# The influence of social network size on speech perception

Shiri Lev-Ari

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Royal Holloway University of London, Egham, UK

Key words: social network; individual differences; speech perception.

* I would like to thank Shawn Bird for assisting with the programming of the simulations in Study 2.

Address correspondence to:

Shiri Lev-Ari

Royal Holloway University of London

Egham TW20 0EX

United Kingdom

Tel: 44-1784276547

shiri.levari@rhul.ac.uk

Abstract

Infants and adults learn new phonological varieties better when exposed to multiple rather than a single speaker. This paper tests whether having a larger social network similarly facilitates phonological performance. Experiment 1 shows that people with larger social networks are better at vowel perception in noise, indicating that the benefit of lab exposure to multiple speakers extends to real life experience and to adults tested in their native language. Furthermore, the Experiment shows that this association is not due to differences in amount of input or to cognitive differences between people with different social network sizes. Follow up computational simulations reveal that the benefit of larger social networks is mostly due to increased input variability. Additionally, the simulations show that the boost that larger social networks provide is independent of amount of input received but is larger the more heterogeneous the population is. Lastly, a comparison of "adult" and "child" simulations reconciles previous conflicting findings by suggesting that input variability along the relevant dimension might be less useful at the earliest stages of learning. Together, this paper shows when and how the size of our social network influences our speech perception. It thus shows how aspects of our life-style can influence our linguistic performance.

The speech signal is inherently variable, and lacks one-to-one mapping. That is, one person's /ɒ/ (as in *hot*) can be someone else's /ɔ/ (as in *caught*). In general, phonemes' articulation varies according to their phonological context, the speech style, the identity of the speaker and so forth. While, for the most part we seem to process speech flawlessly despite this lack of invariance, individual differences exist.. So what makes us better or worse at interpreting speech? As any person who tried to learn a second language knows, experience matters. But experience is not only the amount of input one receives but also its nature. This paper will show how differences in our social networks influence our speech perception by influencing the nature of the input we receive.

People differ in their social networks. For example, Hill and Dunbar (2003) found that some people send Christmas cards to fewer than 25 people while others send Christmas cards to more than 350 people. This paper takes a statistical perspective and tests how interacting with more people influences the nature of the linguistic input one receives, and consequently, one's success in speech perception. In previous work, I have found that having a larger social network improves global comprehension of novel speakers, , as reflected in better comprehension of restaurant and product reviews, and that this effect is causal (Lev-Ari, 2016). In general, people learn language from their environment. Furthermore, an integral part of language learning is achieved via statistical learning. For example, infants are sensitive to phonological transitional probabilities, and use them for speech segmentation (e.g., Saffran, Aslin & Newport, 1996). Similarly, transitional probabilities between words are argued to be used in grammatical acquisition (Thompson & Newport, 2007). The distributional nature of the input can also influence not only rate of acquisition (e.g., Huttenlocher, Haight, Bryk,

Seltzer & Lyons, 1991; Vosoughi, Roy, Frank & Roy, 2010) but also the number of categories one develops and their boundaries. Thus, Maye and colleagues (2002) showed that infants develop two phonological categories, /d/ and /t/, if they are exposed to bi-modal distribution of phones along the Voice Onset Time (VOT) continuum[1], but they develop a single category collapsed over both phonemes if they are exposed to a uniform distribution of these phones.

A key aspect of the input that has been argued to facilitate phonological acquisition is its variability. For example, Lively, Logan and Pisoni (1993) have shown that Japanese speakers, whose native language does not have two distinct categories for /l/ and /ɹ/, are more successful at learning to identify these two English phonemes if they are trained by listening to productions from five speakers rather than a single one, despite not receiving more input from the multiple speakers. This finding has consequently led L2 training on perception and production to habitually use a High Variability Phonetic Training paradigm, in which phonetic contrasts are presented by multiple speakers and in multiple phonetic contexts. Similarly, adaptation to foreign-accented speech improves more with exposure to more speakers. For instance, listening to English speech from four Chinese-accented speakers rather than only one improves one's ability to understand novel Chinese-accented speakers, even when the amount of input is held constant (Bradlow & Bent, 2008). First language acquisition is also better with exposure to multiple rather than a single speaker. Thus, 14-months old infants have been shown to struggle at perceiving /buk/ and /puk/ as two different words, suggesting they do not perceive /b/ and /p/ to be two different phonemes (e.g., Rost &

[1] Voice Onset Time is the time distance between the the release of the consonant (the burst) and the beginning of voicing. It is a feature that contrasts voiced and voiceless stops, such as /d/ and /t/ in English.

McMurray, 2009; Stager & Werker, 1997). Yet when exposure consists of productions of /buk/ or /puk/ from 18 speakers and not only a single speaker, they succeed at differentiating the two words, even though the amount of exposure is identical across conditions (Rost & McMurray, 2009). The benefit that listening to multiple speakers confers is argued to be due to the greater variability in input from multiple speakers than a single speaker. In line with this argument, Sumner (2011) has shown that exposure to multiple tokens from a single speaker also leads to greater adaptation to that speaker than listening to a single token of that speaker for the same number of times. Interestingly, acoustic variability seems to boost not only acquisition at the phonological level, but also vocabulary learning. For example, Barcroft and Sommers (2005) found that English speakers were better at acquiring new Spanish words the more speakers they heard produce these words, even when the total amount of exposure was held constant. Similarly, learning was better when the words were produced in multiple rather than a few or a single speech style (e.g., neutral, excited, whispered). Even in the visual domain, learning of categories has been shown to be better when the input in exposure is noisier, and therefore more variable (Posner & Keele, 1968). Taken together, the literature suggests that first language acquisition, second language acquisition, and even acquisition of visual categories, despite differing in many components of their underlying mechanism, are all influenced by the same statistical principle.

