AN INFORMATION THEORETIC PERSPECTIVE ON PERCEPTUAL STRUCTURE: CROSS-ACCENT VOWEL PERCEPTION

Jason A. Shaw[1], Catherine T. Best[2], Gerard Docherty[3], Bronwen Evans[4], Paul Foulkes[5], Jen Hay[6], Karen Mulak[7]

[1]Yale University, [2]Western Sydney University, [3]Griffith University, [4]University College London, [5]University of York, [6]University of Canterbury, [7]University of Maryland.
jason.shaw@yale.edu, C.Best@westernsydney.edu.au, gerry.docherty@griffith.edu.au, Bronwen.evans@ucl.ac.uk, paul.foulkes@york.ac.uk, jen.hay@canterbury.ac.nz, kmulak@umd.com

ABSTRACT

Analytical tools from Information Theory were used to quantify behaviour in cross-accent vowel perception by Australian, London, New Zealand, Yorkshire and Newcastle UK listeners. Results show that Australian listeners impose expected patterns of perceptual similarity from their own accent experience on unfamiliar accents, regardless of the actual phonetic distance between accents.

Keywords: vowel perception; English accent variation; perceptual assimilation; information theory

1. INTRODUCTION

There is broad consensus that models of speech perception must capitalize on the benefits listeners gain from phonetically detailed exemplars, i.e., episodic memory for speech via encoding continuous phonetic dimensions, reflected in so-called "hybrid" models of spoken word recognition [1-3]. However, identifying the perceptually relevant phonetic dimensions within a particular speech community remains a challenge. Listeners may pay more attention to some dimensions of speech than others, or weight exemplars differently depending on the social context in which they are heard [4]. These factors complicate models of speech categorization based on acoustic/articulatory similarities, as we need to discover empirically the dimensions of relevance to listeners in different speech communities. In this study we present an approach to exploring dimensions of perceptual similarity that makes use of analytical tools from Information Theory [5]. Information Theory offers a general formal framework for quantifying information transfer, which, applied to our experiments, reveals patterns that may elude more traditional analyses in terms of, e.g., categorization accuracy.

The empirical domain of focus is variation in the perceived similarity of vowels across regional accents of English. In a set of nine experimental conditions, listeners from five non-rhotic regional accents of English (Australia [A], New Zealand [Z], London [L], Yorkshire [Y], and Newcastle [N]) categorized 19 vowels, drawing from the lexical sets of [6], in their own accent. Australian listeners also categorized the vowels from each of the other four accents.

Recent work on cross-accent vowel perceptual has shown that listeners of Australian English, when categorizing vowels of other regional accents, show surprising consistency in accuracy patterns [7]. To a greater degree than expected based on sociophonetic descriptions, Australian listener accuracy in categorizing vowels across accents resembles the pattern of accuracy on their own vowels. However, that study provided no baseline on how vowels in other accents are perceived by listeners of those accents, e.g., how London listeners perceive London vowels. In addition, it focused narrowly on the "accuracy" of vowel categorization as opposed to the broader pattern of perceptual confusion represented in a complete confusion matrix. The focus on accuracy is particularly concerning because accuracy tended to be low (well above chance but nowhere near ceiling) even for Australian vowels, the listeners' native accent. The explanation for this pattern in [7] is that the experiment restricted vowel judgments to bottom-up cues, via the use of nonce words.

In this study, we took a more comprehensive approach, using the concept of Entropy from Information Theory to quantify cross-accent speech perception based upon complete confusion matrices. We also contextualised our analysis against new baseline conditions in which listeners of each non-Australian accent categorized their native vowels.

2. APPROACH

To visualize and compare confusion matrices across conditions, we applied a method of hierarchical cluster analysis. Confusion matrices were progressively fused into binary clusters according to an objective function: minimize variance of each cluster, a common technique for clustering [8]. The result is a hierarchical structure representing a series of binary branches, which can be conceptualized as decisions and quantified in bits. The correlation coefficient, Baker's Gamma as implemented in [9], provides a measure of similarity between clustered confusion matrices—a value of 1 indicates two

confusion matrices are identical; a value of 0 indicates no similarity. This analysis provides a basis for comparing across conditions holistically.

