Bayesian Analysis of Phoneme Confusion Matrices

Bayesian Analysis of Phoneme Confusion Matrices

Arne Leijon, *Member, IEEE*, Gustav Eje Henter, *Member, IEEE*, Martin Dahlquist.

***Abstract*—This paper presents a parametric Bayesian approach to the statistical analysis of phoneme confusion matrices measured for groups of individual listeners in one or more test conditions.**

Two different bias problems in conventional estimation of mutual information are analyzed and explained theoretically. Evaluations with synthetic datasets indicate that the proposed Bayesian method can give satisfactory estimates of mutual information and response probabilities, even for phoneme confusion tests using a very small number of test items for each phoneme category.

The proposed method can reveal overall differences in performance between two test conditions with better power than conventional Wilcoxon significance tests or conventional confidence intervals. The method can also identify sets of confusion-matrix cells that are credibly different between two test conditions, with better power than a similar approximate frequentist method.

***Index Terms*—Speech recognition, parameter estimation, mutual information, Bayes methods**

I. Introduction

PHONEME confusions that arise when humans listen to distorted or noise-corrupted speech is a central topic in the field of communication acoustics, as first established in the early 1900s. An important goal of early studies was to determine necessary technical requirements on the telephone speech transmission network [1]–[3], but phoneme confusions continue to be studied also today. The topic is relevant for many practical purposes, since the early studies also showed that phoneme identification of nonsense syllables is positively correlated with understanding the meaning of complete sentences, e.g., [4, Fig. 11].

The present study is aimed at the following experimental scenario: The goal is to find out whether a new signal-processing system provides better phoneme recognition than a state-of-the-art reference system. To this end, phoneme identification ability is measured for a group of participants. For example, each participant might listen to a speech test material using different speech-coding algorithms in a cochlear implant system, or hearing aids adjusted using different fitting principles.

The experimenter wants to answer these research questions:

- Can we safely conclude that the new system improves phoneme recognition, compared to the reference system?
- What degree of improvement is to be expected in the population from which the test participants were recruited?
- For which phonemes, and for which types of identification errors, can we safely conclude that the new system improves performance?

The experimenter must choose a single-number measure to quantify the overall phoneme identification ability. The most obvious choice, used in a vast number of studies, is to simply report the percentage Score of Correct responses (SC) with the given speech test material.

Further insight can be obtained if the recognition results are organized in the form of a confusion matrix, displaying for each stimulus category the number of responses in each of the possible response categories. Typically, a confusion matrix for consonant recognition in noise reveals many errors among acoustically similar consonants such as /p/, /t/, /k/, or among nasals, /m/, /n/, but much fewer confusions between more dissimilar consonants, e.g., /s/ vs. /m/.

Miller & Nicely [5] introduced the Mutual Information (MI) between stimulus and response categories[1] as an interesting overall performance measure. The mutual information quantifies the average amount of information reeived by the listener about the stimulus phonemes [6]. This measure is appealing because information transfer is the fundamental goal of communication.

There are good reasons to use both SC and MI as performance measures. The SC is easy to understand and also has one notable statistical advantage: The observed SC is an unbiased and consistent estimate of the underlying Probability of Correct response (PC). This means that, if the same test would be repeated many more times with other listeners in the population for which the test group is representative, the average test result will converge towards the true probability in the population.

The MI measure does not only consider the number of correct identifications, but also takes into account all the types of identification errors that the listener can make. Theoretically, MI can be maximal while PC is minimal (zero), e.g., if recognized phoneme labels are cyclically permuted compared to the true labels. Thus, MI and PC are two fundamentally different concepts. An acoustic signal transformation that improves the MI might improve the long-term PC, although the PC initially decreases. This result occurs if the listener consistently misinterprets the presented phonemes with the new signal processing, but the identification errors are more systematic with the new system and more random in the reference condition. Such a result would suggest that the new processing actually makes it easier for the listener to distinguish between the different presented sounds, but the

A. Leijon and M. Dahlquist are affiliated with ORCA-Europe/Widex, Stockholm, Sweden. A. Leijon recently retired from the School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. E-mail: leijon@kth.se

G. E. Henter is with the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK.

Manuscript received May 18, 2015; revised Sep 6, 2015, and Nov 29, 2015.

[1] In [5], the MI was calculated not only as a single measure for all consonants, but also separately for five articulatory *distinctive features*. However, this paper uses the MI only as a single-number measure of the total transfer of phonemic information.

listener's *labeling* of the sound categories is still somewhat incoherent.

The MI quantifies the ability to *discriminate* between the test stimuli, which is a necessary but not sufficient condition for correct *identification*, which is quantified by the PC. Therefore, improved MI should be considered as a promising result, even if the PC does not immediately improve. Of course, additional experiments are then still needed to indicate if listeners actually can learn to benefit from the new processing after sufficient acclimatization.

Well-established conventional methods are available for the analysis of the statistical reliability of speech recognition test results with words or sentences [7]–[9]. If the items in a recognition test are statistically independent and equally difficult, the number of correct responses follows a binomial distribution. Even if the stimulus items are not exactly equally difficult, the binomial distribution can still be applied to estimate the test-retest reliability [7].

However, consonant recognition tests with nonsense syllables usually exhibit a very wide range of recognition performance among the different stimulus phonemes, even for listeners with normal hearing [5], [10]–[12]. These variations are even larger among listeners with impaired hearing [13]–[16]. Thus, when consonant identification tests are used to evaluate a new signal processing algorithm, an improvement may be seen only for a small subset of the presented consonants. This variability makes it difficult to apply the classical method to analyze the reliability of the overall SC in consonant recognition data.

The binomial (or multinomial) model is appropriate for data in each cell of a confusion-count matrix, and statistical estimation of the unknown proportion parameter of the binomial distribution has been studied extensively, e.g., [17]–[21]. Phoneme confusions have been measured in a very large number of studies; see, e.g., recent reviews in [16], [22], [23].

However, surprisingly few studies have applied the multinomial model to analyze the statistical reliability of observed confusion matrices. Jürgens et al. [24] assumed a multinomial distribution of response counts and estimated confidence intervals for response probabilities using the Clopper-Pearson method [17], [21]. A similar approach was used in [14], [15]. Multiple joint comparisons for several matrix cells were not considered in these studies. Such comparisons are more complicated, firstly because of the general problem of multiple hypothesis testing, and secondly because the counts of different responses to the same stimulus category are statistically dependent.

Conventional statistical tests for the significance of an observed difference in PC or MI between test conditions must use the observed variations among individual results to estimate the reliability. Parametric test methods such as ANOVA have been applied, e.g., in [22], [25], [26]. When PC and MI results are close to their upper or lower limits, it is obviously questionable to assume that data follow a Gaussian distribution. A non-parametric test such as the Wilcoxon signed-rank test might generally be more appropriate.

Unfortunately, using the MI as an overall performance measure raises an additional difficulty: The conventional MI calculation over-estimates the true individual value, unless the number of test items is very large [5], [27]–[29]. Thus, if individual MI results are first estimated for each test participant, and the results then are averaged across listeners, an overestimation bias remains in the final average.

Many studies of phoneme confusions have simply avoided estimation of individual MI results. Instead, following [5], it has been common practice to pool the individual results by first adding the confusion matrices for all listeners, before estimating a single MI value [29]. Sagi and Svirsky [29] suggested that the biased individual estimates might still be used for statistical comparisons, assuming that the bias is constant. However, the present study shows that this assumption is false. We also show that pooling leads to another type of bias that has previously been overlooked.

The present paper provides several novel contributions that offer a solution to the problems mentioned above:

- A parametric Bayesian framework is presented for the analysis of phoneme confusion data.
- The proposed method shows more accurate results than conventional statistical methods when applied to data from simulated consonant recognition tests.
- Two different kinds of MI bias are explained theoretically and quantified by simulations. The proposed method nearly eliminates both types of bias.
- The proposed approach can also determine the statistical credibility of observed changes jointly for multiple cells in different confusion matrices.

The paper is organized as follows: Section II defines the theoretical approach. Proofs of the two MI bias problems are given in Appendix A. Other mathematical details are deferred to later appendices. Sections III and IV describe simulations showing that the proposed method improves on the performance of the conventional approach. Finally, the new contributions are discussed and summarized in Secs. V and VI.

II. Theory

A. Bayesian vs. Frequentist Hypothesis Testing

Assume that we have collected recognition scores SC_A and SC_B from two different test conditions, A and B, for example using two different signal processing algorithms in a hearing aid. The experimental results might be $SC_A = 66\%$ and $SC_B = 74\%$. We need to determine whether the observed improvement $D = SC_B - SC_A$ is caused by a genuine advantage of B over A, or if the difference might just have been caused by random variations in the phoneme identification test.

The well-known classical frequentist approach would then be to assume a *null hypothesis*, stating that the underlying probabilities of correct recognition PC_A and PC_B are *equal* in both conditions. Using the null hypothesis, the total probability p is calculated for all possible random results as extreme as, or more extreme than, the observed difference. If this probability is smaller than a pre-selected *signficance level* (e.g., 0.05), the null hypothesis is rejected, and the researcher concludes that it is acceptable to act as if the observed difference is known to be real.

By convention, researchers are usually satisfied if they find a significant result indicating that PC_B is better than PC_A, although the classical method actually does not calculate the probability that $PC_A > PC_B$. This and other weaknesses in the frequentist methods are discussed in depth in [30, Ch. 5].

There is always a scientific (and perhaps financial) cost incurred by acting as if an improvement is real when, actually, there is no improvement, and the calculated p-value can be used to estimate the expected value of this cost. However, there is always also a risk that we might incorrectly conclude that there is no real difference between A and B, when there is in fact a difference. The simple classical significance test does not quantify the risk for this opposite "type-II error," although this error may actually be more costly in many scenarios.

