Subspace-based Learning for Automatic Dysarthric Speech Detection

Parvaneh Janbakhshi, Ina Kodrasi, and Hervé Bourlard

Idiap Research Institute
Speech and Audio Processing Group

October 2020

Outline

Dysarthric speech detection

- Disturbances of muscular control on speech production system → dysarthria
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS), and Parkinson's disease (PD)

Dysarthric speech detection

- Disturbances of muscular control on speech production system → dysarthria
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS), and Parkinson's disease (PD)

- Dysarthric speech detection: discriminating between normal and dysarthric speech

Dysarthric speech detection

- Disturbances of muscular control on speech production system → dysarthria
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS), and Parkinson's disease (PD)
- Dysarthric speech detection: discriminating between normal and dysarthric speech

- Subjective screening based on judgement of medical practitioners
  - Labor-intensive
  - Inconsistency
  - Difficulties with early diagnosis
- Automatic and objective detection method
  - Efficient, economical
  - Repeatable
  - Early diagnosis

Initial motivations

- Focusing on connected speech analysis (words and sentences)
  - More challenging than sustained vowel analysis
  - Crucial for assessments of dysarthria
  - More realistic due to the usage in conversations in daily life
- Avoiding the need for any speech segmentations
- Avoiding large-scale feature extraction or selection process
  - Decreasing risks of overfitting

Initial motivations

- Focusing on connected speech analysis (words and sentences)
  - More challenging than sustained vowel analysis
  - Crucial for assessments of dysarthria
  - More realistic due to the usage in conversations in daily life
- Avoiding the need for any speech segmentations
- Avoiding large-scale feature extraction or selection process
  - Decreasing risks of overfitting

Our proposal: Subspace-Based Learning Method

1. Subspace-based speech modeling
2. Incorporating subspace-based discriminant analysis

Main motivations

- Dysarthric speech → atypical changes in spectro-temporal fluctuations (Rosen et al., 2006)
- The dominant spectro-temporal patterns of healthy and dysarthric speech differ

Main motivations

- Dysarthric speech → atypical changes in spectro-temporal fluctuations (Rosen et al., 2006)

- The dominant spectro-temporal patterns of healthy and dysarthric speech differ

1

Subspace-based learning method for dysarthric speech detection

1. Spectro-temporal subspace-based speech modeling of spectrograms
2. Incorporating subspace-based discriminant analysis

Main motivations

- Dysarthric speech → atypical changes in spectro-temporal fluctuations (Rosen et al., 2006)

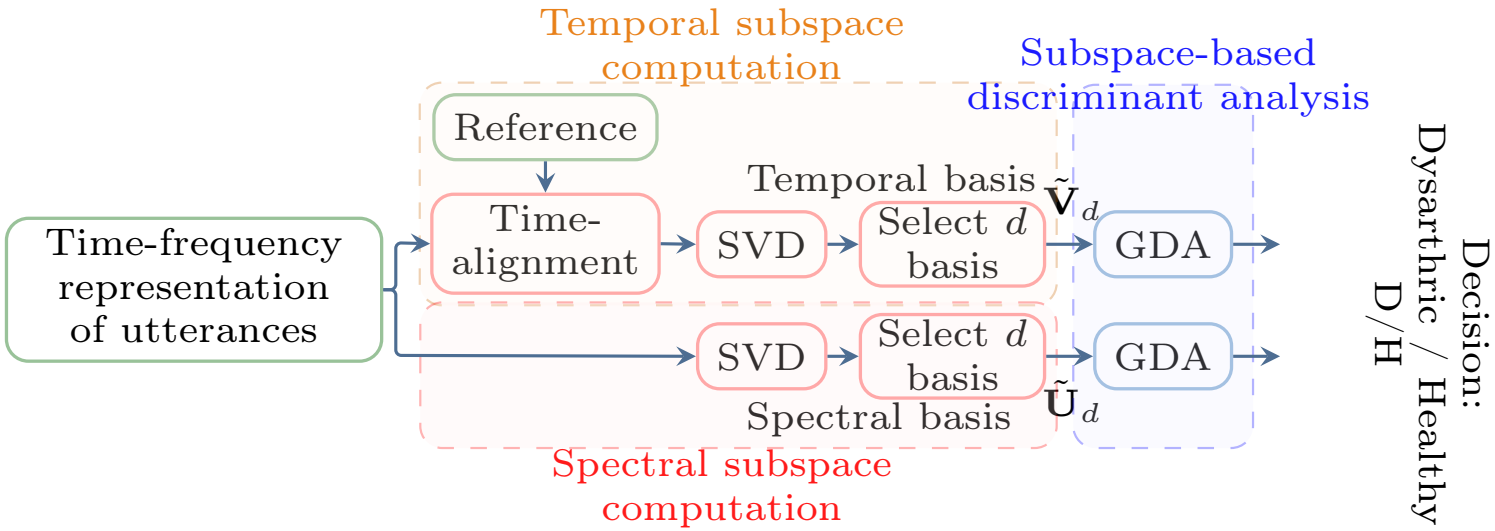
- The dominant spectro-temporal patterns (1) of healthy and dysarthric (2) speech differ

Subspace-based learning method for dysarthric speech detection

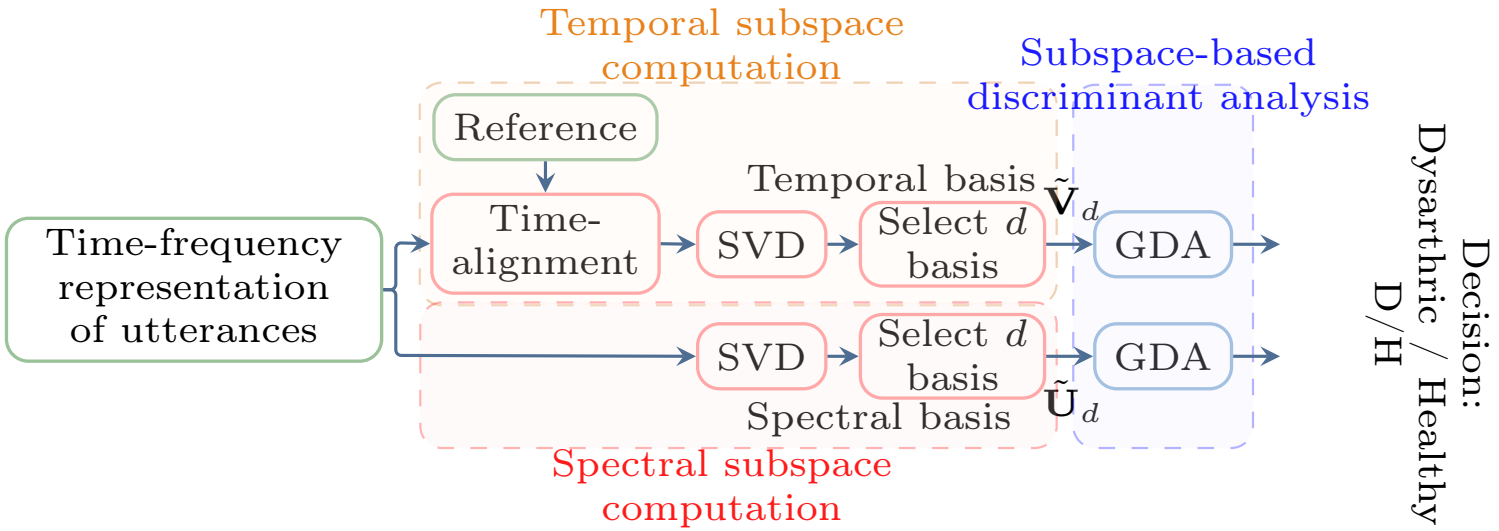
1. Spectro-temporal subspace-based speech modeling of spectrograms
2. Incorporating subspace-based discriminant analysis

