Does English *fish* sound like French *fiche*? Perceptual similarity judgments versus acoustic similarity

Rory Turnbull[1], *Elisa Kiefer*[2], *Sharon Peperkamp*[2]

[1]School of English Literature, Language and Linguistics, Newcastle University, UK
[2]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-PSL / CNRS / EHESS, Paris, France

rory.turnbull@newcastle.ac.uk, elisa.kiefer@sorbonne-nouvelle.fr, sharon.peperkamp@ens.fr

Abstract

We examine the relationship between phonological word similarity judgments from listeners and acoustic measures of similarity. Native speakers of English or French with varying degrees of proficiency in the other language listened to pairs of words, one in French and the other in English. The words were highly phonologically similar but did not overlap in meaning. Participants judged how similar the items sounded. Each item pair was acoustically analyzed with six different acoustic distance metrics. The results demonstrate a weak correlation between the similarity judgments and five of the six acoustic distance measures. Within the experiment, we found an order of presentation effect: when the first word of a pair was in a participant's L1, the pair was rated more similar than if the first word was in the L2. This effect diminished in magnitude with increasing L2 proficiency. We discuss the implications of our results in light of theories of speech representation and processing.

Index Terms: phonology, representation, bilingualism, perceptual similarity, acoustic distance

1. Introduction

Many spoken languages share sounds which are very similar, and may even be transcribed the same way in a phonetic alphabet. Typically these sounds have some small differences, and it has been claimed that "there are no two languages in which the implementation of analogous phonemes is exactly the same" [1, p285]. For example, French and English both have an /i/ phoneme, but this sound is usually shorter in duration and more peripheral in the vowel space in French than it is in English.

Despite these differences, it is uncontroversial to refer to these two vowels as 'the same sound'. Native speakers of both languages will report that these phonemes 'sound the same', and indeed this observation is confirmed in patterns of loanword adaptation and perceptual assimilation [2]. In determining which sounds are 'the same' or 'different', listeners appear to be sensitive to a variety of factors, including phonetic detail, orthography, and a host of psycho-phonological considerations [3, 4, 5].

Attempts have been made to characterize phonological distance through purely acoustic means [6, 7, 8, 9]. These studies have faced normalization difficulties, both in the temporal and spectral dimensions, but advances in speech processing algorithms have led to a gradual increase in their reliability.

An alternative approach, relatively under-utilized in the literature thus far, is to directly ask listeners how similar particular sounds or words are. This method is by its nature more resource intensive than an acoustic comparison but establishes a "ground truth" for perceptual similarity. The goal of the current study is to determine to what extent acoustic measures of similarity can account for listeners' metalinguistic judgments of similarity.

Our motivation here is twofold: there is both theoretical and methodological importance in demonstrating that acoustic measures of similarity are equivalent to human similarity judgments. In terms of theory, this finding would have clear implications for our understanding of the mental representation of speech sounds [10]. In terms of methodology, it would simplify data collection in that we could rely less on human participants and instead make inferences from acoustics alone. Establishing complete equivalence is beyond the scope of the present investigation and we instead seek to establish correlation rather than equivalence.

2. Methods

2.1. Participants

Sixty-one native English speakers living in the US (24 male, 32 female, 5 other gender), age 18–50 years (mean: 33), and twenty native French speakers living in France (15 male, 5 female), age 20–49 years (mean: 31) were recruited via Prolific and tested online.

Thirty of the American participants had learned at least some French, and all of the French participants had learned some English, according to questionnaire data. All participants self-rated their skills in the other language in terms of oral expression, oral comprehension, written expression, and written comprehension. The composite values for these ratings, scaled from 1 to 10, are shown in Table 1.[1]

Table 1: *Means and standard deviations of self-rated L2 language proficiency scores (composite)*

Language background	Self-rated proficiency		
	Mean	SD	N
English L1 French L2	5.68	2.63	30
English L1, no French	1.37	0.76	31
French L1 English L2	8.09	1.54	20

2.2. Stimuli

All 110 stimulus pairs from [11] were selected, with function words removed. The resulting pairs consisted of one English

[1]We opted to combine the oral and written skill scores as they were very similar within each group of participants (English L1 French L2 $M_{\text{oral}} = 5.43, M_{\text{written}} = 5.96, r^2 = .845$; English L1 No French $M_{\text{oral}} = 1.56, M_{\text{written}} = 1.18, r^2 = .278$; French L1 English L2 $M_{\text{oral}} = 7.72, M_{\text{written}} = 8.45, r^2 = .684$).

