Subjective Confidence of Acoustic and Phonemic Representations During Speech Perception

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

Acoustic and phonemic information form the basis for speech perception. We used confidence reports to examine the extent to which 1) both representations contributed to identification performance, 2) whether participants had an awareness of acoustic information, and 3) whether confidence reports were influenced by both acoustic and phonemic representations in an identification task. Our results suggest that participants' response were primarily guided by explicit, phonemic information. We also found that an interpolation between typicality ratings and identification functions yielded an excellent fit to the function produced by confidence reports suggesting that confidence processing.

Responses to Two Alternative Forced Choice (2AFC) identification tasks in which listeners are presented stimuli that vary along an acoustic continuum, such as voice-onset time, produce a logistic function defining two groups of stimuli (i.e., /b/ and /p/). These response patterns have been interpreted as a result of listeners utilizing phonemic representations (e.g., Liberman, Harris, Hoffman, & Griffith, 1957), with evidence from discrimination tasks additionally suggesting that participants cannot detect within-category acoustic differences when stimuli are presented at long ISIs (e.g., Pisoni, 1973). The use of multidimensional scaling techniques has also revealed considerable within-category similarity and between-category dissimilarity (Iverson & Kuhl, 1995). Alone, these findings suggest that phonemes might be the only available representations to identify and discriminate stimuli.

Phonemic representations, however, do not appear to be the only source of information available to participants. Using analysis of response latencies, Pisoni (1973; Pisoni & Tash, 1974) proposed a model wherein acoustic properties determine primary decision response selection under certain conditions. Specifically, participants exhibited short response latencies for pairs of stimuli that were identical acoustically (i.e., ΔVOT = 0 ms) as well as those that were highly dissimilar (ΔVOT $\geq$ 40 ms). Longer response latencies were observed for within-category pairs and between-category pairs which had identical acoustic differences (ΔVOT = 20 ms). These differences in response latencies suggested that participants also had access to acoustic information. Further evidence for the availability of acoustic representations also comes from changes in discrimination performance and identification functions. For instance, Pisoni (1973) found that presenting stimuli at short ISIs resulted in increases in accuracy when participants made within-category comparisons (cf. Werker & Logan, 1985). Psychophysical training studies have also found that participants can learn to identify non-native speech sounds (Pisoni et al., 1982). Pisoni et al. (1982) provided participants with three exemplars and feedback after each response. In the 3-category identification, participants' identification functions produced 3 distinct categories. Still other studies using typicality ratings have obtained results suggesting that participants can detect within-category acoustic difference (Miller & Volaitis, 1989).

The existence of acoustic and phonemic representations suggests that multiple representations can be used as evidence for primary decision response selection. Whereas the role of multiple representations has been examined previously in the context of top-down processing in sine-wave speech (e.g., Remez, Rubin, Pisoni, & Carrell, 1981), it is not entirely clear whether participants

maintain an awareness of these two sources of information. Methods used that examine calibration of subjective awareness are used here to examine this question.

Quantitative approaches to the assessment of subjective awareness require participants to provide a subjective probability (e.g., 50% represents a guess and 100% represents complete certainty) after performing a task. Underconfidence has generally been obtained in perceptual tasks (e.g., Bjorkman, Juslin, & Winman, 1993) leading some to assert our perceptual system is relatively inaccessible (Dawes, 1980). In comparison, confidence reports obtained for general knowledge questions typically produce overconfidence (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991). Challenging these findings, the hard-easy effect (Lichtenstein & Fischhoff, 1977) assumes that participants' subjective bias is determined by task difficulty more generally. For instance, in a line-length discrimination task conducted by Baranski and Petrusic (1994), participants produced overconfident responses for difficult stimulus pairs and underconfident responses for easy stimulus pairs. Similarly, recent studies observed an overconfidence bias in perceptual tasks for stimuli with both perceptual and conceptual properties (e.g., Kvidera, & Koustaal, 2008). In the context of speech perception, we sought to identify overconfident responses due to the presence of acoustic and phonemic representations.