Outline

Subspace-based learning method for dysarthric speech detection

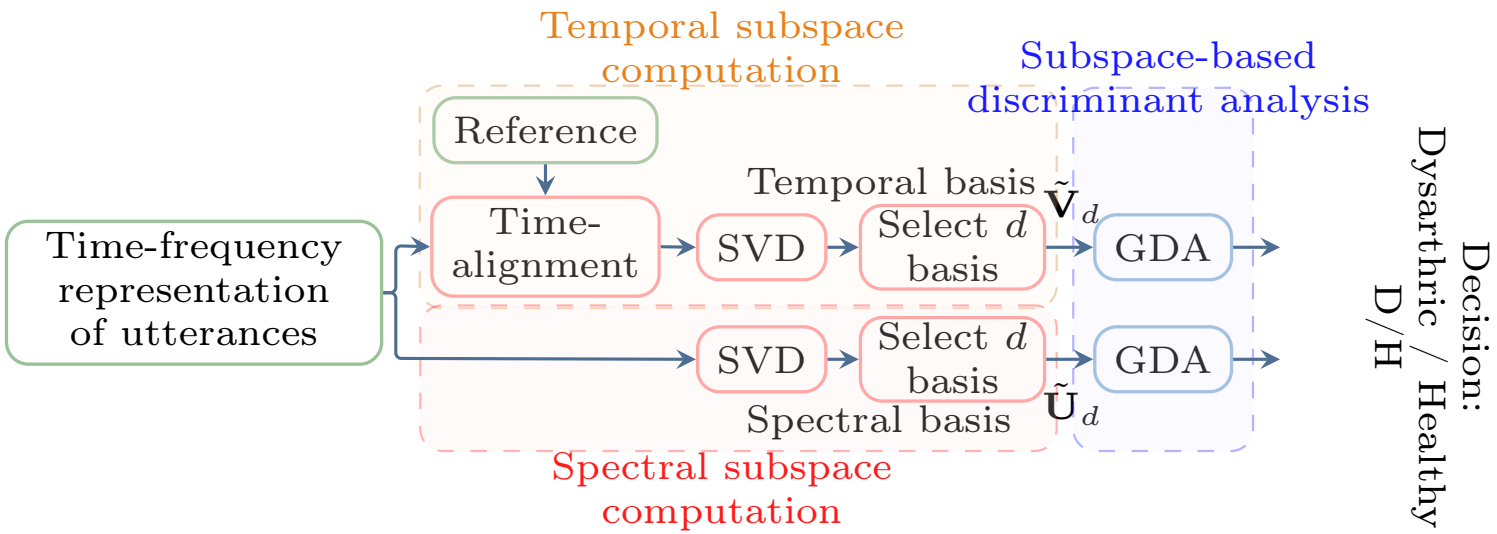


Subspace-based learning method for dysarthric speech detection



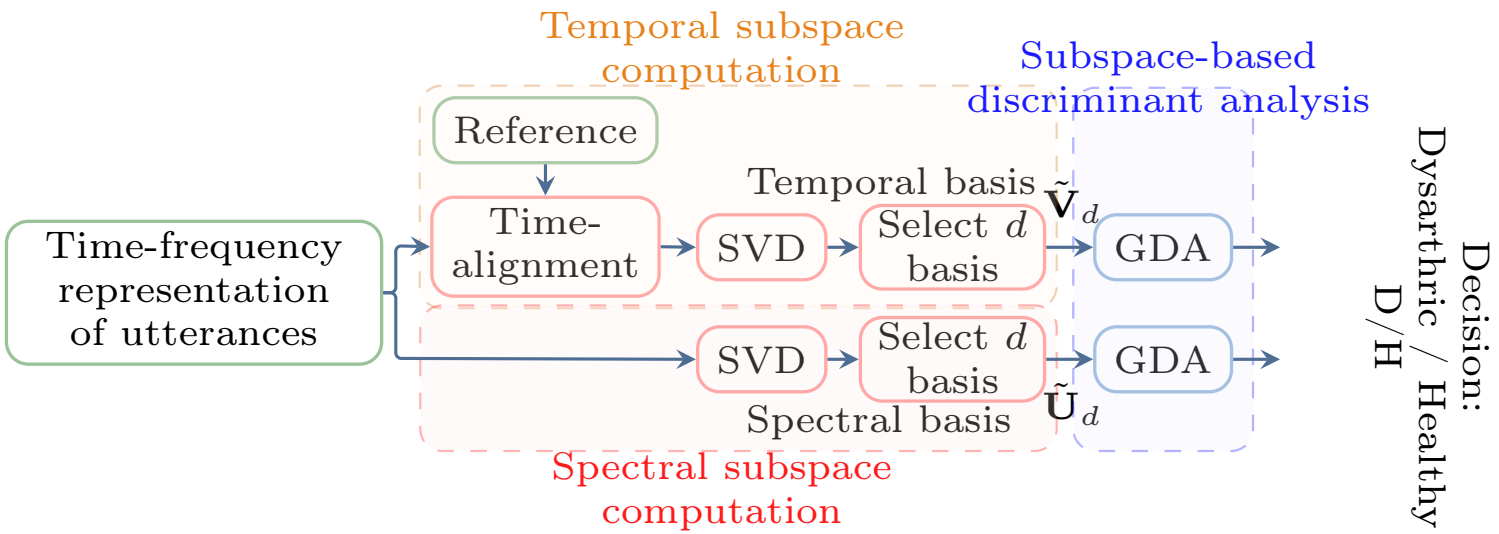
- Extracting spectral/temporal subspaces spanning the dominant spectral/temporal patterns of speech for each speakers
  - Spectral subspaces
  - Temporal subspaces (requires time-alignment)

Subspace-based learning method for dysarthric speech detection

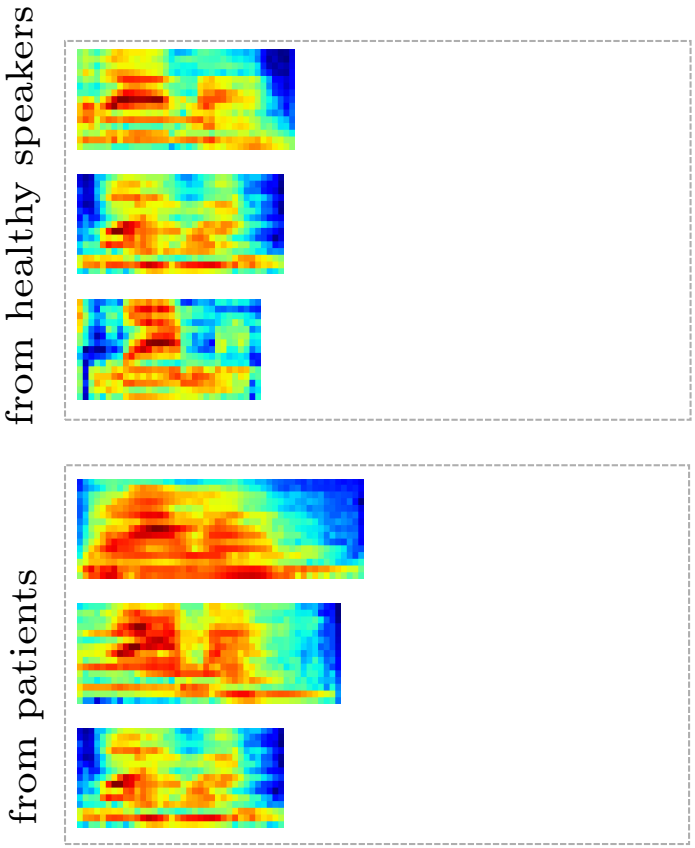


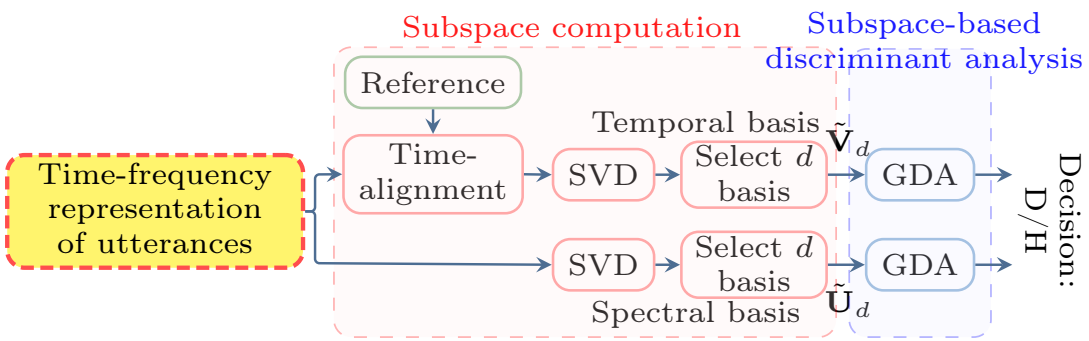
- Extracting spectral/temporal subspaces spanning the dominant spectral/temporal patterns of speech for each speakers
  - Spectral subspaces
  - Temporal subspaces (requires time-alignment)
- Subspaces (representing speakers) lie in a non-Euclidean space called the Grassmann manifold

