Cognition and the Statistics of Natural Signals

Javier R. Movellan and **George Chadderdon**
Departments of Cognitive Science and Computer Science
University of California, San Diego
La Jolla, CA 92093
movellan@cogsci.ucsd.edu and gchadder@netman.orincon.com

Abstract

This paper illustrates how the statistical structure of natural signals may help understand cognitive phenomena. We focus on a regularity found in audio visual speech perception. Experiments by Massaro and colleagues consistently show that optic and acoustic speech signals have separable influences on perception. From a Bayesian point of view this regularity reflects a perceptual system that treats optic and acoustic speech as if they were conditionally independent signals. In this paper we perform a statistical analysis of a database of audiovisual speech to check whether optic and acoustic speech signals are indeed conditionally independent. If so, the regularities found by Massaro and colleagues could be seen as an optimal processing strategy of the perceptual system. We analyze a small database of audio visual speech using hidden Markov models, the most successful models in automatic speech recognition. The results suggest that acoustic and optic speech signals are indeed conditionally independent and that therefore, the separability found by Massaro and colleagues may be explained in terms of optimal perceptual processing: Independent processing of optic and acoustic speech results in no significant loss of information.

Introduction

This paper illustrates how the analysis of the statistical structure of natural signals may provide a rational basis for understanding human cognition. This approach is not new, and in our case it was inspired on David Marr's ideas about the importance of a functional level of analysis, John Anderson's views on rational analysis, and by David Field's work relating early visual processing to the statistics of natural images (Marr, 1982; Field, 1987; Anderson, 1990). In this paper we analyze a regularity found in a wide variety of experiments on audiovisual speech perception.

Research on audiovisual speech perception shows that visual signals modulate the perception of auditory signals. For example, McGurk and MacDonald (McGurk & MacDonald, 1976), showed that when subjects hear "ba" while seeing "ga", they perceive "da", a percept which is jointly influenced by the optic and the acoustic speech signals. Extensive research has been done to understand how optic and acoustic speech signals combine into a unified percept (Massaro & Cohen, 1983; Massaro, 1987; Braida, 1991). For concreteness, consider the following hypothetical experiment, which illustrates a common design in this area of research. Subjects are repeatedly presented with 9 opto-acoustic speech signals obtained by combining, in a fully factorial design, the acoustic articulations */ba/*, */ga/*, and */da/* with optic articulations of the same alternatives. Subjects are then presented with these signals and asked to report what they heard. The responses are then organized into a stimulus-response matrix in which each entry indicates the probability of a particular perceptual response when the subject is presented with one of the 9 possible signal combinations. Let $\{\omega_1, \omega_2, \cdots, \omega_n\}$, represent the response alternatives, ξ^o the optic signal and ξ^a the acoustic signals. Massaro and colleagues (Massaro & Cohen, 1983; Massaro, 1987) have repeatedly shown that in a wide variety of experiments of this type, response probability ratios factorize into independent components one controlled by the acoustic signal and one by the optic signal.

$$\frac{p_r(\omega_i|\xi^o\xi^a)}{p_r(\omega_j|\xi^o\xi^a)} = \left(\frac{F_o(\xi^o,\omega_i)}{F_o(\xi^o,\omega_j)}\right)\left(\frac{F_a(\xi^a,\omega_i)}{F_a(\xi^a,\omega_j)}\right) \qquad (1)$$

where $p_r(\omega_i|\xi^o_k\xi^a_l)$ is the probability of subjects choosing response alternative ω_i when presented with the optic signal ξ^o synchronized with the acoustic signal ξ^a. The term

$$\frac{F_o(\xi^o,\omega_i)}{F_o(\xi^o,\omega_j)}\,, \qquad (2)$$

is interpreted as the relative support of the optic signal ξ^o for the two response alternatives under consideration, and the term

$$\frac{F_a(\xi^a,\omega_i)}{F_a(\xi^a,\omega_j)}\,, \qquad (3)$$

is interpreted as the relative support of the acoustic signal ξ^a for the two response alternatives under consideration.

The crucial aspect of this result is that response probabilities ratios are separable into independent factors. This type of factorization was first noticed by Morton (Morton, 1969) and thus it is at times recognized as *Morton's law*. Movellan and McClelland (Movellan & McClelland, 1995 submitted for publication) showed that Morton's law is the signature of a perceptual system that processes signals as if they were conditionally independent. If the acoustic and optic speech signals were indeed conditionally independent, Morton's law would reflect an optimal processing strategy of multimodal speech.

To investigate this point, we analyze the statistical structure of a small database of audio-visual speech signals. Our goal is to test whether naturally occurring acoustic and visual speech signals are conditionally independent.

