References

Boring, R.L. (2004). *Cognition and Psychological Scaling: Model, Method, and Application of Constrained Scaling.* PhD dissertation, Institute of Cognitive Science, Carleton University, Ottawa, Canada.

Stevens, S.S. (1975). *Psychophysics. Introduction to its perceptual, neural, and social prospects.* New York: Wiley.

Ward, L.M. (1992). Who knows? In G. Borg & G. Neely (Eds.), *Fechner day 92. Proceedings of the eighth annual meeting of the international society for psychophysics* (pp. 217-222). Stockholm, Sweden: International Society for Psychophysics.

West, R.L. (1996). Constrained scaling: Models, methods, and response bias. In S.C. Masin (Ed.), *Fechner day 96: Proceedings of the twelfth annual meeting of the international society for psychophysics* (pp. 423-427). Padua, Italy: International Society for Psychophysics.

West, R.L., & Ward, L.M. (1994). Constrained scaling. In L.M. Ward (Ed.), *Fechner day 94. Proceedings of the tenth annual meeting of the international society for psychophysics* (pp. 225-230). Vancouver, Canada: International Society for Psychophysics.

West, R.L., & Ward, L.M. (1998). The value of money: Constrained scaling and individual differences. In S. Grondin & Y. Lacouture (Eds.), *Fechner day 98. Proceedings of the fourteenth annual meeting of the international society for psychophysics* (pp. 377-380). Quebec City: International Society for Psychophysics.

West, R.L., Ward, L.M., & Khosla, R. (2000). Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias. *Perception and Psychophysics*, *62*, 137-151.

VISUAL ENHANCEMENT OF AUDITORY DETECTION: A THEORETICAL MODEL

Cheng Qian and Willy Wong

Edward Rogers Sr. Department of Electrical and Computer Engineering and Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Canada

Abstract

Audio-visual enhancement is the phenomenon whereby a visual signal can enhance the perception of an auditory signal. This effect has been commonly explored in high-level processes like speech communication yet some aspect of this enhancement can be shown to arise from early sensory processes. Our previous work has shown that the auditory detection threshold of a sinusoidally amplitude modulated tone in quiet is reduced by an average of 2.1 dB when a concurrent, co-modulated visual signal is presented. We report here that the addition of noise does not appear to affect enhancement (average 2.2 dB shift) despite a common notion that noise increases the relative enhancement. We introduce a signal detection model which seeks to quantify the benefit of a co-modulated visual signal. This model is based on the concept of 'matched filters' and can account for the improved detection and spread of the psychometric function observed in experiment, as well as shed some light on the roles of synchrony and modulation frequency in cross-modal enhancement.

In human perception, many modalities contribute to create the full entire sensory experience. How the various sensory modalities interact with each other is known as cross-modal processing. One example of cross-modal interaction/enhancement is in audiovisual speech. The perception of auditory speech can be enhanced or modified through the presentation of a concurrent visual stimulus such as a talking face. 'Speech-reading', as it is commonly referred to, has long been explored at the behavioural level and is thought to be a higher-order process. This is based on the belief that the different sensory modalities, such as audition and vision, are first processed individually and only interact in the later stages of sensory processing. More recently however, cross-modal effects have been repeatedly identified in early sensory processing areas such as the primary sensory cortices (Ghazanfar and Schroeder, 2006). Thus it is entirely possible that some of the benefits of AV interaction in speech or otherwise are, at least partially, derived from early sensory processing. We took inspiration from this point to develop a low-level model to account for the visual enhancement of auditory detection.

The paradigm we are studying involves very simple stimuli. The experimental task is to detect amplitude-modulated pure tones in noise, with or without accompanying visual stimulation. Both the visual and the auditory signals are modulated (i.e. *co-modulated*) by an identical mathematical function thereby allowing us to explore how a visual 'code' or 'imprint' can help pick-up an identically coded sub-threshold auditory signal. While simple in nature, the co-modaulted signals form the atomistic basis to probe more complex audiovisual phenomenon like speech.

In our previous work (Sheena and Wong, 2007), we explored the effect of a co-modulated visual signal on the detection performance of participants in a quiet background. We found that addition of the visual stimulus reduced the average detection threshold by 2.1dB. In the present work, we measure the level of enhancement with the auditory signal embedded in additive white noise. To help understand the experimental results, we also propose a low-level model of audio-visual enhancement based on the theory of signal detectability.

