Automatic estimation of Parkinson's disease severity from diverse speech tasks

J. Kim, M. Nasir, R. Gupta, M. Segbroeck, D. Bone, M. Black, Z. Skordilis, Z. Yang, P. Georgiou, S. Narayanan

Department of Electrical Engineering

University of Southern California

jangwon@usc.edu

Focus

- Four speaking tasks
  - Task classification
  - Task-specific modeling (less variability)
- The level of PD severity is evaluated for individual patients, not for each speech utterance.
  - Clustering test-set utterances based on UPDRS score
  - Rating all utterances of each cluster jointly
- Long speech: repeated syllables, sentences, text, monologue
  - Novel features for stylizing speech rhythm in prosody

System overview

Test set

Individual rating

Rating for each test utterance

Speaking task classifier

Unsupervised clustering

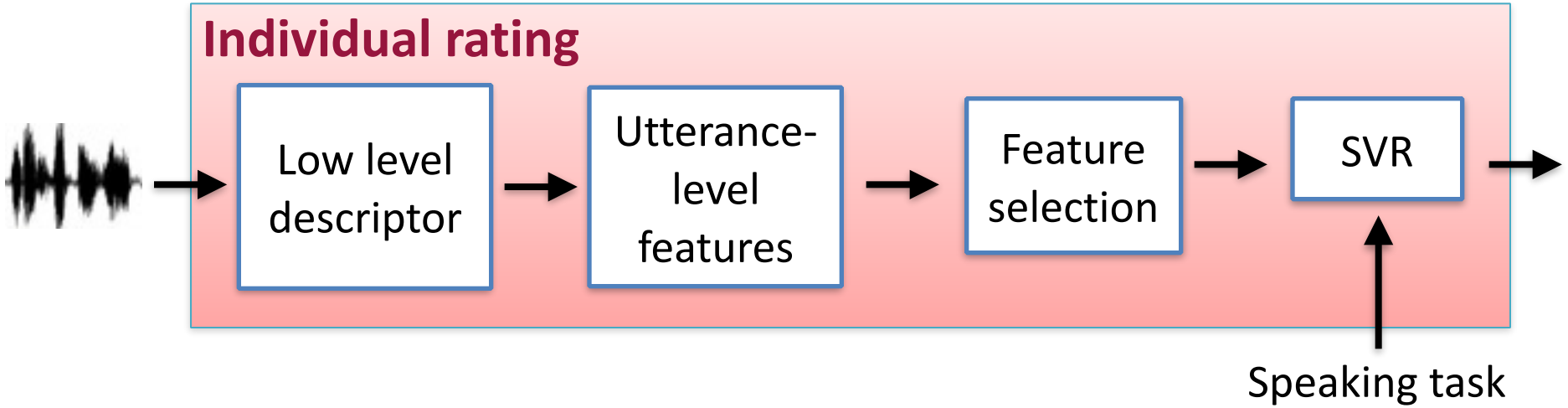
Joint rating

Labels for the test set

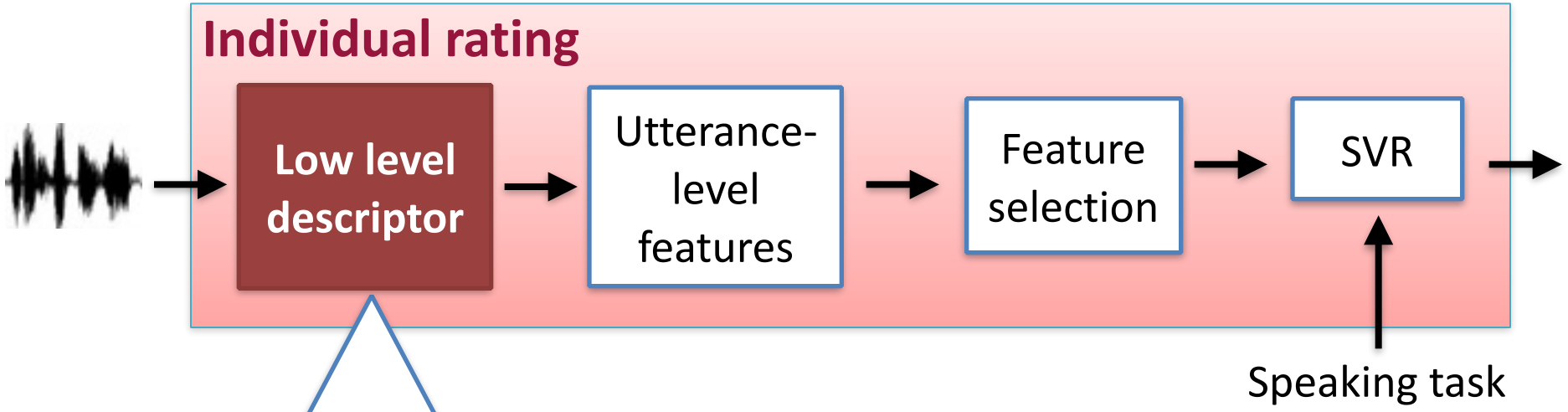
Rating for the test utterances of each cluster

