Infant Ability to Tell Voices Apart Rests on Language Experience

Elizabeth K. Johnson[1], Ellen Westrek[2], Thierry Nazzi[3], and Anne Cutler[2,4]

[1] University of Toronto, Toronto, Canada

[2] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[3] Université Paris Descartes (Laboratoire Psychologie de la Perception, CNRS UMR 8158), Paris, France

[4] MARCS Auditory Laboratories, University of Western Sydney, Sydney, Australia

Address for Correspondence:

Elizabeth K. Johnson

3359 Mississauga Road N.

Dept. of Psychology, University of Toronto Mississauga

Mississauga, ON

Canada L5L 1C6

Email: elizabeth.johnson@utoronto.ca

Abstract

A visual fixation study tested whether seven-month-olds can discriminate between different talkers. The infants were first habituated to talkers producing sentences in either a familiar or unfamiliar language, then heard test sentences from previously unheard speakers, either in the language used for habituation, or in another language. When the language at test mismatched that in habituation, infants always noticed the change. When language remained constant and only talker altered, however, infants detected the change only if the language was the native tongue. Adult listeners with a different native tongue than the infants did not reproduce the discriminability patterns shown by the infants, and infants detected neither voice nor language changes in reversed speech; both these results argue against explanation of the native-language voice discrimination in terms of acoustic properties of the stimuli. The ability to identify talkers is, like many other perceptual abilities, strongly influenced by early life experience.

Keywords: infant speech perception, voice recognition, language discrimination, rhythm

Spoken language simultaneously carries two distinct types of information that are important for human communication. On one level, the speech signal conveys indexical information about the speaker, such as the speaker's identity, age, sex, socio-economic status, and emotional state. On another level, the speech signal carries a linguistic message that is phonologically well-formed and semantically meaningful. It takes many years for children to reach adult-like competency in processing each of these streams of information (e.g. Mann, Diamond, & Carey, 1979; Nittrouer & Lowenstein, 2007). But the acquisition process for both information types starts very early in life. During the third trimester of pregnancy, the fetus begins to receive auditory stimulation from the outside world. This prenatal auditory experience is evidenced by language and voice preferences in newborns; so, newborns recognize the rhythmic structure of the maternal language (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988; Moon, Panneton-Cooper, & Fifer, 1993; Nazzi, Bertoncini & Mehler, 1998), and prefer their mother's voice to the voice of a strange woman (DeCasper & Fifer, 1980; Mehler, Bertoncini, Barrière & Jassik-Gerschenfeld, 1978).

Does initial language experience also support voice discrimination? Note that the indexical and the linguistic information carried by speech signals are logically independent, and are hence thought to be of necessity processed independently of one another. To retrieve linguistic messages, infants must very quickly learn (or be born knowing) which aspects of the speech signal encode linguistic information, so that these can be abstracted from the variation arising from indexical properties. Some research has suggested that in many situations infants indeed readily extract the abstract linguistic message from speech in the face of indexical variation (e.g. Jusczyk, Pisoni, &

Mullennix, 1992; Kuhl, 1979; Van Heugten & Johnson, 2009). Other research, however, suggests that this process is initially difficult for young children (e.g. Schmale & Seidl, 2009; Schmale, Cristia, Seidl, & Johnson, 2010; Houston & Jusczyk, 2000; Singh, Morgan, & White, 2004).

Recently it has become clear that even in adults, the processing of indexical and linguistic information is intertwined in a complex manner, and there may be no straightforward answer to questions of precedence. Indexical information can in some cases be processed in the absence of linguistic processing; thus adult listeners can perform better than chance when asked to identify speakers from samples of backwards or sine-wave speech (Remez & Fellowes, 1997; Van Lancker, Kreiman, & Emmorey, 1985). Conversely, the processing of linguistic information can proceed in the absence of indexical processing, most notably in the case of phonagnosia; people with this condition are unable to identify speakers, but they can still understand language (Garrido et al., 2009).

But at the same time, there is considerable evidence that the processing of linguistic and indexical information is interwoven. On the one hand, voice familiarity can aid linguistic recognition: lexical processing varies across single- versus multiple-talker conditions (e.g. Mullennix & Pisoni, 1990), and difficult processing of linguistic information is made easier when the speaking voice is familiar (e.g. Nygaard, Sommers, & Pisoni, 1994). On the other hand, linguistic knowledge can impact the processing of indexical variation. Thus what counts as indexical can be language-specific – for instance, creaky voice is a property of speakers or a discourse feature in English, but is a manner of articulation encoding phoneme identity in some other languages (e.g., many languages of the North American sub-continent). Likewise, languages differ in how much variation they allow along partly non-linguistic parameters: fundamental frequency (F0), for example, has a wider range in the tone language Cantonese than in English (Chen, 1974), but a wider range in English than in the closely related Dutch (Collins & Mees, 1999). These considerations all make clear that voice recognition capabilities cannot be divorced from linguistic experience. It is thus unsurprising to find that adult voice recognition is most accurate in a familiar language (e.g. Goggin, Thompson, Strube, & Simental, 1991). This effect of language familiarity on voice identification is robust and has been observed in many language pairs including Spanish and English (Thompson, 1987), German and English (Winters, Levi, & Pisoni, 2008), Chinese and English (Perrachione, Pierrehumbert, & Wong, 2009), Spanish and German as well as Chinese and German (Köster & Schiller, 1997).

These adult results motivate important predictions for language development. Children's voice recognition capabilities should develop in pace with increasing competence in processing the linguistic structure of the native language. Although there have been many studies examining infants' recognition of highly familiar voices speaking a familiar language, little work has examined infant recognition of new voices first encountered in the lab (though see Floccia, Nazzi & Bertoncini, 2000). Moreover, the little work that has been done in this area has never examined recognition of voices speaking an unfamiliar versus a familiar language. The existing literature does, however, suggest some cautious predictions concerning when the effect of language experience might first become apparent in infants. Although not specifically designed to test voice recognition, language discrimination studies typically habituate infants to a few voices speaking one language and then present infants with new voices speaking either the same language or a new language. If infants can discriminate the two languages, then the prediction is that they will listen longer to the new speakers speaking the new language rather than the new speakers speaking the habituated language. The logic underlying this study thus assumes that infants will not dishabituate to a simple speaker change (only to a speaker plus language change). Indeed, this is precisely what language discrimination studies have shown. For example, in habituation studies, a neonate habituated with English voices will dishabituate when presented with a new voice speaking Japanese, but fail to notice a change in voice alone (Nazzi et al., 1998). Even 5-month-olds fail to notice a change in voice alone (Nazzi, Jusczyk, & Johnson, 2000).

