HUMAN EVALUATION FOR SPEECH & LANGUAGE AI

Human judgement your *speech* models can learn from.



Automated speaking scores, synthetic voices and speech recognition all need a reliable human reference. I bring fourteen years of rating spoken English on the ACTFL proficiency scale to evaluation work for speech and language teams.

WHERE I FIT

Gold-standard proficiency ratings for training and validating automated scoring; naturalness and intelligibility reviews of synthetic voices; transcript checks on accented and learner English; rubric and rater-guideline writing.

2012
rating spoken English for ACTFL, remotely

OPIc
certified rater on the ACTFL proficiency scale

M.A.
TESOL, Teachers College, Columbia University

EN · FR · IT
native English, working French and Italian

WORKING TOGETHER

Remote and asynchronous
Recorded samples rated from Paris (CET), with batches arranged around your deadlines.

Per speaker or per project
Priced per rated sample, per session or per project, whichever suits your programme.

Confidential
NDAs welcome; audio, data and model details stay private.

Documented
Every rating tied to the published criteria, with written reasons on request.

WHAT I DO

Careful listening, documented and repeatable.

Proficiency-labelled reference sets	Learner speech rated against the ACTFL scale, so automated scores have a trusted benchmark to agree with.
Synthetic voice evaluation	Naturalness, prosody, pacing and intelligibility of TTS output, judged with a theatre-trained ear for delivery.
Transcript and ASR review	Accuracy checks on recognition output, especially for accented, non-native and learner speech.
Rubrics and rater guidelines	Clear criteria and worked examples, so ratings stay consistent across raters and over time.
Error analysis	Not only where a model scores differently from a human, but why: the patterns behind the disagreement.

WHY A PERSON

Software can score a recording. It can't always tell you why it isn't working.

Accent is not error.
Models trained on narrow data penalise accents. A human reference separates accent from real intelligibility problems.

A script versus a speaker.
Memorised answers fool fluency metrics. Experienced raters hear the difference.

Tone and register.
Polite or abrupt, confident or hesitant: qualities that matter to listeners and that metrics often miss.

Labels you can explain.
Every rating tied to written criteria, so disagreements can be traced and resolved.

EXPERIENCE

Fourteen years at the rating desk, twenty-five in the classroom.

2012 – present
Language Consultant, English as an Additional Language
American Council on the Teaching of Foreign Languages (ACTFL), remote

2005 – 2010
Middle School EAL Instructor
Washington Township Schools, Indianapolis

2002 – 2005
Family Literacy EAL & GED Instructor
PS 396, Even Start National Grant, Bronx, NY

2000 – 2002
Adult EAL Teacher
Riverside Language Program, New York

2000 – 2002
EAL Writing Tutor
Teachers College, Columbia University

EDUCATION

M.A., TESOL
Teachers College, Columbia University, 2002

BFA, Film & Theatre
NYU Tisch School of the Arts, 1996

French Literature
University of Pennsylvania, 1988–89

LICENCES

New York Board of Education EAL Teaching Licence

California Adult Education Teaching Licence

LANGUAGES

English
native

Français

Italiano

CONTACT

Building or validating a speech model?

Tell me about the data, the task and the timeline, and I'll suggest how human evaluation can fit.

EMAIL
purpleflowers112@gmail.com

BASED IN
Paris · remote worldwide