speech disorder specific features in terms of Prosodic, Glottal, Phonetic and Articulatory features are relevant in Identifying dysarthric speech patterns.

Identity vector (i-vectors) subspace modelling

- i-vector subspace modelling captures aspect of person's speech, gender, age and intelligibility. (Good for speaker, language and accent recognition)

- maps the high dimensional GMM supervector space to a single total-variability space

- i-vectors using MFCC (iMFCC) and CQCC (iMFCC) are extracted
  - Frame wise 13-dimensional MFCCs and CQCCs, and their first two deltas.
  - The target GMM supervector [dysarthric] (M) is formulated by: **$M = m + T.\omega$**
    m represents the UBM supervector, T is a low dimensional rectangular TV matrix, and w is the resulting i-vector.

- Performance of the different classifiers is analysed by varying the number of mixtures or Gaussian components (Ng) used in building the UBM, and dimension (Niv) of the T matrix used for i-vector extraction.

Feature & Experiment Design

(E1) Analysing MFCCs & CQCCs
13 MFCCs and their first 2 derivatives, for 30ms frames with hop of 10ms. #frames are fixed.

CQCCs extracted in the same way as MFCCs.

DNN, CNN, GRU and LSTMs are used.

(E2) Analysing Speech Disorder Specific Features
36 Glottal, 488 articulatory, 28 phonetic and 103 prosodic features are extracted (655 features).

Dimensionality Reduction is done using Factor Analysis (FA-200).

DNN is used.

(E3) Analysing i-Vectors
i-vectors hold information about the main variabilities describing the data - noise, age, severity dependent factors, intelligibility characteristics, etc.

i-vectors using MFCCs and CQCCs were extracted, their first two deltas.

DNN was trained on these features.

03

Datasets

UA-Speech Universal Access dysarthric Speech Corpus

- 13 healthy & 19 dysarthric speakers
  - Data of only 15 dysarthric speakers available
- 155 common words repeated thrice
  - 155 * 3 * 15 = 6975 total utterances
- 300 uncommon words per speaker
  - Test set
- Sev Levels: Reported by 5 naive listeners. Based on intelligibility.

TORGO

- 7 healthy & 8 dysarthric speakers
- Word utterances
- 80-20 Train Test Split
- Sev Levels: Reported by SLP
  - Based on clinical assessments

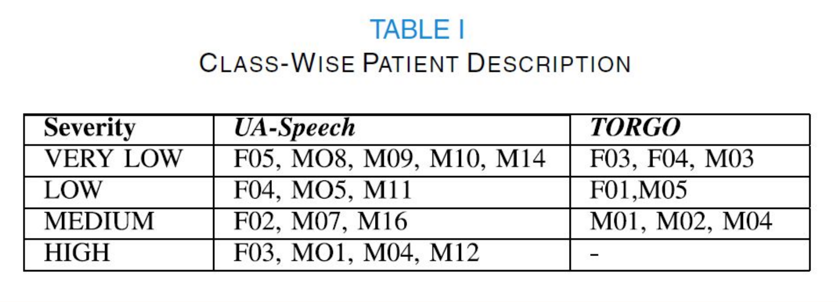


TABLE I
CLASS-WISE PATIENT DESCRIPTION

Severity	***UA-Speech***	***TORGO***
VERY LOW	F05, MO8, M09, M10, M14	F03, F04, M03
LOW	F04, MO5, M11	F01,M05
MEDIUM	F02, M07, M16	M01, M02, M04
HIGH	F03, MO1, M04, M12	-

Aalto University
School of Electrical
Engineering

04

Classifier Design

01. Baseline

- SVM: Linear Kernel with hyperparameter tuning
- RF: nTrees tuned on Validation Set (20% of data)

02. CNN

- n 2D Conv layers (2,2) kernel + ReLU + BatchNorm
- 2D Maxpool (2,2)size + Dropout(0.2)
- n is set using hyperparameter tuning
- Only MFCCs are used, deltas add redundancy.

03. DNN

- n dense layers + ReLU + Dropout (0.4)
- softmax(outputLayer)
- Training:
  - Batch size = 25
  - 120 Epochs
  - LearningRate = 0.001
  - Adam Optimizer
- Hyperparameter tuning for all params, including n

04. LSTM

- Known to capture long range dependencies
- Input, Forget & Output gates for info flow.
- 3 LSTM + 1 dropout + Dense output layer.
- # Hidden units in each layer & scaling factor α were tuned.