But how does variability in input improve learning? Rost and McMurray (2010) have investigated this by, in one study, systematically varying only the critical feature relevant for categorization, VOT, while holding the rest constant, and, in another study, varying all other irrelevant aspects of the speech (e.g., prosody) but keeping the VOT

constant. They found that it was variation along the irrelevant aspects that facilitated learning. According to their findings, variability along the irrelevant aspects allows learners to understand which aspects of the input are relevant for categorization and which ones are not (but see Iverson, Hazan & Bannister, 2005 for evidence that in second language acquisition the distributional patterns in the input might be harder to extract or apply). The same argument has been put forward in the visual domain, where the noise in the input has been argued to facilitate learning by enabling learners realize which aspects of the input are constant within the category and which aspects are allowed to vary within the category (Posner & Keele, 1968). Another possibility that has been raised is that greater variability in the input ensures that more of the sound space is sampled, increasing the odds that upon hearing a new token, the listener has an existing representation to match it to (Sumner, 2011). In Sumner's study, stimuli varied along the critical feature, VOT, and variability influenced learning. One difference between the two proposals is that the former mostly regards the acquisition of new categories, whereas the latter tries to explain tuning of existing categories. It might therefore be the case that different types of variability are useful at different stages of learning. Other mechanisms that have been proposed, but will not be discussed here at length, include the proposal that variability encourages the learner to generalize because it renders it impossible to learn all tokens (Gómez, 2002), and the suggestion that variability boosts learning by increasing the number of connections each type has (Barcroft & sommers, 2005).

It might be worth mentioning that while input variability has been shown to have a facilitatory effect on learning, it makes processing and identification more challenging. That is, processing input from multiple talkers with unpredictable talker

switches leads to poorer identification than processing the same words from a single talker (e.g., Pisoni, 1993). The reason for this detrimental effect is similar to the reason that exposure to multiple talkers is beneficial in the long run – the greater variability. Because speakers differ from one another in the way they produce speech, listeners need to adjust to every new talker, and use the neighboring linguistic context and knowledge about the identity of the speaker to disambiguate and identify the phonemes. Such talker differences, however, as mentioned beforehand, are important for the formation of robust representations. Therefore, input variability might exert additional challenges during processing, but this challenge will improve learning in the long-run.

The goal of this paper is, first, to examine whether having a larger social network, defined here as regularly interacting with more people, leads to better speech perception, in the same way that exposure to multiple speakers facilitates phonological acquisition. This is achieved by testing speech perception skills of people with different social network sizes. At a second stage, this paper uses computational simulations to explore the mechanism by which such an effect can come about, as well as its interactions with other network properties, and its dependence on the stage of learning. These simulations show how network size influences the distributional nature of the input that we receive, and how those changes influence phonological categorization. The simulations additionally show that the same distributional properties can improve performance when the phonological categories are already known, but not at the earlier stage of learning, when the learner still needs to figure out how many categories there are.

## Experiment 1

The goal of Experiment 1 is to test whether individuals who regularly interact with more people are better at speech perception, and in particular, at understanding vowels in noise. Success at identifying vowels in noise is one measure that reflects the robustness of one's vowel categories representations. The decision to focus on vowel perception was due to the fact that even though variation exists at all levels, research shows that variation is much greater across vowels than it is across consonants (Kleinschmidt, 2016), and even more importantly, that variation for vowels is structured by indexical factors, whereas other types of variation, such as for VOTs in stops, is not (Allen, Miller & DeSteno, 2003; Kleinschmidt, 2016). Correspondingly, while past research on vowel production showed its dependence on indexical properties, past research on variability in consonant production has mostly shown its dependence on phonetic context and speech style. For example, vowel production has been shown to be influenced by sex, vocal tract size and shape, and dialect (e.g., Bachorowski & Owren, 1999; Peterson & Barney, 1952). In contrast, Allen et al. (2003) discovered that sex differences in VOTs are eliminated once speech rate is controlled for, and Kleinschmidt (2016) similarly found that indexical properties did not predict VOT production yet did account for variation in vowel production. Exposure to multiple speakers might therefore increase input variability for vowels more than for consonants, and importantly, it will allow listeners to learn the conditioning of this variability, and thus assist in perception of vowels by new speakers. Vowels were embedded in noise, since all participants were adult native speakers, and are therefore expected to perform at ceiling in ideal conditions. Embedding speech in noise is a common practice to test more fine grained differences between participants (e.g. Sidaras, Alexander & Nygaard, 2009).

One potential problem is that people who differ in their social network size might also differ in their cognitive skills, and these cognitive differences might influence speech perception skills. Therefore, all participants were also tested on a host of cognitive measures to ensure that any difference in speech perception performance cannot be explained by cognitive differences.

*Method*

*Participants*. Sixty native Dutch speakers participated for pay. Participants' age ranged from 20 to 57 (M=34; SD=10.6). All reported to have normal hearing.

*Stimuli*. The experiment included a language experience questionnaire from which the main predictors were extracted, a perception of speech in noise task, and four cognitive measures to control for individual differences that might correlate with network size and could influence speech perception (Operation-Span, Auditory Short-Term Memory, Flanker task, Trail making task). Originally, the experiment was designed to test the effect of social network size on two types of skills, the robustness of phonological categorization, as measured by the perception of speech in noise task, and the ability to identify and normalize talkers, as measured by the Coordinate Response Measure, and a multiple talker effect in a phoneme monitoring task. Whereas the test of the robustness of phonological categorization was based on previous literature that shows that exposure to multiple speakers boosts learning of new phonological categories, the tests of talker normalization were more exploratory in nature. Only the results from the speech perception in noise task are described here. The results of the talker normalization tasks did not reveal any effect of social network properties. As it is hard to infer from null results, especially when the test was exploratory, these results are neither discussed in here nor followed up on in the later simulations.

Language experience questionnaire. Before coming to the lab, all participants completed a linguistic experience questionnaire for one week. For seven consecutive typical days (i.e., no holidays, sick days etc.), participants logged in all oral interactions with native speakers that lasted 5 minutes or longer. Participants were instructed to include one-to-one and multi-party face-to-face interactions, as well as phone, Skype and other types of conversations in which interlocutors hear each other. For each interaction, participants listed the identity of the interlocutor, the duration of the interaction, as well as additional qualitative details about the interlocutor (e.g., education, occupation) that were collected for future purposes. Network Size was calculated as the total number of different people with whom participants interacted. Hours of Talk was the sum of all reported hours of interaction. Each interlocutor was counted once regardless of the number and duration of conversations the participant had with them. Participants' social network size ranged from 11 to 74 (M=27; SD=11.7).