Alongside our hierarchical analysis of vowel perception we also computed two quantities based on (Shannon) Entropy. Entropy is a foundational quantity from Information Theory. Generally, Entropy characterizes the amount of uncertainty/ information in a random variable; in our case, the variable is the vowel category. By giving a definite value to a vowel, Entropy (uncertainty) is removed and information is communicated.

The first of our Entropy-based measures is Response Entropy, as in (1), where *v* is a vowel drawn from an inventory of *I* vowels. The theoretical maximum Entropy of a vowel system with 19 vowels is 4.24 bits; this is the case when all vowels occur with equal frequency (which is true of our experiments but not true of naturalistic English corpora). Response Entropy is the Entropy of vowel responses in the vowel perception task. If listeners selected all vowel options equally often, the response Entropy would be 4.24, the theoretical maximum, meaning just over 4 binary choices could distinguish all vowels. If, on the other hand, listeners showed some systematic bias, e.g., choosing one vowel category more often than others, response Entropy would be lower.

$$(1) \qquad H(v) = \sum_{v=1}^{I} -p(v)\log_2 p(v)$$

The second Entropy-based measure is the Conditional Entropy of the stimulus vowel, v_s, given the response vowel, v_r, defined in (2), which is sometimes referred to as the "equivocation of the communication channel" or as "information loss" [10]. Within the context of our task, this quantity describes the uncertainty of the listeners about the vowels they hear. Higher values indicate that responses for a particular vowel are more dispersed across the response categories, i.e., listeners are uncertain about which vowel to choose.

$$(2) \qquad H_{v_r}(v_s) = \sum_{v_r=1}^{I}\sum_{v_s=1}^{I} -p(v_r)\, p_{v_r}(v_s)\log_2 p_{v_r}(v_s)$$

To summarize, we quantified behaviour in our perception experiments with reference to Entropy-based quantities and hierarchical cluster analysis of confusion matrices. The response Entropy quantifies the number of bits, i.e. binary choices, needed to encode participant responses. The hierarchical cluster analysis provides a representation of how binary branches divide the vowels according to similarity. Note that two confusion matrices can have the same response Entropy but different hierarchical clusters, corresponding to different patterns of perceptual similarity; this is something we might expect to see from different groups of listeners responding to the same vowel stimuli, if listener experience shapes perception. Finally, the conditional Entropy of the stimulus vowel given the response vowel or "Information Loss" provides an index of perceptual confusion. This is minimal when listeners reliably choose the same response vowel for a given stimulus vowel and maximized when listeners choose all response vowels equally.

2. EXPERIMENTAL METHODS

The vowel categorization experiments include a total of nine test conditions, which differed in the accent of the stimulus vowels and the accent of the listeners. We refer to the conditions with two letter codes, e.g. "AZ"; the first letter denotes the accent of the listener group and the second denotes the accent of the stimulus items. Five conditions involved Australian listeners: AA, AL, AZ, AY, AN. The other four involved listeners from the other accents listening to their own accent: LL, ZZ, YY, NN.

2.1. Listeners

A total of 139 listeners participated in the study: 12-17 per condition across nine test conditions. Listeners were recruited from local university communities in Western Sydney (A), Christchurch (Z), SE London (L), Newcastle (N), and Yorkshire (Y).

2.2 Stimuli

Listeners were presented with vowels embedded in nonce words. For the target nonce words, 19 English vowels (all monophthongs, diphthongs, and vowels before orthographic <R>, e.g., NORTH) were inserted into the frame /ˈz**V**bə/, which yields no real English words. Twelve speakers each (6f, 6m) from western Sydney, southeast/east/north London, Christchurch, New Zealand, Sheffield/ Leeds, York, Yorkshire, and Newcastle (UK) produced each nonce word six times. Two females and two males of each accent were chosen for the perceptual task stimuli, and two tokens per nonce word per speaker were selected, judged as representative of that accent by a phonetically trained researcher familiar with the accent. Tokens were extracted with 100 ms onset and offset buffers; a ramp and damp were imposed on the initial and final 20 ms. Tokens were normalized to 65 dB.