Subspace-based learning method for dysarthric speech detection

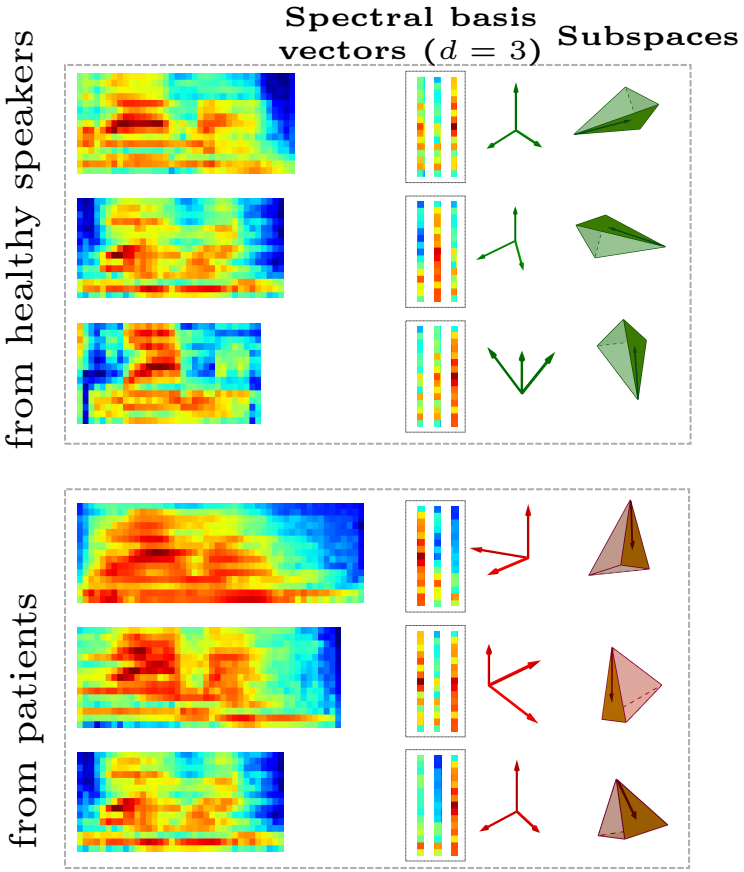


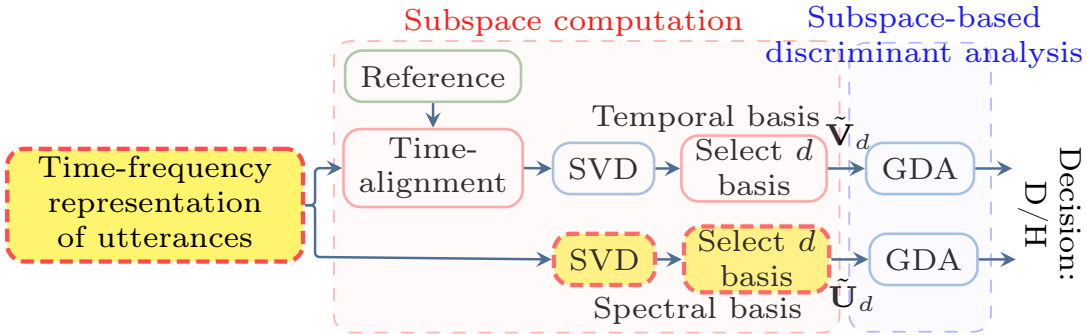
- Extracting spectral/temporal subspaces spanning the dominant spectral/temporal patterns of speech for each speakers
  - Spectral subspaces
  - Temporal subspaces (requires time-alignment)
- Subspaces (representing speakers) lie in a non-Euclidean space called the Grassmann manifold
- Discriminant analysis on Grassmann manifold using kernel LDA with the Grassmann kernels → Grassmann discriminant analysis (GDA) (Hamm and Lee, 2008)





- TF domain: logarithm of the one-third octave band spectrum
- $\mathbf{S}_m$: $(J \times N_m)$–dimensional TF representation from speaker m
- J: number of octave bands, N_m: number of time frames with $J \ll N_m$

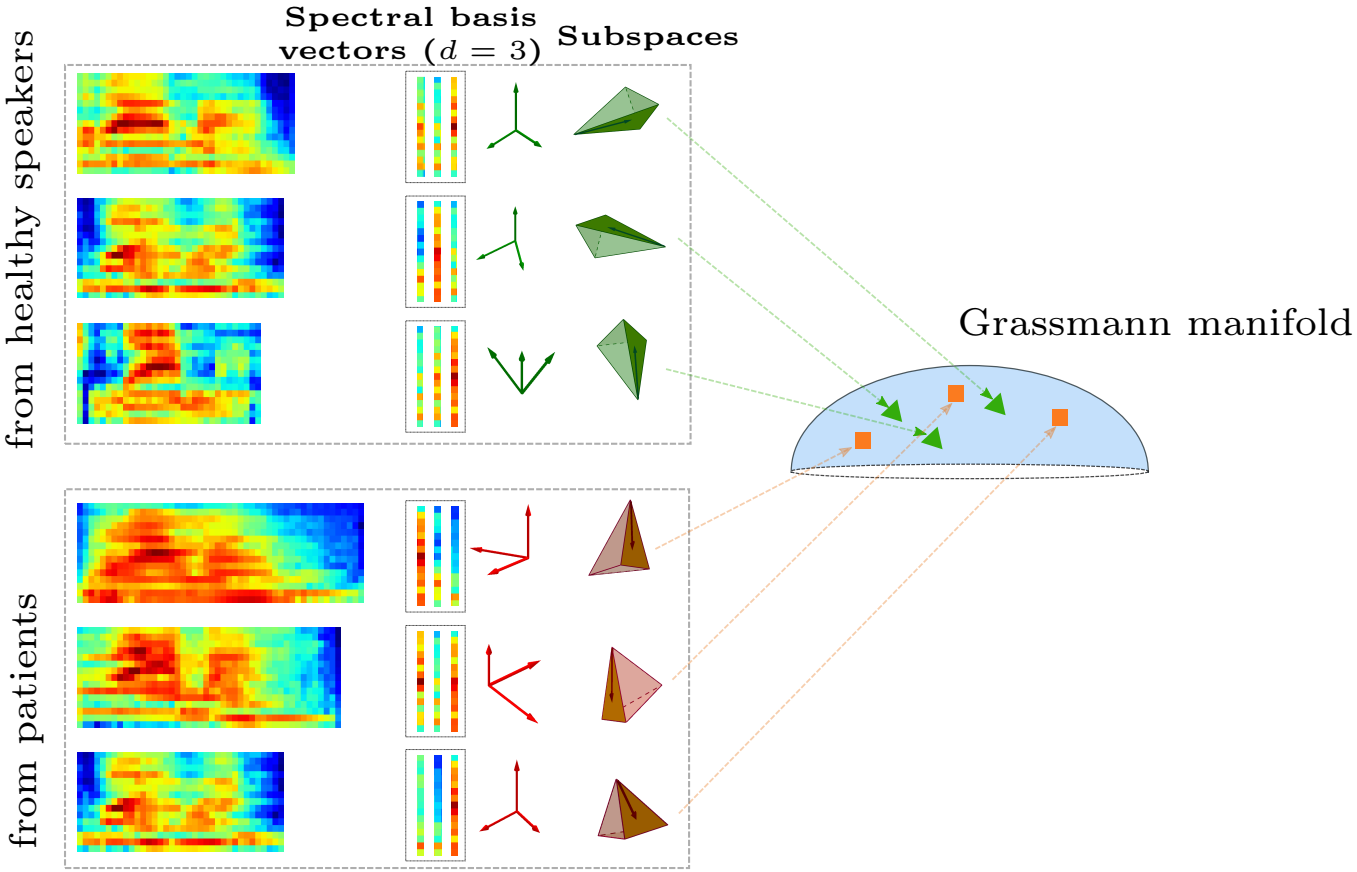


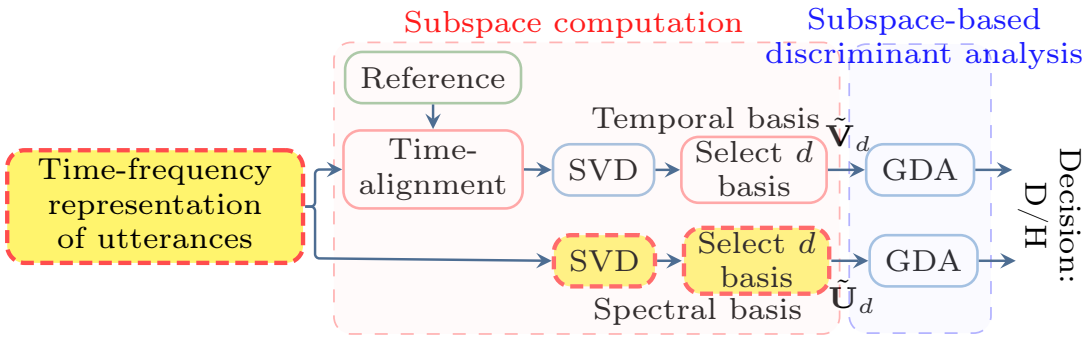


- SVD of TF representations:

$$\mathbf{S}_m = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

- First d dominant spectral basis vectors in $\mathbf{U} \Rightarrow \tilde{\mathbf{U}}_d$
- $\tilde{\mathbf{U}}_d$ spanning the dominant spectral patterns in $\mathbf{S}_m \Rightarrow$ spectral subspace



