Ordinal models of audiovisual speech perception

Andersen, Tobias

Published in:
Proceedings of the International Symposium on Auditory and Audiological Research 2011

Publication date:
2011

Citation (APA):
Andersen, T. (2011). Ordinal models of audiovisual speech perception. In Proceedings of the International Symposium on Auditory and Audiological Research 2011

Ordinal models of audiovisual speech perception

TOBIAS S. ANDERSEN

Informatics and Mathematical Modelling, Technical University of Denmark, 2800 Lyngby, Denmark

Audiovisual information is integrated in speech perception. One manifestation of this is the McGurk illusion in which watching the articulating face alters the auditory phonetic percept. Understanding this phenomenon fully requires a computational model with predictive power. Here, we describe ordinal models that can account for the McGurk illusion. We compare this type of models to the Fuzzy Logical Model of Perception (FLMP) in which the response categories are not ordered. While the FLMP generally fit the data better than the ordinal model it also employs more free parameters in complex experiments when the number of response categories are high as it is for speech perception in general. Testing the predictive power of the models using a form of cross-validation we found that ordinal models perform better than the FLMP. Based on these findings we suggest that ordinal models generally have greater predictive power because they are constrained by a priori information about the adjacency of phonetic categories.

INTRODUCTION

Speech perception in face-to-face conversation is based not only on hearing the acoustic speech signal but also on lip-reading. Observers tend to integrate audiovisual information across the sensory modalities without being aware of it. In the natural, ecological valid situation where the voice and lip-movements are congruent this facilitates speech perception (Sumby & Pollack, 1954). When an incongruent voice is dubbed onto a video of a talking head observers may perceive a fusion type McGurk illusion in which the perceived phoneme differs both from that mediated by the voice and that mediated by the face (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). The typical example of this fusion type McGurk illusion is when a voice saying /ba/ is dubbed onto a face saying /ga/ causing observers to hear /da/. Other types of McGurk illusions include combination illusions in which the observer hears both the phoneme mediated by the voice and the phoneme mediated by the face. An example of a fusion illusion is when a voice saying /da/ is dubbed onto a face saying /ba/ which observers tend to hear as /bda/. Visual dominance illusions is another type of McGurk illusions in which observers hear the phoneme mediated by the lip-movements rather than that mediated by the voice.

Because the influence of vision on hearing in speech perception is so profound understanding how it works may give us fundamental cues to how speech perception work in general. Furthermore, understanding how vision helps audition in speech perception may enable us to develop hearing aids that help audition both when visual information is available and when it is not.

A computational account is fundamental to a good understanding of audiovisual speech perception (Braida, 1991; MacDonald & McGurk, 1978; Massaro, 1998). Part of a full computational account is the internal representation that is the basis of audiovisual integration. In other words: How is auditory and visual speech represented in the brain at the point of cross-modal integration? Additionally, we must also seek to understand the nature of the integration process itself. What is its functional form? How does it depend on the acoustic environment and the state of the observer?

One account of audiovisual integration in speech perception is the Fuzzy Logical Model of Perception (FLMP) (Massaro, 1998). The FLMP can be expressed as

$$P(R_r \mid A,V) = \frac{P(R_r \mid A)P(R_r \mid V)}{\sum_{\tilde{r}} P(R_{\tilde{r}} \mid A)P(R_{\tilde{r}} \mid V)}$$

where *P(Rr|A,V)* is the audiovisual response probability, i.e. the probability that the observer responds in the r^{th} response category given auditory, *A*, and visual, *V*, information. Likewise $P(R_r|A)$ and $P(R_r|V)$ are the unimodal response probabilities given either auditory or visual information.

The FLMP appears to be a parsimonious model with the only free parameters being the unimodal response probabilities, $P(R_r|A)$ and $P(R_r|V)$. This implies that the mechanism of integration is constant uninfluenced by the acoustic environment and the state of the observer. This does not mean, however, that audiovisual perception is uninfluenced by these factors as they can influence unimodal perception prior to integration. Such effects may then propagate through cross-modal integration to reach the integrated percept.

In the FLMP audiovisual integration occurs only after phonetic classification has occurred in vision and audition, i.e. late in the perceptual processing pathway. In a way, one might say that the model lacks a front end describing auditory speech perception and lip-reading. To overcome this problem, the FLMP is typically tested also on unimodal auditory and visual response proportions in addition to audiovisual response proportions.