Analysis of the development of voice recognition capabilities requires knowing what aspect of language knowledge drives adult effects of language familiarity on voice recognition. The locus of this effect in adults is still under contention. Some evidence suggests that listeners need to understand a second language to show improvement in recognizing voices speaking that language (Perrachione & Wong, 2007). Then again, in adults, the ability to comprehend a second language may be confounded with knowledge of sound structure. Support for sensitivity to the sound structure of language as a determinant of voice recognition ability is however not strong. The prosodic structure of a language alone, as retained in reiterant speech, does not support an effect of language familiarity on voice recognition (Schiller et al., 1997). Moreover, linguistic similarity between familiar and unfamiliar languages does not seem to modulate the language familiarity effect on voice recognition; English, Spanish, and Chinese speakers all find it equally difficult to identify German speakers (Köster & Schiller, 1997). These combined results thus do not strongly motivate a claim that knowledge of the native phonology supports the ability to discriminate talkers' voices.

The proposal that language comprehension is required for voice discrimination does, however, make a prediction: a native-language advantage for voice discrimination should not be present in the first year of life, before comprehension is present. In the present study we test for the presence of such an advantage. In our first experiment, we examined whether 7.5-month-olds detect changes in voices better when those voices are speaking a familiar rather than an unfamiliar language. At 7.5 months, infants have started acquiring many aspects of the sound structure of their native language (Saffran, Werker & Werner, 2006). If this growing knowledge of phonology is sufficient to support a native-language advantage in voice discrimination, we should be able to observe it at this age. At 7.5 months infants cannot comprehend novel utterances, however. If the ability to comprehend linguistic input is a prerequisite for the native-language advantage as observed in adults, we should not find it at this age.

Experiment 1

Infants were tested on their ability to detect changes in speaker's voices. Besides testing their ability to notice a change only in voice (Voice Change trials), we established their overall sensitivity to changes by including trials in which both voice and language changed (Voice+Language Change trials). In the latter case, all language pairs we used should be easily discriminated, since previous studies have shown that younger infants readily tell the difference between them. Crucially, in the voice-only case, we compared speech in the native language versus a foreign language; if, as we suggested, voice discrimination ability rests on phonological experience, we predict that even at this young age voices will be easier to tell apart when the native language is being spoken.

The voices we used were chosen for their similarity and, for all languages presented, have not shown evidence of individual differences in discriminability in previous studies in which they have been used (e.g. Nazzi et al., 1998; 2000; Ramus et al., 2000); acoustic measures reported below confirm this. Nonetheless, to test for whether acoustic properties of the overall signals could alone support voice discrimination, we also included conditions in which stimuli were played backwards. Reversing speech retains voice quality cues that have traditionally been implicated in voice identification (F0, F0 range, breathiness, speech rate, etc; see Murray & Singh, 1980, for related discussions) but renders uninterpretable both phonetic information in the speech signal (Van Lancker et al., 1985) and the language-particular rhythmic structure that infants use to tell languages apart (Nazzi et al., 1998).

Method

Participants

Seventy-six monolingual Dutch-learning 7- to 8-month-olds (M=225 days; range=215-239) were tested (41 female). The data from 19 additional infants were discarded due to extreme fussiness (16), parental interference (2), and equipment failure (1). Note that dropout rate due to fussiness in the two reversed-speech conditions (7 infants) was very similar to that in the two corresponding forward-speech conditions (6 infants).

Design

Infants were assigned to one of 5 Habituation Conditions: the native language (Dutch), foreign language 1 (Japanese), foreign language 2 (Italian), the native language reversed (Reversed Dutch), and foreign language 1 reversed (Reversed Japanese; see Table 1 for additional details). In each condition, infants were presented with 2 Voice Change trials and 2 Voice+Language Change trials. Order of presentation of these two types of test trials was counterbalanced across participants.

----TABLE 1----

Stimuli

Stimuli were 12 sets (one set per speaker) of 34 unrelated sentences each. Each set was read, in an adult-directed manner, by a different female speaker (four Dutch, four Japanese, four Italian), who were selected for similar voice quality (average fundamental frequency, speaking rate, breathiness, age, etc). Example sentences are given in the Appendix. Acoustic measurements of these stimuli were carried out, and the average values (mean duration, mean amplitude, mean fundamental frequency, SD of fundamental frequency) are listed (along with those for English stimuli to be used in Experiment 2) in Table 2. It can be seen that although there are cross-language differences (e.g., the Italian voices are slightly lower-pitched than the Dutch and Japanese voices), the four voices speaking any one language are, on these averages, very similar.

----TABLES 2 AND 3----

What is important in this case is the relative amount of within-language variability across languages. To address this, we compared for each language pair the variance across speakers of Language A on a given measure with the variance across speakers of Language B on the same measure. The ratio of the two variances allows us to derive an F value with degrees of freedom (n-1,n-1); in this case the df are (3,3), so that the critical F value for significance at $p < 0.05$ would be 9.1. The F values for the three pairs are listed in Table 3 (the first three lines); those for duration, amplitude and mean F0 are far below 9.1, and the Dutch-Italian comparison for standard deviation of F0 was also insignificant. The two F0 standard deviation comparisons involving Japanese were however significant, in both cases because the Japanese standard deviations were, though higher, more similar than the other two sets (see Table 2). That is, the four Japanese speakers were more similar in how much pitch movement they used, while the four Dutch speakers were less similar to one another on this measure, and the four Italian speakers even less similar to one another[1].

If mean pitch, amplitude or duration are of use to infants in discriminating speakers, then these results suggest no difference in how useful they will be across the three language sets here. If the amount of pitch movement is useful, then these data suggest that infants might be able to make relatively more use of this dimension to discriminate Italian speakers, less use of it to discriminate Dutch speakers, and least use of it in discriminating Japanese speakers.

The Dutch sentences served as native-language stimuli (N), while the foreign stimuli were Japanese (F1) and Italian (F2). Two more stimulus sets were created by reversing the Dutch (RN) and Japanese (RF) recordings. Thirty sentences from each set were used as habituation material; the remaining four sentences formed the test material.

Procedure

We used the Visual Fixation Procedure (VFP), which allows assessment of infants' speech discrimination abilities (Werker et al., 1998). In VFP tests of listening discrimination, infants receive auditory stimuli with coincident visual stimuli. If infants are interested in auditory stimuli, their looks to the visual stimuli increase; as they grow tired of listening, their looks diminish. When looking times decrease to a preset criterion, a new auditory stimulus is presented, together with the old visual stimulus. If looking times increase on presentation of the new stimulus, discrimination between old and new auditory stimuli may be inferred.

