FOLLOWING THE EMBEDDING: IDENTIFYING TRANSITION PHENOMENA IN WAV2VEC 2.0 REPRESENTATIONS OF SPEECH AUDIO

Patrick Cormac English[1], Erfan A. Shams[1], John D. Kelleher[2], Julie Carson-Berndsen[1]
SFI Centre for Research Training in Digitally-Enhanced Reality (d-real),
[1]ADAPT Research Centre, School of Computer Science, University College Dublin, Ireland
[2]ADAPT Research Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland.

Centre for Research Training

Science Foundation Ireland For what's next

Overview

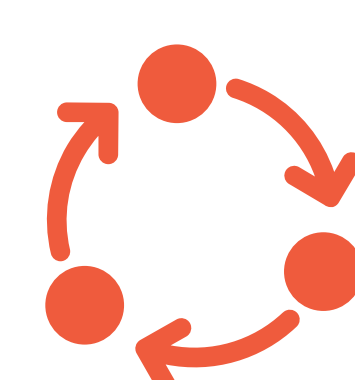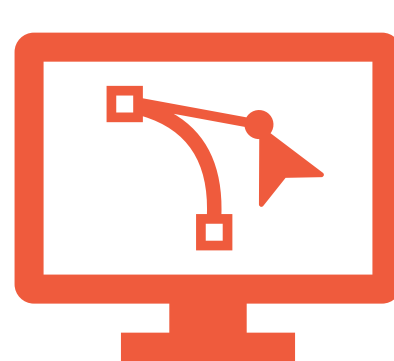## Outline

This study investigates the encoding of speech signal information in wav2vec 2.0 embeddings, focusing on phone intersections to determine how detection ouputs reflect phonological events, and the extent to which the intersection reflects unique phonological data vs reflecting characteristics of neighbouring phones.

Target

By probing the hidden representations of the self-supervised model, the research aims to uncover traces of learned articulatory features, such as manner of articulation (MOA) and place of articulation (POA), at a more granular level than phone annotations.
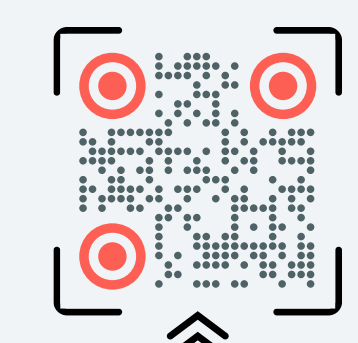
Method

The methodology involves training multilayer perceptron (MLP) models on phone-averaged embeddings to predict articulatory feature presence and then applying these probes to unaveraged time-step embeddings, enabling a more detailed analysis of the model's learned representations.

Visualise

A custom visualisation tool, w2v2viz, was developed to facilitate the analysis and manual inspection of probe outputs, displaying feature probabilities across layers and frames in a 3D terrain format, allowing users to evaluate probe feature probabilities at each time-step in wav2vec 2.0 representations.

Show & Tell Submission #6

w2v2viz: An Interactive Transformer Probe Visualisation Toolkit
GITHUB:

SCAN ME
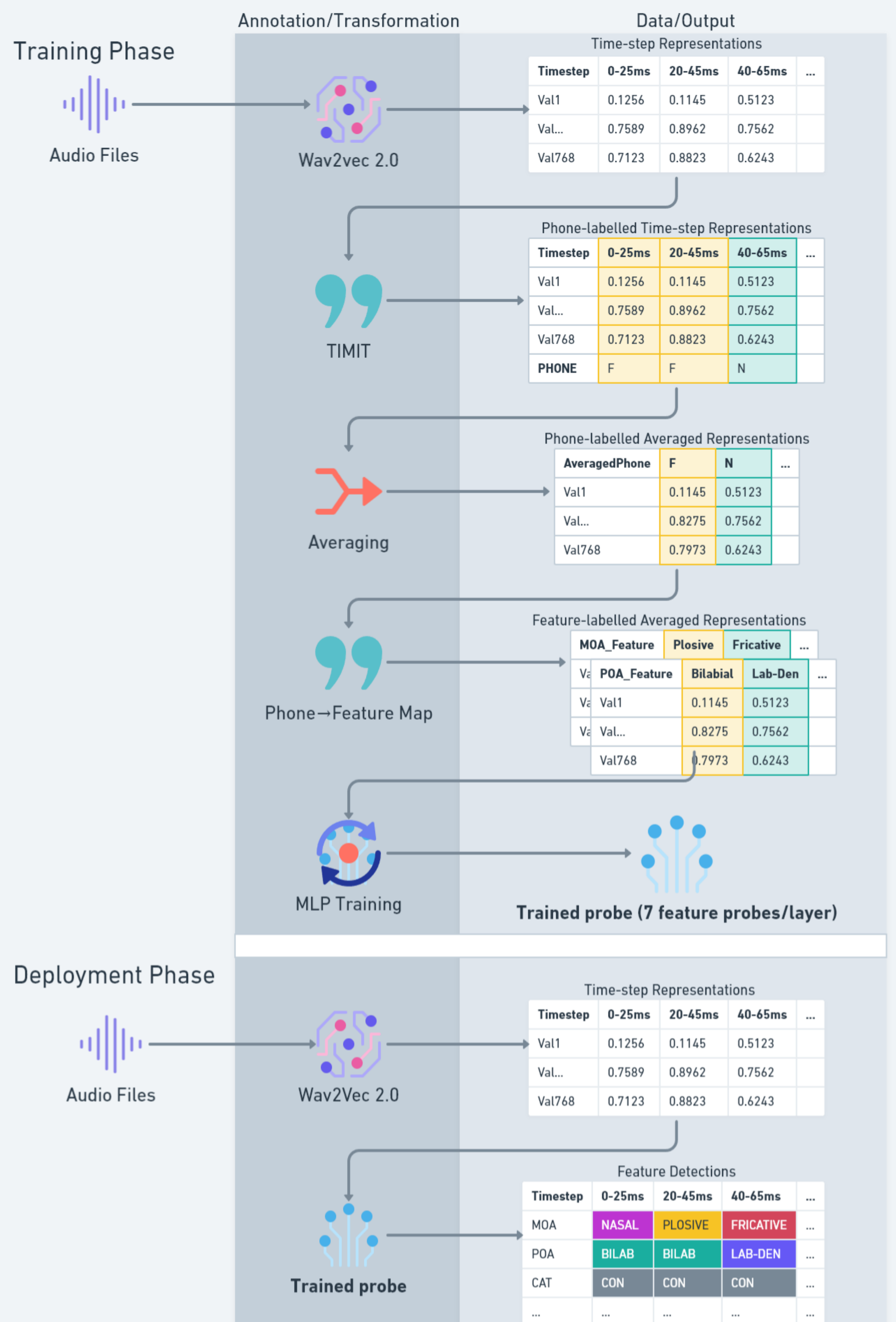https://github.com/erfanashams/w2v2viz

Aims

By providing insights into the structure of speech signal information learned by unsupervised transformer models at phone intersections, this study contributes to the development of more explainable speech processing systems and hopes to advance our understanding of how these models capture phonological events.
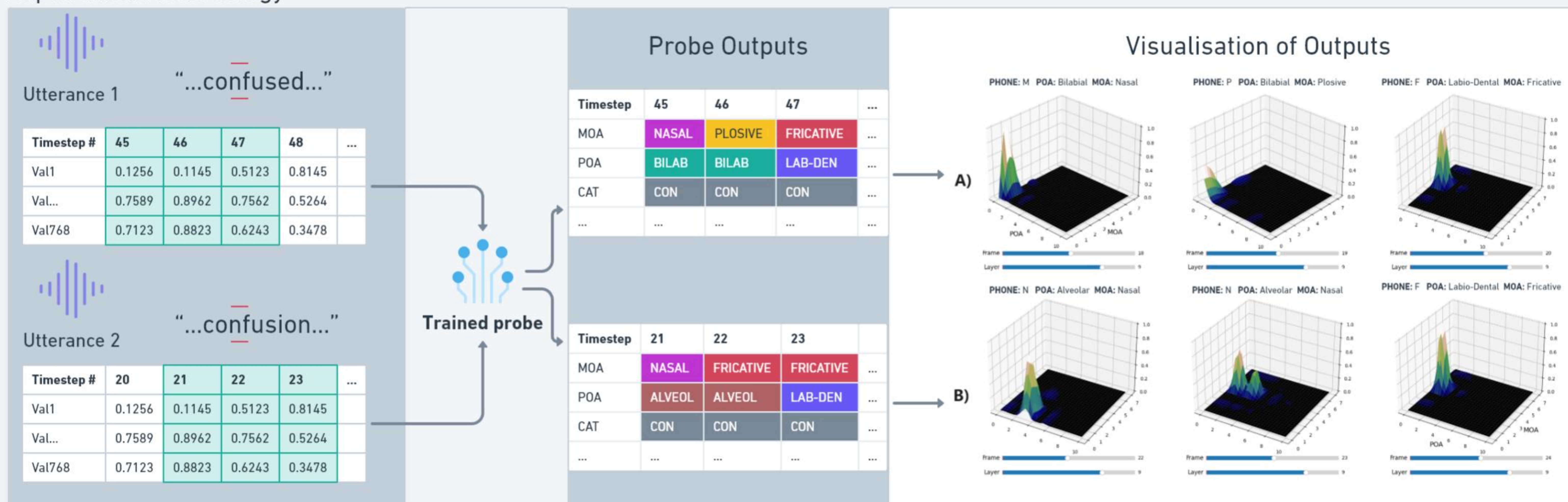