This system performs **two-stage rating**: Initial rating for individual utterances, followed by final rating for all utterances of each cluster.

Individual rating system



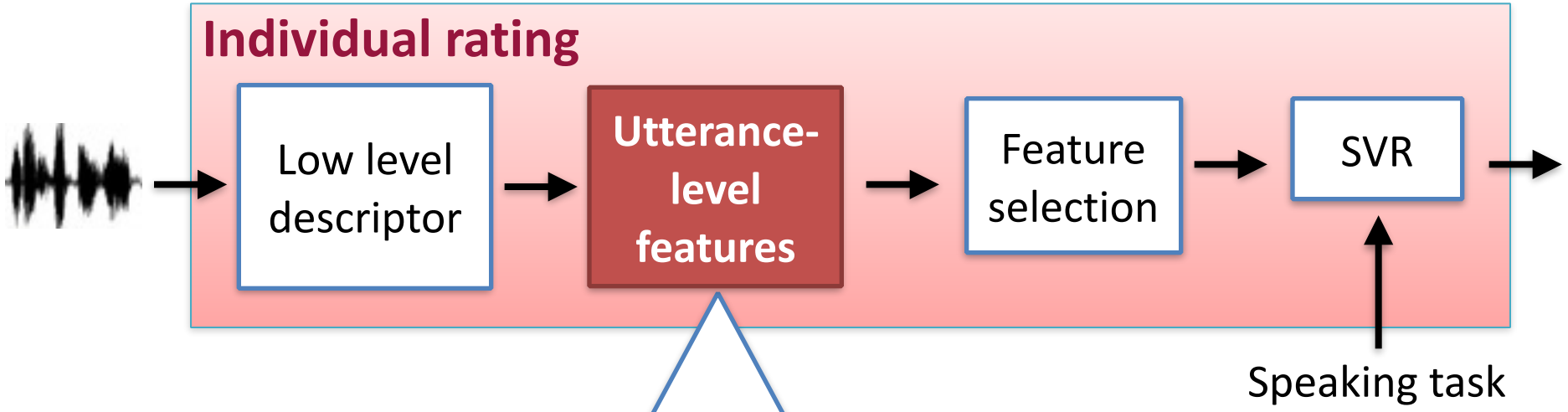
Individual rating system



Spectral	MFCCs, MFBs, Spectral shape, GFCCs, Gabor, Spectro-temporal modulation, long-term spectral variability profile
Prosodic	F0, RMS energy
Voice quality	HNR, jitter, shimmer
ASR feature	ASR-posterior entropy*

**K. Audhkhasi et al., "Theoretical analysis of diversity in an ensemble of automatic speech recognition systems," Transactions on Audio, Speech & Language Processing, 2014.*

