UC Berkeley Phonology Lab Annual Report (2005)

Resonance in an exemplar-based lexicon: The emergence of social identity and phonology

running title: Exemplar-based lexicon

Keith Johnson

Department of Linguistics
UC Berkeley
Berkeley, CA 94720 USA
keithjohnson@berkeley.edu

Abstract: Two sets of data are discussed in terms of an exemplar resonance model of the lexicon. First, a cross-linguistic review of vowel formant measurements indicate that phonetic differences between male and female talkers are a function of language, dissociated to a certain extent from vocal tract length. Second, an auditory word recognition study (Strand, 2000) indicates that listeners can process words faster when the talker has a stereotypical sounding voice. An exemplar resonance model of perception derives these effects suggesting that reentrant pathways (Edelman, 1987) between cognitive categories and detailed exemplars of them leads to the emergence of social and linguistic entities.

Keywords: vowel formants, lexicon, auditory word recognition, gender, normalization, exemplar-based model, resonance

1. Introduction

This paper approaches the topic of sociophonetic variation from the viewpoint of speech perception theory. The connection between these two areas of research arises because acoustic/phonetic variability of the sort that often marks social identity poses a most significant and interesting challenge to theories of speech perception.

Sociophonetic variation is more interesting than anatomical variation because it points the speech perception theorist to what in my opinion is a more nearly correct view of speech perception. We have tended to assume that all acoustic/phonetic variability is lawful and perhaps hardwired in the neural circuitry for speech perception (Sussman, 1986; Fowler, 1986). However, when we recognize sociophonetic variability, with its phonetically arbitrary social signaling values, it becomes necessary to give up on hardwiring as the key to overcoming variability in speech perception.

My own research has tended to focus on the sociophonetic variation found in gender differences. This is because speech perception theorists have typically treated gender variation as (1) a source of substantial phonetic variation, and (2) lawfully derived from typical vocal tract length differences, and therefore a candidate for processing via a hardwired "normalization" algorithm. Recent research (summarized in Johnson, 2005) suggests that the cross-language and within-language phonetic arbitrariness of gender and its implications for spoken language processing are very important for the theory of speech perception. In particular, the arbitrariness of gender differentiation in speech calls into question "normalization" and the unitary abstract phonetic representations usually assumed by normalization procedures.

Perhaps also the research reported here will also be of some interest to researchers whose main focus is on sociophonetics because identifying cross-linguistic arbitrary gender differences supports the idea that people "perform" gender to some extent (see e.g. the articles in Caplan, 1987; and Weeks, 1989).

Section 2 describes some cross-linguistic phonetic differences between male and female talkers, in support of the idea that gender differences are purely due to anatomical differences. Section 3 reviews

Strand's (2000) finding that gender stereotypes influence spoken word processing. Section 4 presents the exemplar resonance model that accounts for these findings.

2. The Cross-linguistic phonetics of gender.

In Johnson (2005), I presented the results of a cross-lingusitic survey that showed that men and women's vowel formants differ from language to language. This exercise built on an earlier small survey reported by Bladon, Henton & Pickering (1984). They found that vowels produced by men are on average lower by about one Bark (critical band of auditory frequency resolution) from vowels produced by women. They also reported that the male/female difference varies from language to language in a sample of six languages.

Figure 1 shows some new data on gender differences in vowel production drawn from published studies of groups of men and women, see appendix 1. The frequencies of F1 and F2 across [i], [e], [a], [o], and [u] (except as noted in the appendix) were converted into auditory frequencies on the Bark scale (Schroeder et al., 1979) and the male/female difference for each vowel was calculated. Then the male/female differences were averaged over the five vowels to produce the "Gender difference" values plotted on the horizontal axis in Figure 1.

The data in Figure 1 support the initial observation reported by Bladon, Henton & Pickering (1984) illustrating that male/female vowel formant differences are variable across languages and dialects of the same language. Together with data on the acquisition of phonetic gender differences (see Johnson, 2005), these data suggest that gender differences are not solely due to vocal tract anatomical differences between men and women.

However, the cross-linguistic differences that we see in Figure 1 could be due to the luck of the draw. If, for example, in Behne et al.'s study of Norwegian vowels, the male participants happened to have unusually long vocal tracts or the female participants unusually short vocal tracts, then the large difference between male and female vowel formants in Norwegian may not have anything to do with Norwegian generally, but only with these particular talkers. Of course, with a large enough random

sample, we expect our findings such as these vowel formant measurements to be representative, but phonetic samples are often not very large or random. But even if the phonetic data are representative, of Norwegian men and women, it might be that Norwegian men and women generally differ more in vocal tract length than do, for example, Danish men and women.

One special feature of the languages surveyed in Figure 1 is that demographic data are available for each population (Tolonen et al., 2000). In particular, we have available data on the average height of men and women in these populations. Because research by anesthesiologists has shown that body height is highly correlated with vocal tract length ($r(292) = 0.79$, Cherng et al., 2002), it seems likely that we could make some predictions about male and female vowel formant differences on the basis of the Tolonen et al. body height data, and then compare these predictions with the data observed in the survey.

For example, the Tolonen et al. (2000) data find that men in Norway average 176 cm in height while women average 161.5 cm. This height difference of 15.5 cm is larger than any other difference among the languages represented in Figure 1. So, vowel formants *should* be more different for Norwegian men and women than for the other languages in the vowel survey.

Figure 1 about here

However, as Figure 2 shows, the relationship between body height differences between men and women and vowel formant differences between men and women is not tight at all. The general trend is as we expect - the larger the height difference in the Tolomen et al. (2000) data, the larger the difference between male and female vowel formants. However, height difference only accounts for between 6% and 40% of the vowel formant difference between men and women. Note that the highest correlation occurs for F2 where we have the greatest degree of separation between male and female formant frequency - a range effect - and the lowest correlation occurs for F3 where we have in addition to a small range of formant frequencies a smaller number of observations because F3 data are not available

for four of the languages. Note also that a nonlinear regression predicting the F2 difference from the log of the height difference resulted in a slightly lower R2 value (0.39). The main point of this figure can be made by looking at the F2 panel. In this data, Danish men and women's vowels differ quite a bit less than we would expect given their height difference, while the Australian English men and women differ more than the linear regression line predicts.

Figure 2 about here

There are two possible sources of the looseness of the relationship shown in Figure 2. First, it may be that the phonetic samples and the demographic samples do not overlap with each other sufficiently to allow for accurate prediction of vowel formants of one group of people from the body heights of another group drawn from the same population. A second possibility is that men and/or women "perform" gender differently in these languages. This was the interpretation given in Johnson (2005).

Figure 3 supports the second alternative. This figure shows the average formant frequency measurements and average body height measurements that are summarized as male/female differences in Figures 1 and 2. The key observations to draw from this figure are that average body height is pretty well correlated with with F3 frequency, but that F1 and especially F2 are not so predictable from height.

Figure 3 about here

Table I quantifies these observations. The correlations for F3, even with this small number of data points (n=14), were reliable for women and nearly so for men, while F1 and F2 could not be reliably predicted from body height. Additionally, the slope of the regression line of F2 for men was positive, indicating a tendency for lower F2 for men with *shorter* vocal tracts! These data support the idea that the phonetic samples are representative of vocal tracts of speakers of these languages - this is indicated

by the F3/height correlations - but that the weak relationship between vocal tract length differences and male/female vowel formant differences (Figure 2) is due to aspects of the performance of gender.

Table I about here

This analysis of overall trends in average vowel formant frequencies neglects a number of possible "gender dialect" features that have been noted in previous research (Fant, 1975), but the overall conclusion is clear. People (perhaps especially men) perform gender. The implications of this result for the theory of speech perception have been explored before (Johnson, 1989, 1997a,b, 2005) - listeners must compensate for gender differences in a way that pays special attention to the typical or expected difference between men and women for a particular speech community. Johnson (1997a) proposed that this community-sensitive evaluation of gender differences could be accomplished in an exemplar-based system of categorization (Nosofsky, 1986; Hintzman, 1986). Before discussing some additional details about how a resonance-exemplar model might accomplish this, the next section describes some recent data that show that listeners are sensitive to gender stereotypes in auditory word recognition.

3. A stereotype effect in auditory word recognition.

We have seen that some of the particulars of speech production are performed differently by men and women. This seems to be a sensible interpretation of how we have F3 predictable from height while F1 and F2 are not.

This section reviews a recent experiment (Strand, 2000) that shows that gender stereotypicality matters in speech perception. The purpose of this section is to tie the performance of gender together with representation of speech in long term memory. If gender is a social category and sounding male or female is a matter of performance, then the perceptual process is likely to be sensitive to stereotypical gender characteristics.