Infants sat on a caregiver's lap facing a 52" TV monitor, which showed a multi-colored flickering checkerboard during all trials. Accompanying this visual stimulus, loudspeakers presented the speech samples. An experimenter monitored infants' fixation behavior on a separate screen, and relayed looking behavior data to a computer via a keyboard. At each trial's end, checkerboard and audio presentation ceased; a blinking light served to center the infant. Once the infant was focused on the light, the experimenter initiated the next trial by a keyboard press. Stimulus presentation was controlled using HABIT 2000, version 2.2.4 (developed by Les Cohen at the Children's Research Laboratory, University of Texas). Both experimenters and caregivers wore close-fitting headphones and listened to masking music mixed with stimuli used in the experiment to prevent them following the stimulus presentation.

The experiment included two phases: habituation and test. During habituation, three voices speaking one language were played in a cyclic manner, each voice repeating two sentences per 16-second trial. The test phase began once infants' looking times had decreased to 65% of their initial duration (calculated over a sliding window of three trials), or infants had completed 15 such trials. The four test trials (two Voice Change, two Voice+Language Change) were identical in structure to the habituation trials. Thus, the experimenter was unaware of test phase commencement. Infants were randomly assigned to condition (16 per condition, except for condition 3 with 12 participants); in each condition, half of the infants heard Voice Change Trials first, and half heard them second. The specific voice used in test trials was counterbalanced across participants.

Results and Discussion

In the forward speech conditions, infants completed on average 11.15 habituation trials (SD=3.3) before proceeding to test. Similarly, infants in the reversed speech conditions completed on average 11.13 habituation trials (SD=3.6) before proceeding to test. An ANOVA comparing number of habituation trials completed per condition revealed no effect of Condition, $F(4, 71) < 1$.

To assess change detection, we compared mean looking time on the last two habituation trials to mean looking time during test trials. If infants detect that a change has occurred, they should dishabituate on test, i.e. they should look longer in the test trials than in the habituation trials. Figure 1 presents differences in mean looking time from habituation for each type of test trial in each condition.

----FIGURE 1----

For Voice+Language Change (Figure 1, left panel), we predicted that infants would discriminate between samples in all unmodified stimuli. In the three conditions involving unmodified stimuli, 29 of 44 infants looked longer overall in the Voice+Language Change Trials (M=8.3; SD=2.9) than the Last Two Habituation Trials (M=6.54; SD=2.4). In the two conditions involving reverse-language samples, 18 of 32 infants looked longer overall in the Change Trials (M=7.6; SD=2.7) than the Last Two Habituation Trials (M=7.3; SD=2.9). One-way paired t-tests revealed that infants significantly increased looking time from habituation to test in all unmodified conditions, but in no reversed condition: N-F1 [t(15)=2.08, p=.02, d=.5], F1-N [t(15)=2.2, p=.02, d=.5], F2-F1 [t(11)=1.74, p=.04, d=.5], RN-RF [p=.71], RF-RN [p=.28].

For Voice Change (Figure 1, right panel), we predicted better performance in the native language. In the unmodified native-language condition, 12 of 16 infants looked longer overall in Voice Change Trials (M=8.3; SD=2.8) than the Last Two Habituation Trials (M=5.93; SD=1.6). In the remaining four conditions, 31 of 60 infants looked longer overall in Voice Change Trials (M=7.4; SD=3.5) than the Last Two Habituation Trials (M=7.08; SD=2.8). One-way paired t-tests revealed that infants significantly increased looking time from habituation to test only with the unmodified native-language samples N-N [t(15)=3.2, p=.003, d=.8]; no other Voice Change condition exhibited discrimination:, F1-F1 [p=.46], F2-F2 [p=.63], RN-RN [p=.32], RF-RF [p=.32].

These results suggest that 7.5-month-olds are more sensitive to voice changes in the native language than in an unfamiliar language. Given the acoustic similarity between the voices used in this study and the results of the reversed speech voice discrimination task, it is unlikely that greater acoustic dissimilarity between the Dutch voices than Italian or Japanese voices could explain our results. However, as a further check we performed an adult voice recognition task in Experiment 2.

Experiment 2

In Experiment 1, Dutch infants readily noticed voice changes in our Dutch recordings, but failed to notice changes in our Italian or Japanese voices. We interpreted this as evidence that language experience shapes voice recognition in early infancy. The control conditions using speech reversal suggested that differences in acoustic distinctiveness were not present in that manipulation. However, it is still possible that the Dutch voice recordings may have been perceptibly more distinct than our Italian or Japanese recordings in some way masked by the reversal. Therefore, in Experiment 2, we subject this alternative explanation to a more rigorous test, by presenting the same recordings to adult listeners.

Given the evidence, reviewed earlier, that adults discriminate voices best in their native language, testing adults with the same native language as the infants would most likely deliver the same results as we found in Experiment 1. We therefore tested English-speaking adults who knew no Dutch. In order to evaluate the hypothesis independently with the present type of material, we also used English stimuli. The test procedure was a voice line-up based on the study of Goggin et al. (1991), referred to in the Introduction.

Method

Participants

Thirty-six English-speaking undergraduates from the St. George and UTM campuses of the University of Toronto were tested (26 females; Mean Age=21.3 years; Range = 18 to 36). Compensation included extra credit or a $10 dollar payment. All participants learned Canadian English before the age of five. Participants who could converse fluently in any of the three Experiment 1 languages were not included. This does not rule out passive exposure; according to recent census polls, in the Toronto context this is likely to be high for Italian but negligible for Dutch and Japanese (census data at http://www12.statcan.ca for Toronto reports that after English, Italian is the mother tongue most widely represented in the city, while neither Dutch nor Japanese appears among the top 30 mother tongues).

Design

Over 16 trials, each participant was tested on ability to recognize each of the four voices in each of four languages (English, Dutch, Italian, and Japanese). Trials were pseudo-randomized with the restriction that the same language could not be presented for more than two trials in a row.

Stimuli

Four similar-sounding native English-speaking females from Southern Ontario region were recorded reading 10 sentences each. Each speaker read different sentences. The sentences were those used for English in Nazzi et al. (2000), and were closely matched in rate of speech and number of syllables to the Experiment 1 sentences. Acoustic measurements were again made and are reported in Table 1; the variance in the recordings was compared to that in the Experiment 1 sets as before. These analyses showed that the English set also differed from the highly similar Japanese set on the pitch movement measure, but they did not differ from the other two languages on this measure. They did not differ from any of the three other languages on mean pitch, again confirming that the choice of speakers with similar-pitched voices had been successful. The English set was more variable in duration, however, than all three other languages, and slightly more variable in amplitude than the Dutch and Italian sets.

