

# The Traces of Inner Life

Lecture 06 (Analysis III)

Alessandro Vinciarelli



# Outline

- Social Signals
- Speech Articulation and Tract Variables
- Articulation as a Dynamic Process
- Conclusions

# Outline

- Social Signals
- Speech Articulation and Tract Variables
- Articulation as a Dynamic Process
- Conclusions

# Social Signals

“[...] acts or structures that influence the behavior or internal state of other individuals.”

# Social Signals

“actions whose function is to bring about some reaction or to engage in some process.”

# Social Signals

“[...] communicative or informative signals which [...] provide information about social facts.”

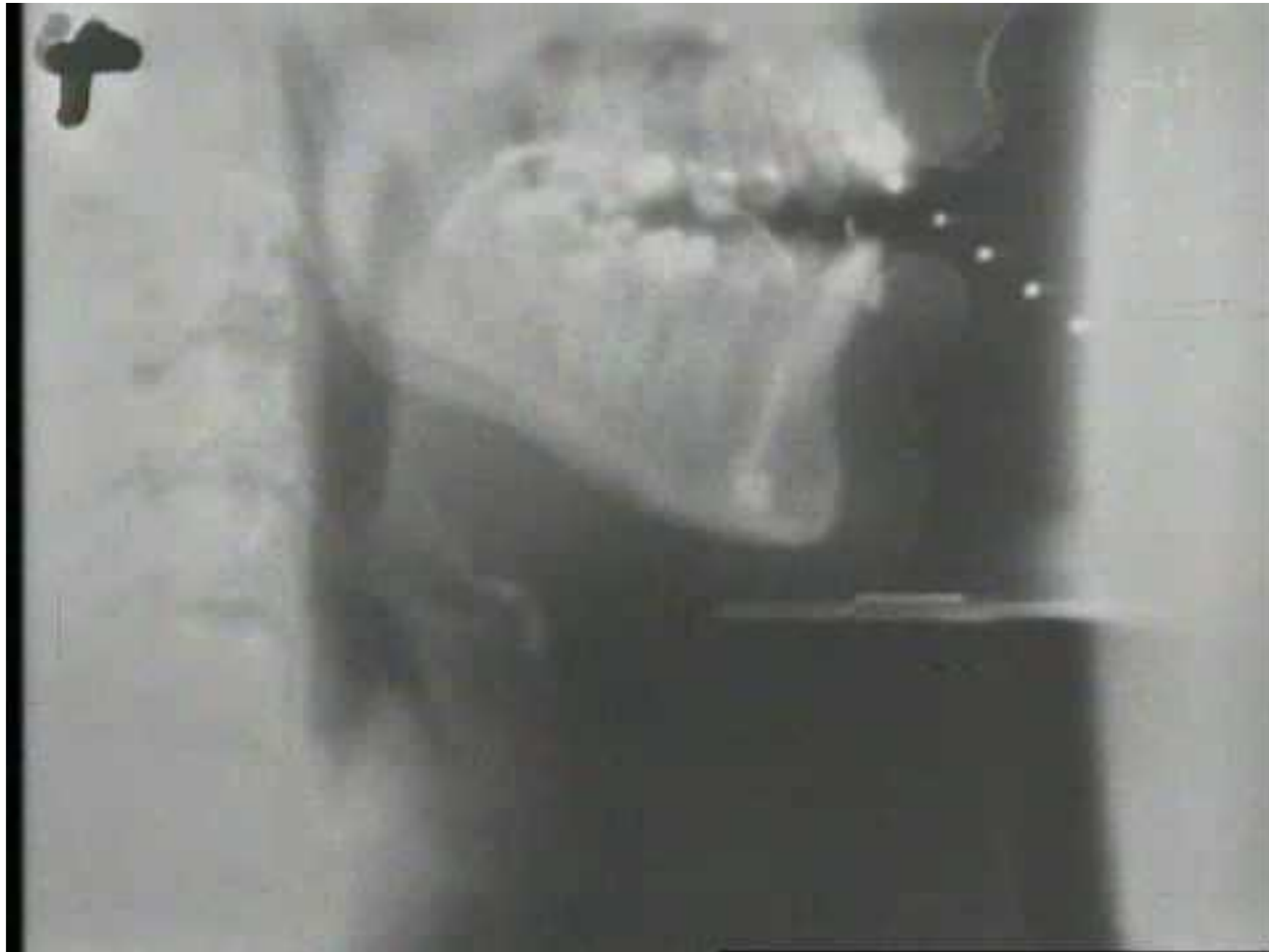
# Social Signals as Physical Traces

- Social signals are the **physical, machine detectable traces** of social and psychological phenomena;
- If this was not the case, it would not be **possible** for them **to influence others or attract reactions**;
- Similarly, it would not be **possible** to **provide information** about social facts;
- In some cases, their identification is simple (e.g., a smile), in others it is more challenging (e.g., the “speaking style” of depression patients).

# Outline

- Social Signals
- Speech Articulation and Tract Variables
- Articulation as a Dynamic Process
- Conclusions

# Articulation



# Articulation

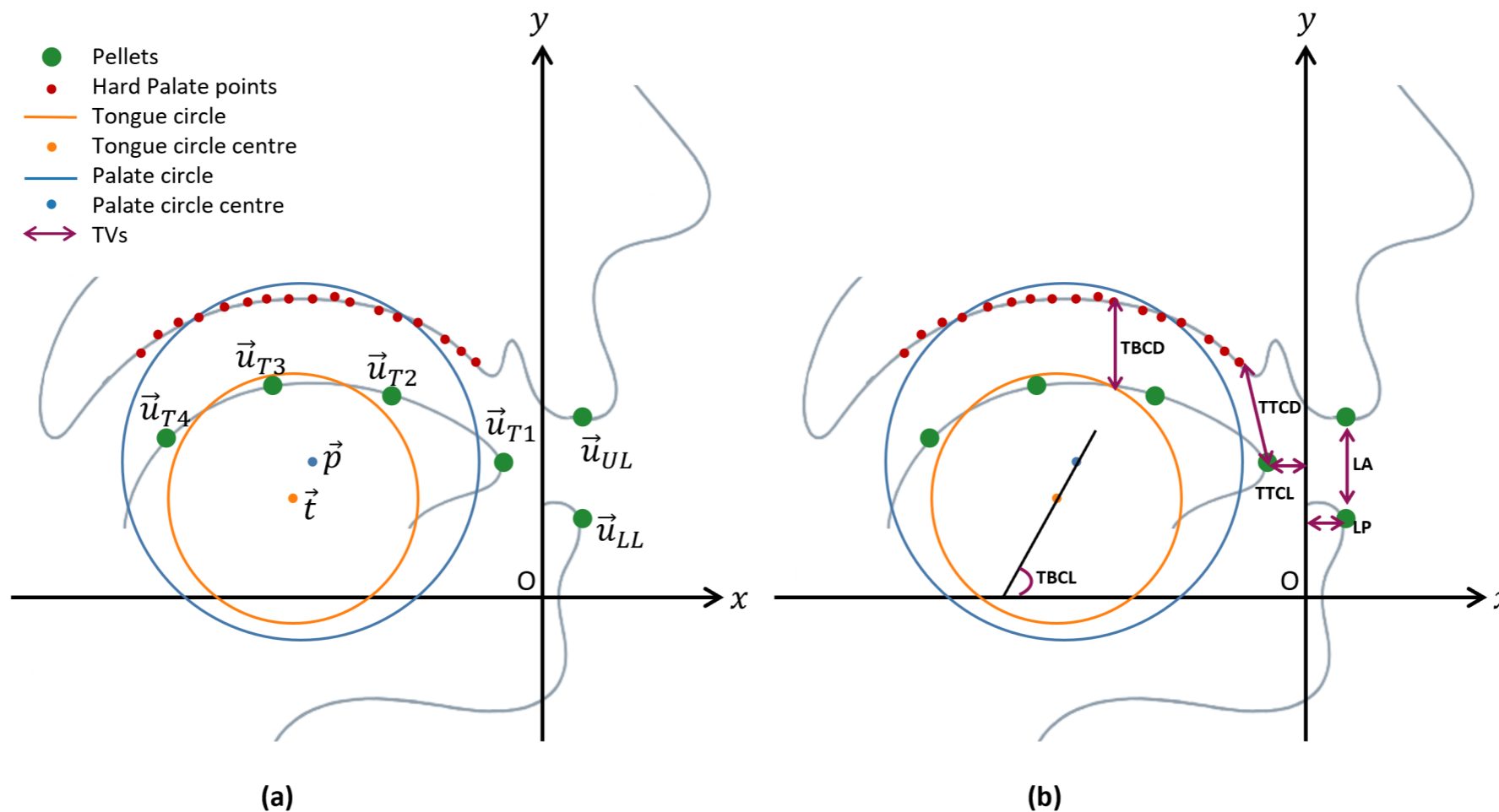


# XRMB Dataset

Speakers	Number	Median Age	Q1 Age	Q2 Age
Female	25	21.3	20.2	24.7
Male	21	20.8	20.0	22.4
Total	46	21.1	20.1	23.9

- The data were collected in the early nineties, before experiments of this type with X-Rays were considered too dangerous;
- The position of the pellets is sampled at 145 Hz, the speech signal is sampled at 22.05 kHz.

