Phoneme Perception as Bayesian Inference with a Narrow-Tuned Multimodal Prior

Abstract

Phoneme Perception as Bayesian Inference with a Narrow-Tuned Multimodal Prior

Seng Bum Yoo

Brain and Cognitive Sciences

College of Natural Sciences

Seoul National University

Statistical inference well describes sensory perception: estimating true status of the world according to newly obtained sensory information and beliefs that are formed by previous experience. Beyond previous attempts of explaining perception by Bayesian framework, distribution of prior is inferred by fitting a Bayesian model to perceptual bias and variability exhibited by observers (Girshick et al., 2011). In current study, we inferred a prior that is being combined with given sensory information in phoneme perception, where the presence of a strong prior is expected.

For inferring the prior, subject performed two distinctive psychophysics experiments: identification and discrimination. The acoustic stimuli varied gradually along the spectrum encompassing three different stop consonant - /ba/, /da/, and /ga/. A significant component of model, which is prior, is estimated as mixture of three normal distribution having the means and variance of which reflect the centers and spread of phoneme stimuli that is most frequently heard by the listener in the past. Likelihood is similarly modeled as normal distribution except having its mean corresponding to given stimuli and variance identical to all types of stimuli. Only with a few numbers of free parameters, the hallmark features of phoneme perception are well explained simultaneously: ‘drastic change of selection category’ in identification task and ‘enhanced discriminability around boundaries of two phonemes’. Further in goodness of fit, our model implementing mixture normal surpassed a model with uniform prior distribution and matched with a model having non-parametric prior.

Suggested Bayesian model provides evidence that human phoneme perception requires a narrow-tuned multimodal prior whose peak exists at prototypical phoneme stimuli.

Keywords: Bayesian Inference, Categorical Perception, Phoneme

Student ID: 2013-22455

Contents

Figures

Introduction

Among the computations for speech perception, one foremost crucial step is to chop acoustic input streams into pre-lexical categories, called phoneme (Obleser and Eisner, 2009). The forming of a limited number of abstract phonological representations at an initial stage of speech processing can help address several fundamental problems in speech perception, including the 'invariance problem (Stevens and Blumstein, 1978; Perkell and Klatt, 1986)'- a task of accomplishing perceptual constancy in a high degree of variability in speech sensory input (Kraljic and Samuel, 2008). Perceptual invariance is critical evidence demonstrating the absence of one-to-one mapping between acoustic features and perceptual categories of speech sound (Hickock, 2012).

The perceptual invariance achieved with pre-lexical units has been discussed with opposing views for its explanation. One extreme description is the motor theory of speech perception, which diminishes the role of acoustic representation by speech signal (Liberman & Mattingly, 1985). Disruption of pre-motor area by Transcranial Magnetic Stimulation (TMS), in fact, influences categorical perception of phoneme (Meister et al., 2007). In contrast, neural response patterns of brain that is intensely organized along phonetic categories without demonstrating sensitivity for gradual acoustic variation within sensory area is counter-evidence for motor theory of speech perception (Chang et al., 2010).

The computational importance of phoneme representation demonstrated by perceptual invariance forced many models of speech perception to adopt, whether implicitly or explicitly, pre-lexical representations as primitive input to their lexical processing system (McClelland and Elman, 1986; Norris et al, 2000). In general, structural and functional properties of inputs can greatly constrain the way any given systems process those inputs to achieve their computational goals. Hence, understanding of pre-lexical representations has a fundamental significance on establishing models for speech perception. Regardless of the significance has been emphasized and many plausible hypotheses given for explaining perceptual variance achieved by pre-lexical units, only limited numbers of the mechanistic model incorporates phoneme representation in detail.

Among many qualities of phoneme, the prominent feature of phoneme that would be focused on when establishing mechanistic model is its inherent variability (Clayards et al, 2010). Phoneme inputs are intrinsically noisy due to variations in their origin (e.g., different speakers) or context (e.g.,

different preceding words or syllables), making it necessary for the brain to make probabilistic judgments. When human makes probabilistic judgments about external stimuli, the information of sensory input is probabilistically represented at the stage of encoding. In parallel, probabilistically represented knowledge accumulated via experience is combined with sensory information. The two information, sensory measurements and knowledge accumulated via experience, can be substituted as 'likelihood' and 'prior' in Bayesian formalism. At the following stage of decoding, 'posterior' is computed as consequence of two antecedents.

Reduced discriminability near prototypical phoneme, which is one of the hallmark behavioral features for phoneme perception, was captured with the model based on Bayesian framework (Feldman et al., 2009; Kuhl et al., 1991). Having statistical inference as underlying the rationale, their model captures perceptual warping by combining probabilistically represented sensory information and accumulated knowledge. To simplify the model, the representation of the likelihood and prior were parametrically approximated with Gaussian distribution having a mean at frequently exposed stimuli with particular variance. Their model suggested feasibility that Bayesian model can account for categorical perception reported from previous studies (Kuhl et al., 1995).

Though the model based on Bayesian framework captures significant aspect of phoneme perception with the plausible assumption, still certain improvement is required. Most of all, their study makes only prediction of result given by other studies with simulated data, but not fitting the model to own empirical data. Since the purpose is rough prediction rather than accurate description, noise characteristics or parameter values in the prior distribution are chosen for computational convenience. Hence, accurate description with precise model fitting is required to be validated. Furthermore, the prediction was separately made for data acquired from each different study. Thus, the model needs to be constrained properly by simultaneous fitting of data from each different study so that suggested model can be acquire validity for wide-range of tasks in phoneme perception.

The purpose of current research is to validate the mechanistic model explaining categorical perception of phoneme based on Bayesian framework. To acquire accurate description beyond simulation prediction, we acquired empirical data from two distinctive psychophysics experiments: classification and discrimination. Then we estimated a prior distribution and likelihood by fitting both of the components of the model to result obtained from behavioral experiment (Stocker & Simoncelli, 2006). Especially, for constraining the model more strictly, we simultaneously fitted components of the

Bayesian model to the result obtained from each types behavior. We further challenged the estimated prior distribution by introducing priors having different shapes, which includes uniform and non-parametrically generated prior distribution. By this approach, we aimed to vary systematic bias caused by prior distribution and figure out appropriate shape of prior in explaining categorical perception. Successful explanation of various behaviors provides direct evidence that humans behave according to the rules of Bayesian inference to perceive the variation of formant for phoneme perception. Especially, the simultaneous fitting disclosed the evident fact that narrow-tuned multimodal prior distribution plays crucial role for explaining categorical perception of phoneme with Bayesian inference.

Method

The Seoul National University review boards approved the experimental protocol, and the subjects gave their informed consent before participating experiment.

Subject Profile 33 subjects participated in screening audibility test with pure tone sound ranging from 500Hz to 4000Hz in 20dB. After audibility test, listeners performed practice trials on three prototypical syllable stimuli (/ba/, /da/, and /ga/), which were identified by a pilot test as those leading to the highest fraction of choice for each phoneme category. If the subjects did not correctly discriminate those prototypical stimuli more than 27 out of 30 trials (> 90%), they did not perform classification task. From the subject who showed categorical perception, 5 subjects participated to discrimination experiment. Within the five subjects, one subject was discarded since she could not discriminate any range of sound presented. Then the remaining participant performed adaptation experiment. All except one subject (the SBY, the author) were naive subjects.

Equipment and System Profile Discrimination task is performed with Psychtoolbox-3 (Brainard et al, 1997) in identical dark room. Auditory stimuli were presented through earphones (Etymotic Research ER-4B), with instructions being displayed on a monitor (HP LP2065). Adaptation experiment was conducted by running the E-Prime on an iMac computer using Window 7 OS implied by commercial software, bootcamp, in a quiet dark room. Listeners' manual responses were recorded with a numeric keypad (SAMSUNG SNK2000). The volume of the device was the identical for all the subjects. To prevent any unwanted visual factors from interfering with task performance or sound localization problem by ear plug movement while providing the listeners with pre-trial cues and post-trial feedback, listeners were asked to maintain their posture by stabilizing their head on a chin rest and fixate their gaze on a computer screen to be warned of an upcoming trial.