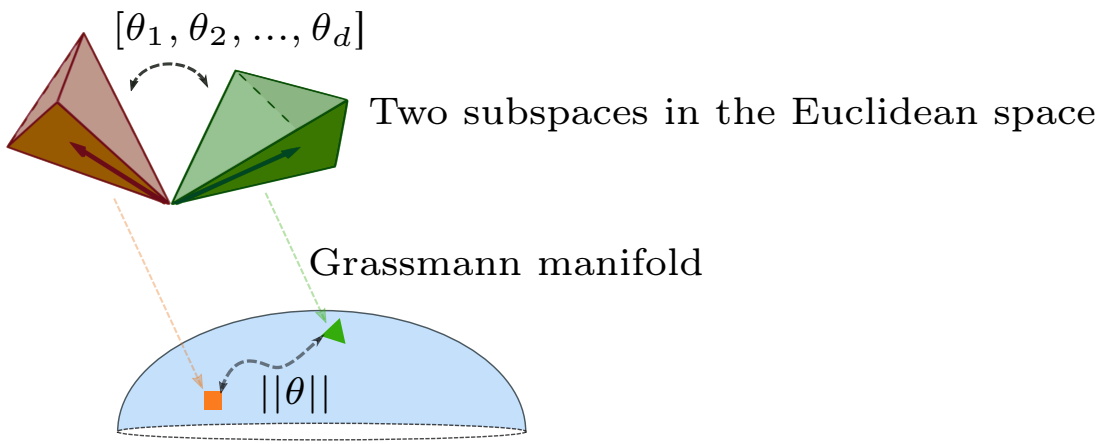


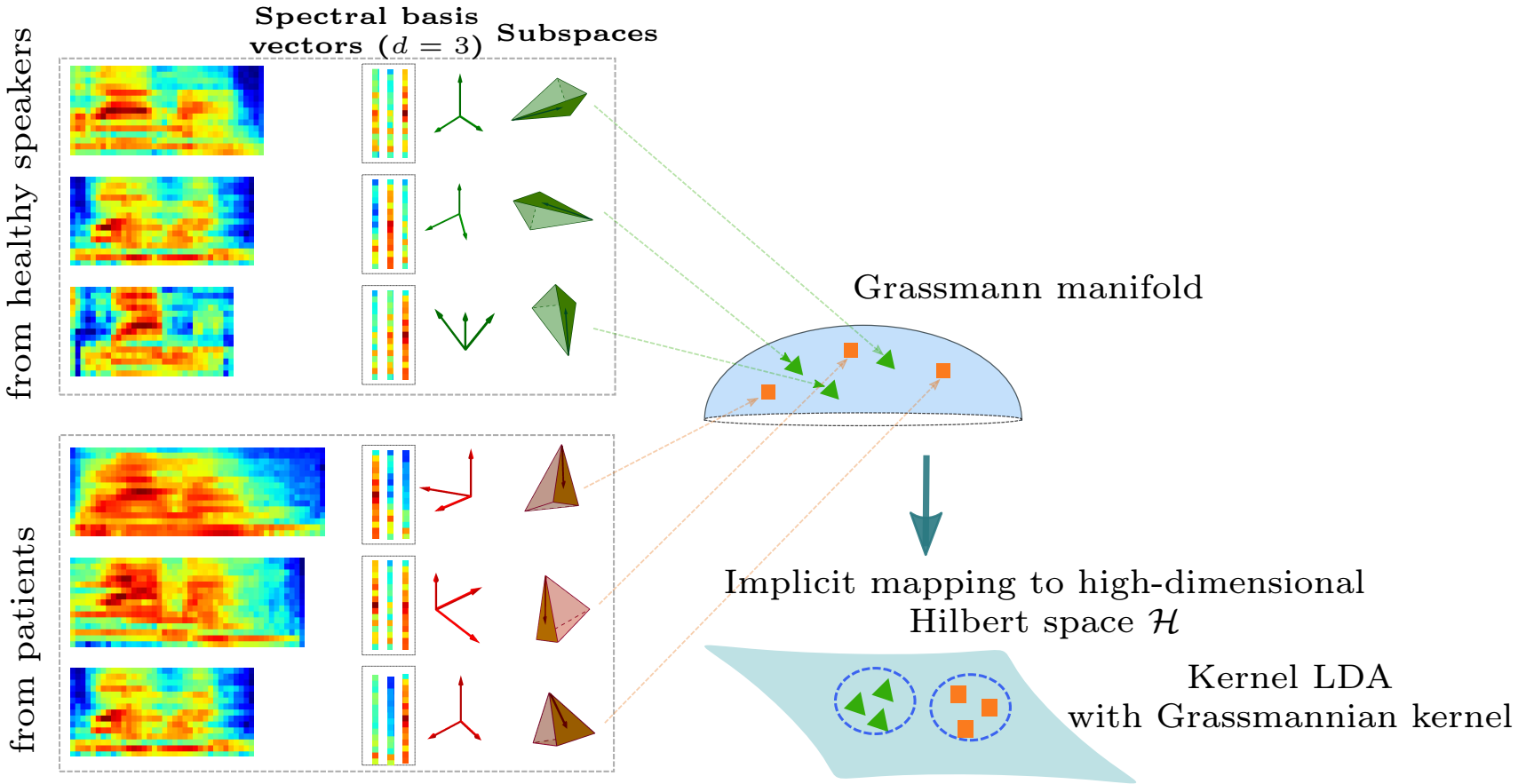
- Subspaces lie in the Grassmann manifold (not obeying Euclidean geometry)
- Healthy and dysarthric speakers are represented by points on the Grassmann manifold

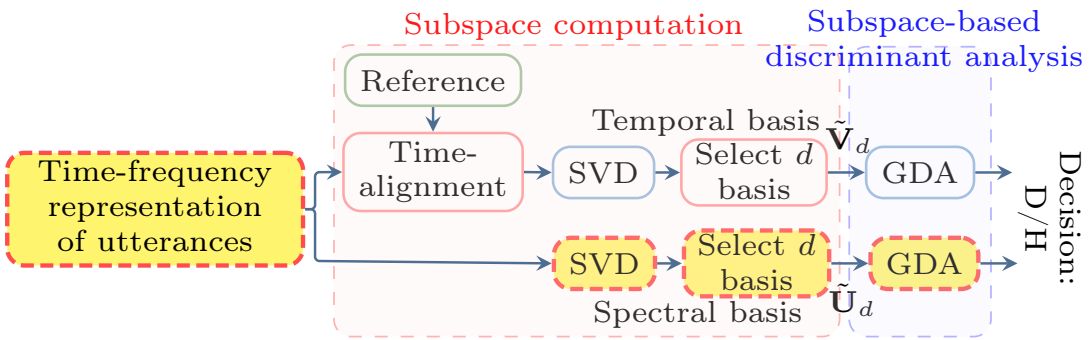
Grassmann manifold

- Subspace distances are defined with particular combinations of the principal angles between subspaces
- Valid metric on Grassmann manifolds → valid kernel function compatible with the metric
- $\mathbf{Y}_p$ and $\mathbf{Y}_q$ being the orthonormal matrices representing the subspaces of speakers p and q, we use the kernel defined (Hamm and Lee, 2008):

