The Effect of Language Ability on Speaker Identification

Sofia Jimenez

15 June 2012

UC San Diego

INTRODUCTION

Spoken language contains a large amount of information, including speech sounds, vocabulary, syntax, and semantics. The speech signal also contains speaker-related acoustic information: listeners not only acquire information about *what* is said but also *who* says it. Compared to our understanding of word recognition and syntactic processing, we know little about how speaker information is processed (though see Creel, in press; Van Berkum et al., 2008), or what factors contribute to proficiency in speaker recognition. In order to understand fully the nature of spoken-language knowledge, we must look into what acoustic and linguistic factors affect listeners' processing of speaker-related properties.

Talker Recognition

Listeners can glean information about the age (Mulac & Giles, 1996), gender (Lass, et al. 1976) and emotion (Scherer, 1981) of the speaker as well as the individual identity. From very early in development, children are not only becoming familiar with the speech sounds of a language, but are also listening to what people sound like and becoming familiar with their unique voice patterns. For example, fetuses react differently to their mother's voice than to a stranger's voice before birth (Kiselevsky et al., 2003; see also DeCasper & Fifer, 1980). Preschool aged children show some capacity to recognize familiar cartoon characters' voices (Spence, Rollins & Jerger, 2002) as well as their own classmates' voices (Bartholomeus, 1973). Voice identification ability continues to improve throughout development along with other perceptual abilities such as face identification (Mann, Diamond & Carey, 1979).

Being able to recognize a speaker by their voice is important for social reasons. Imagine answering the phone and not being able to distinguish if it was your boyfriend on the other end or your father. People usually use different social registers to talk to talk to different people that they know. Even more detrimental would be to mistake a robber's voice for a roommate's, or the other way around!

Acoustic Properties Distinguishing Speakers

There are many variables that affect how good we are at speech recognition. For example, duration of speech sample is important in voice identification, the longer the sample lasts, the better able it is to be identified (Pollack, Pickett & Sumby, 1954). There is also the factor of pitch inflection: tested by having speakers whisper, which gets rid of pitch inflection. It was found that when speakers whispered it was slightly more difficult for listeners to identify them. You need to listen to whispered speech three times as long as normal speech to get the same accuracy of identification (Pollack et al., 1954). If age, race, musical ability, duration and pitch inflection all influence ability to identify speakers then we should determine if language effects speaker recognition.

A lot of information is contained within the individual speech signal. Adults can glean enough information about speaker's voice for us to be able to identify backwards speech to some extent (Van Lancker, Kreiman & Emmorey, 1983). Though speech recognition is better in a familiar language, there is still enough information in an individual's speech signal for us to identify voices across languages: Winters, Levi and Pisoni (2009) found that listeners were able to identify bilingual speakers across languages. Although there is enough language-independent information in speech to be able to identify voices in two different languages, there is still some reliance on language

dependent cues. For example, in this study when the subjects knew English and were trained on English voices then presented with novel German stimuli they performed worse than their German-trained counterparts.

Linguistic Properties Distinguishing speakers

We know that preschoolers are much worse than adults are at learning novel voices even when they vary by prosody and by spectral differences (Jimenez & Creel, 2011; Creel & Jimenez, under review). The heightened ability to recognize voices might correlate with increased verbal intelligence, therefore we would find that as vocabulary size increases so could the ability to recognize differences between voices and learn to associate them with different characters.

There is a trend that indicates that vocabulary size is correlated with phonological awareness in children (Schwarz, Burnham & Bowey, 2006). Phoneme sensitivity and vocabulary size predict each other in 30, 33 and 36 month old children, we can predict that phoneme sensitivity would go along with voice recognition sensitivity. Based on these results, we would like to see if there are vocabulary effects in an older population.

