Supplemental Material S1.

SI Methods

Experimental Protocol

Continuous EEG was collected from Ag/AgCl electrodes at Cz, M1 (left mastoid) and M2 (right mastoid) at a 20 kHz sampling rate (Compumedics, Australia) with CPz as a reference and Fpz as a ground. During EEG recording the inter-electrode impedances were maintained at less than or equal to 1 kΩ. Cz data were re-referenced offline with the average of two mastoids for the subsequent analysis. The analysis was performed on one electrode (Cz) in MATLAB 9.2 (R2017a) with custom-made scripts using built-in functions of the Signal Processing Toolbox and the Statistics and Machine Learning Toolbox. Two different analysis pipelines were used for FFR and LLR extraction from the same raw EEG data. For FFR extraction, the data were filtered with an 8-order Butterworth 80-1500 Hz bandpass filter. The epochs were extracted around the stimulus onset time from -50 to 225 ms using a ± 25 µV artifact rejection criterion after baseline-correction at the 50-ms pre-stimulus interval. For LLR extraction, the data were filtered with a 4-order Butterworth .1-40 Hz bandpass filter and down-sampled to 3 kHz. Other LLR studies have also used this frequency range (e.g., Bishop, Hardiman, Uwer, & Von Suchodoletz, 2007; Paquette et al., 2015; Polich, Aung, & Dalessio, 1988). The epochs were extracted around the stimulus latency time from -100 to 500 ms using a ± 100 µV artifact rejection criterion after baseline-correction at the 100-ms pre-stimulus interval. Seven percent of all EEG recordings were considered as unsuccessful due to the large amount of artifacts.

Experimental Stimuli

Three maximally distinctive native and non-native tones that have been used before (Maggu, Zong, Law, & Wong, 2018) were chosen as the speech stimuli based, attesting to their validity for obtaining robust FFR responses. The stimuli were spoken by a phonetically-trained female speaker, and were duration and amplitude normalized. The Fundamental Frequency (F0) ranges for the Native /ga2/, Native /ga4/ and Non-Native /ga3/ were respectively 182-278 Hz, 152-187 Hz, and 142-177 Hz. These three tones are described as 25, 21, and 214, respectively, according to Chao's nomenclature (Chao, 1930). The stimuli were presented to the participants via Audio CPT module of STIM2 (Compumedics, Australia). One thousand repetitions for each of the three tones were delivered in three separate blocks to both ears of the participant with fixed inter-stimulus interval (ISI) and alternating polarity. The presentation rate was fixed at either ~1.5 stimulus or ~4 stimuli per second, resulting in one of two possible ISIs for each participant; consequently Stimulus-onset-asynchrony (SOA, i.e., ISI plus stimulus duration) was included in classification features (see below).

EEG Measures (Short and Long-Latency Measures)

The 69 EEG measures are derived from 23 measures for each of the three speech stimuli, including 17 short- and 4 long-latency measures as well as 2 general EEG measures. These measures are standard for auditory-evoked EEG (Akhoun et al., 2008; Anderson, Parbery-Clark, White-Schwoch, & Kraus, 2015; Krishnan, Xu, Gandour, & Cariani, 2004; Näätänen & Picton, 1987; Russo, Nicol, Musacchia, & Kraus, 2004; Skoe & Kraus, 2010; Skoe, Krizman, Anderson, & Kraus, 2015).

1. Short-Latency measures
    1.1. Number of good FFR trials was defined as the number of trials after the procedure of artifact rejection that removed every trial with values outside ±25 µV range in case of FFR.
    1.2. FFR Signal-to-noise ratio was calculated as the dB-transformed ratio of the root-mean-square (RMS) power of the 170-ms post-stimulus to 50-ms pre-stimulus interval of the FFR waveform.
    1.3. Noise RMS was the RMS of the 50-ms pre-stimulus interval of the FFR waveform.

1.4. Lower-range Fast Fourier transform (FFT) power was calculated as mean over 120-260-Hz interval of the whole-epoch power spectrum of the averaged across trials FFR in dBmV. This range corresponds to the fundamental frequency (F_0) a.k.a. first harmonic of our stimuli.

1.5. Middle range FFT power was calculated as mean over 260-750-Hz interval of the whole-epoch power spectrum of the averaged across trials FFR in dBmV. This range corresponds to the second and third harmonics of our stimuli.