Frequency and Temporal Profile of Phoneme Stimuli Profiles of stimuli were modulated by MATLAB (Version 8.1, MathWork). Calculated stimuli frequency profiled are imported and generated into phonetic stimuli by the PRATT (freeware provided by Paul Boersma and David Weenink, Phonetic Sciences, University of Amsterdam, The Netherlands). With PRAAT software, we synthesized a set of 500ms voiced stop consonant-vowel syllable stimuli by specifying seven formant components. The each single stimuli used in three studies has identical temporal profiles, transition profile in five background harmonics (F0, F1, F4, F5, and F6), and the steady state frequencies. The

fundamental frequency (F0) was 132Hz at onset time (25ms) and fell to 120Hz in 40ms. The first harmonics (F1) started at 200Hz and monotonically increased, reaching to its steady-state frequency, 720Hz, in 50ms. The frequency values of the fourth, fifth, and sixth formants were remained identical at 3650, 4500, and 4900 Hz, respectively. While keeping these five background harmonics unchanged, we varied only the starting frequencies of the second (F2) and third (F3) harmonics according to achieve goals of individual studies (Figure 1a). Reason selecting two specific formants (F2 and F3) is because they contain large amount of the acoustical energy that is crucial in distinguishing stop-consonant in phoneme. By changing formant frequencies in the acoustical signal, one consonant can be perceived more or less acoustically similar to another consonant. The steady state frequencies of the F2 and F3 were 1240Hz and 2850Hz, respectively, and it took 40ms for the initial frequencies to change to those steady state frequencies. The particular set of constant parameters for the background harmonic components, steady state frequencies, and the transition times was chosen because it turned out best among alternative sets of harmonics for producing categorical phoneme perception in a pilot test. Stimuli were 500ms in duration and presented binaurally with peak amplitude of 80 dB for the syllables and 60 dB for the tones.

Classification Experiment Stimuli Generation We generated a cyclic spectrum of stimuli whose perceived sounds change gradually from '/ba/' to '/da/' to '/ga/' and then go back to '/ba/' by varying starting frequencies of the second (F2) and third (F3) harmonics (Figure 1a) (Steinschneider, 1995). In the pilot test, subjects were unable to categorize the stimulus profile identical to previous study. Thus, we have added 350hz to second and third formant of each stimulus. By this procedure, subjects were able to form categorical perception to the given stimuli. The twenty-one synthesized syllable phoneme stimuli was grouped into three distinctive pairs: /da/ vs /ba/, /ba/ vs /ga/, and /ga/ vs /da/. The order of grouped-phoneme pairs was pseudo-randomized using the Latin Square method. Purpose of pairing was to design experiment corresponding to two-alternative choices so the data can be fitted with sigmoidal psychometric curve. In each block, listeners performed the task within the seven neighboring stimuli that bridge between the two-prototypical phoneme stimuli corresponding to a given pair of alternative categories. Within single stimuli pair, seven neighboring stimuli were located equidistantly on two-dimensional space mentioned above. Individual block consisted of 20 trials per stimuli type, which result in 140 trials in single block. A type of block was repeated twice, resulting in 40 trials per

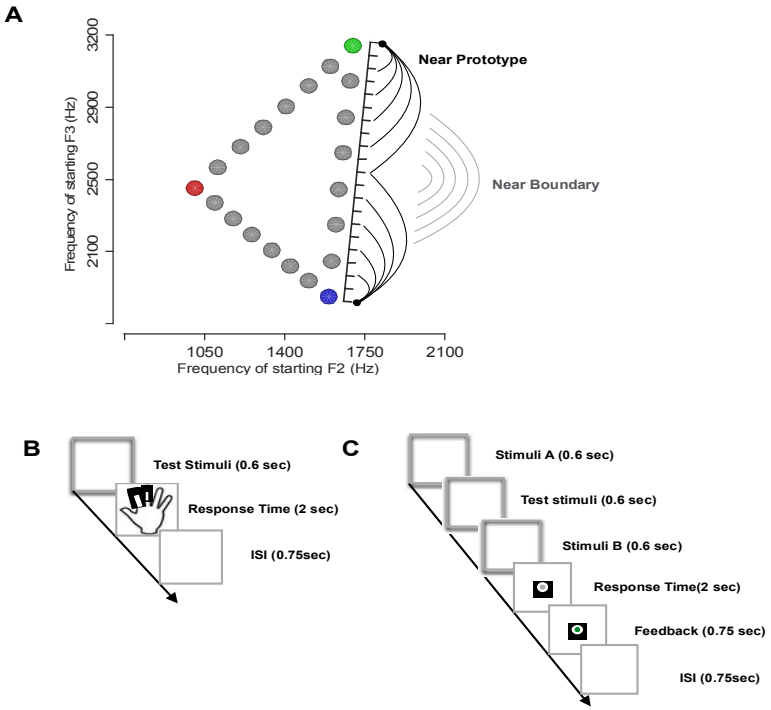

Figure 1. Stimulus and experiment apparatus. **a**) Stimuli profile used for classification and discrimination experiment. Space represents change in starting frequency of second and third formant frequency. Colored circle at each apex of triangle indicates prototypical phoneme frequency (/ba/: red, /da/: green, /ga/: blue). The six circles at each side of the triangle have identical distance between each other. At the range between /da/ and /ga/ prototype, example of discrimination stimuli is depicted. The reference stimuli, which means X stimuli in AXB matched-to-sample experiment paradigm, is located at three different conditions. The two conditions are located near each prototypical phoneme (near prototype: black), and the other one is located near boundary of two phoneme types (near boundary: gray). As the distance between prototypes of phoneme denoted as 1, all difference between the reference stimuli is given as a proportion. In each condition (near prototype and near boundary), two reference stimuli varied its distance from 0.1 to 0.5. Thus, applying the relative distances to all three conditions result in fifteen different stimuli delivery within a single range between two phonemes. This was replicated to all ranges three ranges (/ba-da/, /da-ga/, /ga-ba/). **b**) The classification experiment with 2-AFC paradigm. When the stimulus is given, subjects were required select one of the phoneme category as fast as possible. **c**) The discrimination experiment with 2-AFC paradigm. While stimuli are given, there is gray fixation dot that later are used as dot providing feedback for correctness.

stimulus and 840 trials in total. Once each block ended, subjects were allowed to have 60s of break time at the experiment site. Identical procedure was performed for two separate sessions in different days.

Classification Experiment Procedure The classification experiment was performed with sixteen subjects who met the criteria in the screening test. In the main experiments, listeners performed a two alternative forced choice task (2AFC), in which they heard a single syllable stimulus and classified it into one of the two pre-designated alternative phoneme categories (Figure 1b). Listeners made responses by pressing a button on a keypad using one of the three right-hand fingers (index for /ba/, middle for /da/, and ring for /ga/). To minimize response errors due to key-phoneme mis-assignment, an image of right-hand fingers was shown with the phoneme symbols that corresponded to type of the block. To control for speed-accuracy trade off, listeners were instructed to complete their responding within 2.0 s after stimulus offset and were given a "too late" warning in case not. Once the response is made, the each trial terminated and proceeded to next trial with having 750ms of inter-trial interval. Unlike the screening and practice trials, however, no feedback was provided in the main experiments.

Discrimination Experiment Stimulus Generation Frequency profiles of stimuli used both in training and main experiment of discrimination experiment were generated on range used in classification experiment (Figure 1a). The relative difference between stimuli pairs was scaled proportionally according to starting frequency profile difference between prototypical phonemes. For example, along the two-dimensional coordinate mentioned above, Euclidean distance of prototypical /ba/ to /da/ was assigned with value 1.0. Considering distance between /ba/ and /da/ as standard, distance between stimuli used were assigned proportionally. Identical method was used for generating stimuli in other phoneme ranges.