The Bayesian approach quantifies the probability of the tested hypotheses and both types of error, by estimating a probability distribution for all underlying model parameters, given the observed experimental result. In the example discussed above, the Bayesian analysis would estimate the conditional probability q for the event that $PC_B > PC_A$, given the observed data. This result implies that the probability is $1-q$ for the opposite event that $PC_B \leq PC_A$. The same approach is used for any other performance measure, e.g., the MI or the response probability in any confusion-matrix cell.

How should the researcher interpret this q-value? Obviously, a very large q of 0.99 or so strongly suggests that B is genuinely superior to A, and vice versa if q is very small. However, it is not possible to define a fixed decision threshold. For example, if there is no other information to point in either direction, the researcher should act as if B really is better than A whenever the probability $q > 0.5$.

Interesting philosophical discussions of the Bayesian vs. frequentist viewpoints are given in [30], [31]. Bishop's textbook [32] presents many applications of the Bayesian approach for machine learning.

B. Notation and Definitions

The primary result from a phoneme-identification test, for one listener in one test condition, is a confusion-count matrix $\boldsymbol{x}$ with one row for each stimulus category and one column for each response alternative. Each matrix element x_{sr} is an integer counting the number of events where a stimulus of category s was presented and the listener responded by category r. In many forced-choice experiments the confusion matrix is square, but there may also be one or more additional columns for "no response" or "noise only" as in, e.g., [10]. If allowed in the experiment, such responses are regarded here as legitimate alternatives and treated statistically just like the responses that identify a phoneme. The correct-response counts are then found in the elements along the main diagonal. In some experiments, e.g., [12], [16], different tokens (phones) are presented for each phoneme category, and the same response category is considered correct for these different stimulus variants. Then the responses considered correct are not on the main diagonal. In any case, the number of presented test items from the sth stimulus category is

$$N_s = \sum_{r \in \mathcal{R}} x_{sr} \tag{1}$$

and the total number of presented test items is $N = \sum_{s \in \mathcal{S}} N_s$. Here, $\mathcal{S}$ is the set of all stimulus categories, and $\mathcal{R}$ is the set of all allowed responses.

When all confusion-count matrices from an experiment have been collected, the complete dataset will be denoted $\underline{\boldsymbol{x}} = (\ldots, \boldsymbol{x}_{lt}, \ldots)$, where $\boldsymbol{x}_{lt}$ is the count matrix for the lth listener in the tth test condition.[2]

Here and forthwith, matrices and vectors are written with bold symbols. Random variables are denoted by upper-case letters, while lower-case letters are used for specific outcome values of the random variable.

In both the classical frequentist as well as the Bayesian approach, the observed response counts are considered to be outcomes of discrete random variables. It is assumed that when a stimulus of type s is presented, the listener will respond by alternative r with some probability, here called u_{sr}. This conditional probability is a real number between 0 and 1. Of course, the exact value of each u_{sr} is not known. It is a model assumption that all these probabilities have fixed values for each listener in each test condition.

An observed confusion-count matrix $\boldsymbol{x}$, with rows $\boldsymbol{x}_s$ and elements x_{sr}, is seen as just one outcome of a matrix-valued random variable $\boldsymbol{X}$ with a multinomial probability mass

$$P\left[\boldsymbol{X} = \boldsymbol{x}\right] = \prod_{s \in \mathcal{S}} c(\boldsymbol{x}_s) \prod_{r \in \mathcal{R}} u_{sr}^{x_{sr}}; \quad c(\boldsymbol{x}_s) = \frac{N_s!}{x_{s1}! \cdots x_{s|\mathcal{R}|}!} \tag{2}$$

for any event $\boldsymbol{X} = \boldsymbol{x}$. The responses to all items are assumed to be conditionally independent,[3] given the probabilities u_{sr}. If only two response alternatives are considered, i.e., $|\mathcal{R}| = 2$, expression (2) reduces to a product of binomial distributions.

The expected value of the response count is $E\left[X_{sr}\right] = N_s u_{sr}$ for any stimulus-response pair (s, r), and the variance is $\mathrm{var}\left[X_{sr}\right] = N_s u_{sr}(1 - u_{sr})$. Thus, the expectation and variance for the relative response rates are $E\left[X_{sr}/N_s\right] = u_{sr}$ and $\mathrm{var}\left[X_{sr}/N_s\right] = u_{sr}(1-u_{sr})/N_s$. The counts within each matrix row are correlated, with $\mathrm{cov}\left[X_{sr}, X_{sq}\right] = -N_s u_{sr} u_{sq}$, for $q \neq r$, but elements in different rows are independent.

In the frequentist view, the parameters u_{sr} are simply numbers, and not endowed with a probability distribution. The frequentist model can therefore only calculate the probability of events $\boldsymbol{X} = \boldsymbol{x}$ given $\boldsymbol{u}$.

The Bayesian approach, in contrast, regards the individual response-probability matrix $\boldsymbol{u}$ as an outcome of a matrix-valued random variable, here called $\boldsymbol{U}$. The goal of the Bayesian approach is to estimate a conditional probability distribution for $\boldsymbol{U}$, given the observed outcome $\boldsymbol{X} = \boldsymbol{x}$.

In the Bayesian view, the conditional probability mass for an observed event $\boldsymbol{X} = \boldsymbol{x}$, given an outcome $\boldsymbol{u}$ for the parameters $\boldsymbol{U}$, can now be expressed as

$$p_{\boldsymbol{X}|\boldsymbol{U}}(\boldsymbol{x} \mid \boldsymbol{u}) \propto \prod_{s \in \mathcal{S}} \prod_{r \in \mathcal{R}} u_{sr}^{x_{sr}}. \tag{3}$$

[2]To avoid clutter, the listener and condition indices will be omitted when not essential to the discussion.

[3]This assumption is appropriate, because phoneme identification tests usually use nonsense speech material. This makes the results most sensitive to the acoustic characteristics of the speech signal and eliminates influence from syntactic and semantic context.

Here, the normalization factors $c(\boldsymbol{x}_s)$ in (2) have been omitted, because the observed $\boldsymbol{x}$ is from now on regarded as fixed, and because (3) only will be used to find a conditional probability density for $\boldsymbol{U}$, given the observed $\boldsymbol{x}$.

The random-variable elements U_{sr} of the parameter matrix $\boldsymbol{U}$ are defined as *conditional* probabilities for the random response variable R, given the random stimulus variable S,

$$U_{sr} = P\left[R = r \mid S = s\right]; \quad \sum_{r\in\mathcal{R}} U_{sr} \equiv 1, \quad \text{for all } s, \quad (4)$$

since in most phoneme recognition tests the number of stimulus items N_s in category s is not the result of random choice, but is determined in advance by the phoneme test material.

To complete the Bayesian model specification, a prior probability density function $p_{\boldsymbol{U}}(\boldsymbol{u})$ is required. This density summarizes all information available about the distribution of $\boldsymbol{U}$ in the population of interest, before considering the observed data. In some situations, such prior information can be gleaned from previous studies, but usually the prior density function is not known in advance. Methods to select the prior distribution are discussed in Appendix B.

The core of the Bayesian approach is now to determine the posterior conditional probability density of the parameter matrix $\boldsymbol{U}$, given the observed data $\boldsymbol{x}$, as

$$p_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u} \mid \boldsymbol{x}) \propto p_{\boldsymbol{X}|\boldsymbol{U}}(\boldsymbol{x} \mid \boldsymbol{u})\, p_{\boldsymbol{U}}(\boldsymbol{u}). \quad (5)$$

The precise mathematical form of the density functions $p_{\boldsymbol{U}}(\boldsymbol{u})$ and $p_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u} \mid \boldsymbol{x})$ are defined in Appendix B.

If N_s is small, a wide range of probability values u_{sr} is compatible with the observed event $X_{sr} = x_{sr}$. An example of the result of applying (5) is presented in Fig. 2, showing conditional probability distributions for the response probabilities U_{sr} in three selected matrix cells for one listener.

C. Performance Measures

Once the posterior probability density $p_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u} \mid \boldsymbol{x})$ in (5) has been determined, it is straightforward to estimate the statistical properties of any performance measure that is a function of $\boldsymbol{U}$.

Several interesting such measures have been proposed in the literature: For example, [33] introduced *Confusion Patterns*, i.e., $\log U_{sr}$ as a function of signal-to-noise ratio (SNR), to illustrate phoneme confusions. The overall *log error probability* $\log(1 - PC(\boldsymbol{U}))$ was shown to have an interesting relation to the Articulation Index [15], [33]. The *Response Entropy* [15] quantifies the degree of response randomness.

This paper is focused on *response probabilities* in specific cells, total *probability of correct responses* (PC), and *mutual information* (MI), but the proposed estimation methods can also be used for other measures.

1) Response Probabilities: The posterior distribution of response probabilities can be summarized by the marginal means and/or quantiles for any element U_{sr}, independently of other matrix elements. The q-quantile $\widetilde{u}_{sr}(q)$ is defined from the cumulative posterior density as the solution to

$$P\left[U_{sr} \le \widetilde{u}_{sr}(q) \mid \boldsymbol{X} = \boldsymbol{x}\right] = q. \quad (6)$$

The median is the quantile $\widetilde{u}_{sr}(0.5)$.

2) Probability of Correct Response: The posterior total probability of Correct Response is a weighted average,

$$PC(\boldsymbol{U}) = \sum_{s\in\mathcal{S}} \frac{N_s}{N} U_{s,c(s)}, \quad (7)$$

where $c(s)$ is the response category defined as "correct" for stimulus s; usually $c(s) = s$. Quantiles for the PC are calculated in analogy with (6).

3) Mutual Information: The mutual information between stimulus and response random variables S and R, conventionally written as $I(S;R)$ in information-theory literature, indicates the average amount of information about the stimulus identity received by the listener for each test item. The MI quantifies the reduction in the listener's uncertainty about the stimulus label, achieved by hearing the presented sound.

