MyScript®

CogniTalk

{EPITECH.}

L'ECOLE DE L'INNOVATION ET DE
L'EXPERTISE INFORMATIQUE

ML WEEK

THE MACHINE LEARNING WEEK

Thesis supervisor: *Jean-François Bonastre*

Co-supervisors: *Richard Dufour* and *Nicolas Obin*

The character dimension for the representation of acted voices

Seminar in Nantes Machine Learning Meetup

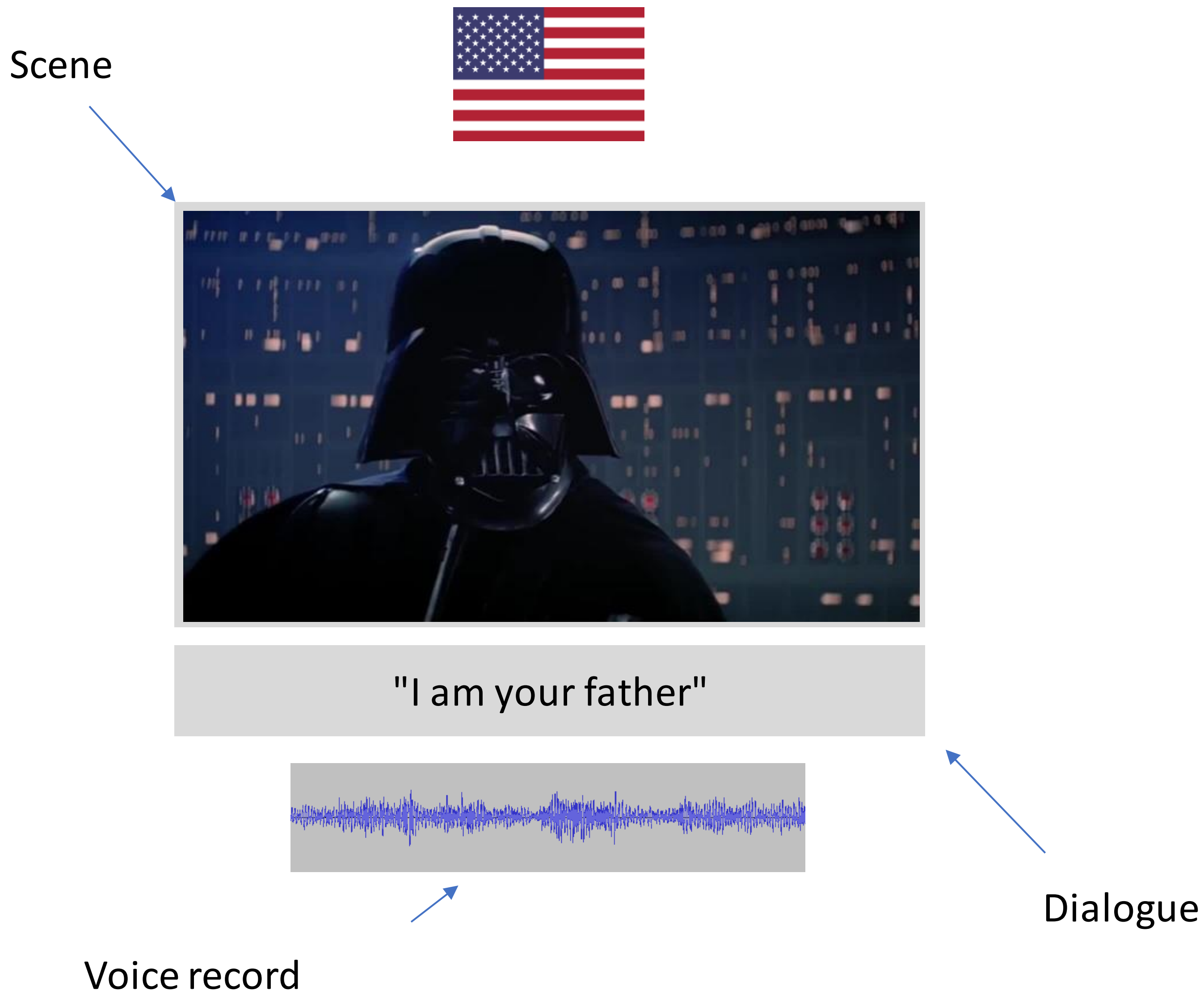
AVIGNON UNIVERSITÉ

LABORATOIRE INFORMATIQUE D'AVIGNON

Mathias Quillot - 2019

Voice Dubbing

Replace the original voice by an other one in a different language/culture

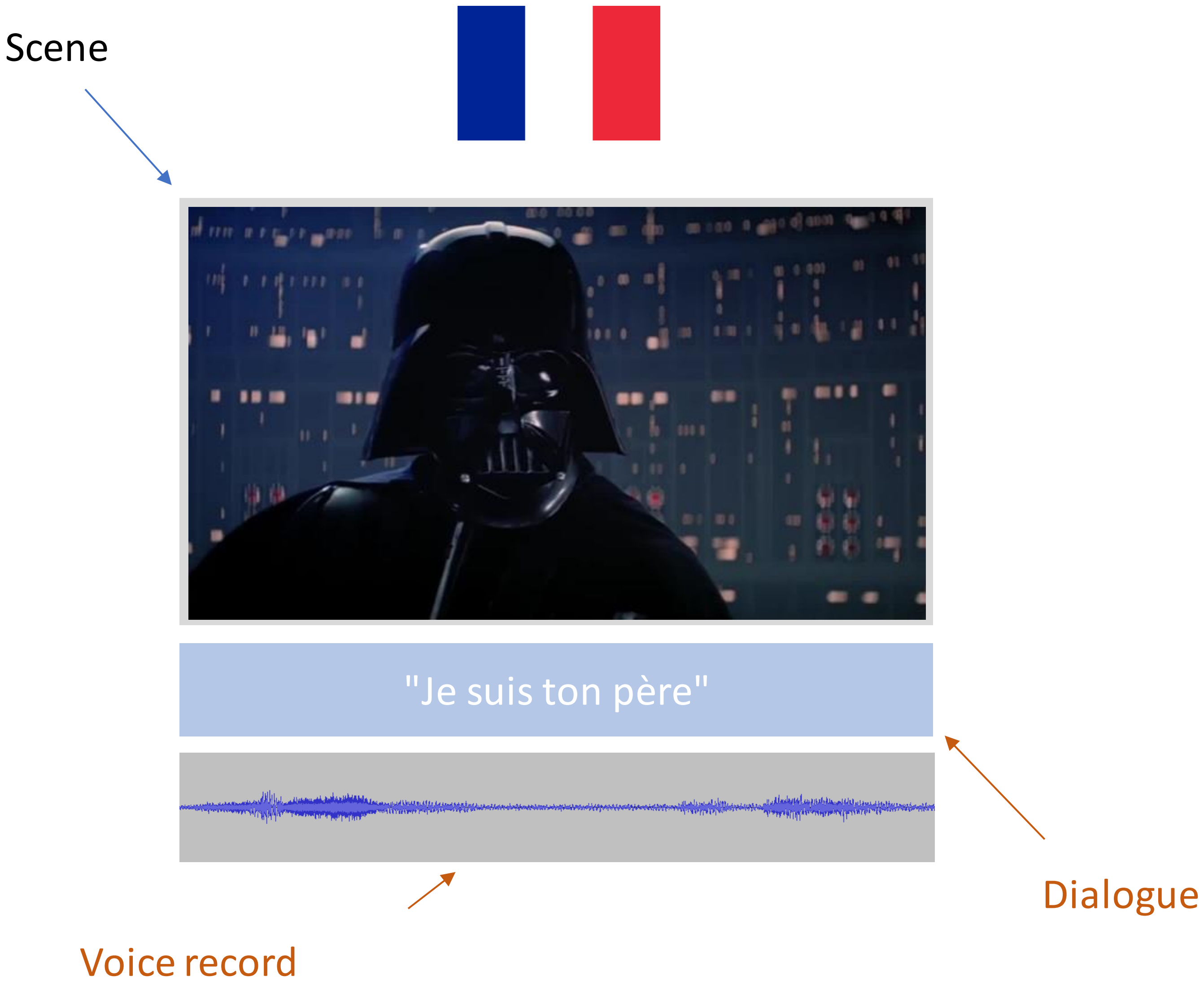


Scene

"I am your father"

Dialogue

Voice record



Scene

"Je suis ton père"

Voice record

Dialogue

Voice Casting

Select the voice that will replace the original one

Voice Casting

STEP OF DUBBING

CHOICE MADE BY ARTISTIC DIRECTOR (AD)

Artistic Director often choose the same performer

Performers become more expensive

Performers become less available

Difficult to find new talents

Create automatic tools can help DA

the Voice

DUBBING BROTHERS

AGENCE NATIONALE DE LA RECHERCHE
ANR

ircam
Centre
Pompidou

LABORATOIRE
INFORMATIQUE D'AVIGNON

ANR Project The Voice

Voice casting tools

Voice recommandation system

Voice Casting – Based on history

Original Character

French Performer

Character History

Voice Casting – Based on character play

Original Character

French Performer

The way he plays the original character

Voice choosing

Artistic Director

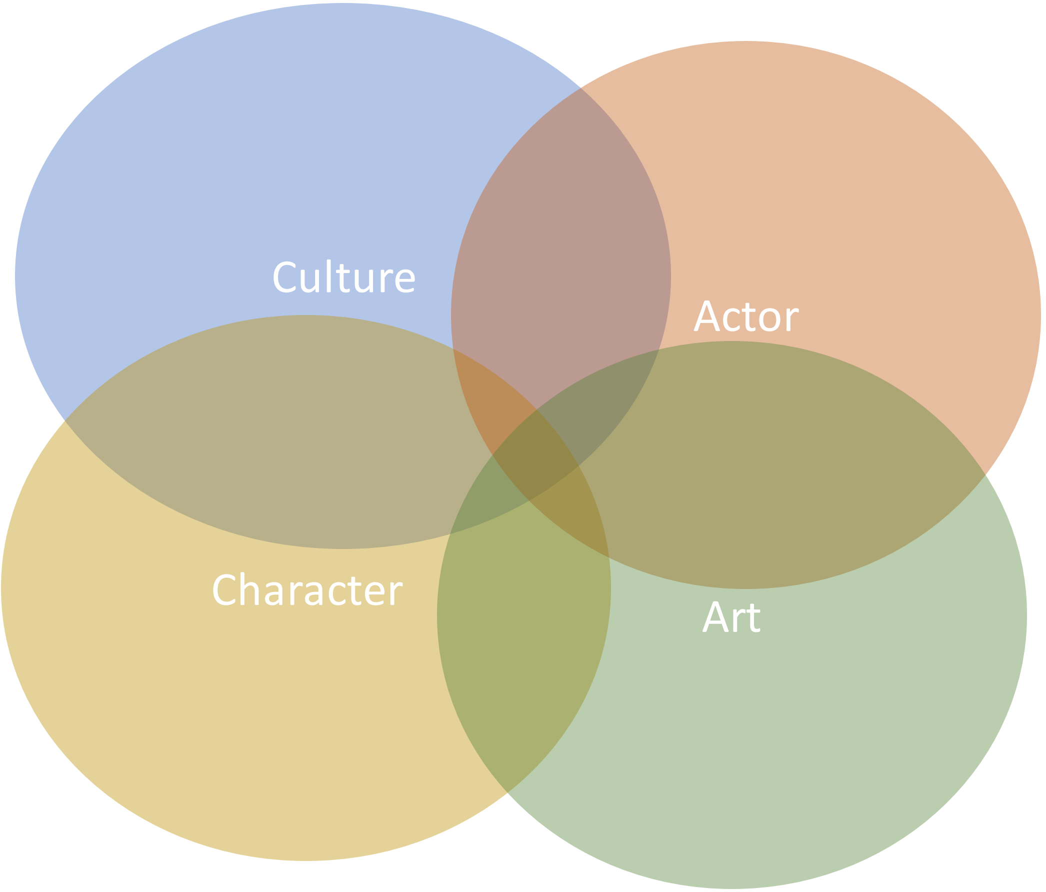
Does the vocal french performer match the original character?