Models of confidence process differ along three dimensions: the locus of confidence processing, the dependency of confidence processing on primary decision processes, and the sources of evidence used to compute confidence. Many early models of confidence assumed a decisional-locus (e.g., Ferrel & McGooey, 1980; Gigerenzer et al., 1991; see also Pleskac & Buseymeyer, 2010) wherein confidence reports are based solely on information used by the primary decision process thereby requiring no additional processing, a post-decisional locus wherein confidence is computed following the primary decision (e.g., Audley, 1960; Vickers & Packer, 1980). A later development was an alterable locus model wherein confidence processing can occur during or after the primary decision depending on speed or accuracy stress and used the total accumulated amount of nondiagnostic evidence to determine certainty (Baranski & Petrusic, 1998). In a study conducted by Baranski and Petrusic (2001) participants were given blocks of trials wherein they were required to simply make a decision or make a decision followed by a post-decisional confidence report. They found that response latencies for the primary decision were significantly longer when confidence was required relative to a no confidence condition indicating an additional set of operations was required to compute confidence. Recent studies have also found that by manipulating the nature of nondiagnostic information available during the primary decision, confidence reports can vary independently of accuracy (Schoenherr, Leth-Steensen, & Petrusic 2010). Applied to phonemic categorization, if acoustic information is available from a perceptual process and phonemic representations are available from the activation of long-term memory representations, then both sources of information should influence confidence reports. Substantial differences in the patterns observed between accuracy and confidence would suggest the existence of acoustic and phonemic representations.

Method

Fifteen and Fifteen listeners from Carleton University students participated in the study for course credit in Experiments 1 and 2, respectively. All participants reported normal hearing and no speech pathologies. Using the paradigm developed by Pisoni and Tash (1974) participants were presented with /b/ and /p/ stimuli that varied along the VOT continuum. Fifteen speech stimuli corresponding to -70 to 70 ms VOT, originally synthesized by Lisker and Abramson (1967), were obtained from the Haskins Laboratories website (HL, 2011). The sounds were originally recorded on reel-to-reel tape and later

converted into AIFF format. Stimuli were pre-processed using a DC offset correction to eliminate clicks present in the AIFF versions and then converted into WAV files. Whereas Experiment 1 only used stimuli from the 0 to 60 ms VOT range to replicate Pisoni and Tash (1974), Experiment 2 used the full stimulus range, with stimuli from the -70 ms end corresponding to the prevoiced phoneme category /p^h/ not used phonemically in English, and the remaining stimuli corresponding to the /p/ and /b/ phoneme categories used in English.

Procedure

Trials in the ID task had one or two components depending upon block. In both blocks of trials participants reported whether the stimulus was a /b/ or /p/ using keys labeled B or P on the keyboard ('V' or 'N' key, respectively). For one block participants also rated the confidence they had in their ID responses using a 6-point scale using the 'E' through 'I' keys, with 50% representing a guess and 100% representing certainty. Participants completed a total of 180 trials in each block of the ID task.

Half of the participants performed the ID task first whereas the other half performed the AX task first. Half of the blocks of trials required participants to provide confidence reports whereas the other half only required participants to complete the ID task alone. Presentation of confidence and no confidence blocks was counterbalanced. The experiment required approximately 30 minutes to complete. Stimuli were presented via headphones using PsychoPy software (Peirce, 2007). The procedure was modified in Experiment 2, to include a typicality rating task following the ID task.

Results and Discussion

The results for Experiment 1 and Experiment 2 are provided in Figure 1a and Figure 1b. Identification responses, typicality ratings, and confidence indices were all subjected to repeated measures ANOVAs.

Experiment 1

Proportion Identification. Participants clearly identified two discrete categories for /ba/ and /pa/, respectively, with a category boundary situated between +20 and +30 ms VOT. This pattern replicates the findings obtained by Pisoni and Tash (1974) as well as other studies (e.g., Experiment 1 in McMurray et al., 2003). The proportion of correct ID responses was analyzed for each VOT stimulus and whether a confidence report was provided or not. The only significant finding observed was the location of the stimuli along the VOT continuum, $F(6,84) = 6.394$, $MSE = .019$, $p = .02$, $\eta^2 = .314$. The absence of a main effect or interaction of confidence reports is important as it suggests that the addition of confidence reports did not significantly affect ID performance thereby permitting a straightforward interpretation of the remaining results.