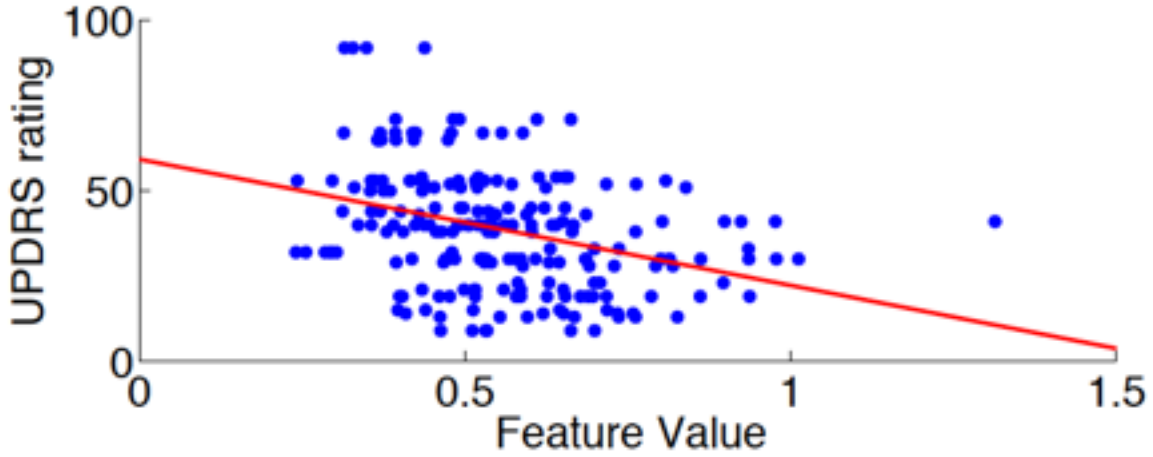
Individual rating system



Functionals (Fnls)	[0.1 0.25 0.5 0.75 0.9] quantiles, interquartile range, kurtosis and skewness	Speech region
		Consonant region + Vowel region
I-vector	UBM-fused total variability modeling technique*	
NTSA	Utterance-level Non-linear time series features	

** M. Van Segbroeck, R. Travadi, and S. S. Narayanan, "UBM fused total variability modeling for language identification," in Proceedings of Interspeech. ISCA, 2014.*

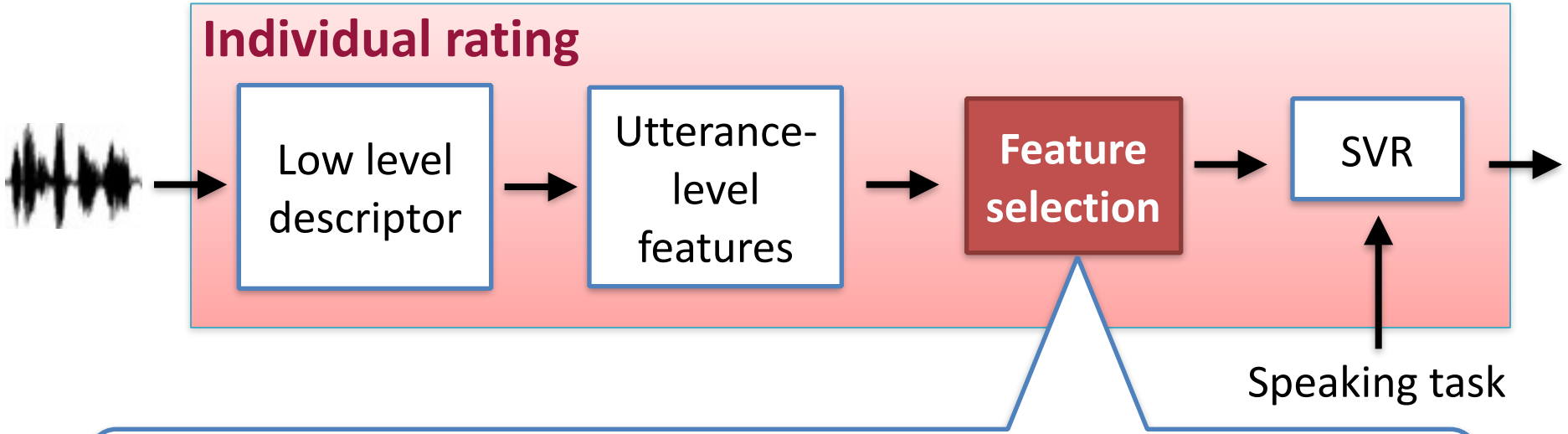
Non-linear time series features



0.9 quantile of LLE from glottal flow derivative (Spearman correlation = -0.34)

- Correlation Dimension (CD), Largest Lyapunov Exponent (LLE), smoothed LLE, Fractal Dimension (FD) in both time and phase spaces
- Prosodic feature streams (f0, RMS energy): Directly computed the 5 NTSA feature streams after smoothing
- Glottal flow and its delta, and speech waveform: functionals of the 5 NTSA features in every 40ms window, 20ms shifting

Individual rating system



- **Step 1 (relevant)**: Choose if spearman > threshold (~0.2)
- **Step 2 (representative)**: Select one among highly correlated features (Spearman >0.98), which follows Normal the best
- **Step 3 (consistent)**: Select if $\mu_{train} - \sigma_{train} < \mu_{dev/test} < \mu_{train} + \sigma_{train}$

This feature selection method helps to reduce regression accuracy mismatch between partitions and folds.