Given any response-probability matrix outcome $\boldsymbol{u}$ of the random matrix $\boldsymbol{U}$, the MI is defined as

$$\begin{aligned} MI(\boldsymbol{u}) &= I(S;R \mid \boldsymbol{U} = \boldsymbol{u}) = \\ &= E_{R,S}\left[\log \frac{p(R \mid S, \boldsymbol{U} = \boldsymbol{u})}{p(R \mid \boldsymbol{U} = \boldsymbol{u})} \;\middle|\; \boldsymbol{U} = \boldsymbol{u}\right] = \\ &= \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}} \frac{N_s u_{sr}}{N} \log \frac{N u_{sr}}{\sum_{z\in\mathcal{S}} N_z u_{zr}}. \end{aligned} \quad (8)$$

If the test material includes different sets of phoneme tokens, e.g., for different speakers, the MI should be calculated to quantify the listener's ability to identify the phoneme class, regardless of the speaker. As the response patterns can differ consistently between token variants of the same phoneme [12], the response counts should not be averaged across variants of the same phoneme. Instead, the MI should first be estimated separately for each token set, and then averaged across sets if a single overall measure is needed.

Once the posterior distribution (5) for $\boldsymbol{U}$ has been determined, given an observed test result, MI quantiles $\widetilde{MI}(q)$ are defined in analogy with (6) as the solution to

$$P\left[MI(\boldsymbol{U}) \le \widetilde{MI}(q) \;\middle|\; \boldsymbol{X} = \boldsymbol{x}\right] = q. \quad (9)$$

The conventional estimate, introduced in [5], is obtained by plugging in the observed relative frequencies $\widehat{u}_{sr} = x_{sr}/N_s$ as point estimates of u_{sr} in (8). This MI estimate then becomes

$$\widehat{MI}(\boldsymbol{x}) = \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}} \frac{x_{sr}}{N} \log \frac{N x_{sr}}{N_s n_r}; \quad n_r = \sum_{z\in\mathcal{S}} x_{zr}. \quad (10)$$

D. Bias of Mutual-Information Estimates

It has been noted in the literature that (10) *overestimates* the true $MI(\boldsymbol{u})$ in (8) for individual listeners, if too few stimulus items are used in the test [5], [27]–[29]. The proof in Appendix A shows that the bias problem is caused by the fact that $MI(\boldsymbol{u})$ in (8) is a *convex* function of $\boldsymbol{u}$.[4]

A common alternative approach [5], [29] is to pool the results across test participants, i.e., to sum the confusion-count matrices for all listeners, before estimating a single MI value for the group of listeners. However, this pooling tends to

[4]Thus, the *Response Entropy* [15] suffers from the opposite bias problem, because it is a *concave* function of response probabilities [6, Theorem 2.7.3].

underestimate the true mean MI in the population from which the listeners were recruited. This paper is the first to identify this persistent underestimation property of the pooled analysis. Appendix A shows that this second type of bias cannot be reduced by increasing the number of test participants, or by increasing the number of test items for each listener, because the bias is caused by the inherent true variability among individuals in the population.

The main advantage of the Bayesian approach is that it estimates the complete posterior distribution of $\boldsymbol{U}$, instead of just a single point value as in the classical method. This makes it possible to reduce both kinds of bias, as shown in Sec. IV.

E. Difference Between Test Conditions

The Bayesian approach is easily extended to comparisons of performance results across test conditions.

1) Single-Number Performance Measure: Assume a pair of confusion-count matrices, $(\boldsymbol{x}, \boldsymbol{y})$, have been collected, e.g., for a single listener using two different transmission systems. A posterior density function for the corresponding pair $(\boldsymbol{U}, \boldsymbol{V})$ of response-probability matrices, $p_{\boldsymbol{U},\boldsymbol{V}|\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{u}, \boldsymbol{v} \mid \boldsymbol{x}, \boldsymbol{y})$, is then estimated using the principle in (5), given the observed pair of confusion-count matrices.

Once this joint posterior probability density for $(\boldsymbol{U}, \boldsymbol{V})$ has been calculated, it is straightforward to characterize the overall difference between test conditions by any measure defined as a function of the response-probability matrix. For example, the probability that the MI is higher in the second condition is

$$q = P\left[MI(\boldsymbol{V}) > MI(\boldsymbol{U}) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right] \qquad (11)$$

The interpretation of this q-value was discussed in Sec. II-A. Quantiles for the difference $MI(\boldsymbol{V}) - MI(\boldsymbol{U})$ are calculated in analogy with (6).

2) Responses in Multiple Cells: If an overall performance measure indicates a credible difference between test conditions, it is interesting to find a subset of stimulus-response pairs showing the most reliable differences. A conventional Bonferroni correction might be used to compensate for the effect of multiple hypothesis testing, but this approach may be too conservative as it does not account for the statistical dependence between cells in the confusion matrices.

Appendix C describes a computational procedure to find a subset of stimulus-response pairs that are all jointly credibly different between test conditions. As the probability is computed jointly for all the pairs, no further corrections for multiple hypothesis testing are needed.

F. Sampling Approximation

All the performance measures discussed in Sec. II-C can be defined as some function $g(\boldsymbol{U})$ of the random probability matrix $\boldsymbol{U}$. However, the exact posterior density function for $g(\boldsymbol{U})$ is usually not computationally tractable. Therefore, the distributions of these result measures are estimated by *sampling*. A large number of random sample matrices $\boldsymbol{u}(n)$, $n = 1, \ldots, N_u$, are drawn from the posterior density (5). The sample average then approximates the exact expectation, as

$$\bar{g} = \frac{1}{N_u} \sum_{n=1}^{N_u} g\left(\boldsymbol{u}\left(n\right)\right) \approx E\left[g(\boldsymbol{U})\right]. \qquad (12)$$

Distribution quantiles $\widetilde{g}(q)$ are estimated as the empirical quantiles in the set of samples, using linear interpolation between sorted sample values [34].

This approach is also applicable to entire datasets. Probability models $\boldsymbol{U}_{lt}$ are fitted for the lth participant in the tth test condition, using the confusion-count matrix $\boldsymbol{x}_{lt}$. Sample matrices $\boldsymbol{u}_{lt}(n)$ are generated from each posterior model $\boldsymbol{U}_{lt}$. The complete set of samples, for all listeners, are compared between test conditions in the same way as before. The results indicate the difference between test conditions for the entire population for which the participants are representative.

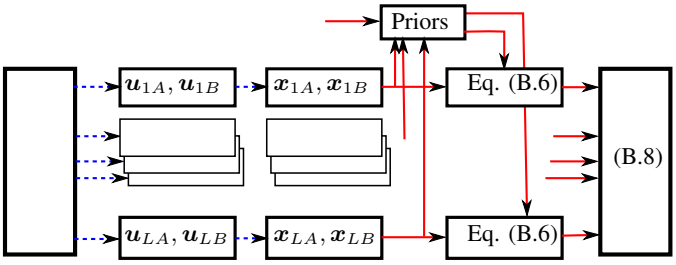

Fig. 1. Block diagram for the simulation of two test conditions, A and B. Pairs of response-probability matrices $(\boldsymbol{u}_{lA}, \boldsymbol{u}_{lB})$ and confusion-count matrices $(\boldsymbol{x}_{lA}, \boldsymbol{x}_{lB})$ are generated at random (dashed lines) for L listeners. The Bayesian analysis (solid lines) first adapts two prior distributions as described in Appendix B, then estimates L individual pairs of response-probability distributions by (B.6), and finally combines all results into an estimated group distribution of $(\boldsymbol{U}_A, \boldsymbol{U}_B)$ using (B.8).

III. Simulation Methods

To evaluate the proposed analysis methods, the calculated results must be compared to known true values, but the true values can only be known for synthetic data, necessitating the use of simulations.

Three levels of distributions were considered: (1) A *population* was simulated with known distributions of response-probability matrices in each test condition; (2) A *listener group* was drawn at random from the population model; (3) A single *confusion-count matrix* was generated at random from each listener model in each test condition. The complete procedure is illustrated in Fig. 1 for two test conditions.

1) Population: The distribution in the population was based on the extensive data in [5, Tables II, III, IV, and V] presenting square confusion-count matrices for 16 English consonants, pooled across five female normal-hearing listeners. The speech was presented with full frequency bandwidth in a background of flat-spectrum noise with nominal signal-to-noise ratios (SNRs) of $-12, -6, 0$, and $+6$ dB. Each matrix shows the result of 4000 test items in total, with 200–300 presentations for each consonant. Accurate mean response probabilities can therefore be estimated from these data, but the inter-individual variability is not known, because the published matrices were pooled across listeners and speakers.

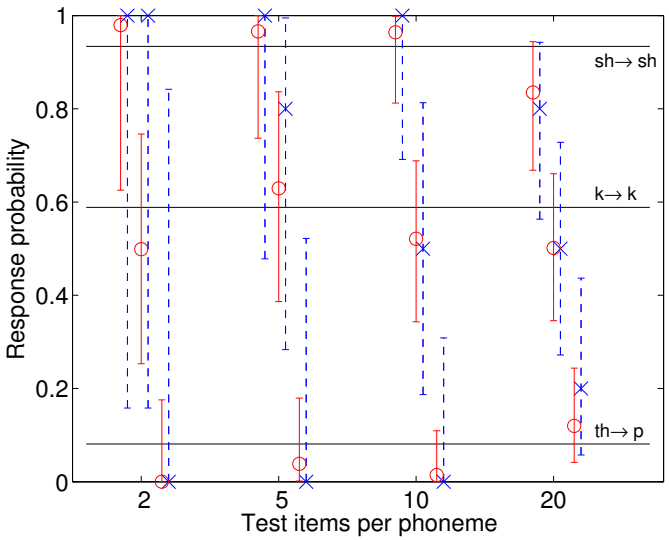

Fig. 2. Response probabilities in three selected confusion-matrix cells, sh→sh, k→k, and th→p, for a single listener selected at random from a group of $L = 20$ simulated listeners, drawn at random from the simulated population shown in Fig. 3. Bayesian estimates are shown by posterior medians (○) and 95-% credible intervals (vertical solid lines). Conventional results are shown by 95-% Clopper-Pearson confidence intervals (adjacent dashed vertical lines) and point estimates $\widehat{u}_{sr} = x_{sr}/N_s$ (×). All estimates should be compared to the known true response probabilities u_{sr} shown by solid horizontal lines.

