Improvements to vowel categorization in non-native regional accents resulting from multiple- versus single-talker training: A computational approach

Sarah M. Wright[1], Jason A. Shaw[1], Catherine T. Best[1], Gerard Docherty[2], Bronwen G. Evans[3], Paul Foulkes[4], Jennifer Hay[5], Karen Mulak[1]

[1]University of Western Sydney, [2]Griffith University, [3]University College London, [4]University of York, [5]NZILBB, University of Canterbury

sarah.wright@uws.edu.au, j.shaw@uws.edu.au, c.Best@uws.edu.au, gerry.docherty@griffith.edu.au, bronwen.evans@ucl.ac.uk, paul.foulkes@york.ac.uk, jen.hay@canterbury.ac.nz, k.mulak@uws.edu.au

Abstract

A computational modeling study was conducted using multinomial logistic regression to predict whether exposure to an unfamiliar regional accent of English would influence vowel categorization in (1) the exposure accent, (2) the native accent, and (3) another unfamiliar accent. We manipulated the number of talkers in the exposure data to determine whether talker variability influenced the efficacy of the training. Results showed a multiple-talker training benefit for the categorization of some vowels. Training also transferred to an untrained accent. Finally, the models predicted that exposure to an unfamiliar accent has a negative impact on vowel categorization in the native accent.

Index Terms: accented speech, computational modeling, vowel categorization, talker variability

1. Introduction

Perceptual studies yield seemingly conflicting results when it comes to the effects of talker variability on speech perception. Mullennix, Pisoni and Martin [1] found that listeners were slower and made more errors when identifying isolated words spoken by 15 different talkers, as compared to a single talker. Conversely, another study [2] found that participants who were trained to understand sentences produced by multiple talkers of a foreign accent tended to understand novel test sentences in the same accent, including those produced by novel talkers, more accurately than those who were trained with a single talker. This indicates that a greater amount of variability in training may lead to more robust learning. However, the benefit of multiple-talker training did not appear to transfer to a novel accent. While these two studies seem to yield opposing results regarding performance in high-variability situations, it is more likely that this is a result of a cost/benefit tradeoff. That is, initial performance is impeded due to the high cognitive demand of processing large amounts of talker variability from the speech signal. The benefit of this initial effort is more robust learning that generalizes better to new talkers, sentences and situations, resulting in better performance in the long-term.

The variable nature of speech requires listeners to be able to identify and understand words regardless of talker differences [3]. This skill is termed *phonological constancy* in recent studies on spoken word recognition across regional accents [4]-[7]. Accented speech features linguistically systematic variation of different types (both subtle and less-subtle) compared to differences between talkers of the same accent [8]. Offering linguistically constrained variation in addition to talker differences [4]-[6], regional accents are an ideal domain to investigate the effects of talker variation on speech perception.

The aim of the present study is to better understand how speech derived from multiple talkers affects the categorization of vowels across regional accents. To achieve this aim, the present study trained multinomial logistic regression (MLR) models to predict vowel identification across three accents of English: Yorkshire (henceforth referred to as York), traditional working-class London (Cockney), and Western Sydney (AusE).

Many of the differences between regional accents of English are due to variations in vowel production [9]. We used the standard, widely used measures of first and second formant frequencies (F1 and F2), and vowel duration.

The modeling experiments were set up with AusE designated as the "native" accent. Cockney vowels tend to be similar to AusE, while York vowels tend to be quite different [9]. Some notable differences between York and AusE include the STRUT, FACE and GOAT vowels (vowel categories are referred to with Wells' lexical sets [9]).

A key characteristic of Northern English dialects, including York, is a lack of split in the FOOT and STRUT categories. Thus for Yorkshire dialect speakers, words such as *put* and *putt* are homophonous, both /pʊt/. In AusE these words are phonetically distinct, /pʊt/ versus /pɐt/. Given this difference, native speakers of AusE would be likely to misinterpret York STRUT vowels as FOOT vowels.

Another difference is that York speakers use monophthong variants for FACE such that it is typically /ɛː/, whereas the AusE FACE vowel is produced as the diphthong /æɪ/. This York pronunciation of the FACE vowel is more similar to the AusE SQUARE vowel /eː/, making it susceptible to being misidentified as SQUARE by AusE listeners.

