Game Theoretic Models of Clear versus Plain Speech

Paul Tupper (pft3@sfu.ca)
Jie Jian (jjian@sfu.ca)
Department of Mathematics, Simon Fraser University
Burnaby, BC, V5A 1S6, Canada

Keith Leung (kwl23@sfu.ca)
Yue Wang (yuew@sfu.ca)
Department of Linguistics, Simon Fraser University
Burnaby, BC, V5A 1S6, Canada.

Abstract

Clear speech is a vocal style used when a speaker wishes to improve comprehension, usually due to the presence of external noise, less-than-optimal listening conditions, or when they are simply instructed to speak clearly. Clear speech has many distinguishing features, including increased duration, pitch, and amplitude, as well as the exaggeration of articulatory movement. We use game theory to model the phenomenon of clear speech, and make predictions of how it changes under different circumstances. We view the behaviours of speakers and hearers when communicating as optimal strategies in communication games. When comprehension becomes more difficult, the optimal strategies of the games shift so that speakers exert more energy to improve the likelihood of accurate communication. We discuss how our models correspond to experimental observations and see what predictions are made for future experiments.

Keywords: phonetics; game theory; communication games; clear speech; Lombard effect

Introduction

In many situations, speakers use *clear speech*: speech whose properties are modified with the intention of being more comprehensible. Contexts in which clear speech occurs include in the presence of external noise, when speaking to the hearing impaired or language learners, or merely when the speaker is instructed to speak clearly. Clear speech is not identical in these different contexts, but there are certain features that remain similar: increased amplitude, higher pitch, and longer duration (Lam, Tjaden, & Wilding, 2012). Another important feature of clear speech is that phonetic differences between phonemes are exaggerated in order to make distinguishing different phonemes easier for the hearer (Lindblom, 1990).

There is a need for a predictive theory of the differences between clear and plain speech. One issue is the multitude of different effects observed in different studies, and the need to have a systematic framework for their organization and study. Another is the fact that many predicted and observed effects in clear speech are in tension with each other. For example, making a larger articulatory gesture (and thus making the individual phoneme easier to hear) may result in making a pair of phonemes more confusable (Leung, Jongman, Wang, & Sereno, 2016). How does the speaker resolve these competing demands, and what paradigms can we use to probe these issues experimentally?

Our basic assumption is that clear speech exists for the purpose of improving the probability of correct transmission of information. Though it is known that clear speech strategies do not always work, (shouting does not help someone understand you when you don't share a language), in many contexts they do. We conjecture that clear speech style developed as an adaption to improve communication in adverse conditions, and use this assumption to make predictions about clear speech styles.

We use the framework of Game Theory to model clear and plain speaking styles (DeVos & Kent, 2016; Jäger, 2008). We imagine a speaker and a hearer who are engaged in a communication game that they play over and over again. The game involves the speaker transmitting one of a few different possible messages to the hearer. The speaker is allowed to emit a continuous-valued signal to communicate the message, and must decide how different messages are encoded as different signals. When the signal is transmitted, its value is perturbed, so that the hearer only receives a corrupted version of it. The hearer then must decide on a strategy for decoding the message from the signal. Two important factors in the game are that different signals have different costs to the speaker, and that the presence of noise in the system leads to the possibility of transmission of the wrong message. It will turn out that the speaker has to strike a balance between the effort expended in communication and the probability that the wrong message will be received, just as in Lindblom's H&H theory (Lindblom, 1990). Our results demonstrate that as the noise level in the game increases, the speaker will devote more effort to emitting the signal in order to increase the probability of correct communication again. This shift in strategy in response to more noise in the system is what we take as our model of clear speech.

