Talker identification as a categorization problem

Casey L. Roark[1,2] **(croark@pitt.edu)**

Gangyi Feng[3,4] **(g.feng@cuhk.edu.hk)**

Bharath Chandrasekaran[1,2] **(b.chandra@pitt.edu)**

[1]Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA 15213 USA
[2]Center for the Neural Basis of Cognition, Pittsburgh, PA, 15213 USA
[3]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[4]Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

Abstract

Learning to identify a person's voice is a key component of speech perception. In this study, we use a categorization framework to provide insights about the mechanisms supporting talker identification. Native Mandarin Chinese listeners learned to categorize sentences in three tasks with different language contexts – native Mandarin talkers speaking Mandarin, native English talkers speaking English, and native Mandarin talkers speaking English. We compared learning when listeners received fully informative or minimal feedback. Using decision bound models, we examined the strategies participants used in each of the three tasks. Regardless of language context, full feedback was initially better for learning than minimal feedback but was no different after the second block. Across tasks, participants often used strategies based on mean fundamental frequency to separate the talkers. These results demonstrate that talker identification is a categorization problem, which enables leveraging existing category learning frameworks to understand the mechanisms of this important ability.

Keywords: categorization; category learning; talker identification

Introduction

Learning who is talking is an important ability that guides spoken communication. Especially in the absence of concurrent visual cues, it can be difficult to identify a speaker based on their voice. This complex ability requires that we identify variable spoken utterances as coming from a single talker. This many-to-one process can be conceptualized as a categorization problem. In this study, we leverage approaches from the category learning literature to provide a better understanding of how people learn to identify talkers.

Prior research has identified language experience as an important factor that influences talker identification. Listeners find it easier to recognize talkers in their native language than a foreign language (Goggin et al., 1991; McLaughlin et al., 2019; Perrachione, 2018; Perrachione, Del Tufo, & Gabrieli, 2011; Perrachione & Wong, 2007), a phenomenon labeled the Language Familiarity Effect (Thompson, 1987). When listening to talkers in one's native language, more cues are available to the listener to help tell different talkers apart. Listeners have both acoustic (i.e., familiar sound patterns) and linguistic cues to guide their decision making (Levi, 2019; Xie & Myers, 2015; Zarate et al., 2015).

Acoustic cues alone can be particularly useful in differentiating talkers. Easily identifiable cues like mean fundamental frequency (F0, e.g., pitch height) of a talker's voice can be a useful marker of talker identity (LaRiviere, 1975; Lavner, Rosenhouse, & Gath, 2001; Perrachione, Furbeck, & Thurston, 2019; Sambur, 1975; van Dommelen, 1990). Other cues like F0 variability and speech rate can also contribute to talker identity (Perrachione et al., 2019; Skoog Waller, Eriksson, & Sörqvist, 2015; Winkler, 2007).

However, cues that signal talker identity may also depend on the talker's native language. Listeners perform just as poorly identifying talkers with accented speech in their native language as identifying talkers speaking a foreign language (McLaughlin et al., 2019; Stevenage, Clarke, & McNeill, 2012; Yu, Schertz, & Johnson, 2012). An effect termed the Other Accent Effect suggests that talkers with the same accent as the listener are easier to recognize than other-accented talkers (Stevenage et al., 2012). This effect may depend on the nature of the other accent. In a recent study, native listeners of Canadian-accented English performed equally well on Canadian-accented and Australian-accented English but showed similarly poor performance on Mandarin-accented English as on foreign Mandarin speech (Yu et al., 2021). These results suggest that identifying talkers with some foreign-accented speech can be just as difficult as identifying talkers in a foreign language.

Less is understood about how listeners identify talkers who speak in a foreign language with the same accent as the listener (e.g., a native Mandarin listener identifying talkers in Mandarin-accented English). However, there is evidence to suggest that same-accented listeners of a foreign language may have some benefits in speech perception (Bent & Bradlow, 2003). For example, studies have shown that native Mandarin listeners have comprehension advantages in Chinese-accented English relative to native English listeners (Bent & Bradlow, 2003; Yuan, Jiang, & Song, 2010). This suggests that some cues that signal talker identity might be similar in one's native language and same-accented foreign speech. For example, it is possible that cues that are available

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

to signal talker identity in native Mandarin speech may also be available in Mandarin-accented English speech.

