The storage of indexical information across segmental length utterances

Richard J. File-Muriel* **y Samuel Turiciano****
Universidad de Nuevo México

Recibido: 23-05-2012 • Aceptado: 31-05-2012

Abstract

This study examines the ability of listeners to store and recall indexical properties in segmental-length utterances, including details regarding the socially-constructed category "gender" as well as information about individual voices. The idea that indexical properties are irrelevant to speech recognition is a core assumption of generative phonological theory, which emphasizes the role of abstraction and categorization in identifying symbolic-like phonemic strings that are serially ordered. On the other hand, Pisoni (1997) and subsequent work have shown convincingly that this is not the case; indexical information is highly relevant and stored in lexical memory, evidenced by the fact that speakers recognize words uttered by familiar voices faster than unfamiliar voices. This study reports on a simple listening task, in which participants heard segmental-length stimuli (a-i-r-l-m-n-z-s) produced by both familiar and unfamiliar voices. Our results show that listeners store information regarding both gender classification and individual gestural behavior on utterances even as small as the segment. Higher correct identification scores are reported for the voiced sound /z/ than the voiceless /s/, indicating that listeners store information regarding individual speakers' fundamental frequency and/or vocal cord physiology. At the same time, the identification scores for the voiceless sound /s/ was well above chance, indicating that listeners also store information regarding the configuration of individual speaker's oral tracts. Our findings contribute to the growing body of research that phonological representation goes well beyond serially-ordered abstract symbols; it is rich and detailed (Pisoni, 1997; Port, 2010). At the same time, our results could also be taken as supportive of the Motor Theory of phonological representation, given that our results indicate storage of physiological differences and gestural properties of meaningless sounds (Galantucci et al., 2006). We also report an unexpected, albeit preliminary finding: Female listeners seem to recall more accurately acoustic detail regarding other female voices, as their identification scores for female stimuli were significantly higher than the scores achieved by male listeners.

Key words: Speech perception, indexicality, gender, phonology.

* Ph.D. in Hispanic Linguistics. Assistant Professor, University of New Mexico. Department of Spanish and Portuguese; College of Arts and Sciences. Correo institucional: richfile@unm.edu

** MA in Hispanic Linguistics. Department of Spanish and Portuguese; College of Arts and Sciences at Indiana University Bloomington. Correo electrónico: motalian@msn.com

Introduction

Speech recognition and how humans process auditory stimuli have been widely studied in Linguistics, Psychology, and the Cognitive Science. For years, the generative tradition (Halle, 1954; Chomsky & Halle, 1968; Kenstowicz & Kissebert, 1986; Prince & Smolensky, 1993; McCarthy, 2008; among many others) has focused nearly exclusively on the representation of abstract phonemic units, with little attention to other details that are regarded as "grammatically irrelevant" (Chomsky & Halle, 1968:3)[1]. According to such a view, speech is processed in similar fashion to an orthographic code of discrete symbols. The symbolic abstractionist model searches for elements of invariance, which are allegedly shared across speakers and allow for recursion to take place: "one component of the grammar must have a recursive property; it must contain certain rules that can be applied indefinitely often, in new arrangements and combinations, in the generation (specification) of structural descriptions of sentences" (Chomsky & Halle, 1968, p. 6). The recursivity of invariant language is what enables strangers to efficiently communicate information on unpredictable topics, using different dialects, speaking rates, and mediums, without much decrease in performance. This abstractionist phonological model is quite tempting, especially given the feasibility of modeling language computationally.

However, there appears to be many more factors at work during communication, which extend well beyond the high level of abstraction proposed by most generative phonologists. This probably explains why "even the most sophisticated state-of-the-art speech-recognition systems cannot compare to the speed and efficiency of the human listener" (Pisoni, 1997). An alternative explanation put forth by Pisoni and colleagues, Port, and others is that lexical representation is *crucially* highly redundant and detailed; indexical properties of the human voice play a huge role in speech perception. Our everyday life experiences remind us that indexical properties are processed and stored in memory, enabling us to deduce information about the speaker (with a fairly high degree of accuracy), such as his/her age, gender, ethnicity, dialect, mood, intention, whether or not s/he's a smoker, inebriated, sick, just woke up, etc.