Discrimination Experiment Task Structure For discrimination experiment, AXB match-to-sample discrimination experiment design was used (Figure 1c). The task consists of three consecutively delivered stimuli. Among the stimuli, the reference stimuli are initial and last stimuli, and the test stimulus is the one delivered in between the reference stimuli. Subjects are required to decide test stimulus identical to which one of the reference stimuli and press corresponding keyboard. White fixation circle was located at the center of the screen with 1 degree in visual angle and subjects were asked to maintain fixation during the experiment. In the fixation circle, the gray inner circle with 0.5

degree in visual angle indicated availability of response after three consecutive phonemes were presented. Fixed 250ms of inter-stimuli intervals (ISI), which was tested empirically, was applied to avoid perceptual masking between stimuli and to allow clear distinction. Particular set of constant parameters including background harmonic components, steady state frequencies, transition times, and duration of single stimulus corresponded with other studies. Feedback indicating correctness and late with color change was provided immediately at the fixation circle after the response with the duration of 750ms. Response deadline was 2s and late feedback appeared in fixation circle when the decision was not made within response deadline. Trials were terminated once performance feedback was given and 750ms of inter-trial interval (ITI) was applied to provide clear sense of distinction between the trials. As single block ends, observers were required to take 30s of break and to relax in same posture within experimental spot.

Discrimination Experiment Procedure To make familiar with the task structure, 3 blocks of training session with identical delivery method was performed in each day before main discrimination experiment. For the training session, difficulty of the task was individually adjusted by applying 3down-1up adaptive staircase method. The minimum of 0.125 and maximum of 1.0 relative distance between reference stimuli was used. Step size in staircase method was 0.125, and thus the relative distance between reference stimuli used was 0.125, 0.25, 0.375, 0.50, 0.625, 0.75, 0.875, and 1.0. In each block of training session, single descending stair from maximum difference or ascending stair starting from minimum difference around arithmetic mean of two prototypical phonemes are applied separately. Numbers of the trials within a block was decided by 5 times reversal of stair, which was considered as flexible way to control numbers of trials used within the block in staircase method. In addition to termination rule above, maximum number of trial within a block was limited to 30 trials in training session. The main discrimination experiment adopted identical stimuli delivery method with training session except the stimuli profile and the numbers of trials in single block differs. Alike how stimuli were generated at training session, each stimuli step used in main discrimination experiment was denoted as proportional Euclidean distance between two prototypical phonemes. The minimum difference applied was 0.1 and maximum difference was 0.5, which generally yield approximately 100% percent correct of performance in training session. Within the range, 0.1 relative distances was single step of difference applied in main experiment. Each blocks consisted of 30 trials as fixed, and 20

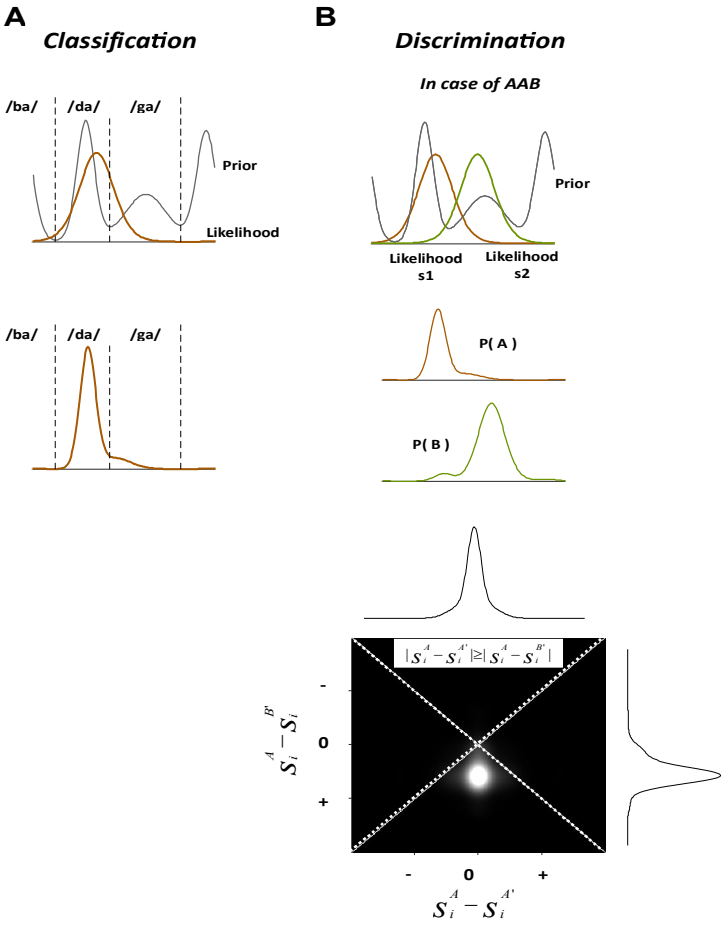

Figure 2. Computation of model for classification and discrimination of stimuli. **a**) Computation for reproducing classification experiment by model. By combining two probabilistic representations of prior (gray) and the likelihood (orange), the posterior distribution is generated (figure in the second row). By calculating the proportion of area belonging to a particular category, percentage of selecting a particular category is decided. The categorical boundary is the location where probability of prior is lowest in between the two peaks. **b**) Computation for reproducing discrimination experiment by model. Identically to classification experiment, posterior is generated as a consequence of two antecedents except there are two different types of likelihood generated according to different stimuli (orange and green). The three consecutively perceived stimuli are not being classified into a particular category. Instead, the posterior distribution is being directly compared. The procedure of comparing the posterior distribution is following: single value is sampled from individual posterior distribution and distances between the samples are compared. As consequence, s^A_i - s^X_i for distance between initial stimulus and middle stimulus is generated as well as s^B_i - s^X_i for distance between last stimulus and middle stimulus. The distances, which are represented as probability, is described as a joint probability matrix. By depicting the diagonal line (white dashed line), we can visualize which area indicates ($|s^A_i - s^X_i| < |s^B_i - s^X_i|$) and ($|s^A_i - s^X_i| > |s^B_i - s^X_i|$). According to the sequence of reference stimuli and types of test stimuli, the area dedicated for correct answers differ.

blocks performed per day, total 600 trials per day, and the individual subject was required to visit five times in total to secure sufficient numbers of observations. The reference stimuli were given with pseudo-randomized constant stimuli method. To prevent biased presentation of stimuli, all possible kinds of combination (AAB, ABB, BAA, BBA) of reference stimuli pair was assigned within every two blocks. The correct percentage for the each performed block was provided at screen in between the blocks, and total percentage correct till particular block was additionally presented.

Modeling Axis definition The initial frequencies of second and third formant are projected into two-dimensional coordinate having individual formant at each axis. By projecting into two-dimensional space, relative location of individual stimuli and Euclidean distance between generated stimuli can be estimated. According to formant frequency of stimuli, individual stimuli can be located in one of the three sides of triangle within the coordinate (Figure 1). By unfolding the triangle, relative position of each stimulus can be defined in single axis. The length of unfolded triangle was measured and each endpoint of was assigned to have -pi and pi. By this procedure, angular values along the cyclic one-dimension axis are assigned to each individual stimuli, and these values was used in model fitting procedure as single dimension.

Maximum likelihood The individual response for a particular stimulus is sampled from a population with particular probability distribution. Here, we assume a binomial distribution is underlying probability distribution since response of experiment was from two-alternative forced choice task. Two parameters that are crucial in a binomial distribution are the number of trials and probability of selecting that response. The parameter for trial number is substituted with the number of trials done for a particular stimulus. The parameter for probability is substituted with the percentage of selecting particular response in classification experiment. In the case of discrimination experiment, percentage for correct discrimination is used instead of the percentage of selecting particular response. From the parameters above, probability density function for the behavioral data is generated. Then we can compute how probable is the percentage value acquired from model within the probability distribution generated by behavioral data. We can acquire optimal parameter of the model for explaining behavioral data by maximizing the likelihood value acquired from individual stimulus.