Not much is understood the strategies individuals use to learn talker identities across different language contexts. In this study, we leverage a category learning perspective to understand how people learn who is talking in three language contexts: one's own native language spoken by native talkers, a foreign language spoken by talkers with the same accent as the listener, and a foreign language spoken by native talkers. We take two approaches from the category learning literature to better understand how people learn talker identities and the information they use to make their decisions over the course of learning: feedback manipulations and decision strategies.

Feedback Manipulations

The type of feedback that individuals receive affects category learning. With full feedback, participants are told whether their response was correct or incorrect as well as the correct category (e.g., "Correct, that was category 1"). With minimal feedback, participants are only told whether their response was correct or incorrect (e.g., "Correct").

Full feedback is superior to minimal feedback for learning categories that can be differentiated by verbalizable rules (Maddox, et al., 2008; Yi & Chandrasekaran, 2016). Full feedback may be beneficial in these cases because it helps test explicit rules about category identity. In contrast, minimal feedback is helpful for learning categories that require integration across dimensions and are difficult for learners to describe with simple verbalizable rules (Maddox et al., 2008). When learning non-rule described categories, including foreign language speech categories, minimal feedback may be just as effective as full feedback (Chandrasekaran, Yi, & Maddox, 2014; Yi & Chandrasekaran, 2016). Other researchers have shown that the presence of a full or minimal feedback benefit in the visual modality may depend on the nature of the type of stimulus mask that is used during learning (Dunn, Newell, & Kalish, 2012).

In the current study, we predict that if a particular talker identification problem can be solved with verbalizable rule-based strategies, then learning should be better for full feedback than minimal feedback. However, if the problem is difficult to solve with simple verbalizable rules, then we may see no difference between the two types of feedback.

Critically, it is possible that the effect of feedback would depend on the specific problem being learned (i.e., one's native language, same-accented foreign language, and native-accented foreign language). Since people are more proficient in identifying talkers in their native language, it is possible that identifying talkers in one's native language can be solved with verbalizable rules, with clearer possible rules about what identifies a specific talker based on their long-term language experience. If this is the case, then full feedback may be superior to minimal feedback for native-language talker categorization.

In contrast, since people are less proficient at identifying talkers in a foreign language, it may be more difficult to find reliable acoustic cues to signal talker identity. As a result, rule-based strategies may be less reliable or effective when learning who is talking in a foreign language. If this is the case, then full feedback may be no different than minimal feedback for foreign-language talker categorization.

Finally, listening to talkers in a native-accented foreign language may align with either of these perspectives or fall somewhere in between. If more native-like cues and rules are available when listening to same-accented foreign speech, then full feedback may be superior to minimal feedback. Instead, if listening to same-accented foreign speech is more like listening to native-accented foreign speech, then minimal feedback may be no different from full feedback.

Decision strategies

Regardless of the type of feedback that participants receive during learning, not much is understood about the cues that learners use to decide who is talking and, critically, how their use of those cues might change over the course of learning. We leverage a commonly used tool for understanding decision strategies in perceptual categorization contexts – decision bound models (Ashby, 1992; Maddox & Ashby, 1993). These models enable understanding of how participants separate categories in multidimensional space. Specifically, these models enable us to go beyond accuracy to understand more about *how* participants learn. This is important because two individuals (or the same individual in two different problems) may have similar accuracies but use different strategies.

At present, decision bound models are restricted two-dimension problems, so we focus on two dimensions that provide reliable information about talker identity (Lavner et al., 2001; Perrachione et al., 2019; Skoog Waller et al., 2015; Winkler, 2007) – mean F0 (e.g., average pitch height across an entire sentence) and speech rate (e.g., number of syllables normalized by duration). While naturalistic speech stimuli are highly complex and multidimensional, we have chosen a set of dimensions that are likely to be informative for talker identity in these three language contexts. Using these models will allow us to assess the types of strategies that participants use to solve these three talker identification problems.