Original Character

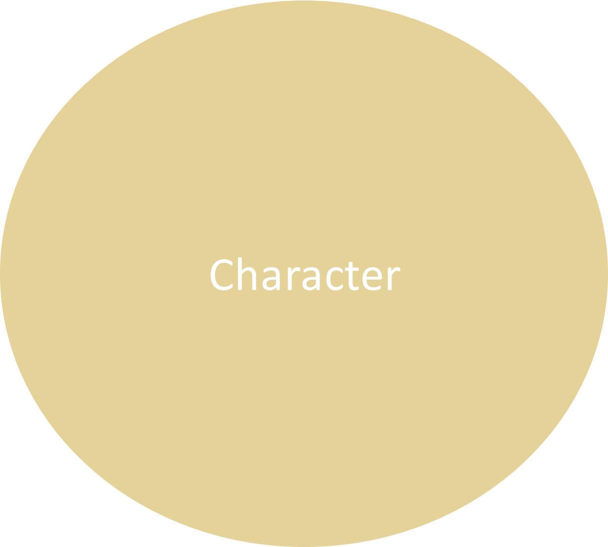
French Performer

Operator's choice doesn't simply involve applying an acoustic similarity

What factors are involved?



Focus on **character dimension**

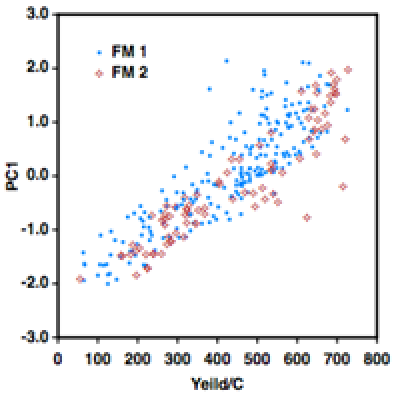

Character

What characterizes the character in the signal?

Summary of my experiments

Signal level

1



Representation level

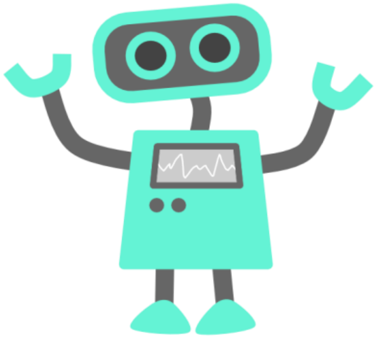
2

Expert comparison

Human level

3

Let me explain my decisions



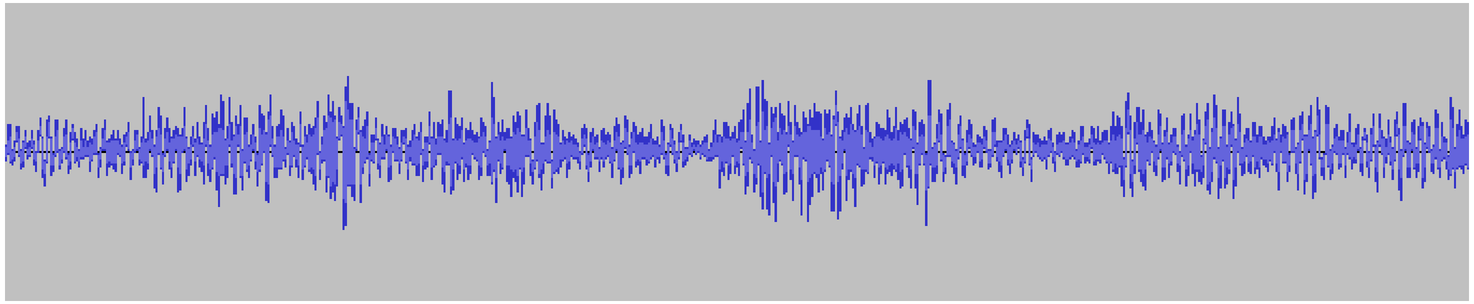
Explainability

4

Are acoustic signs of the character dimension present in the acted voice?