05. GRU

- Simpler version of LSTM – lesser data and computation power, trains faster.
- Architecture similar to the LSTM Model

05

Results

E1. Analysing MFCCs & CQCCs

- DNN & CNN tuned for n
  - As n increases, model has better generalization, but eventually overfits.
  - Similar trends for MFCC and CQCC models, but CQCC has much lower accuracy.
- LSTM & GRU tuned for α
  - There is a clear margin between the MFCC and CQCC accuracies, for LSTMs
  - GRU is the only model that gave comparable performance for both features and datasets.
- The results obtained on TORGO are almost always better than those obtained on UAS.

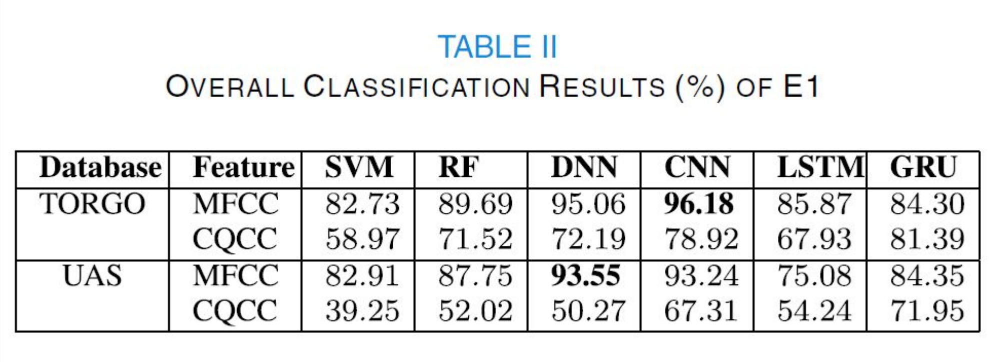


TABLE II

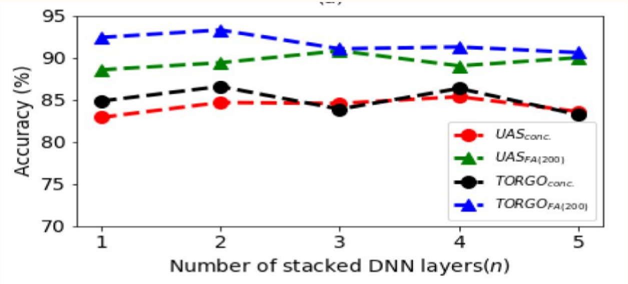
OVERALL CLASSIFICATION RESULTS (%) OF E1

Database	**Feature**	**SVM**	**RF**	**DNN**	**CNN**	**LSTM**	**GRU**
TORGO	MFCC	82.73	89.69	95.06	**96.18**	85.87	84.30
	CQCC	58.97	71.52	72.19	78.92	67.93	81.39
UAS	MFCC	82.91	87.75	**93.55**	93.24	75.08	84.35
	CQCC	39.25	52.02	50.27	67.31	54.24	71.95

05 Results

E2. Analysing Speech Disorder Specific Features

- Severity classification is done using each of the features.
  - DNN outperforms SVM for all features, but RF has some comparable results.
  - They created Confusion Matrices on the other features – they saw that misclassification happens between nearby classes – No signs of overfitting.
  - Concatenated Feature sets are used, with Factor Analysis. 200 factors gave best results for DNN.



(b)

Database	Features	SVM	RF	DNN
TORGO	Concatenated	82.51	82.24	≈ 86.00
	FA(200)	86.71	73.00	93.27
UAS	Concatenated	79.69	89.69	≈ 84.00
	FA(200)	85.35	82.06	90.80

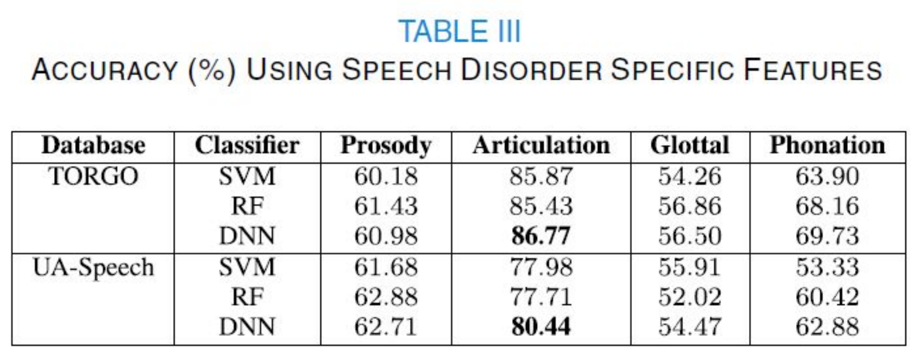


TABLE III

ACCURACY (%) USING SPEECH DISORDER SPECIFIC FEATURES

Database	Classifier	Prosody	Articulation	Glottal	Phonation
TORGO	SVM	60.18	85.87	54.26	63.90
	RF	61.43	85.43	56.86	68.16
	DNN	60.98	**86.77**	56.50	69.73
UA-Speech	SVM	61.68	77.98	55.91	53.33
	RF	62.88	77.71	52.02	60.42
	DNN	62.71	**80.44**	54.47	62.88