Methods

We examine how native Mandarin Chinese listeners learn to distinguish different talkers in three different language contexts – native Mandarin Chinese talkers speaking Mandarin Chinese, native Mandarin Chinese talkers speaking English, and native English talkers speaking English. We compare learning across the three tasks when participants were given full feedback (e.g., "Correct, that was 1") or minimal feedback (e.g., "Correct").

Participants

Participants were 79 students (Full: *N* = 39; Minimal: *N* = 40) recruited from the South China Normal University community, ages 18-26 (Full: *M* = 20.7, *SD* = 2.18, 17M/22F; Minimal: *M* = 20.3, *SD* = 2.32, 17M/23F). Participants were native listeners of Mandarin Chinese. Participants received monetary compensation for their participation. Experimental procedures were approved by the South China Normal University Institutional Review Board and the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee. An additional participant in the Full Feedback condition did not complete all tasks and was excluded from analyses. Participants completed the task in the Gorilla Experiment Builder (gorilla.sc; Anwyl-Irvine et al., 2019).

Participants reported their familiarity and experience with other languages. Among all participants, around half reported knowing English (Full: 20/39; Minimal: 24/40), with an average self-reported proficiency of 4/10 (Full: *M* = 4.10, *SD* = 1.74; Minimal: *M* = 3.88, *SD* = 1.48).

Stimuli

Stimuli were short (1-2.6 sec) Hearing in Noise Test sentences (Soli & Wong, 2008) from the SpeechBox corpus (Bradlow, n.d.-b) and ALLSSTAR corpus (Bradlow, n.d.-a). Sentences were spoken by 12 male talkers and the talkers were unique in each task (4 talkers/task). In the Native Mandarin and Native English tasks, talkers were native speakers of either Mandarin Chinese or American English and spoke sentences in their native languages. In the Mandarin-Accented English task, talkers were native speakers of Mandarin Chinese and spoke sentences in English. As in natural speech, there were some variations in acoustic features like mean F0 and speech rate across talker (Figure 1). In each task, there were 10 training sentences and 10 test sentences. Each sentence was spoken by each of the four talkers in a task for a total of 40 training sentences and 40 test sentences.

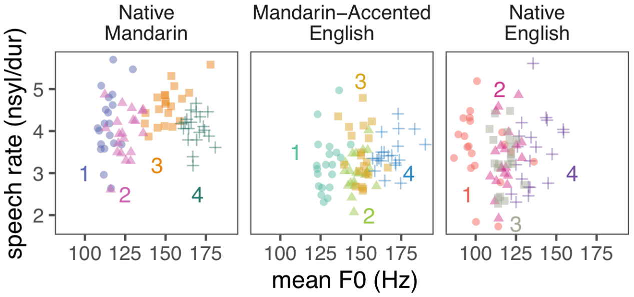

Figure 1: Distributions of all spoken sentences based on mean fundamental frequency (F0) and speech rate (number of syllables divided by total duration in seconds). Each talker is shown in a different color and shape.

Intelligibility measures for Mandarin talkers were available from a previous study that used these stimuli (Bradlow, Blasingame, & Lee, 2018). The Native Mandarin sentences were highly recognizable by native Mandarin listeners (*M* = 88% correct words identified in -4 dB signal-to-noise ratio [SNR] in white noise). The Mandarin-Accented English sentences were recognizable by native English listeners (*M* = 70% correct words identified in 0 dB SNR white noise). We separately tested Native English sentence intelligibility in two naïve native English listeners and the sentences were highly recognizable (*M* = 94%, correct words identified in -4 dB SNR in white noise).

Procedure

Participants first completed a headphone screening to ensure they were using headphones and could hear the sounds (Milne et al., 2020). All participants completed each of the three tasks, with order counterbalanced across participants. The only difference between conditions was the nature of the feedback. In the full feedback condition, participants received fully informative feedback (e.g., “Correct, that was 1”; “Incorrect, that was 2”) and in the minimal feedback condition, participants received minimally informative feedback (e.g., “Correct”; “Incorrect”). Feedback was presented immediately for 750 ms and all were told to use the feedback to improve their performance.

