A Brief Primer on Experimental Designs for Speech Perception Research

Grant McGuire
Department of Linguistics
UC Santa Cruz
Draft: Summer 2010 (comments welcome!)

0 *Preface*

Overall, I hope this is the beginning of a dialog on best way to answer questions in speech perception. There certainly are mistakes in this essay, as well as a few controversies. Please contact me with any corrections or comments you have.

This essay is very much in debt to my lucky experiences as a researcher--especially to Dan Silverman as the first person to expose me to the role of experimental research in phonology. Thanks, of course, also go to my primary advisers, Mary Beckman and Keith Johnson at Ohio State. They have taught me much (well, the most) about science and the nature of experiments. I was very lucky to be in a position to get their advice and be in the presence of their expertise--much of the knowledge on experimentation and analysis I have gained was found in their amazing labs. Similarly, I am very much in debt my fellow graduate students at OSU who simultaneously grounded me and provided a sounding board for more extreme ideas. Of their number, I must especially mention Mike Armstrong, Robin Dautricort, Robin Dodsworth, David Durian, FangFang Li, Jeff Mielke, Misun Seo, Andrea Sims, Giorgios Tserdanelis, and Stephen Winters as especially helpful in technical and academic issues. I also want to thank Robert Fox in Speech and Hearing for access to his lab and advice.

After my time at OSU, I had the pleasure of joining the the Phonology Lab at Berkeley. There I gained a tremendous amount of knowledge and am grateful for both continuing my close contact with Keith Johnson, my now distant contact with Mary Beckman, and my new contact with John Ohala. I owe many thanks to Ron Sprouse for his technical expertise; much of the discussion of programs and stimuli are rooted in his advice to me. Thanks are certainly also due to the graduate students in the lab at Berkeley, especially Christian DiCanio, Shira Katseff, Pawel Nowak, and Ryan Shosted, who provided great insight into my work and many wonderful discussions on linguistics in general. Similarly I owe my collaborators in current research projects, Molly Babel and Sam Tilsen, a great deal of gratitude; these fruitful collaborations taught me much about the nature of academic research, in addition to adding to my technical knowledge.

I also owe a lot to two research labs at Purdue University where I had to the opportunity to both visit and collaborate with the labs of Amanda Seidl and Alexander Francis. Observing their operation and being a part of their work was a major influence. The people I met there provided a great influence on this work.

Finally, I happily thank the University of California at Santa Cruz, its Linguistics Department, and Dean of Humanities Georges Van Den Abbeele for providing me an opportunity to be a lab director and offering many wonderful collaborators. I especially thank Jaye Padgett, both for his early establishment of experimental perceptual research in the department, but also his kindness, criticism of papers/projects, and comments on this specific work. I further appreciate my technical and theoretical discussions with my colleagues Pranav Anand and Matt Wagers and especially being present for the the expansion and creation of their respective labs.

Many have contributed to this work. If I've forgotten you, please forgive (and remind!) me. Any mistakes are my own, and I hope to hear about them.

Grant McGuire
UC Santa Cruz
Summer, 2010

1 Introduction

This is a very brief guide to several common experimental designs used in speech perception research on adults and older children. These are not the only designs available, but are commonly encountered and have fairly well understood perceptual properties. I hope to include (soon!) a database of papers exemplifying these designs as well as some basic E-prime examples for people to play with.

I have divided this into five sections: this introduction, the designs themselves (divided roughly into discrimination and identification), more elaborate experimental concepts that build on the basic designs, some notes on stimulus creation and presentation, brief descriptions of common data measures, a glossary of some components common to many designs, and finally a few notes on navigating the local institutional review board. Where appropriate I've included a few notes on statistical analyses, though this is far from comprehensive.

As to the designs themselves, I have grouped them into two general categories, *discrimination* and *identification*. Though this differs from some classifications[1], I think this is the most intuitive distinction for general experiment design and understanding their use. The primary idea behind my classification is the ultimate goal of the researcher, the kind of information they desire, is either in a subject's ability to discern stimuli or categorize it. I am not beholden to this distinction, as it is fairly artificial, and am open to alternatives. It should further be noted that the names for these designs are not standardized and though I've tried to include this information, *caveat experimentor*.

2 Basic Experimental Designs

2.1 Discrimination

A discrimination experiment measures subject's ability to differentiate stimuli and frequently involve multiple stimulus presentations on a single trial. These designs are excellent for exploring the architecture of perceptual space and how it is affected by differences among subject populations or changes due to learning or exposure. As I have defined them, all discrimination experiments involve correct and incorrect responses from subjects, meaning that these designs all have similar analyses of percent correct data in common, including the use of *d'* (see Macmillan and Creelman, 2005; M&C hereafter).

[1] For example, M&C consider some of what I call “identification” designs “discrimination”. Their classification is based on how the sensitivity analysis should follow from the data.

2.1.1 Same – Different (AX)

The *same – different* design is quite common in speech research. In this design two stimuli are presented in each trial, separated by a specified amount of time (ISI: interstimulus interval). The stimuli are paired such that on any given trial they are either the same or different in some way and the subject's task is to identify which was presented. For example, if there are two stimuli, *A* and *B*, then there are two possible *different* pairs, <AB> <BA>, and two *same* pairs, <AA> <BB>. Generally, the number of same pairs presented to the subject matches the number of different pairs, though if discrimination is difficult then the number of same pairs may be reduced. This is due to the assumption that subjects expect an even number of same and different pairs and this assumption should be met or else the subjects will be alarmed and change strategies[2]. The most common measures from this design include accuracy (percent correct, or a sensitivity measure such as *d'*) and reaction time. Note that the order of the pairs, often ignored for simplification or due to lack of statistical power, can be a consistent effect (see Best et al. 2001, Francis and Ciocca, 2003).