Strand (2000) asked 24 listeners (15 female, 9 male) to use a nine-point scale to rate how different pairs of talkers sound. Half of the pairs to be rated for subjective talker similarity were cross-gender pairs and half were within-gender pairs. There were 10 men and 10 women in the set of talkers to be judged, and the listeners heard all possible pairs of these talkers speaking a word drawn from a list of 24 phonetically balanced, monosyllabic, high frequency words. Another group of 10 listeners (2 male, 8 female) heard the same talkers and words and were asked to simply identify the gender of the talker as "male" or "female". The amount of time that it took the listener to identify the talker's gender was recorded to within the nearest millisecond.

These two sets of data (talker difference, and speeded gender identification) were used to identify stereotypical and nonstereotypical male and female talkers. A multidimensional scaling (MDS) map of the talkers was produced from the talker difference data. In this map the male talkers clustered together leaving only a couple of speakers on the periphery of this cluster of male voices. Similarly, there were female talkers who clustered with each other and a few who were more peripheral to the female group. In the speeded gender classification results male talkers who clustered tightly with other male talkers were generally identified as "male" more quickly than male talkers who were peripheral to the male group. The same relationship held between the perceptual talker space for women and the "female" speeded gender classification response times. Strand selected a stereotypical male as a talker from the male cluster who was identified as "male" more quickly than any of the other male voices, and a nonstereotypical male who was peripheral to the male group in the perceptual talker map, and was identified as "male" more slowly than the other men. Stereotypical and nonstereotypical women's voices were also selected by these criteria.

Another group of listeners (14 female, 10 male) were then asked to "name" words produced by the four talkers identified as stereotypical and nonstereotypical male and female. Each listener heard 48 phonetically balanced, monosyllabic, high frequency words produced by each of the four talkers (4 presentations of each word). These were counterbalanced for order of the blocks with each speaker presented 12 times in each block and no word repeated in a block. The listener's task was to repeat aloud the word that the talker said and the response measure was the amount of time it took to repeat

the word, as measured by a voice-activated switch.

Figure 4 about here

Strand's (2000) reaction time data are shown in Figure 4. The primary finding of this study is that listeners took longer to begin repeating the words produced by the nonstereotypical talkers ($F[1,23]=183.9$, $p < 0.01$]. This voice stereotypicality effect was also significant in an analysis of covariance that took stimulus word duration as a covariant of reaction time ($F[1,187]=4.7$, $p<0.05$). This analysis of covariance showed that the apparent interaction in Figure 4, in which the nonstereotypical male's words were processed more slowly than the nonstereotypical female's words, is due to different stimulus word durations and not a genuine interaction in processing time. The ANCOVA also demonstrates that the voice stereotypicality effect is real and robust, and is independent of any word duration effect.

The general finding is clear. Listeners come to the task of speech perception/auditory word recognition with gender expectations. Linguistic material produced by voices that do not fit those expectations is not processed as efficiently as is material produced by stereotypical voices. The next section presents a resonance-exemplar model that is capable of producing emergent categories of talker, and incidentally linguistic analysis.

4. An exemplar-based model of perception.

The model presented in this section builds on the one described earlier in Johnson (1997b). Resonance (Carpenter & Grossberg, 1987) or reentrant mapping (Edelman, 1987), which has precursors in cognitive science at least since Wiener (1950), is used in this version of the model to provide interaction between auditory, visual, and declarative knowledge representations. Hintzman's (1986) MINERVA used a similar mechanism that he called an "echo".

Before turning to a discussion of resonance, the next section digresses briefly for a discussion of declarative and recognition memory.

4.1. Remembering John Choi: Declarative versus recognition memory

My memories of the late John Choi (1991, 1992, 1995, Choi & Keating, 1991) are based on my personal interactions with him in dinners, road trips, club hopping, meetings, formal presentations, and heart to heart chats. I can describe some of these experiences in words. Such as the loud, rambling discussion we had about phonetics while we were driving from Berkeley to Los Angeles on the I-5, smoking and listening to loud music. I remember watching with awe his impersonation of a Korean soldier (made more believable by the fact that he had served in the Korean army). I remember how fast he could chop garlic, and how reluctant yet determined he was to keep asking hard questions of Donca Steriade at a phonology group meeting.

More to the point of this paper, I have memories of John's physical appearance. The deep brown of his eyes, the fuzzy cut of his hair, the breadth of his shoulders and the spindliness of his legs. And of his voice, the highish pitch and careful enunciation. But the descriptions of these memories is different from the mental images themselves. To a person who didn't know him, my descriptions are probably not enough to pick him out of a set of photographs or recognize his voice among a set of recordings because recognition requires access to the particulars, before they have been filtered through a verbal description.

Psychologists call these two types of memory declarative memory and recognition memory. The fact that I can remember who was the seventeenth president of the United States is a case of declarative memory. I had no experience of Andrew Johnson or his impeachment. These are only facts that I was taught in school. If I were to try to describe the man it would be only from the photograph or two that I have seen of him, and the sketchy stories that I have heard. However, my experience of the thirty-sixth president, while still relatively impoverished is a good deal richer. As with Andrew Johnson, I can recite (or declare) some of the facts that I learned about Lyndon Johnson. But I also have the image of him on television declaring his intention to not run for reelection. Somehow, stored with that black

and white television image I have recorded the texture and color of the shag carpet that I was lying on when I saw the broadcast. Though I can only give a rough description of that experience, I am pretty sure that I would recognize the difference between an authentic recording and an imitation even though I saw that speech more than thirty years ago.

Linguistic descriptions of language are declarative memories. The linguist who experiences the language, who listens to a consultant pronounce words and sentences, produces a description of the experience. Though our descriptions may be carefully accurate, and our analyses clever in discovering and describing the patterns in these descriptions, we nonetheless and by necessity give a sketchy report of the vast richness of the experience. And for most linguists, this experience is itself only a shadow of the linguistic experience of our language consultants. As in records of the recent past, such as the life of John Choi, or of more distant lives such as Andrew Johnson, a descriptive record of language is a very valuable thing.

However, keeping in mind the intrinsic value of grammars and dictionaries, the key idea of the exemplar-based approach is that people remember, as the core of the cognitive representation of language, linguistic episodes not linguistic descriptions. We operate from mental images - detailed memories of specific linguistic experiences - rather than from impoverished descriptions of such experiences.

4.2 An Exemplar-resonance model.

Figure 5 illustrates the exemplar resonance model used in the simulation here. In this model, which builds on the one presented in Johnson (1997b), an exemplar memory retains the auditory/phonetic details of linguistic episodes associated with both words and gender. The computer simulation thus takes as input speech sound files, encodes them as auditory spectrograms in an exemplar-based memory system. Similarity between input speech and exemplars determines the activation of each of the exemplars in response to the input and then connection weights between exemplars and category nodes feeds activation from exemplars to categories. These operations are formalized in Nosofsky's (1986) Generalized Context Model with three formulas.

Figure 5 about here

$$d_{ij} = \sqrt{\sum (x_i - x_j)^2} \qquad \text{(1) Auditory distance}$$

The auditory distance between two sound files *i* and *j* from auditory spectrograms, x_i and x_j, where *x* is an auditory spectrogram.

$$a_{ij} = e^{-cd_{ij}} \qquad \text{(2) Exemplar activation}$$

The amount of activation on exemplar *i* caused by input token *j* is an exponential function of the auditory distance between the exemplar and the input token.

$$E_{kj} = E_{kj} + \sum a_{ij} w_{ki} \qquad \text{(3) Category activation/evidence}$$

The evidence that input token *j* is an example of category *k* is then a sum of the activations of all of the exemplars of category *k*. In the simulations here, the weights *w* between exemplars and categories are equal to 1 if the exemplar was categorized *k* and equal to 0 if the exemplar was not categorized *k* (see Kruschke, 1992 for a back-propagation method of weight training in an exemplar-based model of memory). Initially E_{kj} is set to 0 in these simulations, though one could imagine a scheme in which the initial evidence would reflect topdown expectations. Evidence accumulates over several cycles of resonance, as will be discussed below.

This type of model performs speech perception without speaker normalization (Johnson, 1997a).

Incoming speech is not normalized, removing talker characteristics, prior to matching with abstract phonetic or phonological representations stored in long-term memory. Instead, talker characteristics are retained in long term memory in the set of remembered exemplars. In this way, the system accommodates the "social construction of gender" (Caplan, 1987) where the normalization model does not, because the representation of "male" or "female" can be dictated by the exemplars of voices rather than by a universal vocal tract normalization algorithm.