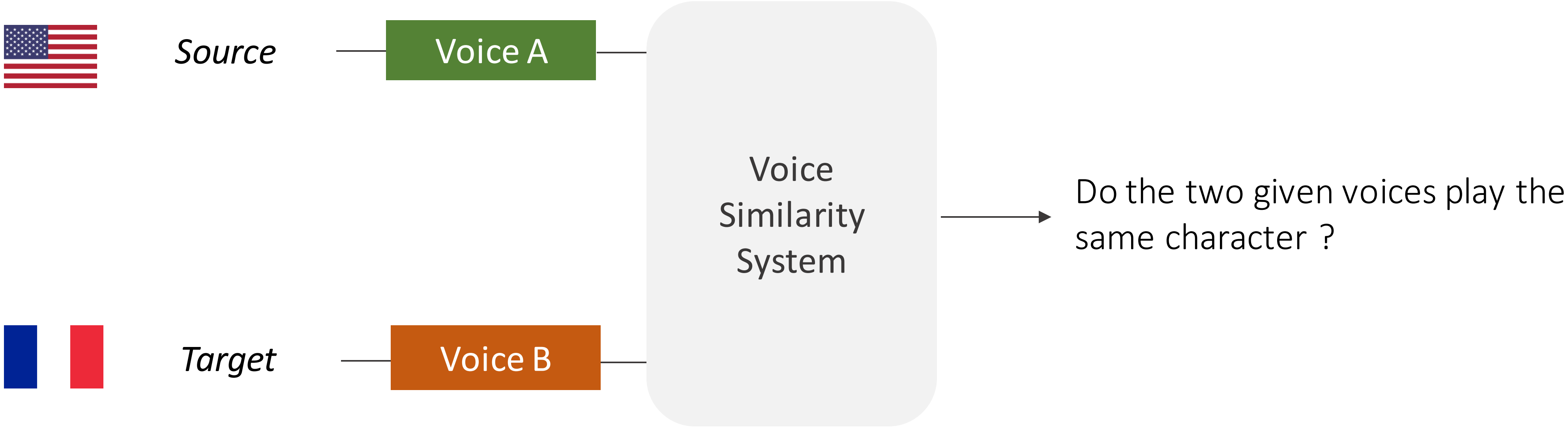
1

Signal Level



To confirm the dimension is present in the signal

Voice Similarity for character dimension



Source

Voice A

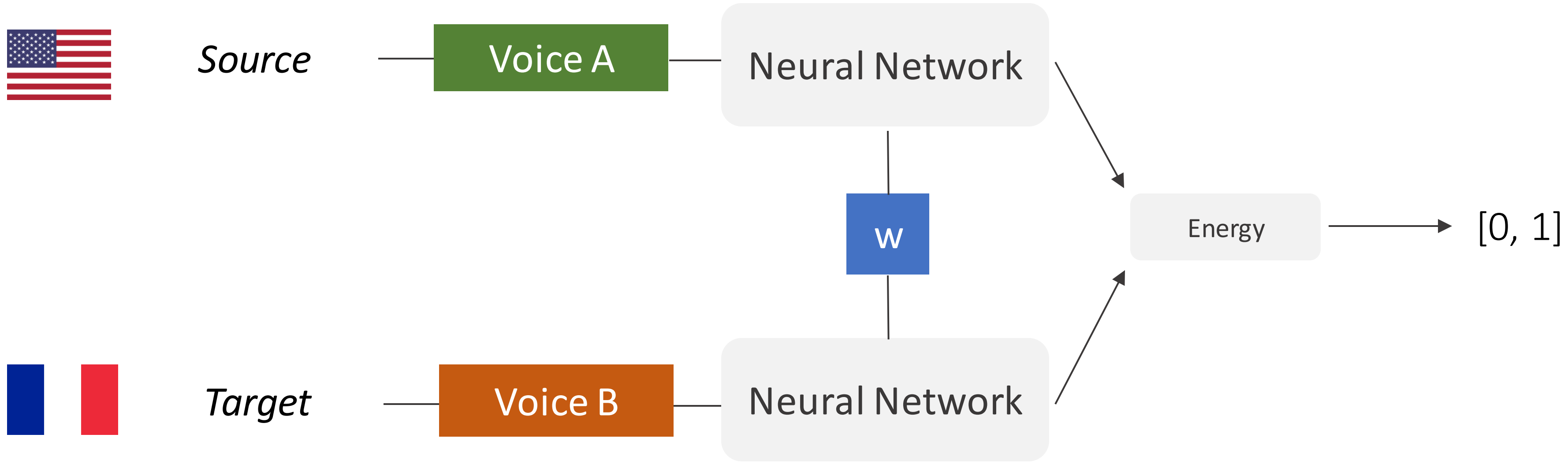
Target

Voice B

Voice Similarity System

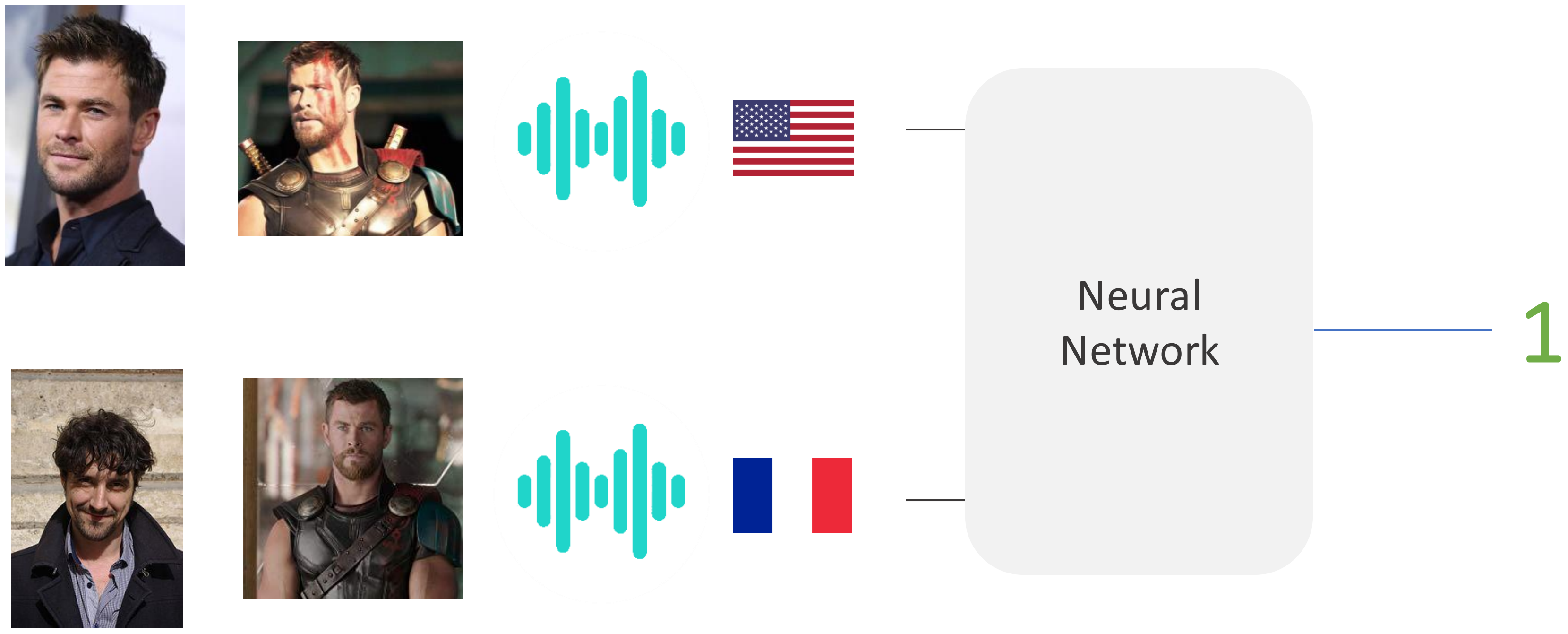
Do the two given voices play the same character ?

Voice Similarity Artchitecture: Siamese Neural Network



SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING, A. Gresse, M. Quillot, R. Dufour, V. Labatut, JF. Bonastre

Voice Similarity: Training



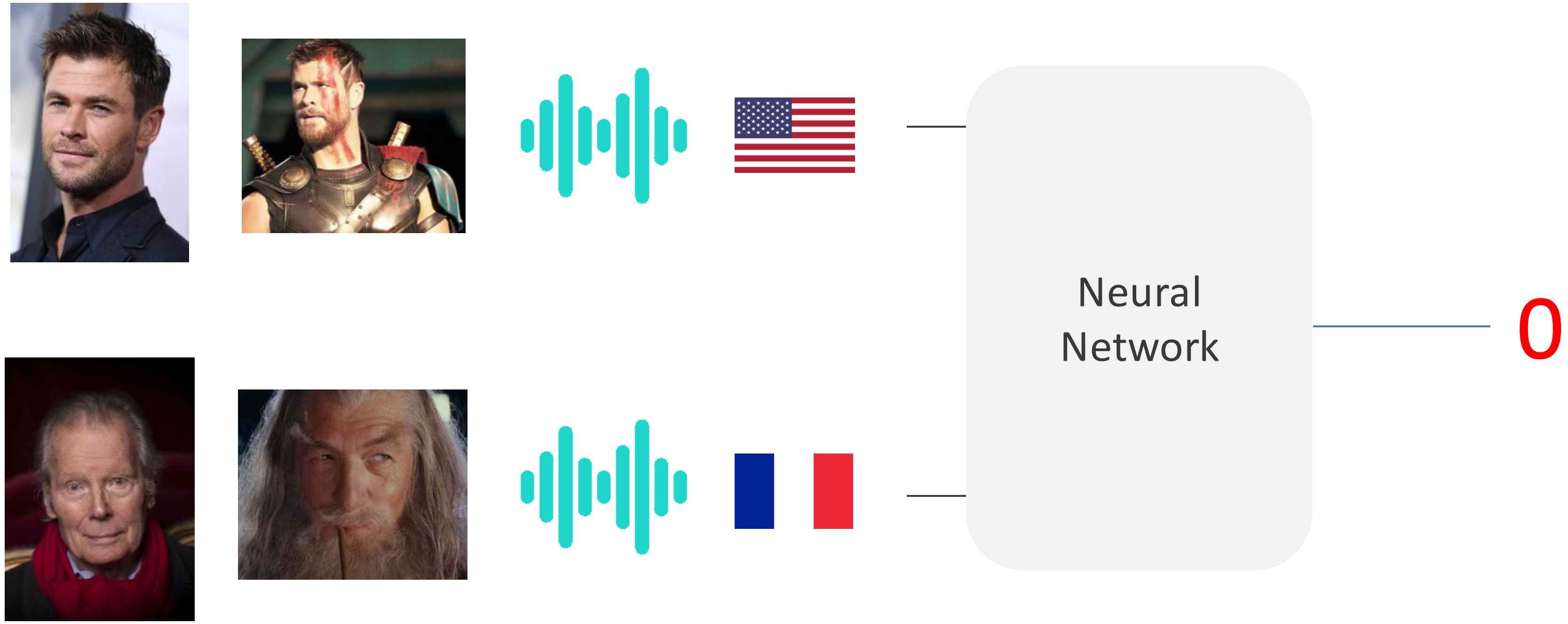
Neural Network

1

Voices play the same character

SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING, A. Gresse, M. Quillot, R. Dufour, V. Labatut, JF. Bonastre

Voice Similarity: Training



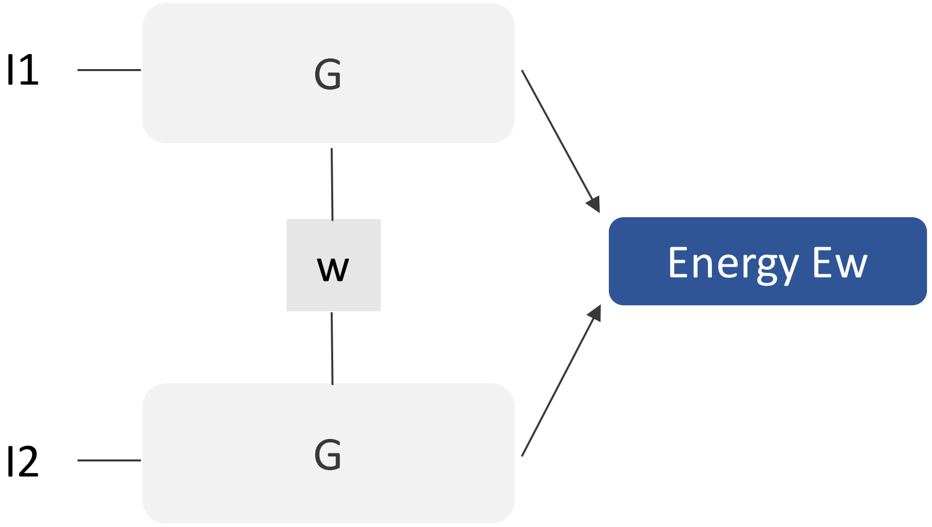
Neural Network

0

Voices don't play the same character

SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING, A. Gresse, M. Quillot, R. Dufour, V. Labatut, JF. Bonastre