Advantages: This design is very simple to explain to subjects. However, it has a further "ease of explanation" advantage: the differences and similarities of the stimuli do not need to be described to the subject. That is, subjects only need decide that two sounds are different in some way and do not need to consciously identify or be told the difference or know a particular label (e.g. compare to the labeling examples below.) Moreover, because subjects obligatorily make decisions based only on the second stimulus, reaction times derived from this paradigm are easy to measure and generally reliable.

Disadvantages: There are three main problems with the *Same – Different* design. The first is that it can encourage bias towards responding *same* when the task is difficult, i.e. the number of erroneous *same* responses becomes greater than the number of erroneous "different" responses. This can cause deeper problems when *different* pairs do not have similar perceptual distances; where the bias to say *same* varies by pair rather than being comparable across all pairs. The second problem is that calculating *d'* from the accuracy scores becomes quite complicated, especially when a *roving* design is used[3]. A roving design means that the first stimulus rotates between all possible stimuli rather than staying fixed. One way around this is to block the stimuli such that each stimulus is the first one (the "*A*" of *AX*) for an entire block. The third complication is that for many analyses the "same" stimuli are thrown out as uninterpretable or uninteresting, removing as many as half the trials.

2.1.2 Speeded Same – Different (Speeded AX)

This is a variant of the *same – different* design that deserves its own section. The basic design is the same, however subjects are directed to base their decision on a highly detailed short-term memory trace of each stimulus. The theory underlying this is that there are two modes of perception, an *auditory mode* consisting of a highly detailed but quickly decaying trace memory, and a *phonetic mode*, consisting of a more abstracted or categorical representation of the sound in question (Pisoni 1973, Durlach and Braida 1969). Under such an analysis, the auditory mode has been seen as being analogous to a non-speech mode of perception and used to compare raw perceptual distance independent of language specific effects (see e.g. Johnson and

[2] This could be addressed in instructions by explicitly telling the subjects that there's an even number of same and different pairs. It could also be argued that accuracy feedback will have the same effect.

[3] According to M & C this also suggests that the "differencing" model of subject decision making should be assumed (p. 221). This is open for debate and seems to depend on the size of the stimulus set and how categorical subjects' knowledge may be. In any event recent advances in computing (see the "psyphy" package for R) may make both analyses equally easy.

Babel 2007). In order to get this effect it is necessary for the task to have a low memory load[4], usually by making subjects respond very quickly, at least less than 800ms and preferably below 500ms (Pisoni and Tash 1974, Fox 1984). This is accomplished in two ways. First, the ISI must be sufficiently short, less than 500ms, with 100ms being a common duration. Note that an ISI of 0 is actually too short and results in slower RTs (Pisoni 1973). Second, the subjects need to be encouraged to respond quickly, typically by giving them an RT goal (such as < 500ms), frequent feedback as to their RTs (usually every trial), and a cut-off time when the RT is too long (e.g. 1500ms).

Advantages: This paradigm provides a way of assessing psychoacoustic distances using speech stimuli, if the assumption of bypassing the speech mode of perception is true. Additionally, RTs, because they tend to be more constrained on their upper bound by the task, usually show less variability than in other designs. Moreover, it is possible to get two measures of performance, RT and accuracy. The speed of the task also allows many trials in a reasonable amount of time.

Disadvantages: Data may be lost due to the difficulty of the task as subjects are encouraged to choose speed over accuracy and the incorrect responses must be thrown out for reaction time analyses. Moreover, different subjects or different groups may choose to respond more accurately or more quickly, independently of each other, complicating or invalidating statistical analyses. For this reason, accuracy feedback is often given in addition to RT feedback and subjects are given a target for both.

2.1.3 ABX (AXB, XAB, Matching-to-Sample)

In an ABX discrimination design three stimuli are presented in a series and the listener compares which stimulus, the *A* or the *B,* is the same or most similar to the *X* stimulus. This is also called *matching-to-sample* as the subject's task is to "match" the *X* stimulus to the sample, *A* or *B*. Other variations have the sample stimuli flanking the target (AXB; Harnsberger 1998 is an excellent example), or less commonly, following (XAB). There are consequently two ISIs, usually the same, though this may be varied such that the sample (*A-B*) interval in ABX or XAB is shorter than the interval to the stimulus to be matched.

Advantages: The ABX task has many of the same advantages as AX discrimination. The primary advantage is that subjects do not need to explicitly know or name the nature of the similarities/differences of the stimuli. Also, explanation of the task is also quite simple, though brief practice may be necessary as it is slightly more complicated than AX. However, the unique advantage of this task is that listeners are comparing a stimulus to two possibilities and know on each trial that either *A* or *B* is the correct answer, removing some of the bias problems inherent in AX discrimination.

Disadvantages: Though it has some advantages over AX discrimination, ABX brings some unique problems of its own. First, as there are three stimuli presented in a temporal order, recency effects due to memory become a consideration. This usually means a bias towards the *B* token as it is more current in memory. This effect can be accommodated by strict balancing of *AB* ordering and treating it as noise or a factor in analyses. However, this also means that this design doesn't have a speeded

[4] Sometimes this speeded design is explicitly contrasted against a "slowed" AX paradigm that may or may not include a white noise burst in the ISI. This white noise burst is designed to destroy the memory trace, see Guenther et al. 1999 for an example.

analogue (except for possibly very short stimuli). This is due to both sample stimuli necessarily being stored in memory and the assumed decay in detail of these memories, making the *A* stimulus more abstracted than the *B* one. A final concern is that, like AX discrimination, calculating *d′* is computationally complex and requires similar considerations when arriving at a decision model.

2.1.4 Two Alternative Forced Choice (2AFC)

In this design two stimuli are presented on each trial and subjects are asked to discern their order. Instructions usually take the form of, "which stimulus came first, A or B?". This design is considered highly valuable as it minimizes bias and can be used for very similar stimuli.