In exemplar-based modeling we have to decide what a linguistic experience is, because exemplars are "experienced instances" of language. This is essentially a question of consciousness. Searle (1998) suggests that non conscious brain states or events are not experiences. That is to say, I do not "experience" neurotransmitter release. Nor do I experience my belief that Roh Moo-hyun is president of South Korea while I am sleeping. Searle doesn't identify a neural mechanism that derives this result, but Edelman (1987) does. For Edelman, conscious experience is generated by the interaction between neuronal maps - for instance interactions between the auditory cortex and frontal lobes may underlie generation of linguistic experiences. Thus, I choose to treat "words" as exemplars because words lie at the intersection of form and meaning and thus generate coordinated patterns of activity in both sensory and higher level areas of cognition. Following Searle's manner of argumentation we can note that when people talk to each other about language we generally talk about words. For instance, the way that naive speakers discuss high vowel tensing in Southern Ohio ([puʃ] *[pʊʃ] "push" and [fiʃ] *[fɪʃ]) is in terms of words not sounds - "you say [puʃ] instead of [pʊʃ]?" instead of "you say [u] instead of [ʊ]?".

A related issue is the fact that not all experiences of a word have the same impact on speech perception. This relates to the logarithmic influence of word frequency in speech perception during language acquisition as in the DRIBBLER exemplar-based model (Morgan et al., 2001; Anderson, et al., 2003, see also Jusczyk, 1993). Searle's (1998) view of experience is relevant on this point as well, because for him the "strength" of an experience in memory is related to the attention that one gives to it during processing. This is an important consideration in any theory of exemplar-based speech processing, but I have not addressed it at all in the modeling work done for this paper. It isn't hard to imagine that attention to particular exemplars could have an important effect on stereotype formation

and on the representation of lexical reduction in conversational speech. Unfortunately though, the necessary background knowledge that might guide exemplar weighting decisions is not available, so the work done here proceeds giving all exemplars equal weight in perception.

Still, the word categories in the model illustrated in Figure 5 are a fudge. For instance, the node labeled "see" is shorthand for all of the non-auditory aspects of the linguistic episodes that include mention of this word. This could include some specifically linguistic information such as gestural intentions of the speaker for self-spoken instances of "see", or the orthographic representation seen in episodes of reading along with someone. Non-auditory aspects of linguistic episodes might also include observed gestures (such as the observation of someone pointing while saying "see?") or other aspects of the visual or conceptual scene experienced simultaneously with the auditory image. Ultimately the conception envisioned is that of Edelman (1987), that the mental representation of a word emerges from correlated multimodal experiences of the word. The correlations/generalizations do not need to be, and perhaps cannot be, abstracted and stored separately from the episodic memory.

However, in practical terms, it is desirable to permit a single node in the simulation to represent this collection of non-auditory episodic experience so the model can be focussed on auditory/perceptual emergent generalizations.

$$a_i = a_i + w_{ki}(E_{kj} / n) \qquad \text{(4) Resonance}$$

Resonance is added to the model by feeding activation down from the category nodes to the exemplars. Whether an exemplar *i* receives increased activation from category node *k* is determined by the weight w between the exemplar and the category *w* (formula 4). The effect of this operation is to spread activation from one activated exemplar to all other exemplars that share category memberships with the activated one. This will then tend to produce a blurring of auditory details, and will tend to centralize the response of the model so that perception is more categorical than it would have otherwise been.

4.3. Simulating results for stereotypical and nonstereotypical talkers.

This simulation used the 960 (20 talkers * 48 words) sound files that were recorded for Strand (2000). To test the perception of the stereotypical female talker, for example, the set of exemplars included the other 19 talkers and the response of the model was then calculated for the 48 words produced by talker F5 (the stereotypical female). This was repeated for each of the four talkers. Table II shows the percent words correctly identified and the percent correct gender identification for each talker in these simulations.

Table II about here

The gender of the nonstereotypical female was usually incorrectly given as "male" - for over 80% of the words - and the nonstereotypical male's words were often misidentified. So the performance of the model is consistent with listener performance, but at much lower rates of correct identification (especially for the gender identification of the stereotypical male talker. Evidently though, this model is not picking up on what it is that makes speaker M5 stereotypical sounding. All that the model knows about male talkers is given in the 432 tokens of male speech given to it during training.

The behavior of the exemplar-resonance model can be illustrated by looking at performance on one of the words in the data set. Figure 6 shows the perceptual space for the word "case". This Figure was generated by taking exemplar activations for every sound file compared with every other sound file. The resulting similarity/activation matrix was then visualized using multidimensional scaling with two dimensions. In this perceptual space male talkers are separated from female talkers (as indicated by the hand-drawn line in the Figure. Interestingly, the nonstereotypical talkers appear close to the line dividing male and female talkers.

Figure 6 about here

Now, if we remove talker F5 from the exemplar space and then present her token of "case" to the model we get the exemplar activation pattern shown in the upper left graph in Figure 7. The location of talker F5 is indicated by the black dot and the magnitude of each exemplar's activation in response to this token is indicated by the size of the plotting symbol. Diamonds show the initial exemplar activations and circles show exemplar activations after three resonance cycles. When the circle is bigger than the diamond this indicates that relative exemplar activation is increasing due to resonance and when the diamond is bigger activation is decreasing. In the case of the stereotypical female, the most active exemplars are of female talkers, and with resonance the activation of male exemplars decreases. A similar pattern was found for the stereotypical male talker.

In the case of the nonstereotypical female talker the most active exemplars are examples of male talkers, and with resonance this pattern only increases. This token was incorrectly identified as "male". The situation with the nonstereotypical male talker is different. The most active exemplar in the map is a token of a woman, but the sum of exemplar activations is slightly greater for male exemplars. Thus, the gender is correctly identified, and the activation of male exemplars increases and female exemplars decreases with resonance, but the overall response to this talker is rather lower than to the other talkers.

Figure 7 about here

---

5. Conclusions

This paper has focussed on the perceptual representation of talker gender, arguing that the language-specificity of male and female vowel formant frequencies (independent of presumed vocal tract length) indicates that talkers perform gender. This suggestion, that talkers perform gender, was extended by data showing that listeners respond more slowly in a word recognition task to talkers who sound less stereotypically male or female. This suggests that listeners' expectations of how men and women

should sound has an impact on auditory word recognition. Finally, we explored some aspects of an exemplar-based model of speech perception that performs "speech perception without talker normalization" (Johnson, 1997a) and found that this model is also sensitive to the nonstereotypicality of talkers. Overall, the suggestion of this paper is that an exemplar-based approach incorporating a mechanism of reentrant mapping, or resonance, may be a fruitful line of investigation.

We can see that for social identity - what is sounds like to be male or female - resonance leads an ambiguous item to be "drawn in" to one category or another, so that listeners may be convinced of the gender of a voice even when that the phonetic characteristics of the speech token are not clearly male or female. Interestingly, the power of suggestion can reverse perceived gender (Johnson, Strand & D'Imperio, 1999). For instance, the actor who voices the popular cartoon character "Bart Simpson" is a woman. With eyes open, watching the "male" cartoon character the voice sounds male, but with eyes closed and reminded of the fact that the actor is female the voice sounds female. Much of animation relies on the perceptual strength of just such suggested realities.

The resonance mechanism proposed here to account for the categoriality of perceived gender also operates to enforce categoriality in speech perception, so that ambiguous stimuli get "pulled in" to existing word categories. The application of this idea to an account of "underlying" unity in allophonic and morphophonemic variation is only beginning to be explored, but perhaps as we posit that gender may be a category that emerges from experienced instances of male and female talkers, so also abstract phonological patterns of relation may emerge from resonance among exemplars that are related to each other by semantic ties. Just how far one could take such notions in the study of language sound patterns is a topic currently being explored.

Acknowledgments. Thanks to Robert Remez and David Pisoni, who commented on the vowel formant difference data in Johnson (2005), to Liz Strand for permitting me to present some of her dissertation data here, to Robert Kirchner for insightful discussions on exemplar-based modeling, to the audience at the 2005 Trilateral Phonology Weekend at Stanford for comments, and to the editors and three reviewers for their very helpful comments.

References.

Adank, P., van Hout, R. and Smits, R. (2004). An acoustic description of the vowels of Northern and Southern standard Dutch. *Journal of the Acoustical Society of America 116(3)*, 1729-1738.

Anderson, J.L., Morgan, J.L., White, K.S. (2003). A Statistical Basis for Speech Sound Discrimination. *Language and Speech 46(2-3)*, 155-182.

Ásta Svavarsdóttir, Halldór Ármann Sigurdsson, Sigurdur Jónsson and Sigurdur Konrádsson (1982). Formendur íslenskra einhljóda: Methaltídni og tídnidreifing. *Íslenskt Mál 4*, 63-85.

Behne, D., Moxness, B. and Nyland, A. (1996). Acoustic-phonetic evidence of vowel quantity and quality in Norwegian. *TMH-QPSR 2/1996*, 13-16.

