Extracting binary features from speech production errors and perceptual confusions using Redundancy-Corrected Transmission

Zhanao Fu
University of Toronto
zhanao.fu@mail.utoronto.ca

Ewan Dunbar
University of Toronto
ewan.dunbar@utoronto.ca

Abstract

We develop a mutual information-based feature extraction method and apply it to English speech production and perception error data. The extracted features show different phoneme groupings than conventional phonological features, especially in the place features. We evaluate how well the extracted features can define natural classes to account for English phonological patterns. The features extracted from production errors had performance close to conventional phonological features, while the features extracted from perception errors performed worse. The study shows that featural information can be extracted from underused sources of data such as confusion matrices of production and perception errors, and the results suggest that phonological patterning is more closely related to natural production errors than to perception errors in noisy speech.

1 Introduction

Phonological features have usually been assumed to be phonetically grounded in addition to explaining phonological behaviour. Yet the sources of phonetic data that have been used to infer the nature of phonological features are largely limited to physical acoustic and articulatory measures. Furthermore, the analytical methods available to infer features that are consistent with phonetic data are limited. This study proposes a new method for automatically inferring binary features from similarity matrices, which lends itself to directly studying data relevant to human phonetic processing: here we study perception and production errors.

Previous work has attempted to infer phonetically-grounded features using clustering (Lin, 2005; Lin and Mielke, 2006; Mielke, 2008, 2012; Shain and Elsner, 2019). For example, Mielke (2008) modelled consonant similarity using hierarchical clustering applied to perceptual confusion data, which combines consonants together into nested clusters.

However, clustering does not directly output features in the usual sense of independent, cross-cutting properties of phonemes. Non-hierarchical clustering applied to phonemes yields a flat set of classes, the equivalent of a single binary or n-ary feature. Hierarchical clustering yields classes that can contain other class divisions (for example, a cluster of vowels can be subdivided into a cluster of high and a cluster of low vowels, and so on). However, in typical approaches to hierarchical clustering, decisions as to how to make sub-clusters are taken independently in each cluster. Features are thus not allowed to have scope over more than one sub-cluster. Not only does this contrast sharply with usual approaches to phonological features which naturally give rise to parallel relations across clusters—the "proportional oppositions" of Trubetzkoy (1969)—it means that any data about similarity between phonemes across clusters is necessarily ignored by such algorithms.

To address these issues, we develop a method inspired by Miller and Nicely's (1955) analysis of confusion matrices, based on an information-theoretic measure of *feature transmission*. We first introduce the algorithm and demonstrate it using an artificial example. Next, we report an experiment where the feature extraction algorithm is applied to phoneme perception and production errors, and the extracted features sets are evaluated based on their utility and efficiency in describing phonological classes. Finally, we discuss the insights yielded for the study of phonological features.

Although the paper infers phonological features from data, our goal is not to argue that phonological features are emergent. This paper analyzes confusion data, and determines what set of features would be most compatible with the data (under certain assumptions). While this could be consistent with a hypothesis that learners infer features based on their own confusions, we tend toward the opposite view: features are primary, and feature

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 95-107.
Amherst, Massachusetts, June 15-17, 2023

similarity is a *cause* of confusions. In any case, our analysis is correlational, and as such it is neutral to what is the cause and what is the result. The question is merely what features best explain the data at hand.

Of course, if we do assume that feature representations are one cause of errors (rather than assuming that features are emergent from error patterns), we must accept that feature similarity is only one cause among others—for example, noise in the audio signal, the nature of that noise, physiological constraints on production, and phonological neighbourhoods (Vitevitch, 2002), among other things. For our purposes, we need to assume that the effect of distinctive features on error patterns is strong enough to be detected in spite of these other factors.