Advantages: Subjects hear two stimuli on each trial and so know that each order is possible. This is assumed to minimize bias as subjects should assume that either order is equally possible (compare to AX discrimination). This results in very simple *d′* calculations (see M&C Chpt. 7), little bias, and is generally considered to be easy for subjects.

Disadvantages: This task only works for binary choices. It also requires some sort of explicit label in order to determine "order". Subjects have to know what makes the stimuli unique in order to do this. This can make instructions difficult, if not impossible to relate to subjects, or requires some sort of subject training.

2.1.5 4-Interval Forced Choice (4IAX)

This design is considered analytically identical to 2AFC and consists of a presentation of four sounds and only a binary choice for the subject. This task has two subtle alternatives in design and subject explanation. In one, subjects are instructed to determine whether the second or third sound is different from the other three. Possible stimulus presentations are limited to <ABAA>, <AABA>, <BABB>, and <BBAB>. This is a very specific version of the "oddity" design (see below) and usually has identical ISIs across all intervals. The first and last stimuli are called "flankers" or "flanking stimuli". The other version of this design is explicitly related to *same – different* discrimination such that two pairs are presented, one a *same* pair and the other *different*. This is usually paired with a somewhat longer medial ISI compared to the first and third ones to more explicitly separate the stimuli into pairs for the subjects. It can be designed identically to the previous description (with slightly different instructions) or further complicated such that <AB AA>, <BA AA> pairs are allowed (though this may only complicate the analysis and is probably not recommended). For both types the subject is really being tasked with deciding the order of given stimuli, just as with 2AFC, so the analysis for *d′* is considered identical.

Advantages: This design is a way to remove the bias issues with AX discrimination. Because both *same* and *different* options are available on each trial, subjects shouldn't be biased towards *same* (both are present on each trial and therefore equally likely). This also means that very difficult discrimination tasks can be performed without assumed bias and pairs differing in similarity should not differ in bias. Moreover, unlike AX discrimination, the very simple *d′* analysis for 2AFC is appropriate for this design.

Disadvantages: The primary drawback to this design is that RTs are difficult to measure. Although theoretically subjects could make a decision upon hearing the second stimulus, it is possible and even likely that subjects will wait until the third or even fourth stimulus for reassurance. More worryingly, subjects could differ dramatically on what decision point they chose, or even differ by pair contrasted, or vary unpredictably trial by trial. Practically, their decision point is unpredictable and RTs are likely to be highly variable and unusable, except for the grossest distinctions.

2.1.6 Category Change (Oddball)

This design was created to be an adult analogue to a common infant perception task in which the infant hears a stream of syllables and responds by looking (in interest) when the stream changes in some detectable way. In the adult version, the stream of syllables are presented and the subject presses a button when they hear the change. The measure in this task is whether a change is detected or not, making it essentially a kind of Yes-No task (see below) where stimuli designed for infant research can be used. Versions of the design are now being used in brain imaging tasks where continuous stimulus presentation is important (eg. fMRI, NIRS, etc.)

2.2 Identification tasks

In identification tasks subjects are presented one or more sounds and asked to give an explicit label to one or more of them. This usually requires a category label, although some designs avoid this, essentially making them the same as the discrimination experiments previously described.

2.2.1 Yes – No

The simplest identification design is one in which a stimulus is compared against one other stimulus and only one is presented per trial. This can take the form of asking the subject whether or not a stimulus was present or whether the stimulus was *x* or *y*. The classic example of this design is the Bekesey hearing test, where a tone at a certain frequency and decibel level is presented and the subject responds whether they heard it or not. Though simple, this design is not as limited as it may seem and offers several options. For example, if there are two categories, <x y>, they can be compared as *x* ~ not *x* and *y* ~ not *y*. This potentially gives analytically a different result from an experiment where the task on each trial is to choose *x* or *y*. Additionally, multiple categories can be compared by blocking, such as *x~y, y~z, x~z*.

Advantages: This design is very simple to explain to subjects: did you hear *x* or not? The calculation of *d'* is very straightforward and no stimuli are thrown out in calculating typical statistics. RTs are easy to calculate and generally reliable.

Disadvantages: This design quickly becomes unwieldy the more comparisons that are made. Also, there are no direct comparisons of stimuli in each trial so difficult contrasts can be at a disadvantage.

2.2.2 Labeling (Identification, Forced Choice Identification)

In the very popular labeling task only a single stimulus is presented each trial and the subject must apply a label to that stimulus, either from a closed set (e.g. two buttons labeled *s* or *sh*) or some open set (e.g. "write what you hear"). This design is generally used to assess categorical knowledge; often ambiguous or continuous stimuli are categorized in this way. The simplest version of this task, having only two possible labels, can be seen as identical to a Yes – No task and may be analyzed accordingly. However, a more complex labeling task having more or unlimited choices may require a more complex analysis, though a simple analysis of counts is often adequate. Analyzing *d'* in such a task is more complicated when not a *yes-*

no design, depending on the specifics (see M&C, Chpt 10).

Advantages: The task is simple and straightforward with generally simple analyses. It can be explained quickly with little likelihood of error. When the response set is small, many trials can be performed in fairly short period with little stress on the subject.

Disadvantages: The primary drawback to this design is that labels are required, generally imposing a categorical decision for the subject and all the ramifications that come with that fact. Larger response sets make many analyses (such as *d'*) difficult or practically impossible.

2.2.3 Oddity

In this design multiple stimuli are presented, one is different from the rest, and the subject determines which is the unique, or odd one. Often limited to three (also known as the "triangular" design) or four stimulus presentations on each trial. Note that 4IAX is a special case of a four interval oddity design. Guenther et al. 1999 offers an interesting example where oddity was used to train subjects. The difficulty of the task was increased over the course of the training by increasing the number of intervals from two (2AFC) to three (triangular) and finally to four.

Advantages: Generally easy to explain to subjects, no explicit label is necessary beyond order, and many of the usual analyses are available. Also, the interval number can be varied without changing the instructions.