Energy formula and contrastive loss



$$E_w(I_1, I_2) = (\| G_w(I_1) - G_w(I_2) \|)^2$$

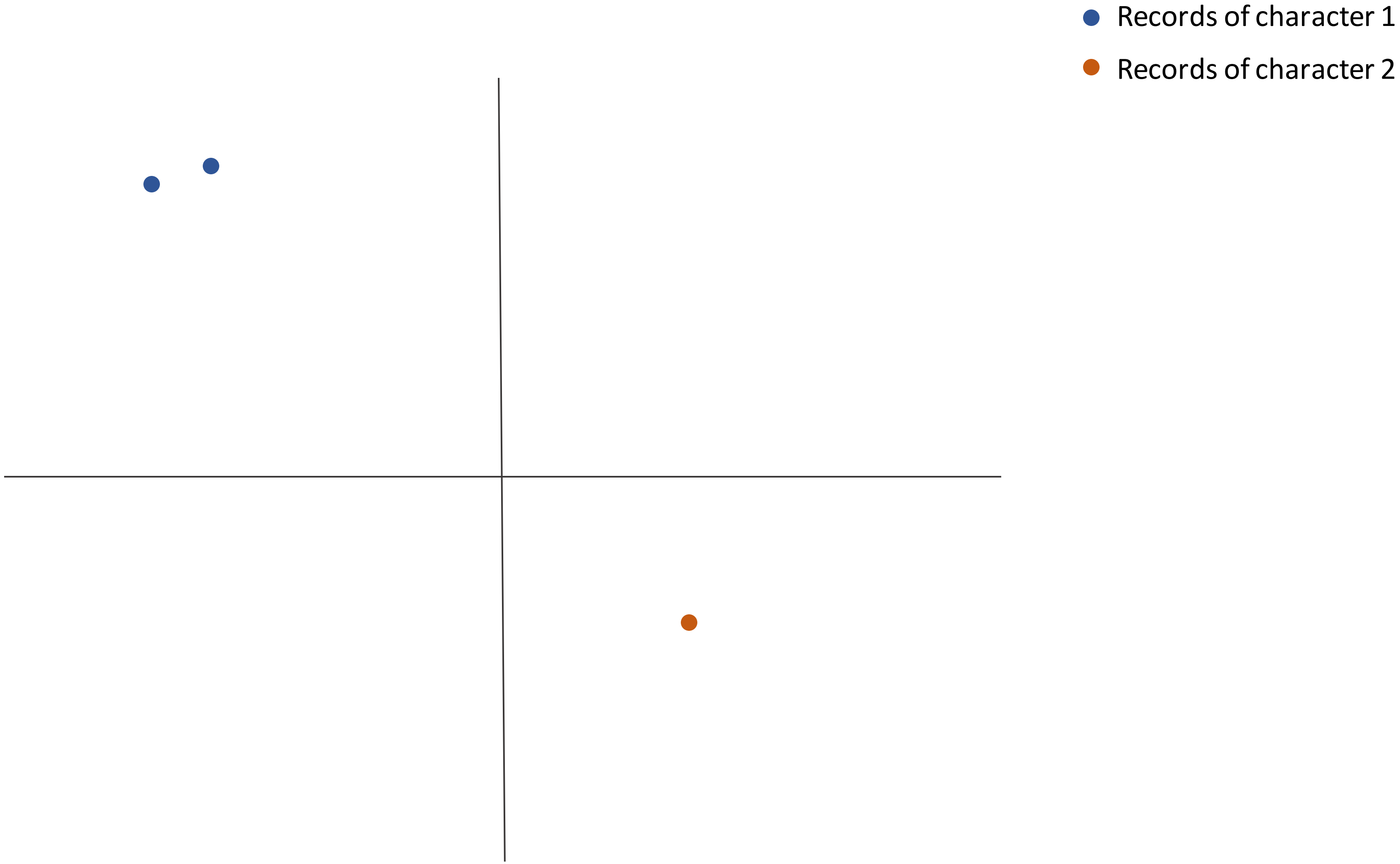
Contrastive Loss

$$L(I_1,\, I_2,\, T) = (1-T) \times E_W(I_1,\, I_2) + T \times max\{0,\, m - E_W(I_1,\, I_2)\}$$

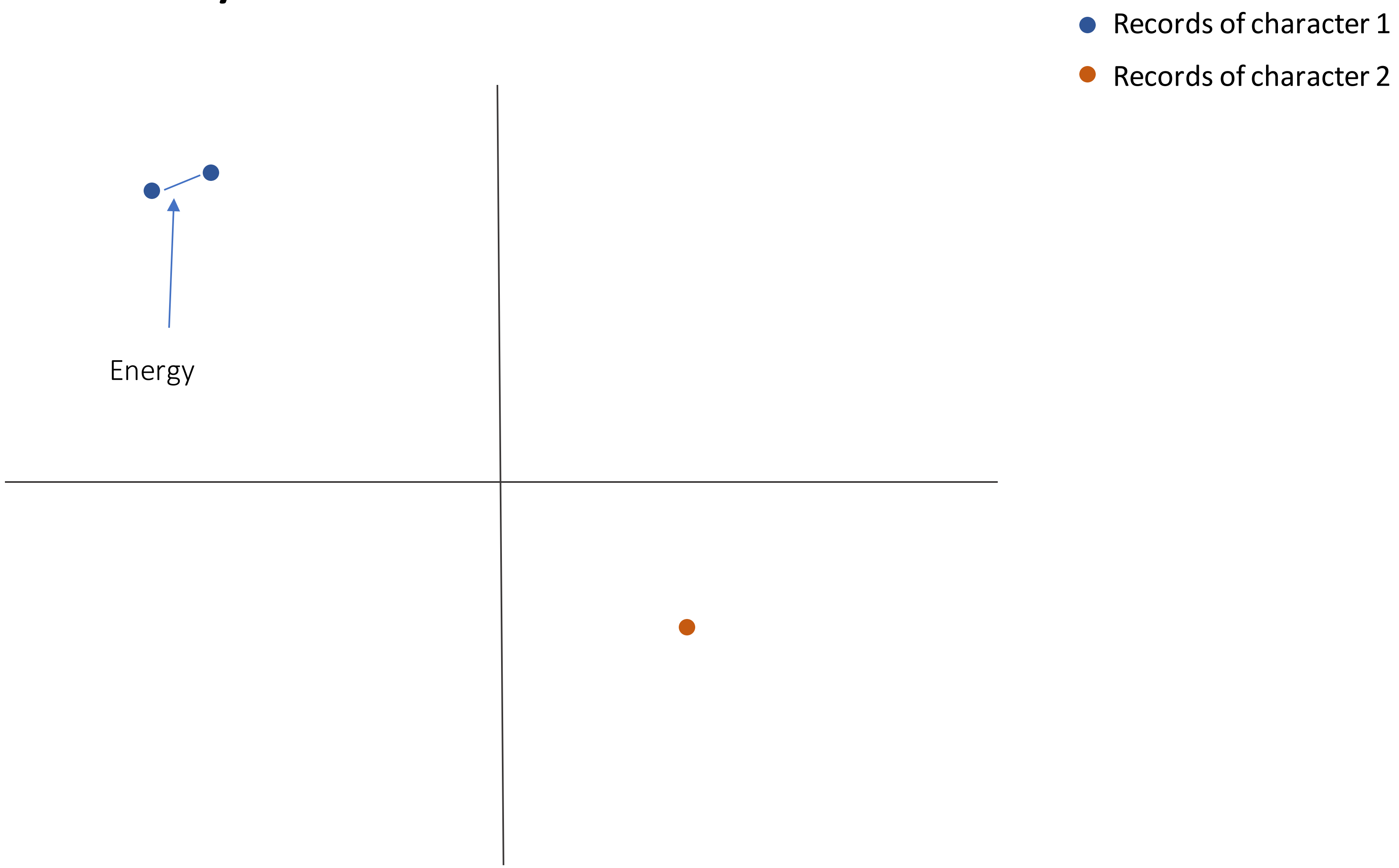
T = {0, 1}
m is the margin

SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING, A. Gresse, M. Quillot, R. Dufour, V. Labatut, JF. Bonastre

