Speech, Hearing and Language: work in progress

Volume 14

A CHOICE THEORY METHOD FOR EVALUATING AUDIOVISUAL PHONEME RECOGNITION

Paul IVERSON

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

A CHOICE THEORY METHOD FOR EVALUATING AUDIOVISUAL PHONEME RECOGNITION

Paul IVERSON

Abstract
This article describes a mathematical method, based on Choice Theory (e.g., Luce, 1963), that can be used to predict audiovisual phoneme confusion matrices from unimodal audio and visual data. The predictions made from this method can be compared to obtained levels of audiovisual processing, for the purpose of identifying individuals whose audiovisual integration processes are not efficient. A reanalysis of Grant et al.'s (1998) audiovisual consonant confusion data is presented to evaluate this method. The results demonstrate that this method is effective at predicting audiovisual phoneme recognition responses, and suggests that Grant et. al's. subjects were highly efficient at integrating audiovisual information. Matlab code used in these analyses is available at http://www.phon.ucl.ac.uk/home/paul/CT/home.htm.

Introduction

Several methods have been developed to predict audiovisual phoneme confusion matrices based on phoneme confusion data collected under unimodal audio and visual stimulus conditions (Blamey, 1989; Braida, 1991; Massaro, 1987). This article describes a new method[1], based on Choice Theory, which serves the same purpose. Compared to other existing methods, the present method offers at least two advantages.

First, this method is designed to estimate audiovisual phoneme recognition under *optimal-processing* conditions (see Braida, 1991; Grant et al., 1998); it estimates the highest level of audiovisual consonant recognition that is possible given the phonetic information available separately through the auditory and visual modalities. Predictions based on optimal processing assumptions are useful, because they can be compared to obtained levels of performance to help identify individual patients whose cognitive/perceptual processes (e.g., processes that integrate the phonetic information from each modality and map the phonetic information onto long-term memory representations for language) are not making efficient use of the available phonetic information (Grant et al., 1998; Grant & Seitz, 1998).

Second, this method is less complex mathematically than the only other proposed method for estimating audiovisual performance under optimal-processing conditions, Braida's pre-labeling model (1991). Braida's pre-labeling model is based on a multidimensional extension of Signal Detection Theory (e.g., Durlach & Braida, 1969; Green & Swets, 1966; Macmillan et al., 1988). It requires fitting the consonants to locations within a multidimensional audiovisual perceptual space, calculating response regions for each consonant within this space, and integrating a multidimensional Gaussian probability function for each consonant over each of these response regions. In contrast, the Choice Theory method used here does not require the consonants to be represented within a multidimensional space (although it is possible to represent Choice Theory coefficients

[1] Matlab code used in these analyses is available at http://www.phon.ucl.ac.uk/home/paul/CT/home.htm.

multidimensionally if this is desired; see Nosofsky, 1985), and involves simpler probability calculations.

1. The Choice Theory Method

The core of the method is the use of Luce's choice rule (1963) to estimate perceptual similarities and response biases from phoneme confusion matrices obtained under audio, visual, and audiovisual stimulus conditions. The rule is mathematically expressed as

$$p(r_i \mid s_j) = \frac{\alpha_{ij} b_i}{\sum_k \alpha_{kj} b_k} \quad (1)$$

where $p(r_i|s_j)$ is the probability that the response phoneme i will be given by the subject when stimulus phoneme j is presented, α_{ij} is the similarity between phonemes i and j, b_i is the bias for giving phoneme i as a response, and the denominator is the sum of the similarity-bias products for all response phonemes. The theory underlying this equation is that the probability of giving a particular response for a stimulus is dependent on the perceptual similarity of the response to the stimulus, the overall bias to give that response, and the combination of biases and similarities for all other possible responses. Equation 1 is used here as a means of estimating perceptual similarities for individual phoneme pairs independent of response bias or the perceptual similarities of other phoneme pairs.