Basic Model

We begin with the most basic version of our model, which we depict in Figure 1. The speaker has one of two messages to convey: either **a** or **b**. We may imagine that the speaker is ordering a beverage at a café, so that **a** means "coffee" and **b** means "tea". We assume that the two messages need to be transmitted equally often, and the consequences of mistakenly transmitting **a** for **b** are the same as for transmitting **b** for **a**. To convey the message the speaker has a single variable whose value they may set and then transmit. For example, suppose the speaker may only emit a single tone of fixed

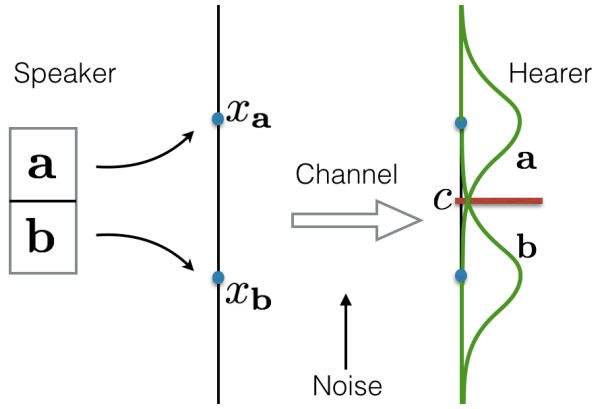

Figure 1: A schematic showing our basic model. The speaker is required to communicate one of two messages: **a** or **b**. They select signal values $x_{\mathbf{a}}$ or $x_{\mathbf{b}}$ which they transmit to the hearer. Noise in the communication channel leads to the hearer receiving a perturbed signal which they classify as either **a** or **b** based on the criterion c.

duration, but is able to select the pitch. Now either in production, transmission, or reception, noise is added to the variable, so that the original value selected by the speaker is perturbed before it is perceived by the hearer. Then the hearer must infer what message was intended by the speaker from this noisy stimulus.

The speaker's strategy is to select two values of the variable x: $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$ for the signals **a** and **b** respectively. Without loss of generality we assume that $x_{\mathbf{a}} < x_{\mathbf{b}}$. Now, depending on the message, **a** or **b**, what is perceived by the hearer is either $y = x_{\mathbf{a}} + \sigma n$ or $y = x_{\mathbf{b}} + \sigma n$ where n is a standard Gaussian random variable, and σ is a noise amplitude. The hearer must make a decision based on the heard signal y. The hearer's task is an instance of the standard model in *Signal Detection Theory* (Macmillan, 2002): the optimal choice is to fix a value c, known as the *criterion*, and choose **a** when $y \le c$ and choose **b** otherwise. The optimal value of c will be the one that maximizes the probability of receiving the correct message. Recalling that each message is equally likely, we can express this probability as

$$\begin{aligned} P(x_{\mathbf{a}}, x_{\mathbf{b}}, c) &= \mathbb{P}(\text{correct transmission}) \\ &= \mathbb{P}(\text{correct}|\mathbf{a})\mathbb{P}(\mathbf{a}) + \mathbb{P}(\text{correct}|\mathbf{b})\mathbb{P}(\mathbf{b}) \\ &= \frac{1}{2}\mathbb{P}(x_{\mathbf{a}} + \sigma n \le c) + \frac{1}{2}\mathbb{P}(x_{\mathbf{b}} + \sigma n > c) \\ &= \frac{1}{2}F\left(\frac{c - x_{\mathbf{a}}}{\sigma}\right) + \frac{1}{2}\left[1 - F\left(\frac{c - x_{\mathbf{b}}}{\sigma}\right)\right] \end{aligned}$$

Here F is the cumulative distribution function of a standard normal random variable with mean 0 and variance 1, as shown in Figure 2 *left*. A simple use of calculus shows that whatever $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$ are, the optimal value for the criterion is $c = (x_{\mathbf{a}} + x_{\mathbf{b}})/2$. So whatever the speaker chooses for $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$, this is the choice the hearer will make in equilibrium.

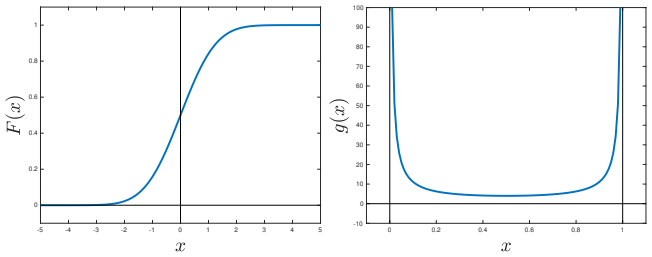

Figure 2: *Left:* F, the cumulative distribution function of a standard normal random variable. *Right:* g, the function describing the cost in our model of emitting a signal with a given phonetic variable.