Transcription of nonwords on noise. To test the robustness of participants' phonological representations, participants transcribed 120 monosyllabic nonwords in noise. Twenty-three nonwords had a CVC structure, 40 had a CCVC structure, 57 CVCC . Participants were informed that the recordings were of non-words. Nonwords were used to minimize any influence of vocabulary or grammatical knowledge. All nonwords were legal words in Dutch and were taken from Janse and Newman (2013). Nonwords were recorded by a native female Dutch speaker. The amplitude envelope of each recorded nonword was extracted with Praat, and white noise was generated to fit this envelope. Then the original recording was combined with the generated white noise using Audacity, creating a file with a Signal to Noise Ratio of 0. The nonwords were presented in a random order, and participants responded at their own pace. Participants' vowel

recognition was scored. Dutch has transparent orthography, such that each vowel and vowel combination can only refer to one vowel or diphthong. The diphthong /ɛi/ can be written in two different manners, 'ij' and 'ei'. Both were scored as correct. On average, participants transcribed 66% of the vowels correctly (SD=5.7).

Working Memory. Unsworth and colleagues' (2005) Operation Span was used with Dutch instructions. Participants evaluated whether equations were correct. Following each equation, participants received a letter to memorize. Following a stretch of between 3 and 7 equation-letter pairs, participants recalled the memorized letters in the correct order. The time provided for solving each equation was adjusted to participants' pace of solving equations during an initial baseline stage to prevent participants from rehearsing the letters during the task.

Auditory Short Term Memory. To measure participants' auditory Short Term Memory (STM), participants heard 30 sequences of 4 non-musical tones. The first three tones in each sequences appeared at an inter-stimulus onset interval of 750ms, followed by a pause of 1000ms, and then the fourth tone. Participants' task was to determine whether the last tone appeared among the first 3 tones. Twelve different tones were used in total. Participants' auditory STM was scored as the proportion of trials they answered correctly.

Selective Attention. Participants' selective attention was measured with the Flanker task (Eriksen, 1995). Participants saw a string of 5 symbols on the screen. The middle symbol was always a chevron (<,>), and participants' task was to indicate in which direction the chevron pointed. On congruent trials, the chevron was flanked by 4 other chevrons pointing in the same direction. On incongruent trials, the flanking chevrons pointed in the opposite direction. On neutral trials, the chevron was flanked by four

hyphens instead of chevrons. The symbols remained on the screen until participants responded or until 1000ms have elapsed. There were a total of 144 trials. The selective attention score was calculated as the ratio between the Response Times (RTs) on the incongruent trials and the RTs on the neutral trials. Higher scores indicate worse selective attention.

Task Switching. Participants' task switching abilities was measured with Reitan's (1958) Trail making task. On this task participants draw a line to connect 25 circles in a set order. In the baseline condition, participants connect circles labeled with increasing numbers. In the critical condition, participants link circles labeled with increasing numbers and letter in alternating order (i.e., "1", "A", "2" etc.). Task switching score is calculated as the ratio between the completion time for the critical trial and completion time for the baseline trial. Higher score indicates worse task switching ability.

*Procedure*. Participants first completed the language experience questionnaire. They were then invited for a lab session that took about one and a half hours. Participants performed the tasks in the following order: Operation Span, Trail making task, transcription of nonwords in noise, Auditory STM, Coordinate Response Measure, Flanker task, and Phoneme monitoring.

*Results*

First, the relation between participants' cognitive abilities and network size was examined. Three participants were missing one task each, so some of the reported correlations were conducted on 59 or 58 participants, and the general analysis was conducted on 57 participants. None of the cognitive measures correlated with social network size (all *r*s<|0.1|; See Table 1). Therefore, people with different network sizes

do not seem to differ in their general cognitive abilities. Additionally, social network size did not significantly correlate with the number of hours of talk (r=0.13, n.s.).

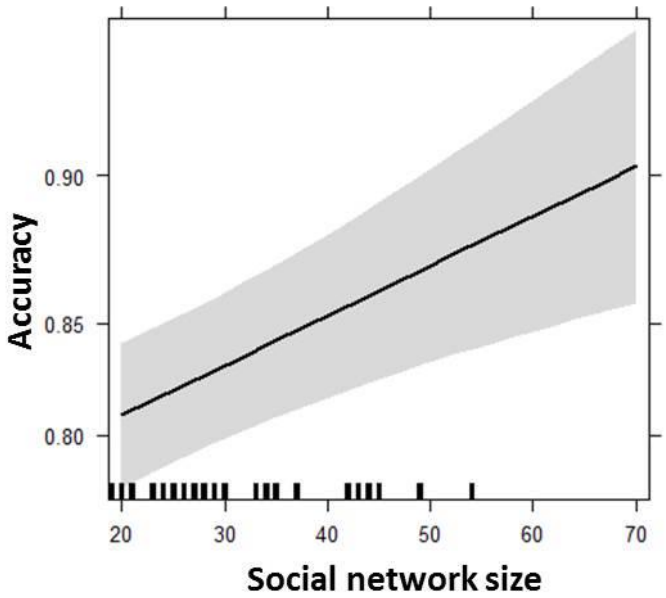



Figure 1. The effect of Social Network Size on vowel perception in Experiment 1. The gray band indicates Standard Error.

| **Cognitive measure** | **Correlation with number of interlocutors** | **Range** | **Mean (SD)** |
|---|---|---|---|
| O-SPAN (Working Memory) | 0.01 | 4-75 | 51.2 (16.37) |
| Auditory STM | 0.07 | 0.43-0.97 | 0.75 (0.12) |
| Flanker task (Selective Attention) | -0.05 | 0.37-1.52 | 1.24 (0.16) |
| Trail making task (Task Switching) | -0.03 | 1.29-4.41 | 2.15 (0.63) |

Table 1. Correlations between the cognitive measures and social network size

To test whether participants' social network size predicted their success at understanding speech in noise, a logistic mixed model analysis with Participants and Items as random variables and Network Size as a fixed factor was run. Despite the lack of correlations between the cognitive measures and Social Network Size, to be

conservative, WM, Auditory STM, Selective Attention, and Task Switching were simultaneously entered into the model as fixed factors. Similarly, to further ensure that any effect of Network Size is not due to greater amount of input, Hours of Talk was also simultaneously entered into the model as a fixed factor. The random structure included intercepts for both random variables as well as slopes for WM, Auditory STM, Selective Attention, Task Switching, and Network Size for the Items variable[2]. Results revealed a significant effect of Network Size ($\beta$=0.02, SE=0.01, z=2.01, p<.05; See Appendix A for the full table of results), such that participants with larger social networks transcribed more vowels correctly[3]. No other effect reached significance[4].

