Fundamental Frequency of Vowels

Georg Khella, Applied Statistics, Project 1

February 2025

1 Introduction

This project is intended to examine the acoustic measurements of vowels from American English speakers, according to the dataset of Hillebrand et al. (1995). The fundamental frequency f_0 is one of the primary features of vowel sounds, establishing their pitch and perceived quality.

The goal of the current study is to estimate a distribution that would enable actual simulation of vowel frequencies f_0, making it possible to analyze the characteristics of these sounds in the time domain. For this purpose, sophisticated statistical methods will be utilized to ascertain the validity of a combination of two log-normal distributions. This provides a sufficient context for the lowest part of the base frequency, shortened as f_0. Scientists claim that this model is especially good at explaining such frequencies, a hypothesis to be examined via data analysis.

To achieve these objectives, the Expectation-Maximization (EM) algorithm with jittering and a Bayesian approach will be employed to estimate the distribution of parameters. Finally, the model will be validated based on parametric acceptance of bootstrap methods, along with goodness-of-fit tests, giving an opinion of the ability of the model to successfully reproduce the compiled information and establish a valid basis for subsequent vowel explorations in frequency structures.

2 Data Overview and Analysis

This section provides an overview of the dataset and an exploratory analysis of its distribution, including some relevant plots.

2.1 Dataset Description

This section contains a thorough breakdown of the dataset as well as an exploratory analysis of its distribution.

The data are due to Hillebrand et al. (1995) and consist of acoustic measurements pertaining to vowel sounds produced by American English speakers, including children, males, and females. The main variable of interest is the root frequency (f_0), in Hertz (Hz), which determines the pitch of vowel sounds.

The information is in binned format, i.e., vowel frequencies are grouped into pre-defined intervals instead of being noted individually. The primary variables are:

- **origin**: Corresponds to the bottom of every frequency bin.
- **endpoint**: The upper limit of every frequency bin.
- **count**: Number of vowel samples measured in each bin.
- **percentage**: The percentage of total samples in each bin.

Since the binning is not equidistant, extra care needs to be taken when modeling and analyzing the data.

2.2 Exploratory Data Analysis

The exploratory data analysis aims to comprehend the nature of the dataset by combination of summary statistics and the appropriate visual displays. Since data are accumulated into bins (instead of single values), we focus primarily on a histogram.

2.2.1 Descriptive Statistics

Prior to conducting the graphical analysis, it is useful to examine descriptive statistics, including the median and interquartile range. These statistical measures give a good description of the central tendency and dispersion of the basic frequency (f_0). In our categorized dataset, the summary statistics are determined by considering the bin limits (starting point and finishing point), as well as the number of observations in each bin (count) and the percentage of the total.

	Min	**1st Qu.**	**Median**	**3rd Qu.**
start_point	90.0	144.0	192.8	233.0
end_point	107.0	154.0	202.4	237.0
count	48.0	64.0	67.0	69.0
percentage	2.878	3.837	4.017	4.137

Table 1: Descriptive statistics for selected variables in the binned dataset.

From Table 1, we observe that the binned frequencies (f_0) start as low as 90 Hz and extend up to about 330 Hz (not shown in this table but computed from the maximum end point). The median start point (192.8 Hz) and end point (202.4 Hz) suggest that a substantial portion of the bins lies roughly between 144 Hz and 233 Hz (the 1st and 3rd quartiles for start point).

In terms of *count* and *percentage*, the data indicates that every bin has, on average, about 67 vowels. The recordings exhibited a variation of a percentage of vowels in any bin from approximately 2.88% to 4.74%. This fairly small range of bin percentages indicates that no individual bin dominates the distribution, although there is a possibility of multiple peaks in the bins.

2.2.2 Histogram of Fundamental Frequency

Because the observations are binned, a histogram is the most direct and informative way to visualize the distribution of the fundamental frequency (f_0). Figure 1 depicts the empirical density, allowing us to identify any multimodal tendencies and assess the overall spread of the data.

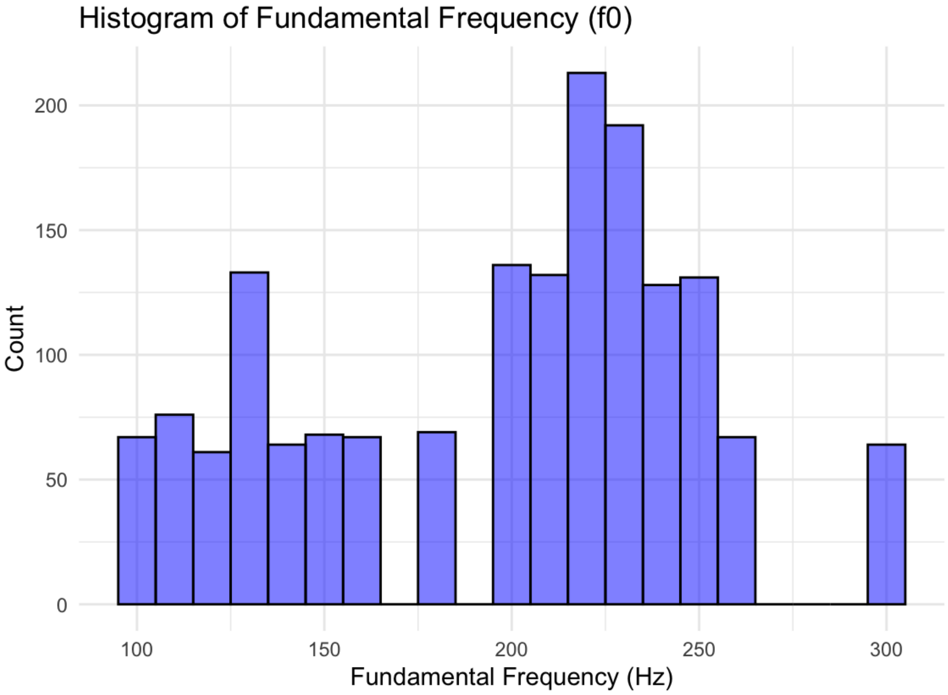

Figure 1: Histogram of the binned fundamental frequency (f_0).

