Supplementary Materials

Contents

I. FEM OF THE IMPLANTED COCHLEA: DETAILED DESCRIPTION OF THE CONSTRUCTION OF THE FEM

A simplified cross-section of the cochlea, including the scala tympani, scala vestibuli, scala media, basilar membrane, Reisner's membrane, stria vascularis, and the osseous spiral lamina, was extruded along a parametric curve that defined the spiral of the cochlea [36]-[40]:

$$R(\theta) = A\, e^{-B\,\theta} + C\, e^{-D\,\theta}$$

$$A = 3.457, B = 0.0382, C = 3.746, D = 0.0013 \qquad (1)$$

$$z = 0.00367 * (\theta - 10.3)$$

R is the radius of the cochlea (measured from the tip of the osseous spiral lamina), θ is the angle in degrees around the central axis, and *A*, *B*, *C*, and *D* are constants used to best fit the piecewise function defined in Clark et al (2011) [39]. Note, the double exponential fit was used rather than the original piecewise function because faces cannot be extruded along discontinuous piecewise functions. The function defined in Clark et al (2011) [39] was derived from the combined anatomical data from a total of 30 CI recipients and 34 cadaver temporal bones [36],[37]. The variable *z* represents the height of the spiral, which changes linearly as a function of the angle θ [38] .

A tapering function (Equation 2) was applied to the cross section according to data from Clark et al (2011) [39], which showed that the area of the cross section of the cochlea reduced from approximately 3 mm^2 at the base to 1 mm^2 at the apex:

$$S = Fe^{-Gx} \qquad (2)$$

S is the scaling factor that modifies the cross sectional area of the cochlea as a function of *x* (in mm) along the spiral. *F* and *G* are constants that fit an exponential function to the tapering data measured in humans, and are set to 0.9667 and -0.0332, respectively [37],[39],[40].

The resulting cochlear spiral was embedded in a sphere representing temporal bone with a radius of 5 cm, and a cone was placed in the center of the sphere to represent the modiolus. The helicotrema, which connects the scala vestibuli to the scala tympani, was also modelled.

Conductivities of the different materials, and the sources where the information was obtained, are shown in Supplementary Table 1. Here, for model simplification, purely resistive materials were used, i.e., the conductivities were real numbers.

Several electrode arrays can be implemented in the FEM model, including lateral wall and perimodiolar electrode arrays, which will be analyzed and reported in subsequent reports. The number of electrodes, size of the electrodes, and spacing between the electrodes can be adjusted to model different implants. For the analyses in this paper, a 16-electrode "generic" array was used, with an electrode size of 0.3 mm and a 0.75 mm spacing between electrodes. These electrode dimensions and spacings were averaged between the Helix (Advanced Bionics), Freedom 24RE (Cochlear, Inc), and CI522 (Cochlear, Inc) electrode arrays.

To validate voltage spread measurements from the FEM, transimpedance matrices were compared between the model and CI listeners [41]. In Figure 1, the CI data represents the

mean and standard deviation of voltage spread calculated from the transimpedance matrices for seven CI listeners, each with a 22-electrode lateral wall array. Note, for this validation, a 22-electrode array was used in the FEM. Voltage spread was calculated from the transimpedance matrices by multiplying the measured impedance between electrodes by the stimulation current of 125 µA, which was the current level used to measure transimpedance matrices in the CI listener data. Paired t-test were performed for every combination of stimulating and measurement electrode. In 450 out of 462 comparisons, no significant difference was found.