The tasks were identical for each set of stimuli (Native English, Native Mandarin, Mandarin-Accented English). The only difference was in the nature of the sentences. Participants were told at the beginning of each task the language of the sentences (e.g., “In this task, you will hear sentences in English”). In each task, participants completed five training blocks, where they heard each of the 40 sentences (10/talker) once. In the generalization test block, participants heard 40 novel sentences spoken by the same four talkers (10/talker) and categorized the sentences without any feedback. Understanding talker identification as a categorization problem relies on the ability of participants to *generalize their knowledge* to these novel situations. All trials were followed by a 1 sec inter-trial interval.

Finally, at the end of each task, participants were given open prompt questions about each talker and were asked to describe how they decided it was this talker who was speaking. Due to the complex nature of this response dataset, analyses of these responses are ongoing.

Table 1: Maximum possible accuracy of different strategies.

Task	Mean F0	Speech Rate	2D Rule	2D Integration
Native Mandarin	85%	25%	75%	91%
Accented English	78%	25%	71%	78%
Native English	59%	24%	56%	45%

To understand the decision strategies that participants used to categorize the sentences by talker, we applied decision bound models (Ashby, 1992; Maddox & Ashby, 1993). These models allow for assessment of how participants use the acoustic dimensions to identify who is talking. We fit a series of models that make different assumptions about the information participants use in their responses (i.e., mean F0, speech rate, both dimensions, neither dimension) and compared how well each of these fit the participant’s response data. The participant’s strategy was selected as the model that best captured the participant's response pattern. Due to variability across talkers, there are some differences in the maximum accuracy that one could achieve if they used mean F0, speech rate, or both dimensions (2 Dimension/2D) to separate the stimuli by talker (Table 1).

Rule-based models Rule-based models assume that participants use verbalizable rules to separate the sentences into talker categories. We fit models that assumed that participants used single-dimension (1D) rules on mean F0 or speech rate or two-dimension (2D) rule models that assume that they use rules along both dimensions. The 1D models have four free parameters: three for boundaries along the key dimension (either mean F0 or speech rate) and one noise parameter. We fit two versions of the 2D rule models – one that assumes that participants use one rule along the mean F0 dimension and another along the speech rate dimension (e.g., categories fall into four quadrants such as top-left [low mean F0 and high speech rate]) and another that assumes that participants use two rules along one of the dimensions and another along the remaining dimension (e.g., highest mean F0 is category 1, lowest mean F0 is category 2, mid-level mean F0 and high speech rate is category 3, mid-level mean F0 and low speech rate is category 4). The two-rule models have three free parameters: two for the boundaries along the mean F0 and speech rate dimensions and one noise parameter. The three-rule models have four free parameters: three for the placement of the boundaries along the two dimensions and one noise parameter.

Integration model The integration model assumes that participants used both mean F0 and speech rate to decide which speaker was talking. Different from the 2D rule models, the integration model reflects a strategy that is difficult for participants to verbalize. The model is implemented as the Striatal Pattern Classifier (SPC; Ashby et al., 1998) and assumes that participants use feedback to learn stimulus-response associations based on the neurobiology of the striatum (Ashby & Waldron, 1999). The SPC model can be thought of as a complex version of an exemplar model (Ashby & Rosedahl, 2017). The SPC model has nine free parameters: eight that determine the location of hypothetical striatal units in perceptual space and one that represents the noise associated with the placement of the units.

Random guessing model The random guessing model assumed that participants randomly guessed the category identity on each trial.

Model fitting and selection We fit each of the models to each block of each participant's learning data and the generalization test. The model parameters were estimated using maximum likelihood procedures (Wickens, 1982) and model selection used the Bayesian Information Criterion (BIC, Schwarz, 1978), which penalizes models with more free parameters. The model with the lowest BIC value was selected as the best fitting model for that block of that participant's data. This procedure was repeated for each block and each participant. The models accurately captured participants' responses with a prediction accuracy of 58%, substantially better than chance (25% +/- 10% across 40 trials, .25 prob. of success, 95% cumulative prob.). Accuracy is expected to be less than 100% due to noise in participant responding stemming from factors such as attention lapses or inconsistent strategy application.