To conclude, the results of Experiment 1 indicate that participants with larger social networks are better at understanding speech in noise, at least in terms of vowel recognition. Importantly, the experiment's results show that this advantage is not due to differences in cognitive abilities between participants who have social networks of different sizes. While it cannot be completely ruled out that there is another factor that was not measured here, that correlates with Social Network Size and is responsible for the effect, such a candidate does not immediately present itself. Similarly, nonwords were used to minimize the influence of any potential differences in linguistic knowledge, but such effects cannot be completely ruled out. Additionally, as with any

[2] A slope for Hours of Talk was not included, because when included, the model failed to converge. Considering that slopes are included to prevent spurious effects, and Hours of Talk did not have a significant effect, its omission does not influence the results.

[3] An identical model using log-transformed Network Size instead of raw Network Size was run to examine whether the effect of network size is logarithmic rather than linear. The effect of Log Network Size was smaller and did not reach significance ($\beta$=0.8, SE=0.51, z=1.60, p=0.11) indicating that the effect of Network Size on performance is linear in nature.

[4] One of the reasons that none of the cognitive measures showed a significant effect is due to the high correlation between the Working Memory measure (O-Span) and Auditory STM (r=0.52, p<0.0001). When each cognitive measure was tested in the absence of others, Auditory STM predicted better transcription of speech in noise ($\beta$=1.7, SE=0.7, z=2.43, p<.02). Social Network Size remained significant in this analysis ($\beta$=0.02, SE=0.01, z=2.24, p<.03).

individual differences study, it cannot be ruled out that the direction of the effect goes in the opposite effect, such that those who are better at understanding speech in noise have larger social networks. Nevertheless, there is no known evidence to suggest such an effect. It seems most likely, then, that, in line with previous work about the facilitatory effect of exposure to multiple speakers on phonological acquisition, having a larger social network improves one's perception in noise. Next, computational simulations were performed to further explore this account.

## Simulation 1

Simulation 1 explores the mechanism underlying the effect of social network size on speech perception. Computational simulations allow an understanding of how social network size changes the nature of the input we receive, and how such changes influence speech perception. Furthermore, the use of computational simulations allows the isolation and crossing of different aspects of the social network that are difficult to isolate and measure in real life. Thus, the computational simulations reported here reveal both how and when social network size, as well as other properties of the network, improve performance. The simulations were run on recognition of Dutch vowels, rendering them maximally similar to the task in Experiment 1. Noise was not modeled in the simulations, as it was only included in Experiment 1 to prevent ceiling effects and allow examination of fine grained differences. As it played no theoretical role, whereas adding it to the simulations would require making different assumptions about how listeners deal with noise, the simulations were of vowel categorization in silence.

### *General method*

The computational simulations used an agent-based model and were run on recognition of Dutch vowels. Each simulation generated a population of 1000 speakers. The linguistic productions in the simulations were sets of 2 formant frequencies, simulating vowel production. The formant values for the population were set according to the averages and standard deviations of Dutch vowels from Adank, van Hout & Smits (2004). Average formant values for each speaker were randomly sampled from this distribution. Networks were then generated by randomly selecting individuals from the population.

Following network generation, meetings between the agent and members of her network were simulated. The agent started out without any tokens of any of the vowels. In each meeting with a member of the network, the agent's interlocutor produced one vowel of each type, by sampling from a distribution centered around the interlocutor's formant means and with a standard deviation of 0.02 of the formant's mean. The agent stored each of these vowels with their appropriate labels[5]. This continued for a pre-defined number of meetings. In the main set of simulations, the number of meetings for agents in both small and large network size conditions was 500. Importantly, in all simulations, the number of meetings was identical across agents in the small and large social network conditions.

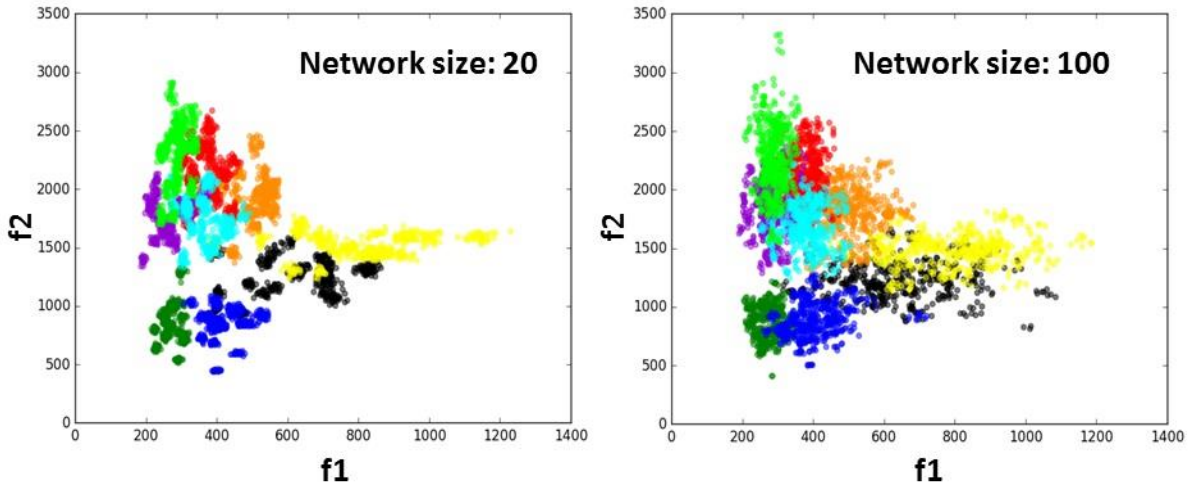
Following the meetings, the agent was tested on recognition of vowels produced by speakers outside of her network. In each test trial, one member of the population that is not in the agent's network was randomly selected. That speaker then randomly

[5] The choice of setting variability to 0.02 of the formants' mean was somewhat arbitrary, due to the lack of large enough corpora that provide information about intra-speaker variability with the same phoneme within the same phonetic context (the simulation ignores variability due to phonetic context and ability to use that information to disambiguate the sound, even though social network size might also improve this ability). Importantly, as reported on p. 19-20, simulations that varied intra-speaker variability showed that the benefit of having a larger social network also extends to both lower and higher levels of intra-speaker variability.

produced one vowel. The agent classified the incoming vowel by calculating the Mahalanobis distance between the vowel's formants and each of the vowel categories in her stored inventory, and labeling it with the label of the vowel category to which it is closest. If correct, the trial was scored as 1, and otherwise, it received a score of 0. In each simulation, the agent was tested on 100 vowels.