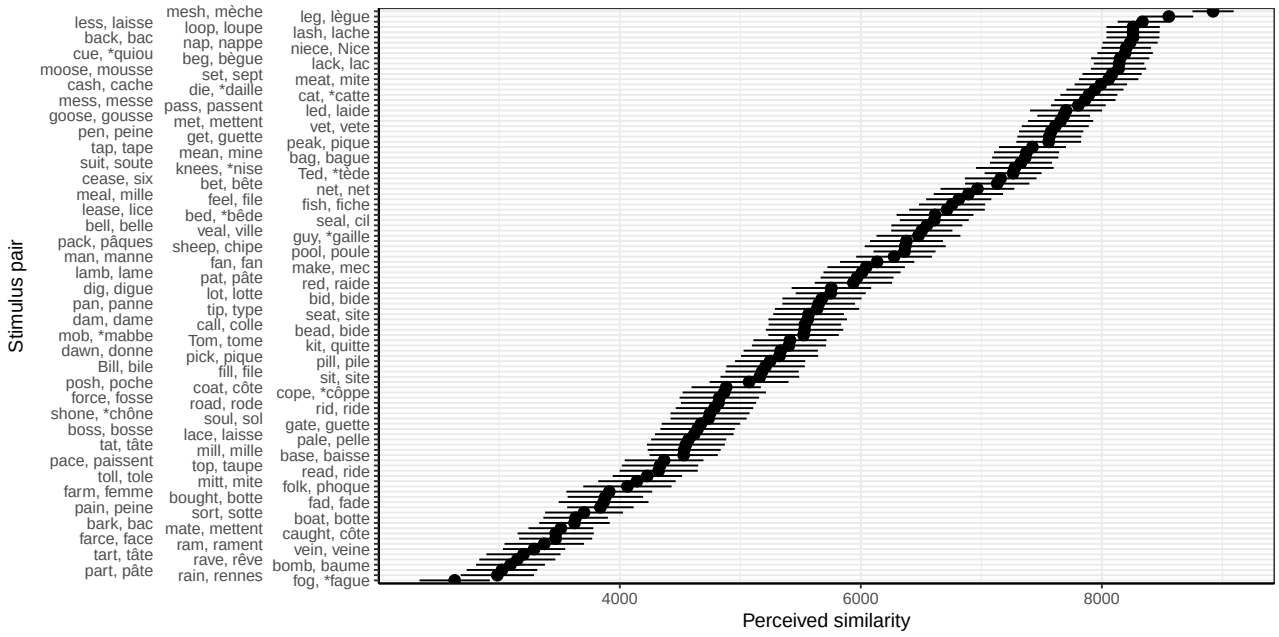

Figure 1: *Mean similarity rating for each stimulus pair. A higher score means a greater perceived similarity. Lines depict standard error. French non-words are preceded by an asterisk.*

and one French item which share a high degree of phonological similarity with no semantic overlap.[2] Ninety-nine pairs involved real words in both English and French (e.g. English *posh*, French *poche* 'pocket'; English *niece*, French *Nice*); the remaining 11 pairs involved an English real word and a French non-word (e.g. English *cope*, French **coppe*). Stimulus items were pronounced by a male native speaker of American English and a female native speaker of French.

2.3. Procedure

The experiment was run in Labvanced [12]. Participants were told that they would hear pairs of more or less similar sounding words, one pronounced by an English and one by a French speaker. They were instructed to pay no attention to the meaning of the words but focus on the sounds only, as they would have to indicate how similar the words in each pair sounded. Item pairs were presented with an ISI of 500ms, and participants indicated their responses on a visual analogue scale by dragging the cursor on a bar whose endpoints were marked "not very similar" and "extremely similar".[3] The scale returned values from 0 (least similar) to 10000 (most similar). After giving their response, they pressed a button to proceed to the next trial, which started 1000 ms later.

The item pairs were presented in two blocks of 55 that differed in language order (i.e. French–English vs. English–French). The ordering of these blocks was counterbalanced between participants. Within each block, items were presented in a random order. Each participant heard each item pair only once in only one order.