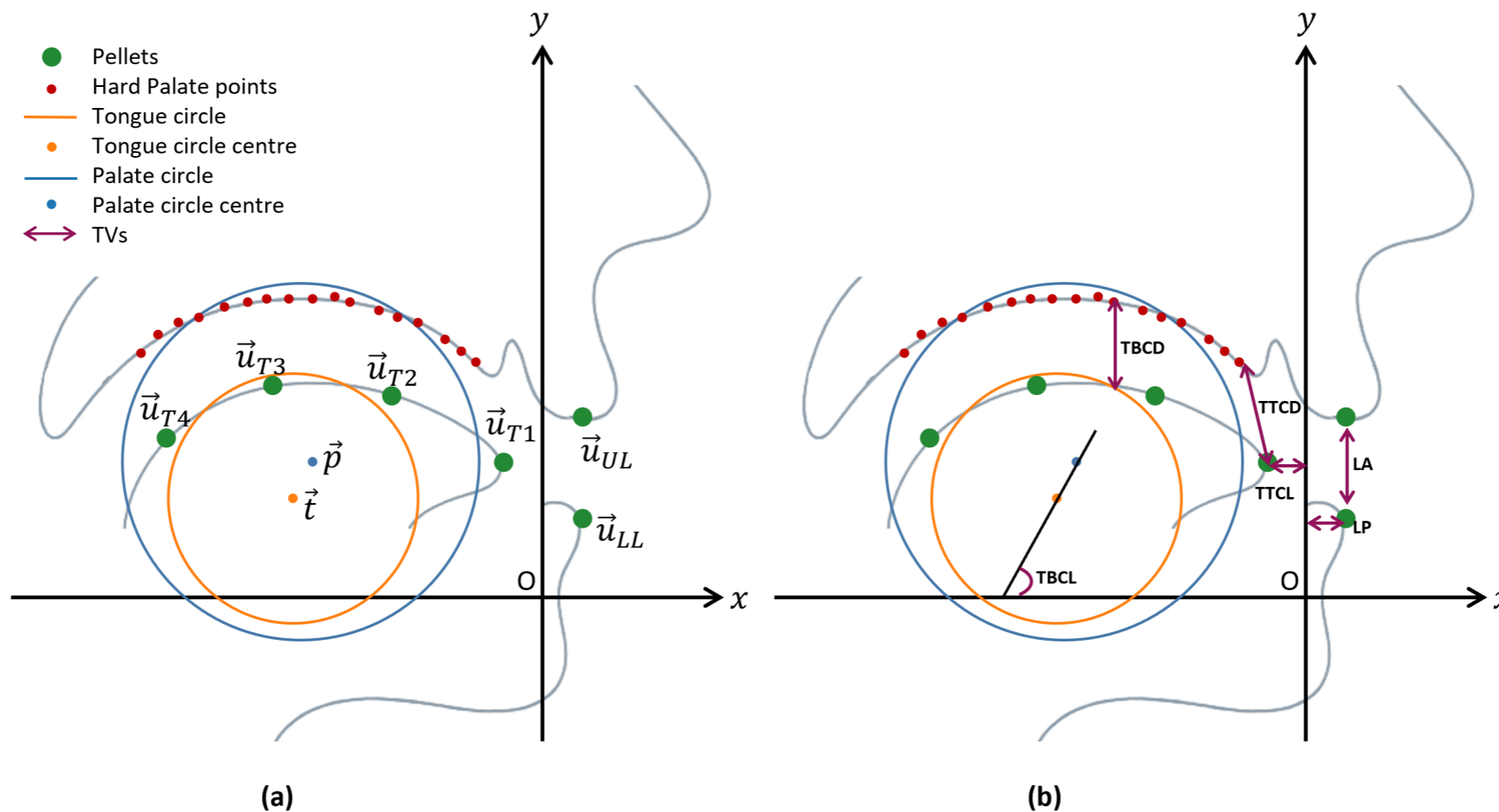
# Lip Aperture



- The Euclidean distance between upper and lower lip:

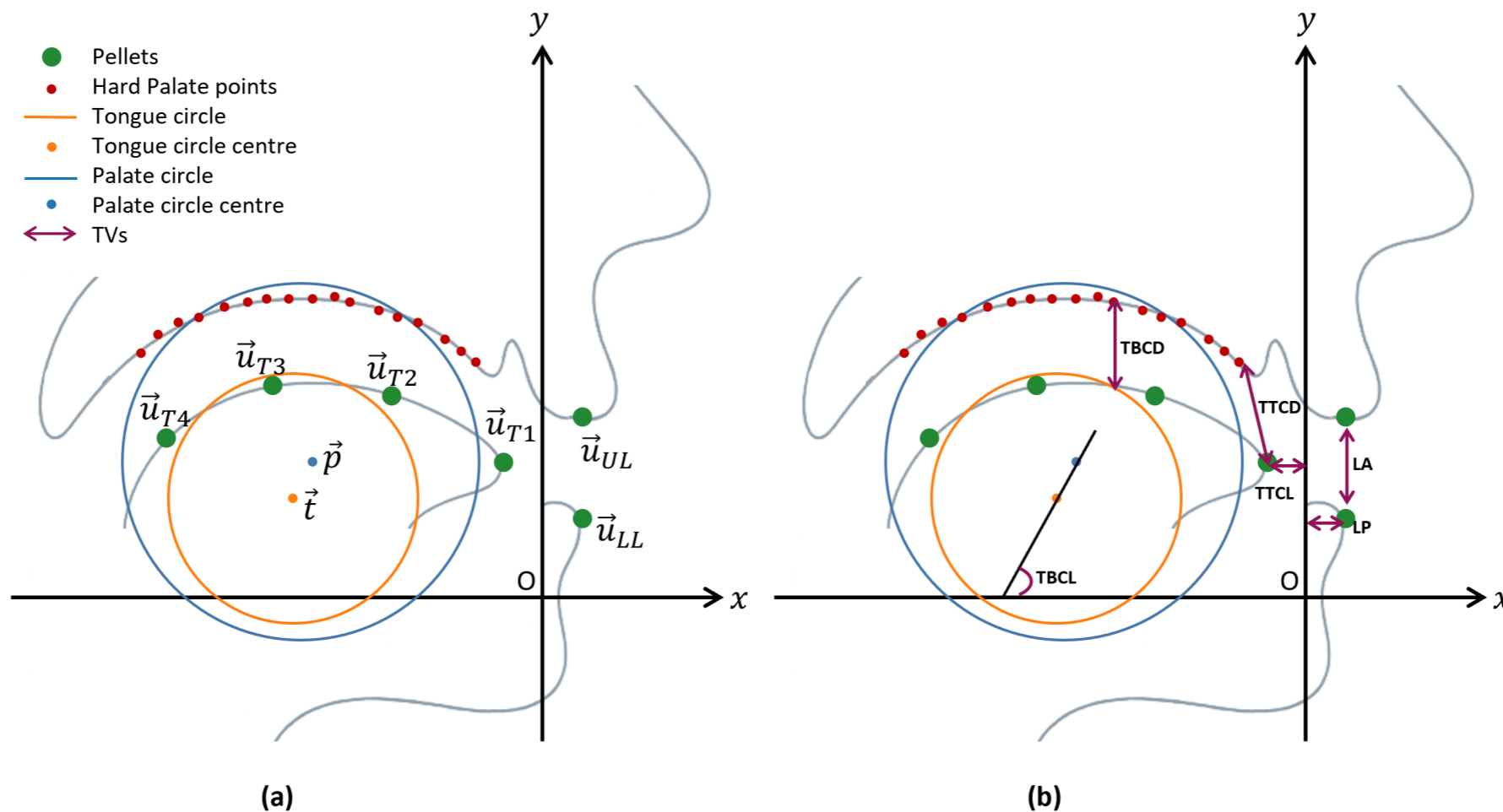
$$\|\vec{u}_{LL} - \vec{u}_{UL}\|_2.$$

# Lip Protrusion



- The horizontal distance of the LL pellet from the origin along the  $x$  axis:  $x_{LL}$ .