Voice Similarity



Voice Similarity

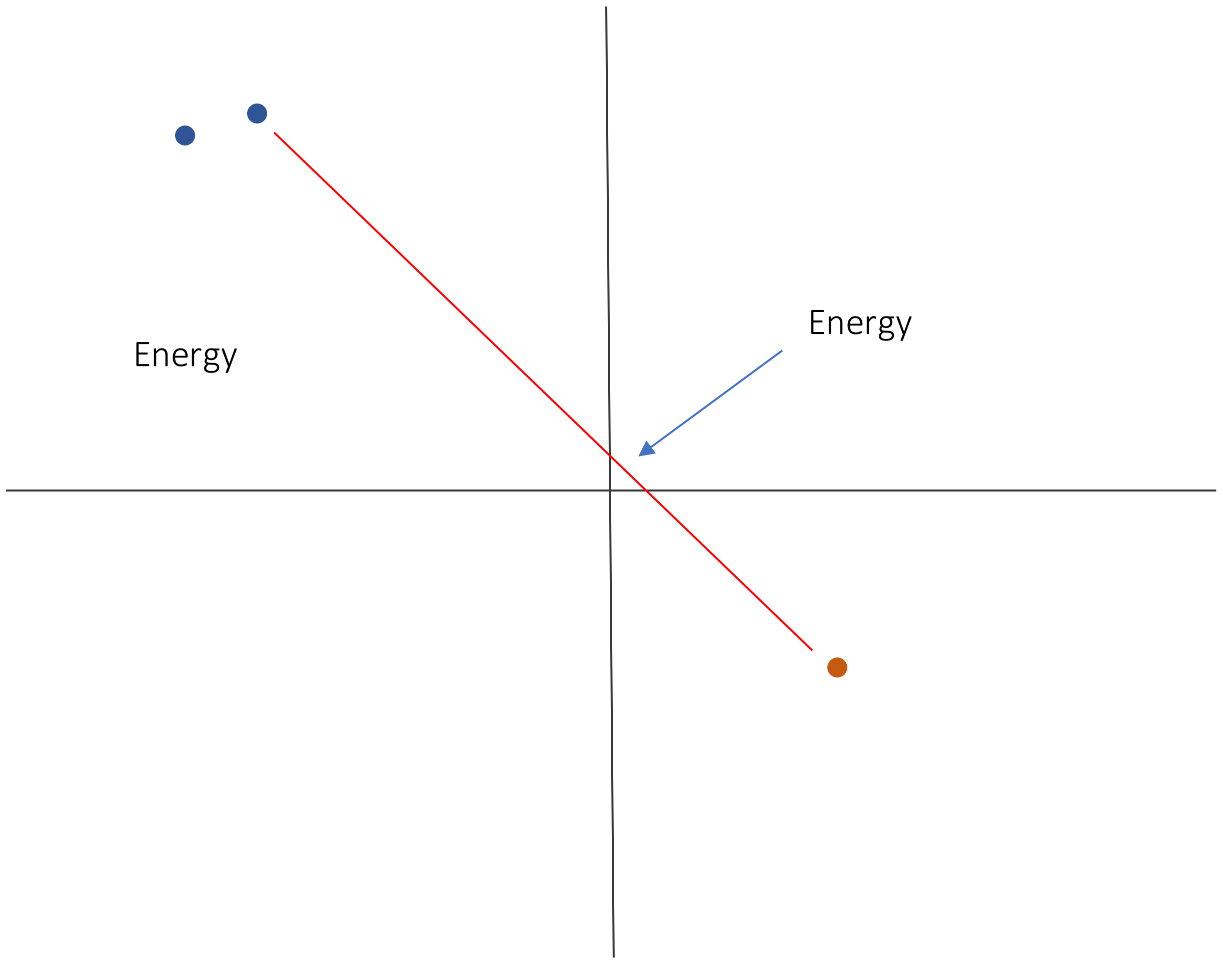


● Records of character 1

● Records of character 2

Energy

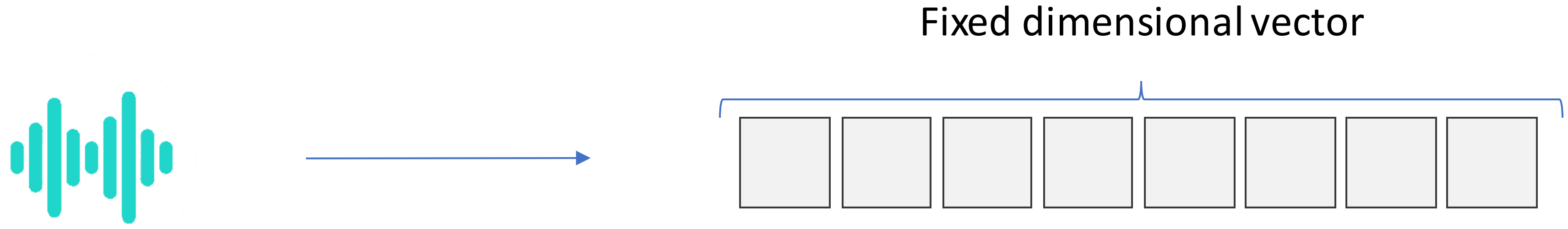
Voice Similarity



● Records of character 1

● Records of character 2

I-vectors



Fixed dimensional vector

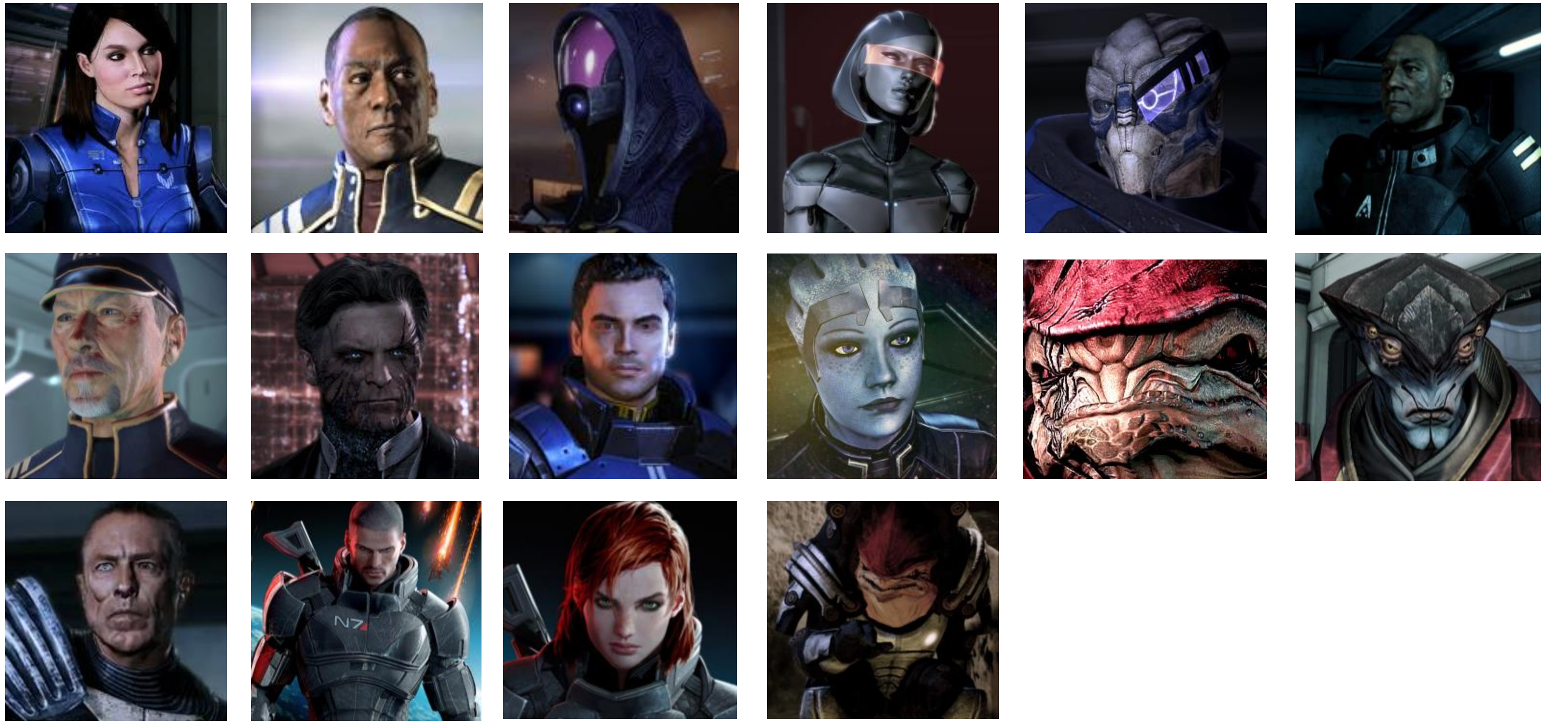
Data

2880 records
Total duration 161 min
16 characters
Cross validation, sets A, B, C and D

train	val	test
1728	432	720
12 chars	12 chars	4 chars
98 min	23 min	39 min
8 min / char	2 min / char	10 min / char

SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING A Gresse, M Quillot, R Dufour, V Labatut, J-F Bonastre

16 characters



SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING A Gresse, M Quillot, R Dufour, V Labatut, J-F Bonastre

Results

	2 in-conc acc	2 in-merge acc	Siamese-net acc
A (test)	0.49	0.52	**0.55**
B (test)	0.49	0.50	**0.59**
C (test)	0.51	0.53	**0.62**
D (test)	**0.53**	0.52	0.50
A (dev)	**0.94**	0.93	0.72
B (dev)	**0.96**	0.94	0.71
C (dev)	**0.93**	0.93	0.70
D (dev)	**0.96**	0.96	0.71

Presence of acoustic signs of the character dimension ***confirmed***

SIMILARITY METRIC BASED ON SIAMESE NEURAL NETWORKS FOR VOICE CASTING A Gresse, M Quillot, R Dufour, V Labatut, J-F Bonastre

How to represent the character dimension of the acted voice?