Procedure

Each trial began with the visual prompt “Remember this voice” appearing on the screen. This prompt was followed by exposure to a pair of sentences repeated once by an individual speaker. After exposure to the voice, participants watched a silent puppet show accompanied by instrumental music for approximately one minute. Participants were then presented with a voice “line-up” in which all four speakers of a language produced a pair of sentences. Participants were presented with the visual prompt “Now listen to the line-up and mark the voice you heard 1 minute ago”. Participants then marked on an answer sheet which speaker they thought they had heard during the exposure period at the beginning of the trial.

For each speaker of each language used in the study, one pair of sentences was chosen for use as exposure and eight pairs of sentences were chosen for use in the voice line-up. The voice line-up on any given trial always consisted of two different sentences produced by each of the four speakers of a given language. The sentences used in the voice line-up were unique for each trial, but the ordering of the voices in the line-up remained constant. That is, if Dutch speaker 1 was first in one line-up, the same speaker was first in all other line-ups. But the sentences spoken by Dutch speaker 1 were different each time the line-up was presented (also, the sentences spoken by Dutch speaker 1 were different from the sentences spoken by any other voice in the line-up in any given trial).

Since there were four voices to choose from, chance performance equaled 25% correct. Participants also marked how confident they were in their choice (1=highly confident; 5=not very confident at all). Each participant was presented once with each of the 16 voices[2]. Participants were given a brief break halfway through the experiment.

Results and Discussion

We first examined participants' mean proportion of correct responses for trials involving each language (see Table 4). T-tests revealed that participants performed above chance with all four language sets: English [t(35)=11.1, p<.0001], Italian [t(35)=9.5, p<.0001], Dutch [t(35)=3.6, p=.001], and Japanese [t(35)=6.0, p<.0001]. However, as predicted by the language familiarity hypothesis, participants performed significantly better in English than in the unfamiliar languages, t(35)=4.43, p<.0001. Paired t-tests revealed that English listeners performed significantly better with English than with Dutch [t(35)=4.6, p<.0001] or with Japanese [t(35)=3.7, p=.0007], but the difference in their performance with English and Italian was not significant [t(35)=.63, p=.54].

Next we examined participants' confidence ratings (see Table 4). Participants were more confident in their ability to identify a voice from a lineup if the voices in the set spoke a familiar language. Paired t-tests revealed that our English-speaking participants were more confident identifying speakers of English from a voice line-up than speakers of all three other languages [English vs. Dutch: t(35)=8.3, p < .0001; English vs. Italian: t(35)=3.4, p=.0019; English versus Japanese: t(35)=4.0, p=.0003].

The voice lineup data set further allows us to assess the effect of acoustic parameters on discriminability by comparing the participants' success in identifying each speaker with the acoustic measures made for that speaker's productions. No acoustic measure (of duration, amplitude or F0) proved to correlate with the Experiment 2 results (all p > .25).

In combination, participants' accuracy and confidence rating measures provide support for the language experience hypothesis. English listeners found English voices the easiest to recognize, closely followed by voices speaking Italian, another language with which, given the population statistics, our participants were likely to have had perceptual experience. Crucially, this adult study has also provided further support for the language experience hypothesis as the correct interpretation of our infant study, since even when presented without distortion, the Dutch stimuli provided no acoustic cues that facilitated their identification relative to the other language stimuli. For these adult English speakers, the Dutch voices were the most difficult to recognize.

General Discussion

Seven-month-old infants who have as yet no comprehension of spoken language can successfully tell new voices apart, but only when the voices are speaking a familiar language. Thus linguistic experience feeds into voice recognition very early in life. Language comprehension is apparently not the crucial contributing factor driving language familiarity effects on voice identification.

Although neonates possess the auditory discrimination abilities necessary to tell the mother's voice from other women's voices, further development of an adult-like ability to recognize individual talkers clearly draws on the language-specific knowledge built up in the course of acquiring the mother tongue. The stimuli we used were spoken in an adult-directed manner, so that we could test for the role of such language-specific factors. Infants prefer infant-directed speech – not for prosodic or any other language-specific reasons, but rather because of the affect expressed in such speech (Kitamura & Burnham, 1998). Using adult-directed speech thus made it less likely that infants would respond primarily to (putatively universal) signals of affect, and also made it more likely that the infants would rapidly habituate to the speech materials. The acoustics of the adult-directed sentences were also easy to match; any differences between the four acoustically similar voices in a particular language set were insufficient to support discrimination between them on these acoustic dimensions alone, as our control condition with reversed speech demonstrated. Where the voices were speaking a foreign language, no significant discrimination was possible. But when the speech input was in the mother tongue, infants could successfully tell even these very similar voices apart.

The results of our adult voice recognition study in Experiment 2 lend further support to this conclusion. Using the same recordings as in Experiment 1, we showed that an explanation in terms of acoustic distinctiveness could not account for why the Dutch infants found our Dutch voice recordings easiest to tell apart. English-speaking adults performed relatively poorly at identifying the voices speaking Dutch, though well at identifying the voices speaking in familiar language.

At seven months, infants do not yet utter words to communicate, nor can they recognize spoken utterances except perhaps for a few familiar names (Bortfeld, Morgan, Golinkoff & Rathbun, 2005; Mandel, Jusczyk & Pisoni, 1995; Tincoff & Jusczyk, 1999). Nonetheless, they have amassed, since birth, a formidable amount of relevant experience with the native language. They have learned to distinguish the native language from similarly structured languages and even other variants of the same language (Nazzi et al., 2000). They are poised to begin the construction of a vocabulary, as is evidenced by their ability to store familiarized word forms and recognize them later in a new context (e.g., Juszcyk & Aslin, 1995; van Heugten & Johnson, 2009; see, however, Houston & Jusczyk, 2000; Singh et al., 2004). This experience, it appears, also pays off in developing ancillary communication skills such as voice recognition.

Our results, in combination with the results from our control study with adults, point to some possible answers as to the nature of the language-specific knowledge that supports this skill. First, it is clear that understanding the linguistic message that speakers are communicating is not a necessary prerequisite for voice discrimination; as the examples in the Appendix illustrate, our 7.5-month-old listeners would not have been understanding the content of any of these sentences. Thus the native-language advantage cannot rest crucially on higher-level processing. Second, it is also clear that language rhythm and gross prosodic structure, arguably the linguistic properties used by both infants and adults to distinguish between languages (Nazzi & Ramus, 2003), likewise do not suffice to support voice discrimination. Although the Dutch stimuli in our Experiment 2 were rhythmically and prosodically more like English than the stimuli in the other two languages, the English listeners performed worst with the Dutch stimulus set. In this respect our Experiment 2 results confirmed the earlier findings of Schiller and colleagues (Schiller et al., 1997; Köster & Schiller, 1997) with other language pairings and different methodologies.