Model

This model seeks to explain how an added visual cue can aid in the detection of an acoustic signal in noise. The target acoustic signal is a pure tone (carrier) multiplied by a sinusoidal (modulation) function, resulting in what is known as a *sinusoidally amplitude modulated tone*. The carrier frequency and the modulation frequency are independent parameters and were taken to be 1kHz and 3Hz respectively. Our model is based on the task whereby a participant is asked to detect the sinusoidally amplituded modulated target tone embedded in an additive noise background. The central premise of our model is that the co-modulated visual stimulus provides information of when the auditory target signal can be expected (Figure 1). By multiplying the auditory signal with the 'expectation function' (i.e. the visual envelope), the auditory stimulus is, in essence, boosted when the signal is present and attenuated when there is only noise. From Figure 1 we can easily intuit that detection performance is enhanced by such a manner of operation. This method of signal detection/enhancement is commonly employed in communications engineering. It can be shown that if a known signal is embedded in additive white noise, the optimal method of detecting the target is to employ what is known as a "matched filter" (Kay, 1998). Our hypothesis therefore is that cross-modal enhancement takes place via a matched filter-like process. Neural mechanisms that would support such a detection procedure are discussed later in this article.

Next, we provide an analytical demonstration that a matched filter process can indeed improve the auditory signal-to-noise ratio, thereby enhancing detection performance when an accompanying co-modulated visual signal is presented.

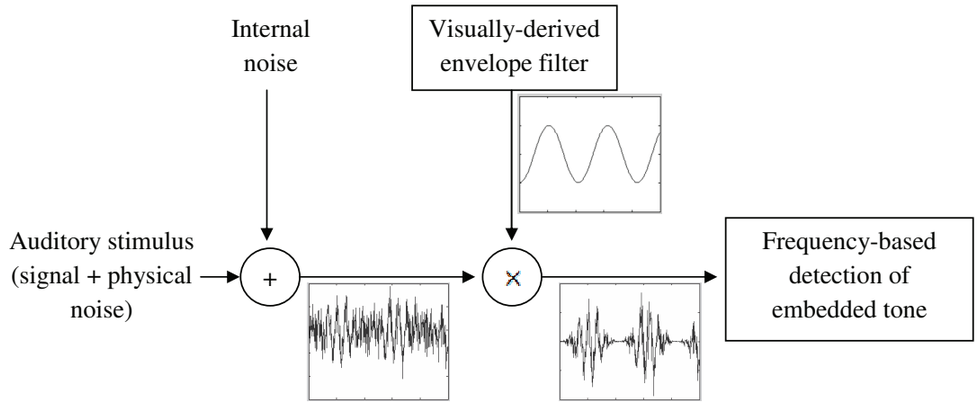

Figure 1. Block representation of model for AV enhancement. Input auditory stimulus is an internal representation to which internal noise is added. If available and congruent, visual information is multiplied in as an envelope filter. Detection occurs in the frequency domain.

Analytical Justification

Consider a sinusoially amplitude-modulated pure tone $x(t)$ added to Gaussian white noise $n(t)$:

$$y(t) = x(t) + n(t) \tag{1}$$

$$x(t) = [\frac{1}{2} + \frac{1}{2}\cos(\omega_m t)]\sin(\omega_c t) \tag{2}$$

$$n(t) \sim N(0, \sigma^2) \tag{3}$$

where ω_m is the modulation frequency and ω_c is the carrier (tone) frequency. We will take the decision statistic to be the energy to noise ratio (ENR) of the stimulus at the carrier frequency ω_c. That is, if the signal energy at the carrier frequency exceeds the noise energy by a set factor, we assume that the signal can be detected by the subject. Defining the Fourier transform of $f(t)$ as $\bar{f}(\omega)$, the detection statistic can be written as

$$\left.\frac{E}{N}\right|_{\omega=\omega_c} = \frac{|\bar{y}(\omega_c)|^2}{|\bar{n}(\omega_c)|^2} = \frac{|\bar{n}(\omega_c) + \bar{x}(\omega_c)|^2}{|\bar{n}(\omega_c)|^2} = 1 + \frac{|\bar{x}(\omega_c)|^2}{|\bar{n}(\omega_c)|^2} \tag{4}$$

with the last equality holding when (4) is averaged across different experimental trials. The ratio $|\bar{x}(\omega_c)|^2/|\bar{n}(\omega_c)|^2$ can be thought of as the physical signal to noise ratio (SNR).