2.2.3 Modeling with a Bi-Lognormal Distribution

Based on the observed distribution in the histogram and prior literature, we fit a **bi-lognormal distribution**, which can capture multiple peaks. The probability density function of a lognormal distribution is:

$$f(x) = \frac{1}{x\,\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right),$$

and the bi-lognormal distribution is a weighted mixture of two such components:

$$f_{\text{bi}}(x) = w\, f_1(x; \mu_1, \sigma_1) \; + \; (1 - w)\, f_2(x; \mu_2, \sigma_2),$$

where w is the weight of the first lognormal component. We use heuristically chosen parameters from the histogram:

- $w = 0.3$
- $\mu_1 = \log(130)$, $\sigma_1 = 0.1$
- $\mu_2 = \log(220)$, $\sigma_2 = 0.1$

To visually validate this assumption, we overlay the estimated bi-lognormal density function on the histogram (Figure 2). This comparison illustrates how closely the model aligns with the observed distribution and serves as a preliminary check before more rigorous statistical testing.

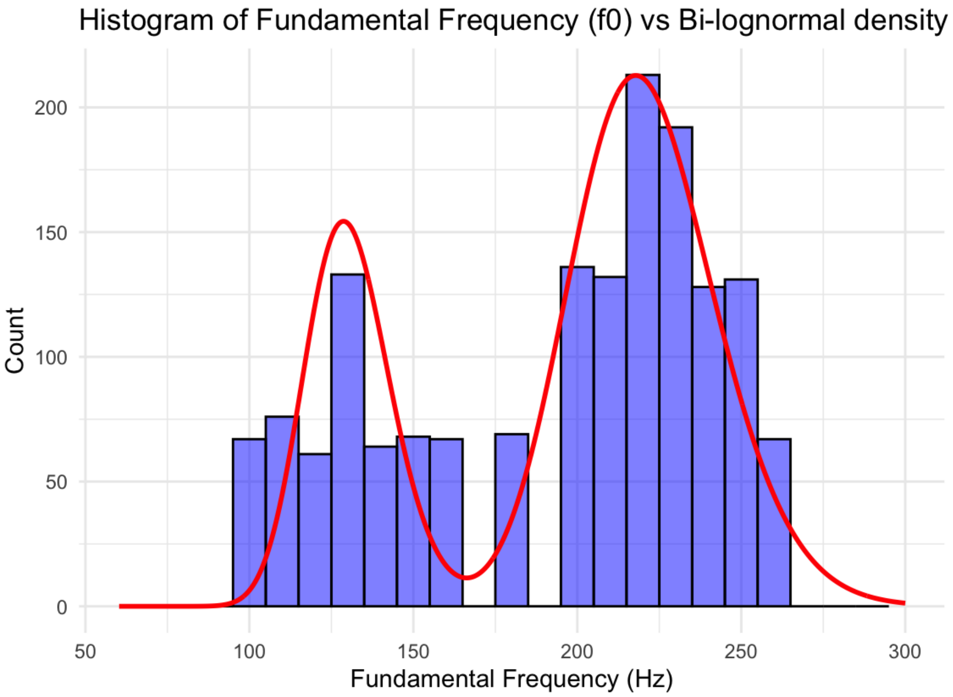

Figure 2: Histogram of fundamental frequency (f_0) vs. bi-lognormal density.

3 Statistical Methods

In this report, we present the statistical methods used to infer the parameters of a bivariate lognormal distribution: the Expectation-Maximization (EM) Algorithm and the Bayesian Approach. These methods provide robust frameworks for parameter estimation, each with its own advantages and applications. The EM Algorithm is particularly useful for handling incomplete data, while the Bayesian Approach offers a probabilistic perspective that incorporates prior knowledge into the inference process. Detailed explanations and implementations of these methods will be discussed in the following sections.

3.1 Expectation-Maximization Algorithm

The Expectation-Maximization algorithm is an iterative method for computing the Maximum Likelihood Estimators in cases where missing or latent data complicate direct maximization of the likelihood function [1].

