Massaro, D. W. (2003). Model Selection in AVSP: Some Old and Not So Old News. In J.L. Schwartz, F. Berthommier, M.A. Cathiard & D. Sodoyer (Eds.), *Proceedings of Auditory-Visual Speech Processing (AVSP'03), ISCA Tutorial and Research Workshop on Audio Visual Speech Processing* (pp. 83-88; also CD-ROM). St Jorioz,

Model Selection in AVSP: Some Old and Not So Old News

Dominic W. Massaro
Perceptual Science Laboratory, Department of Psychology,
University of California, Santa Cruz, Santa Cruz, CA. 95064 U.S.A.
massaro@fuzzy.ucsc.edu

Abstract

We reiterate a paradigm of inquiry in auditory visual speech processing, focusing on appropriate experimental procedures and methods of model selection. Several methods of model selection find evidence in support of the fuzzy logical model of perception (FLMP). We caution investigators to not limit themselves to simply testing the classic McGurk effect because its outcomes cannot distinguish among alternative interpretations.

1. A Paradigm for Inquiry

The study of speech perception by ear and eye has been and continues to be a powerful paradigm for uncovering fundamental properties of the information sources in speech and how speech is perceived and understood. Our general framework documents the value of a combined experimental/theoretical approach. The research has contributed to our understanding of the characteristics used in speech perception, how speech is perceived and recognized, and the fundamental psychological processes that occur in speech perception and pattern recognition in a variety of other domains.

We believe that our empirical work would be inadequate and perhaps invalid without the corresponding theoretical framework. Thus, the work continues to address both empirical and theoretical issues. At the empirical level, experiments have been carried out to determine how visible speech is used alone and with auditory speech for a broad range of individuals and across a wide variation of situational domains. At the theoretical level, the assumptions and predictions of several models have been analyzed, contrasted, and tested. In addition, a general framework for inquiry and a universal principle of behavior has been proposed.

1.1. Critique of the McGurk Paradigm

In AVSP'98 (Massaro, 1998a), I criticized the approach of the many experiments of multimodal speech perception carried out in the context of the McGurk effect, a striking demonstration of how visual speech can influence the perceiver's perceptual experience. The classic McGurk effect involves the situation in which an auditory /ba/ is paired with a visible /ga/ and the perceiver reports hearing /da/. The reverse pairing, an auditory /ga/ and visual /ba/, tends to produce a perceptual judgment of /bga/. Most studies of the McGurk effect, however, use just a few experimental conditions in which the auditory and visual sources of information are made to mismatch. Investigators also sometimes fail to test the unimodal conditions separately so that there is no independent index of the perception of the single modalities. The data analysis is also usually compromised because investigators analyze the data with respect to whether or not there was a McGurk effect, which often is simply taken to mean whether the auditory speech was accurately perceived. Investigators also tend to take too few observations under each of the stimulus conditions, which precludes an analysis of individual behavior and limits the analyses to group averages. Because I realized that the data from the McGurk paradigm were underdetermined, we did not carry out any formal model tests of results from this paradigm. Schwartz (this volume) discovered another huge problem with the McGurk paradigm when he attempted to test the FLMP against results in this task. Before discussing this discovery, we describe our experimental and theoretical approach.

1.2. Varying the Ambiguity of the Modalities

An important manipulation is to systematically vary the ambiguity of each of the source of information in terms of how much it resembles each syllable. Synthetic speech (or at least a systematic modification of natural speech) is necessary to implement this manipulation. In a previous experimental task, we used synthetic speech to cross five levels of audible speech varying between /ba/ and /da/ with five levels of visible speech varying between the same alternatives. We also included the unimodal test stimuli to implement the expanded factorial design.

1.2.1. Prototypical Method.

The properties of the auditory stimulus were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, properties of our animated face were varied to give a continuum between visual /ba/ and /da/. Five levels of audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. In addition, the audible and visible speech also were presented alone for a total of 25 + 5 + 5 = 35 independent stimulus conditions. Six random sequences were determined by sampling the 35 conditions without replacement giving six different blocks of 35 trials. An experimental session consisted of these 6 blocks preceded by 6 practice trials and with a short break between sessions. There were 4 sessions of testing for a total of 840 test trials (35 x 6 x 4). Thus there were 24 observations at each of the 35 unique experimental conditions. Participants were instructed to listen and to watch the speaker, and to identify the syllable as /ba/ or /da/. This experimental design was used with 82 participants and their results have served as a database for testing models of pattern recognition (Massaro, 1998b).