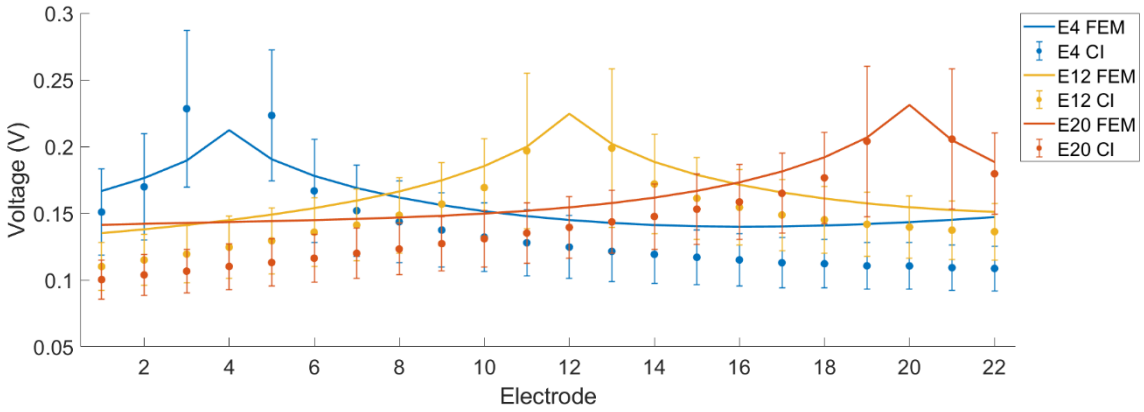

Figure 1. Comparison of voltage spread between the finite element model (FEM) and CI listeners on electrodes 4, 12 and 20 (E4, E12, E20). CI user data is averaged over 7 participants, and error bars represent +/- 1 standard deviation.

II. Computational Model of the Auditory Nerve: Phenomenological Model Description

Below is a description of the computational model of the auditory nerve, including phenomenological models of refractoriness, adaptation, and temporal integration:

A. Refractoriness

To model the absolute and relative refractory periods, the function defined by Bruce et al (1999) [17] was used (Equation 2).

$$\frac{(V_{\text{Thresh}} + V_{\text{Ref}})}{V_{\text{Thresh}}} = \begin{cases} \infty & 0 \le t \le 0.7\ ms \\ 1 + 0.97e^{-(t-0.7ms)/1.32\ ms} & 0.7\ ms < t \le 20\ ms \\ 1 & t > 20\ ms \end{cases} \quad (4)$$

This equation describes the amount that the neural activation threshold (V_{Thresh}, µV) shifts according to the time since the last spike (t, ms). There is an absolute refractory period of 0.7 ms, where no spike can be initiated, followed by a relative refractory period from 0.7 to 20 ms, where the threshold exponentially decreases from double V_{Thresh} to V_{Thresh} with a time constant of 1.32 ms.

B. Adaptation

To model adaptation, the multi-exponential fit to power-law adaptation defined by van Gendt et al (2020) [45] was used (Equation 3).

$$I_{\text{Adapt}}(t) = \alpha \sum_i 0.72\, e^{(t-t_i)/23\text{ ms}} + 0.26\, e^{(t-t_i)/212\text{ ms}} \qquad (5)$$

In this equation, I_{Adapt} represents the amount that the threshold current (μA) is increased based on the spiking history of a particular neuron. The variable α is the adaptation constant, which is set to 1 ± 0.6% of the unadjusted firing threshold for a neuron, in μA. The index *i* iterates over all previous spikes for a neuron through the duration of a sentence, and t_i is the time since spike *i*. Van Gendt et al (2020) [45] also used their model in conjunction with the Bruce et al (1999) [17] model of refractoriness.

C. Temporal Integration

The next step in the computational model of the auditory nerve is to convolve the spike trains of each neuron with a temporal integration window [46][47], and then to resample from 20000 Hz to 500 Hz. The temporal integration window consists of three exponential functions, two which describe forward masking and one which describes backward masking.

$$W(t) = \begin{cases} (1-w)e^{\frac{t}{\tau_{b1}}} + we^{\frac{t}{\tau_{b2}}}, & t < 0 \\ e^{-\frac{t}{\tau_a}}, & t \geq 0 \end{cases} \qquad (6)$$