3.1.1 Steps of the EM Algorithm

Given a set of observed data $\mathbf{X}_{\text{obs}}$, unobserved (latent) data $\mathbf{X}_{\text{miss}}$, and model parameters $\boldsymbol{\theta}$, the complete log-likelihood function is defined as:

$$\ell_{\text{comp}}(\boldsymbol{\theta}) = \log p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{miss}} \mid \boldsymbol{\theta}). \tag{1}$$

The goal is to maximize the observed log-likelihood:

$$\ell_{\text{obs}}(\boldsymbol{\theta}) = \log p(\mathbf{X}_{\text{obs}} \mid \boldsymbol{\theta}), \tag{2}$$

which can be rewritten using the marginalization property:

$$\ell_{\text{obs}}(\boldsymbol{\theta}) = \ell_{\text{comp}}(\boldsymbol{\theta}) - \log p(\mathbf{X}_{\text{miss}} \mid \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}). \tag{3}$$

Since $\ell_{\text{obs}}(\boldsymbol{\theta})$ is often intractable, the EM algorithm approximates its maximization through an iterative procedure consisting of two steps:

- **E-step (Expectation):** Compute the expected complete log-likelihood given the observed data and current parameter estimate $\boldsymbol{\theta}^{(t-1)}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \mathbb{E}_{\mathbf{X}_{\text{miss}} \mid \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(t-1)}} \left[\ell_{\text{comp}}(\boldsymbol{\theta}) \right]. \tag{4}$$

- **M-step (Maximization):** Update the parameter estimates by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$:

$$\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}). \tag{5}$$

This iterative process ensures a non-decreasing likelihood at each step, a property known as the *monotonicity of EM*:

$$\ell_{\text{obs}}(\boldsymbol{\theta}^{(t)}) > \ell_{\text{obs}}(\boldsymbol{\theta}^{(t-1)}). \tag{6}$$

In this project, we use the EM algorithm to estimate the parameters of a bi-log-normal distribution modeling the fundamental frequency (f_0) of vowel sounds. Given that the data are binned while the parameters to be estimated come from a continuous distribution, we apply the EM algorithm with jittering to iteratively estimate the mixing proportions, means, and variances.

3.1.2 Jittering in EM Algorithm

Jittering is a technique used to handle binned data when estimating parameters of a continuous distribution. Given a set of binned observations $\mathbf{X}_{\text{obs}}$, the true underlying values $\mathbf{X}^*$ are unknown but assumed to follow a continuous distribution. Formally, if a binned observation $X_{\text{obs},i}$ falls within the interval $[a_i, b_i]$, the latent continuous value X_i^* is modeled as:

$$X_i^* = X_{\text{obs},i} + \epsilon_i, \quad \epsilon_i \sim \mathcal{U}(a_i - X_{\text{obs},i}, b_i - X_{\text{obs},i}). \tag{7}$$

where ϵ_i represents a uniform perturbation within the bin range. This transformation reconstructs an approximate continuous distribution, reducing bias introduced by discretization [4].

In the EM algorithm, jittering is applied in the E-step, where the expected complete log-likelihood is computed not directly from $\mathbf{X}_{\text{obs}}$ but from the jittered data $\mathbf{X}^*$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \mathbb{E}_{\mathbf{X}^* \mid \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(t-1)}} \left[\ell_{\text{comp}}(\boldsymbol{\theta}) \right]. \tag{8}$$

This improves the estimation of the parameters $\boldsymbol{\theta} = (\pi_k, \mu_k, \sigma_k^2)$ in a mixture model, where each component k is modeled as a log-normal distribution:

$$f_0 \sim \sum_{k=1}^{2} \pi_k \mathcal{LN}(\mu_k, \sigma_k^2). \tag{9}$$

By incorporating jittering, we obtain a smoother approximation of the likelihood surface, leading to more robust parameter updates in the M-step. In this project, this approach is essential for accurately modeling the fundamental frequency (f_0) of vowel sounds, ensuring a better fit of the bi-log-normal distribution to the observed data.

3.2 Bayesian Inference Approach

In the Bayesian framework, all model parameters, represented by the vector θ, are treated as random variables rather than fixed but unknown quantities. This perspective assigns a prior distribution $p(\theta)$, which encodes our knowledge or assumptions about θ before observing any data.

Let X be the observed data. The likelihood function $p(X|\theta)$ measures how likely the observed data X are, given specific parameter values θ. Combining the prior with the likelihood via Bayes' theorem yields the posterior distribution:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}, \tag{10}$$

where

$$p(X) = \int p(X|\theta)p(\theta)d\theta \tag{11}$$

is the marginal likelihood (also called the model evidence). This integral acts as a normalizing constant ensuring that $p(\theta|X)$ is a valid probability distribution over θ.

The posterior distribution $p(\theta|X)$ then represents our updated knowledge about the parameters θ after observing X. However, evaluating the marginal likelihood $p(X)$ directly can be computationally prohibitive in most non-trivial models, since it often requires high-dimensional integration.

One widely used approach to overcome this challenge is Markov Chain Monte Carlo (MCMC). In MCMC, we construct a Markov chain whose stationary distribution is exactly the desired posterior $p(\theta|X)$. By simulating the chain for many iterations, the generated samples $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ can be used to approximate posterior quantities of interest (e.g., posterior means, variances, and credible intervals). This sampling-based method allows us to explore complex posterior landscapes where direct analytical solutions are infeasible [6, 7].

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is a sampling technique that allows us to approximate the posterior distribution $p(\theta|X)$ by constructing a Markov chain. A Markov chain is a sequence of states where the probability of each state depends only on the previous state. In MCMC, we generate samples $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ that, after a certain number of iterations, converge to the desired posterior distribution.

Two common algorithms for MCMC are:

- **Metropolis-Hastings**: Proposes new samples based on a proposal distribution and accepts or rejects the samples based on an acceptance criterion.

- **Gibbs Sampling**: Samples each variable conditionally on the others, iterating through all variables.