Next we consider the role of the visual signal through its envelope function $h(t)$. Under ideal conditions when video and audio are fully correlated, the envelope filter will take the same form as the envelope of $x(t)$. That is, $h(t) = ½ + ½ \cos(\omega_m t)$. However, we also wish to explore the possibility that the visual signal may be phase-shifted relative to the auditory signal in which case we take $h(t)$ to be

$$h(t) = [\frac{1}{2} + \frac{1}{2}\cos(\omega_m t + \theta)] \tag{5}$$

By the matched filter procedure, $y(t)$ is now multiplied with $h(t)$, and ENR is again evaluated at ω_c. Recalling that multiplication in the time domain is equivalent to convolution in the frequency domain, we obtain

$$\left.\frac{E'}{N'}\right|_{\omega=\omega_c} = 1 + \frac{|\bar{h}\otimes\bar{x}(\omega_c)|^2}{|\bar{h}\otimes\bar{n}(\omega_c)|^2} = 1 + AVgain\frac{|\bar{x}(\omega_c)|^2}{|\bar{n}(\omega_c)|^2} \tag{6}$$

where $\otimes$ denotes the convolution operator. AVgain is a constant that quantifies the benefit of the envelope-filter mechanism. To determine this constant, some further analyses are required.

The fundamental difference between x and n is that one is a coherent signal and the other (being white noise) is not. We make use of this to derive an expression for AVgain that has some interesting properties. In the frequency domain, the signal x consists of a carrier peak and two sidebands. When the envelope-filter is applied to x, the sidebands themselves are split and some of the energy is moved back to the carrier frequency. They sum coherently since the signal is deterministic. Next we consider the effect of the envelope on white noise. Since white noise is statistically independent across frequencies, the energy moved to the carrier frequency from the surrounding regions by the envelop-filter does not add coherently. Thus the expected energy of the noise at the carrier frequency sums to a value which is disproportionately lower that that of x.

The derived mathematical expressions can be shown to be

$$|\bar{h}\otimes\bar{x}(\omega_c)|^2/2\pi = \frac{1}{4}\left(1 + \cos\theta + \frac{1}{4}\cos^2\theta\right)|\bar{x}(\omega_c)|^2 \tag{7}$$

$$|\bar{h}\otimes\bar{n}(\omega_c)|^2/2\pi = \frac{3}{8}|\bar{n}(\omega_c)|^2 \tag{8}$$

From this we write the AVgain:

$$AVgain = \frac{|\bar{h}\otimes\bar{x}(\omega_c)|^2}{|\bar{h}\otimes\bar{n}(\omega_c)|^2} \Big/ \frac{|\bar{x}(\omega_c)|^2}{|\bar{n}(\omega_c)|^2} = \frac{2}{3}\left(1+\cos\theta+\frac{1}{4}cos^2\theta\right) \quad (9)$$

That is, the enhancement due to envelope-filtering is dependent on the phase difference. There is a positive benefit which decreases slowly for small amounts of phase asynchronies, more rapidly at higher levels. The maximal theoretical benefit is attained when the two are perfectly aligned ($\theta = 0$). AVgain then takes on a value of 1.8dB. This implies that through matched filtering, we can expect an equivalent gain in the SNR of 1.8dB due to cross-modal enhancement.

Method

To test our ideas, we carried an experiment in audiovisual enhancement. We followed the same 2 alternative forced choice (2AFC) paradigm used previously (Luu and Wong, 2007) to explore how noise affects audiovisual enhancement.

5 participants (3 male, 2 female) were tested with a 2AFC paradigm on a PC platform in a double-walled sound chamber. The acoustic stimuli used were 2s duration 1kHz pure tones modulated by a 3Hz sinusoidal envelope, embedded in a 2kHz low pass filtered white noise at 75dBSPL. White noise was ramped with a 50ms risetime and the sinusoidally amplitude modulated tone followed 500ms afterwards. The entire stimulus duration was 2.5s. The visual stimuli were presented on a CRT computer monitor, and consisted of either a small white fixation cross (for the audio only condition) or a white Gaussian blob that is co-modulated with the acoustic signal (for the AV condition).

Each participant carried out 6 runs of the experiment, 3 runs for AV and 3 runs for audio-only. In each run, there were 100 trials which were randomly roved over 5 different signal levels spanning a range of 18dB. In each trial, there were two intervals, one which had noise only was presented and another where the target tone was presented. In both intervals, the same video was displayed. In total, 60 measurements per participant taken at each signal level for each condition.