Individual Variability in Identifying Speakers

We know that there are a variety of factors that influence the ability to identify speakers by their voice. Another factor may be music ability. Slevc and Miyake (2006) tested Japanese/ English bilinguals on different dimensions of language processing, including receptive phonology, productive phonology, syntax and lexical knowledge. Music ability was also assessed in all participants by various music tasks. They found that receptive phonology in a second language was correlated with music ability, suggesting a relationship between music ability and perception of phonological differences. If music

ability is related to *phonological* perceptual acuity, it may also confer benefits in perceiving non-phonological speech-relevant acoustical differences such as speaker recognition. Another factor is dialect familiarity: people are more accurate at learning and identifying voices when the speaker is the same race—that is, has the same accent—as the listener (Perrachione, Chiao, Wong, 2010).

Identifying speakers is also influenced by language. Since we have an abstract phonological representation of what the languages should sound like it is easier for us to identify voices in our native tongue (Perrachione, Pierrehumbert & Wong, 2009; Johnson, Westrek, Nazzi & Cutler, 2011). This own-language bias is present even for infants as young as seven months: when tested in a visual fixation habituation paradigm, infants noticed a change in speaker only when the speakers used the infants' native language (Johnson, et al., 2011). Many studies have found a correlation between language knowledge and ability to recognize voice. Monolingual English speakers could identify bilingual Spanish-English speakers better when they spoke in English (Thompson, 1987).

Johnson et. al. (2011) used speech discrimination, which has been found to be different than voice recognition by Van Lancker and Kreiman (1983). The recognition of familiar voices is cognitively different from the discrimination of familiar voices this is tested with people who have brain damage: right, left or bilateral. Both sets of brain-damaged subjects were bad at speech discrimination where in the recognition task the right brain damage subjects were worse. Van Lancker and Kreiman show that speech discrimination and voice recognition are not the same, through different abilities of brain-damaged subjects. We will be testing for recognition, which is typically associated with the right hemisphere of the brain. Language knowledge like speech recognition, however, is

thought to be primarily located in the left hemisphere. Therefore when we are testing whether speaker identification is related to language ability, we are possibly working with areas that are processed in two different hemispheres.

Recent studies suggest that language ability and voice perception are related. Perriachone et al. (2011) tested dyslexic and controls on voice recognition in both Mandarin and English. Dyslexic people were no better at recognizing English voice than Mandarin, but the controls were. Data shows that dyslexics have difficulty with voice recognition. The own language effect doesn't help dyslexic people (Perriachone et al., 2011). Since dyslexic people have been found to have impaired phonological representations it is assumed that it is an important part of being able to identify a voice. This may suggest that language ability affects voice recognition, which may suggest the *who* of voice identification and the w*hat* of the content it contains are not separate as they were once thought to be (Kuhl, 2011). We hypothesize that language ability will be somewhat correlated to voice learning even in non-dyslexic adults. In the present study we test this hypothesis by giving participants a voice-learning task and testing them on two different language measures: vocabulary and phonological processing.

METHODS

Participants. Our experiment will use 30 adult subjects from the UC San Diego experimental subject pool. Half reported their native language as English and half reported a different native language. They will be compensated with class credit for their time.

Stimuli. The auditory stimuli consist of eight adult female voices. Each speaker was recorded speaking six different sentences: three phrases used during training and three different phrases during testing. Speakers were all approximately the same age (age range: 19-22 years old) and were from the same dialect region (all native Californian). Therefore, most variation between voices should be due to individual speaker differences. These individual characteristics might include differences in pitch (fundamental frequency), prosody, speech rate (syllables/second), or formant frequency, among other things. The particular phrases that were recorded each contained all three of the vowels /i/, /a/ and /u/ in different orders, for example, "You eat hot dogs". These vowels were chosen to give the listener the most representative exposure to the vowel space (characteristic formant frequencies) of each speaker. The sounds were .wav files and had the sampling rate of 44100.