In the present application, an iterative procedure is used to find values of α and b coefficients that maximize the correlation between the estimated stimulus-response probabilities (i.e., calculated using Equation 1) and the obtained proportion of responses in each cell of a phoneme confusion matrix. The coefficients are fit with the constraints that the similarity between a consonant and itself is equal to 1 (i.e., $\alpha_{ii} = 1$), similarity coefficients for consonant pairs are symmetric (i.e., $\alpha_{ij} = \alpha_{ji}$)[2], and all coefficient values (α and b) are ≥ 0 and ≤ 1. An NxN consonant confusion matrix would be fit using $N(N-1)/2$ similarity coefficients and N bias coefficients.

A simple multiplicative integration rule is then used to predict audiovisual phoneme similarities based on the corresponding audio and visual phoneme similarities. The equation is

$$\alpha_{AVij} = \alpha_{Aij} \alpha_{Vij} \quad (2)$$

where α_{AVij} is the audiovisual similarity between phonemes i and j, α_{Aij} is the auditory similarity between phonemes i and j, and α_{Vij} is the visual similarity between phonemes i and j.

Finally, the predicted audiovisual phoneme similarities are used to create a predicted phoneme confusion matrix. Specifically, Equation 1 is calculated for each cell in the

[2] It is well known that consonant confusion matrices are not symmetric. These asymmetries are modeled using the bias coefficients in Equation 1.

phoneme confusion matrix, using the predicted audiovisual phoneme similarity values and setting all *b* coefficient values equal to 1 (because there is currently no method to make *a priori* predictions of bias coefficient values).

2. Reanalysis of Grant et al. (1998)

The data collected by Grant et al. (1998) was used to evaluate this method. Grant et al. collected consonant confusion data from 29 hearing-impaired listeners under audio, visual, and audiovisual conditions. The audio was presented in noise with a 0-dB signal-to-noise ratio, in both the audio and audiovisual conditions. The stimuli were 18 medial consonants presented in an /ɑ/-C-/ɑ/ context. Each subject identified each consonant 40 times within each condition.

The best-fitting similarity and bias coefficient values (i.e., the coefficient values that resulted in the highest possible Pearson correlation between the predicted and obtained proportions) were found for the audio, visual, and audiovisual confusion matrices for each subject. The percentages of variance fit by the coefficient values (i.e., the squared correlation between the predicted and obtained proportions) ranged from 97.7-99.9% (mean = 99.4%) for audio, 95.1-99.8% (mean = 98.6%) for visual, and 99.7-100.0% (mean = 99.9%) for audiovisual confusion matrices.

To evaluate the predictions of the Choice Theory method, the obtained and predicted results were compared in terms of the mean percentages correct for each subject in the audiovisual conditions (see Figure 1). The correlation between the predicted and obtained percentages correct was $r = 0.89$, which is similar to the correlations obtained by Grant et al. (1998) for predictions made using the other available mathematical methods ($r = 0.89$ for the methods described by Braida, 1991, and Blamey et al., 1989; $r = 0.83$ for a predictive form of the method described by Massaro, 1987).

Although this high correlation demonstrates that the Choice Theory method predicted much of the relative differences between subjects, the absolute levels of predicted accuracy were always higher than that obtained; the mean difference in the predicted and obtained percentages was 11.9 percentage points. Within the Choice Theory framework, there are two possible explanations for a short-fall such as this: The subjects may not have optimally combined the unimodal perceptual similarities (i.e., the perceptual integration processes were not efficient), or the subjects may have had response biases that were not optimal for this experimental task.

To control for effects of response biases in the obtained audiovisual data, new *bias-corrected* matrices were computed by Equation 1 using the α coefficients that were fit directly from the obtained audiovisual matrices and setting all *b* coefficients to 1; the biases for the predicted and bias-corrected matrices were thus equated. The correlation between the predicted and bias-corrected percentages correct was the same as for the obtained percentages correct ($r = 0.89$; see Figure 1). However, the predicted and bias-corrected percentages correct were more similar in absolute levels; the mean difference between the obtained and bias-corrected percentages was 1.6 percentage points. Therefore, much of the difference between the predictions and the original obtained responses can be attributed to response biases; it appears that these subjects as a group were integrating the available phonetic information from each modality quite efficiently.