Estimation of prior Particular position in stimuli feature dimension is experienced more frequently and forms category in phoneme perception (Figure 5a). In order to capture varying degree of exposure within categories, we assumed circular normal distribution with mean μ_p and variance σ_p assigned to each phoneme as prior distribution.

$$P(s) \propto \exp(\frac{1}{\sigma_p^2}\cos(s-\mu_p)-1)$$

In here, s indicates the stimulus dimension in one-axis. Since we have systematically aligned our phoneme stimulus in circular manner ranging from $-\pi$ to π, circular normal distribution was applied instead of simple Gaussian distribution. The mean of circular normal distribution, which has highest probability, is assumed to be position where individual observer has been exposed most frequently. In real situation, human observer experience internal noise or external variability like speaker even when they perceive within-category phonemes. To account those uncertainty experienced in phoneme perception, we have assigned variance as parameter for each phoneme. After describing circular normal distribution for each phoneme category, all distribution was merged as equation# and then was normalized so that area under prior integrated to 1.

$$P(s) \propto \exp\left(\frac{1}{\sigma_{p\_Ba}^2}\cos(s-\mu_{p\_Ba})-1\right)+\exp\left(\frac{1}{\sigma_{p\_Da}^2}\cos(s-\mu_{p\_Da})-1\right)$$

$$+\ \exp(\frac{1}{\sigma_{p\_Ga}^2}\cos(s-\mu_{p\_Ga})-1)$$

Prior distribution comparison In order to secure validity of model applying circular normal distribution as prior distribution, two additional probability distributions were compared. First consideration of varying shape of prior distribution was applying non-parametric cubic spline interpolation. We assumed non-parametric method would allow increased degree of freedom for the shape in comparison with parametric circular normal distribution. In way to generate probability distribution by cubic spline interpolation, we implied six coordinates in stimuli feature dimension. A few constraint applied were that it integrate to 1, that it be periodic with period 360 degrees in stimuli feature dimension, and that the prior secure to be smooth and continuous. To reflect three different phonemes in the shape of non-parametrically generated prior distribution, it is constrained to have three peaks. The other distribution used for validation was uniformly shaped prior, which causes no bias. Contradictory to non-parametric generation of prior distribution, uniform distribution is assumed to

reveal fundamental role of shape in prior distribution. The only constraint in the uniform distribution is that it integrates to 1.

Estimation of likelihood distribution In a given identical stimulus, observed pattern of activity varies and repeated exposure will form particular distribution. Once single observed activity is given, then the stimulus caused that activity is represented as probability distribution as well. The former distribution is referred as conditional distribution of m given s, and the latter distribution referred as likelihood function of m given s. The both likelihood function and conditional distribution can be expressed as of p(m|s) but likelihood function varies s whereas conditional distribution treated as function of m. Since we varied the stimulus in circular axis, p(m|s) is equated as circular normal distribution with having σ_{lkd} as variance-like variable. This variable is leaved as free parameter that can be adjusted according to model fitting procedure.

$$P(m|s) = \exp(\frac{1}{\sigma_{lkd}^2}\cos(m-s) - 1)$$

The likelihood distribution is visualized as matrix having stimulus at one axis and perceived stimulus in the other axis (Figure 5b). After acquire likelihood distribution with idea above, the posterior is calculated by equation following.

$$P(s|m) \propto P(m|s)p(s)$$

Categorical boundary Decisions for particular categories are made by estimating area of posterior distribution relatively to categorical boundary. With this procedure of estimation, the choices in current classification task get dichotomized regardless of subtle differences in posterior location. Like previous researches (Feldman & Griffith, 2009), the formation of categorical boundary in our model derived from prior domain. Concerning that aspects, our assumption of defining categorical boundary was the lowest position of sum between two-phoneme categories in prior distribution before amalgamating total prior. Thus, according to prior mean and variance, the categorical position was flexibly defined and reflect idiosyncrasy of individual observers. Similarly, the lowest positions in between peaks were selected as categorical boundary in non-parametric prior distribution. Non-parametric prior, however, has asymmetry in either sides of peak compare to parametric circular normal distribution. Thus, flexibility in formation of categorical boundary as well as shape of each individual peak was increased. In uniform distribution, categorical boundary was leaved freely as parameter to be fitted so that unlimited freedom in selecting categorical boundary can be allowed.

Method to acquire response proportion in classification experiment The given stimuli requires two components for being categorized: posterior distribution and boundary between categories. By calculating the proportion of area that belongs to particular categories, the percentage of response was modeled. The equation is following.

$$P_{catA} = \frac{\int_{b-\pi}^{b} P(s|\theta)ds}{\int_{b-\pi}^{b+\pi} P(s|\theta)ds}$$

In the equation, P_{catA} means proportion of area belonging to particular category A, and b indicates location of the boundary between category A and category B where A < B in the axis dimension (Figure 2a).

Method to acquire percentage correct in discrimination experiment In our Bayesian model, we supposed that three consecutive (AXB) stimuli are represented in separate probability distribution. The goal of task is to compare similarity of between samples obtained from posterior distributions generated from given stimuli. Once we assume any single value sampled from each posterior distribution as s^A_i, s^B_i and s^X_i, the similarity between s^A_i and s^X_i is inversely proportional to $| s^A_i - s^X_i |$. In here, the i indicates ith trial or sampling among whole numbers of trial. Then we can produce '$Z_A = s^A_i - s^X_i$' having new probability distribution, which indicates probability of s^A_i obtained from posterior A differs from s^X_i obtained from posterior from X. The probability density function of new variable Z_A is acquired by following equation:

$$f_{Z_A}(z) = f_{\mathrm{sA}-\mathrm{sX}}(z) = P(Z_A = z) = \int P(\mathrm{sA} = y)P(\mathrm{sX} = z - y)dy$$

The given form is identical to convolution of independent variables and can be calculated with MATLAB function. The procedure above is also identical between B and X, and we can acquire $f_{Z_B}(z)$. Having $f_{Z_A}(z)$ and $f_{Z_B}(z)$ as marginal distribution for two independent axes, $s^A_i - s^X_i$ and $s^B_i - s^X_i$, their joint probability distribution is generated (Figure 2b). In the two dimensional matrix for probability distribution, two diagonal lines that cross orthogonally generates areas (Figure 2b: white dashed line): top and bottom indicate P ($|s^A_i - s^X_i| < |s^B_i - s^X_i|$) whereas right and left denote P ($|s^A_i - s^X_i| > |s^B_i - s^X_i|$). When the three consecutive stimuli are given as $s^A_i\ s^A_i\ s^B_i$ or $s^B_i\ s^B_i\ s^A_i$, the correct answer should be proportion of volume involved in top or bottom area generated by diagonal line.

Result

Human Psychophysics To examine how well the Bayesian encoding-decoding model accounts for categorical perception shown in phoneme perception, we simultaneously fitted the posterior generated by prior and sensory likelihood to the data acquired by two distinctive behavioral tasks. The hallmark feature of categorical response in classification experiment is a sharp transition of choice fraction while manipulating stimuli near boundary between prototypical phoneme and occurrence of plateau of choice fraction near prototypical position (Figure 3a). The Bayesian encoding-decoding model provides a good account of the hallmark features shown in behavior. The key features of categorical perception are appeared across the subjects with idiosyncrasy in each subject is shown, and the Bayesian model captures the individual differences (Figure 3b). To check how the model fitting results are deviated from observed results in the classification task, we plotted the predicted choice fraction against observed choice fraction (Figure 3c). In this analysis, diagonal line (gray line) indicates model fit result matches with observed result from the experiment. The behavior or model fit results that correspond to 0 or 1 in choice fraction stays closely to the diagonal line whereas results with other value dynamically vary. Across the whole subjects, the Pearson linear correlation between observed data and model result was 0.987. To confirm the model whether possessing any problems in the fundamental assumption, we performed residual analysis for each single stimulus (Figure 3d). The larger residuals, which indicate a discrepancy between behavioral data and model data, are shown near boundary between prototypical phonemes whereas values acquired near prototypical stimuli approximates to zero.

The behavioral data obtained from discrimination experiment was analyzed with the method adopted in classification experiment. Initially, we checked whether suggested model captures key features in the behavioral experiment. The indicative feature shown in discrimination task is decreased of the discriminability when reference stimuli are given near to prototypes whereas increases when reference stimuli are given near boundary of phoneme categories (Figure 4a). The model result well accounted the pattern of the behavior in general. The phenomena are not limited particular sets of phoneme pair being discriminated, but appeared in all range and all subjects. Despite the substantial idiosyncrasy between subjects, the model accounted for considerable fraction of variance shown in all individual subjects (Figure 4b). Compare to classification data and model, the idiosyncrasy and

difference between phonemes are more evident. For discrimination task, comparison between observed percent correct and predicted percent correct was performed (Figure 4c). Likewise, the pattern in the classification task, the behavior or model fit results that correspond to 0 in percentage correct closely stays close to the diagonal line whereas results with other value dynamically varies. In comparison to the classification, not only general scatteredness was increased but also the results that correspond to 1 in percentage correct are not matching with diagonal line. Across the whole subjects, the Pearson linear correlation between observed data and model result was 0.644, which is lower than value obtained from classification experiment. The residuals of the model fit compare to behavioral data are analyzed (Figure 4d). Unlike the residual analysis result in classification experiment, the residuals did not have systematic pattern of deviation from zero. Instead, the residual pattern between conditions, near prototypical phoneme and near categorical boundary, shows systematic difference in their pattern. When the distance between reference stimuli is minute, the pattern of residual is arbitrary. If the distance between reference stimuli increases, however, the residuals in near categorical boundary condition get approximate to zero whereas the residuals in near prototypical phoneme does not change much.