2

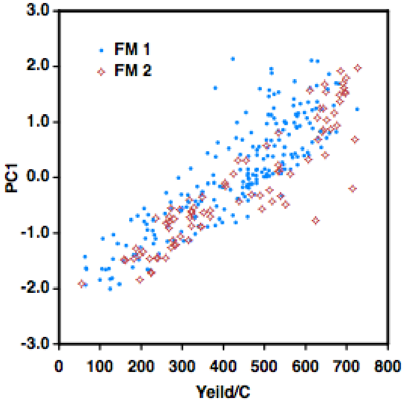
Abstract level

Linguistic content

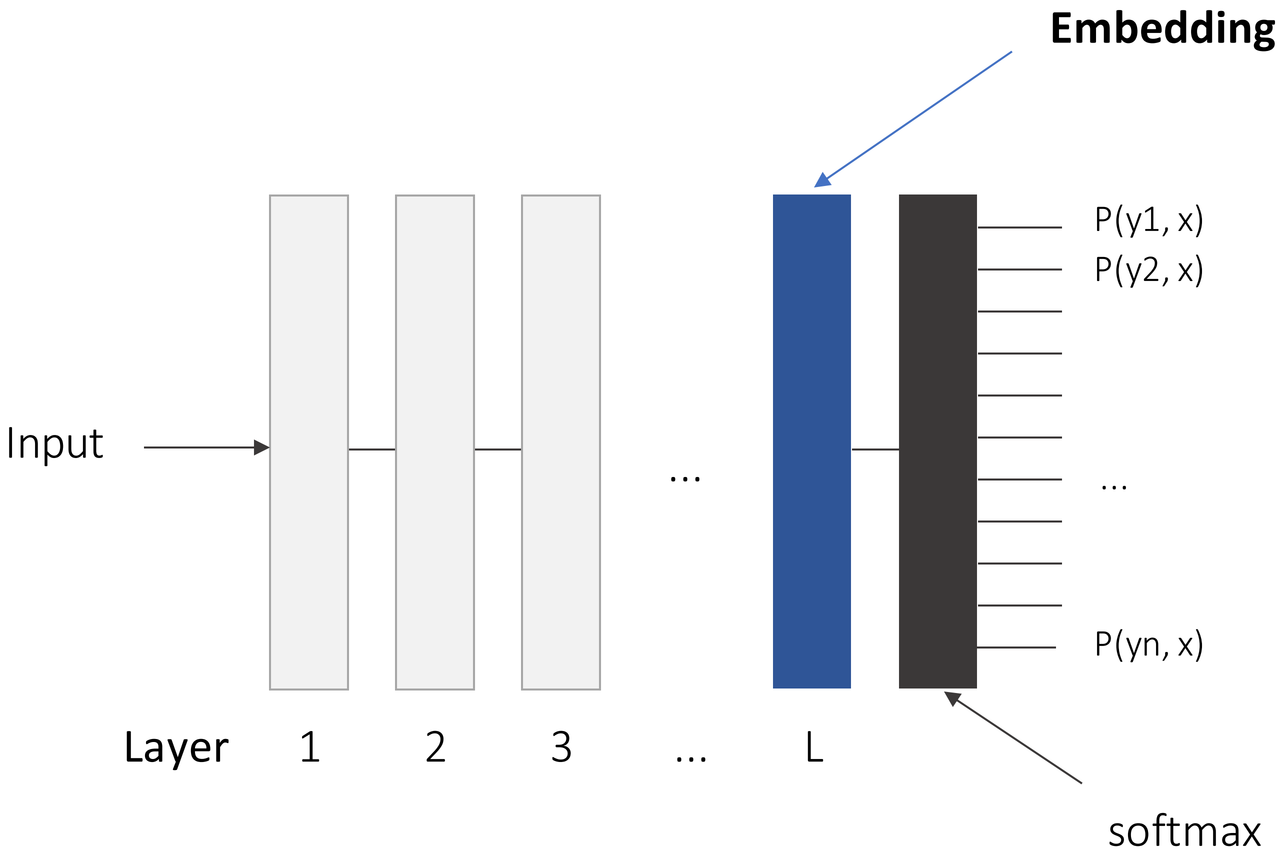
Emotion

Formant

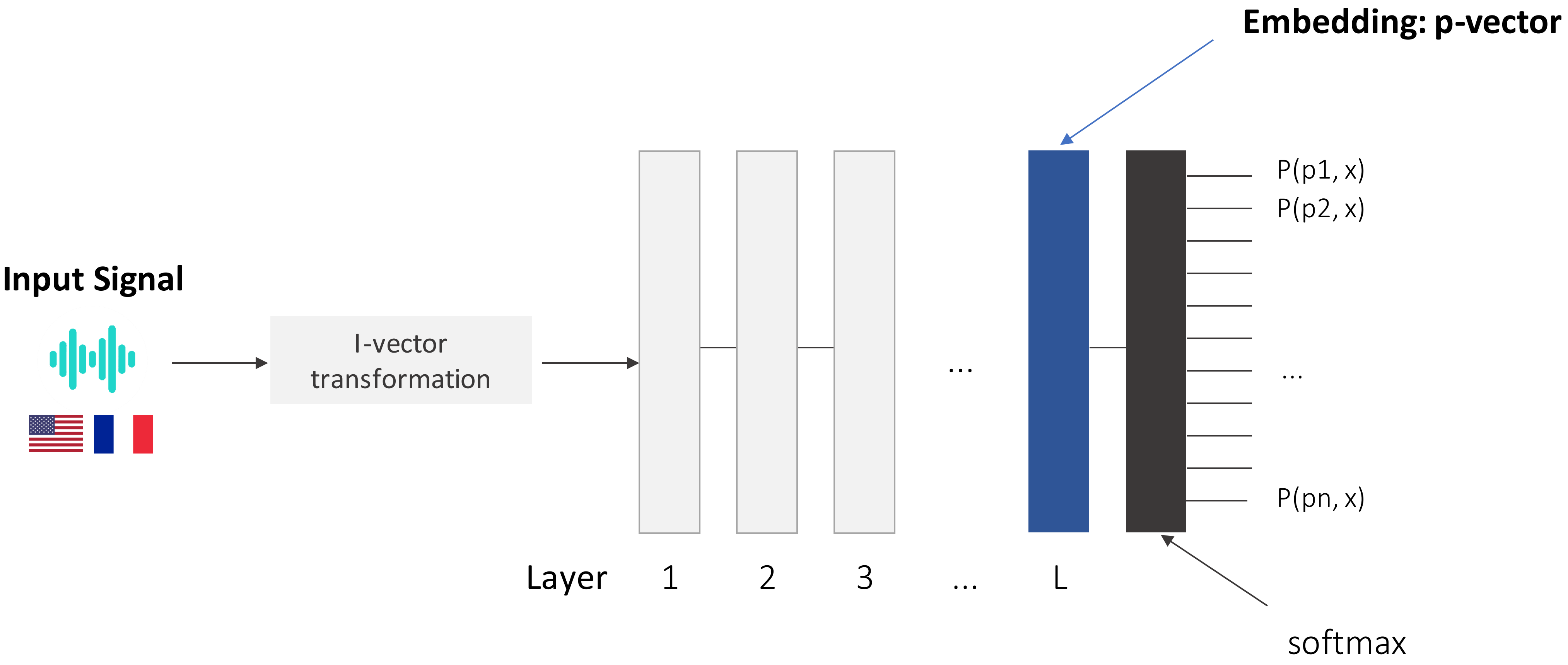
Dimension representation



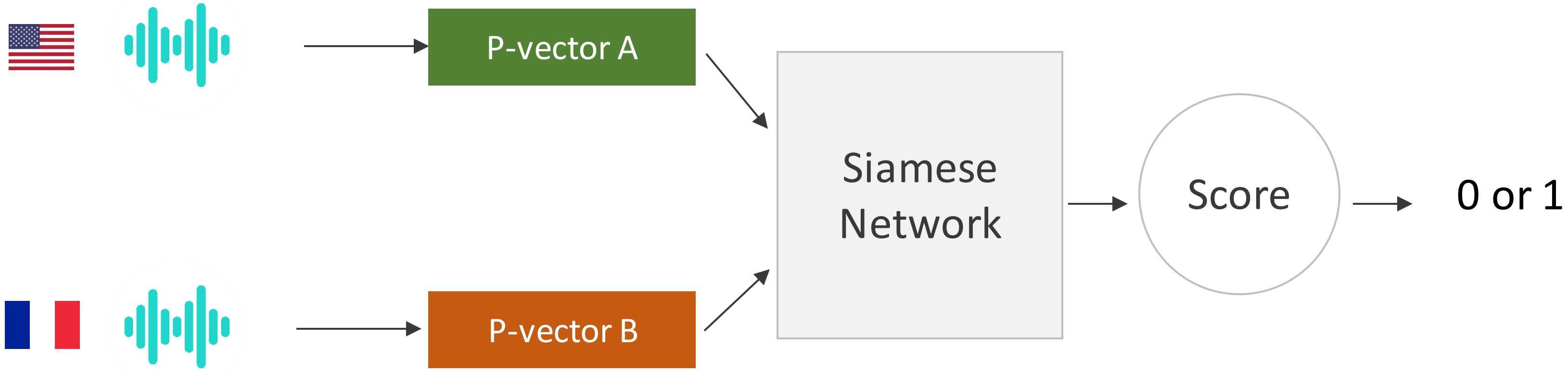
Neural Network - Embedding



Neural Network - P-Vector

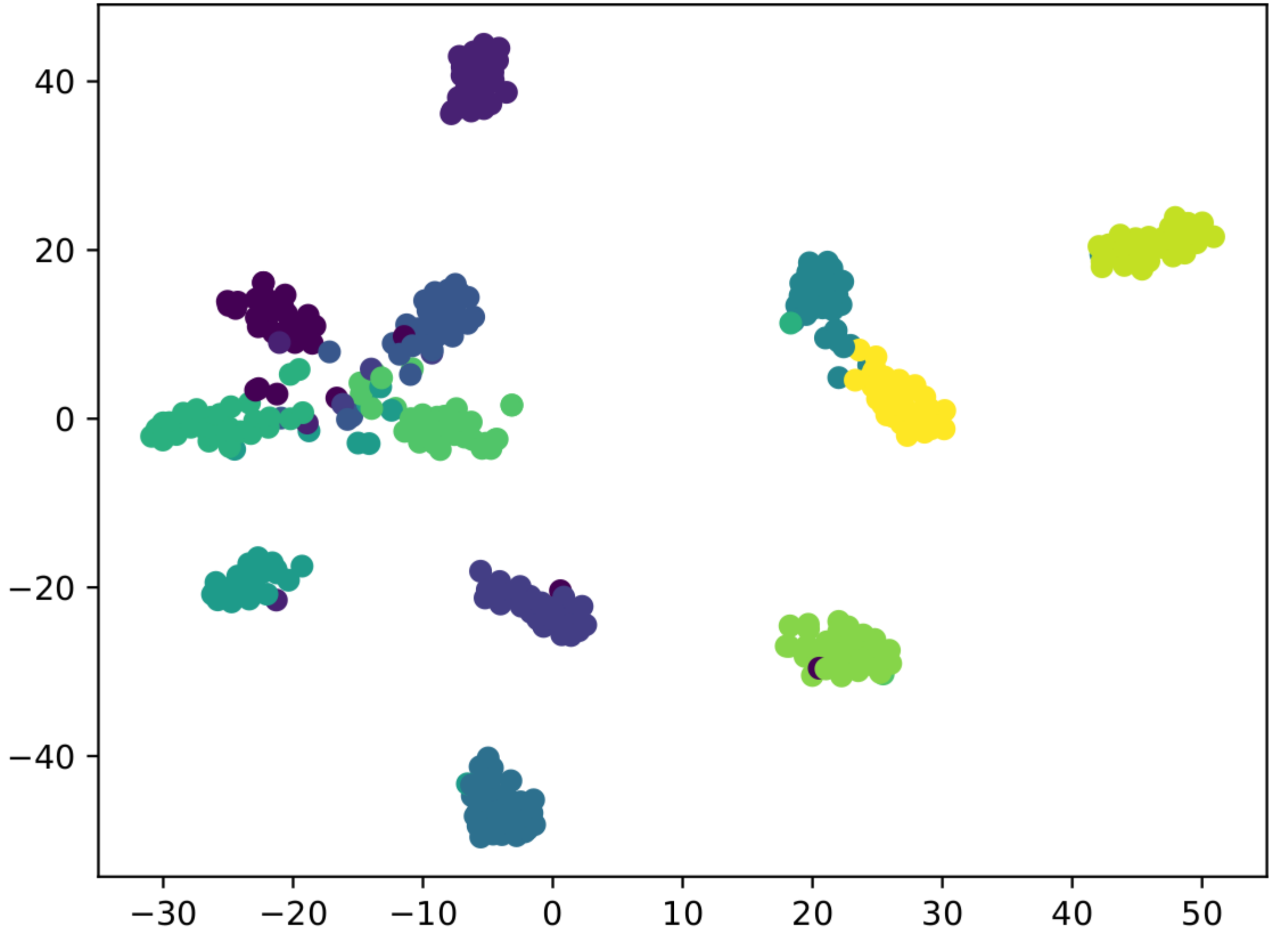


Evaluation



⚠ This evaluation does not ensure that p-vectors model precisely the character dimension

Result: with TSNE



How to go deeper in the representation without meta data?

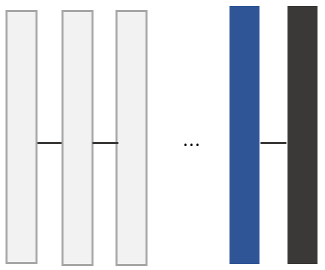
Data refining

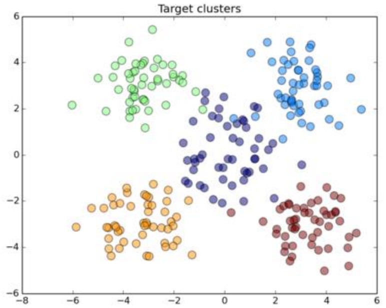
Remove data

Redefine labels

Add neutral label

Data refining: redefine labels





Initial corpus

i-vector	character
{0.2, 0.3, 0.2, 0.3 … 0.4}	1
{0.5, 0.3, 0.3, 0.2 … 0.9}	3
{0.7, 0.4, 0.3, 0.9 … 0.3}	4
{0.6, 0.5, 0.3, 0.2 … 0.6}	1

Replace labels

i-vector	Associated cluster
{0.2, 0.3, 0.2, 0.3 … 0.4}	25
{0.5, 0.3, 0.3, 0.2 … 0.9}	15
{0.7, 0.4, 0.3, 0.9 … 0.3}	13
{0.6, 0.5, 0.3, 0.2 … 0.6}	19

How to choose k?

Results

	A	B	C	D
Baseline	0.63	0.55	0.55	0.55

Baseline learned with teacher/student method

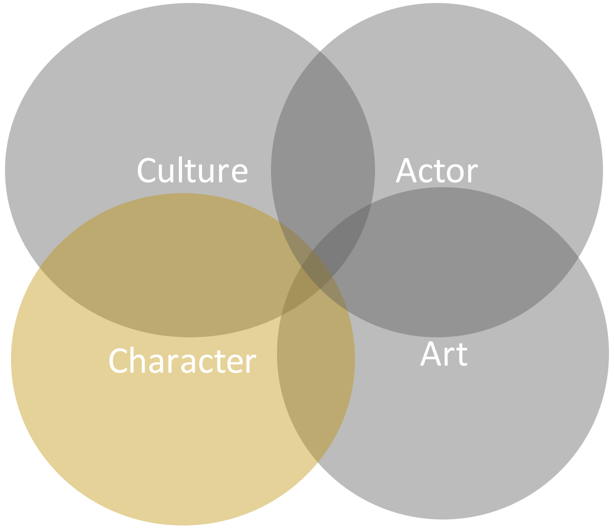
	6	12	24	48	64
Siamese val A	0.80	**0.90**	0.88	0.87	0.87
Siamese val B	0.78	**0.92**	0.90	0.87	0.88
Siamese val C	0.81	**0.92**	0.89	0.87	0.85
Siamese val D	0.74	**0.90**	0.88	0.87	0.85
Siamese test A	0.54	0.51	0.54	0.55	**0.57**
Siamese test B	0.55	0.56	0.53	0.48	0.55
Siamese test C	0.55	0.54	0.55	0.56	0.56
Siamese test D	0.57	0.52	0.54	0.51	0.53

We still keep character information in this refined representation

Next step: How to compare with humans?