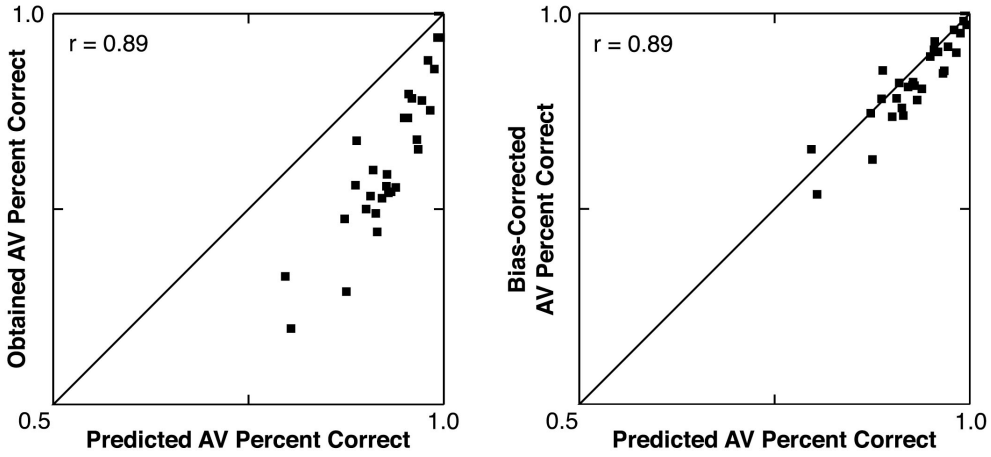

Figure 1. *Scatterplots comparing predicted, obtained, and bias-corrected percentages of correct audiovisual consonant identifications, using the data reported in Grant et al. (1998).*

It is worth noting that the predictions of the Choice Theory method were also highly correlated (r = 0.93) to those obtained by the only other optimal-processing model of audiovisual integration in the literature, Braida's (1991) pre-labeling model (see Figure 2). It has long been known that Signal Detection Theory and Choice Theory provide similar estimates of sensitivity for simple forced-choice experimental designs (e.g., Luce, 1963; Macmillan & Creelman, 1991; Treisman & Faulkner, 1985). The present results suggest that these theories can provide similar predictions of audiovisual integration performance as well. However, in absolute terms, the predicted level of performance was higher for the Choice Theory model than for Braida's pre-labeling model (mean difference = 3.4 percentage points). It is possible that this difference can be attributed to the integration rules that were used[3]

[3] If one assumes that d' values within Signal Detection Theory are equivalent to the natural logarithm of α within Choice Theory (e.g., Luce, 1963; Macmillan & Creelman, 1991), then the multiplicative integration rule used here (Equation 2) is equivalent to adding d' values from the audio and visual modalities. In contrast, Braida's (1991) pre-labeling model uses a Euclidean metric to combine d' values from the two modalities (i.e., the square root of the sum of the squared d' values). Euclidean d' predictions will always be ≤ additive d' predictions.

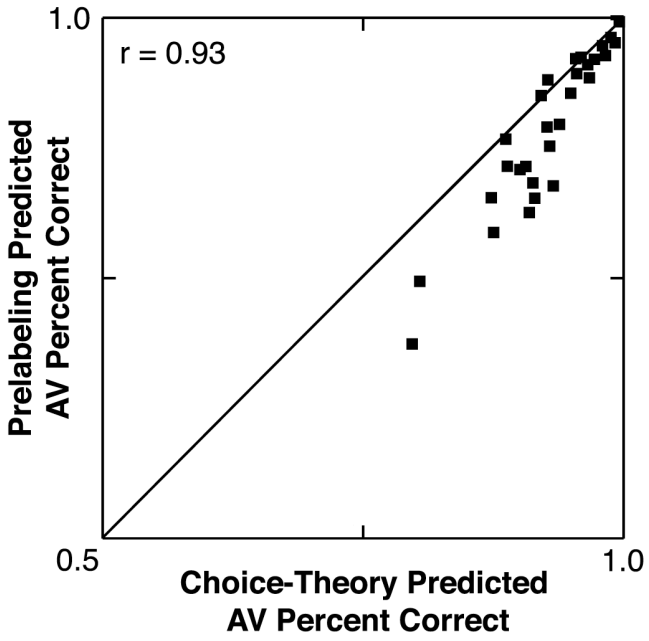

Figure 2. *Scatterplot comparing predictions of the Choice Theory method and Braida's (1991) pre-labeling model, using the data reported in Grant et al. (1998).*