Visual stimuli were eight female cartoon faces downloaded from the Microsoft clip art database (http://office.microsoft.com/en-us/images/?CTT=97). Since this is an experiment on voice identification, not face identification, we did not want to impose perceptual difficulty in discriminating faces. Therefore, face images were chosen to be highly distinct from each other.

Procedure. Participants completed three different tasks in this experiment: two measures of language processing, and one measure of voice learning. First, their verbal intelligence is tested using the Peabody Picture Vocabulary test (Dunn & Dunn, 2007), which provides age-appropriate norms from preschool through late adulthood. The test is started at the appropriate section for the participant's age. On each trial, the participant is asked to point to one of four pictures that a particular word refers to. The words get harder

as the participant moves on. If the participant misses eight or more words in a particular block then the test is finished. This part of the experiment lasts about twenty to thirty minutes depending on the subject.

The second part of the experiment is the Comprehensive Test of Phonological Processing (Wagner, 1999), which is a measure that specifically tests phonological processing and knowledge. Perrachione et al. (2011) found that CTOPP performance was correlated with voice recognition accuracy in dyslexic participants.

This test assesses three different categories: phonological awareness, phonological memory and rapid naming. Each category has two different subtests that determine the subjects' proficiency. Phonological awareness, or the ability to recognize, manipulate and synthesize phonemes in English is tested by the Elsion and Blending Words tasks. The Elison involve repeating a word without a phoneme or cluster of phonemes (e.g. Say powder without the d). In the Blending Words task the subject is asked to listen to a word that is pronounced phoneme by phoneme and report what word it is. (e.g. they hear j-u-m-p over a set of headphones and should reply "jump"). Phonological memory, or the ability to remember sounds of a language is tested by a Memory for Digits and a Nonword Repetition task. Subjects are tested on how well they can repeat numbers that are read to them and made up words that they hear. The third category, rapid naming, was tested by subjects reading a list of digits and a list of letters aloud as fast as they could. This part was not included in the analysis since it did not relate to what we were trying to determine, we were not concerned with production.

The third, final part of the experiment will test subjects' ability to both learn different speakers' voices and then identify them when they utter novel phrases. This

phase of the experiment was programmed in Matlab PsychToolBox 3 (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007) by the author. Participants are seated in secluded testing rooms with a Macintosh Mini computer. Sounds are presented via Sennheiser HD 280 pro headphones. There are two phases: training and testing. During training, listeners are presented with two of the eight faces on the screen, and one of the voices plays. The participant is asked to make a response on a keyboard and press either 'z' if they think the voice belongs to the left face, or 'm' if they think it is the right face. Participants are given unlimited time to respond. After they have made their selection, the correct answer stays on the screen for one second while the incorrect answer disappears. At first, the participant simply guesses which one is correct, until they learn the voice/face pairings and can accurately determine which voice goes with which face.

The training will take place in blocks of 56 trials each. This allows each of the faces to be paired with each one of the other faces on both sides of the screen once. The target face (which is the one that the voice belongs to) is counterbalanced so that it is on the left roughly as often as it is on the right. The two face pairings are randomized so that each participant will receive a different trial order. Each participant will hear each phrase more or less equally often. The participants will move from the training phase to the testing phase when they achieve 85% accuracy in a block of training trials. They may have to repeat the block of training trials multiple times if they fail to meet criterion. If the participant fails to meet criterion in ten training trials the subject is automatically moved to the testing phase, but this may mean that the participant is excluded from the data.

The testing phase (56 trials total) follows the same format as the training trials, presenting two faces and playing a sound; however, no feedback is provided. Further, the

test phase uses three new phrases that the participant has not been trained on. These phrases contain the same three vowels as the training sentences, and are similar in duration and form. On each trial, Matlab records the reaction time, accuracy, faces presented, phrase presented, and the target location in a text file.

Figure 1. Examples of faces participants were asked to match to voices.