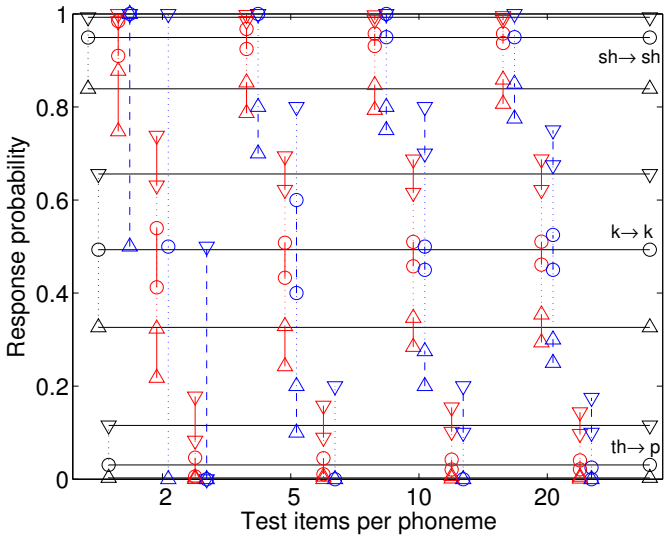

Fig. 3. Group estimates of posterior response-probability distributions in three selected cells (sh→sh, k→k, and th→p) of simulated confusion-count data for groups of $L = 20$ listeners tested with $N_s \in \{2, 5, 10, 20\}$ items per phoneme at a nominal SNR of 0 dB. The population distribution is shown by solid horizontal lines for the medians (○) and 90-% (▽) and 10-% (△) quantiles. These true population quantiles should be compared to the corresponding estimated results marked by the same symbols. Interquartile ranges among $R = 100$ random group replications are shown by vertical solid lines for the Bayesian estimated quantiles and by adjacent vertical dashed lines for the conventional sample quantiles.

The typical confusion patterns of normal-hearing listeners depend on the spectrum of the masking noise [10], [11]. The overall recognition scores as a function of SNR show rather small variations among normal-hearing listeners [10, Fig. 3], but there are notable systematic differences between consonant phonemes and also between different tokens (from different speakers) of the same phoneme category [12]. The variations in overall performance are much larger among listeners with impaired hearing, and the confusion patterns as a function of SNR can also differ systematically from those of normal-hearing listeners [16].

For the present model evaluation, the systematic differences between consonant phonemes are not critical, as they are automatically accounted for by the independent model distributions (B.6) for each stimulus category. The conditional independence assumption between test conditions (B.8) automatically accounts for any systematic differences between different token sets, e.g., as produced by different talkers. However, an attempt was made to create a rather large inter-individual variability in the simulated population, in the following way:

The distribution of response probabilities in the population was simulated by a mixture of $M = 30$ Dirichlet distributions, each representing a slightly different, randomly selected SNR. To simulate a test condition corresponding to a nominal SNR of s_t dB, a set of random SNR deviations $(z_1, \ldots, z_M)$ were first generated from a Gaussian distribution with zero mean and standard deviation $SD = 1$ dB. Mean response probabilities $\overline{\boldsymbol{u}}_{mt}$ were calculated for nominal SNRs $s_t + z_m$ by linear interpolation between the values estimated from confusion counts in [5, Tables II, III, IV, and V],[5] with extrapolation to chance performance at -20 dB SNR.

Dirichlet models $\boldsymbol{U}_{mt}$ were then constructed such that the mean response probabilities were equal to the interpolated means $\overline{\boldsymbol{u}}_{mt}$. The additional random variability around these means was controlled by setting each row sum of Dirichlet concentration parameters in (B.2) as $\sum_{r\in\mathcal{R}} a_{sr} = 15$.[6] The resulting "true" (reference) inter-individual variability can be seen in Figs. 3, 5, and 7.

2) Group: A group of L listeners was generated from the population model. First, L Dirichlet distributions were drawn at random from the M mixture components, and then the individual response-probability matrices $\boldsymbol{u}_{lt}$ were generated at random from the selected distributions. This procedure was repeated for each of R simulated groups. Examples of the "true" individual performance are shown in Figs. 2 and 6.

3) Confusion Counts: For each individual response-probability matrix $\boldsymbol{u}_{lt}$ a confusion-count matrix $\boldsymbol{x}_{lt}$ was generated at random from the multinomial distribution (2) with $N_s \in \{2, 5, 10, 20\}$ simulated presentations per phoneme.

IV. Simulation Results

The results in this section illustrate how the proposed method performs in a scenario where a group of participants has been tested with one confusion-count matrix measured for each listener in each test condition.

[5]Response probabilities were estimated from the published confusion counts by adding a pseudocount of 0.5 to each cell to prevent response probabilities from being exactly equal to 0 or 1.

[6]This implies, for example, that the standard deviation is about 0.125 in a cell with mean response probability 0.5.

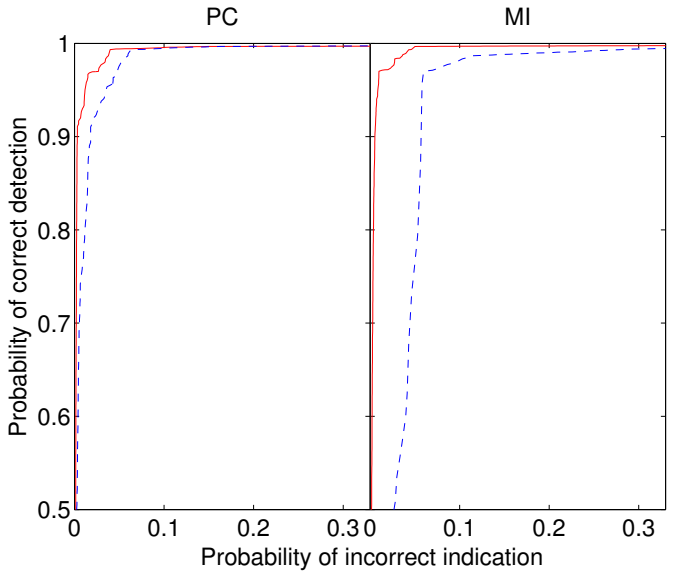

Fig. 4. Empirical Receiver Operating Characteristics (ROCs) for PC (left panel) and MI (right panel), showing the estimated probability of correctly detecting that performance is better in test condition B than in condition A, with nominal SNRs of -6 dB for A and -4 dB for B, vs. the probability of incorrect indication that B is better than A when the conditions are equal, with SNRs $(-6, -6)$ dB. The solid curves show results for the proposed Bayesian method, and the dashed curves show similar results using a one-sided Wilcoxon signed-rank test in a conventional frequentist approach. The tests are simulated with $N_s = 5$ test items per phoneme and $L = 20$ listeners in each group. Each ROC plot is based on $R = 300$ group simulations.

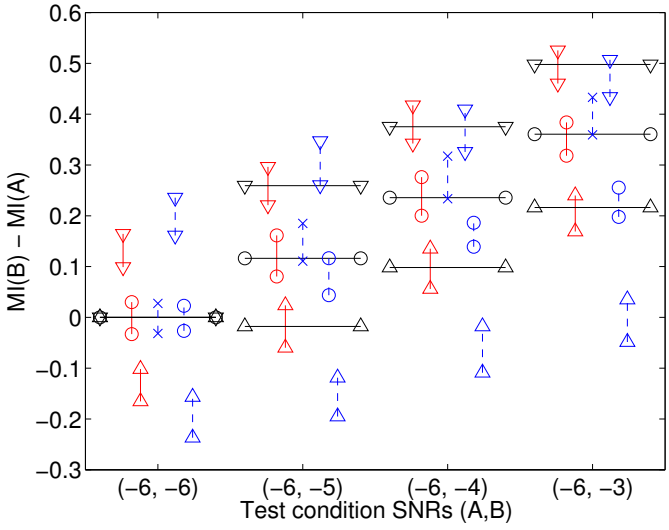

Fig. 5. Estimated improvement of Mutual Information (MI) from test condition A to condition B, simulated by nominal SNR pairs of $(-6, -6)$, $(-6, -5)$, $(-6, -4)$, and $(-6, -3)$ dB. The true MI differences in the population are shown by solid horizontal lines for the medians (○) and 90-% (▽) and 10-% (△) quantiles. The estimated quantiles are marked by the same symbols. Interquartile ranges among $R = 100$ replications are shown by vertical solid lines for the Bayesian quantiles and by adjacent vertical dashed lines for the empirical sample quantiles for the conventional results by Eq. (10). Interquartile ranges for the pooled group estimates are shown by ×. The simulations used $N_s = 5$ test items per phoneme and $L = 20$ listeners in each group simulation.

A. Single-cell Response Probabilities

1) Individual Results: Fig. 2 shows an example of estimated response probabilities in three stimulus-response cells of simulated confusion matrices for one listener randomly selected from a group of $L = 20$ listeners tested at a nominal SNR of 0 dB.

Although the observed response score is an unbiased and consistent estimate of the corresponding true probability, a particular observed result can be quite inaccurate just by chance, if too few test items are used. For example, in the matrix cell for $k \to k$, where the true response probability was about 0.6 for this listener, the observed score happened to be 100% with $N_s = 2$ test items per phoneme and 0.8 for $N_s = 5$. The estimated Bayesian credible interval ranged from 0.4 to 0.85 with $N_s = 5$ test items per phoneme.

The conventional Clopper-Pearson confidence intervals for the response probabilities [17], [21], based on the binomial distribution, are also shown in the figure. These are guaranteed to cover the true value with the specified confidence level (on average in infinitely many replications) but the interval width becomes so large for small N_s that the result is practically useless. Bayesian 95-% credible intervals,[7] while significantly narrower, included the true individual response probability in $\{91, 90, 90, 92\}$% of all 256 cells for $\{2, 5, 10, 20\}$ items per phoneme, across $L = 20$ listeners in $R = 100$ replications.