Disadvantages: Sensitivity can be difficult to compute at larger comparisons (see M&C Chpt. 9). Subjects must hold in memory the number of stimuli that have passed before the oddball, and possibly listen to them all for a difficult contrast, meaning that reaction times are unpredictable and recency effects are notable.

3 Elaborations

Many experiments use multiple designs to either acquire different kinds of data, have the difficulty of the task vary, or use multiple designs to train subjects at some perceptual task. Here are some notes on basic considerations in such experiments.

3.1 Multiple Designs in One Experiment

It is often desirable to acquire different kinds of data from the same subjects and different designs are better for different data. Care must be taken as earlier tasks may affect following tasks. Typically, less categorical tasks precede more categorical tasks, or they are balanced so that an equal number of subjects participate in each possible order. In both cases experiment order is often treated as noise and not analyzed (but this may not be the best tactic).

3.2 Adaptive Testing (Threshold Measurement, Staircase Designs)

An adaptive design is one in which the difficulty or ease of the task is changed based upon the previous trial or trials with the goal of finding the threshold of detection or performance at a specific level (e.g. 75% accuracy). These designs are appropriate to stimuli that are in some sort of continuum such that stimulus comparisons can be made easier or harder by changing the *step size*, which is the distance between the stimuli (increasing step size makes the task easier by increasing the perceptual distance). These can be categorical data that have been made into a continuum (/pa/ to /ba/) or something inherently continuous such as loudness. Note that the aforementioned hearing test is an adaptive design to find the threshold of hearing a tone at a given frequency and dB level. The task in such an experiment is usually some version of discrimination or Yes-No. Although an adaptive experiment may be done for its own sake (like the hearing test), it is more common as a first step in constructing stimuli that must fall into steps and those steps must be of an equal perceptual distance at a certain performance level.

The goals of adaptive test designs are efficiency and accuracy, i.e. the best adaptive test finds the given threshold accurately in the fewest number of trials. Efficiency is important as these tasks can be quite frustrating for subjects: they get more difficult as they go along and the subject may be performing near chance for a considerable number of trials. If a subject "gives up" or "fusses out" (to use the infant research terminology) they will perform poorly and the threshold will be inaccurate.

There are three basic components to an adaptive design that can all interact: performance level, the stepping rule, and the termination rule.

Performance Level: This is the actual threshold desired as the outcome of the experiment, e.g. a step size where discrimination is at 75%, or as in the hearing test, 50% accuracy (chance in a yes-no task). A performance level just above chance is appropriately called the just-noticeable-difference, or JND. Generally if the stimuli are going to be used in another experiment some higher level of performance is desired (or necessary).

Stepping Rule: This is the algorithm for determining when and how much to change the step size. While various algorithms have been proposed, I will describe the Kaernbach (1991) transformed up-down method, but see M&C Chpt. 11 for alternatives. The simplest version is the *staircase* procedure where a correct answer results in a one step increase in difficulty and an incorrect results in a one step decrease in difficulty. This results in only the JND, and no higher performance levels can be determined. To do that a *transformed* method is necessary where a single incorrect response results in a larger step change than an incorrect preceded by one or more correct answers. Kaernbach (1991) determined that a given accuracy level can be determined by having the decrease in step size be x times the increase in step size where $x = p\ /(1\text{-}p)$ and p is the desired threshold. For a 75% threshold the increase should be 3 times the decrease in step size, for 50% it is 1 (i.e. the simple staircase described above). So for performance at 75% the step size should be decreased by three steps after three correct answers (+++), increased by one after two correct answers and an incorrect (++-), increased by two after one correct and an incorrect (+-), and increased by three after a single incorrect (-).

Stopping Rule: This is the algorithm determining the end of the "run". Common ways to end the run are after set number of trials or a set number of *reversals*. A reversal is switch from increasing step size to decreasing step sizes or vice versa. This means that a steady increase or decrease in step size does not contribute to the stopping of the experiment, only a switch in movement. This method is considered a more efficient and accurate way of determining the desired threshold, though less predictable in the length of the experiment. The actual threshold is determined by either the step size at termination, or an average across all or some sample of the trials (again, see M&C Chpt 11 for more information).

3.3 Training Experiments

In a training experiment the explicit goal is to get subjects to perform better at some task, usually with the goal of assessing how perceptual space has changed as a result of this (see Logan & Pruitt 1998 for an overview), though sometimes training is necessary to perform adequately in some further task. They also have, in addition to the training task itself, some sort of testing paradigm. Because such experiments may be highly resource intensive (e.g. weeks of training for just one subject), extreme care should be taken in their design.

Training: Training tasks vary considerably and many of the designs explained above can be used for training, with differing goals. Broadly, discrimination training is used to heighten subjects' sensitivity to differences in stimuli while identification training is used to encourage categorization and minimization of differences within a category. Multiple training tasks can be included to increase subject performance or make their learning well-rounded. Training can be used as goal in and of itself (e.g. how does perception change due to learning) or as an elaborate form of practice so that subjects can perform in some other task that is the primary experiment. Some experiments include a control group that receives no training and is only tested. The length of training can be fixed for all subjects regardless of performance (e.g. 3 sessions, 150 trials, etc.) or can be fixed to some criterion (e.g. >80% correct, 50 trials correct in a row, etc.)

Testing: For most training experiments a pre-test is administered to assess initial performance followed by training, which is followed by a test identical to the first called a post-test. In some experiments only a post-test is administered and performance across different groups with different training conditions is assessed. This minimizes the number of tests that must be administered, but means that within-subject changes in perception can't be observed directly, but must be inferred across groups. An option available in longer training experiments is to administer multiple tests at intervals during training. Any number of possible designs can be used in testing, the best being determined by the goals of the experiment. Multiple types of testing are usually desirable, when possible. Occasionally, pre-tests are used to determine a subject's suitability for further training, where subjects who either perform too poorly for training to be effective (floor effect) or too well for a training effect to be apparent (ceiling effect) are noted and rejected.