In his 1997 article, *Some Thoughts on "Normalization" in Speech Perception*, Pisoni argues against the widely-held belief that speech is understood and represented as abstract units or phonemes, which contain all of the crucial information necessary for comprehensible communication. He labels this the "abstractionist" or "analytic" theory, which searches for invariance in order to identify specific categories that encode the linguistic content of speech into abstract symbolic units: "The assumption of an idealized symbolic representation for spoken language has encouraged researchers to search for simple first-order physical invariants and to ignore the problem of stimulus variability in the listener's environment." He states that within this variability lies the critical information that is needed for communication. The aspect of variability should not be regarded simply "as a troublesome source of 'noise' in the acoustic signal." Instead, it is precisely these instances of variability and fine details like the speakers voice, speaking rate, and environment that are encoded in the memory of the listener and perceived during communication.

Nygaard, Sommers, and Pisoni (1994) carried out a small word recognition (intelligibility) experiment to determine how familiarity with a talker's voice affects the perception of spoken words. Two groups of subjects learned to explicitly identify a set of unfamiliar voices over a 9-day period. After all subjects learned to recognize the voices with a high degree of accuracy, they were

[1] For a more detailed critique, the reader is directed to Port, 2010.

presented with a novel word set. Group 1 heard the samewords produced by new talkers (with no previous exposure), while Group 2 heard the words produced by the familiar voices from the training period. Subjects were required to identify the words, rather than just recognize the voices. They found that the subjects who heard familiar voices were able to recognize the test words more accurately than the subjects who heard unfamiliar voices. This demonstrates, not only that exposure to a talker's voice facilitates subsequent perceptual processing of novel words, but also that listeners store detail about speakers voices that influence speech perception.

Although the present study focuses exclusively on auditory stimuli, it should be noted that speech perception and word recognition are not merely auditory. In their report on a lip-reading experiment, Altieri, Pisoni and Townsend (2011) remind us that that visual speech cues provided by the talker's face, specifically lip reading, has facilitatory effects in terms of accuracy across a wide range of auditory signal-to-noise ratios. They remind us of Sumby and Pollack's (1954) seminal study, in which they report that the obtained benefit from the visual speech signal is related to the quality of the auditory information, with more noticeable gains observed for lower signal-tonoise ratios. Although visual communicative behavior is not addressed in the present study, the findings from these studies, as well as others during the last 50 years, are extremely important as they evidence the richness of language representation, which includes crucial visual information that supports the auditory signal.

We have only touched on a handful of studies looking at word recognition and speech perception; for a more thorough review of the literature and recent advances in this area, the reader is directed to Pisoni and Remez (2007) and the studies within. To our knowledge, the past research in these areas has focused primarily on recognition of the prosodic word and/or phrasal-length utterances. The present study takes segmental-length utterances as a focal point to address three related research questions: 1) Are indexical properties/ individual voice information encoded on utterances as small as the segment?, 2) If so, do differences across segments exist, insofar as the degree to which they encode indexical properties and/or individual voice details? And 3) What is the role (if any) of the sonority hierarchy (Clements 1990)? That is, as sonority increases, does the storage of indexical/individual detail increase?

Methodology

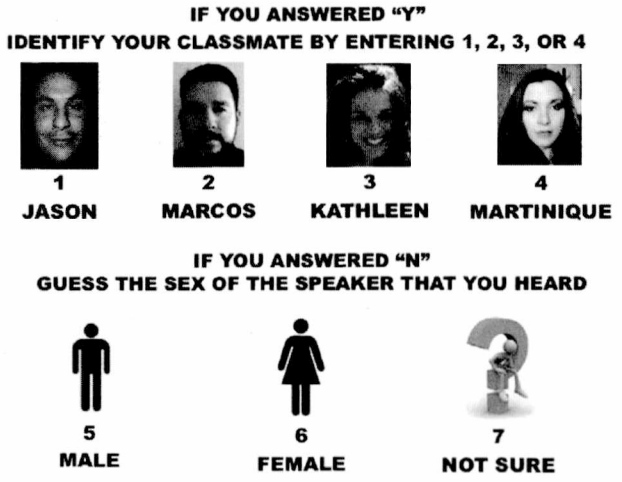
15 participants (eight females; seven males) were recruited from the same beginning-level Spanish class at the University of New Mexico. The class met two times per week for 14 consecutive weeks, with each class period lasting approximately 75 minutes. The experiment stimuli were created by six students. Four students (two males/ two females) from the same class produced the "familiar" stimuli, while two students selected from outside the class produced the "unfamiliar" stimuli. The four students who produced the "familiar" stimuli were rated by the instructor as "active participants." Since the experiment took place in week 14, the participants all had significant previous exposure to their voices.