The FLMP can be interpreted as an optimal model of integration under certain assumptions. To see this we can start by formulating the audiovisual response probabilities in terms of Bayes' rule:

$$P(R_r \mid \mathrm{A,V}) = \frac{P(\mathrm{A,V} \mid R_r)P(R_r)}{\sum_{\tilde{r}} P(\mathrm{A,V} \mid R_{\tilde{r}})P(R_{\tilde{r}})}$$

Assuming conditional independence and a flat prior, i.e. $P(R_{\tilde{r}}) = P(R_r)$ for all $\tilde{r}$ and r, we arrive at

$$P\left(R_r \mid \mathrm{A,V}\right) = \frac{P\left(\mathrm{A,V} \mid R_r\right)P\left(R_r\right)}{\sum_{\tilde{r}} P\left(\mathrm{A,V} \mid R_{\tilde{r}}\right)P\left(R_{\tilde{r}}\right)} = \frac{P\left(\mathrm{A} \mid R_r\right)P\left(\mathrm{V} \mid R_r\right)}{\sum_{\tilde{r}} P\left(\mathrm{A} \mid R_{\tilde{r}}\right)P\left(\mathrm{V} \mid R_{\tilde{r}}\right)}$$

Now we can re-invert the unimodal response probabilites using Bayes' rule once more. If we insert these expressions and assume flat priors on *A* and *V* we arrive at the FLMP.

In addition to being parsimonious and straightforward to interpret the FLMP has been shown to provide excellent fits to a wide array of audiovisual speech perception experiments (Massaro, 1998). These advantages has helped make the FLMP the most studied model of audiovisual integration in speech perception.

The FLMP has, however, also been subject to serious criticism, which has been directed at it not being parsimonious but rather too flexible and able to fit to almost any data sets. Schwartz provided a striking demonstration of this which he called the "0/0 problem" (Schwartz, 2006). For example, for a voice saying /pa/ dubbed onto a face saying /ka/, the observed proportion of P-responses should be close to one for audition and close to zero for vision. If ε_A and ε_V are small we can express this as

$$P\left(R_P \mid A\right) = 1 - \varepsilon_A \quad \text{and} \quad P\left(R_P V\right) = \varepsilon_V$$

If we insert these expressions into the FLMP we arrive at

$$P\left(R_r \mid \mathrm{A,V}\right) = \frac{P\left(R_r \mid \mathrm{A_a}\right)P\left(R_r \mid \mathrm{V_v}\right)}{\sum_{\tilde{r}} P\left(R_{\tilde{r}} \mid \mathrm{A_a}\right)P\left(R_{\tilde{r}} \mid \mathrm{V_v}\right)} = \frac{\left(1-\varepsilon_A\right)\varepsilon_V}{\left(1-\varepsilon_A\right)\varepsilon_V + \varepsilon_A\left(1-\varepsilon_V\right)} \approx \frac{\varepsilon_V}{\varepsilon_V + \varepsilon_A}$$

This expression can take on any value between zero and one while keeping ε_A and ε_V small. Hence, for this particular example the FLMP is certainly too flexible as it will fit any observed data. The reason why this can happen is that the FLMP is highly non-linear in the range of parameter space in which incongruent stimuli fall. The flexibility of non-linear models are not measured well solely by the number of free parameters.

In the literature on model evaluation many ways to take model flexibility into account have been suggested and many of these have been applied to testing the FLMP but with varying results (Massaro, Cohen, Campbell, & Rodriguez, 2001; Pitt, Myung, & Zhang, 2002). One of these is cross-validation, which has the advantage of being conceptually simple and generally applicable to non-linear as well as linear models because it does not rely on counting the models' free parameters. In cross validation the data are divided into test and training sets. The model is fitted to the training set and evaluated on the test set. A model that is too flexible will fit well to the training set partly because it fits not only to the underlying structure of the data but also to the sampling variability inherent to the data in a process known as over-fitting. Since the sampling variability will be different in the test set the model will generally not fit the test set well. One might say that cross-validation thus test a model's ability to predict new data not used in the fitting process. Here we will rely on a particular form of cross-validation in which the data are divided so that the test set consists of response proportions for one particular auditory, visual or audiovisual stimuli (Busemeyer & Wang, 2000; Myung, Pitt, & Kim, 2005). The training set then consists of the rest of the data. The process of splitting the data, fitting the model to the training set and testing it on the test set can be repeated so that the model is evaluated on the entire data set. This gives a cross-validation error which is directly comparable to the conventional model error.