Sessions: Many training experiments have tasks that are difficult enough to require multiple sessions of training. This can be over several consecutive days, several weeks, or even months. The number of training sessions may be the same for all subjects or may be concluded when a certain criterion is reached (in which the number of sessions may vary). In situations with multiple sessions, testing may be done in an entirely different session from training. Also, to assess the long-term effects of training, testing may be repeated at some later interval with no intervening training.

Analysis: The appropriate analysis generally follows from whatever design was chosen for testing and training. The primary exception is that changes at different sequential points in the training may be analyzed to assess training effectiveness. Such analyses are called “time series” and take into account the natural ordering of the data and the fact that each point in time is not independent, but dependent on the previous points, see Brockwell and Davis (2002) for an overview. Note also that whether tests are administered both pre- and post-training can affect the analysis (see above). Moreover, it is common to remove poorly performing subjects from analyses as training can be considered ineffective for them. Criteria for such exclusion should be determined well in advance and be made explicit. Occasionally such removal is justified based on pre-testing/screening.

4 Stimulus Considerations

Just as crucial to the successful answer to a research question as the design of an experiment is the construction of the stimuli used in it. The following are some notes on stimulus construction and use. This primarily concerns auditory stimuli, though I provide a brief discussion of audio-visual stimuli at the end of this section[5].

4.1 Construction Methods

The types of stimuli used in speech perception can be broadly grouped into three types: natural, synthetic, and hybrid.

4.1.1 Naturally-produced stimuli

Natural stimuli are produced by a talker and are not modified further other than by minor adjustments, such as trimming silence or overall amplitude normalization[6]. Because they are produced by an actual talker, they should most accurately represent sounds "in the wild", though at the expense of precise control of the stimulus parameters.

Recording: Usually such stimuli are recorded in a sound attenuated booth or at least a very quiet room using the best recording equipment available. In some cases the stimuli may have been recorded for some other purpose such as acquiring a corpus and the recording conditions are out of the researcher's control.

Talker(s): The producer of the stimuli is usually called a *talker*. Talkers are chosen for their ability to reliably reproduce the linguistic categories needed for the experiment in a controlled and consistent manner. They may speak the same language as the listeners in the experiment or may speak another language in the case of a cross-linguistic perceptual experiment. When more controlled stimuli are necessary, or when a speaker of a language natively having all the relevant contrasts cannot be found, a phonetically trained talker may be used. In such cases the productions can be checked for naturalness by a speakers of languages that have those contrasts. Often several talkers are used to increase the variety of stimuli or to explore the role of individual variability.

Instructions: When recorded for express purposes of an experiment, the talker's instructions are crucial. For example the subject may be asked to produce very careful speech or very natural, conversational speech (possibly difficult given the context unless some deception is involved). A subject may also be given much more explicit instructions, such as "use the same vowel as in *hood*", or a phonetically trained talker may be asked to produce very specific stimuli, such as syllables with only unreleased stops or a variety of "exotic" sounds. In all cases, however, care should be taken to ensure

[5] Tactile stimuli are possible, but beyond the scope of this essay. See Gick et al. for an example. I do not know of any clear examples of gustatory or olfactory stimuli used in speech experiments, but perhaps XXX comes close.

[6] Stimuli with more radical treatments such as cross-splicing portions of natural tokens are often considered "natural".

that productions are consistent, including intonational contours. List intonation is usually acceptable, but the falling tone at the end of a list should be avoided by either repeating the final stimulus in a list (and discarding it) or adding filler stimuli at the end that are not to be used. An alternative strategy is to put the target tokens in a syntactic frame such as, "Please say X again". A good, if more time consuming, strategy is to ensure the stimuli are representative of natural categories by recording many examples, making relevant acoustic measurements, and then selecting tokens that match most closely to previously reported values and rejecting anomalous ones.

4.1.2 Synthetic stimuli

Synthetic stimuli are constructed from scratch using either acoustic/perceptual or production models of speech. Various programs are available for these purposes with varying levels of control or user friendliness. Such stimuli are used when very tight control of specific aspects of the stimuli are necessary, such as when investigating the value of perceptual cues (e.g. what value of F2 onset triggers a given place percept while holding all other parameters constant.) A concern with such stimuli is their ecological validity, e.g. the ability to represent natural categories. Different programs and manipulations result in stimuli that are more or less natural, so care should be taken in choosing a construction method.

Acoustic / Perceptual Models: By far the most widely used program for stimulus construction is the Klatt cascade/parallel synthesizer (Klatt 1980, Klatt and Klatt 1990). Users specify various parameters for source and filter characteristics and build stimuli based on how these parameters change over time. Various more or less user-friendly interfaces have been developed over the years (see McMurry et al. 2008 for an excellent example). The freely available Praat acoustic analysis software (Boersma and Weenink 2010) offers a simple but useful source-filter synthesizer along with reasonable instructions. Usually a naturally produced token is used as a model to create a token from scratch; the parameters measured in the model token are used to make the synthetic token and only certain parameters are further manipulated.

Production Models: Some synthesis programs use models of articulator and vocal tract characteristics to produce an acoustic output. Such models are useful when production aspects are important to the research question at hand or when the stimuli must be producible by a human vocal tract. Examples include the Haskins CASY model and ArtiSynth designed at the University of British Columbia.

4.1.3 Hybrid stimuli

Hybrid stimuli attempt to combine these two methods to produce very naturalistic but controlled stimuli and are sometimes known as "modified naturally produced" stimuli. For such stimuli naturally produced tokens are resynthesized in various ways to change very specific aspects of the signal (changing intonational contours, vowel quality, gender, etc.). The resulting naturalness is dependent on the nature and number of modifications made. Generally, if one stimulus is resynthesized for an experiment, all should be resynthesized as the process does produce noticeable changes. The most common programs used for resynthesis are Praat and Matlab.