For each target vowel, a monosyllabic word was chosen to serve as one of the keywords in print that were presented onscreen in an array of response options. Most were of the form /b**V**d/ or /p**V**d/, unless that gave obscure, ambiguous, or no words (e.g., standard English has no such words for the FOOT

vowel, so we used <h**oo**d>). Nodes in Figure 1 are labelled with the response words.

2.3 Procedure

In all conditions, participants first heard a short story of ~6 minutes told by speakers of their native accent. Following the story, they then categorized nonce word vowels, which were presented, depending on condition, in either their own accent or in another of the test accents.

On each test trial, participants heard a single nonce token and saw the vowel keyword grid displayed on a computer monitor. They clicked on the keyword with the vowel that best matched the stressed vowel in the nonce token, then rated how well the nonce token represented the chosen vowel (1 [poor] to 7 [excellent]). Keyword order on the grid was randomized across participants, but was kept constant for a given participant. To familiarize them to the task and grid, listeners completed 20 randomized training trials with nonce tokens from the same speakers who told the pre-test story, one per nonce word. They then completed the categorization test of 160 trials (20 nonce words x 2 tokens x 4 speakers), presented in random order via E-Prime (v. 2.0.8.22). For all conditions, nonce word speakers were different from the speakers used in the pre-test story.

3. RESULTS

Table 1 shows Response Entropy and Information Loss for each accent in the conditions with Australian listeners. The Response Entropy is similar across accents and approximates the theoretical maximum, indicating that listeners are fairly balanced in their use of the response grid. Information Loss varied across accents. At 37.3 bits, it is lowest for the listeners' own Australian accent and increases gradually as the accent varies from London (38.0) to New Zealand (39.9), Yorkshire (39.9) and Newcastle (42.7).

Table 1: Response Entropy and Information Loss for each accent based on Australian listeners.

Accent condition	Response Entropy	Information loss
Australia (AA)	4.13	37.3
London (AL)	4.12	38.0
New Zealand (AZ)	4.17	39.9
Yorkshire (AY)	4.13	40.7
Newcastle (AN)	4.08	42.7

The high Information Loss (~40 bits) relative to Response Entropy (~4 bits) indicates noisy transmission of vowels from speakers to listeners in the context of this task. In natural listener conditions, redundancy from lexical, contextual, and other factors facilitates accurate transmission of vowels. A key mathematical proof from Information Theory is that redundancy (in bits) equal to the Information Loss ensures accurate message transmission in a noisy channel [5]. Notably, the cline in Information Loss across accents mirrors claims about perceptual similarity between Australian vowels and those of the other accents, which have been based on vowel formant measurements, sociophonetic characterizations of the accents, and cross-accent vowel perception [7].

Figure 1 provides examples of tanglegrams comparing hierarchical clustering of confusion matrices. Since the Response Entropy is just over 4 bits, there are on average, just over four binary branches per vowel. The tanglegrams compare AA-ZZ (top) and AA-AZ (bottom). In both cases, AA is on the left. The ways in which AA differs from ZZ are larger than the ways that AA differs from AZ. These differences are apparent visually as longer lines linking vowels in the top, e.g., for vowels in 'bud' and 'bird', than the bottom part of the figure. The correlation between conditions is quantified using Baker's Gamma correlation coefficient.