Finally, the York GOAT vowel is also usually a monophthong, /ɔː/ or /əː/. This vowel in AusE is produced as the diphthong /əʉ/. The York pronunciation of the GOAT vowel is more similar to AusE THOUGHT, /oː/, or NURSE, /ɜː/ making the York GOAT vowel likely to be erroneously categorized by AusE listeners as their THOUGHT or NURSE vowel.

Given the differences between accents discussed above, we expect models trained on AusE vowels will miscategorize these (and other) York vowels. The outstanding questions that

this study aims to answer are: How much improvement can be observed after exposure to the unfamiliar accent? And, will exposure from a single talker or from multiple talkers be more beneficial? To address these questions, we recorded a corpus of multi-accent speech for training and test of vowel categorization models.

2. Method

2.1. Training data

2.1.1. Talkers

The training data were recorded from 12 speakers: two female and two male monolingual speakers each of Australian-, Cockney-, and York-accented English. AusE speakers were aged between 17.0 and 26.3 years (M = 21.7, SD = 3.9) and were from Greater Western Sydney. York speakers were aged between 19.5 and 31.7 years (M = 24.3, SD = 5.4) and were from the Leeds and York areas of Yorkshire. Cockney speakers were aged between 20.2 and 50.6 years (M = 37.7, SD = 14.3) and were from southeast, east and north London.

2.1.2. Materials

The speakers were recorded producing lists of words in isolation. A target word was chosen for each of the 20 vowel categories, generally /CVd/, e.g., *bead* for FLEECE, *bad* for TRAP. Exceptions were made when this frame resulted in a non-word or an ambiguous pronunciation. The target words were: *bad*, *bard*, *bead*, *beard*, *bed*, *bid*, *bird*, *bored*, *boyd*, *bud*, *code*, *hide*, *hood*, *paid*, *paired*, *past*, *pod*, *proud*, *rude*, and *toured*. Speakers recorded each word six times.

Recordings were made in a sound-attenuated booth using a Shure SM10A head-worn microphone and an Edirol UA-25 external sound card. Single-channel recordings were made at 44,000 Hz using CoolEdit 2000 (Syntrillium Software) on a PC. The single channel was doubled to create a stereo monophonic recording. Words were presented to the participant in random order using OPA 1.0 software developed at MARCS Institute. Each recorded token was saved with a 100 ms buffer at beginning and end and normalized to 65 dB.

2.1.3. Vowel segmentation and formant measurements

For each accent, all tokens of all words were concatenated and the resulting long sound files were uploaded to the FAVE-align website [10] along with a transcript file which returned a TextGrid file. This TextGrid file was then uploaded with the soundfile to the FAVE-extract website [10] and, using the FAAV measurement point method, returned a series of Lobanov normalized formant measurements [11] used in the MLR (multinomial logistic regression) models.

2.2. Vowel categorization models

The vowels from each of the three accents were coded for the 20 lexical sets and associated with phonetic measurements extracted using FAVE-extract. We then fit a series of MLR models to the vowel measurements using the *mlr* function in the *nnet* package in *R* (version 3.0.2). Regression coefficients were optimized via gradient descent to maximize the likelihood of the lexical set category given measurements of the vowels. We experimented with different numbers of phonetic parameters but found that two samples of F1 and F2 taken at 20% and 80% of total vowel duration, together with vowel duration as an independent parameter, accounted for the most variance with the fewest number of free parameters. After optimizing the regression coefficients, we used the models to predict the probability of a lexical set given the phonetic parameters. MLR generalizes the binary equation in (1), where *LexSet* is the vowel category that is predicted and f_1, f_2, and V_{dur} are the phonetic parameters, to multiple categories [12]. The exponential term in the equation is a linear function of the phonetic parameters weighted by the regression coefficients.

$$(1)\quad \boldsymbol{P(LexSet|f_1, f_2, V_{dur})} = \frac{1}{1+e^{\beta_0+\beta_1 f_1+\beta_2 f_2+\beta_3+V_{dur}}}$$

As a baseline, we trained a MLR model on the Australian accent as produced by three Australian talkers. We examined residuals to identify outliers so as to exclude tokens that involved speech errors or that were otherwise not characteristic of the Australian accent. Fifteen Australian tokens were excluded for this reason. After training the baseline model, we trained two additional models, a single-talker model and a multi-talker model. These contained 64% Australian-accented tokens and 36% York-accented tokens. The 64% Australian tokens came from the same three Australian talkers used as the baseline model. In the single talker model, the York tokens came from one talker producing three tokens of each vowel. In the multi-talker model, the York tokens came from three talkers each producing one token of each vowel. Thus, the single-talker model and the multi-talker model received the same number of York-accented tokens. The key difference was whether the tokens came from one talker or multiple talkers.