W(t) is the weight applied to a sample according to time, t. The weighting variable, w, controls the weights of the long and short time constants (τ_{b1} and τ_{b2} , respectively) for forward masking, and τ_a is the time constant for backward masking. Parameters were set according to McKay et al (2013) [47] and Oxenham et al (2001) [46] to τ_a = 3.5 ms, τ_{b1} = 4.6 ms, τ_{b2} = 16.6 ms, and w = 0.17. The temporal integration mechanism has been used to describe many temporal processing phenomena, including amplitude modulation detection [47]-[50], loudness of time-varying stimuli [51], and effects of stimulation pulse rate [52]. In our model, the temporal integration window also provides a practical benefit, in that it allows us to downsample the signal, improving training efficiency for the neural network.

Finally, spatial summation is applied, in which the 1500 simulated auditory neurons are grouped with their adjacent neurons into 150 groups of 10 neurons each. The resulting neurograms thus have 150 frequency channels given by 150 neuron "bundles" whose summed activity is sampled at 500 Hz temporally.

III. Validation of the Computational Model of the Auditory Nerve

To validate the computational model of the auditory nerve, comparisons were made to electrophysiological animal data [57][58], repeating the comparisons made in van Gendt et al (2016)[45]. Spike rate growth functions were compared to Javel and Shepherd (1987)[57] feline data,and spiking patterns were simulated for pulse rates of 100, 200, 300, 400, 600, and 800 pps with a phase duration of 200 us. The overall level was expected to be different, because the Javel and Shepherd (1987)[57] data were measured in an anesthetized cat, with a different stimulating electrode and recording electrode setup than our model. The model replicates the slope of the spike rate growth function, with the discharge rate of the 600 pps stimulus reaching a maximum discharge rate after an approximately 6 dB increase in input current, and the 800 pps stimulus reaching a maximum discharge rate after an increase of 6-7

dB. In both the experimental and model data, an irregularity in the spike rate growth function can be seen at pulse rates above 300 pps at approximately half the stimulus pulse rate, and is due to the interaction between refractory and adaptation effects. The 100 and 200 pps stimulus reach an asymptote more quickly, and do not exhibit the same irregularity as the higher rate responses, consistent with the animal data.