$$k(\mathbf{Y}_p, \mathbf{Y}_q) = ||\mathbf{Y}_p^T \mathbf{Y}_q||_F^2,$$

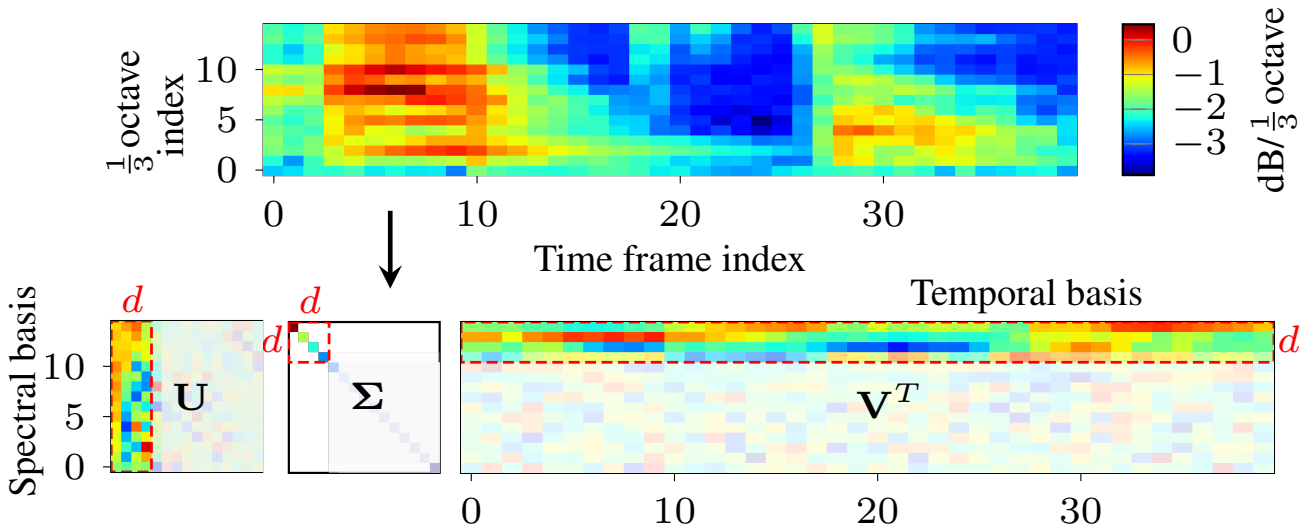






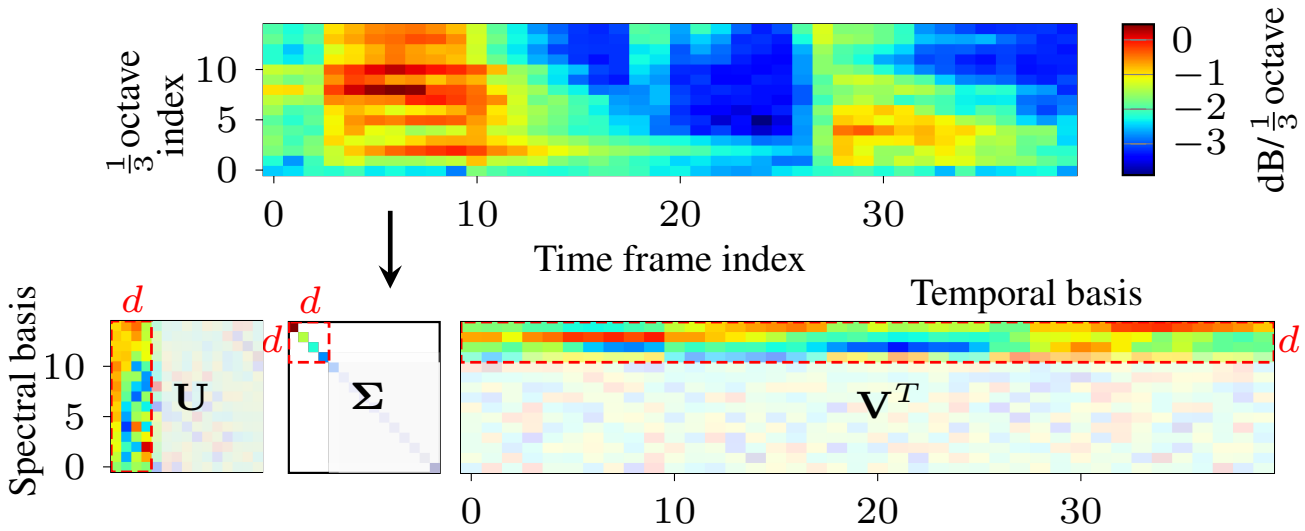
» Applying kernel LDA using a Grassmann kernel respecting the geometry of subspaces on the manifold (GDA) (Hamm and Lee, 2008)

Temporal subspaces computations



$$\mathbf{S}_m = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

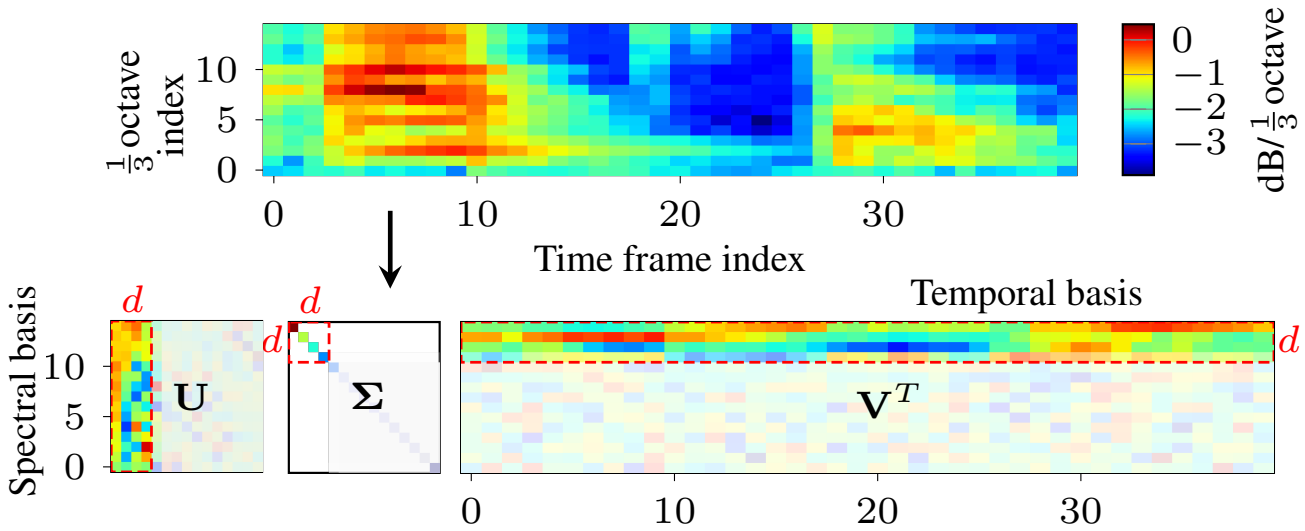
Temporal subspaces computations



$$\mathbf{S}_m = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

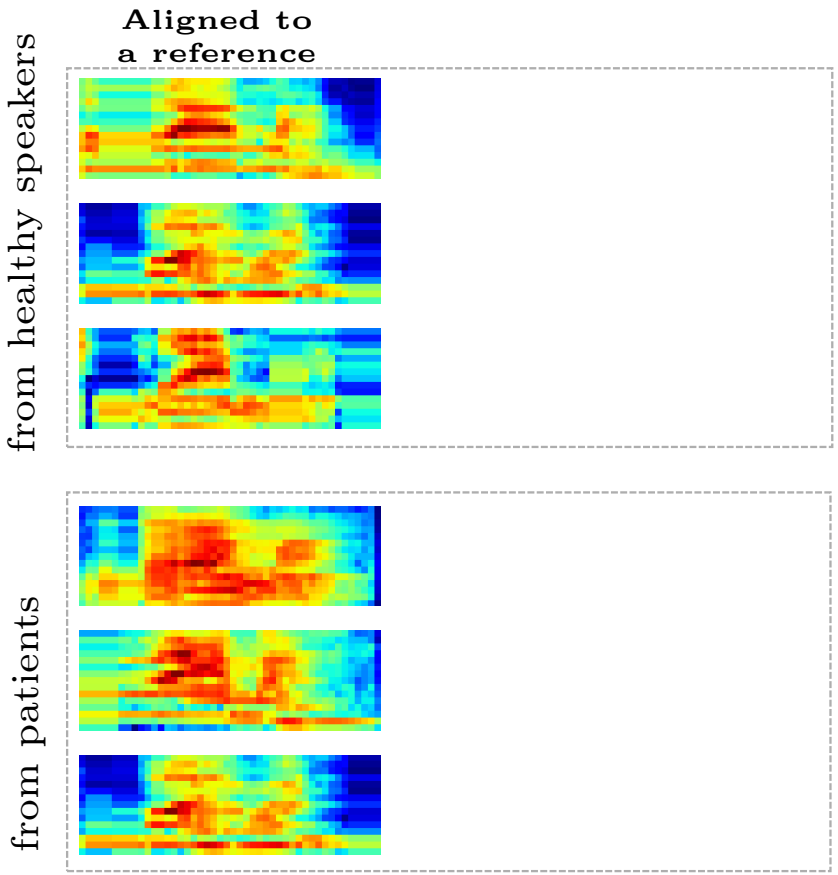
» Temporal basis vectors $\mathbf{V}$ obtained from different speakers cannot be directly compared to each other → unaligned TF representations