Results

Full versus Minimal Feedback

We examined performance across the three tasks (Native Mandarin, Mandarin-Accented English, Native English) when participants were given full or minimal feedback. We examined performance separately during training blocks and in the generalization test.

Training With full or minimal feedback, participants successfully learned who was talking in all three language contexts (Figure 2). We ran a mixed model ANOVA with block (1-5) and task as within-subjects factors and feedback type as a between-subjects factor. The nature of the feedback did not differently affect performance in the three tasks – the interactions between feedback type, block, and task (F(5.82, 447.94) = 0.18, p = .97, η_g^2= .00051) and feedback type and task (F(1.74, 134.06) = 0.46, p = .60, η_g^2= .0010) were not significant. Linear mixed effects models with participant as a random effect were run and give identical results, so we report ANOVAs here for parsimony.

However, we found that the pattern of performance across blocks was different for the full and minimal feedback conditions (F(2.87, 220.7) = 8.83, p < .001, η_g^2= .013). Bonferroni-corrected post-hoc test indicated that full feedback had higher performance than minimal feedback in the first block (p = .0027) with no significant differences in any other block (ps = 1.0). For all three tasks, full feedback gives participants the ability to learn the categories quickly, but the advantages over minimal feedback do not last long.

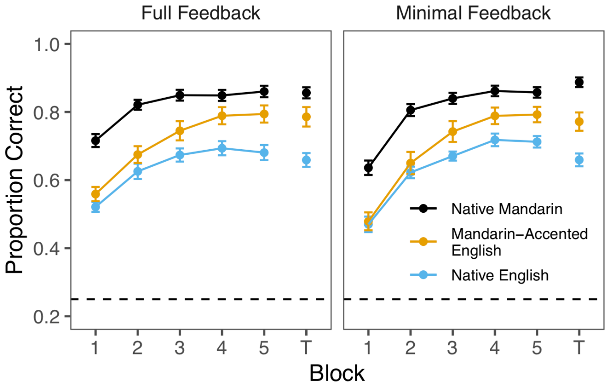

Figure 2: Performance across blocks (1-5) and the generalization test (T) for all tasks. Error bars reflect *SEM.* Dashed line reflects chance-level performance.

The performance in the three tasks was also different across blocks (F(5.82, 447.9) = 5.79, p < .001, η_g^2= .016). According to Bonferroni-corrected post-hoc comparisons, the Native Mandarin task always had significantly higher accuracy than both the Native English (ps < .001) and the Mandarin-Accented English tasks (ps < .004). There were no significant differences between the Mandarin-Accented English task and the Native English tasks in blocks 1 (p = .87) or 2 (p = .26), but Mandarin-Accented English performance was higher than Native English in all other blocks (ps < .003).

These results are consistent with the Language Familiarity Effect (Native Mandarin > Native English) and also suggest that differentiating talkers in a foreign language in the listener's own accent might be easier than differentiating talkers who are native speakers of the foreign language (Mandarin-Accented English > Native English), especially in later stages of learning.

Generalization Test A key component of categorization is the ability to generalize to novel exemplars. Importantly, participants no longer received any feedback in the generalization test block. Participants were able to *seamlessly* transfer their learning from feedback training to novel sentences without feedback (Figure 2-T).

Using a mixed-model ANOVA, we compared the difference between block 5 and test performance across the three tasks in the two feedback conditions. Overall, we found that participants maintained their final-block performance levels in the generalization test. There was no significant differences between full and minimal feedback (F(1, 77) = 0.061, p = .80, η_g^2= .00027) and no significant interaction between feedback type and task (F(1.85, 142.7) = 2.23, p = .12, η_g^2= .019). However, we found that the difference between the final training block and the test block was significantly different across tasks F(1.85, 142.7) = 5.09, p = .009, η_g^2= .042).