Thus language phonology, i.e., the more detailed level of sound structure that infants are acquiring in the second half of the first year of life (Saffran et al., 2006), seems to be the only remaining candidate for the information necessary for successful voice discrimination. Consistent with this is evidence that certain phonemes support voice discrimination better than other phonemes do (Andics, McQueen & Van Turennout, 2007). Furthermore, communicative mastery of the phonology is unnecessary; 7.5-month-olds are acquiring their native phonology at a perceptual level, but they cannot be said to have full mastery of it, nor can they use it as yet in speech production in communicative utterances. In this light consider also the remarkable finding from our adult control study, in which voice discrimination for Italian sentences was not significantly worse than for native-language sentences. None of our participants could converse in Italian, but their residence in Toronto will have likely exposed them to passive perception of this language, spoken by multiple speakers. After some participants in Experiment 2 had volunteered guesses about the languages they heard, the final 16 participants in this experiment were formally asked if they could identify the languages they had just heard. Five mentioned Japanese, only one mentioned Dutch, but 12 mentioned Italian. This could suggest that perceptual familiarity with phonological structure, which we may postulate as holding both for 7.5-month olds acquiring a language and for adults regularly exposed to an environmental language they do not themselves use, is the crucial prerequisite for successful voice discrimination. Further research exploring this suggestion is certainly called for.

Thus our findings indicate that cues to linguistic-particular and speaker-particular information are interwoven, and that perceptual learning from speech shapes voice recognition very early in life. Given that much the same claim has been made for perceptual learning driving the development of both the race and gender effects in early face recognition (Kelly et al., 2000; Quinn et al., 2002), it is tempting to speculate that a common or at least similar underlying learning mechanism may support the development of expertise in different domains (Scott, Pascalis, & Nelson, 2007). Certainly the emergence of language familiarity effects on voice recognition around the middle of the first year of life is consistent with a similar developmental time course. In adulthood, rapid perceptual learning enables adaptation to new talkers, and this learning indeed draws on such a powerful and cognitively general mechanism; the same adaptation first demonstrated in speech perception (Norris, McQueen & Cutler, 2003) is also observable in letter identification (Norris, Butterfield, McQueen & Cutler, 2006) and color categorization (Mitterer & De Ruiter, 2008). In this light, it would be interesting to ask how much and what type of exposure to a second language would suffice to induce early native-like discrimination of voices in a foreign language (see Sangrigoli, et al., 2005, for related discussion on face perception). For example, an infant being raised in an English-dominant country, but receiving daily linguistic input from just a single caregiver speaking another language, might exhibit improved voice discrimination for new speakers of only the caregiver's native tongue, or might not exhibit language specific voice discrimination at all. Future research will be required to establish whether exposure to many speakers of a language is needed to launch voice identification skills.

Finally, the results of the present study demonstrate very sophisticated voice discrimination abilities in infants who have not yet even begun to speak. But at the same time, even school-age children have been argued to have less than adult-like abilities in the domain of voice recognition (Mann et al., 1979). There are several reasons for this apparent inconsistency. First, the effect of language experience reported in this paper reflects months of perceptual learning. Adult-like voice recognition, however, as described in the introduction, draws on many subtle aspects of language-specific structure, and years rather than months may be needed to amass all the required experience. Second, the discrimination we have documented here involves recognition in that a new voice is determined not to be identical to the stored representations of the three voices heard earlier. Mature voice recognition skills, however, require matching heard voices to far richer representations of talker identity. As infants accumulate initial experience listening to different speakers of their native language, they can begin to develop a schema for the parameters, and ranges within those parameters, that most reliably signal talker identity. Regular passive exposure to a language one does not speak may allow exactly this type of knowledge to accrue also. But possession and use of a vocabulary will undoubtedly allow qualitatively and quantitatively more relevant learning.

We conclude that early native-language experience supports more abilities than those which are primarily linguistic; communication skills resting on voice recognition are also facilitated as the native language is acquired.

Acknowledgment: Financial support for Experiment 1 was provided by the Max Planck Society and by the NWO-SPINOZA project "Native and non-native listening" (AC). Conditions 1-3 of Experiment 1 formed EW's graduation project, supervised by EKJ. Support for Experiment 2 was provided by an NSERC awarded to EKJ. A partial report of Experiment 1 was presented to the ISCA Conference on Speech Plasticity, London, UK (June 2005). We thank the families who participated as well as Angela Khadar and Marieke van Heugten for experimental assistance, Holger Mitterer for assistance with our acoustic measures, and Les Cohen for providing the HABIT software.

References

Andics, A., McQueen, J.M., & Turennout, M. van (2007). Phonetic context influences voice discriminability. In J. Trouvain & W.J. Barry (Eds.), *Proceedings of the XVIth International Congress of Phonetic Sciences* (pp. 1829-1832)*,* Saarbrücken, Germany.

Bortfeld, H., Morgan, J.L., Golinkoff, R.M., & Rathbun, K. (2005). Mommy and me: familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16, 298-304.

Chen, G.T. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics*, 2, 159-171.

Collins, B. & Mees, I.M. (1999). The phonetics of English and Dutch. Koninlijk Brill NV, Leiden, NL.

DeCasper, A.J. & Fifer, W.P. (1980). Of human bonding: newborns prefer their mother's voices. *Science*, 208, 1174-1176.

Floccia, C., Nazzi, T., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science*, 3, 333-343.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., & Duchaine, B. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia*, 47, 123-131.

Goggin, J.P., Thompson, CP., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory and Cognition*, 19, 448-458.

Houston, D.M., & Jusczyk, P.W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1570-1582.

Jusczyk, P.W., & Aslin, R.N., (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1-23.

Jusczyk, P., Pisoni, D., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, 43, 253–291.

Kelly, D.J., Quinn, P.C., Slater, A.M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, 18, 1084-1089.

Kitamura, C., & Burnham, D. (1998). The infant's response to vocal affect in maternal speech. In C. Rovee-Collier (Ed.), *Advances in infancy research*, 12, 221-236.

Köster, O. & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition, *Forensic Linguistics*, 4, 18-28.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for perceptually dissimilar vowel categories. *Journal of the Acoustical Society of America, 66*, 1168-1679.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science, 6*, 314-327.

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

Mehler, J., Bertoncini, J., Barrière, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, 7, 491-497.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143-178.

Mitterer, H. & De Ruiter, J.P. (2008). Recalibrating color categories using world knowledge. *Psychological Science,* 19, 629-634.

Moon, C., Panneton-Cooper, R., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, 16, 495-500.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, 47, 379–390.

Murray, T. & Singh, S. (1980). Multidimensional analysis of male and female voices. *Journal of the Acoustical Society of America*, 68, 1294-1300.