To create the stimuli for the experiment, six students were recorded while producing the following eight phones: a-i-r-l-m-n-z-s. These phones were selected using the sonority hierarchy as the guiding principle, which holds that phonological systems are organized based on their level of sonority (Clements, 1990) represented here from highest to lowest: vowels (a i), liquids (r l), nasals (m n), and obstruents (z s). Each sound was produced consecutively and held for at least one second. To avoid any possible coarticulatory effects, the students were instructed to wait for two seconds before proceeding to the next sound.

The recordings took place in a quiet laboratory setting using a Zoom H4N solid-state digital recorder, which was equipped with a head-mounted unidirectional microphone. Using Praat, each sound was delimited to approximately 200 ms in order to create stimuli of equal duration; the sounds were extracted as 48 separate wave files.

The experiment was designed and executed using the software SuperLab 4.5 (Cedrus), which was installed on a MacBook Pro equipped with noise-cancelling head phones. The participants were given a practice trial in order to familiarize them with the experimental design and specific response keys. The eight stimuli (a-i-r-l-m-n-z-s) used in the practice trial were created using one of the researcher's voice, rather than the actual six voices used for the experimental trial in order to avoid potential priming effects. For the actual experimental trial, the stimuli were randomized for each participant. The participants would hear the stimuli and press the "Y" key for "yes, I recognize the voice" or the "N" key, for "no, I don't recognize the voice." They also had the option to listen to the sound as many times as they wanted before making their selection. If "Y" was selected, the participants were then directed to identify the specific classmate that produced the sound. In order to do that, a picture of each of the four "familiar" participants was presented to them with labels 1-4 below each picture. If they did not recognize the voice and selected "N," they were given the opportunity to identify the gender of the speaker, by selecting male (5), female (6), or not sure (7). The experiment lasted anywhere from seven to ten minutes. This format allowed us to assess the participants' ability to make both individual voice judgments about the sounds they heard, as well as gender categorizations. For example, if a participant made an incorrect individual voice judgment, say selecting female 3 when s/he actually heard female 4, we would count this as "incorrect" for individual voice, but "correct" for gender.

Figure 1. Individual voice recognition task



Source: The authors.

Results and discussion

In Table 1, we report the gender categorization results first, as that will help inform our results regarding individual voice recognition. If gender categorization were completely by chance, the %-correct would be straddling the 50-percent line. As we can see, that is clearly not the case. Overall, gender categorization was highly accurate (94.,3-percent), indicating that listeners clearly store indexical properties on utterances as small as the segment. Interestingly, the most notable decrease in accuracy was for the voiceless phone /s/, which we take as suggestive that gender classification is highly dependent on the vocal cord physiology (or fundamental frequency) of the speaker. Even so, gender categorization for the voiceless sound /s/ is still well above chance (67-percent). Given that males generally have longer and more open vocal tracts than females, it is likely that the acoustic manifestations of these physiological differences are stored and generalized. In fact, lower center of gravity / centroid measurements for fricative consonants have been reported for males (c.f. Forrest et al, 1998; Silbert & De Jong 2008; File- Muriel & Brown, 2011).

Table 1. Categorization of gender by phone

Phone	Correct	Incorrect	Total	%- Correct
s	60	30	90	66.,7%
n	85	5	90	94.,4%
m	86	3	89	96.,6%
a	88	2	90	97.,8%
i	87	1	88	98.,9%
z	89	0	89	100.,0%
l	88	0	88	100.,0%
r	90	0	90	100.,0%
Totals	673	41	714	94.,3%

Table 2. Recognition of individual voices by phone

Phone	Correct	Incorrect	Total	%-Correct
s	16	74	90	17.,8%
n	41	49	90	45.,6%
z	44	46	90	48.,9%
l	47	43	90	52.,2%
i	50	38	88	56.,8%
m	52	37	89	58.,4%
r	54	36	90	60.,0%
a	56	34	90	62.,2%
Totals	360	357	717	50.,1%

