The Identification of Non-Native Speech Sounds with Psychophysical Training

Jordan Richard Schoenherr, John Logan, and Cassandra Larose
psychophysics.lab@gmail.com, john_logan@carleton.ca, clarose@connect.carleton.ca
Department of Psychology, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada

Abstract

Many early studies of speech perception assumed that the psychophysical properties of speech sounds were unavailable during stimulus identification. When participants are presented stimuli along an acoustic continuum and perform a 2AFC, identification performance is well-described by a logistic function suggesting two discrete phonemic categories. These phonemic category boundaries are believed to bias the classification of speech sounds from non-native languages, reducing the ability to detect acoustic differences. Using a brief period of psychophysical training and a phoneme identification task, participants were sensitized to differences along a voice-onset time (/b-p/) continuum enabling the identification of a non-native phoneme (/p^h/). Finally, participant confidence reports suggested that they were generally unaware of their capacity to accomplish the task. Results demonstrate that participants could use psychophysical properties of the stimuli to identify non-native speech sounds while subjective confidence reports indicated that they had varying degrees of awareness of the psychophysical properties of the non-native speech sounds.

The acquisition of non-native phonemic categories by listeners can be difficult due to interference from prior linguistic experience. Acoustic differences between speech sounds that listeners might otherwise be capable of detecting become undifferentiated due to the development of native language phoneme categories within the first year of life (Werker, 1989) with some suggesting that after infancy, adults lose the ability to discriminate these sounds (Eimas, 1975). For instance, along the voice-onset time (VOT) continuum, English listeners learn to only differentiate /b/ and /p/ based on the difference between the acoustic cues associated with aspiration and the vibration of the vocal cords. When presented with stimuli from the non-native phoneme category portion of this continuum (VOT < -30), corresponding to the Thai /p^h/ category, adult monolingual English listeners would not be capable of distinguishing these phonemes from the neighbouring /b/ speech sounds. Results such as these were initially taken as suggesting that adult listeners no longer had access to these acoustic properties (e.g., Liberman, Harris, Hoffman, & Griffith, 1957). However, this conclusion was later challenged by findings that response times varied with acoustic differences within categories (Pisoni, 1973; Pisoni & Tash, 1974) suggesting that the amount of evidence that needed to be accumulated was dependent on both the phonemic and acoustic properties of the stimuli. In the present study, a psychophysical training technique developed by Pisoni, Aslin, Perey, and Hennessy (1982) was used to assess listeners' awareness of acoustic differences during the development of the non-native speech category /p^h/.

In an ecologically valid setting of language learning, listeners are presented with a number of natural speech tokens within their linguistic environment. One extensively studied example of this is the acquisition of the /r-l/ liquid distinction by Japanese listeners. Specifically, MacKain, Best, and Strange (1981) required listeners to identify and discriminate (in AXB and oddity tasks) synthetic speech sounds along this continuum, American listeners produced sharper category boundaries and more accurate discrimination performance than Japanese listeners with or without 'intensive training'. Studies using natural tokens produced by native and non-native speakers also produce similarly poor performance

(Sheldon & Strange, 1982). However, training techniques have proven effective in training Japanese listeners by increasing stimulus variability and task demands (Logan, Lively, & Pisoni, 1991).

Attention is required in order to learn to discriminate and identify stimuli along a perceptual dimension (Nosofsky, 1986). Jusczyk (1992) argued that portions of an acoustic continuum that are typically left unattended can be accurately discriminated only when selective attention is allocated to those physical dimensions. Evidence in support of these claims comes from psychophysical training techniques developed by Pisoni, et al. (1982) wherein native English listeners were presented with three exemplar speech sounds from regions along the VOT continuum corresponding to voiceless unaspirated, voiced aspirated, and voiceless aspirated stops (i.e, /p/, /b/, and /p^h/, respectively). Listeners were asked to identify each speech sound and received feedback over multiple days of training. In contrast to earlier studies (e.g., MacKain et al., 1981), Pisoni et al. (1982) observed that listeners could discriminate the non-native /p^h/ category from the neighbouring /b/ category with little to no confusion with the /p/ category. These findings suggest that the focus of selective attention can facilitate acquisition of non-native phonemes and that participants can develop this new category of speech sounds in contrast to the neighbouring category (i.e., /b/). An open question that remains, however, is whether listeners have an awareness of the properties of this new phoneme category or whether instead the learning that has occurred is below the threshold of subjective awareness.