1.3.1. Prototypical Results

We call these results prototypical because they are highly representative of many different experiments of this type. The mean observed proportion of /da/ identifications was computed for each of the 82 participants for the 35 unimodal and bimodal conditions. Figure 1 shows the results for a single participant who can be considered typical of the others in this task.

The points in Figure 1 give the observed proportion of /da/ responses for the auditory alone, the bimodal, and the visual alone conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and

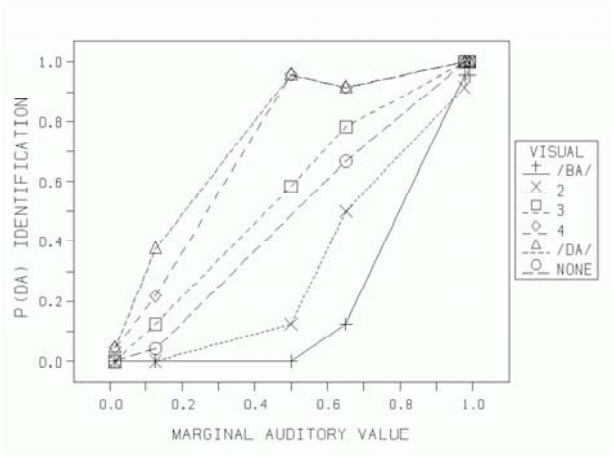

Figure 1. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. Results for participant 9.

/da/. Although this plot of the results might seem somewhat intimidating at first glance, I believe a graphical analysis of this nature can facilitate understanding dramatically. Notice that the columns of points are spread unevenly along the x-axis. The reason is that they are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. This spacing reflects relative influence of adjacent levels of the auditory condition.

The unimodal auditory curve (indicated by the open circles) shows that the auditory speech had a large influence on the judgments. More generally, the degree of influence of this modality when presented alone would be indicated by the steepness of the response function. The unimodal visual condition is plotted at .5 (which is considered to be completely neutral) on the auditory scale. The influence of the visual speech when presented alone is indexed by the vertical spread among the five levels of the visual condition.

The other points give performance for the bimodal conditions. This graphical analysis shows that both the auditory and the visual sources of information had a strong impact on the identification judgments. The likelihood of a /da/ identification increased as the auditory speech changed from /ba/ to /da/, and analogously for the visible speech. The curves across changes in the auditory variable are relatively steep and also spread out from on another with changes in the visual variable. By these criteria, both sources had a large influence in the bimodal conditions. Finally, the auditory and visual effects were not additive in the bimodal condition, as demonstrated by a significant auditory-visual interaction. The interaction is indexed by the change in the spread among the curves across changes in the auditory variable. This vertical spread between the curves is many many times greater in the middle than at the end of the auditory continuum. It means that the influence of one source of information is greatest when the other source is neutral or ambiguous. We address how the two sources of information are used in perception. We formalize two competing models and test them against the results.

To explain pattern recognition, representations in memory are an essential component. The current stimulus input has to be compared to the pattern recognizer's memory of previous patterns. One type of memory is a set of summary descriptions of the meaningful patterns. These summary descriptions are called prototypes and they contain a description of features of the pattern. The features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. To recognize a speech segment, the evaluation process assesses the input information relative to the prototypes in memory. Given this general theoretical framework, we consider whether or not integration of auditory and visual information occurred. It might seem obvious that integration occurred in our experiment because there were strong effects of both auditory and visual speech in the bimodal conditions. In fact, this outcome is logically possible even if integration did not occur. Most experiments using the McGurk effect paradigm were not able to demonstrate conclusively that integration occurred. It is possible, for example, that only the visual speech was used on some of the trials and simply determined the judgments on these trials. This type of nonintegration is the simpler account of pattern recognition and we begin with a formalization of this type of model.

According to nonintegration models, any perceptual experience results from only a single sensory influence. Thus the pattern recognition of any crossmodal event is determined by only one of the modalities, even though the influential modality might vary from trial to trial. Although this class of models involves a variety of alternatives that are worthy of formulation and empirical test (see Massaro, 1998b), we will formulate and test just one for illustrative purposes.