After they have completed all three tasks they are given a language dominance survey to obtain their language background and moving history. Participants can be monolingual or bilingual so we will use this survey to determine their degree of bilingualism as well as their age of acquisition of English and age of acquisition of their other language. Participants will be asked to include all languages spoken, number of years and frequency of each language's use, formal/informal years of schooling in the language and how the language was acquired.

It is not only important to collect the accuracy of the participants but also to see if there are any differences in speed of identification. Since we are testing the participants in regards to two different language measures, vocabulary and phonological processing, we

want to see if these measures are correlated themselves and whether they correlate with the amount of training blocks required to learn the eight voices. The language measures could both be correlated with voice learning or one could be more predictive or it could turn out that neither measures correlate with voice identification. In which case we would conclude that voice recognition ability does not relate to language ability.

RESULTS

Our primary question is whether there is a relationship between language ability and voice recognition. We will test that by determining if there is a correlation between the scores on the PPVT (Vocabulary) and CTOPP (Phonological Processing), and performance in the voice learning experiment. We looked at two factors in regard to the voice learning experiment: 1) how long it took for the participant to learn the voices and 2) their accuracy in the testing trials. Then we tested the significance of the correlation between the performance measure and the score on the PPVT and CTOPP, using linear regression.

Firstly, it is important to note that the wide range of variability found in the training trials required before participants met criterion. This indicates that there are some people that pick up on voice differences and learn them more quickly than others in the typical adult population. The training trial data is what our analysis will be focusing on. It will be the basis of comparison of the language measures. The testing data was not as revealing because of the high accuracy among all participants. This implies that they were able to generalize to the new phrases without much difficulty. Since most of the testing scores

were too high to show a correlation, we will focus on the length it took participants to learn the voices to an 85% accuracy level.

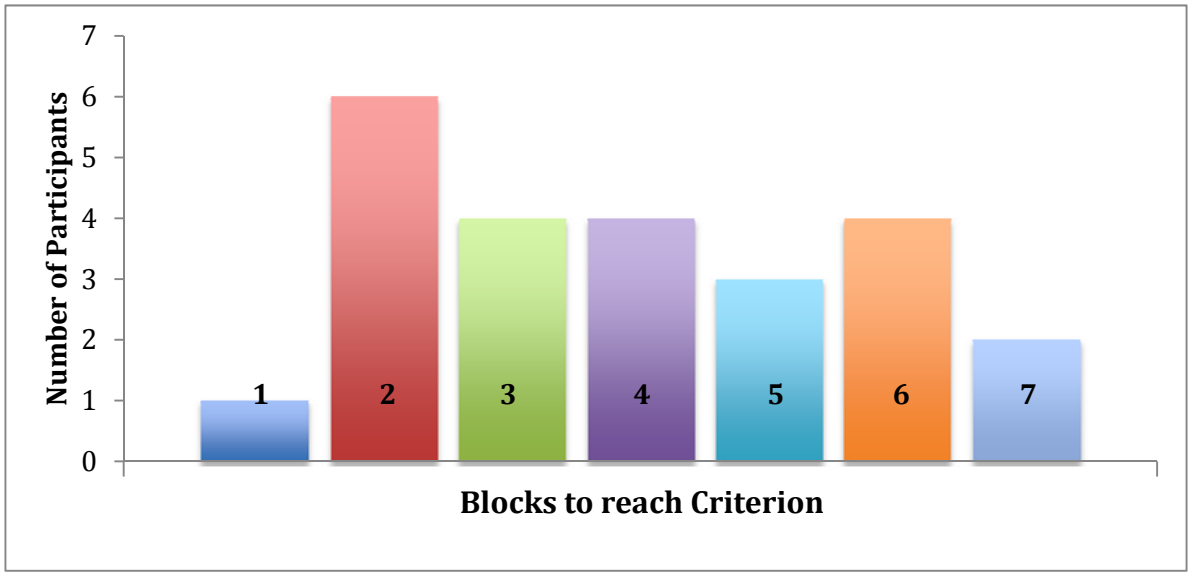

Figure 2. Number of Training blocks for individual participants to reach criterion.