In the Native Mandarin task, participants had 1.33% (95% CI [-0.60, 3.25]) higher accuracy in the generalization test than block 5. In contrast, performance decreased by 3.77% (95% CI [-6.16, -1.37]) in the Native English task and 1.46% (95% CI [-3.86, 0.95]) in the Mandarin-Accented English task. According to Bonferroni-corrected post hoc tests, the Native Mandarin improvement was significantly different from the Native English decrement (p = .0049). There were no significant differences between Native Mandarin and Mandarin-accented English (p = .25) or between Native English and Mandarin-accented English (p = .45). Together, these results suggest that participants generally maintained their performance levels from training to test and this maintenance was better for talkers in one's native language than native speakers of a foreign language.

Decision Strategies

We examined the proportion of participants using different strategies across tasks and blocks (Figure 3). As a reminder, participants could use single-dimension strategies (mean F0, speech rate), two-dimension strategies (2D rule, integration), or a guessing strategy. To understand whether there were differences in strategies across the tasks and feedback conditions, we compared the proportion of participants using different strategies across blocks using Fisher's exact tests.

The most common strategy across blocks was using only the mean F0 dimension, though many participants also used 2D rule strategies. Across all training blocks, there were no significant differences in the strategies participants used in the three tasks with either minimal (ps > .054) or full feedback (ps > .29), with one exception. In block 1 of tasks with full feedback, there was a significant difference in strategies (p = .038), though none of the individual comparisons survived FDR correction in post-hoc tests (ps > .075). There were also no significant differences in strategies used in the full and minimal feedback conditions in any block – including the generalization test – in the Native Mandarin (ps > .068), Mandarin-Accented English (ps > .11), or Native English tasks (ps > .31).

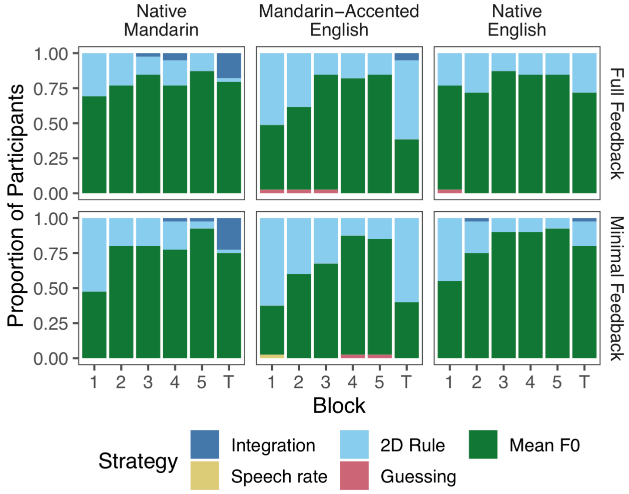

Figure 3: Decision strategies across blocks (1-5) and the generalization test (T) for all tasks.

In contrast, in the generalization test, there were significant differences in the strategies participants used in the three tasks for both minimal (p < .001) and full feedback (p < .001). Specifically, according to FDR-corrected post-hoc comparisons, there were differences in the strategies participants used in all three tasks for both minimal (ps < .003) and full feedback conditions (ps < .0078).

Across feedback conditions, more participants used a 2D rule strategy in the Mandarin-Accented English task (58%) relative to the Native Mandarin (3%) and Native English tasks (23%). More participants used mean F0 strategies in the Native Mandarin task (77%) and Native English tasks (76%) compared to the Mandarin-Accented English task (39%). No participants used a speech rate or guessing strategy in the test block of any task.

Finally, while most participants used mean F0 strategies or 2D rule strategies, some participants used integration strategies in the Native Mandarin task (20%). In contrast, relatively few participants used integration strategies in the Native English (1%) and Mandarin-Accented English tasks

(3%). According to a dual systems perspective of learning (Ashby et al., 1998), an integration strategy indicates these participants used non-verbalizable, procedural strategies in the generalization test. This could indicate that listening to talkers in one's native language encourages using procedural strategies more than foreign speech with any accent.