3. Summary and Conclusions

The Choice Theory method appears to provide accurate predictions of audiovisual phoneme recognition. At least for Grant et al.'s (1998) data, the predictions correlate with obtained data at least as well as predictions generated by the other available methods (Blamey, 1989; Braida, 1991; Massaro, 1987). Furthermore, the Choice Theory method seems successful as a measure of optimal processing in that the predicted levels of accuracy mostly exceeded those obtained; in the few cases where the bias-corrected levels of performance exceeded those predicted, the magnitude of these differences were relatively small (a difference of 6.2 percentage points in the worst case).

Although the Choice Theory method is intended primarily as a mathematical means for predicting audiovisual phoneme intelligibility, the bias and similarity coefficients may provide insights into the underlying perceptual/cognitive processes of the subjects. Choice Theory suggests that phoneme recognition is limited by two factors: The raw perceptual/phonetic similarity of the phonemes and the categorization biases used by subjects when identifying what was perceived. Applied to Grant et al.'s (1998) data, this theory suggests that all of his subjects were efficient at integrating the unimodal perceptual information (see also Massaro & Cohen, 2000), but did not have categorization biases that were optimized to this consonant identification task. The predicted and bias-corrected percentages were close enough to question whether any

subjects in this group were experiencing perceptual integration problems at all; the small amount of subject variance could plausibly be attributed to measurement noise. In addition, the fact that their categorization biases were not optimal for this task is likely not an indicator of problems in cognitive/perceptual processing; the biases that were used here may, in fact, be optimal for recognizing other speech materials, such as words and sentences. In light of these conclusions, it is not surprising that Grant and Seitz (1998) found that differences in obtained and predicted levels of audiovisual consonant identification were not highly correlated with individual differences in other audiovisual integration tasks.

In conclusion, it is still possible that certain individuals (e.g., new recipients of cochlear implants) may have deficits in cognitive/perceptual processing that limit audiovisual phoneme recognition, and that the method described here may prove effective at diagnosing these deficits. However, the present results suggest that any individual differences in the efficiency of these cognitive/perceptual processes are likely to be very small for most subject populations, perhaps too small to be detectable using existing phoneme identification methods.

Acknowlegements
I am grateful to Ken Grant, Richard Tyler, and Andrew Faulkner for comments on this article.

References

Blamey, P.J., Cowan, R. S. C., Alcantara, J. I., Whitford, L. A., and Clark, G. M. (1989). "Speech perception using combinations of auditory, visual, and tactile information," J. Rehab. Res. Dev. 26 (1), 15-24.

Braida, L. D. (1991). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. 43A (3), 647-677.

Durlach, N. I. & Braida, L. D. (1969). "Intensity perception. I. Preliminary theory of intensity resolution." J. Acoust. Soc. Am. 46, 372-383.

Grant, K. W., Walden, B. E., and Seitz, P. F. (1998). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am. 103 (5), 2677-2690.

Grant, K. W., and Seitz, P. F. (1998). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. 104 (4), 2438-2450.

Green, D. M., & Swets, J. A. (1966). Signal Detection Theory and Psychophysics. New York: John Wiley.

Luce, R. D. (1963). "Detection and recognition" in Handbook of Mathematical Psychology (R. D. Luce, R. R. Bush, and E. Galanter, Eds.), 103-187.

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Vowel and consonant resolution: Basic sensitivity and context memory. J Acoust. Soc. Am., 84, 1262-1280.

Massaro, D. W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. (Erlbaum, Hillsdale, NJ).

Massaro, D. W., and Cohen, M. M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception, J. Acoust. Soc. Am., 108, 784-789.

Nosofsky, R. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis Percept. Psychophys., 38, 415-432.

Treisman, M., & Faulkner, A. (1985). On the choice between Choice Theory and Signal Detection Theory, Q. J. Exp. Psych., 37A, 387-405.