2. Single Channel Model (SCM)

Although there are multiple inputs, it is possible that only one of them is used. This idea is in the tradition of selective attention theories according to which only a single channel of information can be processed at any one time. According to the single channel model (SCM), only one of the two sources of information determines the response on any given trial. Given a unimodal stimulus, it is assumed that the response is determined by the presented modality. A unimodal auditory stimulus will be identified as /da/ with probability a_i, and, analogously, the unimodal visual stimulus will be identified as /da/ with probability v_j. The value i simply indexes the ith level along the auditory continuum and j indexes the level of the visual input.

Given that only one of the auditory and visual inputs can be used on any bimodal trial, it is assumed that the auditory modality is selected with some bias probability p, and the visual modality with bias 1 - p. If only one modality is used, it is reasonably to assume that it will be processed exactly as it is on unimodal trials. In this case, for a given bimodal stimulus, the auditory information will be identified as /da/ with probability a_i, and the visual information with probability v_j. Thus, the predicted probability of a /da/

response given the ith level of the auditory stimulus, a_i, and the jth level of the visual stimulus, v_j, is

$$P(/da/ \mid A_i V_j) = pa_i = (1 - p)v_j \quad (1)$$

Equation 1 predicts that a /da/ response can come about in two ways: 1) the auditory input is selected and is identified as /da/, or 2) the visual input is selected and is identified as /da/. This formalization of the SCM model assumes a fixed p across all conditions, an a_i value that varies with the auditory information and a v_j value that varies with the visual information.

We can assess the predictive power of the SCM and other models using the 5 by 5 expanded factorial design. The points in Figure 1 gives the proportion of /da/ identifications for a prototypical participant in the task. Equation 1 is a linear function and it predicts a set of parallel functions with this type of plot, which is clearly contradicted by the data in Figure 1. This mismatch between the observations and predictions illustrates that this model appears to be inadequate. Even so, a formal test is required. Before we present this test of the SCM, it is necessary to discuss estimation of the free parameters in a model.

3. Testing a Model's Predictions

We cannot expect a model's predictions of behavior to be exact or even very accurate without first taking into account what results are being predicted. As an example, we cannot know exactly how often a given person will identify one of the visible speech syllables as a particular alternative. Individual participants give similar but not identical results for the same experiment. We can know that one syllable might be more likely to be identified as /ba/ but we cannot predict ahead of time the actual probability of a /ba/ response by an individual participant. This uncertainty would preclude the quantitative test of models if we were not able to determine (estimate) the values of free parameters. Schwartz (this volume) endorses the strategy of estimating parameters in order to test among models.

When applied to empirical data, most computational or quantitative descriptions have a set of free parameters. A free parameter in a model is a variable whose values cannot be exactly predicted in advance. We do not know what these values are, and we must use the observed results given to find them. The actual performance of the participant is used to set the value of this variable. This process is called parameter estimation. In parameter estimation, we use our observations of behavior to estimate the values of the free parameters of the model being tested. Because we want to give every model its best shot, the goal is to find the values of the parameters that maximize how accurately the model is able to account for the results. The optimal parameter values can be found with an iterative search algorithm to find those parameter values that minimize the differences between the predicted and observed results. The parameters and parameter space must be specified for the search. In the SCM, for example, the parameters are p, a_i, and v_j. These values are probabilities and thus must be between 0 and 1. In our model fitting technique, we usually estimate the free parameters are estimated based on all of the conditions, not just the unimodal ones. We have rationalized why this approach is more optimal, which is accepted by Schwartz (this volume).

Equation 1 predicts P(/da/) for each of the 35 conditions in the expanded factorial experiment. The SCM does not predict in advance how often the syllable in each modality will be identified as /ba/ or /da/. According to the model, there can be a unique value of a_i for each unique level of audible speech. Similarly, there can be a unique value of v_j for each level of visual speech. We also do not know the value of p on bimodal trials, which requires another free parameter. For unimodal trials, we assume that the presented modality is always used. We have 35 equations with 11 free parameters: the p value, the 5 a_i and 5 v_j values. Finding values for these 11 unknowns allows us to predict the 35 observations.