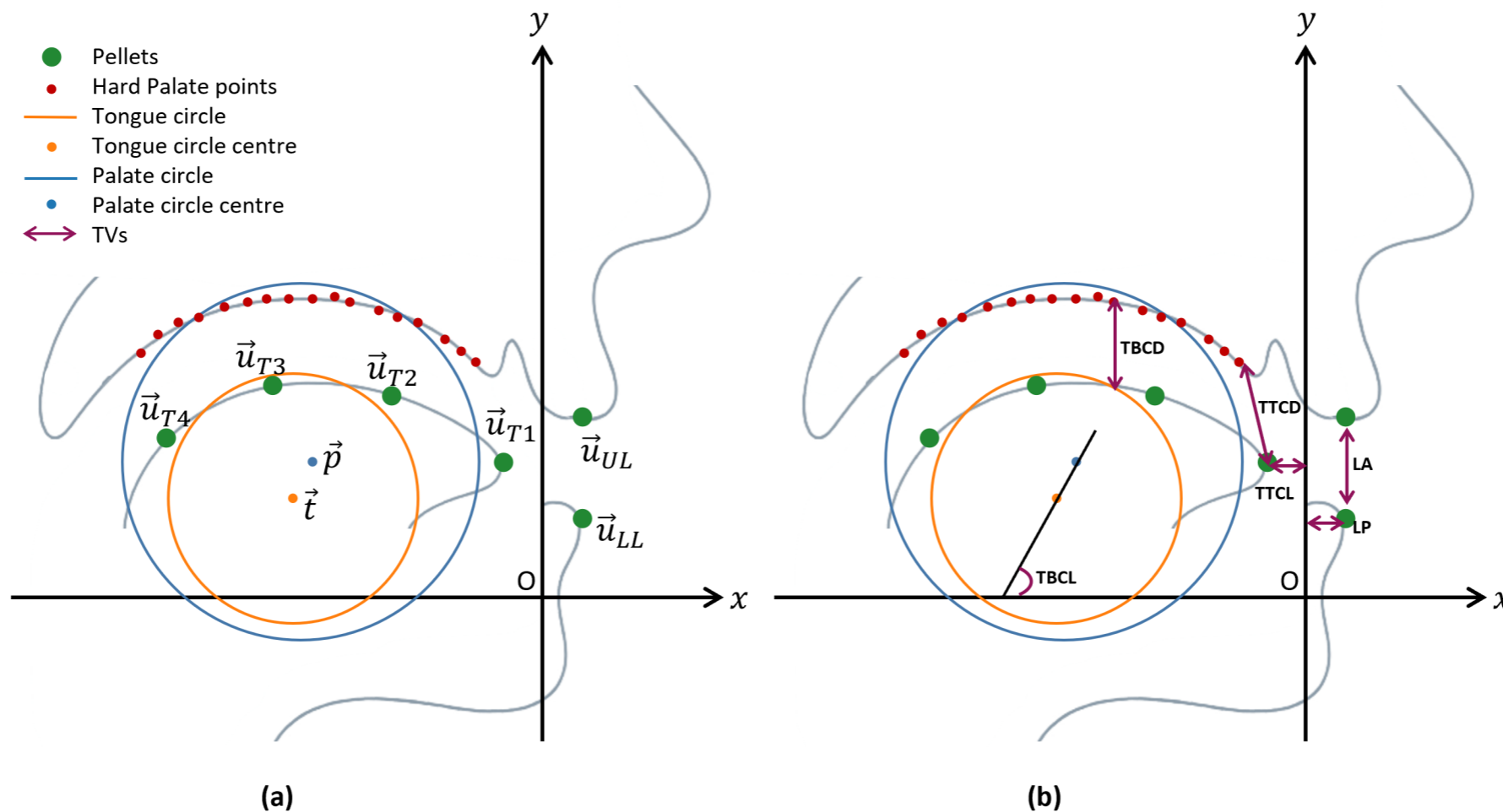
# Tongue Body Constriction Location



- The angle between the  $x$  axis and the line passing through the centres of tongue and palate:  

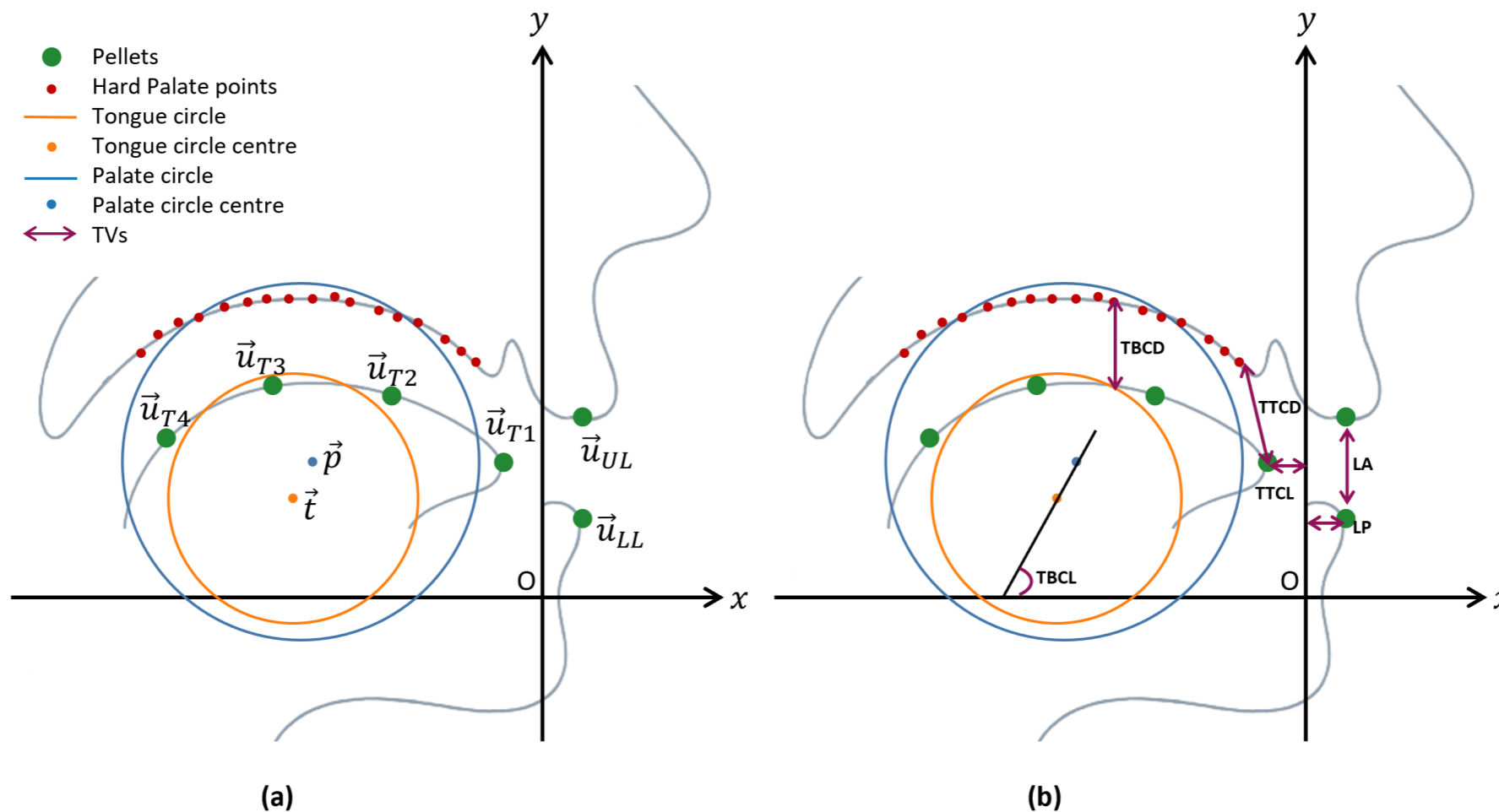
$$\arctan \left[ (y_T - y_P) / (x_T - x_P) \right].$$