4.2 Presentation Methods

Various aspects of stimulus presentation should be considered separately from the experiment designs described above. A common problem is avoiding ceiling and floor effects where subjects perform too well (ceiling) or too poorly (floor) to properly analyze differences in performance.

Gating: Gating is the process of playing only a selected part of a given stimulus. Usually different conditions of such an experiment have stimuli presented in full compared with progressively gated, or shortened, stimuli to better understand the temporal qualities of the signal; this can also be used to vary task difficulty.

Degradation/Simplification: Stimuli can be degraded or simpified in different ways to reduce perceptibility. For example, Choo and Huckvale (1997) resynthesized fricatives with 4, 10, or 22 LPC coefficients, resulting in stimuli greatly varying in their complexity. Similarly, the sampling rate may be changed to reduce the amount of higher frequency information available to the subject (this could possibly be considered "frequency gating").

Embedding in Noise: Stimuli may be embedded in noise to make their perception more difficult. This is done at a specific signal-noise-ratio determined to make the task sufficiently hard (see especially Miller and Nicely (1955) for an example of consonant identification in varying levels of noise). Individual stimuli may be embedded in white or pink noise or played over a continuous stream of noise. Sometimes "cocktail party" noise is used--this consists of many unintelligible voices talking at a low level.

Volume Adjustments: Similar to embedding, stimuli may be played a low volume to make the task more difficult. This avoids possible confounds of adding white/pink noise to the signal, but has the added problem that different people have different hearing thresholds for different frequencies (the classic equal loudness contour is only an average of normal human hearing.) This means the signal to noise ratio cannot be tightly controlled and frequencies will be differentially affected. One option to mitigate this problem is to perform an adaptive test (see section 3.2) for each subject separately before the primary experiment to determine the proper presentation level.

4.3 Audio-Visual Stimuli

As speech is multimodal many examples of audio-visual experiments abound, the most famous being the initial report of the McGurk effect (McGurk and McDonald 1976). In an AV experiment subjects both see and hear the stimuli. Sometimes subjects can only see the stimuli as a way to assess the contribution of the visual component alone. Many of the presentation methods listed above also apply to video stimuli, such as embedding in noise (e.g. Vatikiotis-Bateson et al. 1998), degredation /simplification (Rosenblum and Saldaña, 1996), etc.

To record such stimuli a video camera is obviously needed; most nowadays are digital and specify their resolution in pixels (the more the better). The built-in microphones on such cameras may not be considered adequate for speech research. To get around this, many cameras have an input for a separate mic, though you are still reliant on the internal pre-amp (if any) and electronics of the camera, which may not be satisfactory. Alternatively, a separate audio recording can be made and synched, either live while recording or after the fact. The ability to synch live depends on the software available to the researcher and the camera in question[7].

There are many programs available for editing the audio and video signals. These vary from high-end professional quality ones like Adobe Premiere Pro and Apple's Final Cut Pro which were designed for television and movie editing down to freeware with very basic capabilities for home-video production. Most all programs allow for the cutting and synching of audio and video tracks and the creation of movie files in a variety of formats. Special care should be taken when synching to maintain naturalistic productions, an audio track slightly leading its corresponding video track is more disconcerting than the opposite.

5 Common Data Measures

This section lists some common types of data produced by perception research designs. These are not mutually exclusive and some experiments can produce several types of data. I have included some suggested statistical analyses where appropriate (but again, many more are possible and appropriate!)

Counts: Counts are simply tallies of responses arranged by categories. The *chi-squared* (X^2) statistical test is commonly used on count data, as is linear/logistical regression when categories of counts are related (such as responses to a continuum of stimuli). Skewed data can often be made normal through a square root or cube root transform.

Proportions: A proportion is the number of responses for a given category as a part of some larger (such as the total) number of presentations. It is commonly used for correct/incorrect responses, e.g. percent correct. Many researchers find it necessary to transform proportion data because it is not normally distributed (it has an upper and lower bound, 1 and 0, respectively.) A typical transform to adjust for this is the *arcsine* transform. A logistic regression analysis is usually most appropriate for such data.

***d′* (*d-prime*):** This is a measure of sensitivity derived from proportion / percent correct data. This measure takes into account subject bias in responses by incorporating both accuracy (called "hits") and false positives (called "false alarms") in a single measure. By definition it requires some measure of accuracy, and therefore there must be a possible "correct" response for each trial. It is a well understood and frequently applied measure of sensitivity, considered superior to other measures of accuracy for most applications. Depending on the design used to collect the data, the *d′* calculation varies from the quite simple (*z*-transformed proportions) to computationally complex, requiring table look-ups or special software. Additionally *d′* offers a measure of subject bias, the *criterion* (*c*), though this seems to be rarely reported. It may, however, be used to assign subjects to groups for statistical analyses (see Guenther et al. 1999 for an example). The bible of *d′* is the aforementioned Macmillan and Creelman (2005).

Reaction Times (RT): This measure is the speed of response, usually timed from the presentation of the stimulus[8]. Reaction time is generally easy to calculate assuming hardware having sufficiently

[7] Older video systems that have separate hardware for mixing do this easily. They suffer in being bulky, expensive, and less flexible than modern software.

[8] Some researchers separate reaction time from response time. Here the former records the timing to the initiation of the response and the latter records the timing to the completion of the response. In most speech research such a distinction is not made and the two terms are used interchangeably. This is not true, however, of eye-tracking paradigms (though there is a separate set of jargon in that research.)

accurate timing. Keyboards, mouses, and other peripheral devices should be assumed to have inadequate timing resolution unless shown otherwise. Usually only specially made "button boxes" are appropriate. RTs are often highly variable and the required motor response can add further variability (some researchers control for handedness[9], for example, though this is far from universal.) The start of the RT timing is crucial to accurate interpretation of results and when designing an experiment the subject's hypothesized time-course of decision making must be incorporated (i.e. at what point can a subject make an accurate decision relative to when timing began). Because collected RTs have a lower-bound, they are often skewed right and require a transformation (usually *log*) in order to be normally distributed. Only reaction times to correct answers are analyzed, meaning a considerable loss of data for very difficult tasks (however see Winters 200x for an alternative way at looking at RTs).