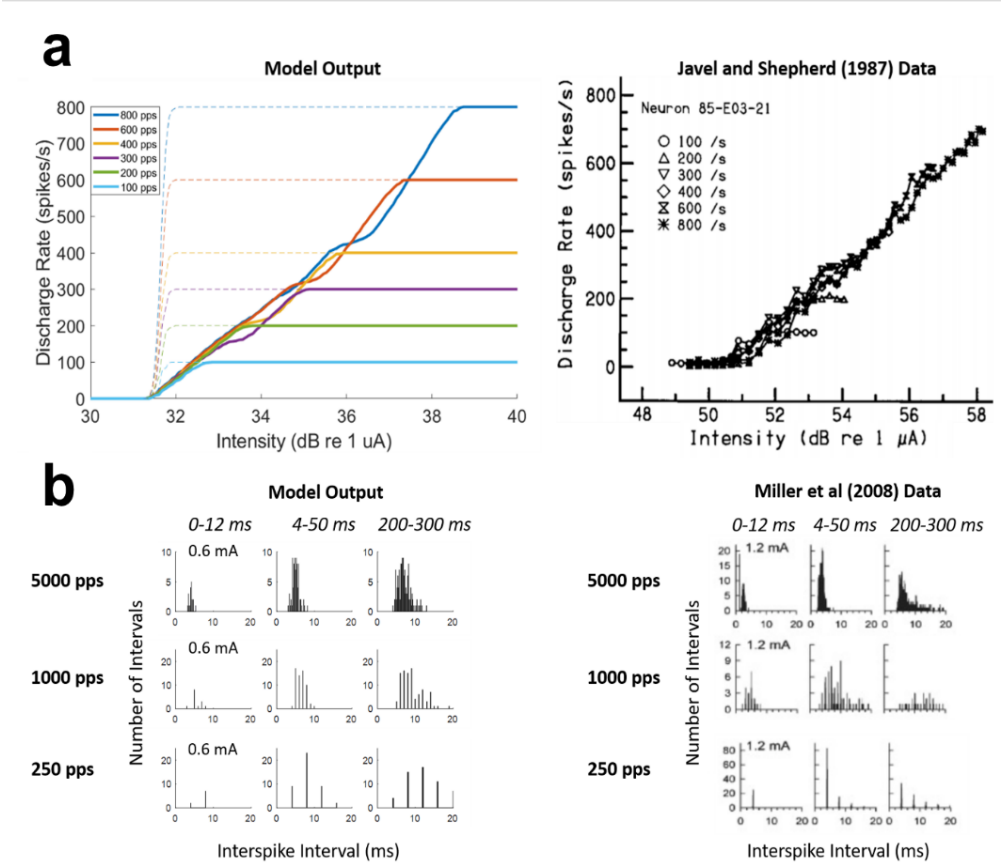

Figure 2. a. Comparison of spike rate growth functions for the model and for feline data from Javel and Shepherd (1987). The dotted lines in the left panel demonstrate spike rate growth functions when refractoriness and adaptation are disabled in the model. b.Comparison of interspike interval histograms between the model and feline data from Miller et al (2008).

Interspike interval histograms were compared to feline data from Miller et al (2008)[58]. Spiking patterns were evaluated for a single neuron in response to pulse trains of 250, 1000, and 5000 pps, and interspike intervals were recorded for time regions of 0-12 ms, 4-50 ms, and 200-300 ms.

The model output shows the combined results of 10 trials for each pulse rate. The model replicates the small distribution of short interspike intervals at the onset of the pulse train (0-12 ms), and the gradually increasing average and standard deviation of interspike intervals in the 4-50 and 200-300 ms time windows as the neuron becomes less responsive. Similar to the spike rate growth validation, the overall levels were expected to be different between the model and the animal data, because the animal data was measured in anesthetised cats with different stimulating and recording electrode setups than our model.

IV. Development of the ASR

The structure of the ASR system was chosen with three goals in mind: (1) Representing the processing in the brain to some extent, (2) fast training, (3) performance close to state of the art. To achieve the first goal, we split the system hierarchically in two parts: (a) a causal neural network that accumulates information for the present phoneme based on the immediate history of the signal, and (b) a non-causal neural network that makes a decision about a phoneme at a given frame based on the information a few hundred milliseconds (and thus a few phonemes) before and after the present phoneme. This roughly resembles the pathway towards higher cortical areas and involving short-term memory. Furthermore, the structure facilitates the second objective, fast training. In our setup the causal network has only two layers with 64 units each, and that is the most interesting part to be trained when varying the input features, for example to explore a different CI processing strategy. By comparison,

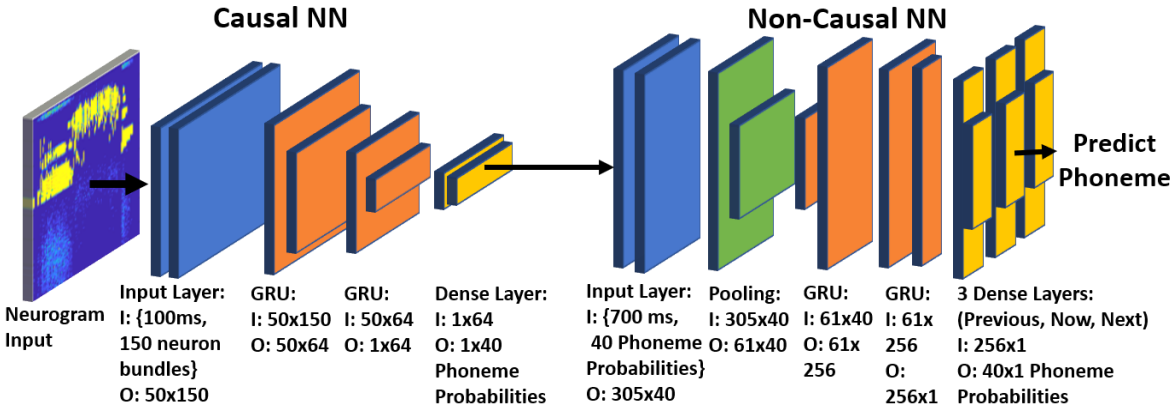

Figure 3. Architecture of the automatic speech recognition neural network.