In Table 2, we report the results for individual voice recognition. It should be pointed out that if individual voice recognition were completely by chance, the %-correct would be roughly 16.6-percent, given that there were six individuals to pick from. However, given our results in Table 1, it's clear that listeners do store information regarding gender classification. At the risk of over-simplifying (i.e. ignoring the /s/ results, as well as the fact that /n m a/ and /i/ were not quite 100-percent accurate for gender), we can roughly place the chance cutoff at 33.2-percent. The assumption is that listeners are able to simplify their individual choices by using gender categorization to reduce their selection to three candidates, rather than six. The results in Table 2 are quite revealing. First, listeners *do* store detail regarding the physiology and/or acoustic details of sounds produced by individual speakers and are able to recall this information from memory. Overall, listeners performed well above chance (50.1-percent), regardless of which cutoff line we choose (i.e. gender-informed = 33.2-percent; or individual only = 16.6-percent). Second, the results regarding the voiceless sound /s/ suggest that individual voice recognition is highly informed by the physiology of the laryngeal cavity, as this was the only sound with little-to-no laryngeal activity.

Our original hypothesis was that the sonority hierarchy (Clements 1990) would be reflected in the storage and recall of individual voices and gender categorization. Tables 1 and 2 together seem to suggest that the sonority hierarchy per se is not relevant, as we'd expect to see an increase in accuracy as we move along the continuum of low-to-high sonority (e.g. s-z-m-n-l-r-i-a). In general, this is not the case. For instance, /m r/ led to more success than the vowel /i/, which is higher on the sonority scale. Another example is that the obstruent /z/ led to 100-percent accuracy in gender categorization, even though obstruents are supposed to be lowest on the sonority scale. On the other hand, vocal cord activity in the laryngeal cavity -—typical of voiced sounds-— does seem to inform individual voice recognition and gender categorization. There appears to be a strong voicing effect: Both individual voice recognition and gender identification are dependent on vocal cord activity. On the other hand, speakers are able to correctly categorize gender well above chance even with the voiceless sound /s/ (67-percent), indicating that listeners do perceive subtle differences of the gestural behavior in the oral tract.

Our third and fourth questions compare differences across gender. Our hypothesis was that there should be no gender difference in the listener's

ability to identify individual voices and/or make gender categorizations. The results reported in Table 3 are quite surprising and warrant further investigation: Female listeners were more accurate than their male counterparts in their recognition of individual voices (females 53-percent; males 47-percent). It is also of interest that the gender gap seems to disappear for gender categorization, as females were only slightly more accurate (females 95.0-percent; males 93.4-percent); the difference of 1.6 did not reach statistical significance.

Table 3. Individual voice recognition by gender

Participant	Correct	Incorrect	Total	%-Correct
Male Listener	158	178	336	47.,0%
Female Listener	202	179	381	53.,0%
Totals	360	357	717	50.,1%

In order to better understand the gender disparity, we asked ourselves: Why did female listeners outperform their male classmates on the individual voice recognition task? At this point, we do not have a conclusive answer to this question. However, it naturally led us to examine the relationship between the listener's gender and the gender of the voice stimuli (henceforth stimuli gender) to determine whether or not this relationship impacts individual voice recognition. For example, would females be more successful in identifying other female voices compared to male voices? Our hypothesis was that there would be no significant difference between males, females, and their ability to recognize individual voices. The results in Table 4 indicate that this is not quite the case.

Along with Table 3, the results reported in Table 4 suggest that, in general, female listeners may store more detail in their episodic memory regarding individual voices. Clearly, further investigation is required given the limited participant pool (15 participants). Perhaps the most striking finding in these data is that male listeners clearly experience more difficulty identifying their female classmates' voices (43.,5-percent correct) than those of their male classmates (50.,0-percent correct). Again, further investigation is needed. One possible explanation is that one or more of the male participants interacted more with the male talkers outside of the classroom than with the female talkers. Unfortunately, this is not verifiable at this point.

Conclusions

The results from this experiment establish clearly that both indexical properties, such as gender categorization, as well as details about individual voices are stored in the memory of listeners, supporting work by Pisoni, Nygaard, Sommers, Port, and others that phonological representation is much richer and detailed than proposed in abstractionist models. At the same time, our results establish that listeners are able to perform these tasks with a high degree of accuracy even on segmental-length utterances, which, unlike the prosodic word, are semantically meaningless. This

Table 4. Individual voice recognition: Comparison of listener's gender and stimuli.

Listener's genderGender- Stimuli Ggender	Correct	Incorrect	Total	&-Correct
Female Llistener- Female sStimuli	101	89	190	53.,2%
Female listenerListener- Male Sstimuli	101	90	191	52.,9%
Male Llistener-Male stimuliStimuli	83	83	166	50.,0%
Male listenerListener-Female-stimuliStimuli	73	95	168	43.,5%
Totals	358	357	715	50.,0%

suggests that such information is not necessarily restricted to lexical storage, but is also accessible through the listener's episodic memory of other individual's gestural behavior (c.f. Gallantucci et al., 2006).