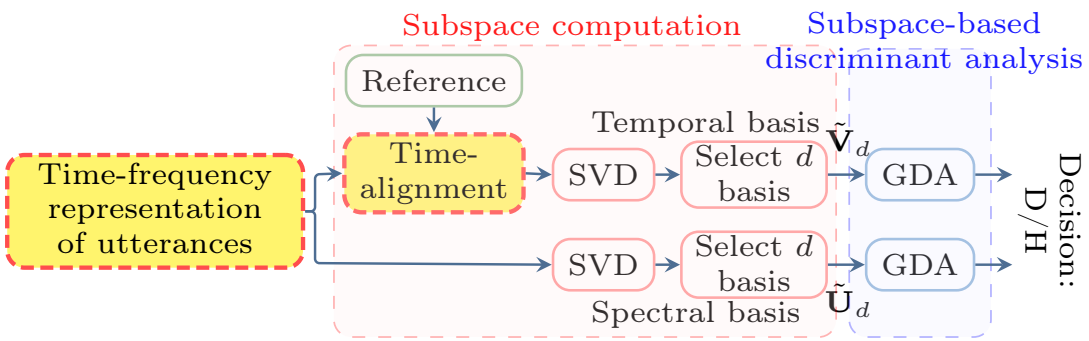
Temporal subspaces computations



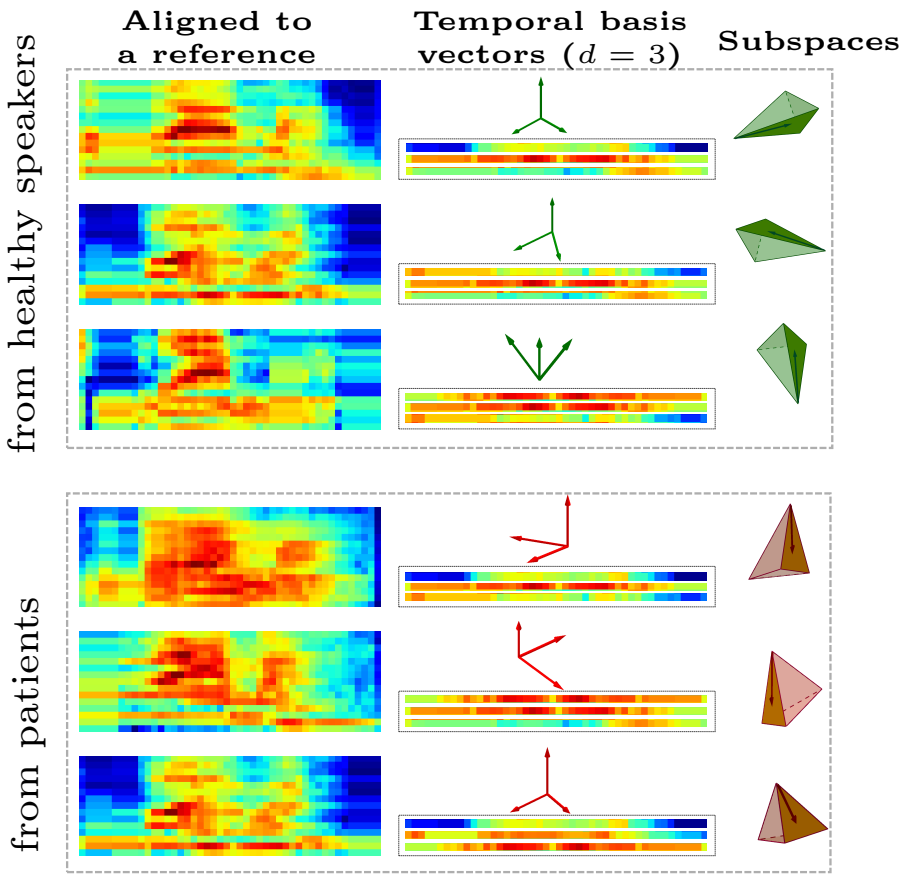
$$\mathbf{S}_m = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

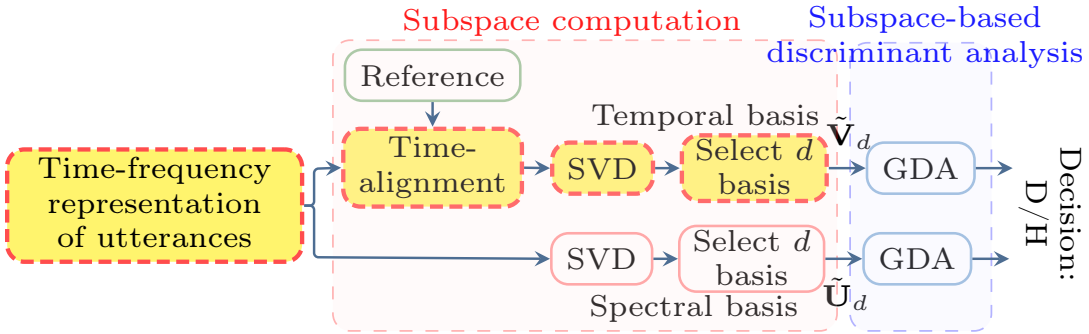
- Temporal basis vectors $\mathbf{V}$ obtained from different speakers cannot be directly compared to each other → unaligned TF representations
- Prior to computing the temporal basis vectors → time-aligning all TF representations
  - Time-aligning to an (arbitrarily selected) healthy reference representations using DTW + temporal averaging (Kodrasi and Bourlard, 2020)





- Time-aligning all TF representations to a reference
- $\hat{\mathbf{S}}_m$: time-aligned representation from speaker m