Not only was there a wide range of scores on the voice-learning task, but also on the language measures that we tested, PPVT and CTOPP (see *Figure 3*), indicating that adults varied in their language knowledge. It is important to note that none of our participants' CTOPP performance qualified as "poor" according to the makers of the test. The CTOPP is one measure used to evaluate Dyslexics and we wanted to ensure that none of our participants were dyslexic-performing so that we could be sure we were, in fact, testing among the typical adult population.

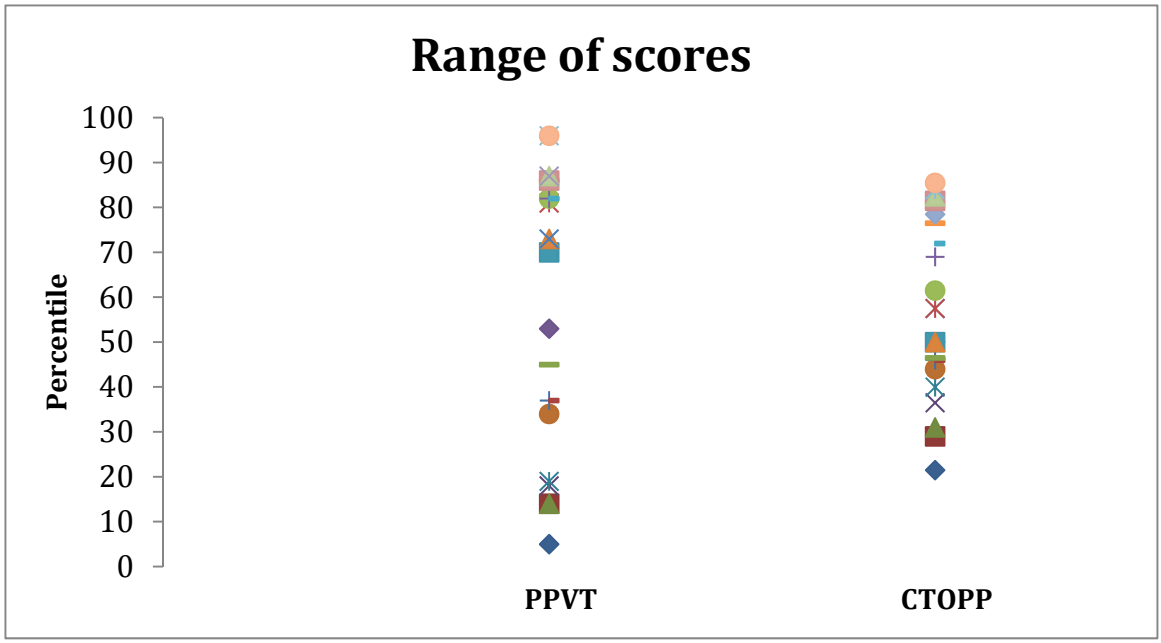

Figure 3. Individual participant scores on language measures: PPVT and CTOPP.

The first language measure that we will explore is vocabulary knowledge. We will see if the participants' score on the Peabody Picture Vocabulary Test predicted the speed of learning voices.