2.4. Acoustic analysis

Acoustic similarity of each stimulus pair was calculated using six methods. Five of these methods were based on cosine similarity. We used Shennong [13] to extract feature vectors and compute cosine similarity between words. Features were extracted with the built-in "Spectrogram", "Filterbank" (mel-filterbank), "MFCC" (Mel-Frequency Cepstral Coefficients), "PLP" (Perceptual Linear Predictive analysis of speech), and "Bottleneck" algorithms. For the sixth method, we calculated acoustic absement, following [14], over euclidean similarities on aligned MFCCs. This method has been proposed to correspond well to mental representations of speech [15, 16].

2.5. Statistical analysis

Six linear mixed-effects regression models were constructed, one for each measure of acoustic similarity (z-scored). The models had participants' rating of each stimulus pair as the dependent variable. Other variables in the model were the listener's L1 (either French or English), order of presentation of the pair (L1–L2 or L2–L1), the listener's L2 proficiency (on a scale of 1–10), and an interaction term between order and proficiency. Random intercepts of participant and stimulus pair were included. The model structure was therefore as follows:

$$\begin{aligned} \text{rating} \sim\ & \text{acoustic\_distance} + \text{listener\_L1} + \\ & \text{presentation\_order} + \text{listener\_L2\_proficiency} + \\ & \text{presentation\_order} : \text{listener\_L2\_proficiency} + \\ & (1|\text{item\_pair}) + (1|\text{participant}) \end{aligned}$$

3. Results

Figure 1 depicts the mean ratings for each stimulus pair. As can be seen, there is substantial variability in the ratings, with some

[2] One exception to this restriction was discovered after data collection: French *fan* is an English loanword and means 'fan' or 'enthusiast', and thus has a partial semantic overlap with English *fan*. Removing this item does not alter the results reported here.

[3] All stimulus pairs are at least somewhat similar. For this reason, the lowest endpoint of the scale was "not very similar", as using "extremely dissimilar" would likely lead to responses that do not use the entire range of the scale.

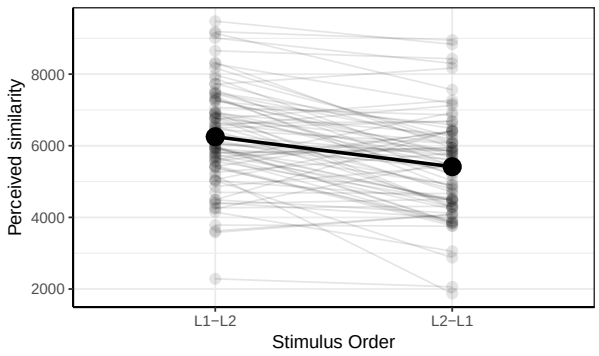

Figure 2: *Grand mean similarity ratings as a function of stimulus presentation order. Lighter points show individual participant means.*

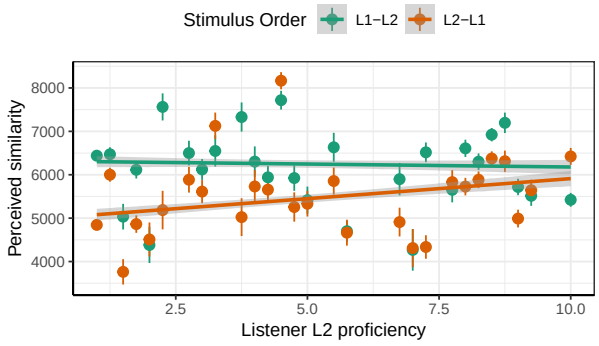

Figure 3: *Similarity ratings as a function of listener L2 proficiency, split by stimulus presentation order. Means and standard errors are depicted with a linear trend overlaid. Note that the apparent order effect is diminished among the more proficient L2 users.*

stimulus pairs being highly-rated and others less so. There is no apparent trend for non-words versus lexical words, although the small number of French non-words in the data does not allow us to assess a potential effect of lexicality.