The three MLR models were then tested on different data sets. The first test was with the vowels of a fourth Australian talker, i.e., a novel talker of the main training accent. The next was a novel York talker. Finally, the three MLR models were tested on a Cockney-accented talker, a novel talker of a novel accent. This last test was to evaluate whether training on one unfamiliar accent (York) would yield improvement on a phonetically differing, untrained, unfamiliar accent.

3. Results

The probabilities of lexical set membership generated by the MLR models were interpreted in a forced choice manner. Category membership was determined by the highest probability for each token. Vowel categorizations were judged as incorrect if the category with the highest probability did not match with the intended lexical set. Exceptions were made where vowel categories in the accent being tested were collapsed or showed a merger in AusE. For example, because BATH and START form a single category in AusE, a token of BATH categorized as START was considered a correct categorization.

The three training models accounted for a large proportion of the deviance in the data. Table 1 shows the residual deviance and Akaike's Information Criterion [13] for each of the training models. Not surprisingly, the AusE only model had the smallest residual deviance, followed by the York Single Talker model, then the York Multiple talker model. This reflects the finding in past work that data from multiple talkers is more variable. Next, we evaluate whether this increased variability leads to benefits in the unfamiliar accent.

The MLR model trained on data from three AusE talkers categorized vowels from the novel AusE talker at an overall categorization accuracy of 82.67%. In contrast, when the

unfamiliar talker was a York speaker, this model's accuracy fell to 40% correct. Accuracy on the unfamiliar Cockney-accented speaker was 57.62% correct. The models that received some York training data improved on the novel York speaker to 65% and improved as well on the unfamiliar Cockney speaker. The overall accuracy rates show little difference between the model trained on one talker and the model trained on multiple talkers (see Table 2). However, this model was based on all vowels. We focus next on the specific vowels of interest in these accents.

Table 1. *AIC and residual deviance for training models*

Training Model	AIC	Residual Deviance
AusE Only	241.98	13.98
AusE and York Single Talker	309.10	81.10
AusE and York Multiple Talker	357.30	129.30

Table 2. *Summary of overall modeling accuracy rates*

Training Model	Test Data		
	AusE	York	Cockney
AusE only	82.67	40.00	57.62
AusE and York single talker	73.33	65.00	64.27
AusE and York multiple talker	73.33	65.00	62.60

3.1. Categorization of York vowels

Models trained on York data in addition to AusE data generally improved in their categorization of York vowels, particularly GOAT, FACE, and STRUT. For each test token of these vowels, the assigned categories and their respective likelihoods can be found in Table 3. (Tables 3-4 list only the top category choice if it was selected $\geq$ .75; for top choices of < .75 the next-highest choice is displayed to show the split).

Table 3. *Categorization likelihoods for Yorkshire GOAT, FACE, and STRUT vowels*

	Model Type		
	AusE Only	York Single	York Multi
GOAT	NURSE (1.00)	GOAT (.51); NURSE (.49)	GOAT (.95)
	NURSE (1.00)	NURSE (.95)	NURSE (.63); GOAT (.37)
	NURSE (1.00)	NURSE (.98)	NURSE (.73); GOAT (.27)
FACE	KIT (.53); NEAR (.46)	SQUARE (.78)	KIT (.73); FACE (.13)
	NEAR (1.00)	NEAR (1.00)	NEAR (.92)
	SQUARE (1.00)	SQUARE (.98)	FACE (.82)
STRUT	FOOT (1.00)	FOOT (1.00)	FOOT (1.00)
	FOOT (1.00)	FOOT (.99)	FOOT (.98)
	FOOT (1.00)	FOOT (.83)	FOOT (.83)

The single-talker York model resulted in improved categorization over the AusE-only model, and further improvement was found for the multiple-talker York model. Improvement can be seen in two ways. First, the number of correctly categorized tokens increased. For GOAT and FACE, the number of correct categorizations increased from zero to one. The second way in which improvement can be observed is in the decreased probability of incorrect categories. For example, the probability of categorizing the second GOAT token as NURSE dropped from 1.00 (AusE) to .95 (York-single) to .63 (York-multi). All three vowels show improvement of this kind on at least some tokens.