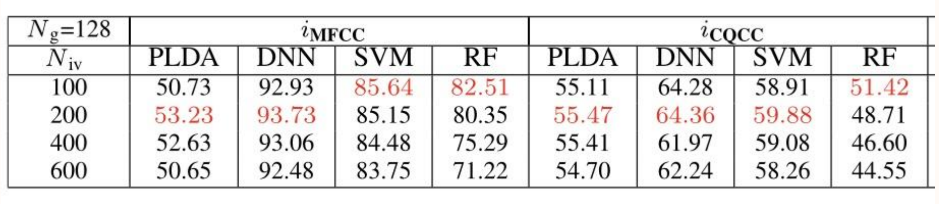
05

Results

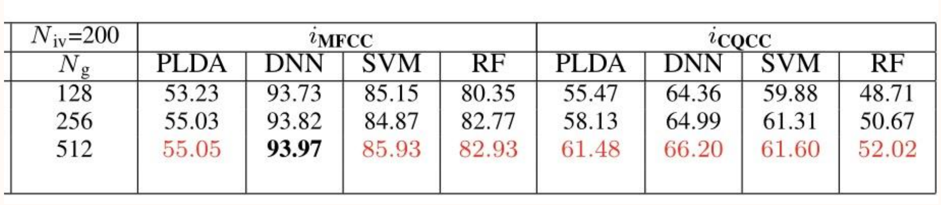
E3. Analysing i-Vectors

- Classification accuracy using I–vectors tuned on UAS
    - Fix Ng, vary Niv and vice versa
    - Using best parameters, I–vectors extracted for TORGO

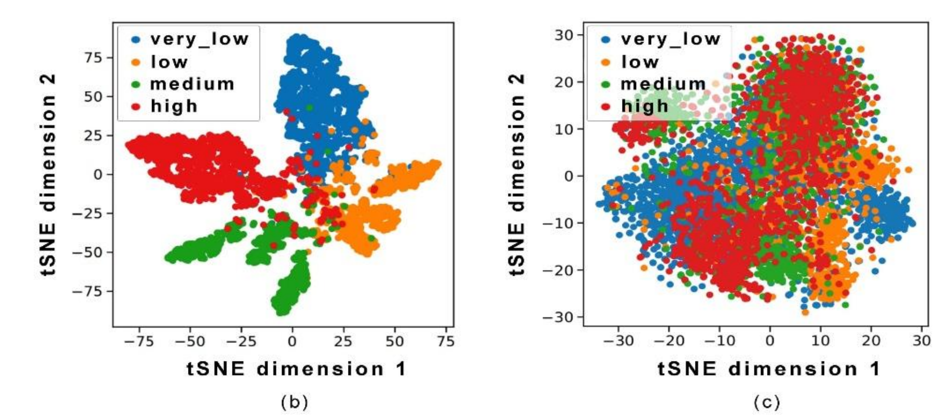
- TORGO accuracies: DNN best 95.29% followed by SVM and RF for iMFCCs. The same reported for iCQCCs in same chronological order (with slightly better acc – DNN 74.22%).



N_g=128	i_{MFCC}				i_{CQCC}			
N_{iv}	PLDA	DNN	SVM	RF	PLDA	DNN	SVM	RF
100	50.73	92.93	85.64	82.51	55.11	64.28	58.91	51.42
200	53.23	93.73	85.15	80.35	55.47	64.36	59.88	48.71
400	52.63	93.06	84.48	75.29	55.41	61.97	59.08	46.60
600	50.65	92.48	83.75	71.22	54.70	62.24	58.26	44.55



N_{iv}=200	i_{MFCC}				i_{CQCC}			
N_g	PLDA	DNN	SVM	RF	PLDA	DNN	SVM	RF
128	53.23	93.73	85.15	80.35	55.47	64.36	59.88	48.71
256	55.03	93.82	84.87	82.77	58.13	64.99	61.31	50.67
512	55.05	**93.97**	85.93	82.93	61.48	66.20	61.60	52.02



t-SNE plots from the last dense layer of DNN using (b) i_{MFCC} (c) i_{CQCC}.