It is important to verify that the chain has reached convergence, meaning that the samples are representative of the posterior distribution. This can be done using convergence diagnostics.

MCMC is powerful because it can handle complex models where analytical solutions are not feasible. However, it requires a large number of samples and can be computationally expensive.

4 Parametric Bootstrap and Goodness of Fit

In this section, we delve into the concepts of parametric bootstrap and goodness of fit, which are crucial for robust statistical inference. These methods provide essential tools for assessing the variability of estimators and the adequacy of statistical models.

4.1 Parametric Bootstrap

The parametric bootstrap is a resampling technique used to estimate the distribution of a statistic by generating multiple samples from an estimated parametric model. This method involves the following steps:

1. Fit the parametric model to the observed data to obtain parameter estimates.
2. Generate a large number of bootstrap samples by resampling from the fitted model.
3. Compute the statistic of interest for each bootstrap sample to form an empirical distribution.

Mathematically, let $\hat{\theta}$ be the estimator of the parameter θ obtained from the original sample. The bootstrap samples $\{\theta_1^*, \theta_2^*, \ldots, \theta_B^*\}$ are generated by drawing from the distribution $F(\hat{\theta})$, where B is the number of bootstrap replications. The empirical distribution of $\{\theta_b^*\}_{b=1}^B$ provides an approximation of the sampling distribution of $\hat{\theta}$ [8, 9].

4.2 Goodness of Fit

Goodness of fit tests evaluate how well a statistical model fits the observed data by comparing the empirical distribution of the data to the theoretical distribution under the model. In this analysis, we employ both the Kolmogorov-Smirnov (KS) test and the Chi-square test, complemented by visual inspection methods, including histogram comparisons and quantile-quantile (Q-Q) plots.

The Kolmogorov-Smirnov test measures the maximum absolute difference between the empirical cumulative distribution function of the sample and the cumulative distribution function of the hypothesized model. It is particularly useful for continuous data and does not rely on binning, unlike the Chi-square test [9].

The Chi-square test, on the other hand, compares the observed and expected frequencies in binned data, making it suitable for categorical and discretized data. It evaluates whether deviations between the observed and expected frequencies are due to random fluctuations or systematic differences, providing an additional perspective on model fit.

In addition to formal statistical tests, we utilize graphical methods to assess the goodness of fit. A key approach is comparing the histogram of the observed data with the probability density function of the hypothesized model. A good fit is indicated when the empirical histogram closely aligns with the theoretical density curve.

Furthermore, the Q-Q plot provides a visual diagnostic tool by comparing the quantiles of the observed data against the quantiles of a specified theoretical distribution. If the points in the plot lie approximately along a 45-degree reference line, the assumed distribution provides a good fit. Deviations from the line suggest discrepancies between the observed and theoretical distributions [10].

By integrating the Kolmogorov-Smirnov test, the Chi-square test, and graphical methods such as histogram comparisons and Q-Q plots, we obtain a robust and comprehensive evaluation of how well the model represents the underlying data.

5 Results

5.1 Model Specification for EM

To model the fundamental frequency f_0 of vowel sounds, we assume that the observed data follow a bi-lognormal distribution, given prior knowledge in acoustic phonetics. The probability density function of a single lognormal component is defined as:

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), \tag{12}$$

where μ and σ are the mean and standard deviation of the underlying normal distribution in the logarithmic domain. A bi-lognormal mixture distribution is then formulated as:

$$f_0 \sim w f_1(x; \mu_1, \sigma_1) + (1 - w) f_2(x; \mu_2, \sigma_2), \tag{13}$$

where $w \in (0, 1)$ represents the mixing proportion between the two lognormal components.

Given the nature of the dataset, where frequencies are binned into intervals rather than recorded as continuous values, direct likelihood maximization is not feasible. Instead, we employ the Expectation-Maximization algorithm, leveraging jittering to handle the binning issue. The likelihood function for the observed data $\mathbf{X}_{obs}$, given the model parameters $\boldsymbol{\theta} = (w, \mu_1, \sigma_1, \mu_2, \sigma_2)$, is expressed as:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} P(X_{\text{obs},i} \mid \boldsymbol{\theta}), \tag{14}$$

where $P(X_{\text{obs},i} \mid \boldsymbol{\theta})$ is computed by integrating the bi-lognormal density over the bin intervals $[a_i, b_i]$:

$$P(X_{\text{obs},i} \mid \boldsymbol{\theta}) = \int_{a_i}^{b_i} f_0(x \mid \boldsymbol{\theta}) dx. \tag{15}$$

Since the likelihood function involves intractable integrals over binned intervals, we employ jittering in the EM framework. In the Expectation step (E-step), we replace the discrete observations with jittered values X_i^*, sampled uniformly from their respective bin intervals:

$$X_i^* = X_{\text{obs},i} + \epsilon_i, \quad \epsilon_i \sim U(a_i - X_{\text{obs},i}, b_i - X_{\text{obs},i}). \tag{16}$$

This transformation allows us to approximate a continuous likelihood function, improving the estimation process in the Maximization step (M-step). The EM iterations proceed by computing the expected complete log-likelihood and updating the parameters iteratively.