3.2. Categorization of native vowels

Although training on York improved categorization of York tokens, it had a detrimental effect on categorization of some native vowels. This was most pronounced for the same vowel categories that showed improvement with training on York tokens (GOAT, FACE, and STRUT). The assigned categories of these vowels and their respective likelihoods are shown in Table 4. With exposure to York, some AusE tokens were miscategorized and, for correctly categorized tokens, predicted probabilities decreased. This drop tended to be greater for multi-talker York exposure than to single talker exposure.

Table 4. *Categorization likelihoods for Australian GOAT, FACE, and STRUT vowels*

	Model Type		
	AusE Only	York Single	York Multi
GOAT	GOAT (1.00)	GOAT (1.00)	GOAT (1.00)
	GOAT (1.00)	GOAT (1.00)	GOAT (.99)
	GOAT (1.00)	NURSE (.93)	FOOT (.90)
	STRUT (.76)	STRUT (.62); GOAT (.37)	STRUT (.64); GOAT (.35)
FACE	FACE (.51); FLEECE (.49)	FLEECE (1.00)	FLEECE (.88)
	FACE (1.00)	FLEECE (.85)	FLEECE (.59); FACE (.41)
	FACE (1.00)	FACE (1.00)	FACE (.97)
	FACE (1.00)	FACE (1.00)	FACE (.90)
STRUT	STRUT (1.00)	STRUT (.97)	STRUT (.95)
	STRUT (1.00)	STRUT (.61); TRAP (.39)	TRAP (.95)
	STRUT (1.00)	TRAP (1.00)	TRAP (1.00)

There were some cases of improved categorization of York vowels due to training on York data, however, that failed to cause decreased performance on AusE vowel categorization. This happened for GOOSE and KIT. Thus, neither the positive nor negative effects of exposure are uniform across vowels.

3.3. Categorization of Cockney vowels

Generally, the AusE-only model categorized Cockney vowels poorly (57.62%), but not as poorly as the York vowels (40%). Categorization of some Cockney vowels improved when the model included York training data. Such vowels include DRESS, FOOT, and GOOSE (see Table 5). In these cases, though, the effect of multiple talkers was rather equivocal. For DRESS and FOOT, multiple talker exposure led to further improvement over single talker exposure but the effect was in the opposite direction for GOOSE. For FOOT, single-talker York exposure interfered (slightly) while multi-talker exposure offered a solid boost.

Table 5. *Percentage of correct categorizations of Cockney vowels*

	Model Type		
	AusE Only	York Single	York Multi
DRESS	7.14% (1/14)	28.57% (4/14)	42.86% (6/14)
FOOT	6.67% (1/15)	0% (0/15)	43.75% (7/16)
GOOSE	41.18% (7/17)	82.35% (14/17)	47.06% (8/17)

4. Discussion

This is, to our knowledge, the first study to investigate the effects of single- and multiple-talker training on cross-accent vowel identification using a computational approach. It provides an important supplement to previous findings from perceptual studies on the same topic. Key findings of the present study are that training on York vowel measurements leads to improvements in categorization of both York and Cockney vowels. Further, the results indicate that for some specific contrasts, training from multiple talkers leads to greater improvements in categorization than training from a single talker. Interestingly, with training on a non-native accent, accurate categorization of native vowels is diminished.

The finding that training on York vowels leads to improved categorization of York vowels is consistent with plasticity of phonological categories, which dictates that category boundaries can shift as increased variability is encountered in their phonetic realization [14]. Similarly, the reduction in accuracy for AusE vowels following York training indicates that exposure to an unfamiliar accent can alter native-accent vowel categories. We note that the approach we have taken treats all training tokens equally, regardless of talker, a naïve assumption given the social relevance of phonetic variation. Whether the model predictions reflect how listeners actually respond to regional accent variation requires perceptual experimentation.