Bladon, R.A.W., Henton, C.G. & Pickering, J.B. (1984). Towards an auditory theory of speaker normalization. *Language & Communication 4(1)*, 59-69.

Caplan, P., Ed. (1987) *The Cultural Construction of Sexuality*. London: Routledge.

Carpenter, G.A. and Grossberg, S. (1987). ART 2: Stable self-organization of pattern recognition codes for analog input patterns. *Applied Optics 26*, 4919-4930.

Cherng, C-H., Wong, C.-S., Hsu, C.-H., Ho, S.-T. (2002). Airway length in adults: Estimation of the optimal endotracheal tube length for orotracheal intubation. *Journal of Clinical Anesthesia 14*, 271-274.

Choi, J.D. (1991). An acoustic study of Kabardian vowels. Journal of the International Phonetic Association. 21 (1) 4-12.

Choi, J.D. and Keating, P. (1991). Vowel-to-vowel coarticulation in three Slavic languages. *UCLA Working Papers in Phonetics 78*, 78-86.

Choi, J.D. (1992). Phonetic underspecification and target interpolation: An acoustic study of Marshallese vowel allophony. *UCLA Working Papers in Phonetics 82,* 1-133.

Choi, J.D. (1995). An acoustic-phonetic underspecification account of Marshallese vowel allophony. *Journal of Phonetics 23* (3), 323-347.

Cosi, P., Ferrero, F.E. and Vagges, K. (1995). Rappresentazioni acustiche e uditive delle vocali italiane", in *Atti del XXIII Cogresso Nazionale AIA*, Bologna,12-14 Settembre, 1995, pp. 151-156.

Easton & Bauer (2000). An Acoustic Study of the Vowels of New Zealand English, *Australian Journal of Linguistics 20 (2)*, 93-117.

Edelman, G.M. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

Eklund, I. and Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica 54*, 1-21.

Fant, G. (1975). Non-uniform vowel normalization. *STL-QPSR 2-3/1975*, 1-19.

Fischer-Jørgensen, E. (1972). Formant frequencies of long and short Danish vowels. In E.S. Firchow et al. (eds.) *Studies for Einar Haugen*. The Hague: Mouton, pp. 189-213.

Fowler, C.A. (1986). An event approach to the study of speech perception. *Journal of Phonetics*, **14**, 3-28.

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics 32* (2), 277-287.

Grossberg, S., Boardman, I., and Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology, Human Perception and Performance 23 (2)*, 481-503.

Hagiwara, R. (1995). Acoustic realizations of American /r/ as produced by women and men, *UCLA Working Papers in Phonetics 90*, 1-187.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review ,93* , 411-428.

Jassem, W. (1968). Vowel formant frequencies as cues to speaker discrimination. *Speech Analysis and Synthesis 1*, 9-41. Warsaw: Panstwowe Wydawnictwo Naukowe. Institute of Fundamental Research, Polish Academy of Sciences.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust.*

Soc. Am. 88, 642-654.

Johnson, K. (1997a). Speech perception without speaker normalization: an exemplar model. In K. Johnson and J.W. Mullennix (eds.) *Talker Variability in Speech Processing*. San Diego: Academic Press (pp. 145-166).

Johnson, K. (1997b). The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics 50*, 101-113, Columbus, Ohio.

Johnson, K. (2005). Speaker normalization. In Remez, R. and Pisoni, D.B. (Eds.) *The Handbook of Speech Perception*.

Johnson, K., Strand, E.A. and D'Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics 27*, 359-384.

Jusczyk, P.W. (1993). From general to language-specific capacities - The WRAPSA model of how speech-perception develops. *J. Phonetics* 21 (1-2): 3-28.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Liénard, J.S. and Di Benedetto, M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America, 106(1)*, 411-422.

Morgan, J. L., Singh, L., Bortfeld, H., Rathbun, K., & White, K. (2001). *Effects of speech and sentence position on infant word recognition*. Paper presented at the Boston University Conference on Language Development, Boston, MA.

Most, T., Amir, O. & Tobin, Y. (2000). The Hebrew vowel system: Raw and normalized acoustic data. *Language & Speech 43*, 295-308.

Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*, 39-57.

Pätzold, M. und A. P. Simpson (1997). Acoustic analysis of German vowels in the *Kiel Corpus of Read Speech*. In A. P. Simpson, K. J., Kohler und T. Rettstadt (Hrsg.), *The Kiel Corpus of Read/Spontaneous Speech — Acoustic data base, processing tools and analysis results*, AIPUK 32, pp. 215–247.

Schroeder, M. R., Atal, B. S. and Hall, J. L., (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America* 66 (6), 1647-1652.

Searle, J. (1998). *Mind, Language and Society*. NY: Basic.

Strand, E.A. (2000). *Gender Stereotype Effects in Speech Processing*. PhD Dissertation, Ohio State University.

Sussman, H.M. (1986). A neuronal model of vowel normalization and representation. *Brain Lang. 28,* 12-23.

Tolonen, H., Kuulasmaa, K., and Ruokokoski, E. (2000). *MONICA Population Survey Data Book* (http://www.ktl.fi/publications/monica/surveydb/title.htm), World Health Organization.

Watson, C.I., Harrington, J. and Evans, C. (1998). An acoustic comparison between New Zealand, and Australian English vowels. *Australian Journal of Linguistics 18(2)*, 185-207.

Weeks, J., Ed. (1989) *Sex, Politics and Society*. 2nd Ed. London: Longman Group.

Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Boston: Houghton Mifflin.

Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics 24 (1)*, 245-261.

Appendix 1. List of references and vowels included in the cross-linguistic vowels and body-height survey.

Language	Reference	Vowel
Australian	Watson et al. (1998)	i,e,a,o,u
Belgian Dutch	Adank, et al. (2004)	i,e,a,o,u
California	Hagiwara (1994)	i,e,a,o,u
Danish	Fischer-Jørgenson (1972)	i,e,a,o,u
Dutch	Adank, et al. (2004)	i,e,a,o,u
French	Lienard & Di Benedetto (1999)	i,e,a,o,u
German	Patzold & Simpson(1997)	i,e,a,o,u
Hebrew	Most et al. (2000)	i,e,a,o,u
Hungarian	Tarnoczy (1964)	i,e,a,o,u
Icelandic	Asta Svavarsdottir	i,e,a,o,u
Italian	Cosi, Ferrero, & Vagges	i,e,a,o,u
Korean	Yang (1996)	i,e,a,o,u
New Zealand	Easton & Bauer (2000)	i,e,a,o,u
Norwegian	Behne et al., (1996)	i,ii,a,aa,o,oo
Polish	Jassem (1968)	i,e,a,o,u
Spanish	Gonzalez (pc)	i,e,a,o,u
Swedish	Eklund & Traunmuller (1997)	i,e,a,o,u

Table I. Regression fits (R2) and the slope of the best fitting regression line, for the regression analyses shown in figure 3 (* $p<0.01$, + $p<0.2$).

		R^2	slope
male	F1	0.072	-0.030
	F2	0.023	0.018
	F3	0.162^{+}	-0.023
female	F1	0.029	-0.030
	F2	0.002	-0.007
	F3	0.34^{*}	-0.046
all	F1	0.345^{*}	-0.045
	F2	0.387^{*}	-0.055
	F3	0.790^{*}	-0.595

Table II. Percent words correct and gender correct in simulations of Strand's (2000) auditory word naming experiment

talker	% words correct	% gender correct
F5 - stereotypical female	79	75
F7 - nonstereotypical female	75	17
M5 - stereotypical male	71	54
M8 - nonstereotypical male	50	79

.

Figure Captions

Figure 1. Average vowel formant difference between vowels produced by men and vowels produced by women, for 17 languages.

Figure 2. The difference in vowel formant frequency between male and female talkers on the vertical axis, as a function of the difference in the average body height between men and women in the countries where these languages are spoken on the horizontal axis. (a) Third formant gender differences, (b) second formant gender differences, and (c) first formant gender differences.

Figure 3. The average vowel formant frequency and average height for female (F) or male (M) speakers of the 17 languages shown in figures 1 and 2. The lowest trend line shows the linear regression fit between the first formant F1 and average height. The middle trend line shows the regression fit between F2 and height, and the highest line shows the fit between F3 and height. Separate fits for male and female F2 frequency are drawn with solid lines.

Figure 4. Results of the speeded word naming task (Experiment 5) in Strand (2000).

Figure 5. Schematic illustration of an exemplar resonance model of speech perception. An incoming utterance is encoded by the auditory system as an auditory spectrogram, which is then compared with a large set of exemplars of similar auditory images. Activation from the exemplars feeds up to linguistic categories and to gender categories. Reentrant loops feed activation back down to the exemplar store setting up a resonance in the system.