Parameter Initialization and Convergence Criteria The initial values for the parameters are set heuristically based on the logarithmic distribution of the jittered data. Specifically, we initialize:

- $\lambda^{(0)} = 0.5$, assuming equal mixing proportions initially.
- $\mu_1^{(0)} = \mathbb{E}[\log X^*] - 0.5$, $\mu_2^{(0)} = \mathbb{E}[\log X^*] + 0.5$, to ensure separation of the two lognormal modes.
- $\sigma_1^{(0)} = \sigma_2^{(0)} = \text{sd}(\log X^*)$, where sd denotes the sample standard deviation.

At each iteration, the parameters are updated based on the posterior probabilities:

$$\tau_i = \frac{\lambda f_1(X_i^*)}{\lambda f_1(X_i^*) + (1-\lambda) f_2(X_i^*)} \tag{17}$$

where τ_i represents the responsibility of the first component for observation X_i^*. Using this, the updated parameters are:

$$\lambda^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_i, \tag{18}$$

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^{N} \tau_i \log X_i^*}{\sum_{i=1}^{N} \tau_i}, \quad \sigma_1^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{N} \tau_i (\log X_i^* - \mu_1^{(t+1)})^2}{\sum_{i=1}^{N} \tau_i}}, \tag{19}$$

$$\mu_2^{(t+1)} = \frac{\sum_{i=1}^{N} (1-\tau_i) \log X_i^*}{\sum_{i=1}^{N} (1-\tau_i)}, \quad \sigma_2^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{N} (1-\tau_i)(\log X_i^* - \mu_2^{(t+1)})^2}{\sum_{i=1}^{N} (1-\tau_i)}}. \tag{20}$$

The algorithm stops when the change in log-likelihood is below a predefined tolerance threshold:

$$|\log \mathcal{L}^{(t+1)} - \log \mathcal{L}^{(t)}| < \varepsilon. \tag{21}$$

where we set $\varepsilon = 10^{-6}$ to ensure numerical stability.

Estimated Parameters from EM Algorithm After running the EM algorithm on the jittered dataset, we obtain the following estimated parameters:

- Estimated mixing proportion: $\lambda = 0.323$.
- First lognormal component: $\mu_1 = 4.86$, $\sigma_1 = 0.163$.
- Second lognormal component: $\mu_2 = 5.43$, $\sigma_2 = 0.118$.

5.2 Model Specification for Bayesian Approach

In the Bayesian framework, all model parameters are treated as random variables, incorporating prior knowledge into the inference process. Given the observed data $\mathbf{X}$, we seek to estimate the parameters $\boldsymbol{\theta} = (w, \mu_1, \sigma_1, \mu_2, \sigma_2)$ that define a bi-lognormal distribution:

$$f_0 \sim w f_1(x; \mu_1, \sigma_1) + (1 - w) f_2(x; \mu_2, \sigma_2), \tag{22}$$

where w is the mixing proportion between the two lognormal components. The likelihood function is given by:

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[w f_1(X_i; \mu_1, \sigma_1) + (1 - w) f_2(X_i; \mu_2, \sigma_2) \right]. \tag{23}$$

Using Bayes' theorem, we combine this likelihood with prior distributions $p(\boldsymbol{\theta})$ to obtain the posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{X})}. \tag{24}$$

Since direct evaluation of the marginal likelihood $p(\mathbf{X})$ is intractable, we employ a Markov Chain Monte Carlo (MCMC) approach to sample from the posterior distribution.

MCMC Implementation Using Gibbs Sampling We implement a Gibbs sampling scheme to iteratively update each parameter while conditioning on the current values of the others. The sampling process follows these steps:

- **Mixing proportion:** The parameter w follows a Beta posterior, updated as:

$$w^{(t+1)} \sim \text{Beta}\left(1 + \sum_i \tau_i, 1 + \sum_i (1 - \tau_i)\right), \tag{25}$$

  where τ_i represents the posterior probability that observation X_i belongs to the first lognormal component, given the current parameter estimates. It is computed as:

$$\tau_i = P(Z_i = 1 \mid X_i, \boldsymbol{\theta}) = \frac{w f_1(X_i; \mu_1, \sigma_1)}{w f_1(X_i; \mu_1, \sigma_1) + (1 - w) f_2(X_i; \mu_2, \sigma_2)}, \tag{26}$$

  where Z_i is the latent variable indicating component membership, and f_1 and f_2 are the lognormal density functions. Since the prior distribution for w is assumed to be $\text{Beta}(1, 1)$, which is the standard non-informative prior for a probability parameter, the posterior update retains the Beta form due to conjugacy with the binomial likelihood.

- **Mean parameters:** Given the latent assignments, the posterior distributions for μ_1 and μ_2 are normal:

$$\mu_1^{(t+1)} \sim \mathcal{N}\left(\frac{\sum_i \tau_i \log X_i}{\sum_i \tau_i}, \frac{1}{\sum_i \tau_i}\right), \tag{27}$$

$$\mu_2^{(t+1)} \sim \mathcal{N}\left(\frac{\sum_i (1 - \tau_i) \log X_i}{\sum_i (1 - \tau_i)}, \frac{1}{\sum_i (1 - \tau_i)}\right). \tag{28}$$

  The normal posterior arises from the assumption that μ_1 and μ_2 follow normal priors and the log-transformed data within each component is approximately normally distributed. Given a normal prior $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and normally distributed observations with known variance, the conjugacy property ensures that the posterior mean follows a normal distribution.