Participants' subjective awareness is assessed by having them provide estimates of the subjective probability that they provided the correct answer. When using numerical scales in a two-alternative forced choice (2AFC) task, 50% represents a response associated with a guess whereas 100% represents absolute certainty in a response. Those listeners that assign their subjective probabilities appropriately to their responses (e.g., p(cor) = 0.5 and mean(conf.) = 50%) are said to be perfectly calibrated. In contrast to this ideal, participants in these experiments are frequently miscalibrated. In perceptual tasks, underconfidence is typically observed with accuracy exceeding confidence (e.g., Bjorkman, Juslin, & Winman, 1993) whereas in tasks assessing general knowledge, overconfidence is typically observed with confidence exceeding accuracy (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991). These findings have been explained by suggesting that underconfidence might result from a lack of awareness of one's perceptual system (Dawes, 1980) or that task difficulty produces differences in subjective calibration (Lichtenstein & Fischhoff, 1977). Overconfidence has also been observed in perceptual tasks. Using a perceptual discrimination task, Baranski and Petrusic (1994) observed overconfidence when participants were presented with pairs of line-lengths that were difficult to discriminate. Such findings also provide evidence against single-process accounts of confidence based solely on the primary decision processes (e.g., Ferrel & McGooey, 1980) while suggesting that additional processes are required to compute a confidence report (for a review, see Baranski & Petrusic, 1998).

The formation of non-native phoneme categories by adult listeners suggests that listeners maintain a capacity to perceive acoustic information from a speech signal (cf. Eimas, 1975) even though their proclivity is to not attend to such information (Werker, 1989). In the present study, we examine whether listeners possess an awareness of these acoustic differences by having them rate their subjective confidence. For the purposes of this study we assume that the information accumulated for one phoneme category (e.g, /p^h/) is contrasted against that accumulated for the neighbouring category along the VOT continuum (e.g., /b/). Thus, when presented with a stimulus the probability of a guess represents 50% thereby allowing the use of a standard 6-point scale from guess (50%) to certainty (100%). If participants are aware of acoustic differences, their subjective confidence should differ across regions of the VOT continuum as evidenced by miscalibration. If underconfidence is evidenced, it suggests that

listeners did not have subjective awareness of a well-defined phonemic category whereas if overconfidence is evidenced, it suggests that listeners believed that they had a better understanding of the phonemic category then they in fact did.

Method

Nine Carleton University students participated in the study for course credit; all were native speakers of English or had extensive experience with English and reported normal hearing and no speech pathologies. Fifteen synthetic speech stimuli were used, obtained from the Haskins Laboratories website (HL, 2011; Lisker & Abramson, 1967). These stimuli varied along the VOT continuum from -70 to 70 ms VOT. As per the method used by Pisoni et al. (1982), listeners were presented with stimuli which corresponded to the prevoiced phoneme category /p^h/ for those stimuli in the negative VOT range, and the /b/ and /p/ phoneme categories for the remainder of the range. The latter categories are present in English while the former is not. The sounds were originally recorded on reel-to-reel tape and later converted into AIFF format at Haskins Laboratories. Stimuli were pre-processed using a DC offset correction to eliminate clicks present in the AIFF versions and then converted into WAV files.

Procedure

Modelled after Pisoni et al. (1982), listeners were presented with a brief training block in which they were presented with three stimuli prior to the identification tasks, one from each region of the VOT continuum (-70, 0, and 70 ms VOT, corresponding to the /p^h/, /b/, and /p/ categories). Ten replications of these stimuli were presented in the order indicated. In the following training identification (ID) task, listeners were provided with feedback. They were presented with a stimulus and then reported whether it was a /p^h/, /b/, or /p/ using the 'V', 'B', or 'N' keys, labeled as '_B', 'B', and 'P', respectively. After they had indicated their response, 'Correct' or 'Incorrect' was presented visually on the screen as per the response accuracy. Listeners completed a total of 80 trials in the training task.

In the following ID task, listeners again identified the stimulus presented as a /p^h/, /b/, or /p/ using the keyboard. In the first block, after they completed each ID trial they also indicated their level of confidence in their response using the 'E' through 'I' keys, on a 6-point scale with 50% representing a guess and 100% representing certainty. In the second block, confidence was not reported. Each block was composed of a total of 150 trials.

The duration of the experiment was approximately 30 minutes. Listeners were presented with the stimuli over headphones using PsychoPy software (Peirce, 2007).

