SUPPORTING DOCUMENTATION
SPEECH AND HEARING SCIENCE 286 and LINGUISTICS 286
(Data Analysis, Specialized Course, Component of the GEC)
Analyzing the Sounds of Language
Margaritis Fourakis and Mary Beckman.

The course titled "**Analyzing the Sounds of Language**" is intended to introduce non-science majors to basic data analysis and quantitative reasoning skills by covering phenomena and concepts in speech acoustics, psychophysics, phonology, and other related areas of the speech sciences. That is, broadly speaking, the course will use spoken language — a phenomenon that is palpably relevant to every hearing human being — as a vehicle for implementing the goals of the Data Analysis GEC. For answers to the more specific questions that are listed in the GEC guidelines, page 9, refer to the attached syllabus.

1. How will the course address the problems of data-gathering, presentation, and interpretation of data?

Analyzing the Sounds of Language is proposed for inclusion under the specialized course subsection of the Data Analysis requirement as defined on pages 9-10 of the Model Curriculum. As is stated in the Course Objectives section of the attached syllabus, the course addresses this question in general and more specifically. The general strategy of the course is to cover one broad general question about speech each week, using that question as the prompt for posing more specific questions that can be answered using data of the sort that (for the most part) the students can gather themselves, in small experiments conducted in the class. For example, the first topic covered involves different measures that can be used to answer the question "We all have an intuitive sense that words can be longer or shorter; where does this subjective sense of variation in word length come from?" The students will participate in two different in-class "experiments" to produce both an integer measure (i.e., a set of subjective counts of the number of phones — consonant and vowel sounds used to compose each word — as judged by each student) and a continuous measure (i.e., length in ms of actual utterances of the target words produced by each student). The students will pool their data for each type of measure in order to be able to make histograms, and also to calculate mean lengths in ms for productions of words of different lengths. They will also make a scatterplot of the mean length in number of phones (averaged over all the students' judgments) for each utterance as a function of the mean length in ms (averaged over all the students' measurements). They will use regression analysis in order to see how much of the variance in the mean length in number of subjective phones can be accounted for by the variation in the physical measures. This will lead to an appreciation of how the numerical test (regression) can be used to interpret the data presented in the scatterplot. A comparable strategy will be used in each module, to introduce the students to the more general problem of how to gather, present, and interpret data that are relevant to some research question. Thus, there will be substantial data gathering, data presentation, and data interpretation all tied together in a meaningful way in every module of the course.

2. How will the students in the course be exposed to graphical and numerical arguments? How will attention be given to problems of measurement in the specific contexts studied?

In each week, the larger phenomenon of interest will be explored in terms of more specific questions that can be addressed by identifying relevant patterns in the graphical presentation of the data and by applying some numerical test, as illustrated above in the description of data

relevant for answering the question posed in the first module of the course. This strategy for addressing the research question is very conducive to imparting an immediate and deep understanding of the problems of measurement. For example, in the first module described above, the students will compare their measurements for the same utterance and see that each student has arrived at a slightly different answer. They will be asked to probe this variation to see how much of it is due to unreasonable precision (the speech analysis software yields times that are constrained only by the 11025Hz sampling rate of the recording, which is orders of magnitude finer than the just noticeable differences in duration that the typical human perceiver can judge) and how much of the variation is due to differences in segmentation criteria across the class.

3. How are statistical ideas to be applied to the course? Please provide typical discussions of the uses and misuses of statistics which will occur in the course.

As illustrated above in our description of the first module, statistical ideas will be applied each week in order to be able to interpret the data that are gathered as a response to the larger question that is addressed. This design is naturally conducive to an appreciation of the misuses as well as the uses of statistics. For example, in the first module, the students will see that the subjective sense of word length does correlate with the objective measure. However, they will see that the objective measure does not "cause" the subjective count. That is, they will see that different utterances of the same word have different objective lengths and that utterances of two different words that have the same subjective length can have vastly different objective lengths (e.g., utterances of "sand" will generally be much longer than utterances of "clip"). They should also see that there is independent variation in the subjective length for the same utterance. That is, different students can arrive at different counts for the same word (e.g., some students will analyze the beginning of the word "tray" as being composed of two consonants — the stop [t] followed by the liquid [r] — whereas others will parse it as three consonants — the two parts of the affricate [tΣ] followed by the liquid [r]). All of these patterns together should lead to an appreciation of the lesson that correlation cannot be used to argue directly for a causal relationship.

4. What topics in the study of probability will be presented in the course?

Probability theory will be introduced both as a foundation for understanding the statistics that are applied and also as a basis for the argumentation in several of the modules. For example, in the first week of the sequence of modules on consonants, the students will be introduced to the notion of "positional constraint" — the term that phoneticians use to describe systemic gaps in the distribution of phones in a language. For example, English has three voiced plosive sounds [b], [d], and [g], which contrast both at the beginnings and at the ends of words, as in "bat", "DAT", "gat" and "lab", "lad", "lag". However, while these three sounds are balanced in frequency at the beginnings of words, 90% of the words that end with any of these three sounds end with [d]. The students will be introduced to the notion of probability as a function of relative frequency and the notion of conditional probability (i.e., the probability that a word begins with [d] is the number of words that begin with [d] divided by the total number of words in reasonably large representative list of English words whereas the conditional probability that a word begins with [d] if it begins with an voiced plosive is larger because the denominator is different). They will then be led through the mathematics of when two probabilities are independent, to see that the contrast among these three sounds is more information-bearing in word-initial position than it is in word-final position. Similar arguments will be used in the last two modules of the course to explore the redundancy that is built into language design.

5. How will the course introduce students to the use of the computer.

The acoustic analysis of speech sounds is done wholly on the computer using task-specific software. The software used in the course is free and will be downloaded to the students' computers so they can use it at home. All the graphical and numerical presentations and all of the statistical analyses will be done on computers, using readily available software (e.g., MS Excel). Use of the computer is such an integral part of the course that it needs to be taught in a specialized classroom such as Derby 029.

6. If proposed as a B.A. course, will the course require only the "Basic Computational Skills" described in A., page 9, or will additional prerequisites apply?

The course is a B.A. course and no additional prerequisites will be required.