- **Variance parameters:** The posteriors for σ_1^2 and σ_2^2 follow inverse gamma distributions:

$$\sigma_1^{2(t+1)} \sim \text{Inv-Gamma}\left(\frac{\sum_i \tau_i}{2}, \frac{\sum_i \tau_i (\log X_i - \mu_1)^2}{2}\right), \tag{29}$$

$$\sigma_2^{2(t+1)} \sim \text{Inv-Gamma}\left(\frac{\sum_i (1 - \tau_i)}{2}, \frac{\sum_i (1 - \tau_i)(\log X_i - \mu_2)^2}{2}\right). \tag{30}$$

The inverse gamma posterior results from assuming an inverse gamma prior on σ_k^2, which is a common choice due to its conjugacy with the normal likelihood when the variance is unknown. Specifically, if $\sigma_k^2 \sim \text{Inv-Gamma}(\alpha, \beta)$ and the likelihood follows a normal distribution, the posterior remains inverse gamma.

The Markov chain is initialized with reasonable starting values (analogously to what was done for EM):

- $\lambda^{(0)} = 0.5$, assuming equal component weights.
- $\mu_1^{(0)} = \mathbb{E}[\log X] - 0.5$, $\mu_2^{(0)} = \mathbb{E}[\log X] + 0.5$.
- $\sigma_1^{(0)} = \sigma_2^{(0)} = \text{sd}(\log X)$.

We run the Gibbs sampler for 5000 iterations, discarding the first 1000 as burn-in.

Estimated Parameters from Bayesian Inference After running the Gibbs sampler, we obtain the following posterior estimates for the model parameters:

- Estimated mixing proportion: $\lambda = 0.326$.
- First lognormal component: $\mu_1 = 4.86$, $\sigma_1 = 0.174$.
- Second lognormal component: $\mu_2 = 5.43$, $\sigma_2 = 0.123$.

These estimates are consistent with the results obtained via Expectation-Maximization, indicating that both methods provide a robust characterization of the bi-lognormal distribution. The Bayesian approach, however, offers additional insights into parameter uncertainty by providing full posterior distributions rather than point estimates.

5.3 Model Validation

To assess the reliability of the estimated bi-lognormal model under both the Expectation-Maximization and Bayesian approach, we perform a combination of statistical tests and graphical diagnostics. The validation consists of:

- A **Chi-square test** to evaluate the agreement between observed and expected frequencies.
- A **Kolmogorov-Smirnov test** to compare empirical and theoretical cumulative distributions.
- **Parametric bootstrap resampling** to compute empirical p-values for both tests.
- **Graphical assessments**, including histogram overlays, Q-Q plots, and density comparisons.

5.3.1 Statistical Validation

Chi-square Test The Chi-square test examines how well the observed frequencies align with those expected under the estimated bi-lognormal model. The test statistic is defined as:

$$\chi^2_{\text{obs}} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}, \tag{31}$$

where O_k and E_k are the observed and expected counts in each bin.

Bootstrap Estimation of the Chi-square p-value Since binning can introduce artifacts, we apply a parametric bootstrap procedure to estimate the empirical p-value:

$$p_{\chi^2} = \frac{1}{B} \sum_{b=1}^{B} \nVdash(\chi^2_b \geq \chi^2_{\text{obs}}). \tag{32}$$

For the EM model, the bootstrap p-value is $p_{\chi^2} = 0.449$, indicating a reasonable model fit. The Bayesian model exhibits an even stronger fit, with $p_{\chi^2} = 0.817$, further supporting its adequacy.

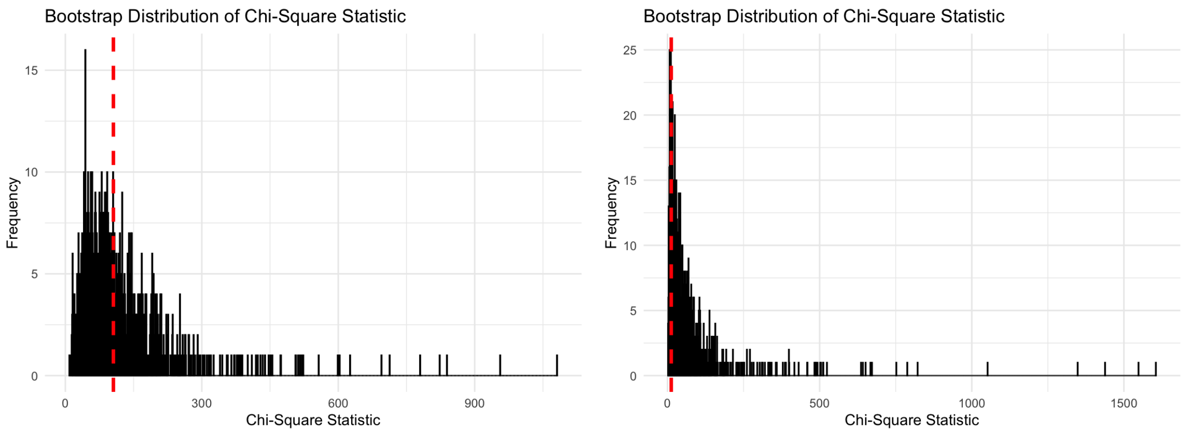

Figure 3: Bootstrap distribution of the Chi-square statistic for the EM (left) and Bayesian (right) models. The red dashed lines indicate observed statistics.