Scaling: Scaling data (also called "rating") consist of subjective evaluations of a given dimension of a stimulus or stimulus pair using numbers on a specified numerical scale (such a scale is also known as a Likert Scale.) For example, subjects may rate the "word-likeness" of a nonce word on a scale of 1 – 5 or the "similarity" of pairs of sounds on a scale from 1 to 10. Because subjects may vary on the way they use a scale, it is common to transform the data into a standardized unit, such as a *z*-score.

Magnitude Estimation: A variant on scaling where there are no absolute endpoints. That is, subjects decide the relative distance of given stimuli using ratios. Typically this takes the form of the presentation of a referent, called the *modulus*, and subsequent stimulus presentations where subjects judge the similarity to the modulus in terms of relative values. Numbers may be assigned such that the modulus is given a defined value and subjects label subsequent stimuli with numbers demonstrating similarity to the modulus. A famous example is the sone scale. S.S. Stevens instructed listeners to adjust a presented tone to be twice or half as loud as a previously presented tone, thus generating a measure of perceived loudness. In general magnitude estimation data is considered more accurate than scaling data at the expense of being more challenging to explain the methodology to subjects and making accurate RT collection more difficult (or impossible).

Eye Tracking (saccades and fixations): Eye tracking is a recent advance in methodology applied to speech perception, largely as a replacement for manual reaction time measurements (though it can be used simply to asses what a subject is looking at in an AV experiment, e.g. Vatikiotis-Bateson et al. 1998). It has two primary advantages. First, the eyes have a much shorter time-course in reacting to stimuli than manual button pressing, reducing noise when assessing processing time. Second, the eyes scanning the visual field allow the researcher to directly observe which alternatives a subject considered before making a decision, rather than inferring such knowledge from comparing reaction times on different trials/blocks. A typical design involves subjects "fixating" on the center of a display screen. A sound is presented which the subject is instructed to identify on the screen (usually by clicking), and several (usually two to four) objects are presented on the screen. The subject's gaze is followed as they scan the possible objects until they "fixate" on the correct object. The two measures usually derived from such experiments are *saccades* and *fixations*. *Saccades* are the eye movements and *fixations* are locations of the pauses. Together they are referred to as the *scanpath*.

[9] This is usually done by balancing handedness or more often (and usually more easily) to balance the buttons.

6 Glossary of Experiment Components

The following is a listing of some basic components found in many experimental designs., most of which were referenced above.

Sessions: A session is a single continuous period of experimentation for a given subject. These usually last one half to one hour. The duration of a session is generally limited by the ability of the subject to concentrate on the task at hand. Longer sessions require brief breaks within them and preferably multiple tasks to reduce stress on subjects and increase subject performance. When more data is needed than can be gained from a subject in a single session, multiple sessions are required, usually on separate days.

Instructions: Written instructions are usually given to subjects before the experiment with minimal verbal input. A computer screen can be used for this purpose or as a reminder. Ideally, instructions give just enough information for the subject to perform the task correctly without extraneous information that might introduce bias. One way this may be accomplished is to mislead the subject as to the experiment's purpose. IRB's often frown on this and usually require *any* deception to be revealed upon completion and the subject again be asked for consent to use their data.

Practice: To ensure subjects understand the task and can perform adequately, a block of practice trials is sometimes added at the beginning of the experiment. This block is usually brief (e.g. as few as four trials) and is either a random selection from the larger experimental set or a carefully selected group of trials that exemplify the expected performance demands placed on the subject (e.g only the easiest and/or most difficult trials). Another option is to give accuracy feedback during practice even if no feedback is given in the main experiment. Of course, any practice may bias the subject and should be included with care.

Inter-stimulus Interval (ISI): This is the time period separating each stimulus from other, temporally adjacent stimuli when multiple stimuli are presented in a single trial. Due to memory effects, shorter ISIs usually result in better performance (though ISIs of 0 *decrease* performance, see Pisoni 1973), and are generally between 100ms and 1000ms. See the discussions of AX Discrimination, Speeded AX, and 4IAX for more on ISIs.

Inter-trial Interval: Analogous to ISI, this is the timing from the end of one trial to the beginning of the next and includes any feedback or instructions. Unlike ISIs, shortening the ITI can increase the difficulty of the task and lengthening can ease a task. Occasionally the ITI is controlled by the subject, i.e. they advance from trial to trial under their own volition..

Feedback: Feedback alerts the subject to their performance in the experiment. Feedback can be used to give subjects a target to reach or expected performance level to hold. It often has the added benefit of increasing or at least holding a subject's attention in a tedious task and may be necessary for adequate performance. Feedback is given after every trial or at the end of a block of trials and often include summarized data. Typical feedback includes accuracy, reaction time, and/or the labels assigned by the subject.

Blocks: A block is a subset of trials in the experiment. Blocks may be used simply to provide breaks for subjects or may be used for more sophisticated means. In a given block of trials a subject may see only a subset of stimuli, for example one block may have a subject discriminate fricatives while in a second block they discriminate stops. In such an experiment the goal of blocking is to create a limited set of contrasts for the subject and presumably changes their expectations. This only works if the subject can quickly identify what is possible, so the ratio of contrasts to trials should be low. One example of the use of blocks in this way is the Garner Paradigm (Garner 1974). This paradigm uses only four stimuli and one basic procedure (yes-no) combined with several different blocking types to tell integral from separable dimensions.