2) Group Results: Fig. 3 shows posterior distributions of response probabilities in three stimulus-response cells for groups of $L = 20$ listeners sampled repeatedly from a population representing a nominal SNR of 0 dB.

The Bayesian estimated quantiles are generally quite close to the corresponding true values for the population. When very few test items are used, the test results can vary widely among listeners. For example, with $N_s = 2$, the median probability for $k \to k$ was about 0.5 in the population, but individual sample scores are often 0 or 100% by chance alone. Of course, with $N_s = 2$ the only possible individual response scores are 0, 0.5, or 1.

The true variability across listeners in the population is the same regardless of how many test items are used. Thus, ideally, the estimated quantiles should not change with the number of test items.

Some of the individual Bayesian medians (○) in Fig. 2 deviate from the observed score (×) towards the average group results shown in Fig. 3. This is caused by the group-dependent adaptation of the prior distribution described in Appendix B-4.

B. Two Test Conditions—Small Differences

For practical purposes, the most interesting situation is when confusion-matrix data have been collected for a group of listeners in two slightly different test conditions. This section considers the problem of how to determine statistically whether the performance is better in the second condition, in relation to the true difference between the conditions.

1) Hypothesis Test: As discussed in Sec. II-A, the conventional method to compare performance results from two test conditions would be to apply a classical frequentist significance test to the measured data. Fig. 4 contrasts the

[7]Estimated interval limits were extended by 0.0001 to allow for sampling inaccuracy in cells with extremely small response probabilities.

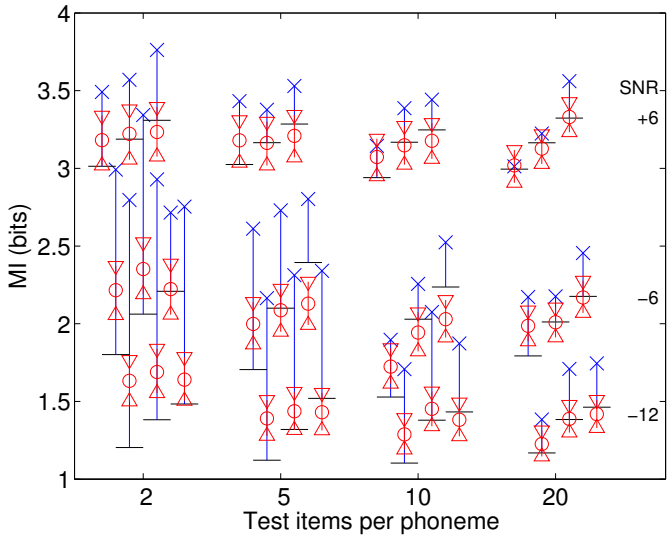

Fig. 6. Individual estimates of mutual information (MI) for simulated confusion-count data with $N_s \in \{2, 5, 10, 20\}$ test items per phoneme, in three different test conditions with nominal SNRs of -12, -6, and $+6$ dB. Posterior medians (○) and 97.5-% (▽) and 2.5-% (△) quantiles may be compared to the true individual values shown by short horizontal lines. Conventional estimates obtained from each confusion-count matrix by Eq. (10) are marked by ×. Data are simulated for groups of $L = 21$ listeners randomly drawn for each test from the population shown in Fig. 7, but results are only plotted for the listeners with the lowest, median, and highest true MI in each group.

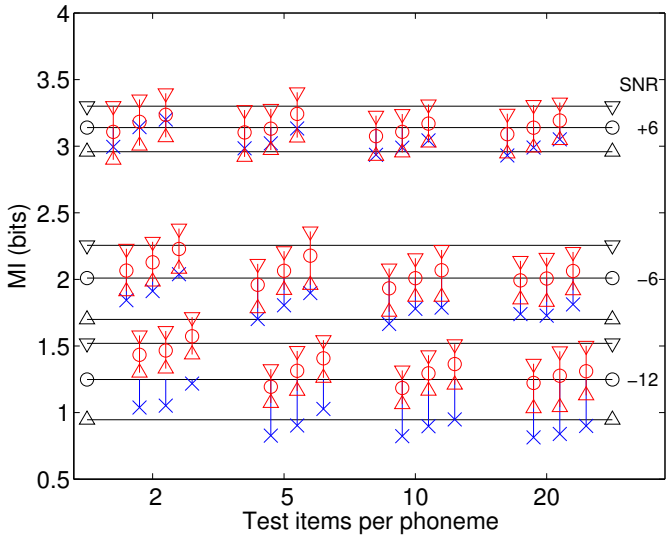

Fig. 7. Group estimates of mutual information (MI) for simulated confusion-count data with $N_s \in \{2, 5, 10, 20\}$ test items per phoneme, in three different test conditions with nominal SNRs of -12, -6, and $+6$ dB. Posterior medians (○) and ranges (vertical lines) between 97.5-% (▽) and 2.5-% (△) quantiles may be compared to the corresponding true medians and quantiles (horizontal lines) in the population. Conventional estimates obtained from the pooled confusion-count results within each group are marked by ×. Data were simulated for $R = 11$ complete group replications, each with $L = 20$ listeners, of which only the group results with the lowest, median, and highest median estimate are displayed.

proposed method for detecting credible population differences between two test conditions against a conventional significance test using the Wilcoxon signed-rank test[8] for simulated groups of 20 listeners drawn repeatedly from the same simulated population. The plotted *Receiver Operating Characteristics* (ROCs) show the proportion of replications where the method correctly indicated a real difference, versus the proportion where the method incorrectly indicated a difference, when the two conditions were equal. The ROC curve shows this pair of estimated probabilities together, plotted as a function of the decision threshold. For MI, the ROC for the proposed Bayesian method is much closer to the upper left corner of the plot, and the curves never cross, which indicates better detection power than the Wilcoxon test for all significance levels. For PC, the Bayesian approach has only slightly better detection power. Of course, with a larger true difference between the test conditions, both methods would perform better.

2) Magnitude of the Difference: When comparing performance results from two test conditions, it is also interesting to estimate how large the improvement might be from condition A to condition B. Fig. 5 shows the estimated MI improvement when there is no real difference and when there is a small true difference corresponding to a nominal SNR improvement of 1, 2, or 3 dB.

The proposed Bayesian method predicts the true population quantiles of the MI difference reasonably well, but the conventional results using the individual plug-in estimates in (10) tend to underestimate the true MI improvement. Conventional 95-% nonparametric confidence intervals [35] for the median MI improvement covered the true value in 83 of $R = 100$ replications with $(A, B) = (-6, -5)$ dB, in 53% of replications with $(-6, -4)$ dB, and in only 26% for $(-6, -3)$ dB. In contrast, the Bayesian 90-% credible intervals included the true median in more than 98% of replications in all pairs of test conditions. This failure of the conventional MI estimation is caused by the bias problem analyzed theoretically in Appendix A and exemplified by simulation results in the following section.

[8]The Wilcoxon results were calculated using the one-sided option of the `signrank` function in MATLAB's Statistical Toolbox.

C. Estimates of Mutual Information

1) Individual Results: Examples of MI estimates are displayed in Fig. 6 for a group with $L = 21$ simulated listeners, tested at nominal SNRs of -12, -6, and $+6$ dB. The Bayesian estimates were typically close to the true value for $N_s \geq 5$ test items per phoneme. The posterior credible intervals included the true individual MI for {37, 75, 83, 90}% of the listeners with $N_s \in \{2, 5, 10, 20\}$. The true value was typically overestimated by about 0.2 bits in the most difficult condition with SNR at -12 dB, when only 2 test items were used for each phoneme.

In contrast, the conventional MI estimates by Eq. (10), marked by × in Fig. 6, severely overestimated the true individual MI values. For example, at a nominal SNR of -12 dB, the true MI was about 1.3 bits per phoneme, but the conventional MI estimates were typically about 2.2 bits per phoneme with 5 test items per phoneme, i.e., close to the true value at 6 dB higher SNR. In difficult test conditions the conventional individual estimates must be considered unreliable even with 20 test items per phoneme.

The amount of overestimation is not constant but varies across test conditions. This explains why systematic errors remain in the conventional estimates of the MI *difference* between test conditions, shown in Fig. 5.

2) Group Results: Fig. 7 shows estimated medians and quantiles for MI in simulated groups of $L = 20$ participants tested at nominal SNRs of -12, -6, and $+6$ dB. The Bayesian estimates of medians and quantiles are generally close to the true values, except that the MI values are slightly overestimated when only 2 test items were presented for each phoneme, especially in the condition with -12 dB. With $N_s \geq 5$ the Bayesian credible intervals covered the true population median at all SNRs in all replications, except for $N_s = 5$ at -12 dB SNR, where the coverage was 91%.

The conventional pooled group estimates systematically *underestimated* the true MI. This bias was most notable (about 0.4 bits) in the most difficult condition with nominal SNR at -12 dB. The underestimation is about the same for any N_s. As explained in Appendix A, this underestimation is caused by the true variability among individuals in the population, and the error therefore persists regardless of the number of test items and the number of listeners in the group.

D. Cell Differences Between Two Conditions

To evaluate the proposed method for identifying a set of stimulus-response pairs showing jointly credibly different response probabilities in two matrices, a pair of population models was constructed for nominal SNRs of -3 vs. -6 dB. Confusion-count matrices were generated with $N_s \in \{5, 10\}$ simulated presentations per phoneme. Thus, each simulation represents a pair of confusion-count matrices that might be measured for $L = 1$ listener in two slightly different test conditions.

A set of credibly different cells was estimated from each simulated matrix pair, as described in Appendix C. For cells to be included in the set, it was required that the response-probability difference in (C.1) was greater than $\epsilon = 0.001$ with a joint probability $q_n \geq 0.95$ in (C.5). For each simulation, it was noted how many confusion-matrix cells were identified as credibly different in the correct direction, as well as how many were identified as credibly different, but in the wrong direction compared to the known true difference. The results are plotted in Fig. 8. For comparison, similar results are also shown for a corresponding sampling-based frequentist method to detect significantly different sets of cells, using a significance level $\alpha = 0.05$. This method is described in Appendix D.