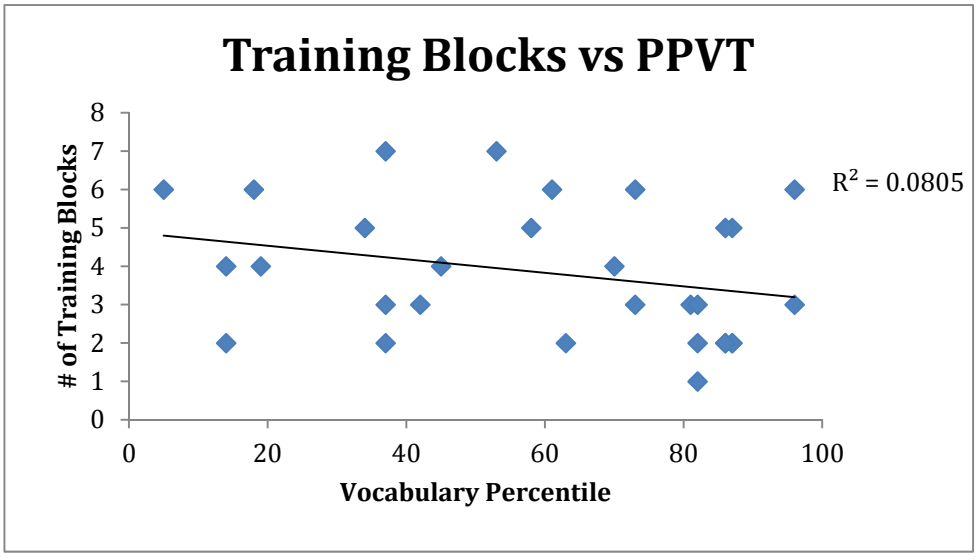

Figure 4. Participants' vocabulary percentile as determined by the PPVT plotted against the number of blocks to reach criterion (85%).

We found that there was no significant correlation between a participants' vocabulary score and the number of trials it took them to learn the voices ($R^2 = 0.0805$). Pearson's R was -0.284 but the correlation was not significant. Vocabulary knowledge does not aid in the learning and identification of voices.

The second language measure, phonological processing, as determined by the Comprehensive Test of Phonological Processing (Wagner, 1999) was also compared against voice learning. Since there are three different measures that are included in the CTOPP each with two tests, we decided to used a combined score to determine a percentile for each category: Phonological Awareness, Phonological Memory, and Rapid Naming. Rapid Naming did not end up correlating with the other two measures or with the voice identification learning. That is why we left it out of the analysis. We also combined the Phonological Awareness and Phonological Memory scores, which were correlated as can be seen in *Figure 6*. *Figure 5* shows a Phonological Processing percentile, which is the average of the Phonological Awareness, and Phonological Memory percentiles.

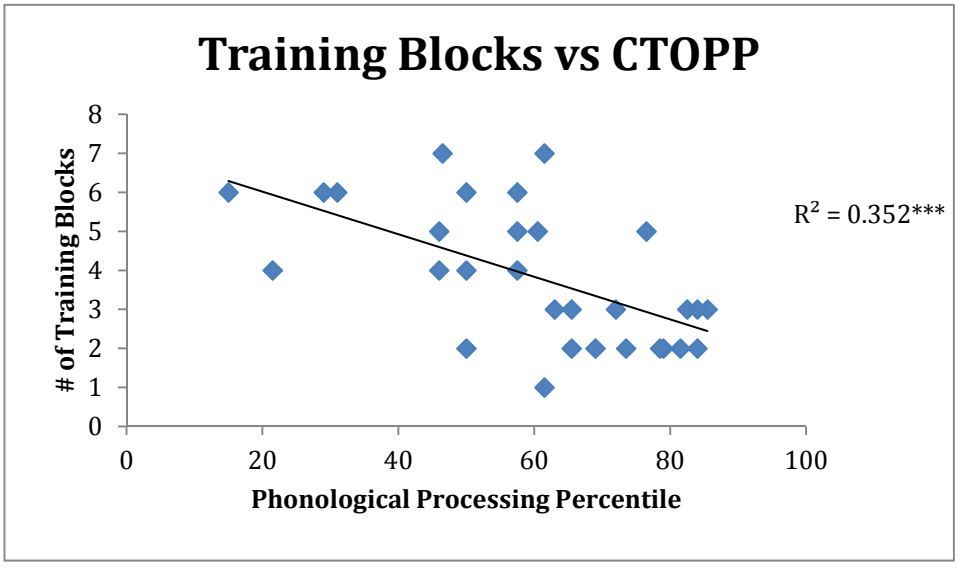

Figure 5. Participants' phonological processing percentile as determined by the CTOPP plotted against the number of blocks to reach criterion (85%).