3.1. RMSD Measure of Goodness-of-Fit

A factor that is often used to maximize the goodness-of-fit is the root mean squared deviation (RMSD) between the predicted and observed values. The best fit is that which gives the minimal RMSD. The RMSD is computed by a) squaring the difference between each predicted and observed value, b) summing across all conditions c) taking the mean, and d) taking the square root of this mean. (Squaring the differences makes all differences positive and also magnifies large deviations compared to small ones.) The RMSD can be thought of as a standard deviation of the differences between the 35 predicted and observed values. The RMSD would increase as the differences increase. In general, the smaller the RMSD value, the better the fit of the model.

The quantitative predictions of the model are determined by using any minimization routine such as the program STEPIT. The model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program maximizes the accuracy of the predictions by minimizing the RMSD. The outcome is a set of parameter values which, when put into the model, come closest to predicting the observed results.

The results for the present model tests come from the results from 82 participants, with 24 observations from each participant under each of the 35 conditions (Massaro, 1998b). The model fit was carried out separately on each participant's results. We have learned that individuals differ from one another and averaging the results across individuals can be hazardous. The free parameters of a model should be capable of handling the individual differences. Fitting a model to single individuals should permit the model to describe individual participants while also accounting for between-participant differences, insofar as they can be captured by the differences among the 11 parameters.

The predictions of the SCM do not capture the trends in the data. The predictions are a set of parallel lines whereas the observations are spread in the middle and narrow at the ends. The RMSD is also used to evaluate the goodness-of-fit of a model both in absolute terms and in comparison to other models. The RMSDs for the fit of the SCM across all 82 participants averaged .097.

4. The Fuzzy Logical Model of Perception (FLMP)

The Fuzzy Logical Model of Perception (FLMP) assumes that multiple sources of information contribute to the

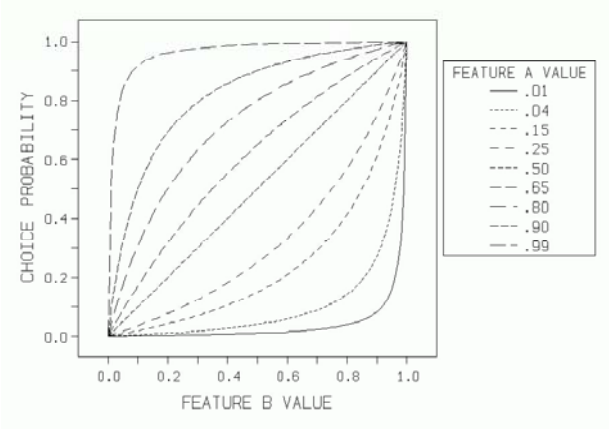

Figure 2. Hypothetical predictions of the FLMP with two-alternatives and two sources of information.

identification and interpretation of the language input. The assumptions central to the model are 1) each source of information is evaluated to give the continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated multiplicatively to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives (Massaro, 1998b). In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1 - a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to its feature value. The predicted probability of a /da/ response given an auditory input, $P(/da/|A_i)$ is equal to

$$P(/da/ \mid A_i) = \frac{a_i}{a_i + (1 - a_i)} = a_i \qquad (2)$$

Similarly, the predicted probability of a /da/ response given an visual input, $P(/da/|V_j)$ is equal to

$$P(/da/ \mid V_j) = \frac{v_j}{v_j + (1 - v_j)} = v_j \qquad (3)$$

For bimodal trials, the predicted probability of a /da/ response given auditory and visual inputs, $P(/da/|A_iV_j)$ is equal to

$$P(/da/ \mid A_iV_j) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)} \qquad (4)$$

Equations 2-4 assume independence between the auditory and visual sources of information. Independence of sources at the evaluation stage is motivated by the principle of category-conditional independence (Massaro, 1998b). Given that it isn't possible to predict the evaluation of one source on the basis of the evaluation of another, the independent evaluation of both sources is necessary to make an optimal category judgment. Although the sources are kept separate at evaluation, they are integrated to achieve perception, recognition, and interpretation. The FLMP assumes multiplicative integration, which yields a measure of total support for a given category identification. This operation, implemented in the model, allows the combination of two imperfect sources of information to yield better performance than would be possible using either source by itself. However, the output of integration is an absolute measure of support; it must be relativized, which is effected through a decision stage, which divides the support for one category by the summed support for all categories. Figure 2 gives hypothetical predictions of the FLMP with two-alternatives and two sources of information. As can be seen in the figure, the curves trace out an American football shape that budges at the extreme top-left and bottom-right corners when extreme feature values are opposed to one another.