Figure 1: (top) tanglegram comparing AA (left) and ZZ (right); (bottom) tanglegram comparing AA(left) and AZ (right)

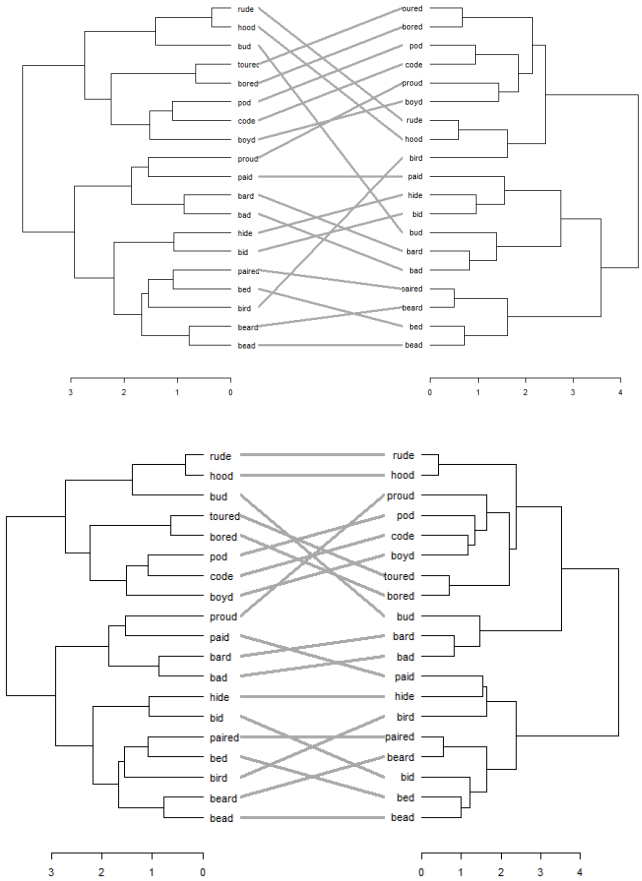

Table 2 reports Baker's Gamma for comparisons across test conditions. We focus on two types of comparison:

1. **Accent differences**: AA is compared with LL/ZZ/YY/NN to probe accent differences in perceptual structure relative to Australian.
2. **Listener effects**: AA is compared with AL/AZ/AY/AN, holding listener group (A) constant while varying the accent of the vowels; relative to the accent baseline, these comparisons reveal the contribution of *listener experience* to perceptual structure.

We first report Baker's Gamma for the accent baseline comparison (1). The perceptual structure that Australians impose on their own vowels was most similar to the structure that Yorkshire listeners (0.436) imposed on their own native-accent Yorkshire vowels, followed by New Zealand (0.420), Newcastle (0.385), and then London (0.270). This pattern is somewhat surprising because it does not match with measures of accent similarity, as reported in past work, or with the pattern of information lost (Table 2). To assess reliability across subjects, we sampled 8 subjects per condition and re-ran the analysis 10,000 times with different samples. The average Baker's Gamma with standard deviation across runs in italics is reported in the second row of Table 2. This analysis presents a different picture. On average, the perceptual structure of listeners is fairly consistent across accents: AA-LL (0.35), AA-ZZ (0.35), AA-YY (0.34), AA-NN (0.39).

Table 2: Baker's Gamma assessing the correlation between hierarchal clustering across conditions.

	LL	***AL***	***ZZ***	***AZ***	***YY***	***AY***	***NN***	***AN***
AA (total)	0.27	**0.63**	0.42	**0.60**	0.44	**0.44**	0.39	**0.44**
AA (samp)	0.35 ***0.22***	**0.45** *0.20*	0.35 *0.19*	**0.44** *0.19*	0.34 *0.12*	**0.47** *0.16*	0.38 *0.17*	**0.45** *0.15*

Comparison testing for listener effects (2) show that Australian listeners tend to impose their native accent perceptual structure on non-native accent vowels; across our entire sample (first row Table 2), the degree of this effect appears to vary across accents. However, the averages from resampling reveal pretty consistent effects: AA-AL 0.45; AA-AZ 0.44; AA-AY 0.47; AA-AN 0.45. Australian listeners' pattern of responses for non-native accents tended to resemble the pattern for their own vowels (AA) regardless of the accent.