1.6. Higher-range FFT power was calculated as mean over 750-1200-Hz interval of the whole-epoch power spectrum of the averaged across trials FFR in dBmV. This range corresponds to the fourth and higher harmonics of our stimuli.

1.7. Lower-range Inter-trial phase coherence (ITPC) was calculated as mean over 120-260-Hz interval of whole-epoch FFR ITPC. FFR ITPC was calculated as an absolute value of the averaged across single-trials FFT that were normalized by their absolute values.

1.8. Middle -range Inter-trial phase coherence (ITPC) was calculated as mean over 260-750-Hz interval of whole-epoch FFR ITPC.

1.9. Higher-range Inter-trial phase coherence (ITPC) was calculated as mean over 750-1200-Hz interval of whole-epoch FFR ITPC.

1.10. Maximal ITPC was the maximal value across the ITPC periodogram. The ITPC periodogram was obtained by combining FFTs in sliding 50-ms Hamming windows in 2-ms steps along the FFR epoch.

1.11. Pitch Strength was calculated as the maximum of the whole-epoch autocorrelation waveform excluding the initial peak at 0 ms delay (Fig. 1C).

1.12. Pitch Tracking was calculated as the correlation between the neural pitch and the auditory stimulus pitch. Neural pitch was extracted as the contour connecting the frequencies of the maxima of FFR autocorrelogram at each time point. Auditory stimulus pitch was a contour connecting the maxima of the stimulus autocorrelogram. Autocorrelograms were built by combining autocorrelation curves from the sliding 50-ms windows moved by a 2-ms step along the FFR or stimulus waveform.

1.13. Pitch Error was calculated as Euclidean distance between the neural pitch and the auditory stimulus pitch (see 1.12 for definitions of neural and auditory stimulus pitch).

1.14. FFR Response Consistency was calculated as a Fisher-transformed averaged correlation between the FFRs from two halves of the randomly split data after 300 iterations.

1.15. Stimulus fidelity was a maximum of cross-correlation between the auditory stimulus and FFR waveforms.

1.16. Stimulus-response delay (also known as neural lag) was the time between the stimulus onset and the maximum of cross-correlation between the auditory stimulus and FFR waveforms.

1.17. FFR peak amplitude was calculated as a global maximum in the 3-25 ms post-stimulus interval on the time domain FFR waveform.

2. Long-Latency measures

2.1. Number of good LLR trials was defined as the number of trials after the procedure of artifact rejection that removed every trial with values outside ± 100 µV in case of LLR. The number of good LLR trials was additionally capped at 600 to avoid the excessive habituation typical for LLR.

2.2. LLR Signal-to-noise ratio was calculated as the dB-transformed ratio of the RMS power of the 500-ms post-stimulus to 100-0 pre-stimulus interval of the LLR waveform (Fig. 2B).

2.3. LLR peak amplitude was calculated as the local maximum on the 500-ms post-stimulus interval on the time-domain LLR waveform.

2.4. LLR latency was calculated as latency of the local maximum on the 500-ms post-stimulus interval on the time-domain LLR waveform.

3. General EEG measures
   3.1. Stimulus-onset-asynchrony (SOA) is time between the onsets of the two adjacent stimuli in the sequence. We used two types of protocols, one with SOA 675 ms and another with SOA 268 ms that correspond to the inter-stimulus intervals of 500 and 90 ms.
   3.2. The total number of trials was usually 1000 per tone, but there were cases that deviated from this number due to unforeseen circumstances (e.g., when a child woke up before the completion of a stimulation block).