*How does network size influence the nature of the input and agent's performance?*

To test whether having a larger social network improves vowel recognition, 100 simulations with a network of 20, and 100 simulations with a network of 100 were run. These network sizes were selected, as they reflect realistic common network sizes located towards the extremes[6]. Replicating the results of Experiment 1, an effect of social network size was found ($t(198)=2.34$, $p<.03$, Cohen's $D=0.33$) such that accuracy was higher in simulations with networks of 100 individuals (M=79.4%) than with simulation with networks of 20 individuals (M=78.0%).




[6] The network size of participants in Experiment 1 tended to be smaller, 11-74, but this is due to the fact that most participants who volunteer for such a time-intensive study do not work full time and engage in relatively few social activities. Earlier pilot studies suggest that 20-100 is a more common range. Furthermore, figures 2-3 show the effects along a range of network sizes.

Figure 2. Illustration of typical input that agents with a network size of 20 (left) and a network size of 100 (right) receive. Axes represent the first and second formant frequencies. Each color represent a different vowel category.

As the effect of social network size was replicated, it is possible to explore its underlying mechanism by examining in what way the input in the two types of networks differs. Previous research suggested that the benefit that exposure to multiple speakers confers is due to the greater variability in the input. Figure 2 illustrates the differences between the typical input that agents with a network of 20 received and the typical input that agents with networks of 100 received. Visual examination suggests that the input that the agent with a network of 100 received is indeed spread out more widely. One way to measure variability is to examine the Standard Deviation (SD) of the vowel categories. Indeed, a comparison of the SDs in the two types of networks shows that the two types of networks differed in the average SD of the vowels' formants for all formants (f1: $t(198)=3.70$, $p<.001$; f2: $t(198)=5.23$, $p<.001$), such that the SD was always higher in networks of 100 speakers. To test whether having greater input variability improves performance, the standard deviations of the formant frequencies were z-scored by vowel category separately for f1 and f2, and then the two scores were averaged to form one Input Variability score, such that, for every simulation, there was one standard measure of variability per vowel category. Additionally, accuracy per vowel category at test was extracted for each simulation. These accuracy scores were then z-scored by vowel category as well, as it is not recommended to use proportions as a dependent measure. A mixed model regression analysis with Simulation and Vowel as random variables, Input Variability as a fixed factor was then run to test whether variability in vowel category predicts accuracy at test for that category. The random

structure included intercepts as well as slopes for Input Variability for both the Simulation and Vowel random variables. Results confirmed that having greater input variability improves accuracy (β=0.33, SE=0.03, t=10.11[7]). To examine whether input variability is the underlying reason that leads larger social networks to improve performance, a mediation test was run using the mediation package in R (Tingley, Yamamoto, Keele & Imai, 2014). This analysis calculates what proportion of the effect of a factor (Network Size) on the dependent measure (Accuracy) is due to a mediator (Input Variability) rather than being a direct effect. It additionally tests whether the factor also has a direct effect on the dependent measure after the mediator has been taken into account. Results indicate that the vowel categories' variability mediates the effect of network size, rendering the direct effect of Network Size nonsignificant. Specifically, 55% of the effect of network size on accuracy is due to input variability. Correspondingly, when both Network Size and Input Variability were entered into the same model, results showed a significant positive effect of Input Variability (β=0.32, SE=0.03, t=9.86), but Network Size was no longer a significant predictor ($\beta=7e^{-4}$, $SE=5e^{-4}$, t=1.22).

To conclude, the simulations revealed that having a larger social network improves speech perception by increasing input variability.

*The interactive effects of network properties on performance*

Next, different parameters of the simulations were modified to examine if and how they influence performance and modulate the effect of social network size. All analyses in the following section compare networks of 20 speakers with networks of

[7] Here and later, significance is determined using the common criterion of t>=2.

100 speakers, as in the previous section. For illustration purposes, the figures also plot results from networks of 10 and 50 speakers.

Most of the literature on the importance of input variability focused on its contribution to learning. As stated earlier, language learning continues throughout our lives. Nonetheless, one may wonder whether input variability plays a larger role when input is scarce, and the more exposure one has, the less of a boost input variability provides. The results of Experiment 1 did not find such an effect, as the total number of hours of talk did not predict performance there. Still, to examine whether that is the case, identical simulations were run in which only the number of meetings the agent had varied between 100, 500, and 5000. Results show that agents that received more input did not perform any better. In contrast, the effect of network size was significant at all input levels (100 meetings: 77.5% vs. 78.9%, $t(198)=2.28$, $p<.03$; 500 meetings: 78% vs. 79.4%, $t(198)=2.34$, $p<.03$; 5000 meetings: 76.9% vs. 78.6%, $t(198)=2.74$, $p<.01$).

Another factor that could potentially modulate the effect of network size is the heterogeneity of the population. Therefore, another set of simulations was run in which the heterogeneity of the population, defined as the standard deviation of the vowel formant frequencies across the population, was either doubled (heterogeneous condition) or cut in half (homogeneous condition)[8]. As Figure 3 illustrates, performance is better the more homogeneous the population is (network size 20: baseline vs. homogeneous: 78% vs. 97%, $t(198)=40.17$, $p<.001$, baseline vs. heterogeneous: 78% vs. 47.25%, $t(198)=43.73$, $p<.001$; network size 100: baseline vs. homogeneous: 79.4% vs.

[8] Thr original means and Standard Deviations of the vowels were taken from Adank, Van Hout & Smits (2004), and are: oe:273 (35),872 (136), ie: 286 (26), 2343 (276), o: 410 (57), 869 (135), a: 668 (139), 1226 (151), aa: 791 (157), 1499 (128), e: 505 (62), 1865 (180), i: 380 (37), 2098 (241), uu: 282 (42), 1826 (187), u: 391 (48), 1713 (171).