# Tongue Body Constriction Degree



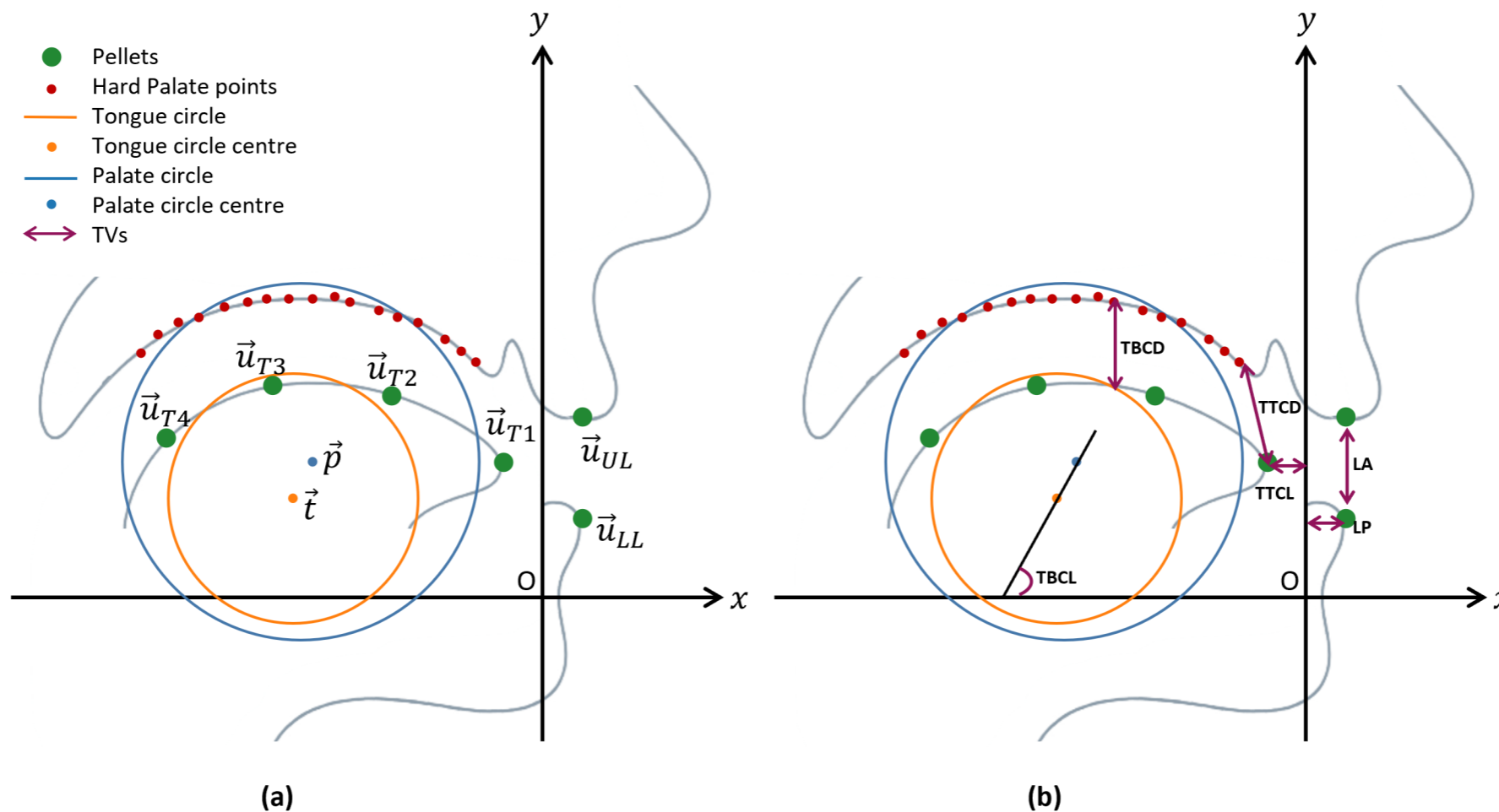
- Shortest distance between any point  $\vec{u}_T$  on Tongue Circle (between  $\vec{u}_{T2}$  and  $\vec{u}_{T4}$ ) and any point on Palate Circle:  $\min_{\vec{u}_P, \vec{u}_T} || \vec{u}_P - \vec{u}_T ||_2$ .

# Tongue Tip Constriction Location



- Horizontal distance between  $\vec{u}_{T1}$  and  $y$  axis:  $x_{T1}$ .

# Tongue Tip Constriction Degree

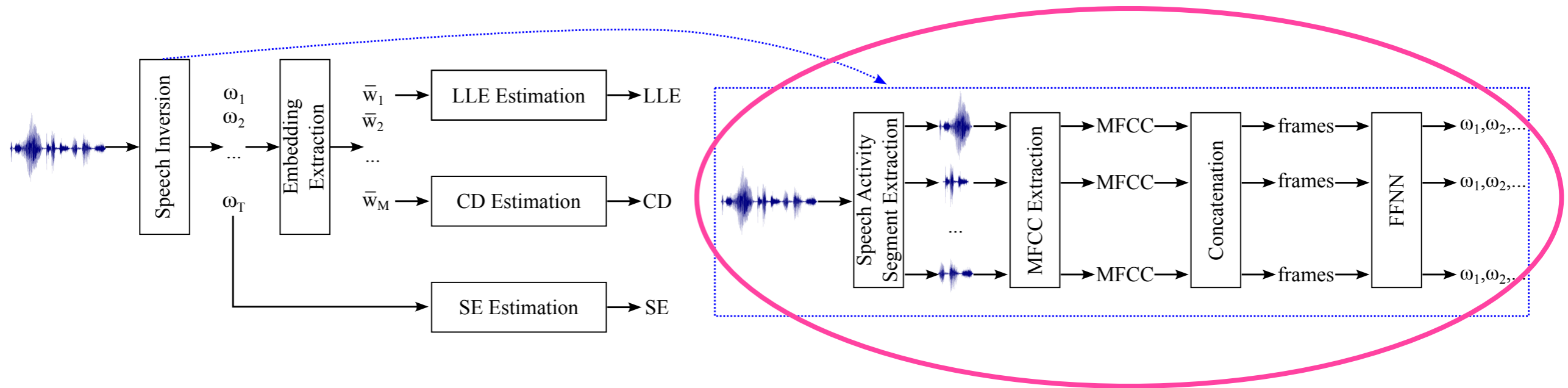


- Shortest distance between  $\vec{u}_{T1}$  and any point on Palate Circle:  $\min_{\vec{u}_P} ||\vec{u}_P - \vec{u}_{T1}||_2$ .

# Tract Variables

- The six measurements of the previous slides are referred to as **Tract Variables**;
- They account for the **configuration of the articulators** during the speech production process;
- The articulators are the **organs that shape the sounds** produced during speech emission;
- The speech production process is the **most complex motor process** performed by the human body;
- **Psychomotor retardation** is likely to impact negatively the speech production process.

# Speech Inversion



- Speech Inversion is the process aimed at automatically mapping speech signals into sequences of Tract Variables;
- The Tract Variables are extracted at regular time-steps of 10 ms from 30 ms long windows (sampling frequency of 100 Hz).

# Inversion Performance

	LA	LP	TBCL	TBCD	TTCL	TTCD	Avg.
Sivaraman 2019	0.856	0.613	<b>0.866</b>	0.745	0.707	0.907	0.782
Attia et al. 2023	0.868	0.590	0.742	0.780	0.597	0.893	0.745
Attia et al. 2024a	0.860	0.710	0.742	0.775	0.742	0.898	0.788
Attia et al. 2024b	<b>0.878</b>	<b>0.724</b>	0.743	0.809	<b>0.787</b>	<b>0.925</b>	0.810
This Work	0.855	0.714	0.816	<b>0.841</b>	0.773	0.897	<b>0.816</b>

- The speech inversion process performs at the level of the state of the art.

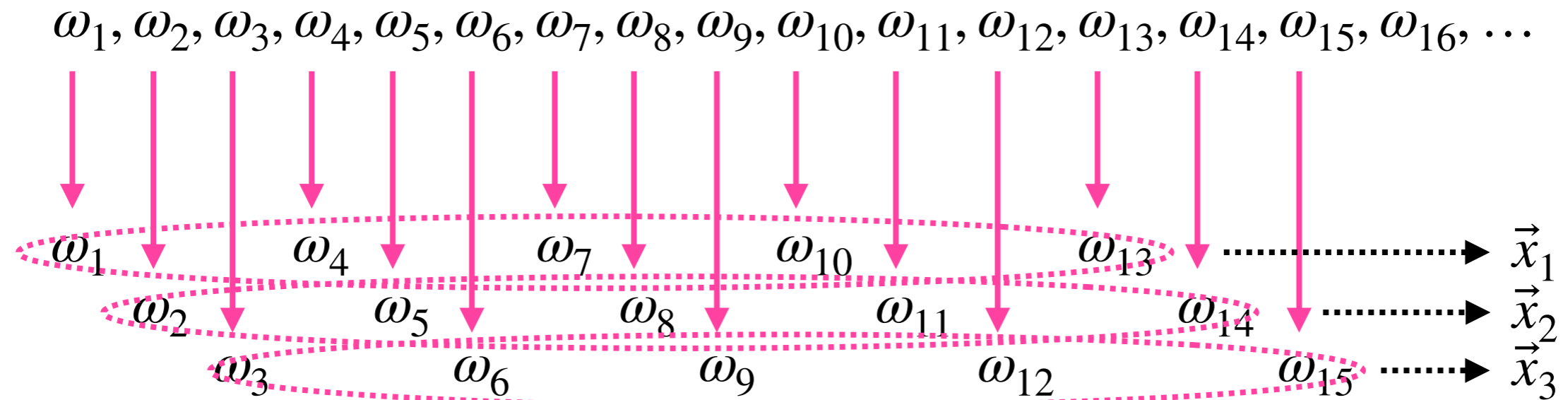
# Outline

- Social Signals
- Speech Articulation and Tract Variables
- Articulation as a Dynamic Process
- Conclusions

# Articulation as a Dynamic System

- The speech production apparatus can be thought of as a dynamical system, that is, a **physical system** that changes state over time;
- The articulators and, correspondingly, the **Tract Variables** can be considered as the **observables of the dynamical system**;
- The **Theory of Dynamical Systems** can be used as a means to model the speech production process and, possibly, to identify differences between depressed and non-depressed speakers.

# Embedding Extraction



- The “distance” between consecutive components in the embeddings is called the **delay**  $\tau$ ;
- According to the Taken’s Theorem, the embeddings  $\vec{x}_i$  are representations of the dynamic system’s state.

# Embedding Space

- The **space of the embeddings** corresponds to **all possible states** a dynamical system can take;
- A **trajectory** in the embedding space corresponds to the **evolution of the system** over time;
- The **properties** of the embeddings' **space** and of the embeddings' **trajectories** provide information about the **properties of the system**;
- The reason for expecting differences between depressed and non-depressed speakers is that the **pathology interferes with language processing in the brain.**

# The Data

Condition	Age	Female	Male	Low Education	High Education
Control	47.6 ± 12.6	42	12	19	33
Depressed	47.6 ± 12.0	37	18	23	31
Total	47.6 ± 12.2	79	30	42	64

- **No differences** between Control and Depressed (diagnosed by professional psychiatrists) in **Gender, Age and Education**;
- **Publicly available** at <https://github.com/androidscorpus/data>.