We now summarize the results of the six models. All models showed the same significant effect of stimulus order ($\beta = -1379$, $t = -16.88$, $p < .0001$ in all models). If the first word in a pair was in a listener's L1, the pair received higher similarity judgments than if the first word was in a listener's L2. Figure 2 shows raw similarity ratings as a function of stimulus presentation order; the difference in similarity rating between the conditions is apparent. Moreover, this effect was modulated by listener's self-rated L2 proficiency, such that participants with higher proficiency showed a smaller order effect than participants with lower proficiency ($\beta = 110$, $t = 7.66$, $p < .0001$ in all models). This interaction is visualized in Figure 3. Note that 'L2' here refers to French for the English listeners, and English for the French listeners.

Table 2 shows the main effects of the different acoustic measures of similarity in each model. As can be seen, all measures except the Spectrogram method showed a significant negative correlation with perceived similarity ratings.

Figure 4 depicts perceived similarity as a function of acoustic similarity. The lack of a correlation with the 'spectrogram' method can be seen in the top left panel. While the linear mixed effects models (Table 2) confirm that statistically significant correlations exist for the other methods, this graph shows that the relationships are not particularly strong. In other words, the participants must be relying on other factors in addition to acoustics when they make their similarity judgments.

Table 2: *Main effects of the different acoustic similarity algorithms in the six different regression models.*

Similarity measure	β	t	p
Cosine similarity			
Spectrogram	-55.76	-0.356	0.723
Filterbank	-453.07	-2.962	0.004
MFCC	-310.22	-2.022	0.046
PLP	-340.17	-2.227	0.028
Bottleneck	-359.60	-2.361	0.020
Absement	-367.92	-2.418	0.017

4. Discussion

4.1. Acoustic measures of distance

All except one of the acoustic measures of distance demonstrated a significant negative correlation with the human similarity judgments. Absement, suggested as a potential candidate for mental representations [14], had the second-largest coefficient after the Filterbank method. The poorest acoustic measure of similarity was the spectrogram method, consistent with prior findings [13].

4.2. Order effects

The order effect was unexpected. Any explanations we can offer are necessarily post-hoc and subject to further validation and testing. Nevertheless, we provide some speculation here.

The interaction between presentation order and proficiency, as depicted in Figure 3, is perhaps the most straightforward to interpret, given the reasonable assumption that higher-proficiency participants have more native-like phonolexical processing of L2 words than lower-proficiency participants. This assumption indeed entails that compared to lower-proficiency participants, higher-proficiency participants have a smaller difference between their L1 and L2 processing. This smaller difference then leads to a smaller order effect.

Turning now to the main effect of order, we first wish to ascertain whether this is a true processing effect or an artifact of particular items. Post-hoc analysis revealed that the mean size of this effect for each item pair is normally distributed ($W = .980, p = .095$), and that there are no apparent patterns in the type of words that have larger or smaller order effects. For example, we observe pairs with /ɛt/ codas in the first, third, and fourth quartile of order effects, suggesting a random rather than principled distribution. However, the mean order effect of each item pair is very weakly correlated with the pair's mean similarity rating ($\beta = -0.089, t = -2.210, p = .029, r^2 = .043$), such that item pairs judged more similar had smaller order effects. At this stage it is unclear what to make of this pattern, but we can conclude that the order effect is not constrained to particular item pairs.

Regarding the processing mechanisms that underlie similarity judgments, it is well-established that L1 phonolexical representations are represented with greater fineness and fidelity

