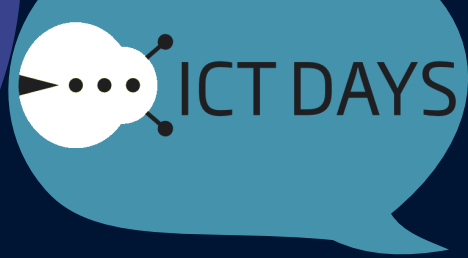


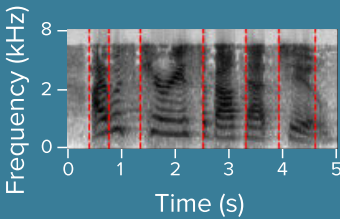
Making speech translation AI explainable: New methods reveal how models (mis)gender speakers



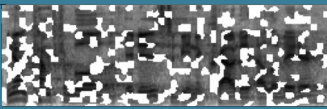
Method

Contrastive Explanations for Speech Translation

Why 'curiosa' (F) and not 'curioso' (M)?



Perturb spectrogram regions

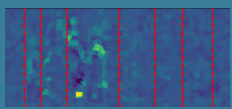


Measure impact on output

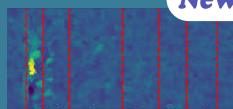


Identify causal features

New!



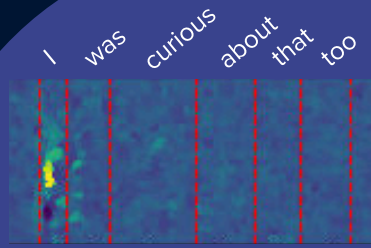
Standard Explanation



Contrastive explanation

Findings

How models assign gender



Time Dimension

'I' functions as an acoustic gendered pronoun, like 'she'/'he' in text translation

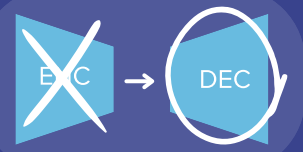
Frequency Dimension

Speaker gender info comes from formants (F1/F2), not just pitch

Training Data



Models don't memorize patterns like 'scientist' → masculine, but learn to use 'masculine by default'



The **Internal Language Model** has a strong masculine bias, but acoustic input can override it



Lina Conti
PhD Student



UNIVERSITÀ DI TRENTO
Department of Information Engineering and Computer Science