Confidence Reports. Figure 1 also demonstrates the effect of confidence measures. Like ID accuracy, we found that subjective confidence varied along the VOT continuum, $F(1,14) = 6.55$, $MSE = 44.11$, $p = .008$, $\eta^2 = .319$. Pairwise comparisons revealed that this effect arose from the difference in confidence between stimuli located at 20 and 30 ms VOT ($p = .035$), which corresponds to the stimuli adjacent to the category boundary. An analysis of subjective calibration revealed only a marginally significant difference across the VOT continuum, $F(6,84) = 3.401$, $MSE = .013$, $p = .085$, $\eta^2 = .195$. This suggests that the greatest difference between subjective awareness and performance occurs for the 20 ms VOT stimulus. Our comparison of over/underconfidence bias did not reveal any significant effects, $F(6,84) = 1.948$, $MSE = .035$, $p = .183$, $\eta^2 = .122$. Together, these findings suggest that participants are only explicitly aware of the phonemic representation.

Figure 1a.

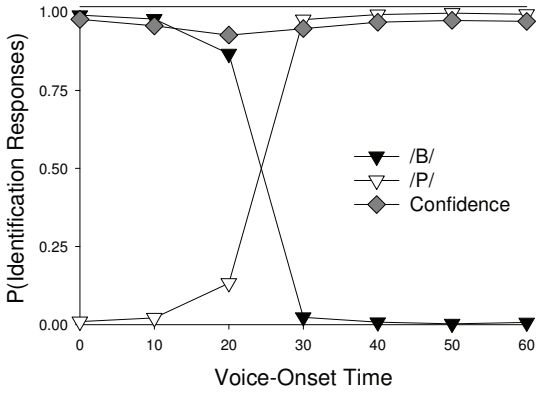

Figure 1b.

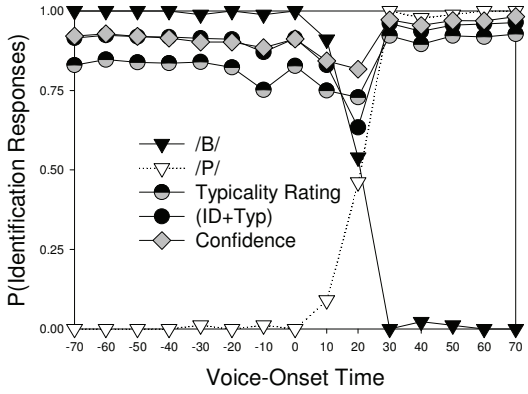

Figure 1a. Mean identification functions, response times for confidence (unfilled circles) and no confidence (filled circles) conditions and mean confidence across VOT continuum. The identification function uses performance in confidence condition to allow comparison with mean confidence. Figure 1b. Mean identification functions, typicality ratings, confidence reports, and interpolation line for Experiment 2.

Experiment 2

Proportion Identification. Replicating the general results of Experiment 1, the location of the stimuli along the VOT continuum significantly affected identification performance, $F(14,112) = 9.149$, $MSE = .124$, $p = .005$, $\eta^2 = .533$. Figure 1b demonstrates, participants had a sharp category boundary between stimuli for the /b/ and /p/ categories. A noticeable difference was evident in the location of the boundary. Whereas in Experiment 1 the boundary was located between VOT 20 ms and 30 ms, a shift such that the boundary was now located at VOT 20 ms with a resulting decrement in performance for VOT 10 ms stimuli. We can take these results as indicative of range effects. In general, these findings permit a straightforward interpretation of the remaining results.

Confidence Reports. Figure 1 also demonstrates the effect of confidence measures. Like ID accuracy, we found that subjective confidence varied along the VOT continuum, $F(1,14) = 6.55$, $MSE = 44.11$, $p = .008$, $\eta^2 = .319$. Relative to Experiment 1, we did observe greater underconfidence in the negative portion of the VOT continuum.