Results

Proportion Identification. Unlike studies that have examined 2 category identification performance using confidence reports (e.g., Schoenherr, Logan, & Larose, 2012), only the /p/ phoneme category showed a sharp identification function (Figures 1a and 1b). In general, however, listeners could consistently identify stimuli associated with the /p^h/ and /b/ category with greater than chance accuracy (i.e., stimuli with VOTs of -70, -60, -50, 0, and 10) indicating that even with a brief period of psychophysical training, listeners can begin to acquire a non-native speech category. Supporting this, we obtained a significant effect for VOT stimulus, $F(14,112) = 7.389$, $MSE = .435$, $p = .001$, $\eta^2 = .480$. Given that we did not obtain a main effect or interaction of confidence reports, it suggests that

confidence reports did not significantly affect ID performance thereby permitting a straightforward interpretation of the remaining results.

Figure 1a. Identification Function without the Requirement of Confidence

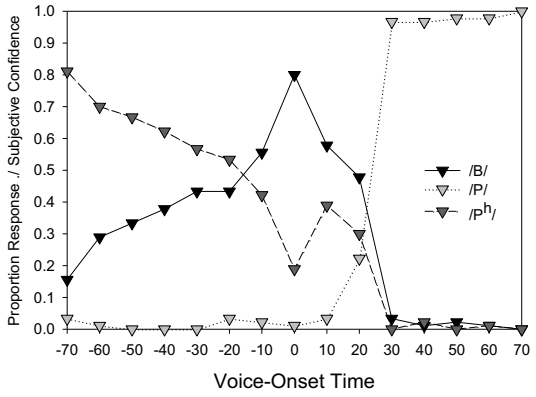
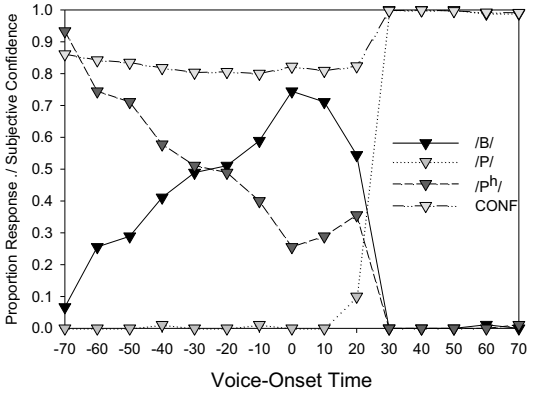


Figure 1b. Identification Function with the Requirement of Confidence



Identification Response Time. Prior to conducting an analysis of the response time data, we collapsed stimuli into regions five regions along the VOT continuum corresponding to two category boundaries (CBs) /p^h-b/ and /b-p/ corresponding to CB_1 (-30, -20) and CB_2 (20, 30), respectively, and equivalent within-category pairs corresponding to /p^h/ (-70, -60), /b/ (0,10), and /p/ (60, 70), respectively. Using the criterion of 3 standard deviations, 4.3% of the responses were identified as outliers and removed from the final analysis.

Table 1. Mean identification response time in ms along the critical regions of the VOT continuum with standard error reported in parentheses.

Confidence Condition	**/p^h/ (-70, -60)**	**/p^h-b/ (-30, -20)**	**/b/ (0, 10)**	**/b-p/ (20, 30)**	**/p/ (60, 70)**
No Confidence	888 (45)	910 (51)	893 (54)	933 (75)	796 (43)
Confidence	1,009 (47)	1,137 (124)	1,072 (113)	992 (83)	855 (41)

An analysis of the remaining responses times revealed a main effect of VOT region, $F(4,32) = 4.45$, $MSE = .041$, $p = .025$, $\eta^2 = .357$. Table 1 indicates that response latencies were longer at category boundaries as well as for the non-native (/p^h/) and modified native (/b/) categories relative to the native /p/ category. A main effect of the requirement of confidence report was also obtained, $F(1,18) = 14.55$, $MSE = .026$, $p = .005$, $\eta^2 = .645$. Again, Table 1 demonstrates longer latencies with the requirement of confidence relative to the no confidence condition. Given that the confidence block always followed the no confidence block, this finding cannot be attributed to automaticity. The interaction of confidence condition and VOT region was only marginally significant, $F(4,32) = 2.724$, $MSE = .019$, $p = .099$, $\eta^2 = .254$.

Confidence Reports. Figure 1a and 1b also demonstrate the effect of confidence measures. Listeners expressed less confidence in their responses to stimuli located within the /p^h/ and /b/ categories. As was the case with ID accuracy, we observed a main effect of the stimulus location along the VOT continuum on mean confidence, $F(14,112) = 6.931$, $MSE = 1011.371$, $p = .018$, $\eta^2 = .464$. Our comparison of

over/underconfidence bias did not reveal any significant effects, $F(14,112) = 2.146$, $MSE = .0354$, $p = .133$, $\eta^2 = .212$. Although it is possible that our small sample size might obscure a significant effect due to the inherent individual differences associated with confidence reports, our findings suggest that listeners might not be fully aware of the processes allowing them to identify stimuli.