Massaro tested the FLMP using a very similar type of cross-validation (Massaro, 1998) and found that the cross-validation error was high compared to the conventional model error. This may indicate that the FLMP over-fits because it is too flexible. If this is the case it should be possible to constrain it or to create another more constrained model that would not over-fit and hence have greater predictive power. One way to constrain the FLMP could be to give it the front-end that it lacks, i.e. incorporate information about the similarities of the speech sounds and lip-movements. This, however, amounts to developing a model of auditory speech perception as well as lip-reading, which is a formidable task, so here we will only incorporate one very simple assumption: combination response categories (e.g. "BG") must be closer to their constituent components (eg. "B" and "G") than to anything else. In order to test the validity of this assumption we conducted an audiovisual speech perception experiment in which observers were presented with P, K or T either as acoustic speech or as a silent video of the talking face. In addition, we presented observers with all the nine possible audiovisual combinations.

METHODS

The stimuli consisted of video recordings of a native Finnish speaker uttering /eke/, /epe/ or /ete/. In the unimodal visual conditions the video was silent. In the unimodal auditory conditions only the audio recording was presented. In the congruent audiovisual conditions both the audio and video recording were presented. In the incongruent audiovisual conditions, an incongruent audio track was dubbed onto the video by aligning the bursts of the stop consonants of the dubbed and original audio recording. All nine combinations of audio and video were presented. The sound level was approximately 20 dB above a noise floor of 30 dB coming from computer ventilation. All stimuli were presented 20 times. The participants were 10 native Finnish speakers who could respond with any consonant or combination of consonants. We ignore double consonants and classify B, D and G as their unvoiced counterparts P, T and K. In classifying combinations responses we ignore the order of the responses. Using these rules, less than 1% of the observers' responses did not fit into the categories K, KP, P, PT, T and TK. We ignore these responses in our analysis.

To order the response categories so that combination response categories fall between the response categories they combine we designed a model where phonetic classification is based on a cyclical internal representation. In order to do so we employ the von Mises distribution, which is similar to a normal distribution but defined on a cyclic continuum. The von Mises probability function is

$$\Phi\left(x \mid \mu, \mathrm{k}\right) = \frac{1}{2\pi}\left(x + \frac{2}{I_0\left(k\right)}\sum_{j=1}^{\infty} I_j\left(k\right)\frac{\sin\left(j\left(x-\mu\right)\right)}{j}\right) \quad , \quad \mu - \pi < x < \mu + \pi$$

In this expression, x, is the internal representation value, μ, is the mean of the distribution, k, is a parameter defining the (angular) variance and the function, I_j, denotes the incomplete Bessel function of the j^{th} order. Since this expression contains an infinite sum it is not possible to calculate exact values so we used an approximation (Hill, 1977).

Dividing the cyclic continuum into six response categories requires six category boundaries, $c_1,...,c_6$, which are free parameters. The unimodal auditory response probabilities can then be defined as

$$P\left(R_r \mid A\right) = \Phi\left(c_r \mid \mu_A, k_A\right) - \Phi\left(c_{r-1} \mid \mu_A, k_A\right) \quad , \quad 2 < r < 6$$

$$P\left(R_1 \mid A\right) = 1 - \sum_{i=2}^{6} P\left(R_i \mid A\right)$$

Similarly, response probabilities can be derived for visual and audiovisual stimuli. This defines what we shall call the *cyclical model without integration*. It needs 6 free parameters to define the response boundaries, 15 free parameters to define the distribution means, which can be different for each of the 15 stimuli, and 1 free parameter to define the model variance parameter, k, which we assume is common to all stimuli. Hence it needs 22 free parameters. We might add that this model does *not* model audiovisual integration. It only models basing phonetic classification on an underlying cyclical continuum and whether the ordering of the categories is reasonable.

We then define the *weighted cyclical model* in which audiovisual integration is modelled as a angular weighted sum of internal representation values so that

$$\mu_{av} = \tan^{-1}\left(\frac{w\cos\left(\mu_a\right) + \left(1-w\right)\cos\left(\mu_v\right)}{w\sin\left(\mu_a\right) + \left(1-w\right)\sin\left(\mu_v\right)}\right)$$

In fact, we did not use the standard arctangent function but the two-argument four-quadrant arctangent function sometimes referred to as atan2. This model does not need free parameters to define the distribution means for the 9 audiovisual stimuli but do need a free parameter to define the weighting factor, w. Hence it has 8 free parameters fewer than the cyclical model without integration, i.e. 14.