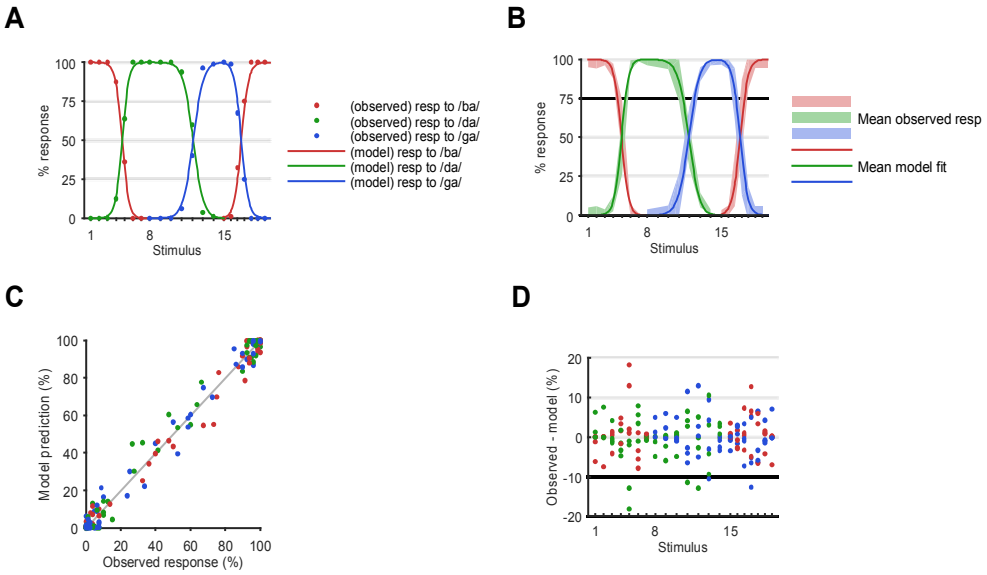

Figure 3. Behavior and model result for classification experiment for both representative subject and population. **a**) Percentage of response selecting particular phoneme in a single subject (S1). Each color indicate individual phoneme as its response (red: /ba/, green: /da/, blue: /ga/, and color indication in Figure 3 is same meaning). The symbol (filled circle) shows behavioral result acquired from experiment and the line shows model fitting result. Each single dot includes 80 trials. **b**) Classification result for the average across the subjects (N=4). The shaded area indicates the standard deviation of behavioral results and line indicate mean of the model results. **c**) Observed data is plotted against model predicted result. Diagonal line designates the perfect correspondence between model prediction and behavior. **d**) Residual between model result and behavioral results are plotted. The value zero specifies model result matches perfectly with the behavioral result.

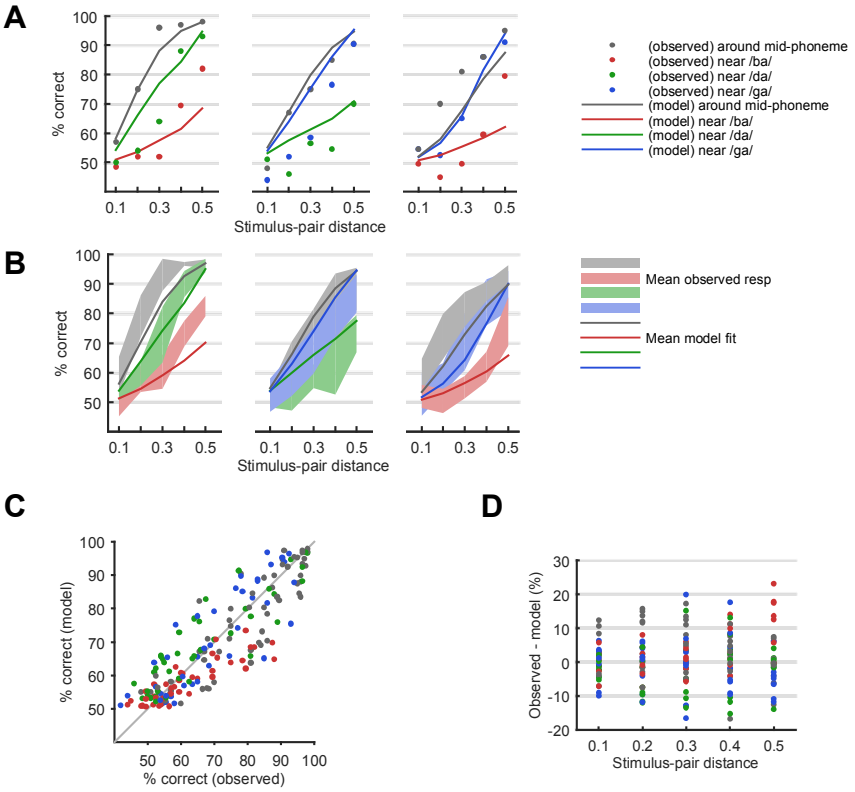

Figure 4. Behavior and the model result for discrimination experiment for both representative subject and population. **a**) Percentage correct for discriminating two reference stimuli in a single subject (S1). Each color indicates each condition where reference stimuli located (red, green, blue: near prototype, black: near boundary, and color indication in Figure 4 is same meaning). The symbol (filled circle) shows behavioral result acquired from experiment and the line shows model fitting result. Each single dot includes 200 trials. **b**) Discrimination result for the average across the subjects (N=4). The shaded area indicates the standard deviation of behavioral results and line indicate mean of the model results. **c**) Observed data is plotted against model predicted result. **d**) Residual between model result and behavioral results are plotted.

Model Fitting The encoded stimuli can be represented as sensory likelihood reflecting diverse source of noise. Because there are multiple sources of noise, we approximated a circular normal distribution to simply capture the noise by variance-like parameter (Figure 5b). Once an encoded stimulus is probabilistically represented, the listener combines prior knowledge with the current stimulus to secure with high precision and temporal efficacy (Figure 5a). The prior is assumed to have multimodal shape with three independent mean positions and variance-like parameters to capture previously experienced variability. To validate the model by whether it robustly accounts for the both phoneme perception tasks, classification and discrimination, we constrained priors and likelihoods by simultaneously fitting the posterior distributions to result from classification and discrimination tasks. Thus, the prior distribution used in classification and discrimination is identical (Figure 5a and d). The probabilistic representation of currently encoded stimuli and prior knowledge is decoded as the posterior distribution (Figure 5c). If the decoder is unbiased, means of the posterior distribution (color dots in Figure 5c) should be aligned in a diagonal line that indicates correspondence between actual and perceived stimuli. The model, however, predicts that systematic bias occurs when generating posterior distribution. The mismatch between the perceived and actual stimuli primarily arises because prior probability strongly influences formation of the posterior distribution. When stimulus is given near categorical boundary, the individual peak of the prior distribution offsets each other's influence in generating posterior distribution.

How brain performs classification and discrimination in the model are determined by method of utilizing internally generated posterior distribution. In the classification task (Figure 5d-e), listener classifies the given stimulus according to the relative location of decoded posterior distribution from the pre-defined categorical boundary that corresponds to least experienced stimuli profile in the linguistic environment. The stimulus given near the peak of the priors are intensely pulled toward one type of prior with forming uni-modal shape of posterior (Figure 5f, single dots in horizontal section). In contrast, when the stimulus is provided near boundary between two different phonemes, then it forms bimodal posterior with difference in area belonging to particular category according to relative strength of prior distribution for particular type of phoneme (Figure 5f, double dots in horizontal section). As consequence, the stimulus given near boundary between prototypical phonemes is less likely to be heard as ambiguous. Instead, it is heard as either one type of phoneme that result in one particular response. The hallmark feature of categorical perception in classification was replicated in Bayesian

model as dot expressed: more large single dots are shown when given stimulus is near prior and small double dots are located near boundary (Figure 5f).