Nazzi, T., Bertoncini, J, & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 756-766.

Nazzi, T., Jusczyk, P.W., & Johnson, E.K. (2000). Language discrimination by English-learning 5-month-olds: effects of rhythm and familiarity. *Journal of Memory and Language*, 43, 1-19.

Nazzi, T. & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication,* 41, 233-243.

Nittrouer, S., & Lowenstein, J. (2007). Children's weighting strategies for word-final stop voicing are not explained by auditory sensitivities. *Journal of Speech, Hearing, and Language Research*, 50, 58-73.

Norris, D., Butterfield, S., McQueen, J. M., & Cutler, A. (2006). Lexically-guided retuning of letter perception. *Quarterly Journal of Experimental Psychology,* 59, 1505-1515.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology,* 47, 204-238.

Nygaard, L.C., Sommers, M., & Pisoni, D.B. (1994). Speech perception as a talker contingent process. *Psychogical Science*, 5*,* 42-46.

Perrachione, T.K., Pierrehumbert, J.B. & Wong, P.C.M (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1950-1960.

Perrachione, T.K. & Wong, P.C.M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia,* 45, 1899- 1910.

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, 31, 1109-1121.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, 288, 349-351.

Remez, R.E., & Fellowes, M. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 651-666.

Saffran JR, Werker J, & Werner L. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In: Siegler R. & Kuhn D,, eds. *Handbook of Child Development*. Wiley; New York, 58–108.

Sangrigoli, S., Pallier, C., Argenti, A.M., Ventureyra, V.A.G., & de Schonen, S. (2005). Reversibility of the other-race effect in face recognition during childhood. *Psychological Science,* 16, 440-444.

Schmale, R., Christià, A., Seidl, A., & Johnson, E.K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy,* 15, 650-662.

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibility of early word representations. *Developmental Science*, 12, 583–601.

Schiller, N. O., Köster, O., & Duckworth, M. (1997) The effect of removing linguistic information upon identifying speakers of a foreign language, *Forensic Linguistics*, 4, 1-17.

Scott, L.S., Pascalis, O., & Nelson, C. (2007). A domain-general theory of the development of perceptual discrimination. *Current Directions in Psychological Science*, 16, 197-201.

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language,* 51, 173–189.

Thompson, C. P. (1987). A language effect in voice identification, *Applied Cognitive Psychology*, 1, 121-131.

Tincoff, R., & Jusczyk, P.W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10, 172-175.

Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, 38, 253-265.

Van Heugten, M. & Johnson E.K. (2009). The robustness of infants' early word representations. *Journal of the Canadian Acoustical Association,* 37, 148-14.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. *Journal of Phonetics*, 13, 9-38.

Winters, S.J., Levi, S.V., & Pisoni, D.B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America*, 123, 4524-4538.

Yamazawa, H., & Hollien, H. (1992). Speaking fundamental frequency patterns of Japanese women. *Phonetica* 49, 128-140.

Footnotes

[1] Note that word prosody in Dutch, Italian and English is based on stress, which is realised only partly in F0, while word prosody in Japanese is based on pitch accent, realised in F0 only. The constraints of the pitch accent system may underlie the higher but more consistent F0 variation in the Japanese voices (Yamazawa & Hollien, 1992; see, however, Van Bezooijen, 1995). We thank an anonymous reviewer for this suggestion.

[2] The data from one trial had to be dropped from each of seven participants' responses because, due to technical difficulties, the speech files played during this trial were presented at a very low volume. Thus, seven participants contributed data from only 15 rather than 16 trials. Means for the language set missing one trial were thus calculated based on three trials rather than four for these participants.

Appendix

Example sentences:

Dutch: Een gevoel van enorme opluchting maakte zich van hem meester.

‘He was overcome by an enormous feeling of relief’

Italian: Il grande ristorante ha chiuso per una settimana.

‘The large restaurant has closed for a week’

Japanese: Kochira no kata wa keiseigeka no senmonka desu.

‘This person is a specialist in plastic surgery’

English: Artists have always been attracted by the life in the capital.

Table 1: Infants were randomly assigned to one of the five experimental conditions listed in Table 1. Abbreviations are as follows: native language (N), foreign language 1 (F1), foreign language 2 (F2), reversed native language (RN), and reversed foreign (RF). Half of the infants in each condition heard the Voice+Language Change trials first; the other half heard the Voice Change trials first. All participants heard a total of five voices during the experiment. During the habituation phase, infants heard three voices speaking the same language. During the test phase, infants heard one new voice speaking the same language heard in the habituation phase, or a second new voice speaking another language.

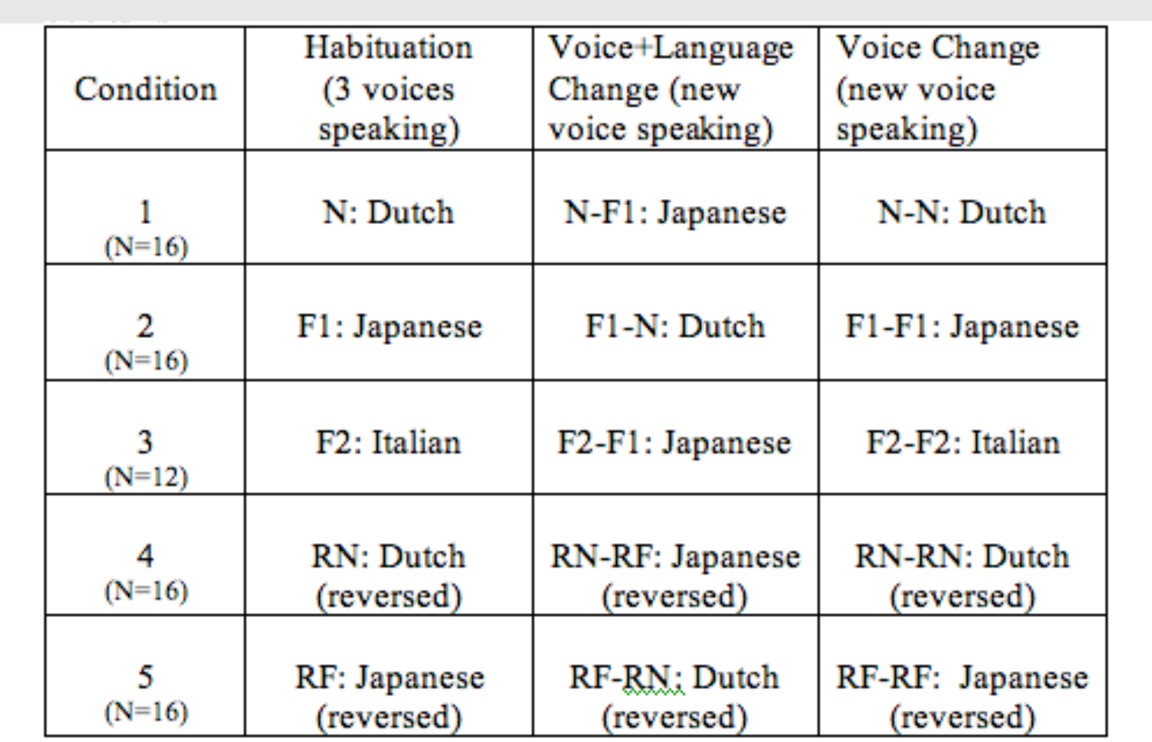
Table 2: Average acoustic measurements for the 12 speakers used in Experiment 1 (Dutch, Japanese, and Italian) as well as the additional 4 voices used in Experiment 2 (English). SD is placed in parenthesis.