Training Phase

Annotation/Transformation — Data/Output

Audio Files → Wav2Vec 2.0

Time-step Representations

Timestep	0–25ms	20–45ms	40–65ms	...
Val1	0.1256	0.1145	0.5123	
Val...	0.7589	0.8962	0.7562	
Val768	0.7123	0.8823	0.6243	

TIMIT

Phone-labelled Time-step Representations

Timestep	0–25ms	20–45ms	40–65ms	...
Val1	0.1256	0.1145	0.5123	
Val...	0.7589	0.8962	0.7562	
Val768	0.7123	0.8823	0.6243	
PHONE	F	F	N	

Averaging

Phone-labelled Averaged Representations

AveragedPhone	F	N	...
Val1	0.1145	0.5123	
Val...	0.8275	0.7562	
Val768	0.7973	0.6243	

Phone→Feature Map

Feature-labelled Averaged Representations

MOA_Feature	Plosive	Fricative	...
POA_Feature	Bilabial	Lab-Den	...
Val1	0.1145	0.5123	
Val...	0.8275	0.7562	
Val768	0.7973	0.6243	

MLP Training → **Trained probe (7 feature probes/layer)**

Deployment Phase

Audio Files → Wav2Vec 2.0

Time-step Representations

Timestep	0–25ms	20–45ms	40–65ms	...
Val1	0.1256	0.1145	0.5123	
Val...	0.7589	0.8962	0.7562	
Val768	0.7123	0.8823	0.6243	

Trained probe

Feature Detections

Timestep	0–25ms	20–45ms	40–65ms	...
MOA	NASAL	PLOSIVE	FRICATIVE	
POA	BILAB	BILAB	LAB-DEN	
CAT	CON	CON	CON	
...	...	...	...	

Exploration Methodology

Utterance 1 — "...confused..."

Timestep #	45	46	47	48	...
Val1	0.1256	0.1145	0.5123	0.8145	
Val...	0.7589	0.8962	0.7562	0.5264	
Val768	0.7123	0.8823	0.6243	0.3478	

Utterance 2 — "...confusion..."

Timestep #	20	21	22	23	...
Val1	0.1256	0.1145	0.5123	0.8145	
Val...	0.7589	0.8962	0.7562	0.5264	
Val768	0.7123	0.8823	0.6243	0.3478	

Trained probe

Probe Outputs

Timestep	45	46	47	...
MOA	NASAL	PLOSIVE	FRICATIVE	
POA	BILAB	BILAB	LAB-DEN	
CAT	CON	CON	CON	
...	...	...	...	

A)

Timestep	21	22	23	...
MOA	NASAL	FRICATIVE	FRICATIVE	
POA	ALVEOL	ALVEOL	LAB-DEN	
CAT	CON	CON	CON	
...	...	...	...	

B)

Visualisation of Outputs

PHONE: M POA: Bilabial MOA: Nasal
PHONE: P POA: Bilabial MOA: Plosive
PHONE: F POA: Labio-Dental MOA: Fricative

PHONE: N POA: Alveolar MOA: Nasal
PHONE: N POA: Alveolar MOA: Nasal
PHONE: F POA: Labio-Dental MOA: Fricative

UCD DUBLIN

ADAPT — Engaging Content Engaging People

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

GAILLIMH 1845 GALWAY

Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

DUBLIN TECHNOLOGICAL UNIVERSITY DUBLIN — Ollscoil Teicneolaíochta Bhaile Átha Cliath