The issue remains of how the speaker should choose $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$. If there are no constraints on x, the probability of success will increase to 1 as the distance between $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$ increases. In any realistic system there is either a finite range of possibilities for x, or there is a disincentive for using large or small values of x. The idea is that more extreme values of x require more effort, and the speaker will make less effort unless there is sufficient benefit to making more effort (Lindblom, 1990).

We combine these ideas by defining a cost for emitting a signal that depends on x, the phonetic variable of the signal. Suppose that the effort required to emit signal x is $k\,g(x)$ where g is defined as

$$g(x) = \frac{1}{x(1-x)}, \text{ for } 0 \le x \le 1, \tag{1}$$

and set $g(x) = \infty$ for $x < 0$ or $x > 1$, as shown in Figure 2 *right*. k is some positive constant we use to parameterize the overall effort in emitting a signal. We chose this form for g because it means that (i) only sounds in the range $(0,1)$ can be emitted and they all have positive cost, (ii) more extreme sounds are more difficult to emit, (iii) effort is close to constant for signals within the middle of range. Other, similar, forms of g give the same qualitative results as we present here.

We now make a fairly strong assumption for the purposes of simplicity: the speaker and the hearer have the same payoff function in the game. So they are equally interested in the correct message being transmitted, and are equally interested in the speaker's effort being minimized. (This is clearly not always a reasonable assumption, and future work will consider different models.) Following this symmetric modelling choice, we assume that the expected payoff to the speaker and the hearer in one round of the communication game is

$$E(x_{\mathbf{a}}, x_{\mathbf{b}}, c) = P(x_{\mathbf{a}}, x_{\mathbf{b}}, c) - \frac{k}{2}\left(g(x_{\mathbf{a}}) + g(x_{\mathbf{b}})\right),$$

that is, the probability of the message being correct minus the average cost to the speaker of transmitting x. We assume the speaker and hearer will act to maximize this quantity. As we have mentioned, the optimal solution will always have $c = (x_{\mathbf{a}} + x_{\mathbf{b}})/2$. The symmetry of $g(x)$ about $x = 1/2$ implies that the optimum will always have $x_{\mathbf{b}} - 1/2 = 1/2 - x_{\mathbf{a}}$

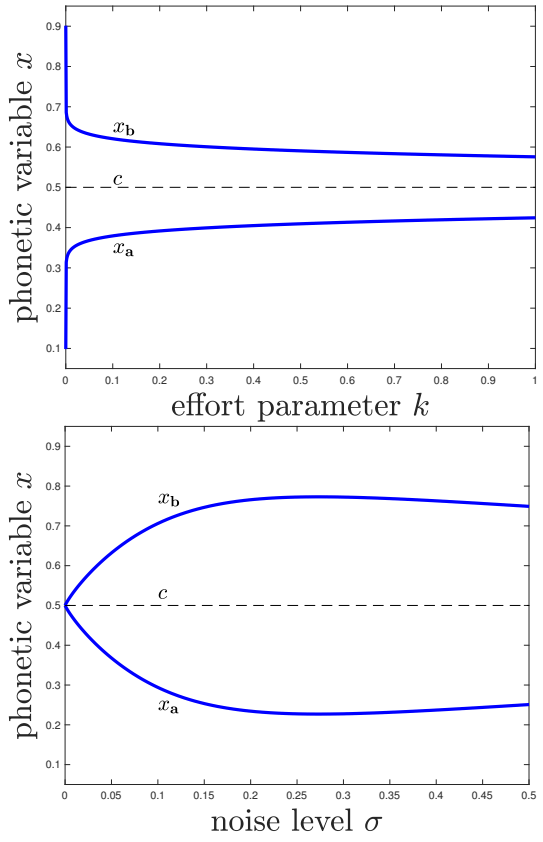

Figure 3: The basic model. *Top:* The value of the optimal $x_{\mathbf{a}}, x_{\mathbf{b}}$ and c for varying k and $\sigma = 0.05$. *Bottom:* The same for $k = 0.05$ and varying σ.

implying $c = 1/2$. So we only need to maximize the function $P(1/2 - \Delta x, 1/2 + \Delta x, 1/2)$ with respect to $\Delta x \geq 0$ to find the optimum of the original problem. This can be performed by a numerical optimization. This and all other optimization problems we study were solved using Matlab's `fminsearch` routine (The MathWorks, Inc., 2016).