Model construction procedures (Fig S3). The raw scores collected from the MCDI were converted into percentile ranks based on the child's age and sex. In a two-way classification, participants were classified into < or > 25th %tile, and for a three-way classification, < 25th, 25th – 75th, or > 75th %tile. Support vector machine (Boser, Guyon, & Vapnik, 1992) was chosen over other machine-learning techniques as it has been used successfully in predicting children's speech outcomes (Feng et al., 2018). The following SVM training parameters were chosen in LIBSVM package (Chang & Lin, 2001): C-SVC type of SVM, radial basis function kernel, gamma of the kernel 1/number of features, cost 100, no shrinking. Note that although the original SVM algorithm was developed for binary classification, the LIBSVM package implements multi-class classification with a "one-against-one" method that we used in case of 3-way classification (Hsu & Lin, 2002). The classification was cross-validated with a 10-fold cross-validation procedure. At each fold of the cross-validation, we trained our SVM classifier on 90% of the data and predicted the outcome with the remaining 10% of the data. Stacking the ten outcome sets together produced a predicted-labels vector that we compared to the true labels. Accuracy was calculated as the percentage of the correctly classified labels to the total number of labels. Sensitivity was calculated as the proportion of the subjects correctly classified as low-performers (< 25th %tile) to the total number of the subjects classified as low-performers. Specificity was calculated as a proportion of the subjects correctly classified as higher-performers (> 25th %tile) to the total number of subjects classified as higher performers. To obtain the receiver-operating curve (ROC), we repeated the training and testing procedure with the same SVM parameters, except that epsilon-SVR was used instead of C-SVC as the model training type. This was done to obtain a smoother ROC. The ROC and its area under the curve were calculated with MATLAB function *perfcurve.* In order to obtain statistical significance, the response predictor matrix was put through 10000 iterations of bootstrapping and permutation. At each iteration we randomly re-sampled the subjects with replacement (MATLAB function *randi*). This kind of resampling with replacement defines bootstrapping in the statistical literature (Efron, 1979). We then randomly permuted the outcomes within the matrix so that the predictors no longer corresponded to the outcomes. Each of the two matrices, bootstrap and permutation, was trained and tested as described above. After 10000 iterations of bootstrapping, permutation and cross-validation, we obtained eight distributions of 10000 values for 4 classification quality parameters (accuracy, sensitivity, specificity and AUC) and for real (bootstrapped) and permuted labels of the subjects. The significance level (p-value) and thus the success of classification was calculated as a proportion of the permuted distribution that fell above the median of the real distribution. If the p-value was below .05, we concluded that the classification results were above chance level. This procedure is a non-parametric analog of a one-tailed t-test. The same procedure was applied for comparison of any other prediction models. It illustrates the advantage of applying bootstrapping in combination with permutation statistics over permutation alone.

Table S2 shows the exact number of subjects falling into the three categories (< 25th %tile, 25th–75th %tile, and > 75th %tile) of performance for each measure of the MCDI. In general, the obtained distribution deviated from the expected distribution (25%, 50% and 25% for the three categories respectively). This for the most part was due to fewer participants in the middle (25th–75th %tile) category. Our study is a community study with no pre-selection criteria other than those stated in the main text. It is possible that those parents who suspected their children to have language difficulties or those who thought their child might excel were more interested in our study.

SI Results

Models with SES as an additional non-neural feature

Socioeconomic status (SES) was measured in this study from all but 3 participant families. We calculated SES as Hollingshead four-factor index (Hollingshead, 2011):

$$SES = 3/2(\mathit{Maternal\ Education} + \mathit{Paternal\ Education}) + 5/2(\mathit{Maternal\ Occupation} + \mathit{Paternal\ Occupation})$$

where education is a number between 1 and 7 and occupation is a number between 1 and 9. In the event that one of the parents was not currently working, the other parent’s occupation alone was entered into the formula. In order to maximize our sample size, the SES value was replaced by the mean for the three participating families with missing SES info. Replacement by mean as an instance of multiple imputation is preferable to complete case analysis for missing data at random situations like ours (Hughes, Heron, Sterne, & Tilling, 2019). SES is an important correlate of child development at the group level (Fernald, Marchman, & Weisleder, 2013), though we are not aware of any research studies that have demonstrated its reliability in making predictions at the individual child level. Nevertheless, we constructed a set of predictive models using SES as an additional non-neural measures, and found that it did not significantly enhance the precision of the models without SES as a measure (Fig. S4).