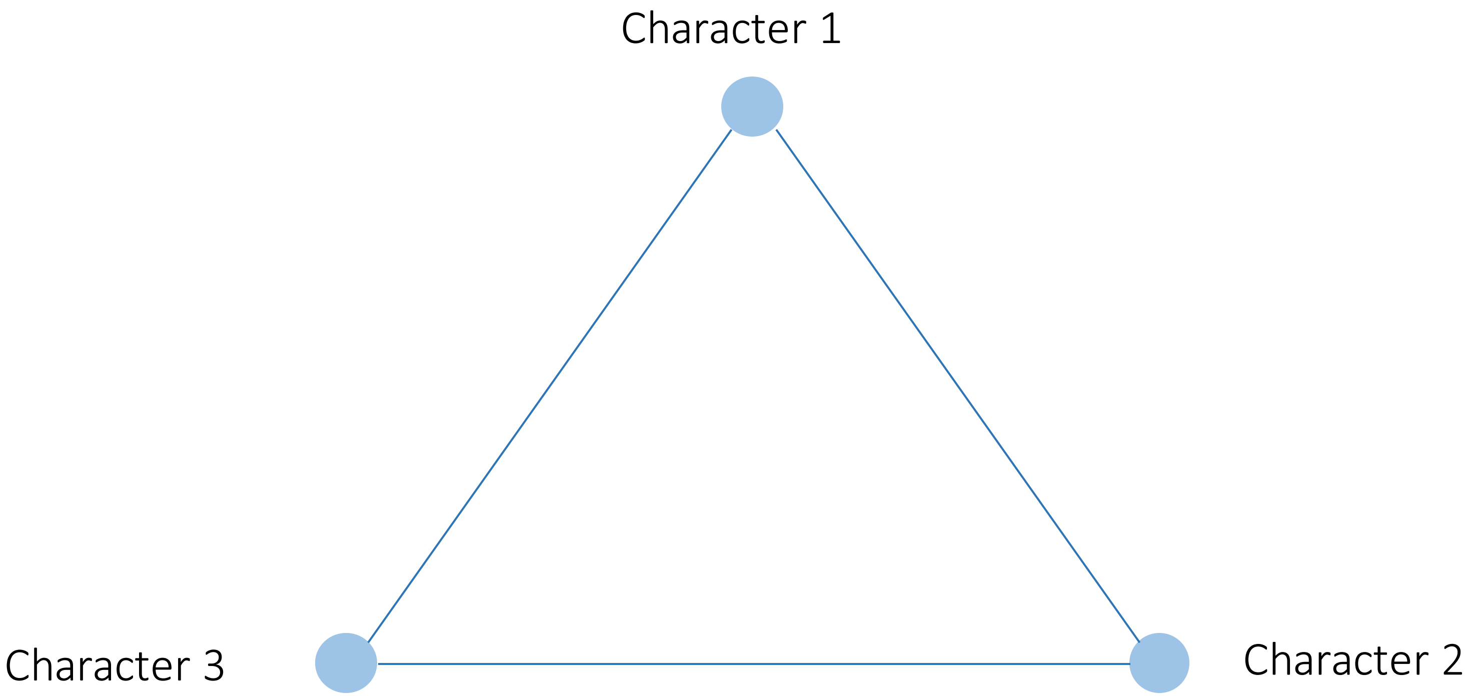
3

Human level



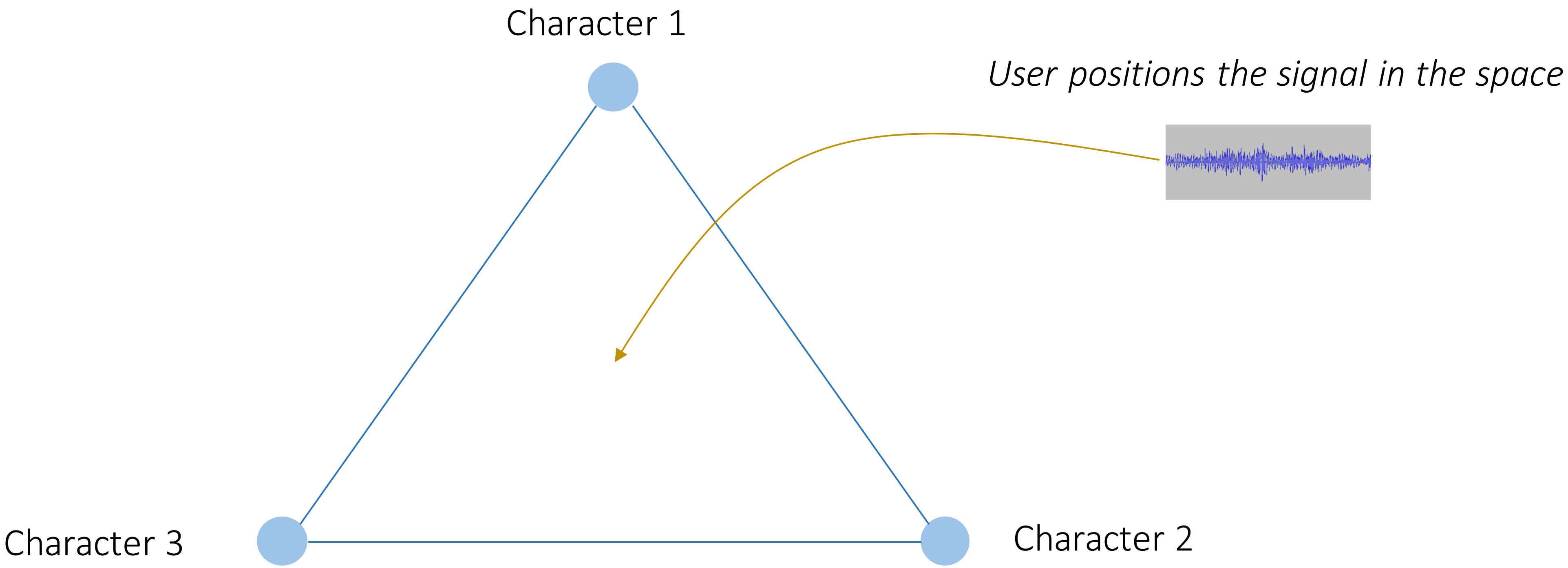
Compare with human experts

Triangular plan



This work will be done with the help of the sociologists of the department of culture and communication of Avignon

Triangular plan



User positions the signal in the space

Triangular plan

Record	User	Character 1	Character 2	Character 3
1	1	0,9	0,2	0,1
2	1	0,1	0,4	0,7
…	…	…	…	…
4	2	0,1	0,7	0,2

Comparable with our p-vectors ?

Can feed machine learning systems to make new representation

How to explain decisions from neural networks?

4

Explainability

Why explainability?

Expert **Doctor** **Learner** **Researcher**

Different kind of explainability

Sensitivity Analysis (SA)

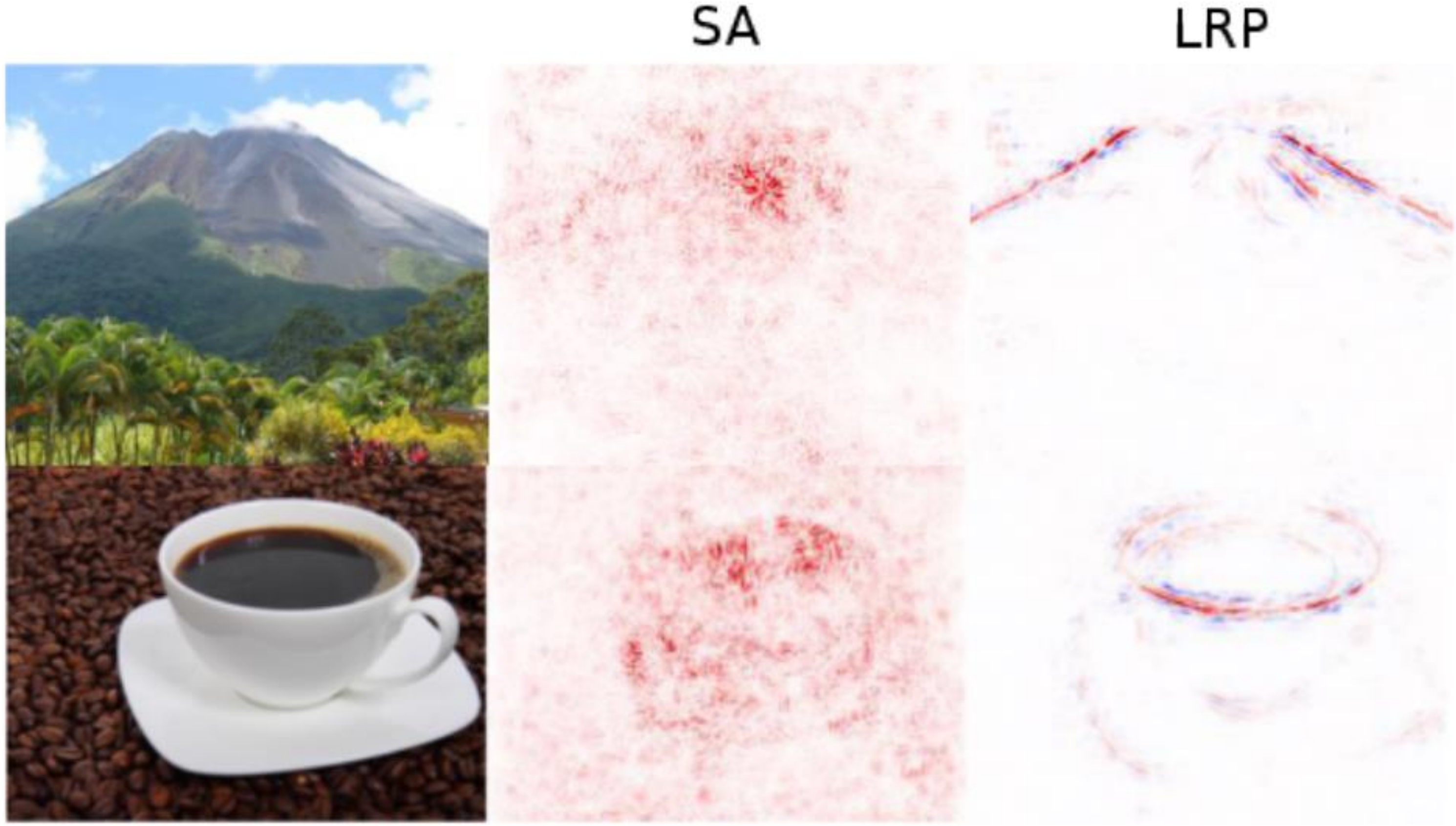
$$R_i = \left|\left|\frac{\partial}{\partial x_i} f(\mathbf{x})\right|\right|.$$