4. DISCUSSION

We applied analytical tools from Information Theory to quantify cross-accent speech perception. Across accents, Australian listeners exhibited similar Response Entropy, which approximated the theoretical maximum. This indicates a largely unbiased selection of the 19 choice words offered as response categories. The measure of Information Loss converged with conclusions about accent similarity (of vowels) based on patterns of categorization accuracy—past work showed that Australian listeners were most accurate categorizing their own vowels, followed by London, New Zealand, Yorkshire, and Newcastle [7]. Information Loss takes into account the entire set of responses, not just accuracy, but converges on the same characterization of accent similarity. In this case, less certainty (more response variation) goes hand in hand with lower accuracy.

Response Entropy shows that the different vowel categories can be encoded with just over four bits (or binary choices). Empirical clustering of responses in the confusion matrices revealed how the bits are distributed in each condition. In large part, listeners from the different accents imposed similar structure; more precisely, to roughly the same degree across accents, the perceptual structure that Australian listeners impose on Australian vowels was similar to the structure that other listeners impose on their native accent.

A second finding was that Australian listeners tended to confuse vowels in similar ways across accents, even as the phonetic realization of those vowels and, by corollary, phonetic similarity to other vowels, varied. This was indicated by higher correlations between, e.g. AA-AX, where X stands for any of the other accents, than for AA-XX. In other words, the errors that Australian listeners make categorizing vowels in unfamiliar accents was similar to the errors that they make on their own accent but different from the errors that native listeners of X make on their own accent) This was true for all accents, even as Information Loss varied.

Taken together, the analyses presented here allow us to contextualize cross-accent perceptual results reported in past work [7, 11-14]. The progressive decrease in vowel categorization accuracy by Australian listeners from Australian to London to New Zealand, Yorkshire and Newcastle accents is related to progressive increases in uncertainty (Table 1). This follows in part from listeners' persistence in imposing expected patterns of perceptual similarity even in the face of stimulus variability. The tools of Information Theory provide succinct quantification of this pattern. More broadly, tacit recognition of these conditions, in particular increased Information Loss, driven by, e.g., the lexical level [15-17] may be preconditions for perceptual learning—an issue we plan to pursue in future work.

5. REFERENCES

[1] Pierrehumbert, J. B. (2016). Phonological representation: beyond abstract versus episodic. *Annual Review of Linguistics, 2*, 33-52.

[2] Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory phonology, 10*, 91-111.

[3] Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proc. 16th ICPHS*, Saarbrücken, 49-54.

[4] Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of phonetics, 34*(4), 409-438.

[5] Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*, 379-423.

[6] Wells, J. C. (1982). *Accents of English, vol.2: The British Isles*. Cambridge: Cambridge University Press.

[7] Shaw, J. A., Best, C., Docherty, G., Evans, B. G., Foulkes, P., Hay, J., et al. (2018). Resilience of English vowel perception across regional accent variation. *Laboratory Phonology, 9*(1), 1-36.

[8] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58*(301), 236-244.

[9] Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics, 31*(22), 3718-3720.

[10] Pierce, J. R. (2012). *An introduction to information theory: symbols, signals and noise*: Courier Corporation.

[11] Best, C. T., Shaw, J. A., Docherty, G., Evans, B. G., Foulkes, P., Hay, J., et al. (2015). From Newcastle MOUTH to Aussie ears: Australians' perceptual assimilation and adaptation for Newcastle UK vowels. *Interspeech*, Dresden, 1932-1936.

[12] Best, C. T., Shaw, J. A., Mulak, K. E., Docherty, G., Evans, B. G., Foulkes, P., et al. (2015). Perceiving and adapting to regional accent differences among vowel subsystems. *Proc. 18th ICPHS*, Glasgow, paper 561.

[13] Ying, J., Shaw, J. A., & Best, C. T. (2013). L2 English learners' recognition of words spoken in familiar versus unfamiliar English accents. *Interspeech*, Lyon, 2018-2112.

[14] Best, C. T., Shaw, J. A., & Clancy, E. (2013). Recognizing words across regional accents: the role of perceptual assimilation in lexical competition. *Interspeech*, Lyon, 2128-2132.

[15] Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology, 47*(2), 204-238.

[16] Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543-562.

[17] Van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1483.