For the discrimination task, distances of posteriors distribution generated by distinctive stimuli are directly compared. Once the stimulus was given near prototypical phonemes, then the distance between posterior distributions becomes closer than given stimuli since the single strong prior pulls both likelihoods towards it. In this case, the posterior distribution will possess overlapping portion due to very close distance in between (Figure 5h, black line). In contrast, the stimulus given around the boundary between two distinctive phonemes generates posterior distribution that is farther than the given stimuli. This is because two distinctive peaks of prior for each phoneme pull separately, and that extends the distance between posterior so that discrimination can be easier (Figure 5h, dotted line). In latter case, since the posterior distribution is less overlapped, it is easier to be discriminated. The discriminability of the posterior distribution in the model was quantitatively described with area under ROC curve (Figure 5I). As seen in figure 5I, the fluctuation of discriminability according to location and distance of reference stimuli is well predicted by the model along whole phoneme ranges.

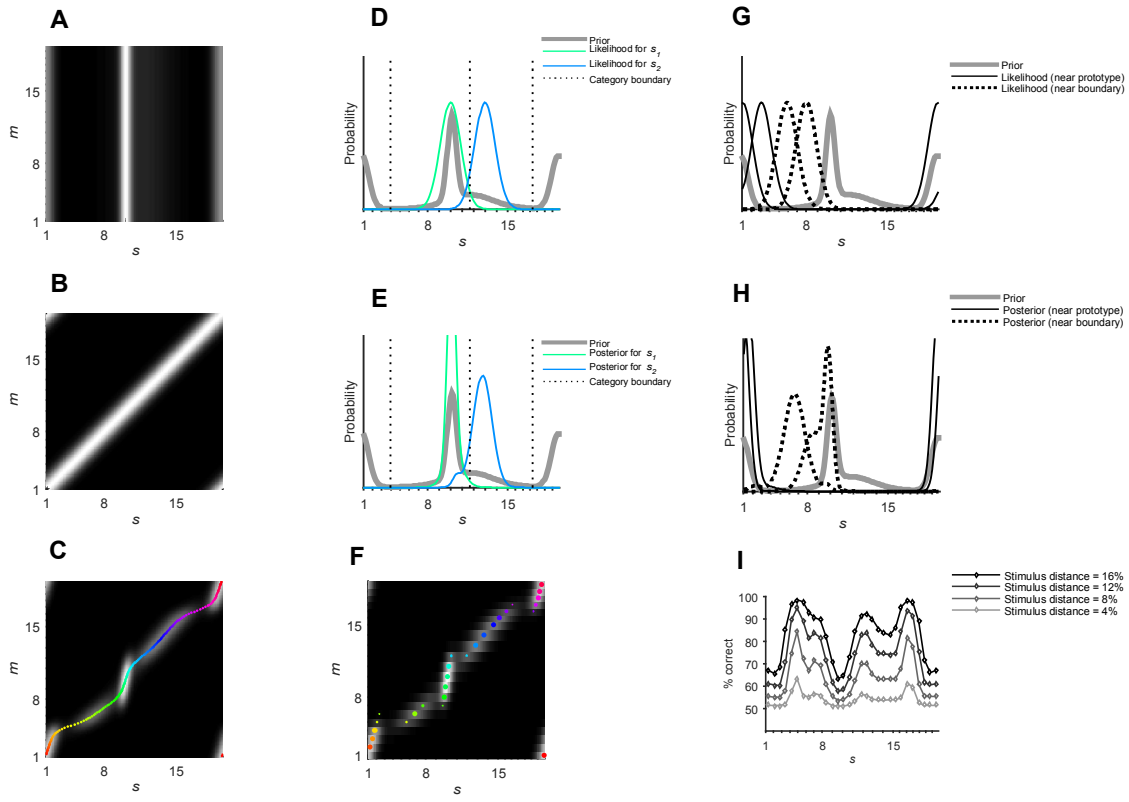

Figure 5. Each component of the model fitted to behavior data and resulting posterior distribution. **a**) The prior distribution of a representative subject (S1) as a result of fitting model to behavioral data. **b**) Likelihood function of the identical subject (S1). s denotes the physical stimuli that is given and m denotes perceived stimuli, and this denotation is identical in figure 5c and f. **c**) The posterior distribution is computed according to Bayes' rule, as the normalized product of the prior and likelihood. Mean of posterior for stimuli differing one degree is shown as color gradient dots. **d**) The prior distribution and likelihood of the subject performing classification experiment for two different stimuli. **e**) Posterior distribution generated as a result of two different stimuli. **f**) Shape of the posterior distribution is plotted. Each dot indicates how much area belongs to single modal of the peak. Thus, is a perception of stimuli result in uni-modal, there will be a single dot whereas double dot per single stimuli when bi-modal shaped posterior distribution is generated. **g**) Posterior distribution generated as a result of the likelihood of two different conditions: two identical colored distribution for near prototype and the other two for near boundary. **h**) Posterior distribution generated as a result of stimuli given in two different conditions. **i**) Discriminability as result of the area under ROC curve, which corresponds for percentage correct. Each colored line indicates different distance between the stimuli. The dots in each line are the mean location between two reference stimuli.

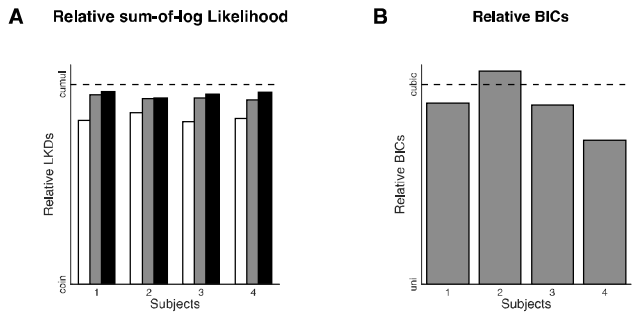

Figure 6. Goodness-of-fits comparisons using three different priors: uniform, circular normal, and cubic spline. **a**) Relative sum-of-log likelihood compared with the coin-flipping and cumulative normal distribution fitting model. Coin flipping, which means the result of all stimuli is 0.5, is denoted as 0 and cumulative normal fitting model is denoted as 1. Each color indicates different types of model (white: uniform, gray: circular normal, black: cubic spline interpolation). **b**) The BIC value resulted from prior with circular normal distribution is relatively compared with other priors. The value 0 is the value acquired from uniform prior, and 1 is the value acquired from the cubic spline prior distribution.

Model Comparison To challenge the prior distribution, which is key for the assumption of the model, we quantitatively compared the sum of log-likelihood obtained from model with differently shaped prior distribution (Figure 6a). To remove any biases introduced by strong prior, we tested the uniform prior that is identical to no prior. In addition, we provided more freedom for the shape than circular normal distribution by generating prior distribution with cubic spline interpolation. To set the ground and ceiling values, coin flipping and cumulative Gaussian model fitting was compared as well. For each subject, the Bayesian encoding-decoding model using the estimated prior distribution and likelihood widths performs as good as the individual cumulative Gaussian fits. Considering the number of parameters that allows a degree of freedom in the model, the result is notable since Bayesian encoding-decoding model has less parameter. In the case of model adopting circular normal distribution as prior, there are only seven parameters in comparison to 60 parameters in total used in cumulative Gaussian fit performed independently for experiment conditions in classification and discrimination. To compare the goodness-of-fit with considering number of parameters used in the model, we computed relative Bayesian Information Criterion (BIC) value of circular normal prior distribution (Figure 6b) (Schwarz, 1978). The BIC value acquired from uniform prior distribution was set as 0 and BIC value of cubic spline prior distribution as 1. Though there are individual differences across subjects, relative BIC values resulted in a similar pattern: the value is always greater than 0 and close to 1. The result is a quantitative indication showing that shape non-uniformity of the prior distribution accounts for categorical perception of phoneme perception. Furthermore, only the shape

non-uniformity can capture a large portion of the categorical perception in phoneme perception regardless of degree of freedom in the shape.