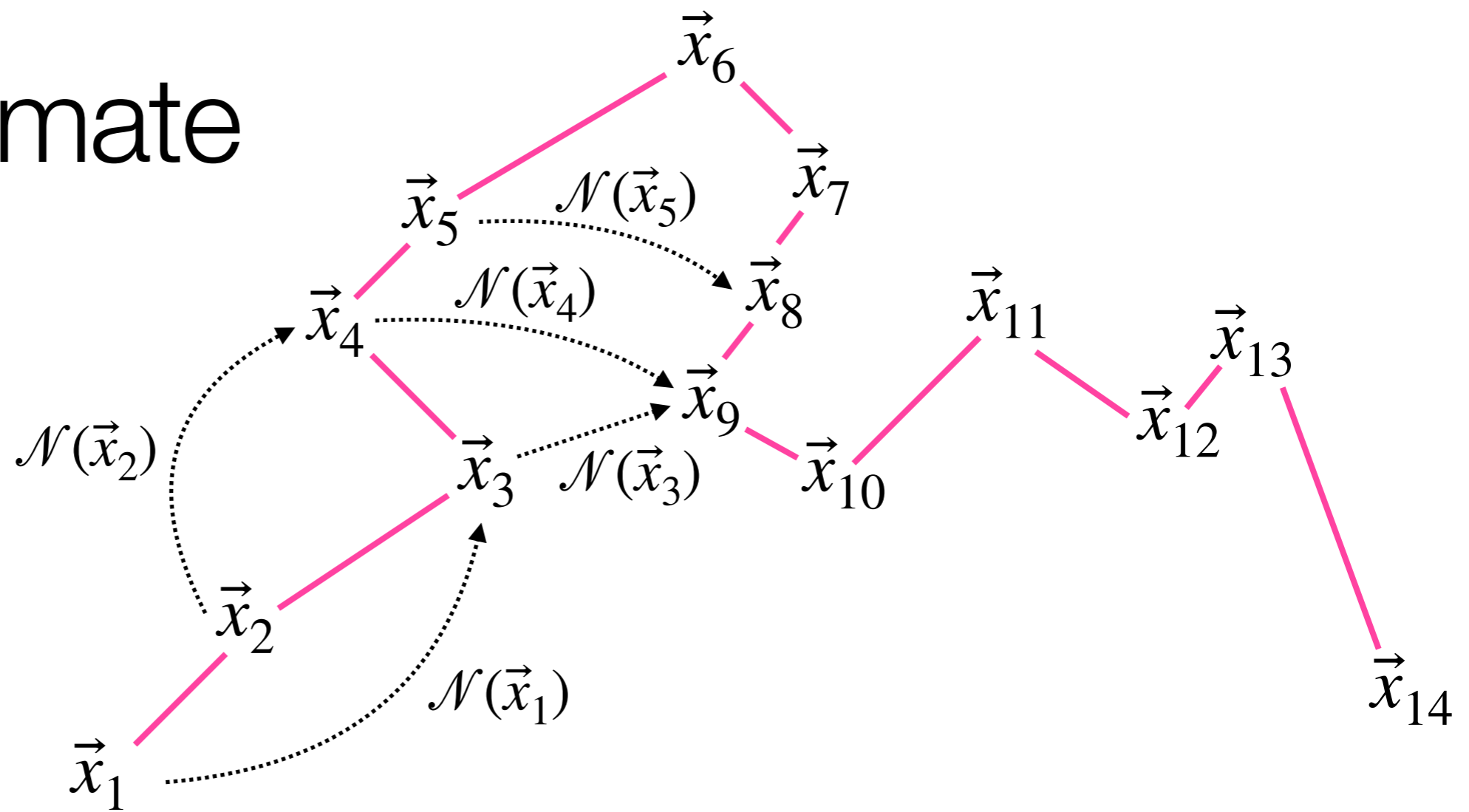
#1



#2

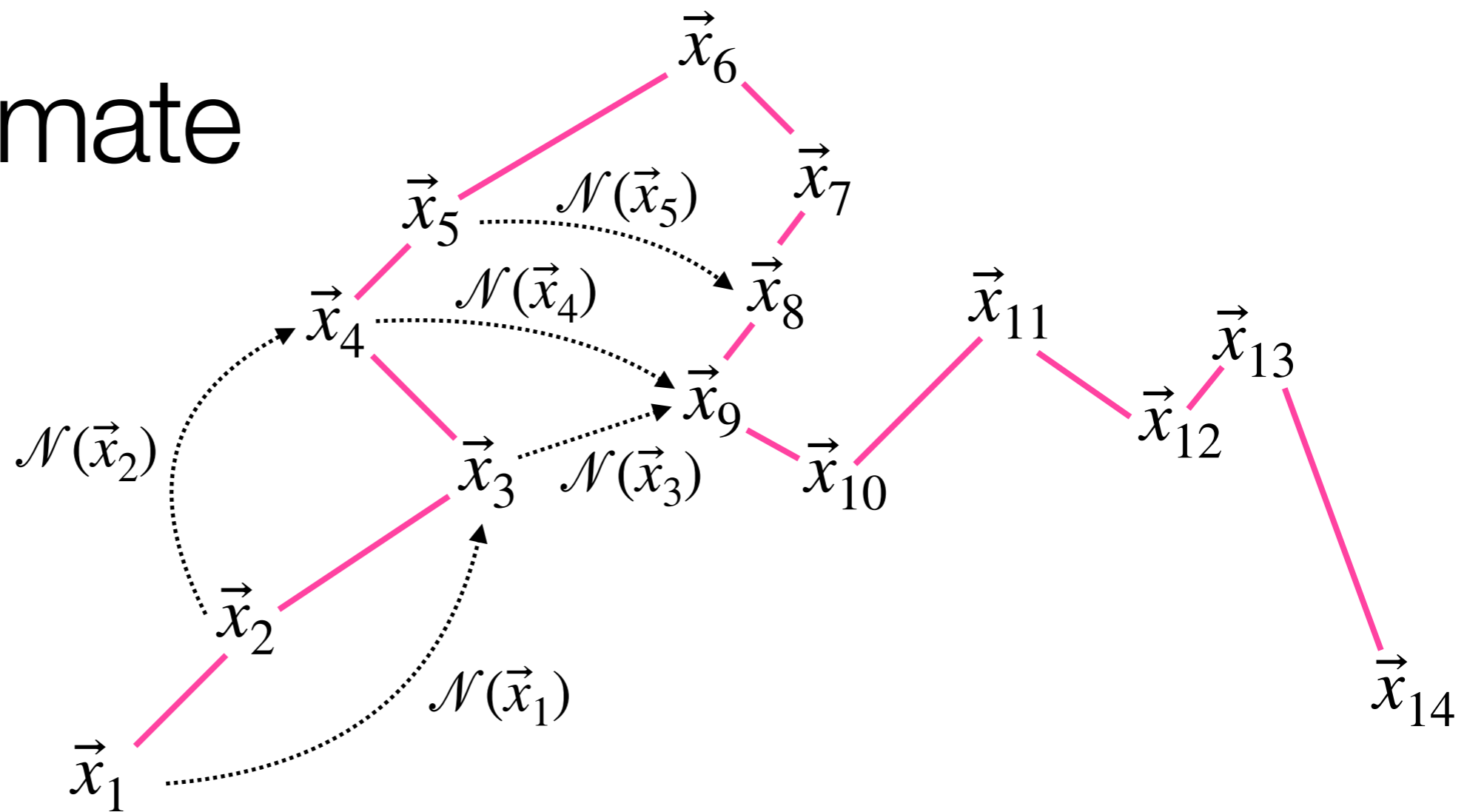
- All 109 participants were asked to **read the same text** (“The North Wind and the Sun” by Aesop);
- The goal is to **eliminate variance** resulting from differences in **spoken content**;
- The depression patient is **record #2**.

# LLE Estimate



- $\mathcal{N}(\vec{x}_j)$  is the **nearest embedding**;
- $d_j(0) = ||\vec{x}_j - \mathcal{N}(\vec{x}_j)||_2$  is the **distance** between embedding  $\vec{x}_j$  and its nearest neighbour (with  $\Delta t > 1/F$ , where  $F$  is the sampling frequency).

# LLE Estimate



- $d_j(i) = ||\vec{x}_{j+i} - \mathcal{N}(\vec{x}_{j+i})||_2$  accounts for how the distance  $d_j(0)$  evolves after  $i$  steps.
- Under the exponential divergence assumption,  $d_j(i) \simeq d_j(0)e^{\lambda_1 i \Delta t}$ , where  $\Delta t = 1/F$ .