In contrast to vocabulary knowledge, phonological knowledge did correlate with ability to recognize voices ($R^2 = 0.352$). Pearson's R was -0.59 in a linear regression analysis this correlation was found to be significant with a p-value of 0.0005. When the linear regression was done without including the four participants that were below the 40th percentile in terms of Phonological Processing, the correlation held (Pearson's R = -0.53, p-value = 0.004). This indicates that the correlation is not just driven by four outliers and remains significant when we exclude them from the linear regression data analysis.

Another area of interest was the correlation of different measures against each other. *Figure 6* shows a summary of how the different measures correlated with each other. While the most relevant to our study was how the "Blocks" (i.e. the number of training blocks it took to reach criterion) correlated with the other measures. However, it is also interesting to note how all the language measures were significantly correlated with each other.

	Blocks	**PPVT**	**CTOPP Memory**	**CTOPP Awareness**
Blocks	1	-0.28	**-0.51**	**-0.52**
PPVT		1	**0.52**	**0.43**
CTOPP Memory			1	**0.48**
CTOPP Awareness				1

Figure 6. Correlation matrix of each test compared against each other. (Bolded items indicate significant correlation)

DISCUSSION

People are known to have varying degrees of language ability and varying degrees of ability to learn and identify speakers, but with this research we found the extent to which these abilities are related.

We are assuming people who took less time to reach 85% accuracy on the voice-learning task are in fact, better at identifying voices. A better score on the Phonological Awareness category of the CTOPP might indicate that the participant is better at distinguishing the different voices from each other. This would give them an advantage in actually learning the voices over people who could not distinguish the voices from each other as well. However, even if a participant can distinguish the voices from each other very well, they may have a difficult time identifying them if they cannot remember which voice goes with which face. Associating the voices with the faces requires a certain amount of learning and memory.

Conversely, a high score on the Phonological Memory category of the CTOPP might give a participant an advantage in learning the voices that they can distinguish. Since phonological memory determines how well a participant can keep sounds in their working memory it could be an indication of a faster association of voice with face. However, if a participant cannot distinguish the voices, their ability to form the associations will be undercut no matter how good their memory is.

As Perrachione et al. (2011) found there may be a correlation between ability to recognize voices and language ability, specifically phonological processing knowledge. That

study was done with dyslexic participants, so we sought to expand upon this and see if the effect would pervade in a population with normal variation in language ability. As we found, the participants that we tested varied greatly in the two measures of language ability and in the time that it took them to learn voices. Even within this typical population of adults the Phonological ability correlated with voice identification. This shows that there is a continuum of voice recognition abilities that corresponds to the continuum of phonological ability. This is not just apparent in the dyslexic/non-dyslexic divide.

We also can say more about the developmental relationship of voice identification. Looking back to Jimenez and Creel (2012) we found that adults were better at learning to identify speakers than preschool aged children. We speculated that this might be due to increased language knowledge in adults as compared to Pre-school children. While we found that this is partially true, it is not all language knowledge that makes a difference in being able to identify voices. Phonological processing knowledge acquired over development does seem to influence talker knowledge.

This is also congruent to what we would expect in regards to Perrachione's (2011) study done with dyslexic subjects. People only qualify as dyslexic if they do not have any intellectual impairment, there is no data that states that dyslexic people have smaller vocabularies, presumably because they would be able to learn words just as easily as anyone else. However, there has been evidence that dyslexic people have impaired phonological processing compared to the typical population. Our findings suggest that the relationship between phonological processing and voice recognition are wider-scale than postulated in Perrachione's research.

REFERENCES

Bartholomeus, B. (1973). “Voice identification by nursery school children.” *Canadian Journal of Psychology*, 27, 464-472.