97.5%, t(198)=41.78, p<.001, baseline vs. heterogeneous: 79.4% vs. 50.2%, t(198)=46.91, p<.001). The effect of Network Size, however, is significant at all levels of population heterogeneity (homogeneous: t(198)=2.55, p<.02, Cohen's D: 0.36; heterogeneous: t(198)=4.02, p<.001, Cohen's D: 0.57), though the effect size is numerically larger when the population is more heterogeneous. In other words, predictably, it is more difficult to understand novel speakers if the population is very heterogeneous. When speakers are very similar to each other, there is less ambiguity in the signal and it is easy to generalize from one to the other. The more speakers there are, the more likely there is to be ambiguity in the signal due to category overlap across speakers, and the less representative each speaker is. Thus, it precisely when such variability exists that having a larger social network is most helpful. When speakers differ from one another, there is need to encounter more of them in order to understand the speech patterns in the population and the structure of the speech categories.

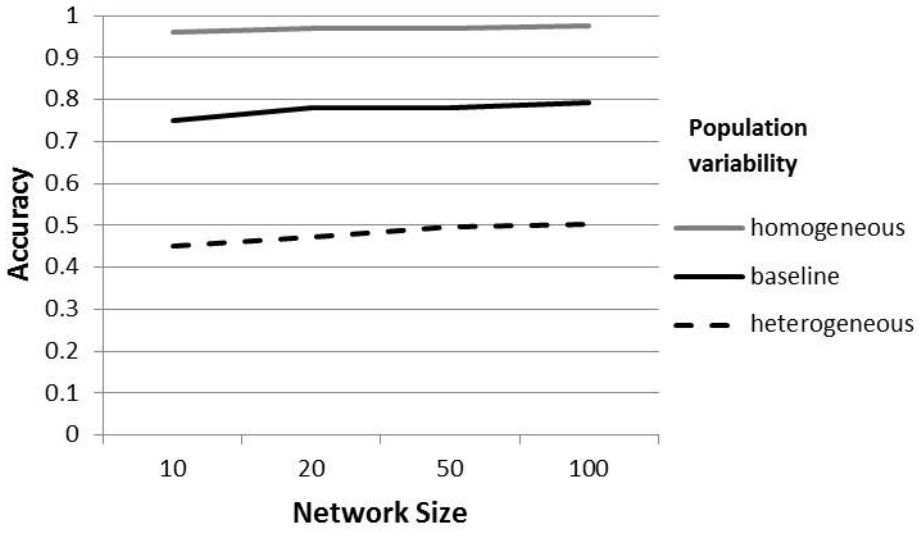



Figure 3. The effect of population variability on accuracy at different network sizes

Similarly, one may wonder how the effect of network size depends on individuals' consistency within themselves. Results show that increasing the standard

deviation of the productions of each speaker led to a drop in accuracy (network size 20: baseline vs. inconsistent: 78% vs. 74.4%, $t(198)=5.79$, $p<.001$, inconsistent vs. highly inconsistent: 74.4% vs. 66.9%, $t(198)= 11.84$, $p<.001$; network size 100: baseline vs. inconsistent: 79.4% vs. 76%, $t(198)=5.66$, $p<.001$, inconsistent vs. highly inconsistent: 76% vs. 68.7%, $t(198)=11.36$, $p<.001$;). This result is also predictable as lower intra-speaker consistency, similarly to greater inter-speaker heterogeneity, increases the ambiguity in the input. Even when individual consistency was lower, however, having a larger network led to better performance (inconsistent: $t(198)=2.59$, $p<.02$, Cohen's D: 0.37; highly inconsistent: 68.71% vs. 66.9%, $t(198)=2.71$, $p<.01$, Cohen's D: 0.38).

To conclude, the simulations reveal that having a larger social network can indeed causally improve speech perception, and that it achieves this by its increased variability. The simulations further show that the positive effect of network size holds across different levels of amount of input, and holds even if speakers are less consistent within themselves. At the same time, its beneficial effect seems to depend on the community's heterogeneity. Having a larger social network seems to be particularly helpful when the population is variable.

The results of these simulations are in line with Sumner's (2011) results, which show that input variability along the relevant dimension boosts learning. They are at odds though with the results of Rost & McMurray (2010), who found that the facilitatory effect of multiple speakers is due to the variability they provide along the irrelevant, rather than relevant, dimension. As mentioned earlier, one potential reason for the difference between the studies is that Rost & McMurray(2010) studied infants, who are yet to establish categorical distinction along the VOT continuum, whereas Sumner

(2011) tested adults who already use VOT to categorize phonemes, but needed to adjust their category boundary.

The situation in Simulation 1 is more similar to the adult than the infant case, as the learners knew how many categories there are, and to which category each token that they received belonged. In this case, learners benefit from having a wide spread in each category, as it assists in correctly categorizing atypical tokens. In contrast, if the learner's task is still to figure out how many categories there are and where in the space they are located, then category spread might hinder performance. In this case, input that is characterized by scattered clusters, as is the case when the social network is small, might be more useful for separating categories. Simulation 2 takes a first stab at this hypothesis by simulating learners who are unaware of the number of categories and the identity of each incoming token. Instead, these learners try to figure out the number of categories that there are from the distribution of the input they receive from either small or large networks.

## Simulation 2

Simulation 2 examines whether having greater input variability along the dimension that is critical for categorization is less helpful at the earliest stages of learning, when the categories are not known yet and still need to be learned.

### *General method*

To test the effect of input variability at the earliest stage of learning, the simulations from Simulation 1 were repeated with the following change: the input that the agent received was not labeled. As in Simulation 1, social networks included either 20 or 100 interlocutors, and the agent met with people from her network a predefined

number of times, each time receiving one token of each vowel. This time, those tokens were not labeled and the agent stored all of them together. After all meetings have concluded, a cluster analysis was run on the input that each agent received using Mclust package in R, which uses Gaussian mixture models (Fraley, Raftery, Murphy & Scrucca, 2012). Separate simulations were run for different number of meetings to examine whether the results differ depending on amount of input. Therefore, for each network size, 5 simulations were run for each of the following number of meetings: 100, 300, 500, and 1000. The number of simulations per condition was kept low as the results were highly consistent within each combination of network size and meetings.