In sum, native Mandarin listeners primarily rely on the mean F0 cue to differentiate the talkers in their native language and native speakers of a foreign language but use rules on both mean F0 and speech rate when listening to same-accented speakers of a foreign language.

Discussion

Overall, our results indicate that full feedback has a brief benefit over minimal feedback for initial learning of talker categories across language contexts but does not substantially impact overall learning outcomes. As a novel approach in this domain, we used computational models to assess learners' strategies to understand how they use different information to make decisions about who is talking. Using approaches from the category learning field provides critical insights about how listeners use acoustic cues to decide who is talking.

Initial learning was superior for full feedback than minimal feedback. Critically, the observed effects of feedback did not differ based on the language context. This suggests that full feedback helps in each of these language contexts by enabling participants to learn somewhat accurate rules quickly but that none of these problems can be completely solved with verbalizable rules.

However, the similarity of full and minimal feedback conditions throughout the rest of training suggests that these tasks may require complex, multidimensional strategies that are difficult to verbalize. This interpretation is consistent with findings that categories that cannot be separated by simple, verbalizable rules demonstrate no differences between full and minimal feedback (Chandrasekaran et al., 2014; Maddox et al., 2008; Yi & Chandrasekaran, 2016), but categories separated by verbalizable rules show clear benefits for full over minimal feedback (Maddox et al., 2008, but see Dunn et al., 2012). Our results indicate that full feedback may highlight verbalizable rules that are somewhat useful for learning (e.g., mean F0) and boost initial performance. Instead of encouraging different strategies during learning, full feedback may encourage more *accurate* initial strategies than minimal feedback.

Our results are also consistent with prior demonstrations of the Language Familiarity Effect (e.g., Perrachione, 2018; Thompson, 1987) – performance was better for the Native Mandarin task than the Native English task. Our results add to prior literature on the Other Accent Effect (Stevenage et al., 2012) that show that talkers are harder to identify when they have a foreign accent than one's own native accent in their native language. Specifically, our results extend previous findings by demonstrating that listeners are better at identifying talkers when listening to same-accented foreign speech than native-accented foreign speech. This suggests that the Other Accent Effect may also exist when hearing speech in a foreign language.

While initial learning of Mandarin-Accented and Native English talkers was not significantly different, with more training, performance was better for Mandarin-Accented English than Native English. This pattern of results may indicate that when listening to own-accented foreign speech, listeners may be able to access native language cues that enable better talker identification relative to native-accented foreign speech. This finding is aligned with prior work that demonstrates that talker identification is enhanced when listeners hear talkers produce pseudo-words that have similar sound patterns of their native language than when listening to foreign speech (Perrachione et al., 2015; Xie & Myers, 2015; Zarate et al., 2015). Similar talker sound patterns, regardless of language context, may aid talker identification.

Our results also provide insights about how listeners decide who is talking and, specifically, how they use acoustic cues in different language contexts. We found that, regardless of task, native Mandarin listeners primarily use single-dimension rules on mean F0 or two-dimension rules on mean F0 and speech rate to separate the talkers. In each of these tasks, mean F0 is a reliable, but not perfect, cue for signaling talker identity, as has been demonstrated in prior work (Lavner et al., 2001; Perrachione et al., 2019). This suggests that listeners may use similar strategies across language contexts.

Because mean F0 and speech rate together cannot fully separate these talkers in any language context (Table 1), it is also likely that other dimensions contribute to talker identification. Indeed, the complex and naturalistic problem of talker identification is likely a high dimensional problem. We selected mean F0 and speech rate as target dimensions as they appeared to be likely candidates to aid in talker identification based on prior literature (Lavner et al., 2001; Perrachione et al., 2019; Skoog Waller et al., 2015; Winkler, 2007). We selected specifically *two* dimensions because of the application of the decision bound models is currently limited to two-dimensional problems. A multitude of other dimensions have been examined and future work should focus on how a combination of these dimensions relates to how participants categorize these stimuli by talker.