Using our computed solutions to the optimization problem, we study how $x_{\mathbf{a}}$, $x_{\mathbf{b}}$, and c depend on σ (the noise amplitude) and k (the effort parameter). Figure 3 *top* shows how $x_{\mathbf{a}}, x_{\mathbf{b}}$, and c depend on k for a fixed value of $\sigma = 0.05$. We see that as k goes to 0, $x_{\mathbf{a}}$ goes to 0 and $x_{\mathbf{b}}$ goes to 1, which makes sense, since in this limit, there is no penalty for making the gestures as large as possible. Likewise, as k goes to infinity, $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$ both go to 1/2, the cheapest possible signal, since the cost of emitting a signal becomes large compared to the benefit of accurate communication.

More interesting is the case of fixed k and varying σ. Figure 3 *bottom* shows how $x_{\mathbf{a}}, x_{\mathbf{b}}$ and c depend on σ for a fixed value of $k = 0.05$. For small values of σ, as we might expect, $x_{\mathbf{a}}$, $x_{\mathbf{b}}$ go to 1/2 as σ goes to zero. This makes sense, as when there is no noise, even the slightest difference between $x_{\mathbf{a}}$ and $x_{\mathbf{b}}$ gives perfect communication, and setting both to 1/2 minimizes effort. Less expected is what happens as σ increases. Initially, as σ increases from 0, gestures become more extreme in order to improve the probability of correct communication. This is the standard clear speech effect, and is a key part of Lombard speech, that is, speech in the presence of noise (Brumm & Zollinger, 2011). See, for example, (Ferguson & Kewley-Port, 2002) for this effect in F2 in English vowels and (Tartter, Gomes, & Litwin, 1993) for duration and amplitude.

Surprisingly, past a certain noise level the effect reverses itself, and gestures become less extreme in our model. This occurs because, if the noise is large enough, the probability of communication regardless of the signals used is so low that it is no longer worth the effort to make the more extreme gestures that were worthwhile for a lower level of noise. We know of no observations of this phenomena, but predict that it will be observed for human subjects with sufficiently large amplitudes of noise. Indeed, (Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988) observes speech amplitude increasing with a decreasing rate as noise level is increased, and a reduction in amplitude may be observable if an even larger noise level is tried. A similar phenomenon *has* been observed in domestic fowl (Brumm, Schmidt, & Schrader, 2009). The chickens studied varied the frequency with which they repeated their calls in the presence of different amounts of noise. It was observed that for lower levels of noise the birds increased call frequencies with increasing noise, and it is conjectured that this is an adaption to improve the probability of communication by expending more effort. However, interestingly the authors noted that after the noise was increased past a certain point, the birds *decreased* the frequency of their calls, as would be predicted by our model.

Four-Message Model

The basic model of the previous section can also be extended to an arbitrary number of messages analogously. Here we just consider the case of four messages, as we will build on this case subsequently. Now there are messages $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$, again corresponding to four distinct meanings. The strategy of the speaker is to choose signals $x_{\mathbf{a}} < x_{\mathbf{b}} < x_{\mathbf{c}} < x_{\mathbf{d}}$ to represent them. Whatever these values, the optimal choice for the hearer is to use criterion points $c_{\mathbf{ab}} = (x_{\mathbf{a}} + x_{\mathbf{b}})/2$, $c_{\mathbf{bc}} = (x_{\mathbf{b}} + x_{\mathbf{c}})/2$, and $c_{\mathbf{cd}} = (x_{\mathbf{c}} + x_{\mathbf{d}})/2$. When the hearer receives signal $y = x + \sigma N$, they select message $\mathbf{a}$ if $y \leq c_{\mathbf{ab}}$, $\mathbf{b}$ if $c_{\mathbf{ab}} < y \leq c_{\mathbf{bc}}$, $\mathbf{c}$ if $c_{\mathbf{bc}} < y \leq c_{\mathbf{cd}}$, and $\mathbf{d}$ if $y > c_{\mathbf{cd}}$. We do not explicitly state the expressions for the probability of correct transmission and the expected cost, since they are unwieldy and completely analogous to those given for the basic model. Again the payoff function $E(x_{\mathbf{a}}, x_{\mathbf{b}}, x_{\mathbf{c}}, x_{\mathbf{d}}, c_{\mathbf{ab}}, c_{\mathbf{bc}}, c_{\mathbf{cd}})$ is the the difference between the probability of correct transmission and the expected cost, and both speaker and hearer act to maximize it.