Figure 6. An auditory perceptual map of Strand's (2000) "case" stimuli. The stereotypical talkers are outlined with squares and the nonstereotypical talkers are outlined with ovals.

Figure 7. The growth of exemplar activation over three cycles of resonance in response to tokens of Strand's (2000) stereotypical and nonstereotypical female and male voices saying "case".

Figure 1

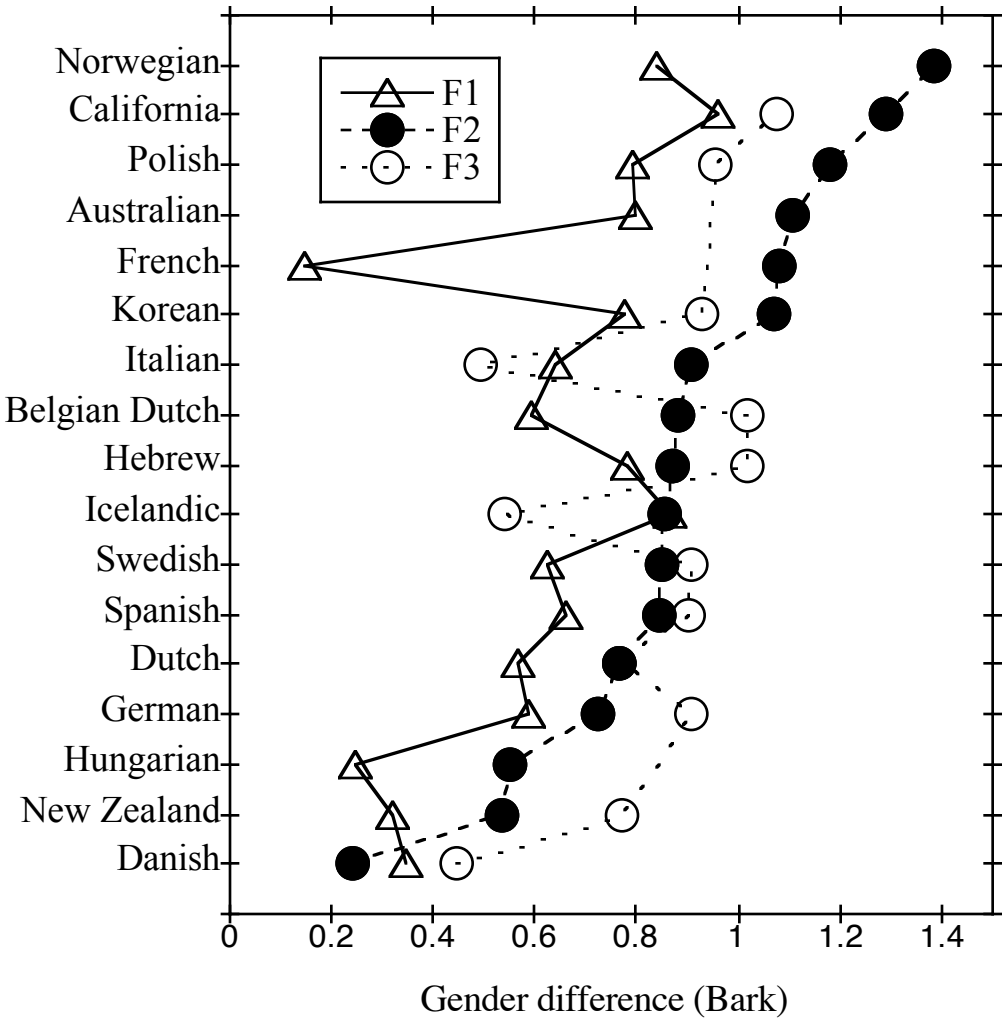

Norwegian
California
Polish
Australian
French
Korean
Italian
Belgian Dutch
Hebrew
Icelandic
Swedish
Spanish
Dutch
German
Hungarian
New Zealand
Danish
F1
F2
F3
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Gender difference (Bark)

Figure 2(a)

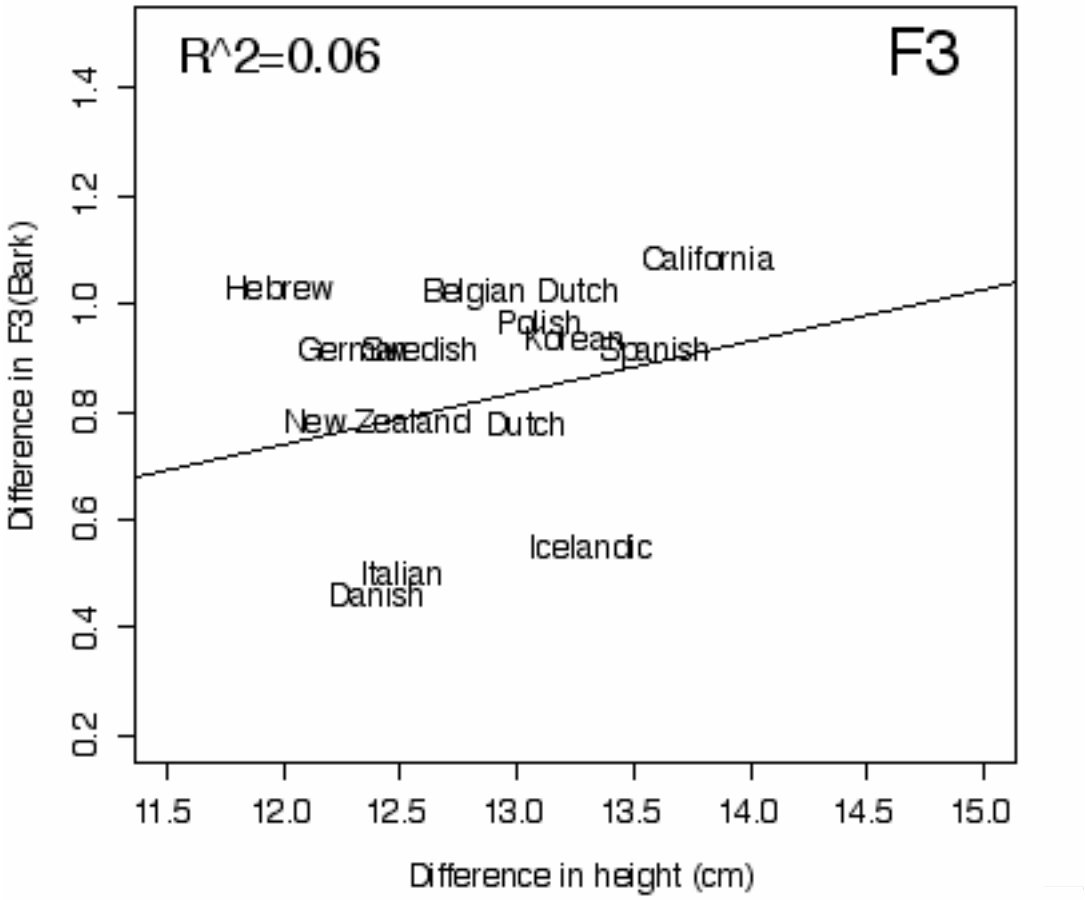

R^2=0.06
F3
Difference in F3 (Bark)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0
Difference in height (cm)
California
Hebrew
Belgian
Dutch
Polish
Korean
Spanish
Swedish
New Zealand
Dutch
Icelandic
Italian
Danish

Figure 2(b)

