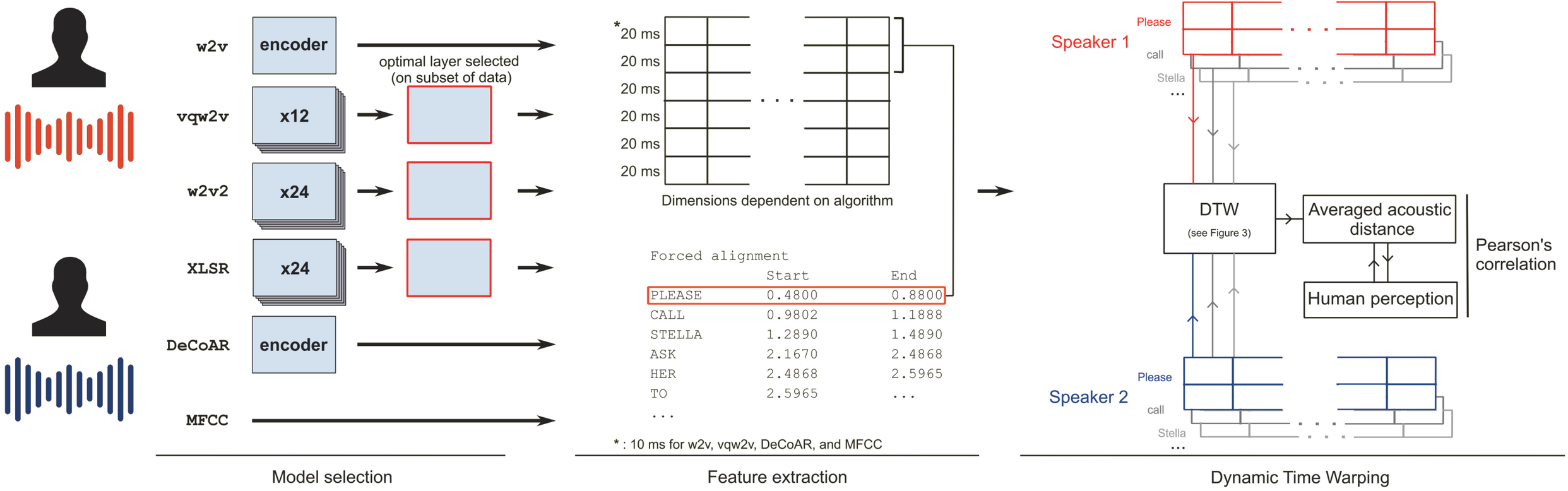
w2v
encoder
optimal layer selected
(on subset of data)
vqw2v
x12
w2v2
x24
XLSR
x24
DeCoAR
encoder
MFCC
Model selection
*20 ms
20 ms
20 ms
20 ms
20 ms
20 ms
Dimensions dependent on algorithm
Forced alignment
Start
End
PLEASE 0.4800 0.8800
CALL 0.9802 1.1888
STELLA 1.2890 1.4890
ASK 2.1670 2.4868
HER 2.4868 2.5965
TO 2.5965 ...
* : 10 ms for w2v, vqw2v, DeCoAR, and MFCC
Feature extraction
Speaker 1
Please
call
Stella
DTW
(see Figure 3)
Averaged acoustic distance
Human perception
Pearson's correlation
Speaker 2
Please
call
Stella
Dynamic Time Warping