Speaking task classifier

- 4 task categories: isolated words, repeated syllables, long speech (text & monologue), sentences
- Acoustic model: triphone HMM trained on Mexican Spanish Broadcast News corpus
- Language model: unigram trained on list of prompt
- Tuning k (best k=1) by by four-folds cross-validation on the training set
- Weighted classification accuracy: **98%** on the devel. set

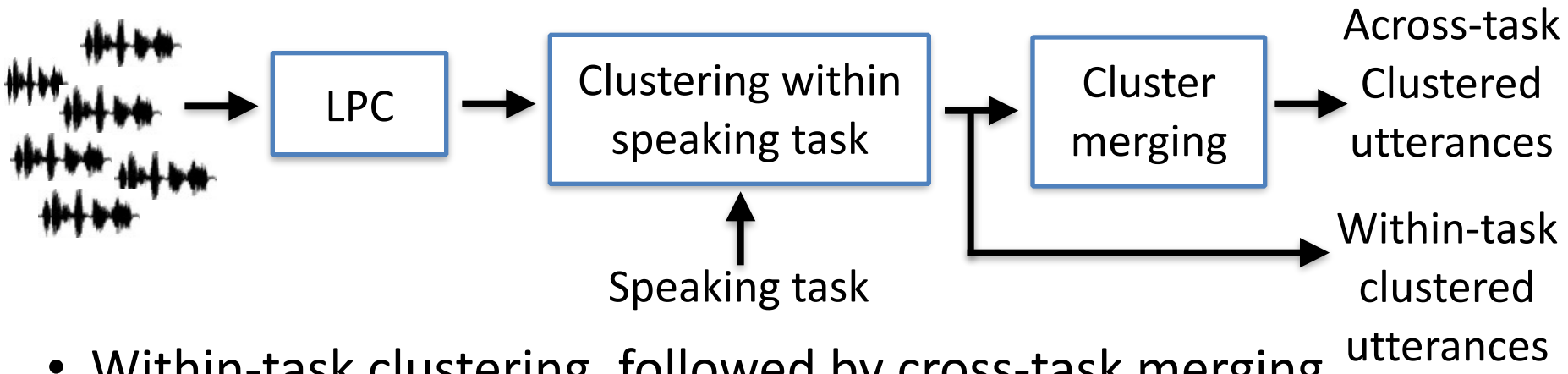
Within-task results

	Words	Repeated syllables	long (text + monologue)	sentences
openSMILE	0.53	0.54	0.20	0.47
NTSA	0.32	0.56	0.37	0.43
Fnls	0.36	0.60	0.25	0.37
C Fnls+V Fnls	0.38	0.61	0.23	0.41
i-vector	0.36	0.59	0.34	0.42

Spearman correlation of individual subsystem on each speaking task data in the devel. set. SVR parameter C (for regularization) is tuned on the training set.

- Best performing subsystem varies depending on the speaking task.
- Functionals computed on consonant and vowel regions separately shows higher Spearman correlation than Functionals on all voice region, except long speech task.

Two-level UPDRS clustering



- Within-task clustering, followed by cross-task merging
- Benefit: Less clustering error due to lexical similarity
- A single Gaussian based bottom-up agglomerative hierarchical clustering*
- Inter-cluster distance: Generalized likelihood ratio**
- Majority ratio: (devel. set)

All-task	Within-task	Two-level
0.70	0.90	0.85

Results: UPDRS Rating accuracy

	Devel set	Test set
Baseline ($C=10^{-3}$)	0.49	0.24
Baseline ($C=10^{-5}$)	0.37	0.39
Feature-level fusion	0.51	——
Joint rating 1	0.50	**0.43**
Joint rating 2	**0.53**	0.42

Spearman correlation of different systems on the development set and the test set. Joint rating 1: Within-task within-cluster joint rating. Joint rating 2: Merged-cluster joint rating

- 4-folds cross-validation on the training set for fusion

The proposed fusion systems (with and without the joint rating schemes) shows higher Spearman correlation than the baselines.

Results for Eating Challenge

	Train CV	Test set
Baseline	61.3	65.9
Functionals	65.3	——
I-vector	75.7	——
System fusion	76.2	74.6

Spearman correlation of different systems on the development set and the test set.

- Leave-one-speaker-out cross-validation for parameter tuning
- SVM classifier with fifth-order polynomial kernel
- Fusion of i-vector system with the SVM posteriors from the functionals

- I-vector system shows better performance than functionals.
- System fusion shows the best performance.

This work is supported by NSF, NIH, DARPA.

Lessons

- Effective features vary depending on the nature of the specific speaking task.
- Results show potentials of Non-linear time series analysis for capturing atypicality in Parkinson's speech.
- It is useful to check the cross-fold/partition behavior of features.