Layer-wise Relevance Propagation (LRP)

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon} R_k$$

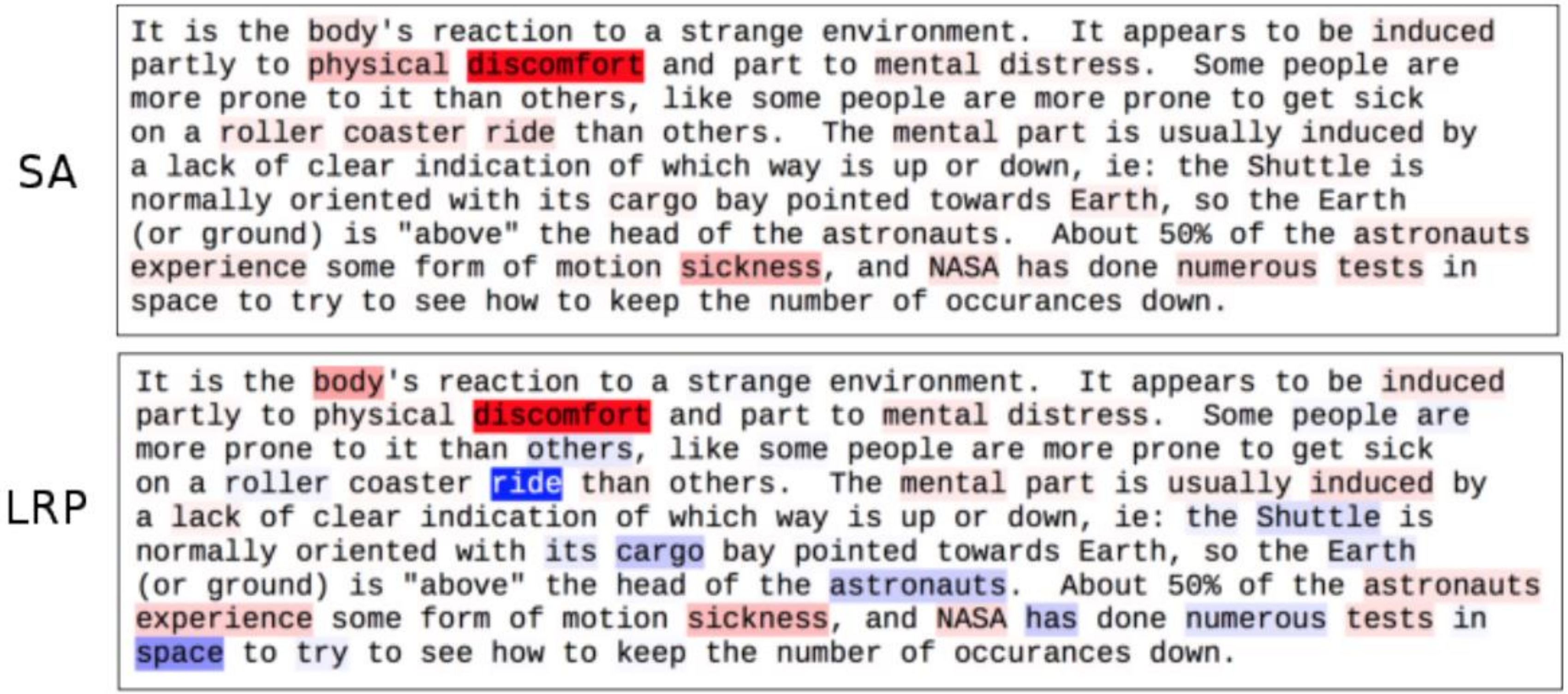
EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS (2017)
W Samek, T Wiegand, KR Müller

Image classification



EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS (2017)
W Samek, T Wiegand, KR Müller

Text document classification



SA

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurances down.

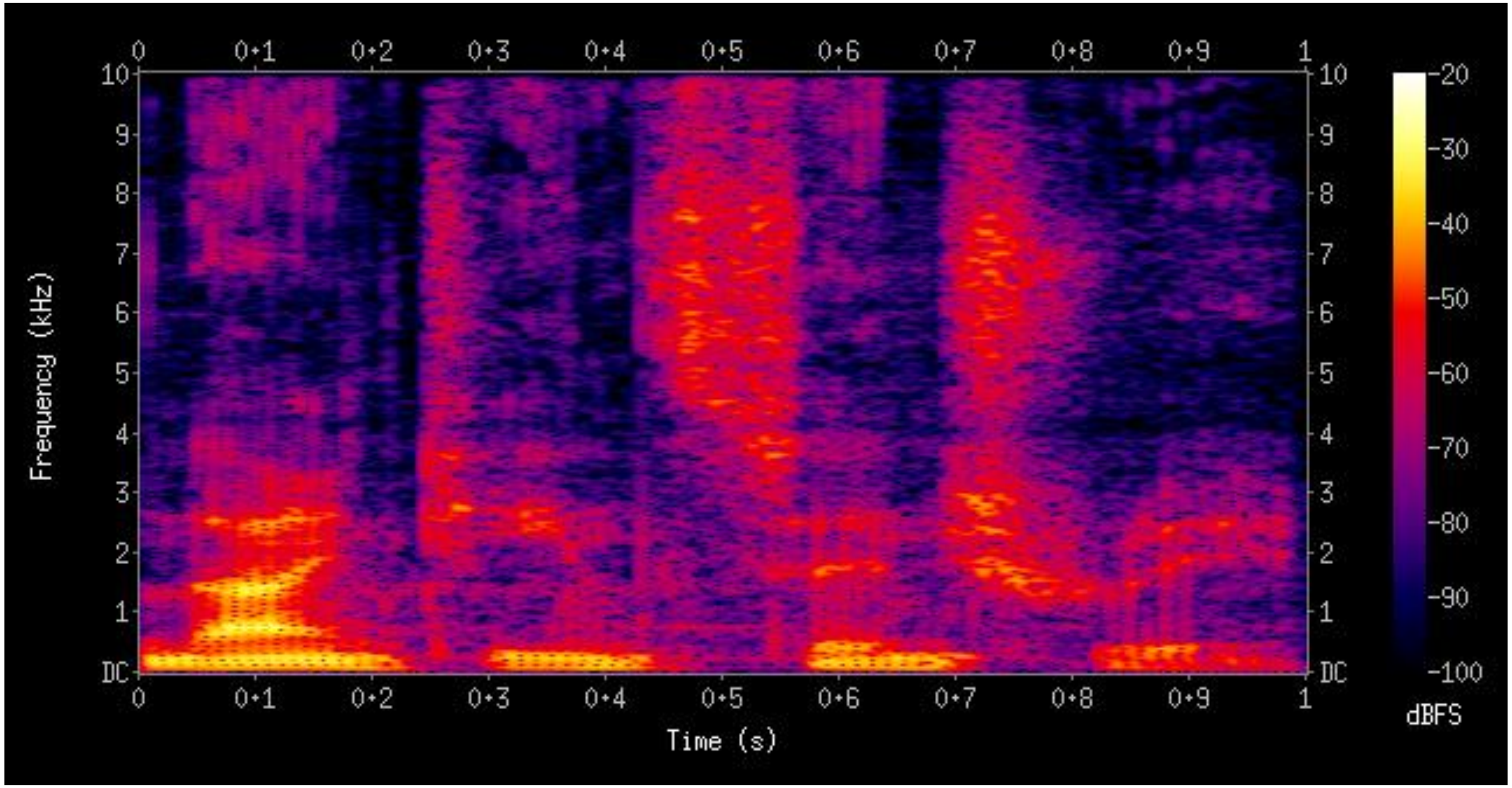
LRP

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurances down.

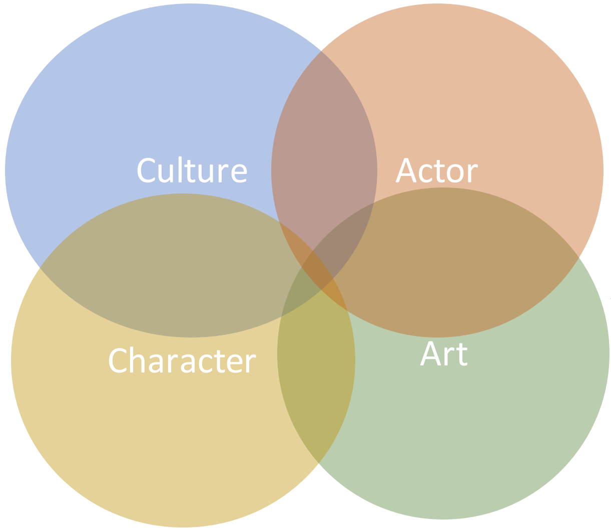
EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS (2017)
W Samek, T Wiegand, KR Müller

And for sound task?

Work with spectrogram



Conclusion



3 Compare with human experts

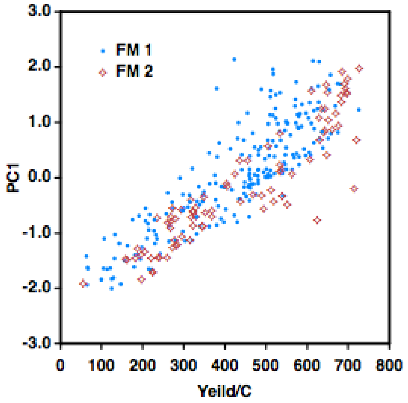
4 **Explainability**

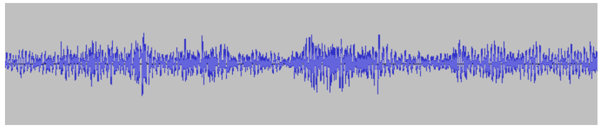
Linguistic content

Emotion

Formant

2 Dimension representation





1 Confirm the dimension is present in the signal

Work difficulties and future

Difficult to generalize the task

Build new corpus for cinema

Improve validation set

Subjective experiments and explainability

Thank you for your attention

mathias.quillot@univ-avignon.fr

Claude Chantal

Christophe Le Moine