Ravanelli et al. (2019) [53] also tested the performance of a GRU network, which was end to end, on TIMIT. According to the configuration files of their source code, this network had five bidirectional layers with 550 units each. The performance was slightly better, 17% phoneme error rate, and could be improved further to 14% phoneme error rate when doing more regularisation and using specifically designed (but remarkably also simpler) Light GRU units. However, for the purpose of the present paper we preferred less layers with less units for faster training, and found that 64 units in the causal network and 128 in the non-causal yielded close enough accuracy on the validation set. We did not follow an approach like wav2vec [54] that yielded a phoneme error rate of 8% but needs extensive pre-training on unlabelled data, which would need to be done for each setting of input features.

V. Analysis of Information Transmission: Detailed Description

Throughout this paper, the unit of information will be bits. To assist in the interpretation of bits as a unit, we will briefly go through an example using the voicing feature. The voicing feature is binary, because there are only two values that the voicing feature can take: voiced or unvoiced. Intuitively, if this binary feature is evenly distributed, we would expect this feature to convey 1 bit of information, which can take a value of 1 (voiced) or 0 (unvoiced). If we want to calculate the amount of voicing information that was transmitted, we can create a voicing submatrix. Because the voicing feature only has two values, the voicing submatrix will be 2 by 2. The unconditional voicing submatrix is shown in Table I:

Table I. An example stimulus-response matrix for the voicing feature.

		RESPONSE	
		Voiced	Unvoiced
STIMULUS	Voiced	3958	558
	Unvoiced	738	4788

The first cell corresponds to voiced stimuli with voiced responses, the second cell in the first row corresponds to voiced stimuli with unvoiced responses, and so on and so forth. The stimuli will be referred to as the rows *x*, and the responses will be referred to as the columns *y*. To calculate the total information (also known as entropy) for the voicing feature, the following equation can be used:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2(p(x_i)) = 0.992 \text{ bits}$$

where the probability of x_i , $p(x_i)$, refers to the probability of a voiced or unvoiced stimulus. To calculate the amount of information transmitted, we calculate the entropy in the stimulus, H(X), minus the entropy in the stimulus *given* the entropy in the response, the conditional entropy H(X|Y). For the above matrix, the conditional entropy H(X|Y) calculation is as follows:

$$H(X|Y) = -\sum_{i,j=1}^{n} P(x_i, y_j)\, log_2 \left(\frac{P(x_i, y_j)}{p(y_j)}\right) = 0.551 \text{ bits}$$

and the amount of information transmitted, or *mutual information*, is:

$$I(X;Y) = H(X) - H(X|Y) = 0.992 \text{ bits} - 0.551 \text{ bits} = 0.441 \text{ bits}$$

The same process is used to calculate information transmission for the other consonant and vowel features, but the size of the sub-matrix depends upon the number of options for a particular feature.

The three measures reported here for IT are Transmitted/Input, Transmitted/Total Information, and Proportion Correct. Transmitted/Input is the ratio of

transmitted information to the amount of information available for a particular feature.

$$\frac{Transmitted}{Input\ Information} = \frac{I(X;Y)}{H(X)}$$

Transmitted/Total Information is the ratio of transmitted information to the total information summed across all features, and shows the relative importance of particular features for identification:

$$\frac{Transmitted}{Total\ Information} = \frac{\sum_{F=1}^{N} I(X;Y)_F}{\sum_{F=1}^{N} H(X)_F}$$

where F represents each feature.