Table 3: Comparison of acoustic variability of speakers for each language pair used in Experiment 1 and 2, in F values calculated as the ratio of the two variances; critical F value for significance at $p < 0.05$ would be 9.1. The first three lines compare stimuli used in Experiment 1 and the last three lines compare stimuli used in Experiment 2.

Table 4: Average accuracy and confidence ratings in Experiment 2, broken down by language.

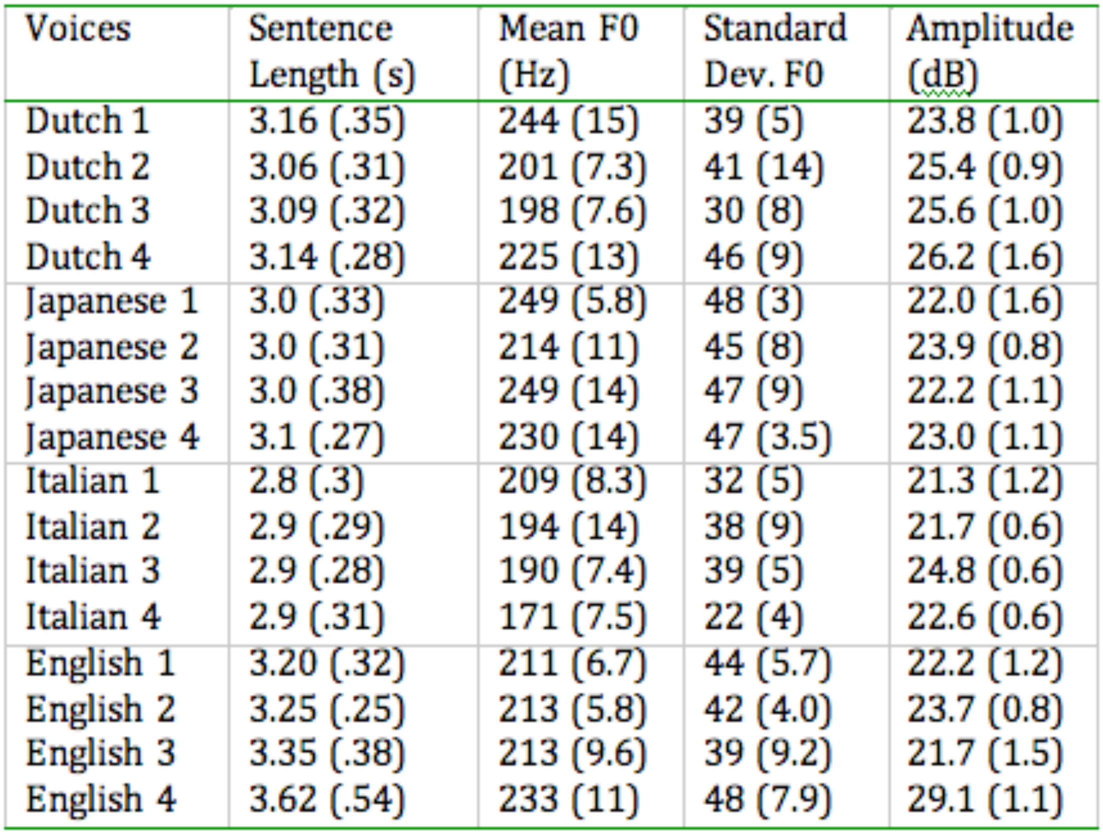
Figure 1: The left panel displays differences in mean looking time to the Voice+Language Change Trials compared to the Last Two Habituation Trials. Conditions are: Native Language changed to Foreign Language 1 (N-F1); Foreign Language changed to Native Language (F1-N); Foreign Language 2 changed to Foreign language 1 (F2-F1); Reversed Native changed to Reversed Foreign (RN-RF); Reversed Foreign changed to Reversed Native (RF-RN). The right panel displays differences in mean looking time to the Voice Change Trials and the Last Two Habituation Trials. Conditions are: native language voices changed to a new native language voice (N-N); foreign language 1 voices changed to a new foreign language 1 voice (F1-F1); foreign language 2 voices changed to a new foreign language 2 voice (F2-F2); reversed native language voices changed to a new reversed native language voice (RN-RN); reversed foreign language 1 voices changed to a new reversed foreign language 1 voice (RF-RF). Error bars represent SE; stars mark conditions differing significantly from zero.

TABLE 1



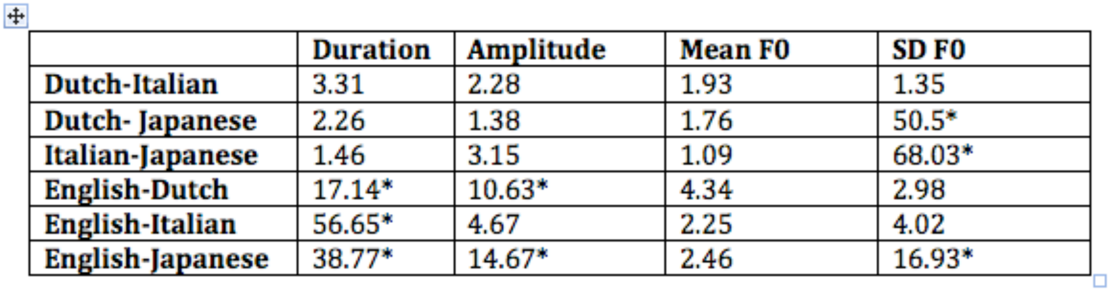
Condition	Habituation (3 voices speaking)	Voice+Language Change (new voice speaking)	Voice Change (new voice speaking)
1 (N=16)	N: Dutch	N-F1: Japanese	N-N: Dutch
2 (N=16)	F1: Japanese	F1-N: Dutch	F1-F1: Japanese
3 (N=12)	F2: Italian	F2-F1: Japanese	F2-F2: Italian
4 (N=16)	RN: Dutch (reversed)	RN-RF: Japanese (reversed)	RN-RN: Dutch (reversed)
5 (N=16)	RF: Japanese (reversed)	RF-RN: Dutch (reversed)	RF-RF: Japanese (reversed)