At a formal level, conditional independence is defined as follows,

$$p(\xi^o\xi^a|\omega_j) = p(\xi^o|\omega_j)p(\xi^a|\omega_j) \quad (4)$$

indicating that the likelihood of each perceptual alternative ω_j, is separable. Intuitively, conditional independence tells us that if we analyze signals belonging to a perceptual category ω_j, we will find that the acoustic and optic signals within that group are statistically independent. From a Bayesian point of view the likelihood is the only source of data-driven information about the perceptual alternatives and thus, conditional independence allows separable processing of the optic and acoustic signals.

Due to the large dimensionality of the opto-acoustic signals we analyze their statistical structure in an indirect manner, by modeling the speech signal using hidden Markov models (HMM), the most successful models for automatic speech recognition. We train HMMs to recognize audiovisual speech. Some of these models are constrained to assume conditional independence some are not. The constrained models are a restricted version of the unconstrained models. We then optimize the entire family of constrained and unconstrained models. If the audio and visual speech signals are conditionally independent, the best constrained model should perform about as well as the best unconstrained model. Otherwise, the best unconstrained models should outperform the best constrained models.

Database

We used Tulips1, a database compiled by Movellan (Movellan, 1995) and consisting of 9 male and 3 female undergraduate students from the Cognitive Science Department at the University of California, San Diego. For each of these, two samples were taken for each of the digits "one" through "four". Thus, the total database consists of 96 digit utterances. The audio sampling rate is 11.1 kHz, and each sample has an 8-bit representation. Each frame in the video track of a movie is an 8-bit grey-scale, 100x75 pixel image, and each movie is sampled at a visual frame rate of 30 frames per second. The subjects were asked to center and align their lips in the camera during the sampling.

Signal processing.

Our signal processing philosophy is to preserve, as much as possible, the information in the original frames. Each frame from the video track is symmetrized along the vertical axis, and a temporal difference frame is then obtained by subtracting the previous symmetrized frame from the current symmetrized frame. The symmetrized and differential symmetrized frames are then low-pass filtered and soft-thresholded (Movellan, 1995), and the left side of the former and the right side of the latter

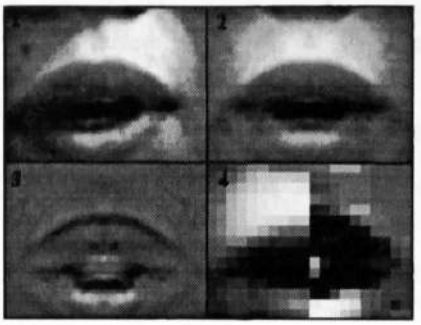

Figure 1: Visual Preprocessing: 1) Raw Image. 2) Symmetrized Image. 3) Difference Image. 4) Final Composite.

are combined to form the final feature frame. Each of these final frames has 300 dimensions (20x15 pixels). No hard feature detection procedures are used to avoid loss of potentially important information. The approach is illustrated in Figure 1.

LPC/cepstral analysis is used for the auditory front-end. This is a fairly standard technique which parameterizes an estimate of the human vocal tract's transfer function. First, the auditory signal is passed through a first-order emphasizer to spectrally flatten it. Then the signal is separated into non-overlapping frames at 30 frames per second. This is done so that there are an equal number of visual and auditory feature vectors for each utterance, and these will be in synch with each other. On each frame we perform the standard LPC/cepstral analysis. Each 30 msec auditory frame is characterized by 26 features: 12 cepstral coefficients, 12 delta-cepstrals, 1 log-power, and 1 delta-log-power. Each of the 26 features is encoded with 8-bit accuracy. The cepstral coefficients are a compact representation of the local power spectrum of the speech signal. The local phase spectrum is lost in this representation. However, loosing local phase does not affect the intelligibility of the acoustic signal.

Statistical modeling of the speech signal.

We model the speech signal using hidden Markov models (HMMs), one per word category, independently trained on signals from the corresponding word categories (see Figure 2). The HMMs were continuous density left-to-right models with a fixed number of states. The probability distribution generated by a state is modeled as a mixture of multivariate Gaussian distributions. A diagonal covariance matrix is used, with the variances of each Gaussian in a particular state tied together. The number of states and number of Gaussian mixtures per state were systematically varied to find the best combination of states and mixtures.

After each model is trained on exemplars from its corresponding word category, classification of an unknown observation proceeds by calculating, for each model, the log-likelihood of the model given the observation. Then the classification corresponding to the model with the highest log-likelihood is chosen as the winner.