# LLE Estimate

$$\log d_j(i) = \log d_j(0) + \lambda_1 i \Delta t$$

- The expression suggests that  $\log d_j(i)$  grows linearly with the time;
- The best fit of the points  $d_j(0), d_j(1), \dots, d_j(M - j)$  has the Largest Lyapounov Exponent as slope.

# LLE (Read Speech)

TV	Control	Depressed	P-value	Cliff's Delta
LA	0.021±0.005	0.023±0.006	0.104	0.18
LP	0.020±0.005	0.025±0.007	<b>&lt;0.001</b>	0.38
TTCL	0.022±0.006	0.026±0.007	<b>0.001</b>	0.34
TTCD	0.020±0.004	0.024±0.006	<b>0.002</b>	0.28
TBCL	0.022±0.006	0.029±0.008	<b>&lt;0.001</b>	0.41
TBCD	0.025±0.008	0.030±0.009	<b>&lt;0.001</b>	0.37
Lips	0.021±0.004	0.024±0.005	<b>0.003</b>	0.31
Tongue	0.022±0.004	0.027±0.006	<b>&lt;0.001</b>	0.51
All	0.022±0.003	0.026±0.005	<b>&lt;0.001</b>	0.50

- Statistically significant differences for most TVs and TV groups.

# LLE (Spontaneous Task)

TV	Control	Depressed	P-value	Cliff's Delta
LA	0.029±0.005	0.032±0.007	0.023	0.23
LP	0.032±0.006	0.033±0.006	0.234	0.09
TTCL	0.028±0.005	0.030±0.006	0.035	0.22
TTCD	0.028±0.005	0.032±0.008	0.013	0.26
TBCL	0.033±0.007	0.035±0.008	0.118	0.15
TBCD	0.034±0.005	0.036±0.007	0.045	0.19
Lips	0.021±0.004	0.024±0.005	<b>0.049</b>	0.19
Tongue	0.022±0.004	0.027±0.006	<b>0.030</b>	0.23
All	0.022±0.003	0.026±0.005	<b>0.024</b>	0.23

- Statistically significant differences only for TV groups, none for individual TVs.

# Cliff's Delta (I)

$$\delta = \frac{|n(x_i > y_j) - n(y_j > x_i)|}{N_x N_y}$$

- A set of values  $\mathcal{X} = \{x_1, \dots, x_{N_x}\}$  and a set of values  $\mathcal{Y} = \{y_1, \dots, y_{N_y}\}$ ;
- $n(x_i > y_j)$  is the number of times a value in  $\mathcal{X}$  is greater than a value in  $\mathcal{Y}$ .

# Cliff's Delta (II)

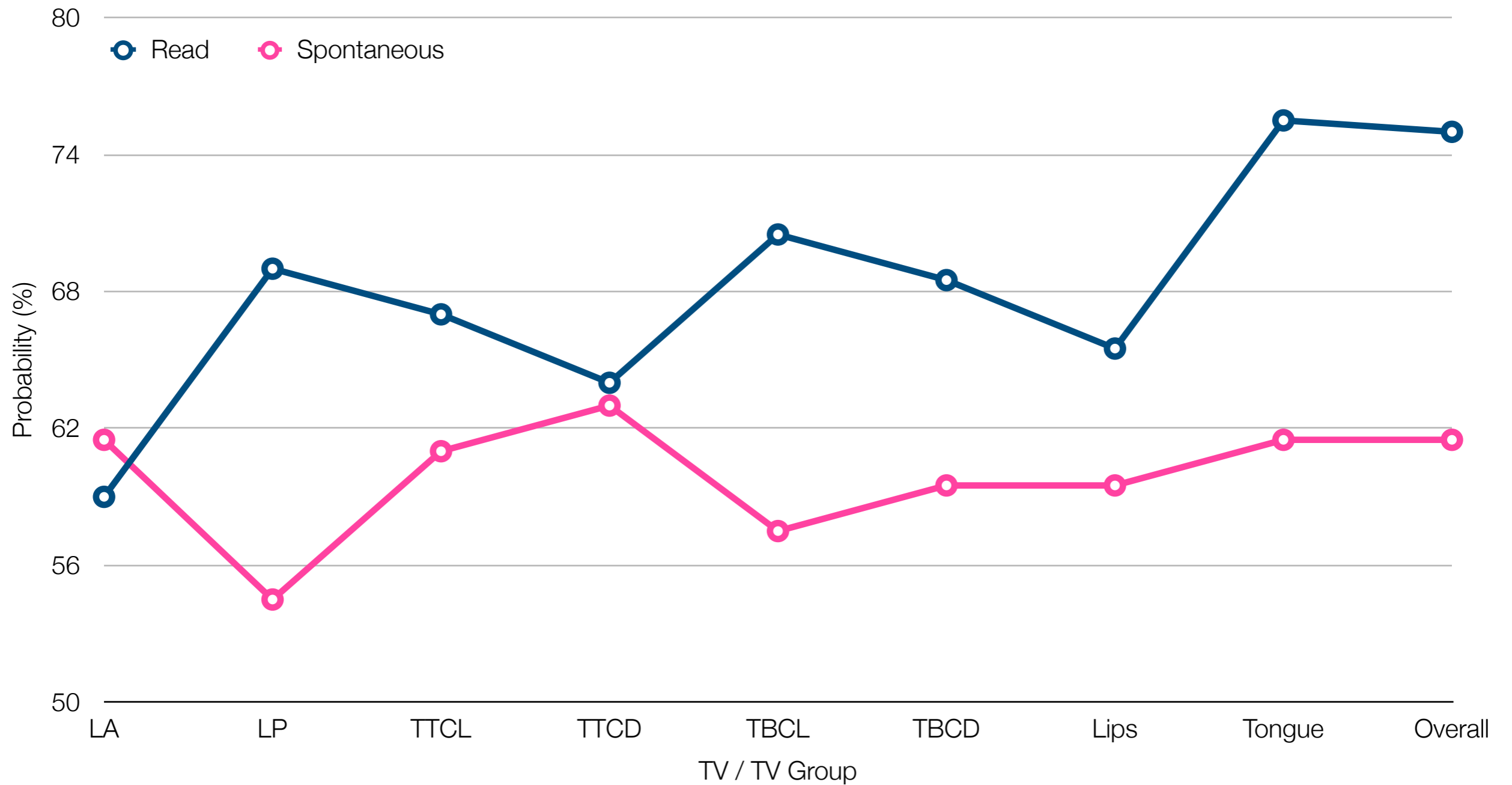
$$p(x_i > y_j) = \frac{\delta + 1}{2}$$

- When  $n(x_i > y_j) > n(y_j > x_i)$ .

$$p(y_j > x_i) = \frac{\delta + 1}{2}$$

- When  $n(x_i > y_j) < n(y_j > x_i)$ .

# Probability of Comparison Outcome



# LLE Differences

- The LLE difference is statistically **significant for all Tract Variables except LA** and for all groups of Tract Variables for **read speech**;
- For spontaneous speech, only groups of Tract Variables show statistically significant differences;
- It appears to be a **good biomarker** for read speech;
- The LLE accounts for the **predictability** and it appears to be higher for depressed speakers (**lower predictability**) in read speech;
- The literature confirms and shows that the feature correlation tends to be lower.

# Correlation Dimension

$$C(\rho) = \frac{2}{M(M-1)} \sum_{i=1}^M \sum_{j=i+1}^M \Theta \left( \rho - \|\vec{w}_i - \vec{w}_j\|_2 \right)$$

- The **intrinsic dimensionality**  $D$  of the embedding space is estimated through the fraction of pairs that lie within distance  $\rho$  from each other;
- The relationship  $\ln C(\rho) \propto D \ln \rho$  allows one to estimate the value of  $D$ ;
- Higher dimensionality, means that the dynamical system has more degrees of freedom.

# CD (Read Speech)

TV	Control	Depressed	P-value	Cliff's Delta
LA	3.66±0.32	3.47±0.31	<b>0.001</b>	0.33
LP	3.62±0.29	3.47±0.27	<b>0.008</b>	0.26
TTCL	3.42±0.33	3.34±0.37	<b>0.022</b>	0.22
TTCD	3.35±0.36	3.16±0.36	<b>0.004</b>	0.29
TBCL	3.16±0.33	2.94±0.31	<b>&lt;0.001</b>	0.37
TBCD	3.45±0.36	3.24±0.38	<b>0.001</b>	0.34
Lips	3.64±0.22	3.47±0.23	<b>&lt;0.001</b>	0.40
Tongue	3.35±0.23	3.17±0.25	<b>&lt;0.001</b>	0.40
All	3.46±0.17	3.28±0.21	<b>&lt;0.001</b>	0.49