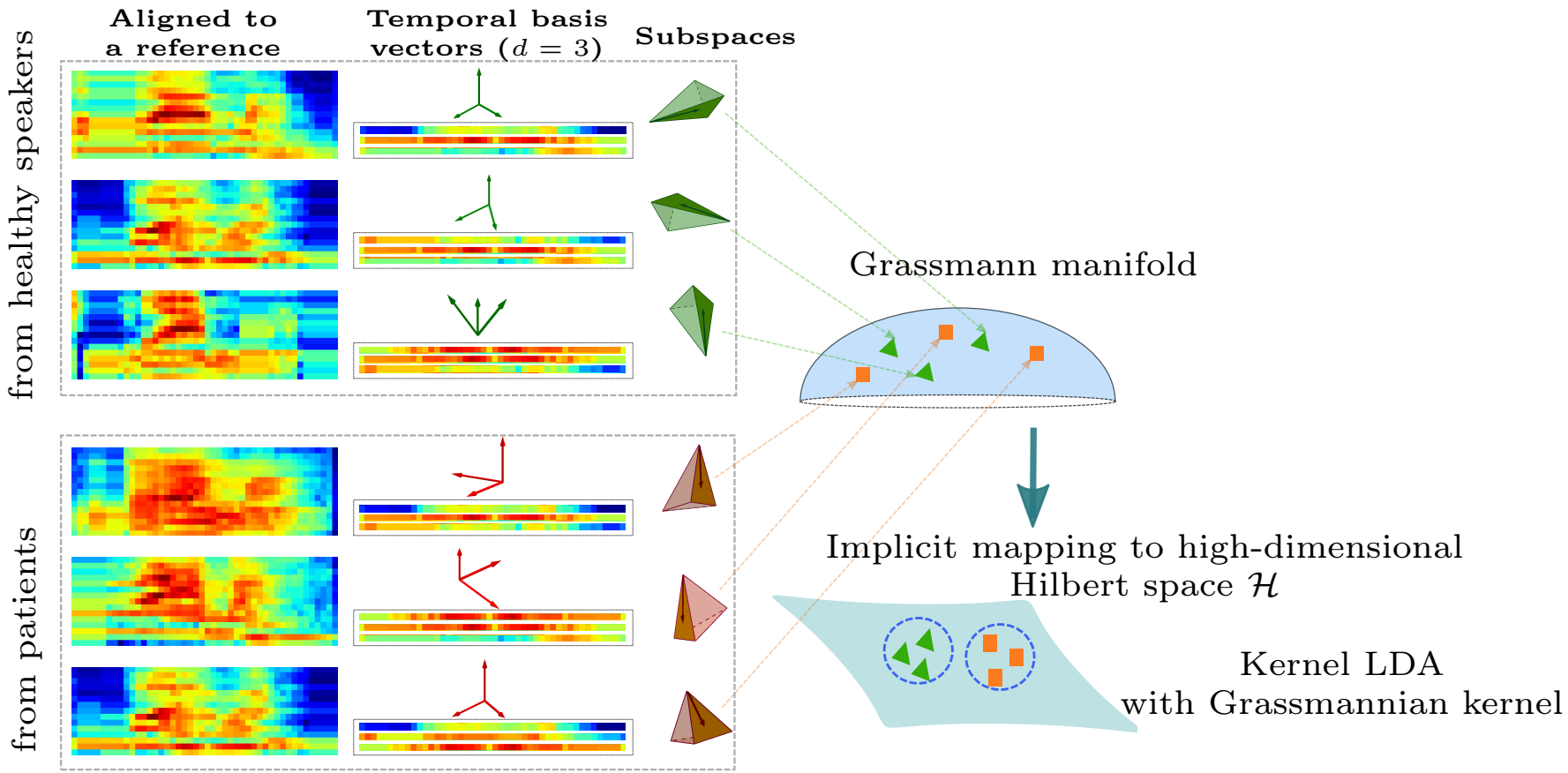


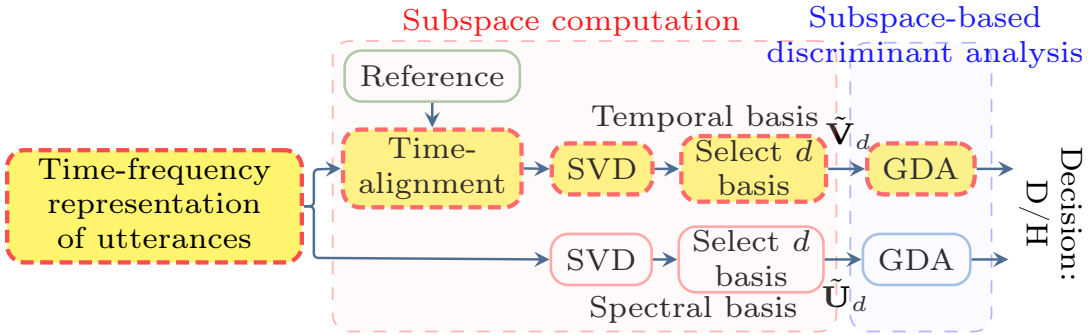


- SVD of $\hat{\mathbf{S}}_m$:

$$\hat{\mathbf{S}}_m = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^T$$

- First d dominant temporal basis vectors in $\hat{\mathbf{V}} \Rightarrow \hat{\mathbf{V}}_d$
- $\tilde{\mathbf{V}}_d$ spanning the dominant temporal patterns $\Rightarrow$ temporal subspace





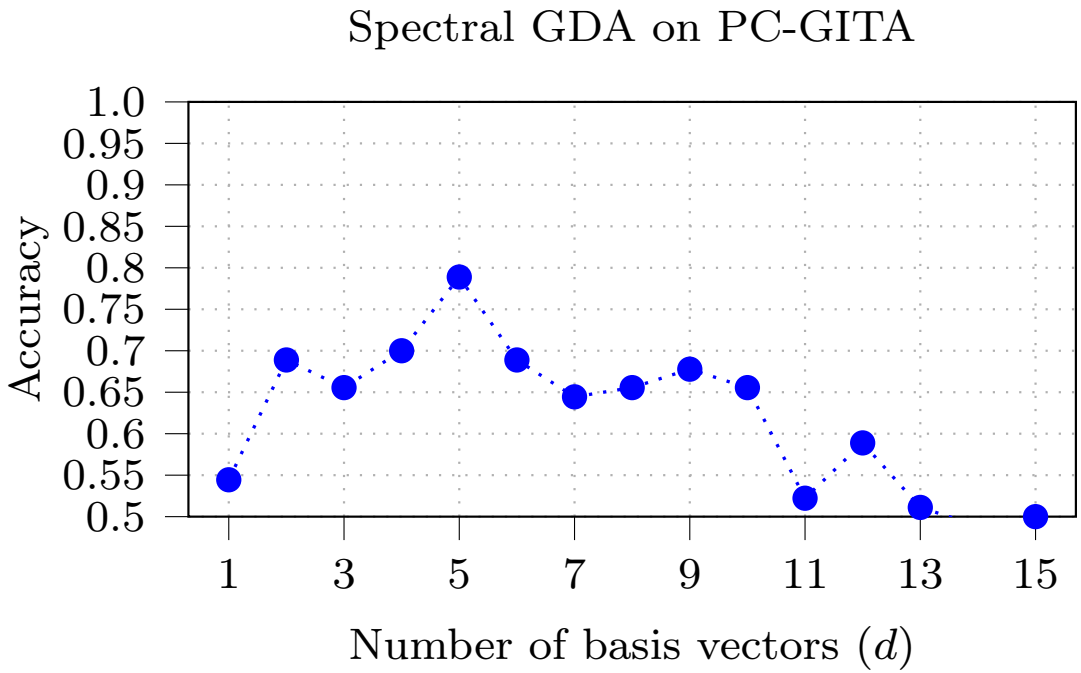
» Applying GDA on subspaces to discriminate dysarthric and healthy speakers

Outline

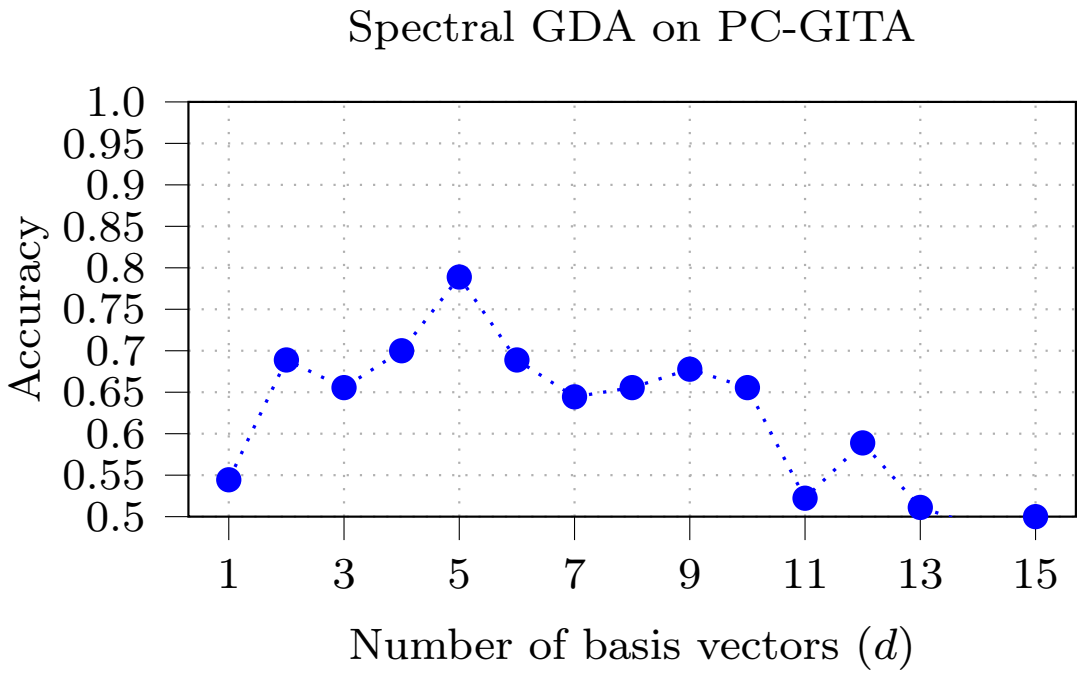
Experimental results

- Datasets
  - **Spanish** PC-GITA database (Orozco et al., 2014)
    - 45 **PD patients** vs. 45 healthy speakers (9-fold CV paradigm)
    - 5 healthy speakers for time-alignment (i.e., references)
  - **English** Universal Access database (Kim et al., 2008)
    - 15 **CP patients** vs. 12 healthy speakers (Leave-one-out paradigm)
    - 1 healthy speaker for time-alignment
  - **French** MoSpeeDi database
    - 20 **dysarthic patients** vs. 20 healthy speakers (4-fold CV paradigm)
    - 10 healthy speakers for time-alignment
- State-of-the-art techniques
  - SVM classifier with a radial basis kernel function using features:
    - Statistics of the F0, jitter, shimmer, Harmonics-to-Noise-Ratio (HNR), and MFCCs
    - Temporal sparsity parameter β_k (Kodrasi and Bourlard, 2018)
- Evaluation
  - Detection accuracy: percentage of correctly classified speakers

