IEEE ICASSP 2021, Paper ID: 4772

Automatic Dysarthric Speech Detection Exploiting Pairwise Distance-Based Convolutional Neural Networks

Parvaneh Janbakhshi[1,2], Ina Kodrasi[1], Hervé Bourlard[1,2]

[1]Idiap Research Institute, Martigny, Switzerland [2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

{parvaneh.janbakhshi, ina.kodrasi, herve.bourlard}@idiap.ch

Aim

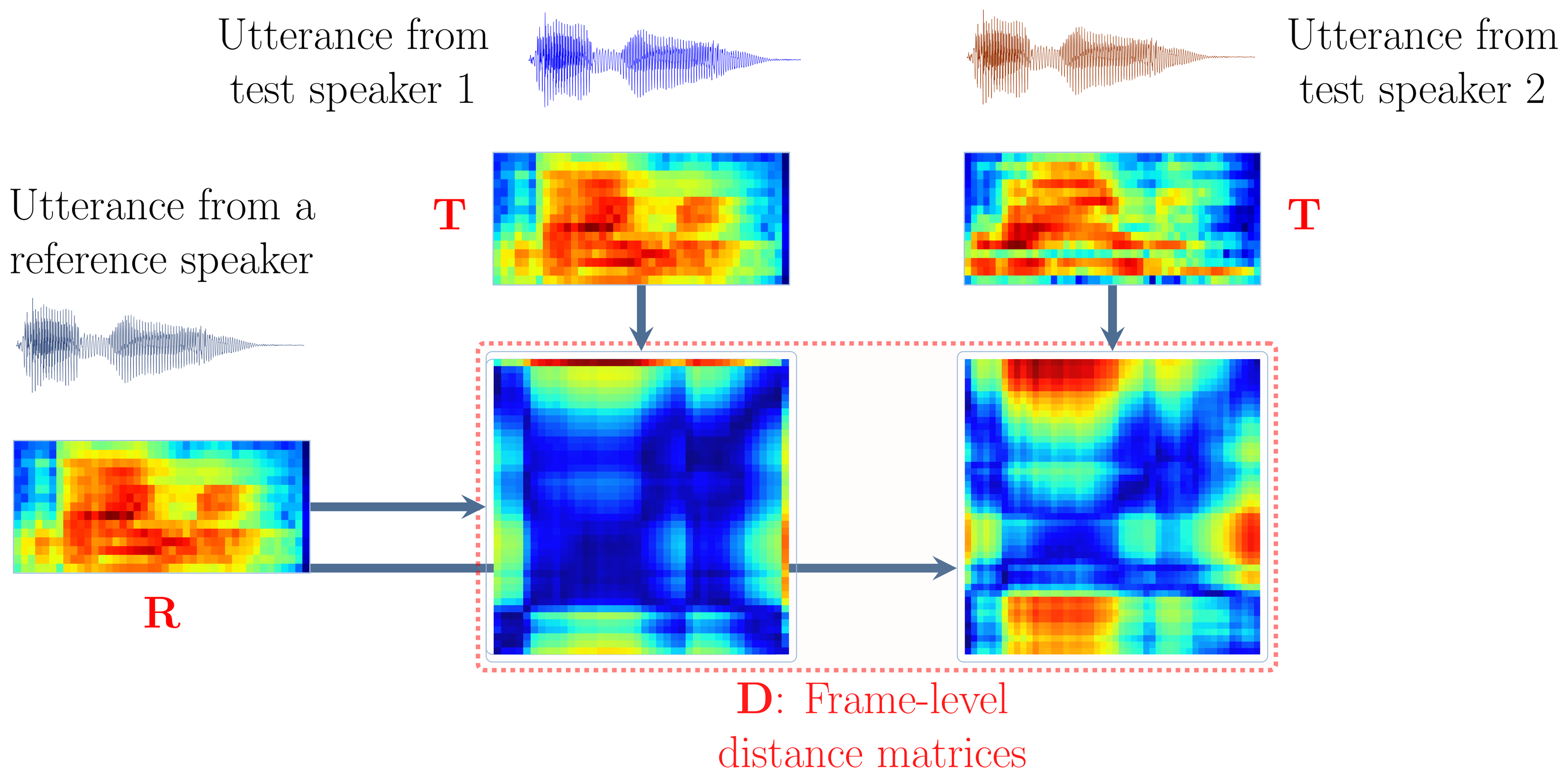
- Dysarthric speech detection
- Exploiting deep learning approaches while alleviating problems
  - Overfitting (due to limited training dysarthric data) and lack of robustness to unrelated speaker variabilities

Objectives

- Exploit pairwise training: advantageous for limited training data while guiding the network to extract robust features
- Use a single network for different utterances

Hypothesis

- The frame-level distance matrix between two healthy utterance representations has a different pattern than between a healthy and a pathological utterance representation
  - Analysing pairs of phonetically-balanced representations; one from a healthy speaker (i.e., the reference) and one from a test speaker