Results

The data were fitted to a logistic function, $f(x) = 0.5\left(1/\left[1+e^{-\frac{x-\theta}{\sigma}}\right]\right)+0.5$ where θ is the threshold defined by the 75% detection, and σ represents the spread of the psychometric curve. A threshold reduction of 2.2dB and a spread increase of 0.95 were observed for the AV condition over the A condition (Figure 2a). The threshold shift is essentially the same as the previously observed for experiments in quiet (2.1 dB, Luu and Wong, 2007).

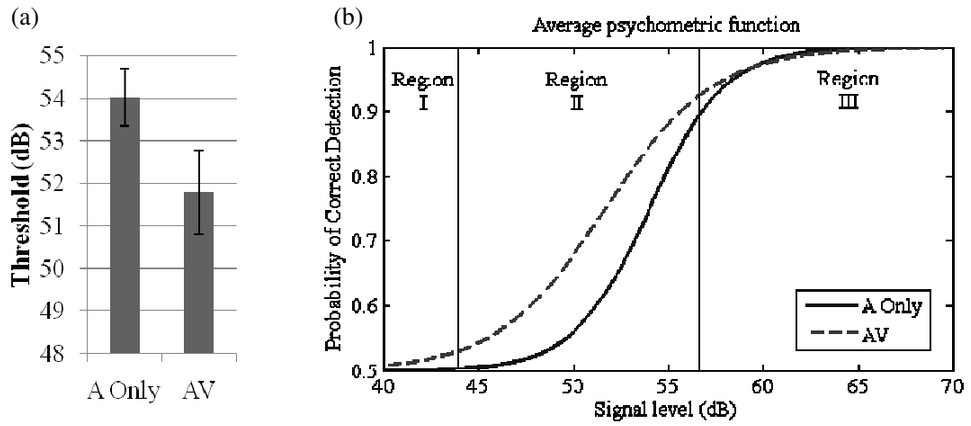

Figure 2. (a) Mean detection thresholds (75% correct point after fitting) for A Only and AV (b) Plot of mean psychometric curves for A Only and AV using mean of threshold and spread parameters from logistic function.

Discussion

No significant difference was observed in the enhancement between noisy and quiet test conditions. This is a surprising result given what has been reported in other studies. Typically it has been found that cross-modal enhancement is most effective when the target signal is least detectable. This effect has been coined the 'principle of inverse effectiveness' and was originally discovered in the neurons of the superior colliculus which are known to code audiovisual events (Meredith and Stein 1986). The lack of difference in enhancement between quiet and noisy conditions can however be understood from the context of our model. Originally we had introduced $n(t)$ to be purely external noise. However to be exact, $n(t)$ is actually a sum of both external and internal noise (sensory or otherwise). AVgain was derived to be 1.8dB irrespective of the nature of $n(t)$ indicating that the matching envelope-filter had the same effect both in quiet and in noisy conditions. By the model, AV enhancement would remain unchanged. The numerical prediction (1.8dB) also seems to correlate well with the observed enhancements in the noisy (2.2dB) and quiet (2.1dB) test conditions.

The model also makes specific predictions regarding synchrony dependence of AV enhancement. If the envelope filter does not line up with the envelope of signal, enhance of SNR will suffer. This is evident in the final mathematical expression of AVgain, which falls off slowly for small θ but much more rapidly as θ increases. AVgain reduces to unity with a phase difference of $\theta = \pi/2$. As the SNR can be degraded further at even larger phase differences, there is likely some cognitive mechanism which will ignore the influence of the visual information when the asynchrony is too large, i.e. the brain assumes that the video and audio streams are unrelated. However, we do know empirically that there is some ability for the brain to time-shift information in one stream relative to another in order to preserve a coherent perceptual experience. One example would be watching a movie where there is a constant lag of the sound relative to the image or vice versa. Within limits, we are still able to process the two information streams together. Mechanisms which underlie synchrony perception and perceptual grouping are likely to be linked. In fact, recent work showing that the brain can recalibrate synchrony between the visual and the auditory streams (Fujisaki et al, 2004) supports this notion quite well.

It is also clear from our model that enhancement should be a function of modulation frequency. For a constant phase difference θ, time shift is related to the phase difference by $\Delta t = \theta / 2\pi f_m$ where f_m is the modulating frequency. When the modulation frequency increases, there is a corresponding decrease in tolerance of the AVgain to cross-modal asynchrony. By this argument, it is clear that at higher modulation frequencies the visual information requires more and more precise synchronization to effectively improve auditory processing. Thus there must be a frequency limit to the effect of visual enhancement.