We show in Figure 4 the results of the optimization for this case for a range of σ and k. Similarly to the two-message case, with both decreasing cost parameter k and increasing but low noise level σ, the speaker uses more extreme signals

to convey the same message. In contrast to the two-message case though, where the hearer adopted the same strategy for all values of k and σ, in the four-message case, the speaker must adjust the criterion points $c_{\mathbf{ab}}$ and $c_{\mathbf{cd}}$ in response to the change in the speakers strategy. As the dashed lines in Figure 4 show, when the speaker uses more extreme signals, the hearer must compensate.

We note in passing that in the four-message game, the speaker uses a larger portion of the phonetic space. This can be seen by comparing the range of x used for the signals for any particular k and σ, as in Figures 3 and 4. For example, when $k = 0.1$ and $\sigma = 0.05$, in the two-message case, the signals used range from about 0.4 to 0.6, whereas in the four-message case they range from about 0.25 to 0.75. This is in general accordance with the Theory of Adaptive Dispersion (Liljencrants & Lindblom, 1972; Lindblom, 1986), which postulates that the more phonemes needed to fit into a space, the more dispersed they will be. One way of describing it is that the basic model has a smaller range of x than the four-message model because that smaller range already provides sufficient contrast between two messages. When two more messages are added, and the number of contrasts that needs to be made rises to three, there is a need for expanding the phonetic space. But because more extreme signals x are more costly, and signals less than 0 or greater than 1 are impossible, the range cannot be tripled. The net effect is that with more messages to transmit, the speaker expands the phonetic space used, while decreasing the spacing between the signals for distinct messages.

Conflicts Between Clearness and Comprehensibility

In the two-message case we saw that the speaker needs to modify their strategy in the presence of greater noise in order to maximize payoff. But in this case the hearer does not need to make any adjustment to their strategy in response to the speaker's speech style, as shown by the flat dashed line labeled c in Figure 3.

In the four-message case, we see from the dashed lines in Figure 4 that the hearer *does* have to adjust their strategy as well when k or σ is changed, in order to optimally respond to the speaker's change of strategy. Such changes in hearer's strategies in interpreting stimuli based on context (such as for accent, gender, or identity of the speaker) are well attested, and known to be an important part of comprehension (McMurray & Jongman, 2011).

However, a problem for our language users can arise if the speaker is using a clear speech style, with its exaggerated signals, but the hearer thinks the speaker is using a plain speech style. If the hearer is using the plain speech criterion points (low noise) to interpret a clear speech signal (high noise), they may end up with the wrong message after decoding. For a concrete example, take our four-message model, and see Figure 4 *bottom*. Suppose $k = 0.05$ and σ takes one of two values: $\sigma_1 = 0.01$ or $\sigma_2 = 0.05$. Suppose the speaker believes