Breaks: Breaks are frequently used to reduce stress on subjects. There are no hard rules on how many or how frequently they should be inserted. They may be strictly timed or the subject can advance the experiment on their own. If the experiment procedure is separated into blocks it's common to include breaks between them, otherwise they may be included at specified time intervals. Another option is to give subjects an escape button or key to pause the experiment at their discretion.

Follow-up questionnaire: After the conclusion of an experiment its common to give a questionnaire to the subject asking questions about the experiment itself. The goals of such questionnaires are usually to find out the strategies used by the subject to complete the experiment, their impressions of the stimuli (naturalness, for example), or any general impressions. Though not often reported as data, they are often useful when interpreting the data and designing subsequent experiments.

Compensation: Subjects are usually compensated for their time in money or course credit. A common going rate is $10 an hour, with more or less given based on the demand of the task or the difficulty of recruiting the subjects. In some instances the amount of compensation varies based on the subject's performance. Generally IRB's are uncomfortable with this for various reasons, but can be persuaded to allow it if the subjects are told their performance affects compensation but in fact all are compensated equally (deception), or all are compensated equally to a given amount with additional compensation available under strict conditions.

Pilot Experiment(s): This is trial run of an experiment used for the purpose of evaluating the experimental design before committing further resources to the experiment. Any possible confounds can be explored and proposed analyses can be tested. Pilot work is especially useful for testing stimuli, subject instructions, and the variability of subject performance. This last point is useful in doing a power calculation of the proposed statistical analyses as you can get some idea of the variability of the data you will collect.

7 *The Good, the Bad, and the IRB*

Perception research requires subjects, and this requires permission from your Institutional Review Board (IRB)[10]. Getting this permission can be a frustrating experience given the generally benign nature of our research, but keeping a few things in mind can streamline the process. You should always keep in mind that the goal of the IRB is to protect human subjects from abuse and embarrassment, mental and physical. The regulations originated in the horrors of Nazi and Japanese experiments on live, non-consenting humans in WWII as well as subsequent experiments by the US Government, like the Tuskegee Syphilis Study (not to mention University-sanctioned studies like the “Monster Study” on stuttering conducted in my hometown).

[10] This section is primarily applicable to the US, which tends to have some of the strictest regulations when it comes to human subjects.

Of course, speech perception experiments are far from Mengelian in concept, but when human subjects are involved, erring on the side of caution is always a good way to go.

In general, IRBs want to know that the subject is fully aware of what is required in an experiment and having that knowledge they fully consents to participate, with no coercion. Only adults 18 and older with normal mental capabilities can consent to participate (and this is the usual population for our studies.) You should be aware that some IRBs will grant exceptions to review of speech perception research if they consider it an "educational test (cognitive / diagnostic)". My current IRB allows most of my research under this rubric; my previous institution did not. Certainly any invasive production studies do not usually qualify (i.e. don't try and convince your IRB that you can do tracheal punctures without review.) When review is necessary, most speech perception research qualifies for the less onerous expedited review process.

When writing a protocol, write it as broadly as possible and let the IRB tell you when you are overreaching. For example, avoiding full, written consent forms and substituting verbal consent is desirable, both for avoiding the paperwork as well as avoiding having a subject's signature on record. Similarly, if you do not want to encrypt the data on a password protected file server, then do not say you will do that. If the IRB requires it, then you have to do it, but make them tell you. Always keep in mind ways you may want to use the data you collect. For example, if you are collecting voice recordings, ask consent to use those recordings in further experiments, even if it's only a remote possibility as you cannot go back and get retroactive permission to do so. View the process of getting a protocol approved as a dialogue between you and the IRB, that is, you may gently argue with them about aspects you view as onerous. Don't be afraid to include supplementary documentation supporting your point.

There are four main issues that come up in speech research that IRBs may balk at. One particularly frustrating one is some may consider voice recordings as "identifiable". While having a dubious theoretical basis (I blame Hollywood and "voice print identification"), it may be best to simply concede the point and mention in a consent form that there is a possibility someone participating in an experiment may identify them based on their voice; if they consent, great, if not you need a new participant. A second issue that often comes up is the possibility that a subject producing spontaneous speech, such as in a sociolinguistic interview situation, may reveal something they do not want recorded for posterity. If your IRB feels this is a serious concern, the best way to handle it is by using a post-interview consent form where the subject consents post-hoc for you to use their data. First they consent to participate, then after participating they consent for you to use what they had said. This is essentially the way you also handle deception, another frequent concern of an IRB.

"Deception" is operative any time you obviously lie to a subject about the experiment. In these cases you must reveal the deception afterward and get permission to use the data. Withholding information is a trickier situation, as we always withhold some amount of information from the subject to avoid bias. A good rubric for whether this is deception as the IRB would define it is whether there is a reasonable possibility that some subject would not consent to you using their data if they knew what you had withheld.

The final common issue that IRBs may be concerned with is performance-based compensation, i.e. where a subject makes more money[11] based on having higher performance. The problems with basing compensation on performance are that it can be considered unfair. Even more problematically, especially from the IRB's perspective, it can be seen as financial coercion. Your IRB should give you guidance on this,

[11] Course credit can never be given based on performance for rather obvious fairness reasons.

but one option that should make it more palatable is to give all subjects some baseline amount of money with performance adding to that total.

8 References

Best et al. 2001

Boersma, Paul & Weenink, David (2010). Praat: doing phonetics by computer [Computer program]. Version 5.1.43, retrieved 4 August 2010 from http://www.praat.org/

Choo Huckvale

Durlach and Braida 1969

Fox 1984

Francis and Ciocca, 2003

Guenther et al. 1999

Kaernbach, 1991

Logan & Pruitt 1998

MacMillan Creelman 2003

McGurk and McDonald 1976

Miller Nicely

Pisoni 1973

Pisoni and Tash

Rosenblum and Saldaña 1996

Brockwell and Davis *Introduction to Time Series and Forecasting*