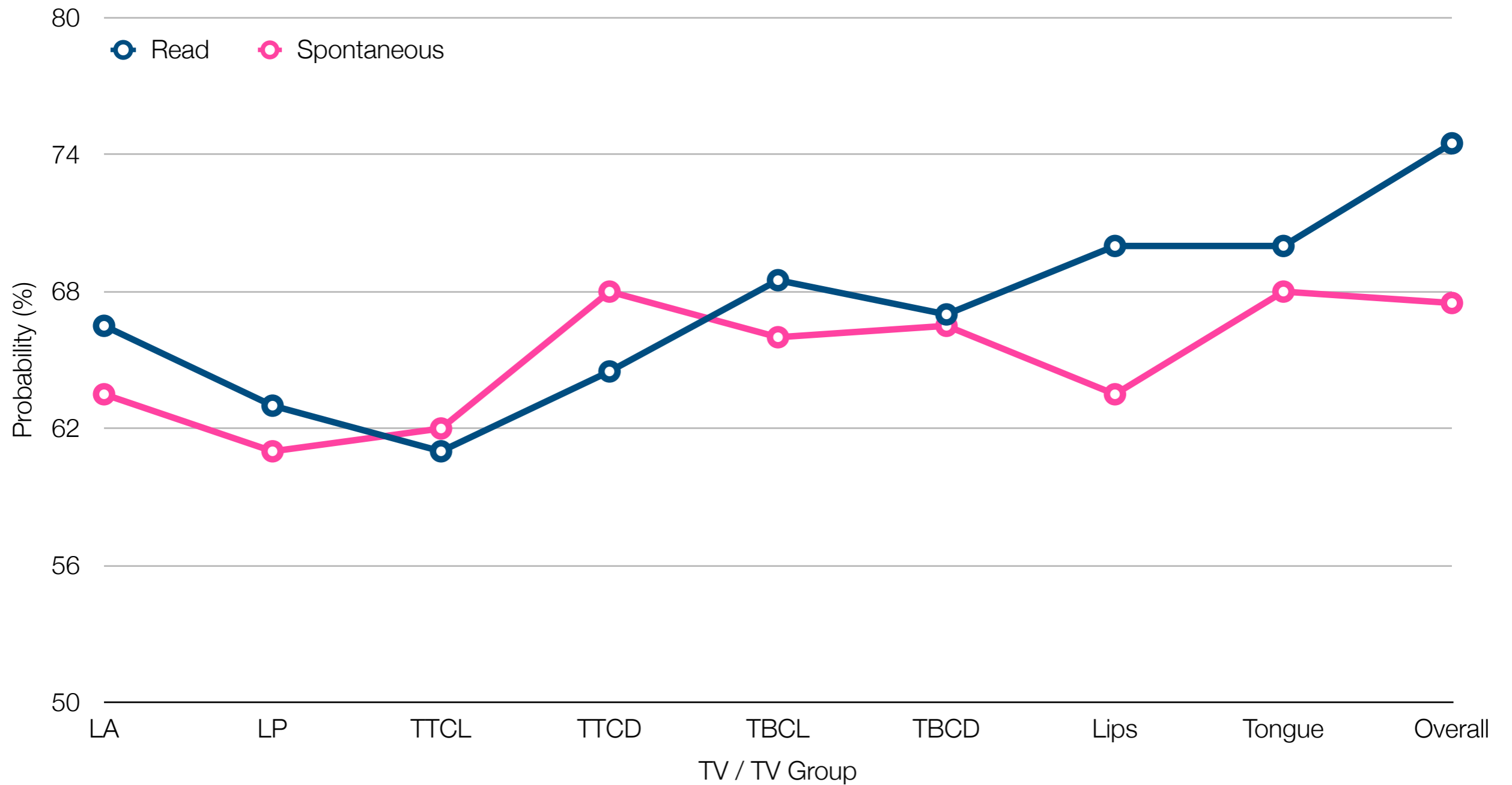
- Statistically significant differences for all TVs and TV groups.

# CD (Spontaneous Speech)

TV	Control	Depressed	P-value	Cliff's Delta
LA	3.19±0.23	3.10±0.28	<b>0.006</b>	0.27
LP	3.30±0.21	3.21±0.31	<b>0.020</b>	0.22
TTCL	3.20±0.24	3.06±0.36	<b>0.013</b>	0.24
TTCD	2.90±0.21	2.73±0.30	<b>&lt;0.001</b>	0.36
TBCL	2.63±0.18	2.50±0.28	<b>0.002</b>	0.32
TBCD	2.89±0.21	2.74±0.27	<b>0.001</b>	0.33
Lips	3.64±0.22	3.14±0.25	<b>0.007</b>	0.27
Tongue	3.35±0.23	2.77±0.25	<b>&lt;0.001</b>	0.36
All	3.46±0.17	2.91±0.21	<b>0.001</b>	0.35

- Statistically significant differences for all TVs and TV groups.

# Probability of Comparison Outcome

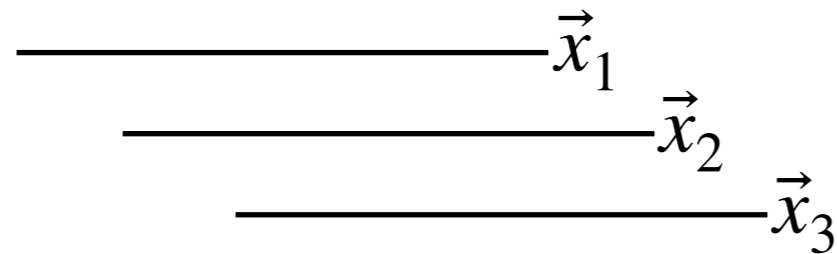


# CD Differences

- The CD difference is **statistically significant for all Tract Variables and for all groups of Tract Variables**;
- It appears to be a **good biomarker** for both read and spontaneous speech;
- The CD accounts for the **degrees of freedom** of the speech production process, it is the dimensionality of the attractor, the set of all possible states;
- The CD tends to be **higher for the control speakers**, suggesting higher complexity;
- This is confirmed by the literature (**depressed speakers tend to show lower variability**).

# Sample Entropy

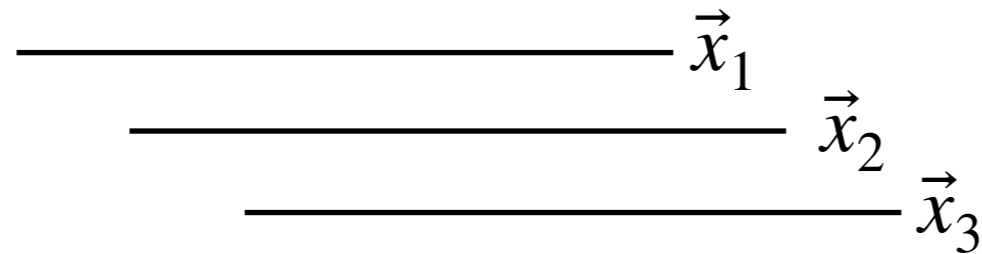
$\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{11}, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}, \omega_{16}, \dots$



- It is possible to count the number  $B$  of pairs  $(\vec{x}_i, \vec{x}_j)$  such that  $dist(\vec{x}_i, \vec{x}_j) < r$ , where  $dist(\cdot)$  is a distance function;
- In this work,  $dist(\vec{x}_i, \vec{x}_j) = \max_{k=[1,D]} |x_{ik} - x_{jk}|$ , where  $D$  is the dimensionality of the  $\vec{x}_i$ .

# Sample Entropy

$\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{11}, \omega_{12}, \omega_{13}, \omega_{14}, \omega_{15}, \omega_{16}, \dots$



- It is possible to count the number  $B$  of pairs  $(\vec{x}_i, \vec{x}_j)$  such that  $dist(\vec{x}_i, \vec{x}_j) < r$ , where  $dist(\cdot)$  is a distance function;
- The dimensionality of the vectors has been increased from  $D$  to  $D + 1$ .

# Sample Entropy

$$SE(D, r, K) = -\ln \left( \frac{A}{B} \right)$$

- The parameter  $K$  is the total number of Tract Variables extracted from a recording;
- The SE does not measure the properties of the dynamical system or, at best, it does it only when  $\tau = 1$ .

# SE (Read Speech)

TV	Control	Depressed	P-value	Cliff's Delta
LA	0.26±0.06	0.23±0.06	<b>0.001</b>	0.33
LP	0.22±0.04	0.19±0.04	<b>&lt;0.001</b>	0.39
TTCL	0.24±0.06	0.23±0.05	0.202	0.09
TTCD	0.26±0.05	0.25±0.06	0.119	0.13
TBCL	0.18±0.04	0.18±0.05	0.487	0.00
TBCD	0.16±0.04	0.14±0.04	<b>&lt;0.001</b>	0.30
Lips	0.24±0.04	0.21±0.05	<b>0.002</b>	0.42
Tongue	0.21±0.03	0.20±0.03	<b>0.024</b>	0.22
All	0.22±0.03	0.21±0.03	<b>&lt;0.001</b>	0.37