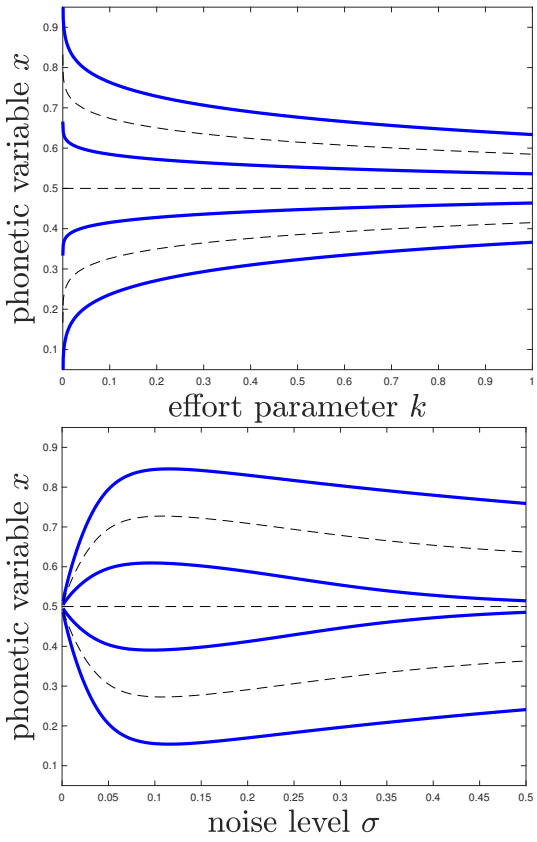

Figure 4: The four-signal model. *Top:* The value of the optimal phonetic variable values $x_{\mathbf{a}}, x_{\mathbf{b}}, x_{\mathbf{c}}, x_{\mathbf{d}}$ (solid blue line) used by the speaker and criteria $c_{\mathbf{ab}}, c_{\mathbf{bc}}, c_{\mathbf{cd}}$ (dashed black line) used by the hearer for varying effort parameter k and noise level $\sigma = 0.05$. *Bottom:* The same for $k = 0.05$ and varying σ.

$\sigma = \sigma_2$ and intends to transmit message **b** and so utters a signal near $y = 0.4$. If the hearer thinks $\sigma = \sigma_1$, they will decode this as message **a**, the wrong message. The problem is that a plain (σ_1) **b** has a similar signal to a clear (σ_2) **a**, and so the hearer cannot tell them apart if they don't know whether a plain or clear style is being used.

An example of how this might occur in a natural language is in the difference between tense and lax vowels in English, as exemplified by the minimal pairs "keyed – kid" (/i – ɪ/), "cod – cud" (/ɑ – ʌ/), "cooed – could" (/u – ʊ/). In each pair the tense vowel is longer and the lax vowel is shorter. This distinction has been shown to be perceptually important in spoken English (though there are other important contrastive features between these vowels as well) (Klatt, 1976). So suppose that a speaker wishes to be clearer by lengthening the vowels of their production. This may create a problem for the hearer, since a long "kid" might be difficult to distinguish from a normal "keyed". There is a conflict between a non-phonemic speech clarity effect (i.e. lengthening) and a phonemic contrast (Leung et al., 2016).

A similar phenomenon may occur in animal communication systems. Many animals lengthen their alarm calls in the presence of external noise (Brumm et al., 2009). But Brants' whistling rat uses a duration of alarm call to indicate the level of threat (Le Roux, Jackson, & Cherry, 2001). It is suggested that they may not lengthen their calls in the presence of noise for this reason (Brumm et al., 2009), though this has not yet been experimentally investigated.

So what is the optimal strategy in the case where the speaker is aware of σ, but the hearer is not, and must use one set of criterion points for both noise levels? One possibility is that the speaker decides it is not worth it to have two different sets of signals, one for each level of noise. The other extreme possibility is that the speaker just assumes that the hearer does know what the noise level is, even though they don't. As we will see, the predicted behaviour is a compromise between these two simpler strategies.

To explore the predictions of our model for these contexts we consider three different cases for the speaker's and hearer's knowledge of the noise level σ. The speaker is communicating one of four different possible messages, with $k = 0.05$ and where in each trial σ takes the value of either $\sigma_1 = 0.01$ or $\sigma_2 = 0.05$ with probability $1/2$. Each of messages $\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}$ occur with probability $1/4$. To simplify notation we let $X = (x_\mathbf{a}, x_\mathbf{b}, x_\mathbf{c}, x_\mathbf{d})$ and $C = (c_\mathbf{ab}, c_\mathbf{bc}, c_\mathbf{cd})$ be strategy vectors for the speaker and hearer respectively. For each condition we state the optimal values for X and C as computed by numerical optimization.