References

Akhoun, I., Gallégo, S., Moulin, A., Ménard, M., Veuillet, E., Berger-Vachon, C., … Thai-Van, H. (2008). The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme /ba/ in normal-hearing adults. *Clinical Neurophysiology*. https://doi.org/10.1016/j.clinph.2007.12.010

Anderson, S., Parbery-Clark, A., White-Schwoch, T., & Kraus, N. (2015). Development of subcortical speech representation in human infants. *The Journal of the Acoustical Society of America*, *137*(6), 3346–3355. https://doi.org/10.1121/1.4921032

Bishop, D. V. M., Hardiman, M., Uwer, R., & Von Suchodoletz, W. (2007). Maturation of the long-latency auditory ERP: Step function changes at start and end of adolescence. *Developmental Science*, *10*(5), 565–575. https://doi.org/10.1111/j.1467-7687.2007.00619.x

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT ’92* (pp. 144–152). New York, New York, USA: ACM Press. https://doi.org/10.1145/130385.130401

Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: a library for support vector machines. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chao, Y. R. (1930). ə sistim əv “toun-letəz.” *Le Maître Phonétique*, *8*(30), 24–27.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Feng, G., Ingvalson, E. M., Grieco-Calub, T. M., Roberts, M. Y., Ryan, M. E., Birmingham, P., … Wong, P. C. M. (2018). Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(5), E1022–E1031. https://doi.org/10.1073/pnas.1717603115

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. https://doi.org/10.1111/desc.12019

Hollingshead, A. B. (2011). Four factor index of social status (Unpublished Working Paper, 1975). *Yale Journal of Sociology*, *8*, 21–52. Retrieved from http://elsinore.cis.yale.edu/sociology/yjs/yjs_fall_2011.pdf#page=21

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. https://doi.org/10.1109/72.991427

Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, *48*(4), 1294–1304. https://doi.org/10.1093/ije/dyz032

Krishnan, A., Xu, Y., Gandour, J. T., & Cariani, P. A. (2004). Human frequency-following response: representation of pitch contours in Chinese tones. *Hearing Research*, *189*(1–2), 1–12. https://doi.org/10.1016/S0378-5955(03)00402-7

Maggu, A. R., Zong, W., Law, V., & Wong, P. C. M. (2018). Learning two tone languages enhances the brainstem encoding of lexical tones. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 1437–1441). https://doi.org/10.21437/Interspeech.2018-2130

Näätänen, R., & Picton, T. W. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, *24*(4), 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Paquette, N., Vannasing, P., Tremblay, J., Lefebvre, F., Roy, M. S., McKerral, M., … Gallagher, A. (2015). Early electrophysiological markers of atypical language processing in prematurely born infants. *Neuropsychologia*, *79*, 21–32. https://doi.org/10.1016/j.neuropsychologia.2015.10.021

Polich, J., Aung, M., & Dalessio, D. J. (1988). Long Latency Auditory Evoked Potentials: Intensity, Inter-Stimulus Interval, and Habituation. *Electroencephalography and Clinical Neurophysiology*, *79*(5), S1. https://doi.org/10.1016/0013-4694(91)90210-u

Russo, N. M., Nicol, T., Musacchia, G., & Kraus, N. (2004). Brainstem responses to speech syllables. *Clinical Neurophysiology*, *115*(9), 2021–2030. https://doi.org/10.1016/j.clinph.2004.04.003

Skoe, E., & Kraus, N. (2010). Auditory brainstem reponse to complex sounds : a tutorial. *Ear and HearingHear*, *31*(3), 302–324. https://doi.org/10.1097/AUD.0b013e3181cdb272.Auditory

Skoe, E., Krizman, J., Anderson, S., & Kraus, N. (2015). Stability and Plasticity of Auditory Brainstem Function Across the Lifespan. *Cerebral Cortex*, *25*(6), 1415–1426. https://doi.org/10.1093/cercor/bht311

Table S1. Summary of participants' demographic information.

	M	Range
Gestational Age(weeks)	38.8	34-41
Birth Weight(kg)	3.1	2.18-4.25
Sex(M/F)	61/57	-
Age of EEG Testing(months)	3.8	.8-12.4
Age of Outcome Testing(months)	12.5	8-18
Age gap between EEG and Outcome Testing (months)	8.7	2.8-15.8

Table S2. The distribution of subjects across the percentiles of the normed MCDI outcomes. χ^2 statistics confirmed that for most measures, the obtained distribution deviates from the expected distribution.

	N subjects by percentile				
MCDI Outcome	> 25%	25% < < 75%	> 75%	χ^2	*p*
Early Gestures	41	43	34	9.2	.0099
Later Gestures	64	31	23	53.1	$3 \cdot 10^{-12}$
Vocabulary Comprehension	44	38	36	15.7	.0004
Vocabulary Production	36	59	23	2.4	.29