Typicality Task. The analysis of typicality ratings also obtained a significant result of stimulus location along the VOT continuum, $F(14,112) = 5.820$, $MSE = .3.295$, $p = .009$, $\eta^2 = .421$. Unlike accuracy, but like mean confidence, typicality ratings appeared to be more responsive to the acoustic properties of the stimuli. Participants considered stimuli in the /b/ and /p^h/ range as less typical then stimuli in the /p/ range even though they exhibited equal accuracy. Moreover, within-category ratings exhibited more graded responses.

Interpolated Function. The similarities in patterns observed in confidence and typicality suggested a potential relationship between these two functions. As Figure 1b suggests, mean confidence ratings are situated between accuracy in the identification task and typicality ratings in the typicality task. Pearson's correlations revealed the strongest relationship between confidence and typicality ratings, $r^2 = .960$, $p < .001$. The correlations between identification responses and mean confidence was also significant, $r^2 = .446$, $p = .007$, although the correlation between identification and typicality was only marginally significant, $r^2 = .261$, $p = .051$. These findings suggest that confidence is associated with both identification accuracy and typicality ratings but that identification accuracy and typicality ratings are only weakly related.

In order to examine the relationship between accuracy, typicality, and confidence ratings we converted typicality to a proportion, summed it with proportion correct, and produced an interpolated function. A paired-samples t-test revealed that the mean confidence function and the interpolated function did not significantly differ from one another, $t(14) = .309$, $p = .762$. This suggests that confidence reports were closely associated with information from both identification accuracy (associated with phonemic representations) and typicality ratings (associated with acoustic information). All other paired-sample t-tests were significant (all ts > 3.283, ps < .005) indicating that different sources of information contributed to response selection for each dependent measure.

General Discussion

In general, we obtained results consistent with previous studies of speech perception and confidence processing. A 2AFC identification task using stimuli from the prevoiced-unvoiced VOT continuum produced responses suggesting two phoneme categories were involved in response selection. When that range was restricted to voiced and unvoiced stimuli (Experiment 1), participants only exhibited overconfidence around the category boundary. This finding supports the claim that overconfidence can be obtained in perceptual tasks (e.g., Baranski & Petrusic, 1994). Moreover, overconfidence in this context also suggests that a phonemic representation is used to identify stimuli. Extending the range to prevoiced stimuli (Experiment 2) resulted in underconfidence, suggesting that acoustic properties of these stimuli were in fact available to participants even if they were not used in primary decision response selection. Thus, the results of the present study support the availability of two kinds of stimulus representations - acoustic and phonemic – that influence response selection (cf. Pisoni, 1973; Pisoni & Tash, 1974).

An important result obtained in the present study was the relationship between typicality ratings and confidence reports. Typicality ratings were provided by participants because previous studies obtained results suggesting that these ratings were affected by acoustic information (e.g., Miller & Volatis, 1989). We initially assumed that confidence judgments might be influenced by acoustic information, producing more continuous responding. In Experiment 1, we observed a trend suggesting that confidence was only affected by phonemic information. In contrast, the results of Experiment 2 acoustic properties reduced subjective confidence and typicality ratings in the prevoiced portion of the continuum. When an interpolated function was obtained for identification accuracy in an identification task and typicality ratings in a typicality task, it yielded an excellent fit to the confidence function in the identification task. In the absence of another explanation, it seems reasonable to suggest that confidence reports are the product of both acoustic and phonemic representations. These results could be taken as support for a doubt-scaling model of confidence processing (Baranski & Petrusic, 1998). In the present task, participants were presented with /p^h/ stimuli which that are outside the range of their native phoneme categories (although Lisker & Abramson, 1964, report that stops produced by some English listeners incorporate prevoicing, it is not used phonemically in English). Namely, although the identification task only requires the use of phonemic representations that are activated in long-term memory as a result of accumulated acoustic evidence, that acoustic information in the prevoiced region increases the amount of uncertainty as to category membership. Thus, while primary decision accuracy has reached a performance asymptote, subjective confidence is reduced due to the availability of this evidence.

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.

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 /.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America, 97,* 553–562.

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.

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 inthe 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.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947-949.

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.