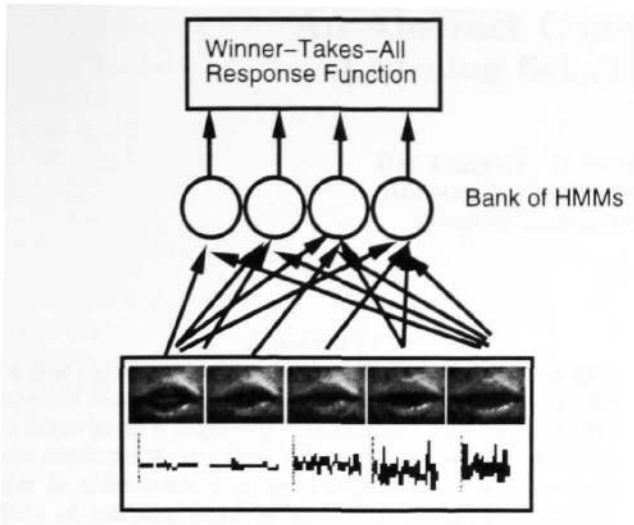

Figure 2: The signal is modeled with a bank of HMMs.

We use two types of models, which we will refer to as *constrained* and *unconstrained*. The constrained models implement the assumption of conditional independence. Each constrained model consist of two independent sub-models, one auditory and one visual. These are trained on their respective data sets, and during testing, the combined response is obtained by adding the log-likelihood of the auditory and visual models given the observations. This operation enforces the assumption of conditional independence. Classification proceeds by picking the model with the highest combined log-likelihood.

The unconstrained models are trained on the combined opto-acoustic signal. These models are more general than the unconstrained models. If the signals are conditionally independent these models should learn to combine their outputs additively thus performing as well as the previous models. However, if the signals are not conditionally independent these models should perform better.

Results

We tested the two types of models with a variety of signal-to-noise ratios (SNR) in the acoustic signal. Training was done with clean auditory samples, and testing with a variable (SNR). No noise was added to the images either for training or testing. The jack-knife method was used for obtaining each generalization performance estimate. Training was done leaving out the utterances of one of the 12 subjects, and testing was done on the utterances of the excluded subject. This was repeated 12 times, leaving out a different subject each time. Jack-knife estimates are based on the average generalization obtained with these 12 samples.

For each of the two types of models (constrained and unconstrained), we systematically tested 45 different architectures by varying the number of states (2,3,4,5,6) and the number of Gaussians per state (2,3,4,5,6,7,8,9,10). We chose the best 2 constrained and the best 2 unconstrained architectures.

	Acoustic Signal to Noise Ratio				
Model	0 dB	6 dB	12 dB	18 dB	Clean
Unconstrained	94.2	95.2	94.5	93.9	95.5
Constrained	92	95.2	96.4	98.7	98
Auditory Only	75.4	84.6	89.1	92.6	92.3
Visual Only	89.4	89.4	89.4	89.4	89.4

Table 1: Performance at different signal to noise ratios.

Table 1 shows the performance of the best constrained and unconstrained models. For completeness we also show the performance of the best Auditory-only and the best Visual-only models. In most cases the constrained architecture performs marginally better than the unconstrained architecture. Thus, assuming conditional independence does not result in loss of information,

Conclusions

The results of this exploratory study suggest that the optic and acoustic speech signals are indeed conditionally independent. Thus, the emergence of Morton's law in audiovisual speech perception experiments, may reflect an optimal functional organization of the perceptual system.

We need to be cautious about these results since our analysis has important limitations: 1) Our work is based on a small database and it is unclear whether it would generalize to other databases. 2) Since our database is small, it is possible that the potential bias introduced by the assumption of conditional independence may be compensated by the fact that it allows a significant reduction in the number of training parameters. 3) We model the speech signal using HMMs and it is possible, that different approaches would have produced different results. 4) Different results may perhaps be obtained using different signal processing strategies (e.g. feature detectors, Gabor filters ...) 5) Our results do not clarify at which level independence holds. It is possible that the independence obtained when conditioning on words is due to the existence of lower level independence (e.g. when conditioning over sub-word units). Our results can only be used as evidence that independence holds at some level but we cannot specify where this level is located.

We are currently working to overcome these limitations but we believe this exploratory work illustrates how current techniques on artificial pattern recognition may be used to analyze the structure of natural signals and to establish a statistical approach to human cognition.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Earlbaum.

Braida, L. (1991). Crossmodal integration and the identification of consonant segments. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 43(3)*, 647–677.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical

cells. *Journal of the Optical Society of America, 4*, 2379–2394.

Marr, D. (1982). *Vision*. New York: Freeman.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological research*. Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Cohen, M. M. (1983). Integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 753–771.

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

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76*, 165–178.

Movellan, J. R. (1995). Visual speech recognition with stochastic neural networks. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems*. Cambridge,Massacusetts: MIT Press.

Movellan, J. R., & McClelland, J. L. (1995, submitted for publication). *Stochastic interactive processing, channel separability and optimal perceptual inference: an examination of morton's law* (Technical Report PDP.CNS.95.4, Available at http://cogsci.ucsd.edu/~movellan/publications.html). Carnegie Mellon University.