On accuracy rates, which are admittedly a coarse-grained measure of model performance, the differences between single- and multiple-talker training were negligible. When we focus on vowels known to differ across accents, however, we see a clear result in both accuracy and predicted probabilities. Multi-talker training provides a distinct advantage over the single-talker training for the York GOAT, FACE, and STRUT vowels. This finding supports the claim that multi-talker training facilitates category learning. Interestingly, as the models themselves are naïve to the number of talkers included in the data, it would appear that the superior performance of the multi-talker model is due to the structure of the data itself, signifying that increased talker *variability*, rather than increased talker *numbers,* is what provides more robust sampling of vowel categories.

The improvements on categorization of Cockney vowels for both the single- and multi-talker models over the AusE-only model indicates that training may indeed transfer across accents, in contrast with the findings of Bradlow and Bent [2]. The reason for this discrepancy in findings may be due to the types of accents used. While Bradlow and Bent [2] used foreign accents (Chinese- and Slovakian-accented English), the present study used two regional English (native language) accents. Regional accents tend to be more phonetically similar to each other than two foreign accents. This is consistent with our finding that Cockney DRESS, FOOT, and GOOSE experienced increases in accuracy, as these vowels are either the same (FOOT-/ʊ/; DRESS-/e/ or /ɛ/) or similar (GOOSE-/u/ in York, and /ʉ/ in Cockney) across the two accents. This indicates that training may transfer across accents if the two accents are sufficiently phonetically similar.

5. Conclusion

The modeling results suggest that exposure to a non-native regional accent allows for more accurate vowel categorization for that accent, and that this training may transfer to other phonetically similar accents. Furthermore, exposure may have a detrimental influence on native vowel categorization, though further investigation is required to test model predictions with human listeners. Finally, training with multiple talkers appears to provide more robust categorization than single talker training, particularly when vowel contrasts are susceptible to confusability across accents.

6. Acknowledgements

This research was funded by ARC grant DP120104596 and a MARCS Institute internship to the first author.

7. References

[1] J. W. Mullennix, D. B. Pisoni, and C. S. Martin, "Some effects of talker variability on spoken word recognition," *J. Acoust. Soc. Amer.*, vol. 85, no. 1, pp. 365-378, 1989.

[2] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707-729, 2008.

[3] M. Joos, "Acoustic phonetics," *Language*, vol. 24, no. 2, pp. 5-136, 1948.

[4] C. T. Best, "Devil or angel in the details? Complementary principles of phonetic variation provide the key to phonological structure," in *Sounds, representations and methodologies: Essays on the phonetics-phonology interface* (Current Issues in Linguistic Theory book series), J. Romero and M. Riera, Eds. Amsterdam, The Netherlands: John Benjamins, in press 5/2014.

[5] C. T. Best, J. A. Shaw and E. Clancey, "Recognizing words across regional accents: The role of perceptual assimilation in lexical competition," in *Interspeech*, Lyon, France, 2013, pp. 2128-2132.

[6] C. T. Best *et. al.*, "Development of phonological constancy: Toddlers' perception of native- and Jamaican-accented words," *Psych. Sci.*, 20, no. 5, pp. 539-542, 2013.

[7] K. E. Mulak *et. al.*, "Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify familiar words spoken in a non-native regional accent," *Child Dev.*, vol. 84, no. 6, pp. 2064-2078, 2013.

[8] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," J. Acoust. Soc. Amer., vol. 116, no. 6, pp. 3647-3658, 2004.

[9] J. C. Wells, *Accents of English*, Cambridge, UK: Cambridge University Press, 1982.

[10] I. Rosenfelder, J. Fruewald, K. Evanini and J. Yuan, *FAVE (Forced Alignment and Vowel Extraction) program suite*, http://fave.ling.upenn.edu, 2011.

[11] B. M. Lobanov, "Classification of Russian vowels by different speakers," *J. Acoust. Soc. Amer.*, vol, 49, no. 2B, 606-607, 1971.

[12] B. McMurray and A. Jongman, "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," *Psych. Rev.*, vol. 118, no. 2, 219-246, 2011.

[13] H. Bozdogan, "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345-370, 1987.

[14] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cog. Psych.*, vol. 47, no. 2, pp. 204-238, 2003.