Case i: Neither Oblivious. Both speaker and hearer know the value of σ in each trial. Thus both their strategies can be modified with respect to noise level. So we determine $X^{(1)}$ and $C^{(1)}$ for noise level σ_1 and $X^{(2)}$ and $C^{(2)}$ for noise level σ_2 as in the previous section. Equivalently, $X^{(1)}, X^{(2)}, C^{(1)}, C^{(2)}$ together maximize

$$\frac{1}{2}E(X^{(1)}, C^{(1)}, k, \sigma_1) + \frac{1}{2}E(X^{(2)}, C^{(2)}, k, \sigma_2)$$

Here, because the σ_1 vectors and the σ_2 vectors are in separate terms, they can be solved for independently.

Case ii: Oblivious Hearer. The speaker knows the value of σ in each trial but the hearer does not. Speaker has strategies $X^{(1)}$ and $X^{(2)}$ depending on the noise level, but the hearer only has C. $X^{(1)}$, $X^{(2)}$ and C together maximize

$$\frac{1}{2}E(X^{(1)}, C, k, \sigma_1) + \frac{1}{2}E(X^{(2)}, C, k, \sigma_2).$$

Case iii: Both Oblivious. Neither the speaker nor the hearer know the value of σ for each trial. In this case X and C are determined by maximizing

$$\frac{1}{2}E(X, C, k, \sigma_1) + \frac{1}{2}E(X, C, k, \sigma_2).$$

In Figure 5 we show the optimal strategies for the speaker and hearer in the four-message, two noise-level game, in each of the three cases. The signals $x_\mathbf{a}, x_\mathbf{b}, x_\mathbf{c}, x_\mathbf{d}$ are shown as blue circles, and the criterion points $c_\mathbf{ab}, c_\mathbf{bc}, c_\mathbf{cd}$ are shown as red $\times$s. Lines connecting corresponding signal values and criterion points are added to aid comparison. Leftmost in the figure, we show the strategies for each level of noise when both speaker and hearer are aware of the level of the noise (Neither Oblivious). As we expect from the previous section and Figure 4, the speaker uses more extreme signal values when the noise is greater, and the hearer is able to take this into account in the setting of the criterion points. The payoff achieved in this case is 0.750. In the middle of the figure, we show the strategies when only the speaker is aware of the level of the noise (Hearer Oblivious). The speaker still emits more extreme signals when the noise level is higher, as in the Neither Oblivious case. But this effect is relatively muted: when the noise level is σ_1 the range of the signals is greater than in the Neither Oblivious case, and when the noise level is σ_2 the range of the signals is less than in the Neither Oblivious case. The fact that the hearer is unaware of the noise level means the speaker cannot deploy this strategy to full effect. The payoff achieved is now 0.741, and hence a cost is paid for the hearer's ignorance. Finally, for purposes of comparison, on the right we show the effect of two noise levels when neither speaker nor hearer is aware of the level in a particular trial (Both Oblivious). In this case, the speaker and hearer use a compromise of the strategies in the other cases, leading to a payoff of 0.729, worse than either of the other two cases.

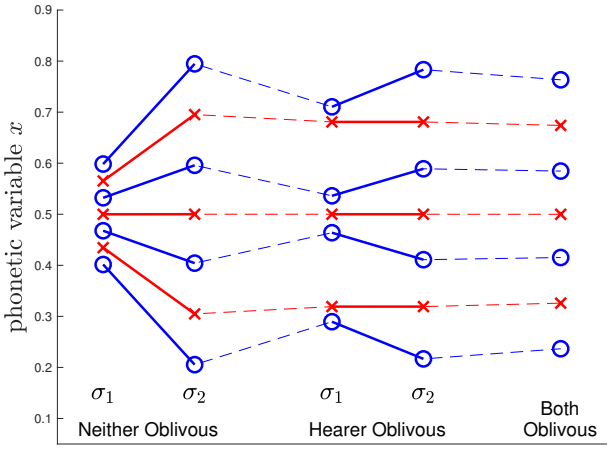

Figure 5: The optimal strategies for the speaker and the hearer in the four-signal model with two levels of noise. The condition is indicated along the x-axis. The y-axis indicates the value of the signal used for each of the four messages by the speaker (in blue), and the three criterion points used by the hearer (in red). Lines are present merely to guide the eye.