Experimental results



Experimental results



- Automatic selection of number of spectral/temporal basis vectors (d)
  - Tuning $d \rightarrow$ a grid search with $d \in \{1, ..., J\}$ based on the accuracy of nested cross-validation on the training set

Experimental results

- Analyzing the sensitivity of the temporal GDA to the reference speaker selection
  - Computing the performance using different reference speakers → mean and standard deviation of the temporal GDA accuracy across different reference speakers

Experimental results

Method	PC-GITA	MoSpeeDi	UA-Speech
T-GDA	**82.0** $\pm$ 3.5	**80.5** $\pm$ 4.7	**96.3**
S-GDA	61.1	75.0	85.2
SVM using f_0	48.9	62.5	55.6
SVM using jitter	52.2	55.0	88.9
SVM using shimmer	52.2	60.0	77.8
SVM using HNR	46.7	45.0	81.5
SVM using MFCCs	75.6	55.5	92.6
SVM using μ	72.2	67.5	88.9

S-GDA $\rightarrow$ spectral subspace-based approach, T-GDA $\rightarrow$ temporal subspace-based approach

Experimental results

Method	PC-GITA	MoSpeeDi	UA-Speech
T-GDA	**82.0** $\pm$ 3.5	**80.5** $\pm$ 4.7	**96.3**
S-GDA	61.1	75.0	85.2
SVM using f_0	48.9	62.5	55.6
SVM using jitter	52.2	55.0	88.9
SVM using shimmer	52.2	60.0	77.8
SVM using HNR	46.7	45.0	81.5
SVM using MFCCs	75.6	55.5	92.6
SVM using μ	72.2	67.5	88.9

S-GDA $\rightarrow$ spectral subspace-based approach, T-GDA $\rightarrow$ temporal subspace-based approach

- Temporal GDA performs better than spectral GDA
  - Temporal patterns $\rightarrow$ higher discriminative power than the spectral patterns

Experimental results

Method	PC-GITA	MoSpeeDi	UA-Speech
T-GDA	**82.0** $\pm$ 3.5	**80.5** $\pm$ 4.7	**96.3**
S-GDA	61.1	75.0	85.2
SVM using f_0	48.9	62.5	55.6
SVM using jitter	52.2	55.0	88.9
SVM using shimmer	52.2	60.0	77.8
SVM using HNR	46.7	45.0	81.5
SVM using MFCCs	75.6	55.5	92.6
SVM using μ	72.2	67.5	88.9

S-GDA $\rightarrow$ spectral subspace-based approach, T-GDA $\rightarrow$ temporal subspace-based approach

- Temporal GDA performs better than spectral GDA
  - Temporal patterns $\rightarrow$ higher discriminative power than the spectral patterns
- Temporal GDA is not highly sensitive to the reference speaker selection

Experimental results

Method	PC-GITA	MoSpeeDi	UA-Speech
T-GDA	**82.0** $\pm$ 3.5	**80.5** $\pm$ 4.7	**96.3**
S-GDA	61.1	75.0	85.2
SVM using f_0	48.9	62.5	55.6
SVM using jitter	52.2	55.0	88.9
SVM using shimmer	52.2	60.0	77.8
SVM using HNR	46.7	45.0	81.5
SVM using MFCCs	75.6	55.5	92.6
SVM using μ	72.2	67.5	88.9

S-GDA $\rightarrow$ spectral subspace-based approach, T-GDA $\rightarrow$ temporal subspace-based approach

- Temporal GDA performs better than spectral GDA
  - Temporal patterns $\rightarrow$ higher discriminative power than the spectral patterns
- Temporal GDA is not highly sensitive to the reference speaker selection
- Temporal GDA outperforms state-of-the-art acoustic features across three databases; generalisable across languages and pathologies

Outline

Summary

- A subspace-based approach to automatically discriminate between dysarthric and healthy speech
  - Representing speakers through spectral or temporal subspaces spanned by the dominant spectral or temporal basis vectors of the octave band representation of speech
  - Applying subspace-based discriminant analysis

- Experimental results on three databases
  - Compared to spectral subspaces, temporal subspaces are more successful in characterizing dysarthric speech
  - Subspace-based approach using temporal subspaces outperforms state-of-the-art features

Outlook

Completed analyses ✓

» Further investigations on the subspace-based approach using spectral subspaces

- Using other representations e.g., MFCCs
- Modifying the TF representations to yield joint spectra-temporal subspace analysis
- Using non-linear, i.e., kernel, subspace analysis of speech instead of linear subspace + kernel GDA

Outlook

Completed analyses ✓

- Further investigations on the subspace-based approach using spectral subspaces
  - Using other representations e.g., MFCCs
  - Modifying the TF representations to yield joint spectra-temporal subspace analysis
  - Using non-linear, i.e., kernel, subspace analysis of speech instead of linear subspace + kernel GDA

In future ✗

- Would have been nice to investigate
  - Applicability of spectral GDA in phonetically-unbalanced scenarios (spontaneous conversation)
  - To what extent, subspace characterisations can capture other information e.g., speaker identity, emotions, ...

Thank You

Reference

Hamm, J. and Lee, D. D. (2008). Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proc. 25th International Conference on Machine Learning*, pages 376–383, Helsinki, Finland.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Proc. 9th Annual Conference of the International Speech Communication Association*, pages 1741–1744, Brisbane, Australia.

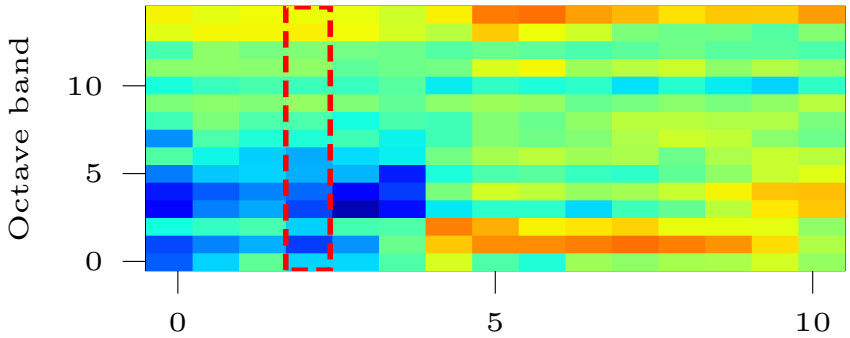
Kodrasi, I. and Bourlard, H. (2018). Statistical modeling of speech spectral coefficients in patients with parkinson's disease. In *Proc. 13th ITG conference on Speech CommunicationSpeech Communication*, pages 1–5, Oldenburg, Germany.

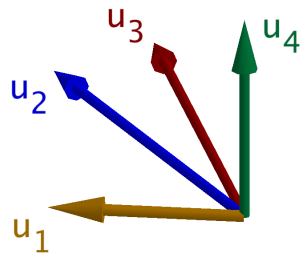
Kodrasi, I. and Bourlard, H. (2020). Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28(1).

Orozco, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J., González-Rátiva, M., and Noeth, E. (2014). New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Proc. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

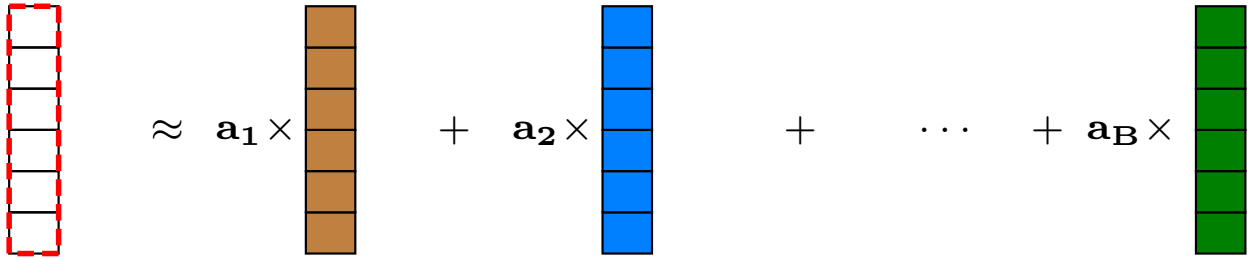
Rosen, K. M., Kent, R. D., Delaney, A. L., and Duffy, J. R. (2006). Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers. *Journal of Speech Language and Hearing Research*, 49(2):395–411.

Spectral basis vectors





Spectral bases of speech



Spectral vector (at time frame i = 3) **Spectral basis 1 $\mathbf{u_1}$** **Spectral basis 2 $\mathbf{u_2}$** **Spectral basis B $\mathbf{u_B}$**

Subspace-based intelligibility (SBI) measure