Discussion

In current study, we have shown that categorical perception of phoneme can be explained by Bayesian encoding-decoding model with a limited number of parameters. Previously suggested model based on Bayesian framework hypothesized importance of systematic bias in explaining categorical perception (Feldman et al., 2009). Beyond prediction made by simulated data in the previous study, we performed two distinctive psychophysics experiments within a single subject and simultaneously fitted the model with the data acquired from each different task. By constraining the model by fitting, we could not only validate the model but also estimate components of the model with accurate values of parameters. As result, Bayesian encoding-decoding model captured hallmark features of categorical perception shown in both behavioral tasks. Especially, the estimated prior distribution indicates that narrow-tuned multimodal shape inducing systematic bias is crucial in explaining categorical perception of phoneme.

Although the previous study has suggested excellent model based on Bayesian and captured categorical perception, still the exact values for the parameters of model components were not obtained since the model was not fitted to empirical data. By fitting the model with empirical data in the current study, we extracted the accurate values of the parameter for the estimated prior and likelihood. By the quantitative value of the parameter, we can claim strong perceptual bias by prior distribution is necessary for explaining categorical perception. Furthermore, individual difference in categorical perception became possible to be compared in a quantitative manner as well.

Challenging the prior distribution by uniform shape emphasized superiority of non-uniform prior over uniform prior in capturing categorical perception. This denotes necessity of the accumulated knowledge, which is the core component of the model. A point that should be focused on is that the superiority is evident when the data from two distinctive tasks were simultaneously fitted. In classification task, both models with uniform and non-uniform shaped prior distribution capture categorical perception. Thus, it is difficult to provide evidence that Bayesian framework with systematic bias caused by prior is appropriate to explain categorical perception. Only when the model is constrained with two distinctive types of task, however, the superiority of non-uniform prior over uniform became more evident. This indicates sufficient numbers of behavioral experiments should be fitted simultaneously for constraining the model appropriately.

Non-uniformly shaped prior distribution was further validated by comparison with non-parametrically generated prior distribution. If the fitting result of non-parametrically estimated prior distribution differs highly with circular normal prior distribution, then Gaussian approximated prior distribution in our model and previous study would be invalid. Comparison between circular normal prior and non-parametrically generated prior implies that assumption for the prior distribution is valid (Figure 6).

Hallmark features of categorical perception are less evidently shown in some subjects compare to other subjects. The result can be attributed to unfamiliarity with the synthetic stimuli. Natural sounds involve a large number of variations when articulation particular sound, for example, timbre or pitch other than formants. In our experiment, only with F2 and F3 formants was manipulated whereas other things were identical between stimuli. Though synthetic sound did not deteriorate the categorical perception, unnatural modulation of acoustic features would be a possible reason particular people showed weak categorical responses. Some findings suggest that people from different linguistic environments show dissimilar categorical response even when stimuli were identical (Bonasse-Gahot and Nadal, 2012). The case can be analogized to our subjects. Since we have adopted stimuli from studies done at Western countries, our results would show different patterns of response.

Some studies distinguish *categorization* rather than *categorical perception* (Holt et al, 2010). The former indicates selecting one particular category from multiple possible categories, and latter one indicates drastic change of perceptual category rather than gradual conversion of perceptual category. In fact, we face multiple alternatives other than /ba/ or /da/ when we make decisions for phoneme perception. Our model, however, has its foundation on forced choice test between only two distinct categories, which requires consideration of statement given above. So far, studies of the mechanisms of decisions between more than two alternatives have, gained less attention at the field of decision-making (Churchland and Ditterich, 2012). Current progress in field of neuroscience and behavioral economics, new plausible models are proposed. For example, multi-alternative decision field theory (MDFT) was proposed (Brown et al, 2009). Gist of this model is that decision is not only based on uncertainty level but lateral inhibition according to desirability. By adopting blooming paradigms in decision-making, the model will evolve to be more realistic.

The Bayesian inference has been successfully adopted in numbers of computational models that explains human behaviors including multisensory integration (Ernst and Banks, 2002),

sensorimotor integration (Kording and Wolpert, 2004), or interval time perception (Sohn and Lee, 2013). Especially, the study that reverse-engineered the prior distribution of orientation sensitivity in vision had a seminal impact (Girshick and Simoncelli, 2011). In their study, the discriminability is enhanced near peak of the prior distribution whereas reduced in between the peaks. The result is contradictory with our findings that discriminability is reduced near peak of the prior distribution. The contradictory result could be attributed to the strength of the prior distribution. Since all orientation is informative in the visual system, the contrast between peak and valley of prior is not as drastic as the prior distribution acquired in phoneme perception.

Suggested Bayesian model can be extended to various ways. Primary consideration is physiological instantiation. The encoding stage of auditory stimulus naturally can be associated with populations of neurons that are selective for the dynamic change of acoustic feature. In the case of phoneme perception, the feature dimension for distinguishing stop consonant is spectral-temporal dynamics in frequency harmonics during the initial 40ms period after stimulus onset. Massive numbers of single-cell studies on animals reported that a substantial fraction of neurons in the primary auditory cortex has a spectral-temporal receptive field ('STRF') structure. STRF indicates neurons are tuned for temporal modulation of the frequency spectrum (Theunissen et al., 2001; Mesgarani et al., 2008, Gill et al., 2006, Singh and Theunissen, 2003). Recently, Mesgarani and the colleagues have reported existence of STRF in the human brain by using ECoG in epileptic subjects (Mesgarani et al., 2014), and fMRI study mimicking the methods in single cell revealed STRF in normal human's brain (Schönwiesner and Zatorre, 2009). Our Bayesian encoding-decoding model posits that the neurons having STRF might be candidates for operating the proposed computation at the encoding stage. As shown previously (Seung and Sompolinsky, 1993; Pouget et al., 2003), the likelihood can be computed from population response of neurons once their tuning functions are known. At following decoding stage, neurons read-out the activities from population of neuron at encoding stage and estimates what information is given. By applying the decoding method called maximum likelihood estimation (Deneve et al., 1999; Ma et al., 2006, 2013), the decoding activity of neural population can be mimicked. One thing, which we can consider when implementing neural component, is the non-uniformities of neural tuning curve, which is location and tuning width. A theoretical study posits that the non-uniformities of neural populations reveal a strategy by which the brain allocates its resources (neurons and spikes) so as to encode stimuli optimally (Ganguli et al., 2012). Previously, the neural model for categorical

perception of phoneme based on information theory and population coding paradigm also posited non-uniformly distributed neural resources (Bonasse-Gahot & Nadal, 2008, 2012). In their model, more neurons are allocated near categorical boundary so that rapid variation of information should be precisely captured. Empirical results, however, show that more neurons are tuned to frequently exposed frequency for auditory tone discrimination task in mouse primary auditory cortex (Han et al., 2007). These results provide possibility that the model suggested by Bonasse-Gahot is unnatural because they assume that more neurons are devoted to infrequently exposed stimuli.

Additional extension to consider with current Bayesian model is finding plausible candidate for the locus of the prior distribution. Primary candidate, which is not exclusive with other candidate, is the correlated variability and decoding read-out weight. While stimuli are preceded in feed-forward hierarchies, fluctuations in correlated sets of neurons to the repeated presentation of identical stimuli builds up correlational neural structure. Previous study has reported correlation structure results in read-out weight (Haefner et al., 2013). Idea above can be one possible method of implementation of prior. Repeated presentation of categorized external phoneme stimuli results in fluctuation in certain correlated neurons. According to the neural plasticity, the read-out weights get adjusted to the optimal way for decoding identical stimuli. That sequential procedure could be hypothetically assumed as a source of the prior distribution. Further plausible hypothesis assumes that modulated spontaneous activity by the degree of cortical reorganization due to frequent exposure to particular frequency profiles is locus of the prior distribution (Köver et al., 2010). The model suggested by the study has captured the difference in tone discrimination behavior by non-specifically increased spontaneous activity in auditory cortex that eventually influences the read-out of stimulus-evoked activity. With the result, they suggest that optimal integration of prior and sensory information may be achieved by adjusting the levels of baseline activity rather than by integrating neural activity generated by top-down signal. This idea can be linked with the findings from Chang and his colleagues (Chang et al., 2010). According to the study, the categorization of the perceived stimuli is achieved within 110ms, which is relatively short period. This short period indicate that the prior distribution lies at the locus of perception with particular built-in neural architectures rather than top-down signal in large scale. The other candidate for phoneme prior would be a phoneme somatotopic arrangement linked with superior temporal gyrus (STG), which the idea corresponds to the theory of motor. Unlike any other sensory modalities, auditory sensation has a close relationship with motor performance called articulatory. For