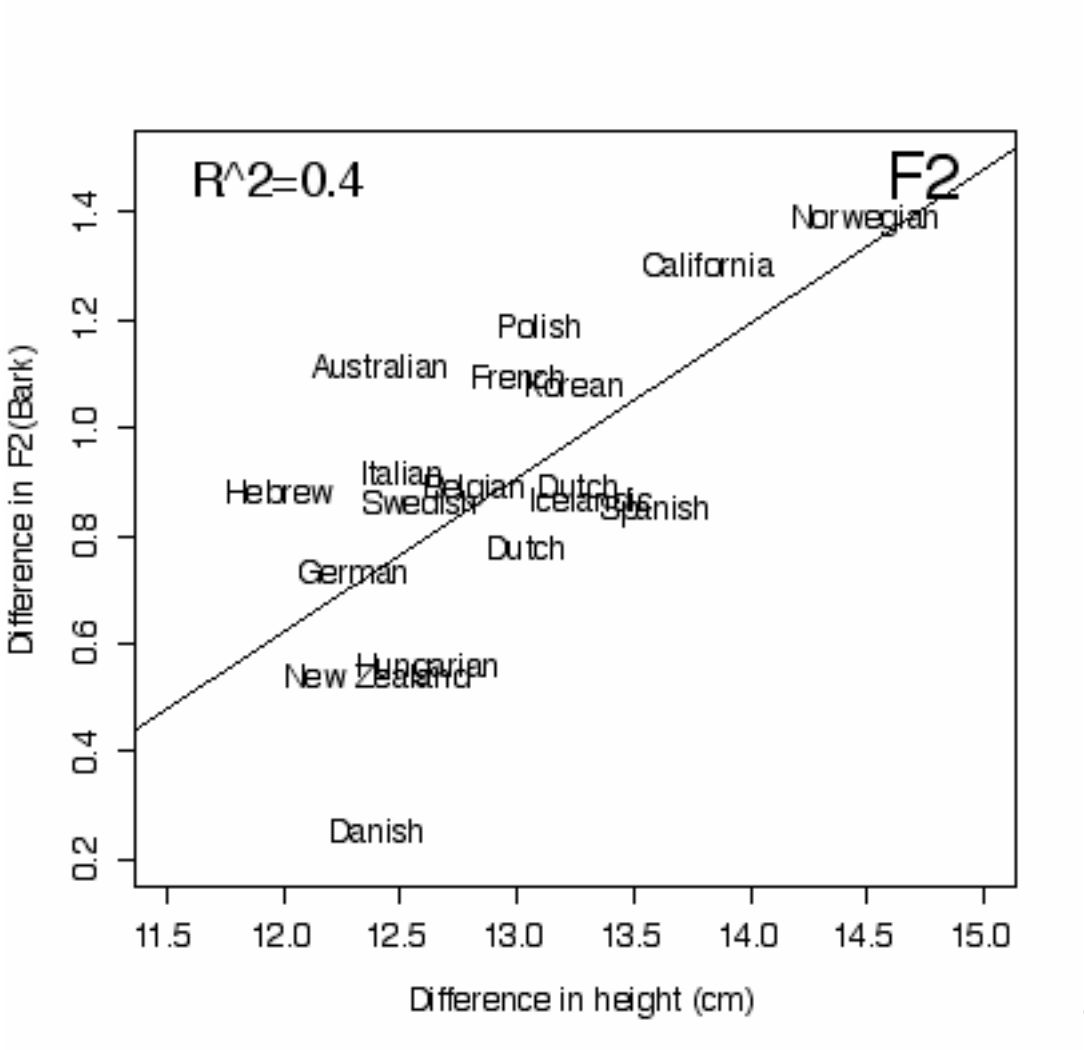

R^2=0.4
F2
Norwegian
California
Polish
Australian
French
Korean
Italian
Hebrew
Swedish
Dutch
Spanish
Dutch
German
Hungarian
Danish
Difference in F2(Bark)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0
Difference in height (cm)

Figure 2(c)

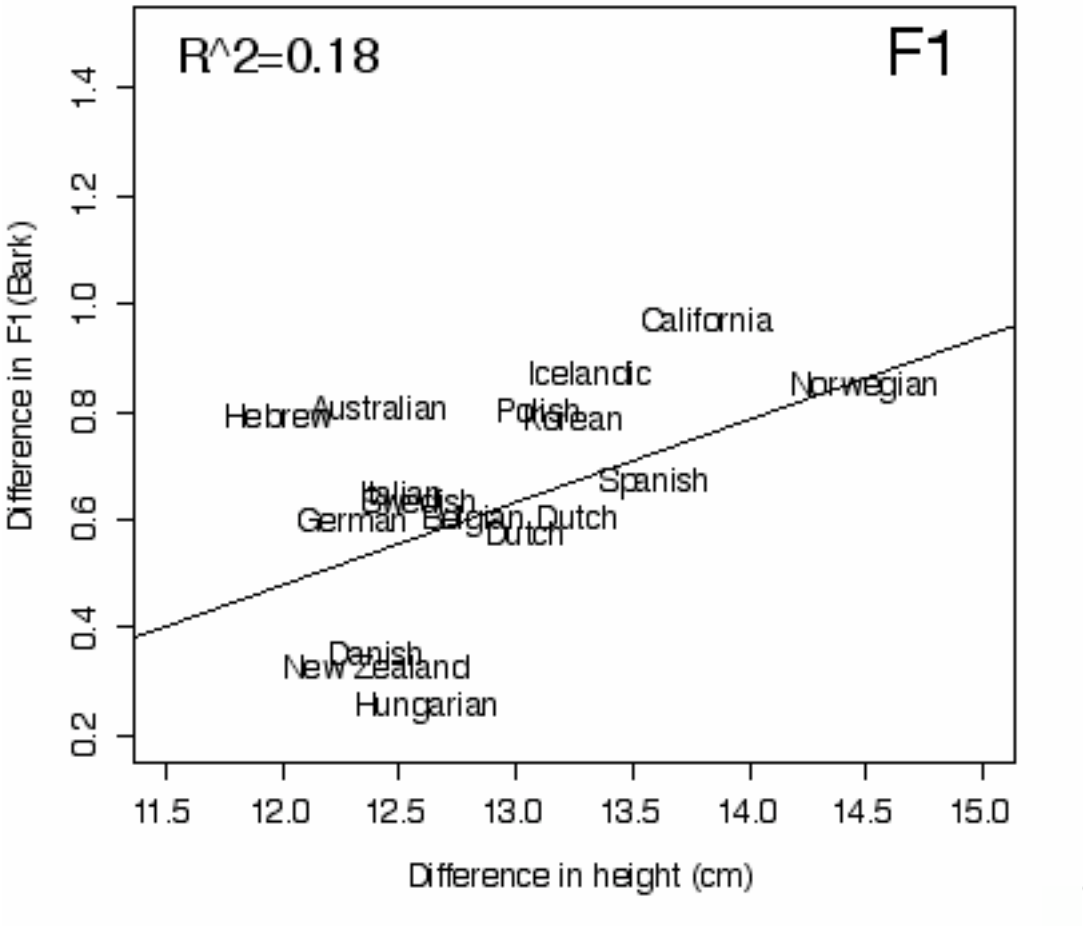

R^2=0.18
F1
Difference in F1 (Bark)
0.2
0.4
0.6
0.8
1.0
1.2
1.4
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0
Difference in height (cm)
California
Icelandic
Norwegian
Hebrew
Australian
Polish
Korean
Spanish
German
Dutch
Danish
New Zealand
Hungarian