Proportion correct is simply percent accuracy, the ratio of correct responses to the total number of presentations for each feature.

VI. Probing Information Transmission Through The CI Signal Processing Pipeline

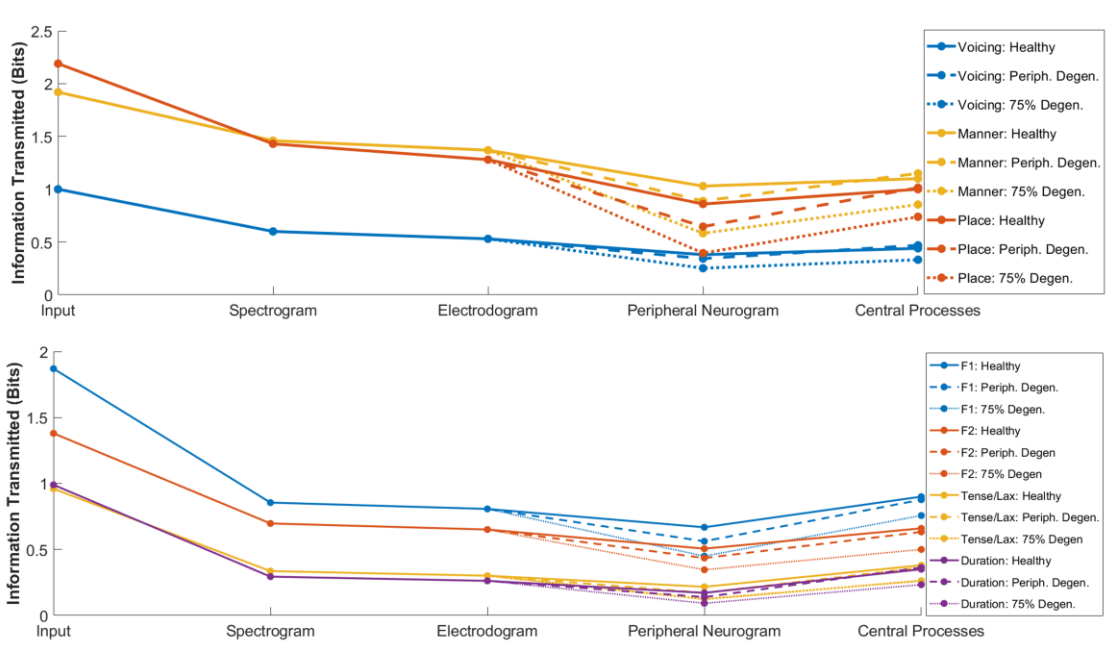

Figure 4. IT Results for different points in teh CI signal processing pipeline for consonants (upper panel) and vowels (lower panel). Solid lines, dashed lines, and dotted lines represent the healthy condition, peripheral degeneration condition, and the 75% degenerated condition, respectively.

VII. SUPPLEMENTARY TABLES

A. Supplementary Table I. Conductivities of different materials in the finite element model of the cochlea.

Anatomical Element	Conductivity (S/m)	Source
Perilymph	1.43	Baumann et al. (1997)[1]
Endolymph	1.68	Misrahy (1958)[2]
Stria Vascularis	0.0053	Briaire and Frijns (2000)[3]
Basilar Membrane	0.375	Finley et al (1990)[4]
Reissner's Membrane	0.0006	Finley et al (1990)[4]
Temporal Bone	0.016	Potrusil et al (2020)[5] , Malherbe et al (2015)[6]
Modiolar Bone	0.0334	Potrusil et al (2020)[5]
Platinum	9.4E6	Serway (1989)[7]
Silicone	0.001	Saba et al (2012)[8]