Alternatively, audiovisual integration can be based on the FLMP after the unimodal response probabilities have been derived from a cyclical continuum. We call this model the *hybrid FLMP/cyclical model*. Since it does not include a weighting factor it has 1 free parameter less than the weighted cyclical model, i.e. 13.

Finally, the audiovisual integration could be based on the conventional FLMP, which needs 30 free parameters to define the 5 independent response probabilities for each of 3 auditory and 3 visual stimuli.

For each model we found the across subject average of the root mean squared error (RMSE) using the nonlinear least squares minimization in the Matlab^{TM} optimization toolbox. To ensure that optimization had not stranded in a local minimum we ran the optimization routine 100 times with random initial conditions. To test the predictive power of the models we conducted cross-validation of the integration models, i.e. the weighted cyclical model, the FLMP and the hybrid FLMP/cyclic model, by leaving out observers' responses to one stimulus from the fit and using those responses left out to compute the cross-validation error again fitting the models 100 times. By repeating this for all 15 stimuli and summing over the error we calculated the cross-validation RMSE, which is directly comparable to the RMSE.

RESULTS AND DISCUSSION

The mean RMSE of the cyclical model without integration was 0.018. This is a low number and indicates that it is possible to base phonetic classification on a cyclic continuum. This result is not trivial. The cyclic model only works if observers use only neighbouring response categories for a given stimulus. Also note that even though this model does not model audiovisual integration it still employs 8 free parameters less than the FLMP.

Having established that using the cyclic continuum is reasonable we proceed to ask the question of which mechanism of integration that seems to describe the data better. The RMSE of the weighted cyclical model was 0.034. The corresponding cross-validation RMSE was 0.23. The RMSE of the hybrid model was 0.023 and the corresponding cross-validation RMSE was 0.23. Hence these two models have very similar cross-validation RMSE and thereby very similar predictive power. This finding is thus not decisive on which mechanism of integration that underlies audiovisual integration of speech. One might say that since the hybrid model has one free parameter less than the weighted model it is the most parsimonious account of audiovisual integration in speech perception.

The FLMP was the model with the lowest RMSE of 0.009. This is not surprising as it was also the model employing the highest number of free parameters. The FLMP was also the model with the highest cross-validation RMSE of 0.30. This indicates that the good fit was due to over-fitting allowed by the high number of free parameters.

In summary, our findings show that constraining models of audiovisual integration in speech perception can help us find models with greater predictive power. Even though the cyclic continuum we employed is probably only a crude approximation of the continuum that the brain truly employs it did constrain the models in a meaningful way that increased the models' predictive power. In this study, we distinguish between the problem of determining the internal representation from the problem of determining the mechanism of integration. We tested two different mechanisms of integration. One is the weighted sum of the cyclic internal representation, the other is the FLMP applied to the response probabilities derived from the cyclic model. Here the FLMP actually fared better indicating that it might be a truer description of audiovisual integration in speech perception although this is not apparent when testing it in its conventional form.

It is our hope that testing quantitative models of audiovisual speech perception such as those described here may help us in shedding light on the true mechanisms of audiovisual integration in speech perception. We believe that the current study show that it may be important to distinguish this problem from the problem of determining the underlying internal representation of speech.

REFERENCES

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

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*(1), 171-189.

Hill, G. (1977). Algorithm 518: Incomplete Bessel Function I0: The Von Mises Distribution. *ACM Transactions on Mathematical Software, 3*(3), 279-284.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Percept Psychophys, 24*, 253-257.

Massaro, D. W. (1998). *Perceiving talking faces*. Cambridge, Massachusetts: MIT Press.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychon Bull Rev, 8*(1), 1-17.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model Evaluation, Testing and Selection. In K. Lamberts & R. L. Goldstone (Eds.), *The Handbook of Cognition*. London: Sage Publications.

Pitt, M. A., Myung, I. J., & Zhang, S. B. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3), 472-491.

Schwartz, J. L. (2006). The 0/0 problem in the fuzzy-logical model of perception. *J Acoust Soc Am, 120*(4), 1795-1798.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215.