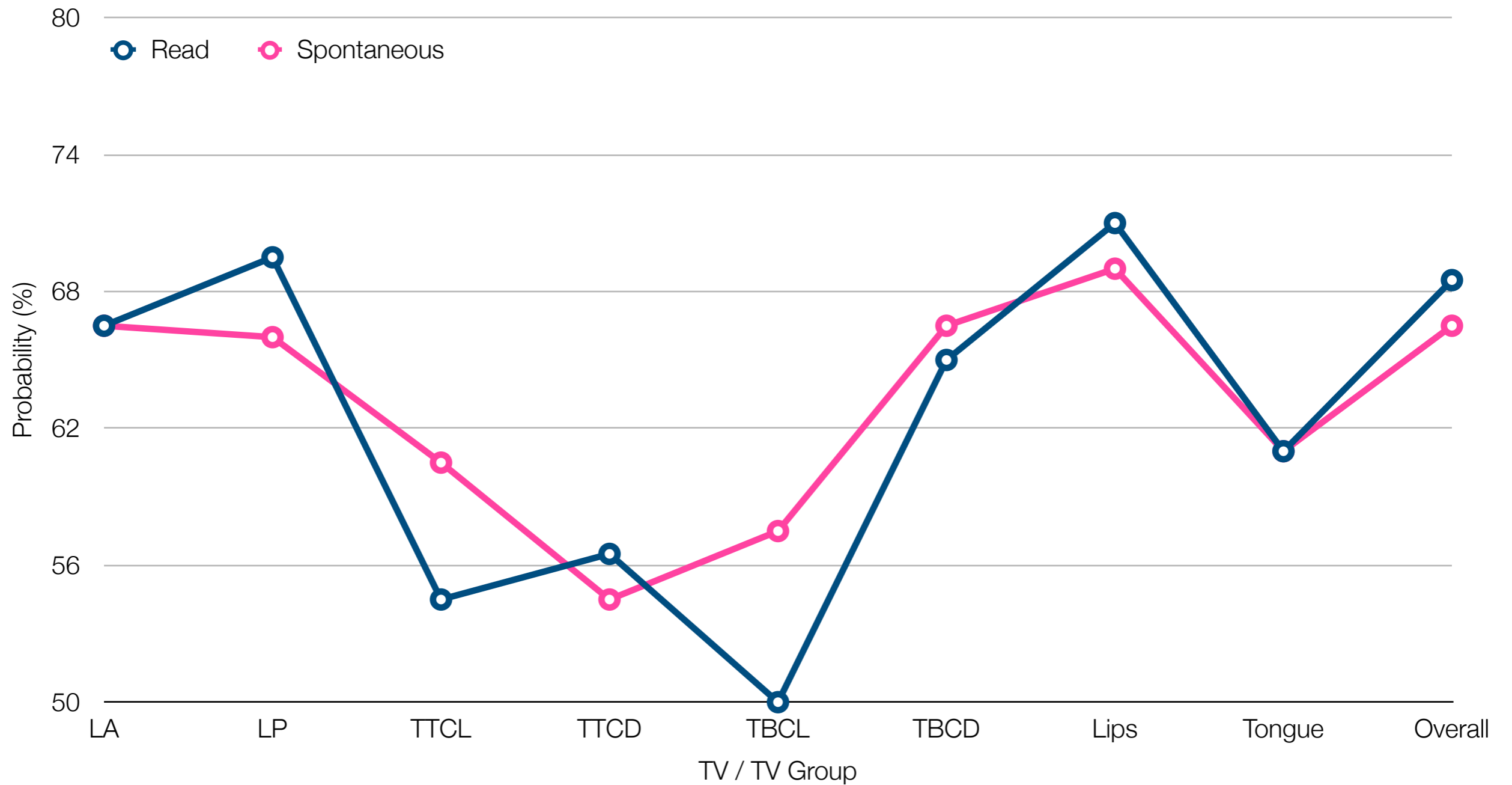
- Statistically significant differences for all TV groups, but only some of the individual TVs.

# SE (Spontaneous Speech)

TV	Control	Depressed	P-value	Cliff's Delta
LA	0.22±0.03	0.20±0.03	<b>0.001</b>	0.33
LP	0.19±0.03	0.17±0.03	<b>0.002</b>	0.32
TTCL	0.24±0.03	0.23±0.04	<b>0.024</b>	0.21
TTCD	0.24±0.03	0.23±0.04	0.213	0.09
TBCL	0.13±0.02	0.12±0.04	0.087	0.15
TBCD	0.13±0.02	0.11±0.02	<b>0.001</b>	0.33
Lips	0.20±0.03	0.18±0.03	<b>&lt;0.001</b>	0.38
Tongue	0.20±0.02	0.19±0.03	<b>0.022</b>	0.22
All	0.20±0.02	0.21±0.02	<b>0.001</b>	0.33

- Statistically significant differences for all TV groups and most individual TVs.

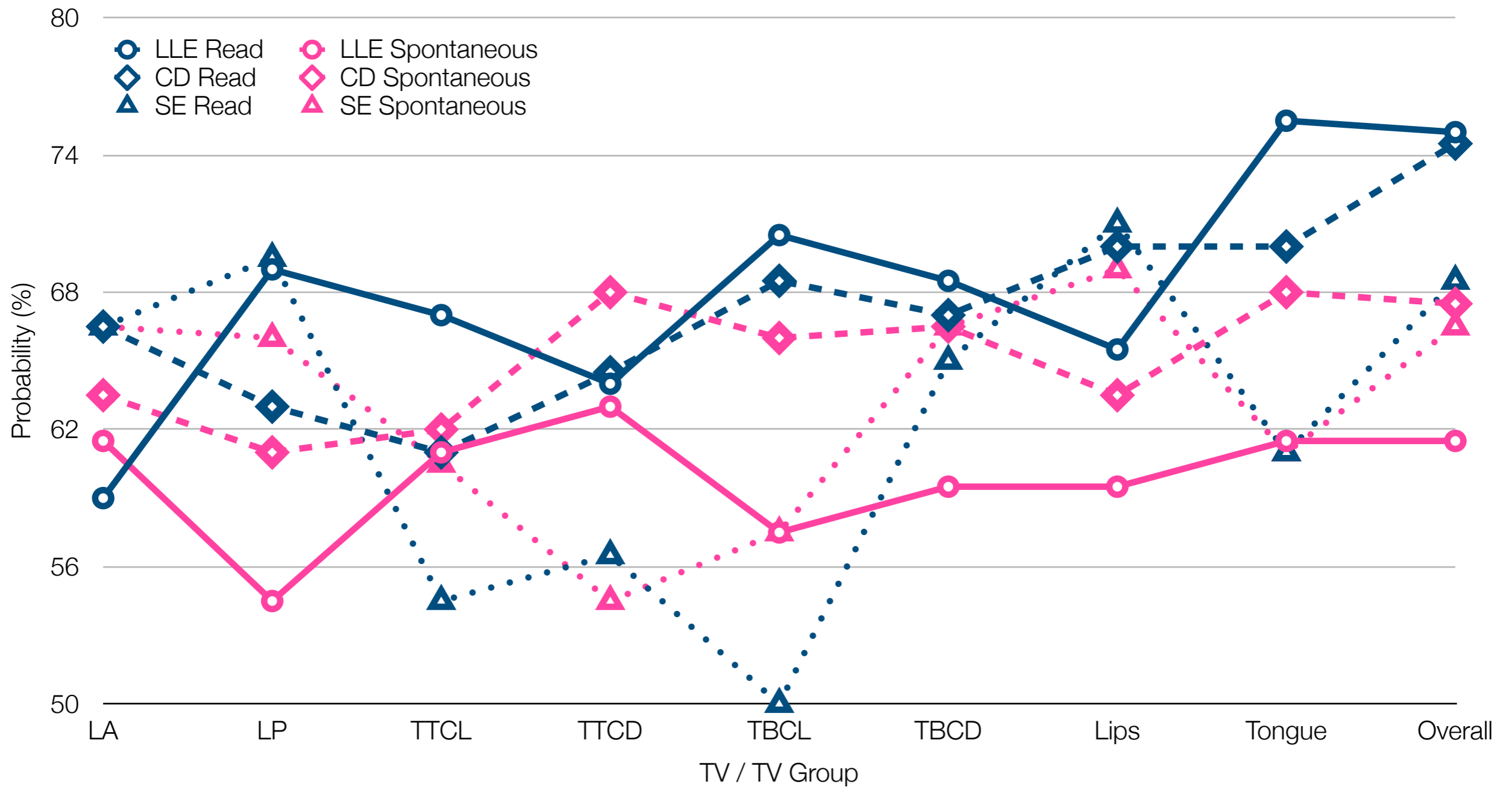
# Probability of Comparison Outcome



# SE Differences

- There are SE differences **statistically significant for several Tract Variables** and for all groups of Tract Variables;
- It appears to be a **good biomarker** for both read and spontaneous speech;
- The SE accounts for the **repetitiveness** of the speech production process, it measures how many new patterns appear;
- The SE tends to be **lower for the depressed speakers**, suggesting higher repetitiveness.

# Probability of Comparison Outcome



# Outline

- Social Signals
- Speech Articulation and Tract Variables
- Articulation as a Dynamic Process
- Conclusions

# Conclusions

- LLE, CD and SE can be used as depression biomarkers, that is, they can be used to discriminate between depressed and non-depressed speakers;
- Depressed speech appears to be less predictable (greater LLE), more constrained (lower CD) and more repetitive (lower SE);
- These effects explain a large number of observations presented in the literature;
- The biomarkers help clinicians to check the consistency between their judgment and measurable aspects of patients' behaviour.

# Thank You!