It is important to note that our participants had a variety of experience and proficiency with English. Second language experience is very likely going to influence performance in these tasks. While we did not examine this directly here, future research should investigate how the extent of experience in English influences native English and same-accented English talker identification.

Overall, our results apply a categorization perspective to understanding talker identification by providing information about how listeners use dimensions to make decisions during learning and by contrasting full and minimal feedback conditions. This work has implications for understanding talker recognition in humans and artificial machine listening contexts.

Acknowledgments

This research was supported by the National Institute on Deafness and Other Communication Disorders (R01DC013315A1 to B.C. and F32DC018979 to C.L.R.) and grants from the Research Grants Council of Hong Kong (14619518 and 14614221 to G.F.).

References

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2019). Gorilla in our Midst: An online behavioral experiment builder. *Behavior Research Methods*, 438242.

Ashby, F. G. (1992). *Multidimensional models of categorization* (F. G. Ashby, Ed.; pp. 449–483). Lawrence Erlbaum.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Rosedahl, L. (2017). A Neural Interpretation of Exemplar Theory. *Psychological Review*, *124*(4), 472–482.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363–378.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610.

Bradlow, A. R. (n.d.-a) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar

Bradlow, A. R. (n.d.-b) SpeechBox. Retrieved from https://speechbox.linguistics.northwestern.edu

Bradlow, A. R., Blasingame, M., & Lee, K. (2018). Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1).

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*, 488–495.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458.

LaRiviere, C. (1975). Contributions of fundamental frequency and formant frequencies to speaker identification. *Phonetica*, *31*(3–4), 185–197.

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The Prototype Model in Speaker Identification by Human Listeners. *International Journal of Speech Technology*, *4*(1), 63–74.

Levi, S. V. (2019). Methodological considerations for interpreting the Language Familiarity Effect in talker processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(2), e1483.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49–70.

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: feedback effects in perceptual category learning. *Cognition*, *108*(2), 578–589.

McLaughlin, D. E., Carter, Y. D., Cheng, C. C., & Perrachione, T. K. (2019). Hierarchical contributions of linguistic knowledge to talker identification: Phonological versus lexical familiarity. *Attention, Perception, & Psychophysics*, *81*(4), 1088–1107.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, 1–12.

Perrachione, T. K. (2018). Speaker recognition across languages. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception*. Oxford University Press.

Perrachione, T. K., Furbeck, K. T., & Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, *146*(5), 3384–3399.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human Voice Recognition Depends on Language Ability. *Science*, *333*(6042), 595–595.

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*(8), 1899–1910.

Sambur, M. R. (1975). Selection of Acoustic Features for Speaker Identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*(2), 176–182.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Skoog Waller, S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology*, *6*, 978.

Soli, S. D., & Wong, L. L. N. (2008). Assessment of speech intelligibility in noise with the Hearing in Noise Test. *International Journal of Audiology*, *47*(6), 356–361.

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology*, *24*(6), 647–653.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131.

van Dommelen, W. A. (1990). Acoustic Parameters in Human Speaker Recognition. *Language and Speech*, *33*(3), 259–272.

Wickens, T. D. (1982). *Models for Behavior: Stochastic Processes in Psychology*. W. H. Freeman.

Winkler, R. (2007). Influences of pitch and speech rate on the perception of age from voice. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1849–1852.

Xie, X., & Myers, E. B. (2015). General Language Ability Predicts Talker Identification. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2697–2702). Cognitive Science Society.

Yi, H.-G., & Chandrasekaran, B. (2016). Auditory categories with separable decision boundaries are learned faster with full feedback than with minimal feedback. *The Journal of the Acoustical Society of America*, *140*(2), 1332–1335.

Yu, M. E., Schertz, J., & Johnson, E. K. (2021). The Other Accent Effect in Talker Recognition: Now You See It, Now You Don't. *Cognitive Science*, *45*(6), e12986.

Yuan, J., Jiang, Y., & Song, Z. (2010). Perception of Foreign Accent in Spontaneous L2 English Speech. *Proceedings of Speech Prosody*.

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1), 11475.