Thus according to our model, uncertainty in the speaker about whether a clear speech style is being used diminishes the speaker's distinction between plain and clear speech. One way to investigate this effect experimentally when eliciting clear speech is to vary the instructions, sometimes explaining that the intended hearer will be aware that the speech is clear,

and sometimes not. Alternatively, both a speaker and hearer could be studied together, with noise of different levels being played on separate headphones. Whether the speaker knows if the hearer has the same noise level or not can be manipulated, allowing this effect to be investigated.

Predictions of our Model

Here we summarize the predictions our model makes for the contrast between clear and plain speech.

1. Phonetic features used to distinguish phonemes will be exaggerated (produced with more extreme values) in clear speech.
2. In noise-induced clear speech, as the noise level is increased, signals will first be exaggerated but then eventually become less exaggerated past a certain noise level.
3. Increasing the number of phonemes contrasted by a single variable will cause the phonetic range to increase while the phonemes are packed closer together within it.
4. In noise-induced clear speech, if both speaker and hearer are aware of the noise level, differences between clear and plain speech will be greater than in the case where the speaker believes the hearer is unaware of the noise level.

Future Directions

There are many ways to extend our models to explore further aspects of clear speech. We list some of them here.

1. **More and Less Probable Messages.** We assumed that all messages were equally likely to be transmitted. This is not at all necessary for our model, and in future we will study how making a message more frequent changes the position of its signal in phonetic space.
2. **Multiple Phonetic Variables.** We considered a communication game in which the speaker can vary only one phonetic variable in the signal. In real speech many different dimensions of a signal can be controlled. Our models can be expanded to handle more signal dimensions. This will allow us to model and study the effects of variables like amplitude (i.e. loudness or intensity) which are typically not used contrastively, as well as how speakers decide among multiple variables which to use as a contrastive one.
3. **Asymmetric Payoffs.** We have assumed that the speaker and hearer have the same payoff function. In reality the speaker bears the cost of articulation alone, and the speaker and hearer may not be equal in how important it is to them for the message to be transmitted correctly. These factors may be important in determining how clear speech is used in a social context.

Acknowledgments

The authors thank Gavin Jenkins for comments on an earlier draft of this manuscript. PT, JJ, KL, YW were supported by NSERC Discovery Grants. PT was also supported by an NSERC Research Accelerator Supplement, and held a Tier II Canada Research Chair.

References

Brumm, H., Schmidt, R., & Schrader, L. (2009). Noise-dependent vocal plasticity in domestic fowl. *Animal Behaviour*, *78*(3), 741 - 746.

Brumm, H., & Zollinger, S. A. (2011). The evolution of the lombard effect: 100 years of psychoacoustic research. *Behaviour*, *148*(11-13), 1173–1198.

DeVos, M. J., & Kent, D. A. (2016). *Game theory: A playful introduction*. American Mathematical Society.

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *112*(1), 259–271.

Jäger, G. (2008). Applications of game theory in linguistics. *Language and Linguistics compass*, *2*(3), 406–421.

Klatt, D. H. (1976). Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research*, *55*(6), 1807–1821.

Le Roux, A., Jackson, T. P., & Cherry, M. I. (2001). Does Brants' whistling rat (parotomys brantsii) use an urgency-based alarm system in reaction to aerial and terrestrial predators? *Behaviour*, *138*(6), 757–773.

Leung, K. K. W., Jongman, A., Wang, Y., & Sereno, J. A. (2016). Acoustic characteristics of clearly spoken english tense and lax vowels. *The Journal of the Acoustical Society of America*, *140*(1), 45-58. doi: 10.1121/1.4954737

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 839–862.

Lindblom, B. (1986). Phonetic universals in vowel systems. *Experimental phonology*, 13–44.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Macmillan, N. A. (2002). Signal detection theory. *Stevens' handbook of experimental psychology*.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917–928.

Tartter, V. C., Gomes, H., & Litwin, E. (1993). Some acoustic effects of listening to noise on speech production. *The Journal of the Acoustical Society of America*, *94*(4), 2437–2440.

The MathWorks, Inc. (2016). *Matlab and statistics toolbox release 2016b.*