Supplementary Note 1: Translations of the "Overall impression" question and MOS scale of the speech rating experiment in Part I.

English

Overall impression – How do you rate the quality of the speech of what you just heard? (ignore the noise)

- Excellent
- Good
- Fair
- Poor
- Very Poor

Dutch

Algehele indruk - Hoe beoordeelt u de kwaliteit van de spraak van wat u zojuist hebt gehoord? (negeer de ruis)

- Uitstekend
- Goed
- Redelijk
- Matig
- Slecht

French

Impression générale - Comment évalueriez-vous la qualité de la parole de ce que vous venez d'entendre? (ignorez le bruit)

- Excellente
- Bonne
- Passable
- Médiocre
- Mauvaise

Spanish

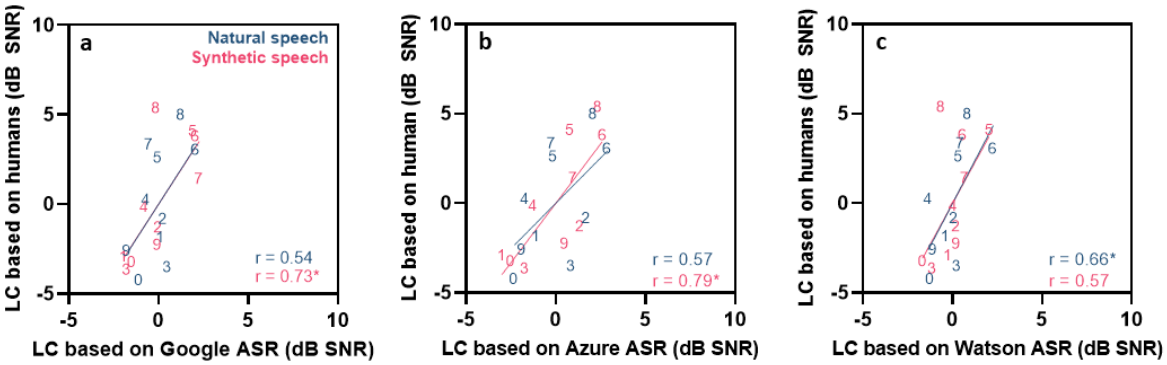
Impresión general - ¿Cómo calificaría la calidad del discurso que acaba de escuchar? (ignore el ruido).

- Excelente
- Buena
- Regular
- Mediocre
- Mala

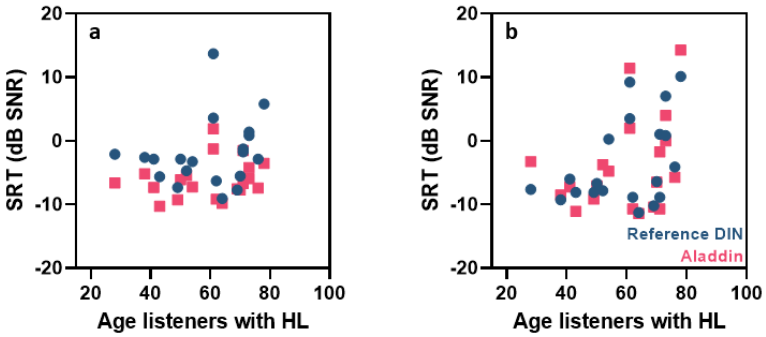
Mandarin

总体印象 - **您如何**评价刚刚听到的语音质量？（**忽略背景噪音**）

- **很好**
- **好**
- **一般**
- **差**
- **很差**



Supplementary Figure 1: Human-based vs. ASR-based level corrections (LC). Graphs represent the LC of the natural male (blue) and synthetic female (pink) Dutch digits derived from the listening experiment with human listeners from Part I, as a function of the LC derived from Google ASR (**a**), the Microsoft Azure ASR (**b**) and the IBM Watson ASR (**c**) (Part III). Three significant correlations ($p < 0.05$) were found, indicated by *.



Supplementary Figure 2: Scatterplots of SRTs as a function of age for the DIN and Aladdin test (Part IV). Scores of the Dutch tests are presented in (**a**) and of the English tests in (**b**). Participants aged 60 and above displayed significantly greater SRT variability than listeners under 60 for the English reference DIN (independent t-test, equal variances not assumed, t(15.168) = -2.082 , *p = 0.05*).