Brainard, D. H. (1997). *The Psychophysics Toolbox, Spatial Vision , 10:443-446.*

Dunn, L.M., & Dunn, D.M. (2007). *PPVT-IV: Peabody Picture Vocabulary Test – Fourth Edition.* Circle Pines, MN: American Guidance Service.

Jiménez, S. J., & Creel, S. C. (2012). “Factors affecting talker recognition in preschoolers and adults.” *Proceedings of the 36th annual Boston University Conference on Language Development*.

Johnson, E. K., Westrek, E., Nazzi, T. & Cutler, A. (2011). “Infant ability to tell voices apart rests on language experience,” *Developmental Science*, 14(5), (1002-1011).

Kisilevsky, B. S., Hains, S. M. J., Lee, K., Xie, X., Yhang, H., Ye, H. H., et al. (2003). “Effects of experience on fetal voice recognition.” *Psychological Science*, 14 (3), 220–224.

Kleiner M, Brainard D, Pelli D (2007), "What's new in Psychtoolbox-3?" *Perception 36 ECVP Abstract Supplement.*

Kuhl, P. K. (2011). “Who’s talking?” *Science* 333, 529.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). “Speaker sex identification from voiced, whispered, and filtered isolated vowels,” *J. Acoust. Soc. Am*. 59, 675–678.

Mann, V. A., Diamond, R., & Carey, S. (1979). “Development of voice recognition: Parallels with face recognition.” *Journal of Experimental Child Psychology*, 27, 153–165.

Mulac, A., Giles, H (1996) "'You're Only As Old As You Sound:' Perceived Vocal Age and Social Meanings." *Health Communication*, 8 (3), 199-215.

Pelli, D. G. (1997). *The VideoToolbox software for visual psychophysics: Transforming numbers into movies, Spatial Vision 10:437-442.*

Perrachione, T. K., Chiao, J.Y., and Wong, P. C. M. (2010). "Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices." *Cognition* 114 (2010) 42–55.

Perrachione, T. K., Pierrehumbert, J.B., and Wong, P. C. M. (2009). "Differential Neural Contributions to Native- and Foreign-Language Talker Identification." *Journal of Experimental Psychology – Human Perception and Performance*. Vol. 35, No. 6, 1950 – 1960.

Perrachione, T. K., Tufo S.N, Gabrieli, J.D.E. (2011). "Human voice recognition depends on language ability" *Science* 333, 595.

Pollack, I., Pickett, J. M., and Sumby, W. H. (1954). "On the identification of speakers by voice," *J. Acoust. Soc. Am*. 26, 403–406.

Scherer K. R. (1981). "Speech and emotional states." In: Darby JK, ed. *Speech evaluation in psychiatry.* New York: Grune & Stratton, 189--220.

Schwarz, I.C., Burnham, D., & Bowey, J.A. (2006). "Phoneme sensitivity and vocabulary size in 2 ½ - 3-year-olds." Proceedings of the 11th Australian International Conference on Speech Science & Technology, 42-147.

Slevc, L. R. & Miyake, A. (2006). Individual differences in second language proficiency: Does musical ability matter? *Psychological Science, 17(8),* 675-681.

Spence, M. J., Rollins, P. R. and Jerger, S., (2002).“Children’s Recognition of Cartoon Voices”, *Journal of Speech, Language, and Hearing Research*, 45(1), 214-222, 2002.

Thompson, C. P. (1987). “A language effect in voice identification,” *Appl. Cognit. Psychol*. 1, 121–131.

Van Lancker D., Kreiman, J., & Emmorey, K. J. (1983). “Recognition of famous voices forwards and backwards.” *J Acoust. Soc. Am*. 74, S50.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing.* Austin, TX: Pro-Ed.

Winters, S. J., Levi, S. V., Pisoni, D. B., (2008). “Identification and discrimination of bilingual talkers across languages,” *J Acoust Soc Am.* 2008 June; 123(6): 4524–4538.