The hypothetical predictions can be used to illuminate the observations that the classic McGurk conditions cannot be used to test the FLMP (see Schwartz, this volume). Consider the predictions when the Feature B value is near 0 and the Feature A value is .99. This corresponds to the classic McGurk effect when a visual /da/ is paired with an auditory /ba/, for example. The bimodal choice probability, however, can value dramatically with very small changes in the value of Feature B. It follows that a small change in the value of Feature B can accurately describe just about any bimodal choice probability, which makes the FLMP nonfalsifiable in this selected test. It should be noted, however, that the same might be true for other models such as the SCM. In this case, a change in the probability of using a particular modality would accomplish the same outcome.

One lesson to be learned from this is that it is necessary to test models against data that require feature values in the middle of the football. Here it can be seen that a large change in parameter value is needed to predict changes in the choice probability. The lesson learned from this analysis is that a valid test of the FLMP cannot be limited to data that have only unimodal judgments near 0 or 1 but must also include data whose unimodal judgments are in the middle range of response probabilities. One way to assess whether your data set is valid in discriminating two models is to 1) fit the models to the results, 2) cross-fit each model to simulated data generated from the predictions of the other rmodel, 3) and find that the two models are equally good at fitting their own simulated data and equally poor at fitting the simulated data from the other model. To see if these conditions hold for Schwartz (this volume) results, the FLMP was fit to the Schwartz results with 11 response alternatives and also when reduced to just 6 response alternatives. The RMSDs were .015 and .017, respectively. The fit of the SCM was much poorer with RMSDs of .083 and .112. When simulated data were created by the SCM from its predictions and then fit by the FLMP, the fit was much worse than the FLMP fits to the original data with RMSDs of .036 and .047. Given slightly larger flexibility of the FLMP relative to the SCM for the Schwartz results, the experiment underdetermines any test between the models on this particular data set.

The individual results of the 82 participants in the 5 by 5 expanded factorial design do qualify as a data set that can distinguish between the models (Massaro, 1998b). The FLMP was fit to the results using Equations 2-4 with 10 free parameters. Like the SCM, the FLMP also requires 5 a_i and 5 v_j values. In the FLMP, however, these are not probabilities but fuzzy truth values between 0 and 1 indicating the degree to which the information supports the alternative /da/ (see Equations 2-4). The RMSD for the fit of the FLMP for the participant shown in Figure 1 was .051, and the RMSDs for the fit of the FLMP for the 82 individual participants averaged .051.

As in all areas of scientific inquiry, it is important to replicate this task under a broader set of conditions. These basic findings hold up under a variety of experimental conditions (Massaro, 1998b, Chapter 6). In one case, participants were given just two alternatives, and in the other the same participants were allowed an open-ended set of alternatives. When tested against the results, the FLMP gives a good description of performance, even with the constraint that the same parameter values are used to describe performance when the number of response alternatives is varied (see Massaro, 1998b, pp. 265-268).

Given the delicate nature of testing among quantitative modes, we have explored alternative methods of model testing. The first involves the match between the goodness-of-fit of a model and a benchmark measure that indexes what the goodness of fit should be if indeed the model was correct. Because of sampling variability, we cannot expect a model to give a perfect description of the results. Second we have used a model selection procedure suggested by Myung and Pitt (1997; Massaro et al., 2001), and by Schwartz (this volume).

5. Model Selection using Bayes Factor

This Bayes factor method of model selection seeks to handicap models to the extent they can predict a large range of outcomes with changes in their parameter values. If a model predicts a large range of outcomes with changes in parameter values, then its ability to predict a single data set with all possible parameter values will be very poor. On the other hand, a model that predicts only a small range of outcomes across changes in its parameter values will do much better if the data to be predicted are within that small range of outcomes predicted by the model. This is the logic of handicapping models based on their flexibility.