Both methods show very few cases of incorrect detection, but the proposed Bayesian method appears to correctly identify more cells as credibly different.

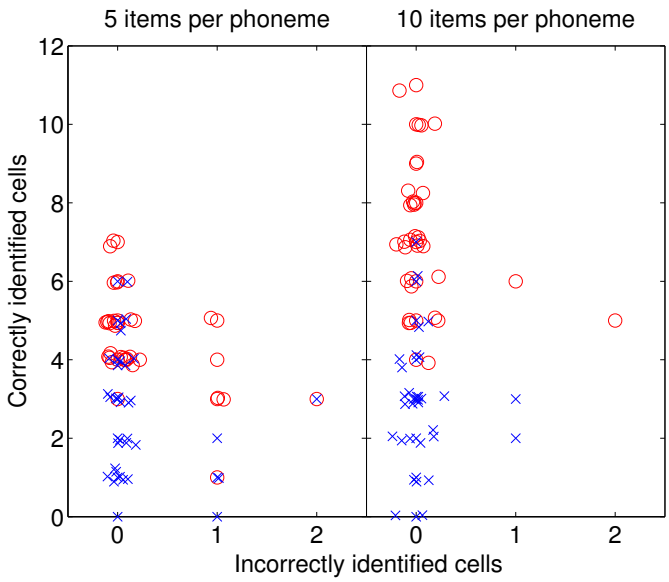

Fig. 8. Number of correctly identified cells with jointly credible differences between two simulated test conditions with nominal SNRs $(-6, -3)$ dB, plotted versus the number of cells identified as credibly different in the wrong direction, in confusion matrices with $16 \times 16 = 256$ cells. Results are shown for $N_s = 5$ (left panel) and $N_s = 10$ (right panel) test items per phoneme. Results for $R = 40$ simulations, each with $L = 1$ listener, are shown by $\circ$ for the Bayesian method in Appendix C, while $\times$ show results from the alternative frequentist method in Appendix D. (Overlapping points have been jittered slightly for clarity.)

V. Discussion

The simulation results indicate that reasonably good estimates of listeners' phoneme recognition performance and the difference between test conditions can be obtained from confusion-count matrices using the proposed parametric Bayesian estimation method, even when only five test items are presented for each stimulus category and listener. Some useful results may be obtained with even fewer test items.

The simulation results in Figs. 6 and 7 showed that the bias problems in frequentist MI estimation can be substantial. As shown in Figs. 4 and 5, the bias tends to weaken any conventional statistical analysis of the MI difference between test conditions. The proposed Bayesian method, in contrast, produced results that were quite close to the true values, even with a very limited number of test items per phoneme.

A key element of the Bayesian approach (Appendix B) is the use of a prior distribution for the model parameters in the analysis. In the absence of prior knowledge about the population from which the listeners are recruited, the proposed hierarchical method first adapts the prior to the test results for all listeners, before individual models are adapted to each participant.

The prior distribution has a regularizing influence on the estimation results. This feature of the proposed approach can thus be seen as a kind of well-tempered compromise between the classical methods of using only the individual results or using only the pooled results across all listeners. This compromise utilizes the fact that the experimental group usually is pre-selected to have some general characteristics in common in scenarios where the main goal is to identify overall group effects for the selected type of participants.

However, this adaptation of the prior also implies that the estimates for individual listeners are partly influenced by the average results among other participants. This method may therefore not be appropriate if the main purpose of the experiment is to reveal systematic individual differences

between listeners.

Although the Dirichlet distribution (Appendix B) is a well-established model for the probability parameters of any multinomial-distributed data, we have not seen this model applied to the analysis of phoneme confusion matrices in previous work. Trevino and Allen [36] used the Hellinger distance to quantify the dissimilarity between matrices to find clusters of similar confusion matrices. Jürgens et al. [24] used Pearson's ϕ^2, which is a similar measure of the dissimilarity. Another related approach was proposed in [37] to evaluate the ability of various models to explain observed confusion data, using a measure closely related to the Kullback-Leibler divergence to quantify the dissimilarity between observed confusion counts and model-predicted probability vectors. All these methods account for the dependencies within each matrix row in a similar manner to that of the Dirichlet distribution, but none are based on Bayesian probability.

Numerous studies have followed the approach in [5] and analyzed the pattern of consonant confusions by characterizing each consonant phoneme by distinctive features which reflect different dimensions of similarity between phonemes; see, e.g., a recent review in [23]. It may be possible to extend the Bayesian approach to derive an empirical set of feature dimensions that are most likely to explain the observed confusion patterns, but such extensions are left for future work.

VI. Conclusion

This paper proposed and evaluated a parametric Bayesian approach for the analysis of phoneme confusion matrices. Two different bias problems in conventional estimation of mutual information were analyzed and explained theoretically. Estimating the overall probability of correct response has no such bias problems.

Simulations indicate that the proposed Bayesian method can give satisfactory estimates of mutual information and response probabilities, even for tests using as few as five presentations for each phoneme.

The proposed method was capable of revealing small overall differences in performance between two test conditions with greater power than conventional significance tests or conventional confidence intervals. The method could also identify sets of confusion-matrix cells that are jointly credibly different between two test conditions, with better power than a similar approximate frequentist method.

Appendix A
Proofs of Mutual-Information Estimator Bias

A. Individual Data

The overestimation of the conventional MI calculation in (10) happens because any random variations in the observed confusion-count matrix is interpreted as a variation in the underlying response probability matrix, and any such variability tends to increase the calculated MI estimate. Formally, this tendency can be expressed as

$$E\left[MI(\boldsymbol{U})\right] \geq MI(E\left[\boldsymbol{U}\right]). \tag{A.1}$$

This relation follows from Jensen's inequality, because the logarithmic transformation $MI(\boldsymbol{u})$ in (8) is a *convex* function of the response probabilities in $\boldsymbol{u}$ [6, Theorem 2.7.4]. Now, assume the listener is characterized by a fixed matrix $\boldsymbol{u}$ of response probabilities and the relative response rates $\widehat{u}_{sr}(\boldsymbol{X}) = X_{sr}/N_s$ are plugged in to calculate $\widehat{MI}(\boldsymbol{X}) = MI(\widehat{\boldsymbol{u}}(\boldsymbol{X}))$ as in (10). The response rates are unbiased estimates of the true response probabilities, i.e., $E\left[\widehat{\boldsymbol{u}}(\boldsymbol{X})\right] = \boldsymbol{u}$. Nevertheless, (A.1) implies that the expected value of the MI estimate, across all possible test results $\boldsymbol{X}$, is biased as

$$E\left[\widehat{MI}(\boldsymbol{X})\right] = E\left[MI(\widehat{\boldsymbol{u}}(\boldsymbol{X}))\right] \geq \\ \geq MI(E\left[\widehat{\boldsymbol{u}}(\boldsymbol{X})\right]) = MI(\boldsymbol{u}). \tag{A.2}$$

The overestimation bias decreases as the number of test items increases, because the variance of $\widehat{u}_{sr}(\boldsymbol{X})$ is inversely proportional to N_s.

B. Pooled Group Data

Assume that phoneme identification has been tested for L different listeners. The random confusion-count matrix is $\boldsymbol{X}_l$ for the lth participant, who is characterized by some fixed response-probability matrix $\boldsymbol{u}_l$. The participants are recruited at random from the population of interest, so the matrices $\boldsymbol{u}_l$ are samples drawn from the distribution of $\boldsymbol{U}$ in this population. Therefore, all individual matrices $\boldsymbol{u}_l$ deviate somewhat around the population mean $E\left[\boldsymbol{U}\right]$. The individual confusion-count matrices are pooled as $\overline{\boldsymbol{X}} = \sum_{l=1}^{L} \boldsymbol{X}_l$. This sum is then used to calculate a point estimate for the group. Since the pooled $\widehat{\boldsymbol{u}}(\overline{\boldsymbol{X}})$ is a consistent estimate of $E\left[\boldsymbol{U}\right]$, the pooled point estimate converges as

$$E\left[\widehat{MI}(\overline{\boldsymbol{X}})\right] = E\left[MI(\widehat{\boldsymbol{u}}(\overline{\boldsymbol{X}}))\right] \overset{L\to\infty}{\longrightarrow} MI(E\left[\boldsymbol{U}\right]) \leq \\ \leq E\left[MI(\boldsymbol{U})\right]. \tag{A.3}$$

Thus, the pooled estimate remains an *underestimate* of the population mean, also in the limit of infinitely many participants, because of the inequality in (A.1).

The overestimation and underestimation tendencies might neutralize each other, if the confusion counts are first pooled within smaller subgroups of listeners, and the MI estimates from the subgroups are then averaged. However, since the underestimation bias depends on the unknown true variability in the population, it would be difficult in practice to select an appropriate size of such subgroups.

Appendix B
Dirichlet Response-Probability Model

This section defines the precise mathematical model for the distribution of each response-probability matrix $\boldsymbol{U}_{lt}$ which determines the distribution (3) of observed confusion counts $\boldsymbol{X}_{lt}$ for the lth listener in the tth test condition. The model involves the *Dirichlet* distribution which is well established in the machine-learning literature, e.g., [32], as a convenient model for the distribution of the probability parameters of any multinomial distribution. The Dirichlet is a conjugate prior for the multinomial.

1) Prior Density: We define the prior distribution for the parameter matrix $\boldsymbol{U}_{lt}$ as a product of independent Dirichlet distributions for all row vectors $\boldsymbol{U}_{s,lt}$, as

$$p_{\boldsymbol{U}_{lt}}(\boldsymbol{u}) = \prod_{s\in\mathcal{S}} p_{\boldsymbol{U}_{s,lt}}(\boldsymbol{u}_s), \tag{B.1}$$

with Dirichlet probability densities for each row vector,

$$p_{\boldsymbol{U}_{s,lt}}(\boldsymbol{u}_s) = C(\boldsymbol{a}_{s,t}) \prod_{r\in\mathcal{R}} u_{sr}^{a_{sr,t}-1}. \tag{B.2}$$

The normalization factors are

$$C(\boldsymbol{a}_{s,t}) = \frac{\Gamma\left(\sum_{r\in\mathcal{R}} a_{sr,t}\right)}{\prod_{r\in\mathcal{R}} \Gamma\left(a_{sr,t}\right)}, \tag{B.3}$$

where $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$ is the gamma function.