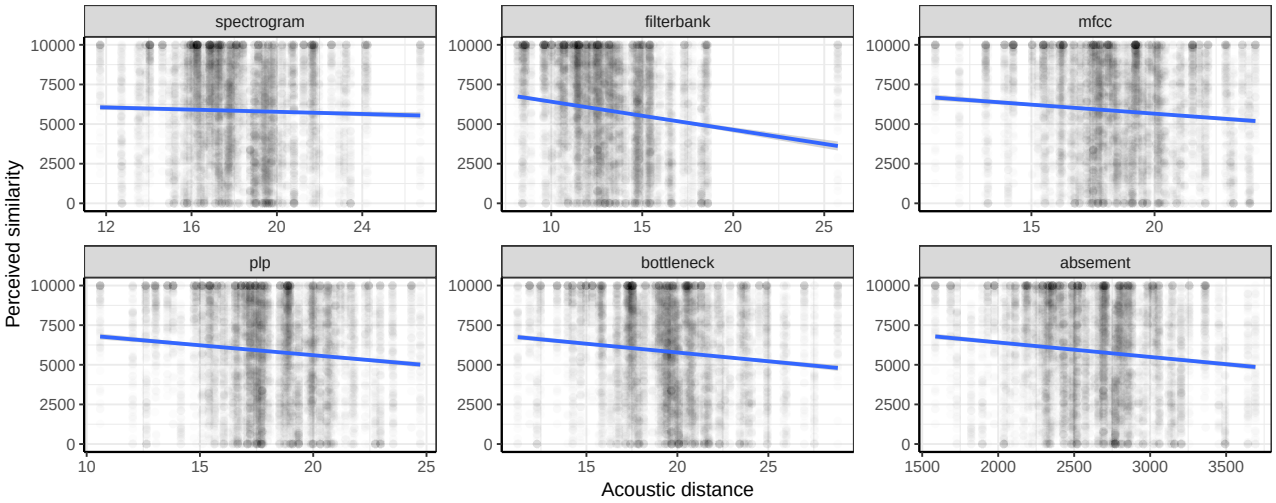

Figure 4: *Similarity ratings as a function of acoustic similarity, by similarity algorithm. A linear trend is overlaid.*

than L2 phonolexical representations [17, 18]. Listeners, upon encountering speech in their L1, are able to encode this speech more effectively and efficiently than they can speech in their L2. This fact motivates two possible explanatory mechanisms for the observed order effect.

One possible mechanism concerns the calculation of distance between two representations. We assume here that the stimulus items are encoded as a multidimensional distribution of features, and that the L2 item necessarily has greater variance than the L1 item. After the listener perceives and encodes the first item, the second item is presented and encoded in turn. In order to judge the similarity of the items, the listener compares the first item to the second [19]. One operationalization of this comparison process could be the mean Mahalanobis distance from the first item's points to the second item. Other methods, such as Kullback-Leibler divergence, will achieve conceptually similar results. If the first item is a finely-encoded L1 representation, then it will align fairly well to a coarsely-encoded L2 representation that is presented subsequently, leading to a high similarity rating. By contrast, if the first item is a coarsely-encoded L2 representation, then it will map poorly on to a finely-encoded L1 representation that follows, leading to a low similarity rating. In other words, the fineness or coarseness of the representation of the second word is like a dartboard: you are more likely to hit the dartboard if it is bigger.

The second possible mechanism treats the first item as a 'prime' or 'attractor'. Again on the understanding that L1 items are encoded more efficiently and effectively than L2 items, these enhanced representations will serve as more effective primes [20]. This explanation predicts faster reaction times in the L1–L2 presentation order than in the L2–L1 order. While reaction times were not captured with the current experiment, this prediction would be straightforward to confirm or disprove in future work.

4.3. One potential task effect

We wish to note here a potential task effect that arose due to the experimental design. All French stimuli can be plausibly interpretable as an English word pronounced with a French accent. For example, French *fiche* can be perceived as a French-accented production of English *fish*. However, the converse is not true—not all English stimuli can be perceived as French words pronounced with an English accent. This is due to the presence of a small number of pairs where the French stimulus is either a nonword (e.g. **coppe*) or a low-frequency word (e.g. *paissent* '(they) graze') that may be hard to recognize as an existing word when presented in isolation. This makes a particular task-specific strategy available to the English L1 listeners that is not available to the French L1 listeners—simply perceiving as if everything is English and then assessing how well the wordforms sound like each other. However, there's no clear prediction for how this strategy would manifest itself as a pattern of results. Indeed, the lack of any meaningful effects of participant L1 within the regression models suggests either that participants did not make use of this strategy, or that the strategy had no meaningful consequences for participants' responses.

5. Conclusion

In this paper we examined relationships between six acoustic measures of similarity and human judgments of phonological word similarity. Five out of the six acoustic measures showed a significant negative correlation with human judgments—in other words, the more acoustically similar a pair of items is, the less similar it was judged to be by human listeners. Despite these correlations, substantial variation in responses was not explainable by acoustics alone, confirming theoretical claims that phonological similarity is influenced by non-acoustic factors [3]. We cannot, therefore, rely solely on acoustics to determine phonological similarity. We also observed an unexpected order of presentation effect; we speculate that this effect may be due to a priming mechanism, but this explanation requires empirical confirmation. Because our experimental task involved two languages, these results are also of interest to studies of interlingual perceptual assimilation. Future work should examine the role of abstract phonological factors in determining similarity.