TABLE 2



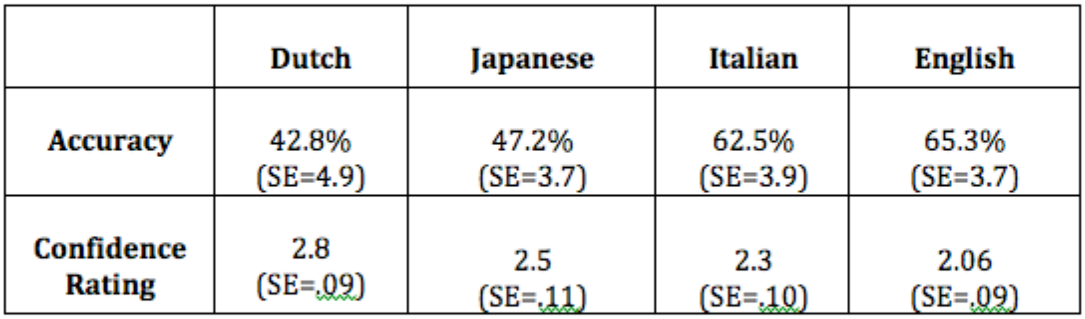
Voices	Sentence Length (s)	Mean F0 (Hz)	Standard Dev. F0	Amplitude (dB)
Dutch 1	3.16 (.35)	244 (15)	39 (5)	23.8 (1.0)
Dutch 2	3.06 (.31)	201 (7.3)	41 (14)	25.4 (0.9)
Dutch 3	3.09 (.32)	198 (7.6)	30 (8)	25.6 (1.0)
Dutch 4	3.14 (.28)	225 (13)	46 (9)	26.2 (1.6)
Japanese 1	3.0 (.33)	249 (5.8)	48 (3)	22.0 (1.6)
Japanese 2	3.0 (.31)	214 (11)	45 (8)	23.9 (0.8)
Japanese 3	3.0 (.38)	249 (14)	47 (9)	22.2 (1.1)
Japanese 4	3.1 (.27)	230 (14)	47 (3.5)	23.0 (1.1)
Italian 1	2.8 (.3)	209 (8.3)	32 (5)	21.3 (1.2)
Italian 2	2.9 (.29)	194 (14)	38 (9)	21.7 (0.6)
Italian 3	2.9 (.28)	190 (7.4)	39 (5)	24.8 (0.6)
Italian 4	2.9 (.31)	171 (7.5)	22 (4)	22.6 (0.6)
English 1	3.20 (.32)	211 (6.7)	44 (5.7)	22.2 (1.2)
English 2	3.25 (.25)	213 (5.8)	42 (4.0)	23.7 (0.8)
English 3	3.35 (.38)	213 (9.6)	39 (9.2)	21.7 (1.5)
English 4	3.62 (.54)	233 (11)	48 (7.9)	29.1 (1.1)

Table 3



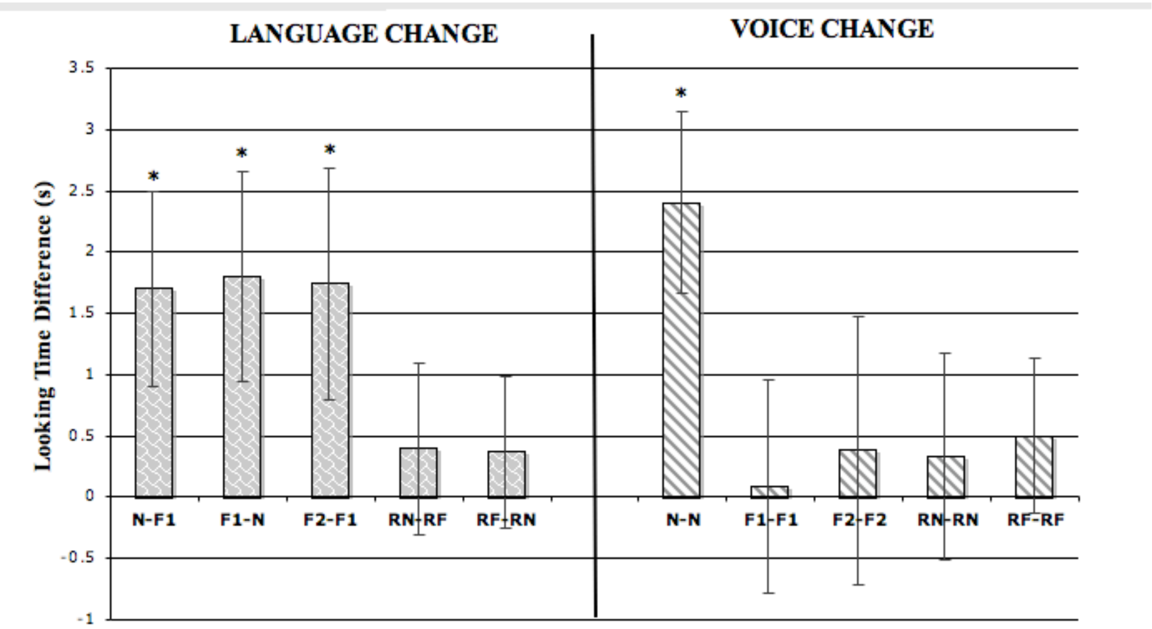
	Duration	Amplitude	Mean F0	SD F0
Dutch-Italian	3.31	2.28	1.93	1.35
Dutch- Japanese	2.26	1.38	1.76	50.5*
Italian-Japanese	1.46	3.15	1.09	68.03*
English-Dutch	17.14*	10.63*	4.34	2.98
English-Italian	56.65*	4.67	2.25	4.02
English-Japanese	38.77*	14.67*	2.46	16.93*

Table 4



	Dutch	Japanese	Italian	English
Accuracy	42.8% (SE=4.9)	47.2% (SE=3.7)	62.5% (SE=3.9)	65.3% (SE=3.7)
Confidence Rating	2.8 (SE=.09)	2.5 (SE=.11)	2.3 (SE=.10)	2.06 (SE=.09)

Figure 1


LANGUAGE CHANGE
VOICE CHANGE
Looking Time Difference (s)
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
*
*
*
*
N-F1
F1-N
F2-F1
RN-RF
RF-RN
N-N
F1-F1
F2-F2
RN-RN
RF-RF