Discussion

The results of the present study replicated findings obtained by Pisoni, et al. (1982) and Baranksi and Petrusic (1994, 1998). With minimal psychophysical training, we were able to induce listeners to perceive a non-native speech category in the voiceless, unaspirated portion of the VOT continuum (i.e., /p^h/) (Pisoni et al., 1982). Importantly, however, listeners' identification functions were not as sharp as those obtained by Pisoni et al. (1982): listeners' performance was lower for stimuli within the /b/ category. One possibility is that the stimuli used in the present study (HL, 2011) might not have provided appropriate acoustic cues to allow participants to acquire the non-native speech category. Alternatively, and more plausible, the reduced quantity of training provided in the present study might have resulted in phonemes that were not as well-defined to the listeners. This provides further support for claims that participants cannot only discriminate acoustic differences within a category boundary (e.g., Iverson & Kuhl, 1995; Miller & Volatis, 1989; Pisoni 1973; Pisoni et al., 1982; Pisoni & Tash, 1974) but that phoneme categories are somewhat plastic in adult listeners (cf. Eimas, 1975; Werker, 1989). Consequently, we can proceed to interpret confidence ratings directly.

When asked to report confidence, listeners took additional time to perform the primary decision. This additional requirement did not however affect performance. Listeners' accuracy in identifying stimuli did not significantly differ between confidence and no confidence conditions, even though the block in which listeners reported confidence always followed the no confidence condition leading to the possibility of training effects. An examination of mean confidence revealed that listeners expressed less confidence in the /p^h-b/ portion of the VOT continuum, suggesting that they had less certainty in their responses. Moreover, the overconfidence expressed by listeners across that portion of the continuum suggests that even if the acoustic properties of the auditory signal were available to listeners, they relied on phonemic representations when reporting the level of certainty in their response.

An instructive comparison can also be made between the results of the present study and those of a comparable 2AFC variant of the task: Schoenherr et al. (2012) observed a small decrease in confidence around the /b-p/ category boundary when only the voiced and voiceless portions of the continuum were presented to listeners. When the prevoiced portion of the continuum was additionally presented, Schoenherr et al. (2012) observed lower confidence in the /p^h-b/ portion of the continuum. When compared to identification accuracy, this pattern of responses leads to underconfidence in comparison to the overconfidence observed in the present study. Taken together with our results, this suggests that training does result in the allocation of attention to newly relevant acoustic properties of the stimuli in this region of the VOT continuum, thereby reducing certainty. Although listeners are somewhat more conservative in their confidence reports, they are still overconfident suggesting that the phonemic representations that they are subjectively aware of are less accurate than the acoustic information necessary to identify the stimuli.

References

Abramson, A., & Lisker, L. (1965). Voice onset time in stop consonants: Acoustic analysis and synthesis. Proceedings *of the Fifth International Congress on Acoustics*, Liege, A51.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgements. *Perception & Psychophysics*, *55*, 412-428.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 929-945.

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception & Psychophysics, 18,* 341-347.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.

Haskins Laboratories (2011). Abramson/Lisker VOT Stimuli. Retrieved 01/12/2011. From http://www.haskins.yale.edu/featured/demo-liskabram/index.html /.

Jusczyk, P. (1989). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications*. Monkton, MD: York.

Kvidera, S., & Koustaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual *Decisions. Journal of Behavioral Decision Making*, *21*, 253–281.

Liberman, A.M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358-368.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how, much they know? *Organizational Behavior and Human Performance*, 20, 159-183.

Lisker, L, & Abramson, A. S. (1967). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences.* Prague: Academia.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*, 505-512.

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

Peirce, J. W. (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8-13.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253-260.

Pisoni, D. B., Aslin, R. N., Percy, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 297-314.

Pisoni, D. B., & Tash, J. B. (1974) Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15,* 285-290.

Schoenherr, J. R., Logan, J., and Larose, C. (2012). Subjective confidence of acoustic and phonemic representations during speech perception. *Proceedings of the 28th Annual Meeting of the International Society for Psychophysics*, Ottawa, Ontario, Canada.

Sheldon, A. & Strange, W. (1982). The acquisition of /r/ and /1/ by Japanese listeners or English: Evidence that speech production can precede speech perception. *Applied Psycholingusitics*, 243-261.

Vickers, D., & Packer, J. S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica*, 50, 179-197.

Werker, J. F. & Logan, J. S. (1985). Cross-language evidence for three-factors in speech perception. *Perception & Psychophysics*, *37*. 35-44.