Figure 3

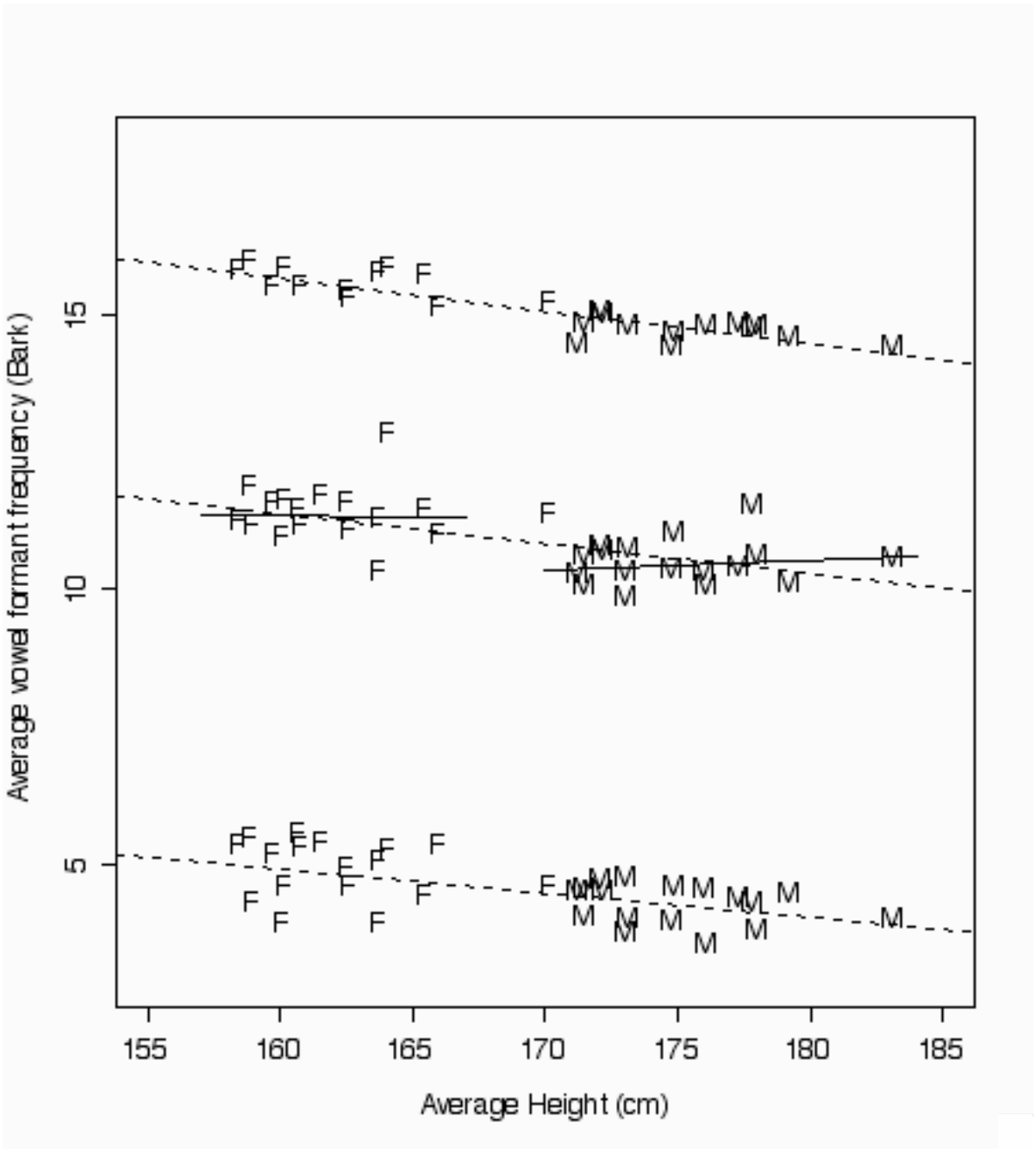

Average vowel formant frequency (Bark)
5
10
15
155
160
165
170
175
180
185
Average Height (cm)

Figure 4

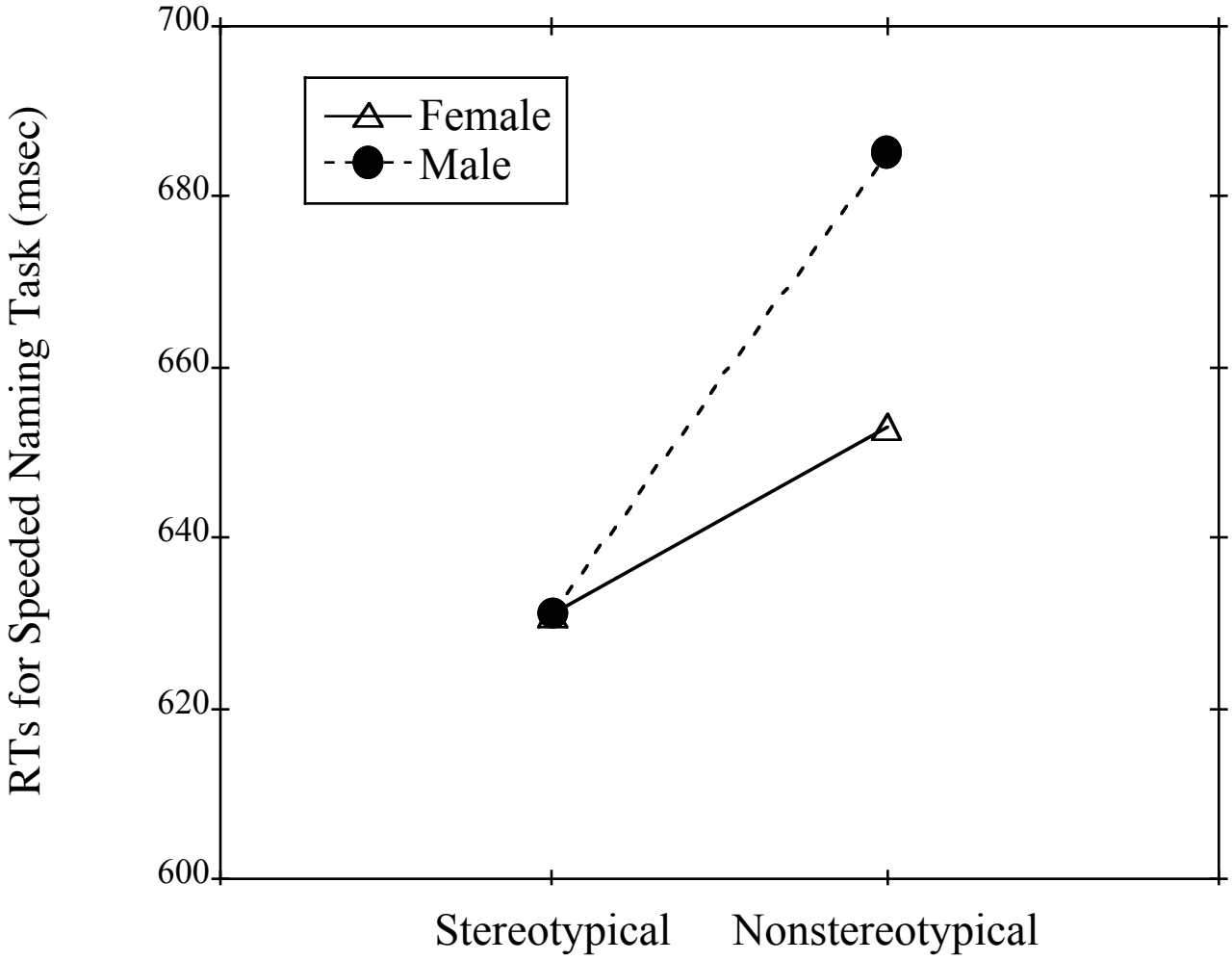

700
680
660
640
620
600
RTs for Speeded Naming Task (msec)
Female
Male
Stereotypical
Nonstereotypical