The results from this study reveal a strong voicing affect, as gender identification and voice recognition are highly dependent on vocal cord activity. The voiceless sound /s/ presented listeners with the most difficulty insofar as their ability to identify individual voices and their ability to categorize the speaker's gender. In short, listeners were unable to reliably identify individual speakers for this sound. At the same time, they were still able to correctly identify the speaker's gender well above chance (67-percent), indicating that subtle differences in the oral tract are also stored and utilized for gender categorization.

Finally, the differences that we report across genders require further investigation. This study suggests that female listeners may store more individual voice detail than male listeners, as the female listeners significantly outperformed the males on this task. Given the limitations on the participant pool (i.e. only 15 participants), further investigation is necessary. A more robust finding, however, is that male listeners exhibited significant difficulty in correctly identifying their female classmate's voices (50-percent for male voices vs. 43.,5-percent for female voices). Although we offer a feasible sociological explanation, namely that the male listeners may have interacted more with their male classmates (either outside of class or in group activity), it clearly lacks empirical support and will need further investigation.

References

Altieri, N. A., David, B. Pisoni & James T. (2011). Some normative data on lip-reading skills. *Journal of Acoustical Society of America*, 130, 1-4. Townsend.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Clements, G. (1990). *The role of the sonority cycle in core syllabification*. Papers in Laboratory Phonology 1: Between the Grammar and Physics of Speech, ed. by J. Kingston & M. Beckman, 288-333. New York: Cambridge University Press.

Forrest, K., Weismer, G., Milenkovic, P. & Ronald N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of Acoustical Society of America,* 84, 115-123.

Gallantucci, B., Carol, A., Fowler & M. T. Turvey. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, 13, 361-377.

Halle, M. (1954). The strategy of phonemics. *Word,* 10, 197-209.

Hankamer, J., & Aissen, J. (1974). The sonority hierarchy. *Papers from the Regional Meetings*, Chicago Linguistic Society, 131-145.

Kenstowicz, M. & Kissebert, Ch. (1986). *Generative Phonology: Description and Theory*. New York: Emerald Group Publishing Limited.

Koenig, L. (2000). Laryngeal Factors in Voiceless Consonant Production in Men, Women, and 5- Year-Olds. *Journal of Speech, Language, and Hearing Research,* 43, 1211-1128.

Lekach, A. (1979). Phonological markedness and the sonority hierarchy. MIT. *Working Papers in Linguistics*, 1, 172-177.

Levitt, A., Healy, A., & Fendrich, D. (1992). Syllable-internal structure and the sonority hierarchy: Differential evidence from lexical decision, naming, and reading. *Haskins Laboratories Status Report on Speech Research*, 109-110(-), 73-88.

Lucero, J. & Koenig, L. (2005). Phonation thresholds as a function of laryngeal size in a two-mass model of the vocal folds (L). *Journal of Acoustical Society of America* 118.2798-801.

McCarthy, J. (2008). *Doing Optimality Theory*. Oxford: Blackwell.

Ohala, J. (1990). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. *Papers from the Regional Meetings, Chicago Linguistic Society*, 2, 319-338.

Pisoni, D. B. (1997). Some thoughts on normalisation in speech production. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.

Pisoni, D. B. & Robert, R. (eds) 2007. The Handbook of Speech Perception. *Blackwell Handbooks in Linguistics*. Malden, MA: Blackwell Publishing.

Pisoni, D., Paul A. Luce, (1987). Acoustic-phonetic representations in word recognition, *Cognition*, Volume 25, Issues (1-2), March1987, Pages 21-52.

Port, Robert F. 2010. Rich memory and distributed phonology. *Language Sciences* 32, 43-55.

Prince, A. & Paul, S. (1993). Optimality Theory: Constraint interaction in generative grammar. Rutgers: Rutgers Optimality Archive, 537-0802.

Silbert, N. & Kenneth de Jong. (2008). Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production. *Acoustical Society of America* 123, 2769-2779.

Stephen, P. (2002). Quantifying the sonority hierarchy (January 1). *Electronic Doctoral Dissertations for UMass Amherst*. Paper AAI3056268.

Sumby & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society of America*, 26, 12-15.