*Results*

To examine whether clusters are harder to perceive when input variability is high, and whether this difference depends on amount of input, a t-test was run comparing the number of estimated clusters in the small and large social network conditions for each number of meetings. Results show that, as predicted, for each number of meetings, the number of estimated clusters was significantly higher in the small network condition than in the high network condition (100 meetings: M=18.4, SD=0.89 vs. M=4, SD=0, $t(8)=36$, $p<.0001$; 300 meetings: M=26, SD=2.45 vs. M=10.4, SD=1.14 , $t(8)=12.91$, $p<.0001$; 500 meetings: M=29.6, SD=0.89 vs. M=12, SD=2, $t(8)=17.96$, $p<.0001$; 1000 meetings: M=29.2, SD=1.79 vs. M=21, SD=4.58, $t(8)=3.73$, $p<.01$).

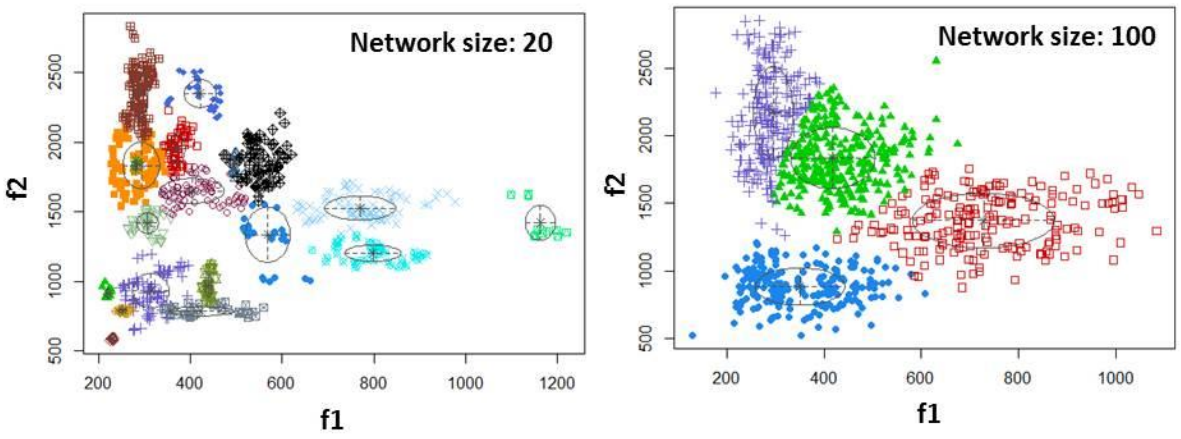



Figure 4. Illustration of estimated number of clusters after 100 meetings. The left panel demonstrates the case with a network of 20, and the right panel demonstrates the case of a network of 100.

Figure 4 illustrates what the clusters in each type of network look like after 100 meetings. As can be seen, when the input comes from a small network, often a single real category is divided into several categories, and outliers often form separate categories as well. Interestingly, the estimated number of categories was always larger than the real number of vowels when the social network was small. In contrast, when the input is provided by a large network, there are large categories, each comprised of several real categories.

Research on second language acquisition indicates that it is hardest to acquire a new distinction if it requires you to divide one category you have into two (e.g., Best, McRoberts & Goodell, 2001). This suggests that in the process of learning it might also be easier to merge distinct categories into one category than to split existing categories into several smaller ones. Therefore, at the earliest stages of learning, input variability along the relevant dimension might indeed not be useful but instead, it might even hinder the acquisition of the categories. At the same time, input from more speakers

often increases variability along both the relevant and irrelevant dimensions. Therefore, receiving input from multiple speakers might still be useful also at the earliest stages of learning, as Rost & McMurray (2009) found, but for a different reason. Having a larger social network then might boost performance via different mechanisms at different stages of learning.

## General discussion

The goal of this paper was to understand how individual differences in social networks can influence speech perception. Previous research suggested that phonological acquisition is influenced by the distributional nature of the linguistic input (Maye et al., 2002). In particular, it has been proposed that learning is better when the input is more variable, and input variability was often manipulated by increasing the number of speakers one is exposed to (Bradlow & Bent, 2008; Lively et al., 1993; Rost & McMurray, 2009, 2010). As people differ in the number of people they regularly interact with (Hill & Dunbar, 2003), this paper examined whether individual differences in people's social network size influences speech perception abilities.

Experiment 1 tested this hypothesis exploiting the natural variation in social network size. Results indicated that individuals with larger social networks are better at understanding vowels embedded in noise. Importantly, participants were tested on several cognitive abilities, and the beneficial effect of social network size on vowel perception was not driven by differences in any of the tested cognitive abilities among participants with different social network sizes, suggesting that the effect of social network size might be causal.

Simulations 1 systematically explored the mechanism underlying the beneficial effect of social network size as well as its interaction with other network properties. The results indicated that having a larger social network increases the variability in the

input, and this greater input variability leads to better phoneme categorization. Simulation 2 showed that this positive effect of input variability might not apply at the earliest stages of learning, as it renders the categories harder to distinguish.

These results reconcile Sumner's (2011) results with those of Rost and McMurray (2010). They suggest that different types of variability are useful at different stages of learning. During the initial stage, when the learner still needs to learn which properties to attend to and how to categorize them, variation in the irrelevant aspects of the input is more useful, but once the learner has learned what she should attend to and what the categories in the language are, it is the distributional properties of the relevant aspect of the input that are crucial for improving ability to classify input from new speakers. Thus, in Rost and McMurray's study (2010) infants did not benefit from variation along the relevant dimension. Simulation 2 suggests that at that stage, input variability at the relevant dimension makes it harder to distinguish between categories. In contrast, once the categories are already known and the tokens can be identified when processed, as is the case of adult native speakers, input variability increases learning and category robustness. Therefore, having larger social networks had a beneficial effect in both Experiment 1 and Simulation 1, and increased input variability had a positive effect in Sumer's (2011) study.