05 Results

Evaluating Speaker-Dependency of the Models

- Different classifiers with their best tuned settings from previous experiments are picked.
  - LOSO cv on UAS Dataset
- 2 experiments, to evaluate Speaker Independency
  - 4 class classification of severity
  - Binary classification – low and high severity
- CQCCs outperform MFCCs
  - CQCCs can identify the same class speakers
  - Dysarthric characteristics specific to the speaker are found by MFCCs
- LOSO was also performed on E2 & E3 setups.
  - i-Vectors with MFCCs performed the best in SID: 49.22%
  - Best SID Sev Detection in literature: 53.90%

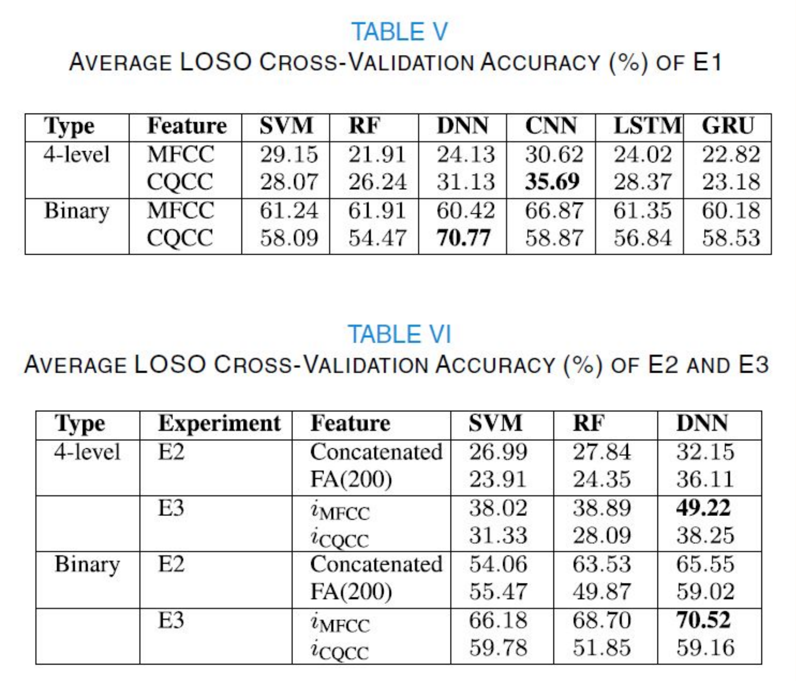


TABLE V
AVERAGE LOSO CROSS-VALIDATION ACCURACY (%) OF E1

Type	Feature	SVM	RF	DNN	CNN	LSTM	GRU
4-level	MFCC	29.15	21.91	24.13	30.62	24.02	22.82
	CQCC	28.07	26.24	31.13	**35.69**	28.37	23.18
Binary	MFCC	61.24	61.91	60.42	66.87	61.35	60.18
	CQCC	58.09	54.47	**70.77**	58.87	56.84	58.53

TABLE VI
AVERAGE LOSO CROSS-VALIDATION ACCURACY (%) OF E2 AND E3

Type	Experiment	Feature	SVM	RF	DNN
4-level	E2	Concatenated	26.99	27.84	32.15
		FA(200)	23.91	24.35	36.11
	E3	i_{MFCC}	38.02	38.89	**49.22**
		i_{CQCC}	31.33	28.09	38.25
Binary	E2	Concatenated	54.06	63.53	65.55
		FA(200)	55.47	49.87	59.02
	E3	i_{MFCC}	66.18	68.70	**70.52**
		i_{CQCC}	59.78	51.85	59.16

06

Discussion & Conclusion

- Comparison of ML models vs DL Models performances on E1, E2 and E3.

- MFCCs outperformed CQCCs in the SD test case, but CQCCs promise better SID models by showing less speaker-overfitting.

- iMFCCs performed best in the SD case.

- Among Speech Disorder specific features, articulation feature set performed the best among these.

Questions

(Q1) Given the difference in accuracies of models trained on the 2 datasets, which dataset would you pick? Why?

(Q2) Why were deltas and delta-deltas not used in the training of the Deep Learning Models?

(Q3) What can you infer about severity classes and speaker variability from Fig.4(b)?

References

- Joshy, Amlu & Rajan, Rajeev. (2021). Automated Dysarthria Severity Classification Using Deep Learning Frameworks. 116-120. 10.23919/Eusipco47968.2020.9287741.

  https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9762324

Slide Template from: https://slidesgo.com