example, pre-lexical unit /ba/ can be both perceived and articulated by individual. In contrast, orientation stimulus perceived in the visual modality cannot be replicated by motor performance. Regarding the uniqueness of phoneme perception, we can make the assumption that the articulatory map in the motor area might influence parameter of phoneme perception prior. Recent study has suggested evidence of speech map in the human ventral sensorimotor cortex with electrode implanted to epilepsy subjects (Conant et al., 2014). Their findings enable us to speculate that prior partially be influenced by phoneme articulatory maps. Further supportive finding is Transcranial Magnetic Stimulation (TMS) stimulation to the premotor cortex, which resulted perceptual change in categorical perception of phoneme in human (Meister et al., 2007; D'Ausillio et al., 2009). Consistent with that, it is testable whether area related with phoneme articulatory like Orofacial area in sensorimotor cortex might change phonemic perception.

Measuring spectro-temporal profile of individual's articulate phonemes and comparing spectral-temporal profile with acquired prior parameters from the model would be one method for validation. The correspondence between estimated prior of orientation sensitivity and probability of orientation in the external environment upholds validity of the estimated prior in the study (Girshick and Simoncelli, 2011). Likewise the previous study, we can compare our estimated prior distribution with the spectro-temporal profiles of articulated phonemes from each subject. If the spectro-temporal profiles of articulated phoneme match with the prior distribution estimated by the model, we can claim that validity of our assumption on the prior distribution.

References

Bonnasse-Gahot L, Nadal J-P (2008) Neural coding of categories: information efficiency and optimal population codes. J Comput Neurosci 25:169–187.

Bonnasse-Gahot L, Nadal J-P (2012) Perception of categories: from coding efficiency to reaction times. Brain Research 1434:47–61.

Brown S, Steyvers M, Wagenmakers E-J (2009) Observing evidence accumulation during multi-alternative decisions. Journal of Mathematical Psychology 53:453–462.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nature Neuroscience 13:1428–1432.

Churchland AK, Ditterich J (2012) New advances in understanding decisions among multiple alternatives. Current Opinion in Neurobiology 22:920–926.

D'Ausilio A, Pulvermüller F, Salmas P, Bufalari I, Begliomini C, Fadiga L (2009) The motor somatotopy of speech perception. Current Biology 19:381–385.

Deneve S, Latham PE, Pouget A (1999) Reading population codes: a neural implementation of ideal observers. Nature Neuroscience 2:740–745.

Feldman NH, Griffiths TL, Morgan JL (2009) The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychological Review 116:752.

Ganguli D, Simoncelli EP (2014) Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. Neural Computation 26:2103–2134.

Gill P, Zhang J, Woolley SMN, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21:5–20.

Girshick AR, Landy MS, Simoncelli EP (2011) Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nature Neuroscience 14:926–932.

Haefner RM, Gerwinn S, Macke JH, Bethge M (2013) Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nature Neuroscience 16:235–242.

Han YK, Köver H, Insanally MN, Semerdjian JH, Bao S (2007) Early experience impairs perceptual discrimination. Nature Neuroscience 10:1191–1197.

Holt LL, Lotto AJ (2010) Speech perception as categorization. Attention, Perception, & Psychophysics 72:1218–1227.

Iverson P, Kuhl PK (1995) Psychophysical procedure and the perceptual magnet effect: Comparisons of fixed and roving AX discrimination of /i/. The Journal of the Acoustical Society of ….

Körding KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427:244–247.

Köver H, Bao S (2010) Cortical plasticity as a mechanism for storing Bayesian priors in sensory perception. PLoS ONE 5:e10497.

Kuhl PK (1991) Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Perception & Psychophysics 50:93–107.

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nature Neuroscience 9:1432–1438.

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognitive psychology 18:1–86.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science 343:1006–1010.

Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123:899–909.

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: Feedback is never necessary. Behavioral and Brain Sciences 23:299–325.

Perkell JS, Klatt DH (1986) Invariance and variability of speech process. Hillsdae, NJ: Lawrence Erlbaum

Pouget A, Beck JM, Ma WJ, Latham PE (2013) Probabilistic brains: knowns and unknowns. Nature Neuroscience 16:1170–1178.

Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. Annu Rev Neurosci 26:381–410.

Schönwiesner M, Zatorre RJ (2009) Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proceedings of the National Academy of Sciences 106:14611–14616.

Schwarz G (1978) Estimating the dimension of a model. The annals of statistics 6:461–464.

Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. Proceedings of the National Academy of Sciences 90:10749–10753.

Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 114:3394–3411.

Stevens KN, Blumstein SE (1978) Invariant cues for place of articulation in stop consonants. J Acoust Soc Am 64:1358–1368.

Stocker A, Simoncelli EP (2006a) Sensory adaptation within a Bayesian framework for perception. Advances in neural information processing systems 18:1289.

Stocker AA, Simoncelli EP (2006b) Noise characteristics and prior expectations in human visual speed perception. Nature Neuroscience 9:578–585.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 12:289–316.

국문 초록

협소한 사전분포를 갖는 베이시언 추론을 통한 음소지각의 이해

유 승 범

뇌인지과학 전공

자연과학대학

서울대학교 대학원

지각작용은 주어진 표본의 정보를 통해 원래 모집단의 상태를 추론하는 통계적 유추 과정과 상당히 유사하다. 이 과정에서 지각기관을 통해 현재 주어진 자극의 표상과 선행경험을 통해 형성된 믿음이 결합된 정보가 이용된다. 최근의 시각 신경과학 연구에서는 지각현상을 베이시언 프레임 내에서 해석하려는 과거의 시도를 뛰어넘어, 모델 피팅을 통해 실제 선행경험이 표상하는 확률분포인 사전분포의 모양과 특성값을 유추하는 것이 시도되었다 (Girshick et al., 2011) 본 연구는 시각 신경과학에서 사전분포의 특성을 유추하는 연구를 음소 이해에 적용하였다. 특히 범주화되어 경험되는 음소의 특징상 협소하고 강한 사전분포 예상해볼 수 있다.

사전분포를 유추하기 위하여, 주어진 자극을 특정 범주로 구분시키는 실험과 주어진 서로 다른 자극을 구별해내는 실험을 진행하였다. 실험에 쓰인 자극은 특정 포먼트를 변화시켜 /바/, /다/, /가/의 범주 내에서 변화하였다. 사전분포는 개인이 가장 빈번하게 경험하는 곳에서 평균값을 갖고 피험자가 기존에 경험한 다양성에 따라 분산값을 갖는 정규분포 세 개를 결합시킨 확률분포의 형태로 표상하였다. 감각 우도값(Sensory likelihood)은 주어진 자극의 값을 평균으로 갖고 자극에 따른 분산값은 동일하도록 표상하였다. 선행 실험에서 나타난 음소이해의 대표적 특징은 자극의 선형변화에도 불구하고 지각은 비선형적으로 급격히 변하는 것이며, 후행 실험에서 나타는 대표적 특징은 양 범주의 중간에서 주어지는 자극에 대한 구별력이 높다는 것이다. 소수의 패러미터만을 변화가능하도록 설정한 모델피팅은 두 개의 다른 실험에서 나타난 음소이해의 대표적 특징을 잘 설명해내었다. 본 연구에서 쓰인

정규분포를 결합시킨 선행경험의 확률적 표상이 균일확률분포로 치환했을 때보다는 현상을 더 잘 설명하고 비모수적으로 생성한 확률분포와 비슷한 결과를 보였다.

위의 결과를 통해, 인간의 음소의 이해를 설명할 때 가장 빈번하게 경험된 음소자극 근처에서 협소하게 형성된 사전분포가 필요하다는 것을 알 수 있다.

지시어: 베이시언 추론, 범주화, 음소

학번: 2013-22455