6. Acknowledgments

This research was supported by the ANR (ANR-17-CE28-0007-01). We are grateful to Lauren Ackerman, Jeff Holliday, and the Newcastle University Phonetics & Phonology Research Group for comments and advice.

7. References

[1] J. Pierrehumbert, M. E. Beckman, and D. R. Ladd, "Conceptual foundations of phonology as a laboratory science," in *Phonological knowledge: Conceptual and empirical issues*, N. Burton-Roberts, P. Carr, and G. Docherty, Eds. Oxford: Oxford University Press, 2000, pp. 273–304.

[2] S. Peperkamp, "Phonology versus phonetics in loanword adaptations," in *The Phonetics-Phonology Interface. Representations and Methodologies*, J. Romero and M. Riera, Eds. Amsterdam: John Benjamins Amsterdam, 2015, pp. 71–90.

[3] C. B. Chang, "Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity," in *The Segment in Phonetics and Phonology*, E. Raimy and C. E. Cairns, Eds. Chichester, UK: Wiley, 2015.

[4] Y. Kang, "Loanword phonology," in *The Blackwell Companion to Phonology*, M. van Oostendorp, C. Ewen, E. V. Hume, and K. Rice, Eds. Malden, MA: Wiley Blackwell, 2011, pp. 2258–2282.

[5] I. Vendelin and S. Peperkamp, "The influence of orthography on loanword adaptations," *Lingua*, vol. 116, no. 7, pp. 996–1007, 2006.

[6] M. Bartelds, C. Richter, M. Liberman, and M. Wieling, "A new acoustic-based pronunciation distance measure," *Frontiers in Artificial Intelligence*, vol. 3, p. 39, 2020.

[7] W. Heeringa, K. Johnson, and C. Gooskens, "Measuring Norwegian dialect distances using acoustic features," *Speech Communication*, vol. 51, no. 2, pp. 167–183, 2009.

[8] J. Mielke, "A phonetically based metric of sound similarity," *Lingua*, vol. 122, no. 2, pp. 145–163, 2012.

[9] K. Yoneyama, "Phonological neighborhoods and phonetic similarity in Japanese word recognition," Ph.D. dissertation, The Ohio State University, 2002.

[10] M. Redford and M. Baese-Berk, "Acoustic theories of speech perception," in *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press, 2023.

[11] R. Turnbull and S. Peperkamp, "Across-language priming in bilinguals: does English *bet* prime French *bête*?" in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 1367–1371.

[12] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: a unified Javascript framework for online studies," in *International Conference on Computational Social Science (Cologne)*. University of Osnabrück Cologne, 2017.

[13] M. Bernard, M. Poli, J. Karadayi, and E. Dupoux, "Shennong: A python toolbox for audio speech features extraction," *Behavior Research Methods*, vol. 55, no. 8, pp. 4489–4501, 2023.

[14] M. C. Kelley, "Acoustic absement in detail: Quantifying acoustic differences across time-series representations of speech data," in *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023, pp. 679–683.

[15] M. C. Kelley and B. V. Tucker, "Using acoustic distance and acoustic absement to quantify lexical competition," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 1367–1379, 2022.

[16] C. H. Redmon, "Lexical acoustics: Linking phonetic systems to the higher-order units they encode," Ph.D. dissertation, University of Kansas, 2020.

[17] S. V. Cook, N. B. Pandža, A. K. Lancaster, and K. Gor, "Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings," *Frontiers in Psychology*, vol. 7, p. 1345, 2016.

[18] M. Llompart and E. Reinisch, "Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts," *Bilingualism: Language and Cognition*, vol. 22, no. 5, pp. 1085–1100, 2019.

[19] M. J. Hautus, N. A. Macmillan, and C. D. Creelman, *Detection Theory: A User's Guide*, 3rd ed. New York: Routledge, 2021.

[20] R. Turnbull and S. Peperkamp, "The asymmetric contribution of consonants and vowels to phonological similarity: Evidence from lexical priming," *The Mental Lexicon*, vol. 12, no. 3, pp. 404–430, 2017.