The mean of each row vector $\boldsymbol{U}_{s,lt}$ is determined by the parameter vector $\boldsymbol{a}_{s,t}$ with nonnegative elements $a_{sr,t}$ as

$$E\left[\boldsymbol{U}_{s,lt}\right] = \frac{\boldsymbol{a}_{s,t}}{\sum_{r\in\mathcal{R}} a_{sr,t}}. \tag{B.4}$$

Increasing the parameter values reduces the variance.

The prior distributions are assumed identical and independent for all listeners in each test condition t, since these distributions represent the general prior knowledge about the performance of any listener drawn at random from the population. Therefore, the prior parameters $a_{sr,t}$ are the same for every listener l.

2) Posterior Density: The posterior distribution of the matrix $\boldsymbol{U}_{lt}$ for the lth listener in the tth condition, given the confusion count matrix $\boldsymbol{x}_{lt}$, is obtained by combining the prior density (B.1) with (3) in (5), yielding

$$p_{\boldsymbol{U}_{lt}|\boldsymbol{X}_{lt}}(\boldsymbol{u} \mid \boldsymbol{x}_{lt}) \propto \prod_{s\in\mathcal{S}} \prod_{r\in\mathcal{R}} u_{sr}^{x_{sr,lt}} u_{sr}^{a_{sr,t}-1}. \tag{B.5}$$

This posterior distribution is obviously again a product of Dirichlet density functions. Including the normalization factors, it can be written as

$$p_{\boldsymbol{U}_{lt}|\boldsymbol{X}_{lt}}(\boldsymbol{u} \mid \boldsymbol{x}_{lt}) = \prod_{s\in\mathcal{S}} C(\boldsymbol{\alpha}_{s,lt}) \prod_{r\in\mathcal{R}} u_{sr}^{\alpha_{sr,lt}-1}, \tag{B.6}$$

where the new posterior parameter vectors $\boldsymbol{\alpha}_{s,lt}$ have elements

$$\alpha_{sr,lt} = a_{sr,t} + x_{sr,lt}. \tag{B.7}$$

Thus, the prior parameters $a_{sr,t}$ play a similar role to the pseudocounts sometimes used in conventional estimates of probabilities.[9]

Finally, the *marginal posterior distribution* of response probabilities $\underline{\boldsymbol{U}} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T)$ for the complete *group of listeners* in the T test conditions is obtained by averaging across individual listeners. The result is a mixture density

$$p_{\underline{\boldsymbol{U}}|\underline{\boldsymbol{X}}}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \mid \underline{\boldsymbol{x}}) = \frac{1}{L} \sum_{l=1}^{L} \prod_{t=1}^{T} p_{\boldsymbol{U}_{lt}|\boldsymbol{X}_{lt}}(\boldsymbol{u}_t \mid \boldsymbol{x}_{lt}). \tag{B.8}$$

The only remaining problem is how to choose the prior parameters $a_{sr,t}$.

[9]For example, in [12, Eq. (6)] a pseudocount of 0.5 was added to each observed response count.

3) Informative Prior: If data from earlier experiments are known for the same or an arguably similar group of listeners, those results may be used to set the prior parameters.

4) Hierarchical Prior: In the absence of prior information, the present proposed approach is to adapt the set of parameters $a_{sr,t}$ to maximize the total *marginal likelihood* (B.10) of the observed data for all listeners.[10] This hierarchical approach was used in all the simulations in the present study.

As a consequence, the posterior distribution for any individual listener becomes somewhat influenced by data for all the other listeners in the group. This approach tends to interpret the most extreme individual confusion-count results as an effect of the variability in the test method for $\boldsymbol{X}_{lt}$ rather than as a true extreme individual deviation of $\boldsymbol{U}_{lt}$ from the common group characteristics.

To ensure that all $a_{sr,t} > 0$, with minimal influence of prior assumptions, all these parameter values are assumed to be drawn from a *nearly uniform hyper-prior distribution* composed of independent and identical gamma density functions

$$\gamma(a_{sr,t}) \propto a_{sr,t}^{c-1} e^{-d a_{sr,t}} \tag{B.9}$$

with fixed shape parameter c and inverse scale d. These hyperparameters are chosen as $c \approx 1.05$ and $d \approx 0.105$, such that the distribution mode is $(c-1)/d = 0.5$ and the mean is $c/d = 10$. Thus, before any data have been observed, the most likely parameter value is $a_{sr,t} = 0.5$,[11] and the values can vary from near zero to much greater than ten.

Given any set of prior parameters $\underline{\boldsymbol{a}}_t = (\boldsymbol{a}_{1,t}, \ldots, \boldsymbol{a}_{|\mathcal{S}|,t})$ for the tth test condition, the *marginal likelihood* is the probability of all observed data $\underline{\boldsymbol{x}}_t = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Lt})$ in this test condition, when all other model variables have been integrated out:

$$\begin{aligned} p_{\underline{\boldsymbol{X}}_t}(\underline{\boldsymbol{x}}_t \mid \underline{\boldsymbol{a}}_t) &= \prod_{l=1}^{L} \int p_{\boldsymbol{X}_{lt}|\boldsymbol{U}_{lt}}(\boldsymbol{x}_{lt} \mid \boldsymbol{u}) p_{\boldsymbol{U}_{lt}}(\boldsymbol{u} \mid \underline{\boldsymbol{a}}_t)\, d\boldsymbol{u} = \\ &= \prod_{l=1}^{L} \prod_{s\in\mathcal{S}} c(\boldsymbol{x}_{s,lt}) \frac{C(\boldsymbol{a}_{s,t})}{C(\boldsymbol{a}_{s,t} + \boldsymbol{x}_{s,lt})}. \end{aligned} \tag{B.10}$$

Here, the Dirichlet normalization factors $C(\cdot)$ are defined in (B.3), and $c(\boldsymbol{x}_{s,lt})$ is the multinomial factor defined in (2).

The parameters in $\underline{\boldsymbol{a}}_t$ are adapted numerically to maximize the objective function

$$\mathcal{L}(\underline{\boldsymbol{a}}_t) = \log p_{\underline{\boldsymbol{X}}_t}(\underline{\boldsymbol{x}}_t \mid \underline{\boldsymbol{a}}_t) + \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}} \log \gamma(a_{sr,t}). \tag{B.11}$$

5) Fixed Non-informative Prior: One might consider assigning a fixed prior designed to exert minimal influence on the results. Setting $a_{sr,t} = 0$ is not possible, because then the posterior response probability distribution becomes improper (not normalizable) in case exactly 0 or N_s counts are observed in some cell, which frequently happens by chance

[10]This hierarchical approach can be seen as a "data-dependent" method to choose the prior distribution. However, as the hyper-prior distribution (B.9) is data-independent, the complete hierarchical model is still rigorously Bayesian. This method is well established in machine learning, e.g., [32]. By analogy with conventional maximum likelihood estimation, it is usually called *"Type-II Maximum Likelihood,"* when the hyper-prior distribution is exactly uniform.

[11]The value 0.5 is the non-informative *Jeffreys prior* [38] for the Dirichlet distribution, equivalent to the pseudocount 0.5 that is often used.

alone. Assigning $a_{sr,t} = 1$ may seem reasonable[12] as this makes the prior distribution of $U_{sr,lt}$ uniform. However, when N_s is small, any predetermined $a_{sr,t} > 0$ can have an undesirably strong effect on the result. Therefore, the more flexible hierarchical model is preferred for the present purpose.

APPENDIX C
RESPONSE DIFFERENCES IN MULTIPLE CELLS

The Bayesian procedure estimates a joint posterior probability distribution for the pair $(\boldsymbol{U}, \boldsymbol{V})$ of response probability matrices for two test conditions, given the observed confusion counts $(\boldsymbol{x}, \boldsymbol{y})$ in those test conditions. For a single stimulus-response pair (s, r), the probability that the response rate is higher in the second condition can be calculated as

$$q_{sr} = P\left[V_{sr} > U_{sr} + \epsilon \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right], \tag{C.1}$$

where ϵ may be set to a small positive number to avoid including tiny differences. If the resulting q-value is close to 0.5, the observed difference was likely just a result of random variability.

Since a comparison of two matrices involves multiple tests, it is necessary to calculate the probability for multiple differences jointly. Let C be the set of all elementary candidate hypotheses to be tested, including $V_{sr} > U_{sr} + \epsilon$ and $U_{sr} > V_{sr} + \epsilon$, for all s, r. We want to find an ordered sequence of hypothesis *combinations*, including one, two, etc., of the elementary events in C, that are jointly highly likely. First, find the single most likely event

$$h^* = \underset{h \in C}{\operatorname{argmax}}\, P\left[h \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right]. \tag{C.2}$$

Store this event as $H_1 = h^*$, remove it from the set of candidates, $C = C \backslash h^*$, and record the corresponding probability $q_1 = P\left[H_1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right]$. Then, to find a highly likely[13] set of joint hypotheses with $n = 2, 3, \ldots$ combined events, continue recursively as

$$h^* = \underset{h \in C}{\operatorname{argmax}}\, P\left[H_{n-1} \cap h \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right] \tag{C.3}$$

$$H_n = H_{n-1} \cap h^*; \quad C = C \backslash h^* \tag{C.4}$$

$$q_n = P\left[H_n \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right] \tag{C.5}$$

as long as the joint probability q_n is above some predetermined threshold. Since one additional hypothesis is included at each step, the joint probabilities will form a decreasing sequence $q_1 \geq q_2 \geq \ldots \geq q_n$.