Figure 5

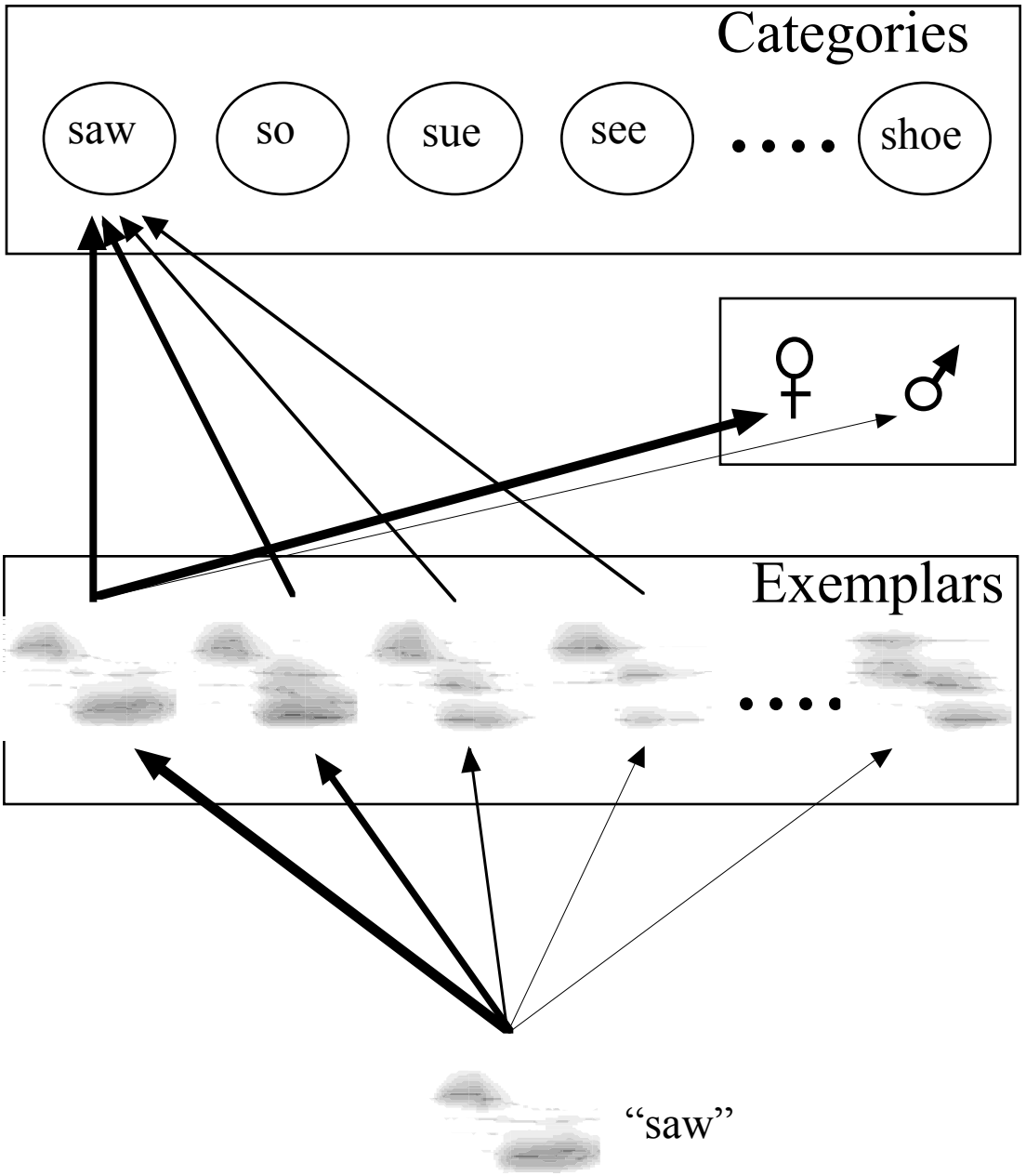


Figure 6

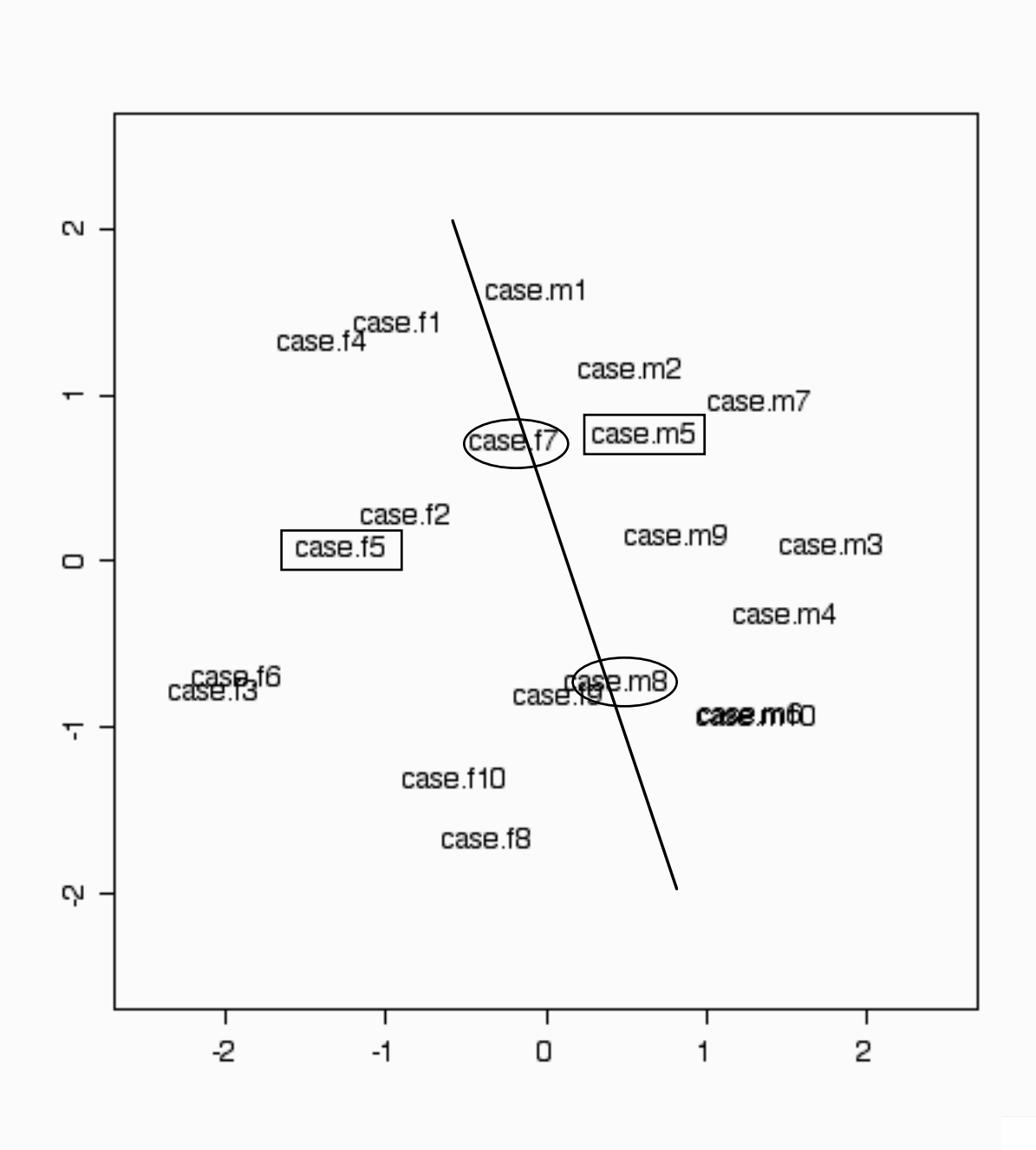
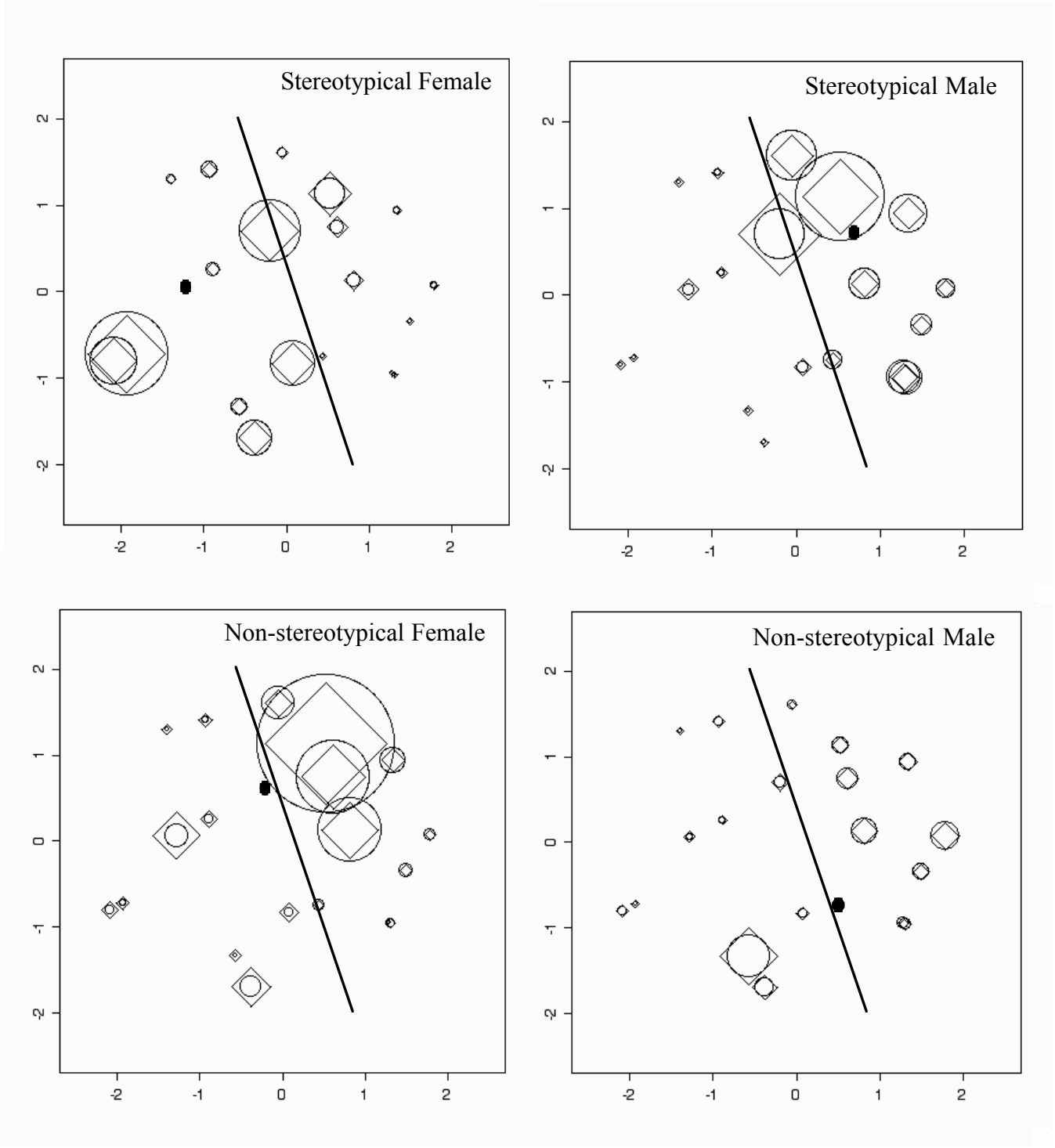

case.m1
case.f1
case.f4
case.m2
case.m7
case.f7
case.m5
case.f2
case.f5
case.m9
case.m3
case.m4
case.f6
case.f3
case.m8
case.m6
case.f10
case.f8
2
1
0
-1
-2
-2
-1
0
1
2

Figure 7


Stereotypical Female
Stereotypical Male
Non-stereotypical Female
Non-stereotypical Male