The Bayes factor adjusts a model's goodness-of-fit index by the model's ability to describe a large range of different data configurations. One model capable of fitting a broader range of data configurations than another is not necessarily the better model. We desire a model to have good taste and to predict only a constrained set of data outcomes—if any configuration of data can be predicted, it is not falsifiable. The Bayes factor handicaps a model to the extent that it can predict a broad range of data configurations other than the observed data, by simply different parameter values. According to the assumptions underlying Bayes factor, a better model is one that predicts only data close to the data actually observed, regardless of the parameter values.

This important analysis and potential solution provided by the Bayes factor alerted us to the possibility that our previous model tests may have led us to incorrect conclusions. In many experiments, the FLMP has been found to provide a significantly better fit than alternative models. The demonstration of Myung and Pitt reveals that our conclusions might have been invalid given the potentially more flexibility of the FLMP to fit results, even results that were not generated by that model. There were several aspects of the Myung and Pitt simulation, however, that did not mirror our prototypical experimental situations. First, the authors simulated data from an unweighted averaging model (LIM) rather than a weighted averaging model (WTAV) that we have tested in all of our research (Massaro, 1998b). The WTAV is mathematically equivalent to the SCM, even though they are formalizations of very different assumptions about the nature of auditory visual speech processing. The FLMP always gave a significantly better fit than the WTAV even though the WTAV also had one additional free parameter compared to the FLMP, and we did not adjust the RMSD measures to reflect this difference in number of parameter values. We expect the WTAV to be more flexible than the LIM, which would influence the outcome of model selection using Bayes factor.

Weighted averaging is more psychologically realistic than unweighted averaging in that it is unlikely that each influential factor contributes equally to performance in pattern recognition tasks. A weighting parameter allows that a .7 scale value from one factor might make a different contribution than a .7 value from another factor. Differential weighting in the FLMP and TSD descriptions emerges from the nonlinear combination of the two sources of information corresponding to the two factors. Second, the authors simulated data from a highly asymmetrical factorial design whereas we usually carry out symmetrical expanded factorial designs. The latter are much more efficient than the former in discriminating among different models. A symmetrical design has the highest ratio of independent observations relative to free parameters, and the expanded design provides an additional set of data points whose expected values are predicted by the same parameter values. Third, the authors used only three hypothetical sets of parameter values to generate hypothetical data whereas we have contrasted the models in literally dozens of independent tests.

To directly insure that the Bayes factor does not revise our conclusions in past work, we tested the FLMP against the WTAV model using the Bayes factor for our prototypical design. These two models were fit to the observed data of the 82 subjects. Using the Bayes factor selection method, the FLMP fit better than the WTAV for 80% of the subjects. Although this difference was statistically significant, 80% wins is still somewhat short of the 94% wins using the RMSD selection method. Because of this discrepancy, we repeated the Bayes factor with 5,000,000 rather than 500,000 iterations in the computation of the marginal likelihoods. The FLMP now fit better than the WTAV for 94% of the subjects. These results support the idea that the RMSD measure yields similar conclusions to the Bayes factor for the conditions of our prototypical design.

We explored the same question when the number of response alternatives was eight. We replicated our basic 5 by 5 expanded factorial design with 8 rather than just 2 response alternatives. This basic task was carried out in 4 different experiments to give a total of 36 subjects in this data set (Massaro, 1998b, Chapter 10). The observed responses of the 36 subjects served as the data for the Bayes factor. The FLMP fit with a log likelihood of –163.5 while the WTAV model performed worse with a log likelihood of –180.2. Using the Bayes factor, the FLMP fit 97% of the subjects better than the WTAV model. These results are consistent with those obtained with the RMSD measure and further support our claim that the RMSD is an accurate measure of model performance for our prototypical design. Thus, we can conclude that previous model tests using RMSD as a measure of goodness of fit provided a valid selection of the FLMP over competing models.

The advantage of the FLMP over the SCM and other competing models holds up under these alternative

procedures of model testing (Massaro, 1998, Chapter 10; Massaro et al., 2001). Thus, the validity of the FLMP holds up under even more demanding methods of model selection. We propose that investigators should make use of as many techniques as feasible to provide converging evidence for the selection of one model over another. More specifically, both RMSD and the Bayes factor can be used as independent metrics of model selection. Inconsistent outcomes should provide a strong caveat for the validity of selecting one model over another in the same way that conflicting sources of information create an ambiguous speech event for the perceiver.