The effect of social network size in both Experiment 1 and Simulation 1 was significant but small. This is partly due to the fact that the participants, as well as the simulated agents, were adult native speakers who are proficient in the language. It is therefore impressive but its contribution at this point is more theoretical than practical. Future research should extend the study of the effect of social network size to populations with poorer linguistic performance, such as children with a language gap, second language learners etc., where social network size might account for a greater

part of performance. Similarly, future research should examine which other phonological aspects the effect extends to, and how this relates to their role in providing indexical information about the speakers.

Simulation 1 also examined the role of other properties of the network, as well as investigated their moderating role on the effect of social network size. Thus, it was discovered that, in line with the experimental results of Experiment 1, receiving more input can only be of limited help, if any, and having a larger social network improves performance independently of amount of input received. In contrast, the heterogeneity of the population plays an important role, and moderates the effect of social network size. Having a larger social network is more useful the more heterogeneous the population is.

These results also raise new questions. First, one may wonder whether social network size causally influences speech perception. After all, Experiment 1 did not manipulate social network size but exploited the natural variation in it. As is the case with any individual differences study, non-causal explanations cannot be ruled out completely. Several factors, however, make such alternative explanations unlikely. First, participants were tested on a host of cognitive measures, and these did not correlate with social network size, as well as were controlled for. Even more importantly, Simulation 1 replicated the positive effect of social network size. While computational simulations only show what's possible rather than necessarily the processes that take place, they show that having a larger social network should influence the distribution of the linguistic input one receives in a manner that facilitates later phoneme categorization. Lastly, these studies were inspired by experimental results that showed that exposure to multiple speakers leads to better phonological acquisition. Therefore, a causal explanation for the role of social network size seems the most plausible one.

Another potential caveat is that the actual variability in the input that participants in Experiment 1 received was never measured. Therefore, while the results of Simulation 1 suggest that input variability accounts at least partially for this effect, it is theoretically possible that the benefit of exposure to multiple speakers is due to another aspect of the input rather than its variability. That said, Experiment 1 was inspired by studies that varied the number of speakers with the goal of manipulating variability (e.g., Bradlow & Bent, 2008; Rost & McMurray, 2009). These studies assumed that increasing the number of speakers increases the variability of the input, but did not measure it. The results of the simulations reported here support that assumption, as they show that larger networks provide more variable information even when all speakers speak the same dialect.

The simulations examined the influence of several different network properties. At the same time, there are additional network properties whose role has not been simulated. For example, network density, that is, the interconnectivity of network members, might play a role as well. Future research should therefore measure individuals' network density, and include simulations that allow network members to interact with each other, and thus, influence each other. Additionally, future research should examine not only how many members people have in their social network but how the interaction with them is distributed. For example, social network size might play a different role if individuals interact a similar amount of time with most members of their network, than if the interactional pattern is skewed, such that they interact with a few for the large majority of the time, and very little with everyone else.

To conclude, this paper shows that the nature of our social network can influence the nature of the input we receive, and consequently, our speech perception. It thus

opens the door for research on how aspects of our life-style can influence our linguistic performance.

References

Adank, P., Van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical society of America, 116, 3*, 1729-1738.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106, 2*, 1054-1063.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27, 3*, 387-414.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America, 109, 2*, 775-794.

Bradlow, A. R. & Bent, T. (2008) Perceptual adaptation to non-native speech. *Cognition*, *106*, 707-729.

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*, 804–809.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*, 2-3, 101-118.

Fraley, C. Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation *Technical Report No. 597, Department of Statistics, University of Washington*.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13, 5*, 431-436.

Hill, R. A., & Dunbar, R. I. (2003). Social network size in humans. *Human nature*, *14*,1, 53-72.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental psychology, 27, 2*, 236.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. *The Journal of the Acoustical Society of America, 118, 5*, 3267-3278.

Janse, E., & Newman, R. S. (2013). Identifying nonwords: Effects of lexical neighborhoods, phonotactic probability, and listener characteristics. *Language and Speech, 56, 4*, 421-444.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*, 485–499.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a Cocktail Party Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice. *Psychological science*, *24*, 10, 1995-2004.

Kleinschmidt, D. F. (2016) *Perception in a Variable but Structured World: The Case of Speech Perception* (Unpublished doctoral dissertation). Rochester, NY, USA.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the

familiar, generalize to the similar, and adapt to the novel. *Psychological review, 122*, 2, 148.

Lev-Ari, S. (2016). How the size of our social network influences our semantic skills. *Cognitive Science*. *40*, 2050-2064.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94,* 3 Pt 1, 1242.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 2, 391.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, 3, B101-B111.

Peterson, G. E., and Barney, H. L. (1952). 'Control methods used in a study of vowels. *Journal of The Acoustical Society of America, 24*, 175–184

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech communication, 13, 1*, 109-125.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353 – 363.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills, 8*, 271–276.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological

processing in early word learning. *Developmental Science*, *12*, 2, 339-349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*, 6, 608-635.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274,* 5294, 1926-1928.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*, 5, 3306-3316.

Stager, C. L, & Werker J. F. (1997) Infants listen for more phonetic detail in speech perception than in word learning tasks. *Nature*, *388, 6640*, 381–382.

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*, 131-36.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1, 1-42.

Tingley, D., Yamamoto, T., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59, 5*, 1-38.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505.

Vosoughi, S., Roy, B. C., Frank, M. C. & Roy, D. (2010). Contributions of prosodic and distributional features of caregivers' speech in early word learning. In S. Ohlsson & R. Catrambone (eds), *Proceedings of the 32nd Annual Cognitive Science Conference*, Portland, Oregon, 1822–27. Austin, TX: Cognitive Science Society.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37-48.

Appendix A – Full table of results for transcription of nonwords in noise task

| | β | SE | Z | p-value |
|---|---|---|---|---|
| (intercept) | 1.1 | 0.33 | 3.36 | <0.001 |
| Social Network Size | 0.02 | 0.01 | 2.01 | <0.05 |
| Hours of Talk | 0.001 | 0.005 | 0.21 | 0.83 |
| WM | 0.001 | 0.007 | 0.2 | 0.84 |
| Auditory STM | 1.37 | 0.87 | 1.58 | 0.11 |
| Selective Attention | -0.42 | 0.62 | -0.69 | 0.49 |
| Task Switching | -0.03 | 0.16 | -0.2 | 0.84 |