As the probability q_n is calculated *jointly* for the set H_n of combined hypotheses, these results can be used without any further corrections for multiple hypothesis testing.

APPENDIX D
FREQUENTIST DETECTION OF CELL DIFFERENCES

In order to perform a conventional frequentist significance test of differences between cells in two observed confusion-count matrices, we must first formulate a *null hypothesis* modeling the case when there are no differences. Given this null hypothesis, the test should estimate the conditional probability for a random result to be as extreme as, or more extreme than, the observed result. The following approximate method was used for comparison with the Bayesian approach in Appendix C:

Given a pair of observed confusion-count matrices $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, the null hypothesis assumes that the response-probability matrix equals the averaged estimate $\overline{\boldsymbol{u}}$ with elements

$$\overline{u}_{sr} = \frac{x_{1,sr} + x_{2,sr} + 1/\left|\mathcal{R}\right|}{2N_s + 1}, \tag{D.1}$$

where, as before, $|\mathcal{R}|$ is the number of response alternatives, and N_s is the number of test stimuli for the ith phoneme. The pseudocount term $1/|\mathcal{R}|$ prevents estimated probabilities from being exactly 0 or 1.

Using $\overline{\boldsymbol{u}}$, we sample $N = 1000$ random confusion-count matrix pairs $(\boldsymbol{y}_1(n), \boldsymbol{y}_2(n))$, $n = 1, \ldots, N$, with N_s test stimuli per phoneme in each matrix. For each matrix pair, a standardized difference $z_{sr}(n)$ is calculated for each cell by scaling with the standard deviation of the difference, as

$$z_{sr}(n) = \frac{y_{1,sr}(n) - y_{2,sr}(n)}{\sqrt{2N_i \overline{u}_{sr}(1 - \overline{u}_{sr})}}, \tag{D.2}$$

such that the standardized values have zero mean and unit variance for all matrix cells, under the null hypothesis.

Using the complete set of random data, an empirical cumulative distribution for the difference is estimated as

$$F(d) = \frac{1}{N\left|\mathcal{S}\right|\left|\mathcal{R}\right|} \sum_{n=1}^{N} \sum_{s \in \mathcal{S}} \sum_{r \in \mathcal{R}} \mathbb{I}\left[z_{sr}(n) \leq d\right], \tag{D.3}$$

where $\mathbb{I}[\cdot]$ is the indicator function which equals one if the argument is true. Since the distribution function $F(d)$ is tabulated with high resolution, it approximates the cumulative probability for the event that the standardized difference is $z_{sr} \leq d$ for *any* cell in the matrix pair.

Finally, corresponding standardized differences $\widehat{z}_{sr}$ are calculated from the observed matrix pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, just as in (D.2). All cells where $F(\widehat{z}_{sr}) < \alpha/2$ or $F(\widehat{z}_{sr}) > 1 - \alpha/2$ are considered as jointly significantly different at the α level.

[12] A pseudocount of 1 was motivated in a different way in [14, Sec. 2.5].

[13] This algorithm is not guaranteed to find the *globally most likely* set of $n > 1$ hypotheses, but a highly likely set is sufficient for practical purposes.

ACKNOWLEDGMENT

This work was supported mainly by ORCA-Europe/Widex, a subsidiary of Widex A/S, Denmark, and to some extent by KTH Royal Institute of Technology, Sweden. The authors would also like to thank Karolina Smeds and Morten Løve-Jepsen for valuable discussions and contributions to the presentation. We also thank two reviewers for constructive critical comments on earlier versions of the paper.

REFERENCES

[1] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.
[2] H. Fletcher and R. Galt, "The perception of speech and its relation to telephony," *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 89–151, 1950.
[3] J. B. Allen, "Harvey Fletcher's role in the creation of communication acoustics," *J. Acoust. Soc. Am.*, vol. 99, no. 4 pt 1, pp. 1825–1839, 1996.
[4] H. Fletcher and J. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, vol. 8, pp. 806–854, 1929.

[5] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, no. 2, pp. 338–352, 1955.
[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, 2006.
[7] B. Hagerman, "Reliability in the determination of speech discrimination," *Scand. Audiol.*, vol. 5, no. 4, pp. 219–228, 1976.
[8] A. Thornton and M. Raffin, "Speech-discrimination scores modeled as a binomial variable," *J. Speech Hear. Res.*, vol. 21, no. 3, pp. 507–518, 1978.
[9] A. Boothroyd and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.*, vol. 84, pp. 101–114, 1988.
[10] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2312–26, Apr. 2007.
[11] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1220–1233, 2008.
[12] J. C. Toscano and J. B. Allen, "Across and within consonant errors for isolated syllables in noise," *J. Speech, Lang., Hear. Res.*, vol. 57, pp. 2293–2307, 2014.
[13] S. A. Phatak, Y.-S. Yoon, D. M. Gooler, and J. B. Allen, "Consonant recognition loss in hearing impaired listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2683–2694, 2009.
[14] W. Han, "Methods for robust characterization of consonant perception in hearing-impaired listeners," Ph.D. dissertation, Univ. of Illinois, Urbana-Champaign, IL, USA, 2011.
[15] R. Singh and J. B. Allen, "The influence of stop consonants' perceptual features on the articulation index model," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 3051–3068, 2012.
[16] A. Trevino and J. B. Allen, "Within-consonant perceptual differences in the hearing impaired," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 607–617, 2013.
[17] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.
[18] R. G. Newcombe, "Two-sided confidence intervals for the single proportion: comparison of seven methods," *Stat. Med.*, vol. 17, pp. 857–872, 1998.
[19] ——, "Interval estimation for the difference between independent proportions: comparison of eleven methods," *Stat. Med.*, vol. 17, pp. 873–890, 1998.
[20] W.-S. Lu, "Improved confidence intervals for a binomial parameter using the Bayesian method," *Commun. Stat. Theory*, vol. 29, no. 12, pp. 2835–2847, 2000.
[21] K. Krishnamoorthy and J. Peng, "Some properties of the exact and score methods for binomial proportion and sample size calculation," *Commun. Stat. Simulat.*, vol. 36, no. 6, pp. 1171–1186, 2007.
[22] B. T. Meyer, T. Jürgens, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3126–3141, 2010.
[23] Y.-S. Yoon, J. B. Allen, and D. M. Gooler, "Relationship between consonant recognition in noise and hearing threshold," *J. Speech, Lang., Hear. Res.*, vol. 55, no. 2, pp. 460–473, 2012.
[24] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2635–2648, 2009.
[25] J. D. Robinson, T. Baer, and B. C. J. Moore, "Using transposition to improve consonant discrimination and detection for listeners with severe high-frequency hearing loss," *Int. J. Audiol.*, vol. 46, pp. 293–308, 2007.
[26] F. Kuk, D. Keenan, P. Korhonen, and C.-C. Lau, "Efficacy of linear frequency transposition on consonant identification in quiet and in noise," *J. Am. Acad. Audiol.*, vol. 20, no. 8, pp. 465–479, 2009.
[27] A. J. M. Houtsma, "Estimation of mutual information from limited experimental data," *J. Acoust. Soc. Am.*, vol. 74, no. 5, pp. 1626–1629, 1983.
[28] J. Castelloe and G. Woodworth, "Jackknife and bootstrap bias correction for single-subject information transfer in audiological testing," Univ. of Iowa, IA, USA, Tech. Rep., 1996.
[29] E. Sagi and M. A. Svirsky, "Information transfer analysis: A first look at estimation bias," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2848–2857, 2008.
[30] C. Howson and P. Urbach, *Scientific reasoning: The Bayesian approach*, 3rd ed. Chicago et La Salle, Ill., USA: Open Court, 2006.
[31] S. E. Fienberg, "When did Bayesian inference become 'Bayesian'," *Bayesian Analysis*, vol. 1, no. 1, pp. 1–40, 2006.
[32] C. M. Bishop, *Pattern recognition and machine learning*. New York, NY, USA: Springer, 2006.
[33] J. B. Allen, "Consonant recognition and the articulation index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2212–2223, Apr. 2005.
[34] E. Parzen, "Quantile probability and statistical data modeling," *Stat. Sci.*, vol. 19, no. 4, pp. 652–662, 2004.
[35] A. D. Hutson, "Calculating nonparametric confidence intervals for quantiles using fractional order statistics," *J. Appl. Stat.*, vol. 26, no. 3, pp. 343–353, 2010.
[36] A. Trevino and J. Allen, "Systematic groupings in hearing impaired consonant perception," in *ISAAR*, 2013.
[37] T. S. Bell, D. D. Dirks, H. Levitt, and J. R. Dubno, "Log-linear modeling of consonant confusion data," *J. Acoust. Soc. Am.*, vol. 79, no. 2, pp. 518–525, 1986.
[38] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *P. Roy. Soc. Lond. A Mat.*, vol. 186, no. 1007, pp. 453–461, 1946.

Arne Leijon is professor em. in hearing technology at KTH Royal Institute of Technology, Stockholm, Sweden. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these devices. He received the M.Sc. degree (Civilingenjör) in engineering physics in 1971, and a Ph.D. degree in information theory in 1989, both from Chalmers University of Technology in Gothenburg, Sweden.

Gustav Eje Henter is a research fellow at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, United Kingdom. His research interests include parametric and nonparametric statistical modeling, particularly for statistical speech synthesis. He received the Ph.D. degree in electrical engineering (telecommunications) in 2013 and the M.Sc. degree (Civilingenjör) in engineering physics in 2007, both from KTH Royal Institute of Technology in Stockholm, Sweden.

Martin Dahlquist is a research engineer at ORCA-Europe/Widex, Stockholm, Sweden. His main research interests include evaluation methods for hearing instruments and methods for individual fitting of these devices. He received the M.Sc. degree (Civilingenjör) in electrical engineering in 1978 from KTH Royal Institute of Technology in Stockholm, Sweden. He has contributed to several audiological research projects at KTH and at the Karolinska University Hospital, Stockholm, Sweden.