6. Standard Statistics versus Model Tests

One difficulty for consumers of our approach has to do with understanding apparent differences between statistical tests and model tests. The important point to understand is that the statistical significance of independent variables and their interactions do not align themselves with the validity of a model. This follows because significant performance differences do not necessarily mean that there are underlying information processing differences. An easy way to see this is to simply generate different sets of predictions of a model by allowing the free parameters to take on different values. The resulting outcomes would certainly be statistically different from one another even though the same model generated the results. The FLMP assumes that the auditory and visual sources of information are evaluated independently of one another. When these two sources of information are manipulated independently of one another in a factorial design, there is usually a statistically significant interaction. This interaction, however, does not necessarily mean that the two sources were not evaluated independently of one another. Specific model tests are required to address the goodness-of-fit of models; statistical analyses of performance differences are not really appropriate to test quantitative models.

The limitations of statistical analyses of performance differences can be seen in a recent study of Mandarin and English perceivers (Chen & Massaro, submitted). Both groups were tested in our standard 5 by 5 expanded factorial design to address the issue of whether they process speech differently from one another. A statistical analysis of the factorial data showed significant effects of the auditory and visual continua and an auditory-visual interaction. The difference between language groups was significant, as well as the 3-way interaction between language groups, visual, and auditory levels. However, these significant differences do not necessarily reflect differences in information processing. In fact, there were significantly more overall /da/ response-judgments for the Mandarin (mean = .58) speakers than for the English (mean = .44) speakers. This difference highlights the complexity of cross-linguistic research, which precludes accepting statistical differences as differences in the nature of information processing. We cannot expect the same speech continuum to be treated equivalently by talkers of two different languages (or even two talkers of the same language). It appears that our speech continuum was biased toward the labial place of articulation for the English talkers and biased away from the labial place for the Mandarin talkers.

7. Conclusions

We have made significant progress in our understanding of speech perception by ear and by eye. It was only less than two decades ago that researchers believed there was a "Preferred Modality for Speech Perception." (Seewald et al., 1985). This interpretation was based on the observation that the amount of influence of visible speech was positively related to a child's hearing loss. We now know, however, that all persons use both auditory and visual speech and the degree of influence of a modality is a function how informative that modality is and its ambiguity relative to other modalities.

We are attracted to bimodal speech perception as a paradigm for psychological inquiry for several reasons. It offers a compelling example of how processing information from one modality (vision) appears to influence our experience in another modality (audition). Second, it provides a unique situation in which multiple modalities appear to be combined or integrated in a natural manner. Third, experimental manipulation of these two sources of information is easily carried out in pattern recognition tasks. Conceptualizing speech as crossmodal has the potential for valuable applications for individuals with hearing loss, person with language challenges, learners of a new language, and for other domains of language.

8. Acknowledgements

The research and writing of the paper were supported by grants from National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Michael M. Cohen's contribution to this enterprise is greatly appreciated, along with the contributions of Christopher Campbell and Trevor Chen. I also appreciate the efforts of Jean-Luc Schwartz to illuminate aspects of the FLMP.

9. References

[1] Chen, T. H., & Massaro, D. W. (submitted). The Relationship between Unimodal and Bimodal Speech Perception of Mandarin Speakers. *Perception & Psychophysics*, submitted.

[2] Massaro, D. W. (1998a). Illusions and Issues in Bimodal Speech Perception. *Proceedings of Auditory Visual Speech Perception '98*. (pp. 21-26). Terrigal-Sydney Australia, December, 1998.

[3] Massaro, D. W. (1998b). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, Massachusetts: MIT Press.

[4] Massaro, D.W.; Cohen, M.M.; Campbell, C.S.; Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8*, 1-17.

[5] Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

[6] Schwartz, J.-L. (2003). Why the FLMP should not be appliced to McGurk data: Or how to better compare models in the Bayesian framework. (this volume).

[7] Seewald, R. C., Ross, M., Giolas, T. G., & Yonovitz, A. (1985). Primary modality for speech perception in children with normal and imparired hearing. *Journal of Speech and Hearing Research, 28*, 36-46.