Kolmogorov-Smirnov Test The Kolmogorov-Smirnov (KS) test measures the maximum absolute difference between the empirical and theoretical cumulative distribution functions:

$$D_N = \sup_x |\hat{F}_N(x) - F(x|\hat{\boldsymbol{\theta}})|. \tag{33}$$

As before, a bootstrap-based estimation is applied:

$$p_{\text{KS}} = \frac{1}{B} \sum_{b=1}^{B} \nVdash(D_N^{(b)} \geq D_N^{\text{obs}}). \tag{34}$$

For the EM model, the bootstrap p-value is $p_{\text{KS}} = 0.118$, suggesting an acceptable model fit. The Bayesian approach performs even better, yielding $p_{\text{KS}} = 0.839$, indicating a closer alignment with the empirical data.

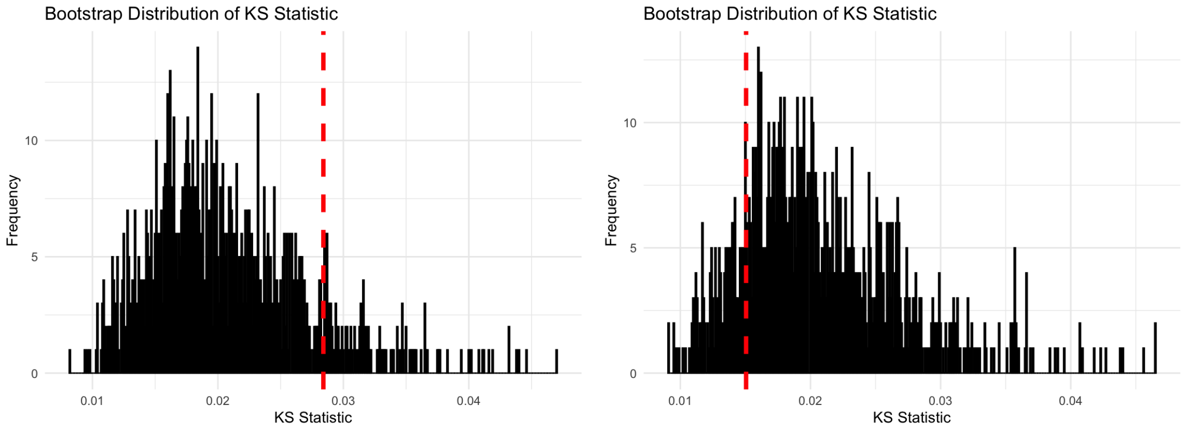

Figure 4: Bootstrap distribution of the KS statistic for the EM (left) and Bayesian (right) models. The red dashed lines indicate observed statistics.

5.3.2 Graphical Diagnostics

Histogram and Density Comparison A key graphical diagnostic is the histogram overlay, where the empirical data distribution is compared against the estimated bi-lognormal density. As seen in Figure 5, both models closely capture the bimodal structure of the data.

Quantile-Quantile (Q-Q) Plot The Q-Q plot compares empirical quantiles of the observed data with those generated from the fitted models. As shown in Figure 6, both approaches exhibit a strong

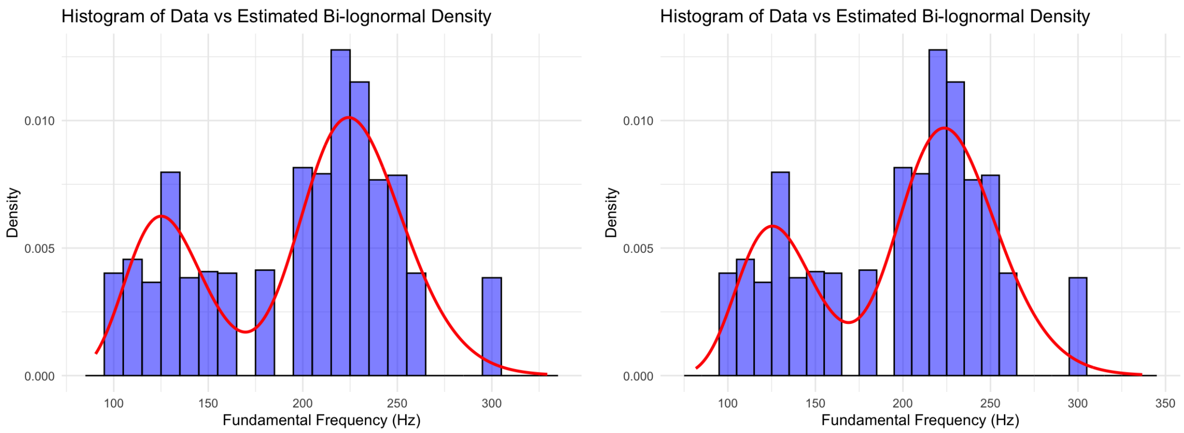

Figure 5: Histogram of the observed data with estimated bi-lognormal density for EM (left) and Bayesian (right) models.

agreement, though the Bayesian model appears slightly more aligned. However, a notable discrepancy emerges in the upper tail of the distribution, where both models exhibit deviations from the theoretical quantiles. This suggests a potential limitation in capturing the right-tail behavior of the empirical distribution, which may explain some of the imperfections observed in the goodness-of-fit statistics. The overestimation or underestimation of tail probabilities could contribute to slight deviations in test statistics, despite the overall adequacy of the models.