Thus far little has been mentioned regarding the biological plausibility of the model presented here for audiovisual enhancement. Although this is not an issue that can be resolved definitely here, we do mention in passing that there are a number of ways in which the system in Figure 1 could have evolved as part of the neural processing mechanism. Here is one example: given that neurons tend to encode intensity through a logarithmic transformation, two neurons – one from the visual stream and the other from the auditory stream – both feed to a third neuron which encodes audiovisual information. To find the response of the third neuron, we add the outputs of the two input neurons. This is equivalent to multiplying the two input signals before taking the logarithm (i.e. log a + log b = log ab). While this argument is primitive at best, it does provide a starting point for further experimental studies.

In future work, we wish to refine our model and to make it made more physiologically accurate in several ways. One change would be the inclusion of auditory filters as part of the calculation of the detection statistic. Currently we have calculated enhancement at the carrier frequency only, but it is clear that detection occurs across a band of frequencies. Further experiments are also planned to explore the effect of reduced correlation between auditory and visual streams, and to use complex envelopes that contain multiple modulation frequencies. Such experiments would allow us to test our model more rigorously and to generalize to more complex audiovisual stimuli.

Acknowledgements

This work was sponsored by Defence Research and Development Canada and CIHR.

References

Erber, N. P. (1975). Auditory-visual Perception in Speech. *Journal of speech and hearing disorders. 40*: 481-492.

Fujisaki, W. (1), et al. (2004). Recalibration of Audiovisual Simultaneity. *Nature neuroscience 7.7*: 773-8.

Ghazanfar, A. A., & C. E. Schroeder (2006). Is Neocortex Essentially Multisensory? *Trends in cognitive sciences 10.6*: 278-85.

Kay, Steven M. (1998). *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. New Jersey: Prentice Hall.

Luu, S. and Wong, W. (2007). Crossmodal Enhancement: Investigating the Role of Timing in Audio-visual Enhancement. *Proceedings of ISP Fechner Day 2007*.

Meredith, M. A., and B. E. Stein (1986). Visual, Auditory, and Somatosensory Convergence on Cells in Superior Colliculus Results in Multisensory Integration. *Journal of neurophysiology 56.3*: 640-62.

SCALING CONFIDENCE CATEGORIES: EQUAL SPACING?

William M. Petrusic[1], Joel A. Lucas[1] & Joseph V. Baranski[2]

1. Department of Psychology, Carleton University, Ottawa, Ontario, Canada
2. Collaborative Performance and Learning Section, Defence Research and Development Canada – Toronto, Toronto, Ontario, Canada

Abstract

On each trial in a psychophysical comparison experiment, participants used the confidence categories "50", "60", "70", "80", "90", and "100 to indicate how certain they were that they had made a correct decision. We applied Case D of Torgerson's (1958) Law of Categorical Judgment (LCJ) to estimate the mean locations of the confidence category boundaries. Scale values for the confidence category boundaries were equally spaced on the underlying subjective probability scale and were identical in the speed and the accuracy stress conditions of the experiment. Excellent goodness of fit of the LCJ to the data was obtained in each condition.

Invariably when confidence ratings are taken as subjective probabilities, in studies examining how closely the confidence ratings correspond to the actual accuracy of the decisions rendered (see Baranski & Petrusic, 1994 for the definitional formula for the calibration index), it is tacitly assumed the objective confidence category labels represent equally spaced underlying numerical probabilities. It is not at all clear that this is the case.

Whether the actual confidence category labels can be taken at face value or not is an issue of some importance in how analyses are to proceed. For example, parametric analyses of mean confidence, as in ANOVAs, require the assumption that the confidence ratings are equally spaced and, in fact, can be viewed as a linear scale. Indeed, most notably, the tacit assumption is that the confidence category labels define points on an equally spaced scale of subjective probabilities. It is not at all well established that they do.

Typically, in probability assessment studies, confidence judgements, viewed as subjective probabilities, are rendered by participants selecting a value from the set, {50, 60, 70, 80, 90, 100}, with "50" denoting a guess and "100" complete certainty. Our approach to permitting a determination of the scale properties of such a set of confidence rating categories is to apply Torgerson's (1957) Law of Categorical Judgment (LCJ) to the matrix of frequencies of confidence category use associated with each of the stimuli in the experiment.

The LCJ posits scale values for both the stimulus items and the midpoints of the rating categories. Torgerson (1957) fully developed the most general forms of the LCJ permitting Gaussian variability in the representations of both the stimuli and the rating category boundaries as well as procedures for obtaining these scale values.