Sources

1. Baumann, S. B., Wozny, D. R., Kelly, S. K., & Meno, F. M. (1997). The electrical conductivity of human cerebrospinal fluid at body temperature. *IEEE transactions on biomedical engineering*, *44*(3), 220-223.
2. Misrahy, G. A., Hildreth, K. M., Shinabarger, E. W., & Gannon, W. J. (1958). Electrical properties of wall of endolymphatic space of the cochlea (guinea pig). *American Journal of Physiology-Legacy Content*, *194*(2), 396-402.
3. Briaire, Jeroen J., and Johan HM Frijns. "Field patterns in a 3D tapered spiral model of the electrically stimulated cochlea." *Hearing research* 148.1-2 (2000): 18-30.
4. Finley, C. C., Wilson, B. S., & White, M. W. (1990). Models of neural responsiveness to electrical stimulation. In *Cochlear implants* (pp. 55-96). Springer, New York, NY.
5. Potrusil, T., Heshmat, A., Sajedi, S., Wenger, C., Chacko, L. J., Glueckert, R., ... & Rattay, F. (2020). Finite element analysis and three-dimensional reconstruction of tonotopically aligned human auditory fiber pathways: a computational environment for modeling electrical stimulation by a cochlear implant based on micro-CT. *Hearing Research*, *393*, 108001.
6. Malherbe, Tiaan Krynauw, Tania Hanekom, and Johannes Jurgens Hanekom. "The effect of the resistive properties of bone on neural excitation and electric fields in cochlear implant models." *Hearing research* 327 (2015): 126-135.
7. Serway, Raymond A., and John W. Jewett. "Sound waves." *Physics: For Scientists and Engineers. 3rd ed. Philadelphia, Pa: Saunders College Publishing* (1990): 455-457.
8. Saba, R. (2012). *Cochlear implant modelling: stimulation and power consumption* (Doctoral dissertation, University of Southampton).

B. Supplementary Table II. Consonant feature matrix

Consonant (TIMIT symbol / IPA symbol)	**Voicing**	**Place**	**Manner**
b	Voiced	Bilabial	Plosive
d	Voiced	Alveolar	Plosive
g	Voiced	Velar	Plosive
p	Voiceless	Bilabial	Plosive
t	Voiceless	Alveolar	Plosive
k	Voiceless	Velar	Plosive
jh (dʒ)	Voiced	Postalveolar	Affricate
ch (tʃ)	Voiceless	Postalveolar	Affricate
s	Voiceless	Alveolar	Fricative
sh (ʃ)	Voiceless	Postalveolar	Fricative
z	Voiced	Alveolar	Fricative
f	Voiceless	Labiodental	Fricative
th (ϴ)	Voiceless	Dental	Fricative
v	Voiced	Labiodental	Fricative
dh (ð)	Voiced	Dental	Fricative
m	Voiced	Bilabial	Nasal
n	Voiced	Alveolar	Nasal
ng	Voiced	Velar	Nasal
l	Voiced	Alveolar	Lateral approximant
w	Voiced	Labial + Velar	Approximant
y (j)	Voiced	Palatal	Approximant
hh	Voiceless	Glottal	Fricative

C. Supplementary Table III. Vowel feature matrix.

Vowel (TIMIT symbol / IPA symbol)	**F1**	**F2**	**Tense/Lax**	**Duration**
er / ɝ (bird)	**1**	**1**	**1**	**1**
aa / ä (bott)	**0**	**1**	**1**	**0**
ah / ʌ (but)	**1**	**1**	**0**	**0**
uw / u: (boot)	**2**	**0**	**1**	**1**
uh / ʊ (foot)	**2**	**0**	**0**	**0**
ih / I (bit)	**2**	**2**	**0**	**0**
iy / i (he)	**2**	**2**	**1**	**1**
eh / ɛ (bet)	**1**	**2**	**0**	**0**
ae / æ (bat)	**0**	**2**	**0**	**0**
ey / eI (bay)	**1**	**2**	**1**	**1**
aw / aʊ (out)	**3**	**1**	**1**	**1**
ay / aI (bite)	**3**	**1**	**1**	**1**
oy / oI (oyster)	**3**	**0**	**1**	**1**
ow / oʊ (boat)	**1**	**0**	**1**	**1**