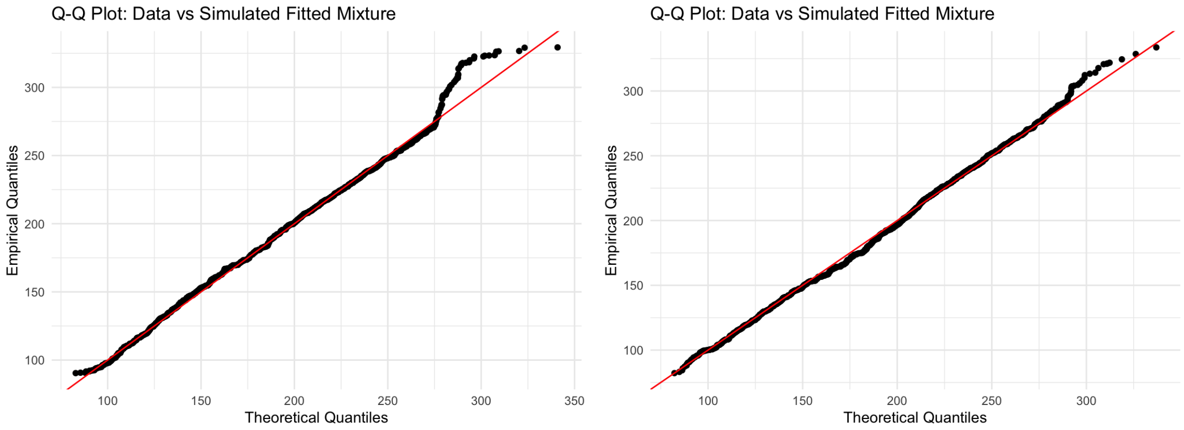

Figure 6: Q-Q plot comparing empirical quantiles to those of the fitted bi-lognormal model for EM (left) and Bayesian (right) approaches.

5.3.3 Comparison of EM and Bayesian Approaches

Both the EM and Bayesian models provide an adequate fit to the data, as indicated by the statistical and graphical analyses:

- The **Chi-square test** does not reject the null hypothesis for either method, but the Bayesian model achieves a higher p-value ($p_{\chi^2} = 0.817$ vs. $p_{\chi^2} = 0.449$), suggesting a stronger overall fit.
- The **Kolmogorov-Smirnov test** similarly confirms model adequacy, with the Bayesian approach again exhibiting a higher p-value ($p_{\mathrm{KS}} = 0.839$ vs. $p_{\mathrm{KS}} = 0.118$).
- The **graphical diagnostics** further validate these findings, with the Bayesian model displaying a closer adherence to the empirical data distribution.
- However, the **Q-Q plot** analysis reveals a systematic discrepancy in the right tail, which is more pronounced in the EM model. This suggests that while the bi-lognormal assumption is effective

in modeling the central distribution, the EM approach may be less robust in capturing extreme values compared to the Bayesian model.

While both models are effective in capturing the bi-lognormal structure, the Bayesian approach demonstrates slightly superior performance, likely due to the incorporation of prior knowledge, leading to more stable parameter estimates.

6 Conclusion

This study rigorously validated statistical modeling of vowel fundamental frequency using a bi-lognormal distribution, a firm theoretical hypothesis in acoustic phonetics. We accomplished this by employing two complementary approaches: the Expectation-Maximization algorithm with jittering and Bayesian inference. Both methods accurately estimated the parameters of the bi-lognormal distribution and validated its applicability in representing the empirical distribution of vowel frequencies.

The EM algorithm produced maximum likelihood estimates by iteratively refining the parameters, handling the discretized data effectively through jittering. This approach was computationally effective and robust in parameter estimation. Meanwhile, the Bayesian method, carried out with Markov Chain Monte Carlo, offered a probabilistic characterization of the parameters, producing posterior distributions that quantified estimation uncertainty.

Validation tests like the Kolmogorov-Smirnov and Chi-square goodness-of-fit tests confirmed that the bi-lognormal model is a statistically suitable representation of the data. The Bayesian method yielded slightly superior goodness-of-fit statistics, closer to the observed distribution, particularly the nuances of the data structure. Graphical inspections, including histogram overlays and Q-Q plots, also corroborated these results, further confirming the bi-lognormal assumption.

In conclusion, this study validates the theoretical suitability of the bi-lognormal distribution to describe vowel fundamental frequency. The efficiency of both EM and Bayesian methods in parameter estimation for this use demonstrates them as suitable for this task. These findings contribute to statistical speech analysis a robust and tested methodology that can be applied in future phonetic studies and applications to speech processing.

References

[1] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

[2] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

[3] McLachlan, G. J., & Krishnan, T. (2008). *The EM Algorithm and Extensions* (2nd ed.). John Wiley & Sons.

[4] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

[5] Diggle, P. J., & Gratton, R. J. (1987). *Monte Carlo methods of inference for implicit statistical models.* Journal of the Royal Statistical Society: Series B (Methodological), 49(2), 193-227.

[6] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis (3rd ed.).* Chapman & Hall/CRC, 2013.

[7] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd ed.).* Springer, 2004.

[8] Davison, A. C., & Hinkley, D. V. (2009). *Bootstrap Methods and their Application.* Cambridge University Press.

[9] Wasserman, L. (2005). *All of Nonparametric Statistics.* Springer.

[10] Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap.* Springer.