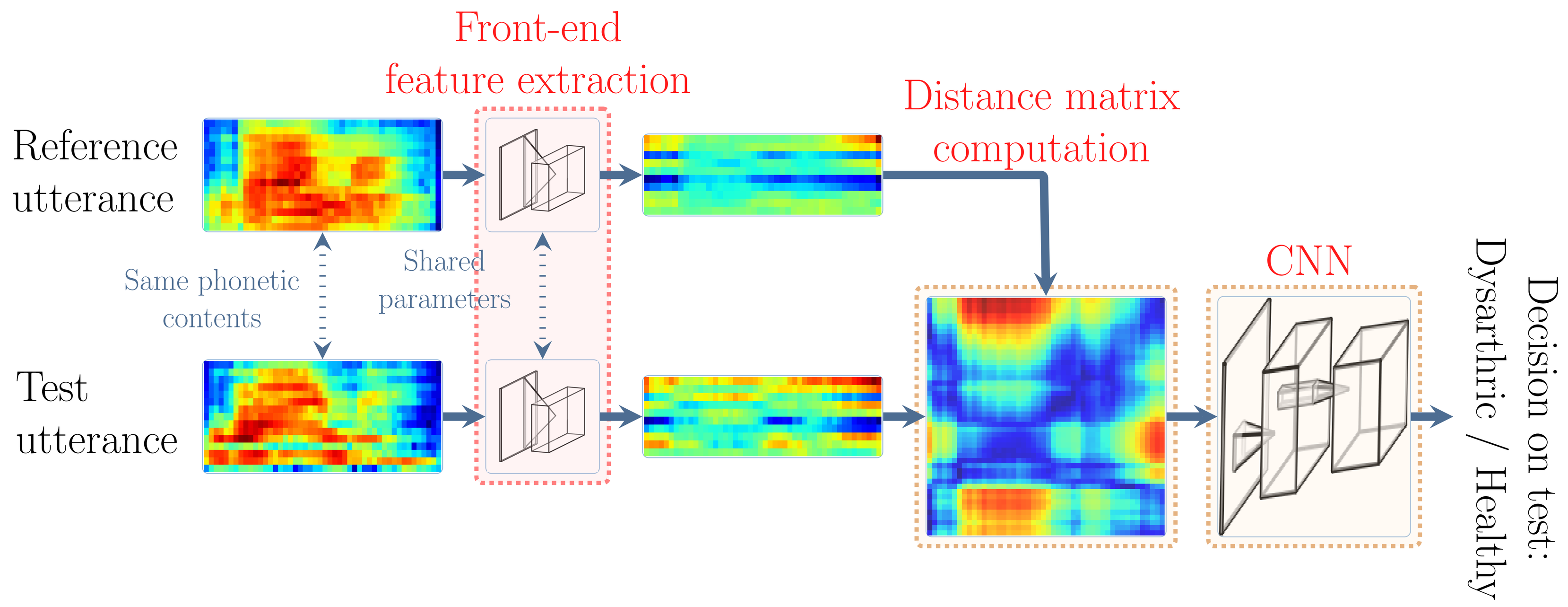


$\mathbf{r}_i$ and $\mathbf{t}_j$: reference and test feature vectors at time frame i and j

Distance matrix $\mathbf{D}$ with (i, j)–th entry: the distance d between $\mathbf{r}_i$ and $\mathbf{t}_j \rightarrow \mathbf{D}_{i,j} = d(\mathbf{t}_i, \mathbf{r}_j)$

Proposed method

1. Converting utterances to articulatory posteriors (APs) representations
   - Advantageous in comparison to short-time Fourier transform (STFT): AP representation characterises articulation, is robust to noise, has multilingual and crosslingual portability
2. Considering pairs of phonetically-balanced representations, one from a healthy speaker (i.e., the reference) and the other from the test speaker
   - Resizing to the same temporal dimension
3. Extracting features using a front-end convolutional layer
4. Computing the frame-level distance matrix from pairs of extracted feature representations
5. Analysing the computed distance matrices by the CNN classifier
   - Predicting whether the test speaker used for the distance matrix computation is healthy or dysarthric



The neural network blocks 3 4 5 are jointly optimized in an end-to-end framework

Evaluation

Databases

- 50 English-speaking patients with Parkinson's disease vs. 50 healthy speakers (10-fold cross-validation paradigm)
- 20 French-speaking patients with Amyotrophic Lateral Sclerosis and Parkinson's disease vs. 20 healthy speakers (5-fold cross-validation paradigm)

Performance evaluation

- Detection accuracy: percentage of correctly classified speakers and AUC: area under ROC curve

Baseline networks

- B-CNN$_1$: CNN trained on representations of short (i.e., 160 ms) segments of speech
- B-CNN$_2$: CNN trained on distance matrices computed directly from pairs of input representations (i.e., without using the front-end feature extraction layer)

Results

- Classification results using B-CNN$_1$ and two input representations

Database	Input representation	AUC	Accuracy [%]
Spanish PC-GITA	STFT	0.56	53.67
Spanish PC-GITA	AP	**0.75**	**72.00**
French MoSpeeDi	STFT	0.64	52.50
French MoSpeeDi	AP	**0.73**	**60.83**

- Classification results using AP representations

Database	CNN	AUC	Accuracy [%]
Spanish PC-GITA	Baseline B-CNN$_1$	0.75	72.00
Spanish PC-GITA	Baseline B-CNN$_2$	0.78	68.33
Spanish PC-GITA	Proposed	**0.83**	**77.67**
French MoSpeeDi	Baseline B-CNN$_1$	0.733	60.83
French MoSpeeDi	Baseline B-CNN$_2$	0.77	70.83
French MoSpeeDi	Proposed	**0.84**	**76.67**

- AP representations perform better than STFT independently of the language or diseases
- Proposed approach outperforms baseline systems for different databases

Conclusion

- The proposed automatic dysarthric speech detection system using a pairwise distance-based CNN operating on speech AP representations is feasible while being generalizable across languages and outperforming state-of-the-art CNN-based systems