2 Extracting feature with Redundancy-Corrected Transmission

2.1 Background

Miller and Nicely (1955) analyzed confusion matrices from an identification task in which participants heard a CV syllable in noise (a consonant followed by /ɑ/) and had to provide a phonemic label for the onset consonant. They developed an information-theoretic measure of feature transmission in a confusion matrix, using it as part of an argument that listeners use distinctive features in speech perception.

Miller and Nicely assumed that speech processing works by transmitting information over a fixed number of channels (features). They used five features to analyze English consonants (voicing, nasality, affrication, duration, and place). By analyzing the confusions between consonants with opposing values for each feature separately (e.g., between voiced and voiceless sounds), they measured the amount of information faithfully transmitted for each feature under various amounts of additive noise. They argue that the result of this analysis suggests that each of these five features is perceived by listeners independently of the others, since the sum of the information transmitted for these five features is close to the the amount of transmitted information measured if phonemes are not organized into features—little information is lost by analyzing phonemes into independent features.

For our purposes, it is not this argument that matters but their transmission measure itself, which can be seen as a measure of how "consistent" a hypothetical feature is with a given confusion matrix. In particular, a hypothetical feature which is consistently extremely poorly transmitted is clearly not implicated in perception. In what follows, we develop this intuition further and show its limitations, and motivate the use of a further term penalizing feature **redundancy**. We show that, despite its limitations, the idea of discovering a set of features with high transmission and minimal redundancy leads to satisfactory results in an artificial example.

2.2 Developing the algorithm

We use a hypothetical phoneme mapping process within a 4-phoneme inventory [ABCD] to illustrate these ideas. We assume these phonemes are transmitted via some noisy process (for example, perception or production) whose goal is accurate transmission—in other words, to faithfully map an input phoneme to itself. Table 1 summarizes a possible outcome from repetitions of this transmission process with different input phonemes.

		Input			
		A	B	C	D
Output	A	10	8	2	0
	B	8	10	0	2
	C	2	0	10	8
	D	0	2	8	10

Table 1: A confusion matrix of the hypothetical mappings in a four-phoneme system with two features.

Furthermore, we assume that, in this hypothetical process, the phonemes are transmitted by transmitting the values of two underlying features *f1* ([AB | CD]) and *f2* ([AC | BD]). As features are often transmitted with different degrees of degradation (Miller and Nicely, 1955), we make it so that *f1* is maintained better than *f2*, resulting in more confusions between phoneme pairs that are differentiated by *f2* (such as A and B) than between phoneme pairs differentiated by *f1* (such as A and C). Our goal of feature extraction is to infer the true underlying features (*f1* and *f2*) based only on the confusion matrix. To achieve this, we consider all potential features, i.e., all binary groupings (While nothing prevents the algorithm we develop here from being used with n-ary features, we restrict the current paper to binary features.). We examine how well each potential feature is transmitted by collapsing the confusion matrix according to that feature. We show this in Table 2 for the feature that splits the inventory into [AB | CD] (which happens to be one of the true features used in transmission).

		Input	
		+	−
		AB	CD
Output	+ AB	36	4
	− CD	4	36

Table 2: Collapsed confusion matrices for the example in Table 1 according to the feature that splits the inventory into [AB | CD].

Higher counts on the diagonal represent more faithful transmissions in the collapsed confusion matrix. Thus, even at first glance, the feature in Table 2 is a good candidate for a feature which is transmitted faithfully. To quantitatively evaluate how well a feature is preserved in the output, we calculate the *transmission* of a signal from the input (I) to the output (O) with Equation 1 as defined in Miller and Nicely (1955).

$$T(I;O) = \sum_{o \in O} \sum_{i \in I} p(i,o) \log \frac{p(i,o)}{p(i)p(o)} \qquad (1)$$

When the confusion matrix is collapsed based on a potential feature f, $T(X_f;Y_f)$ evaluates the how much information about the feature is transmitted.

The transmission alone can capture how much information is transmitted, but it is not sufficient to evaluate how well a feature is transmitted. This is because the transmission value is influenced not only by how well information from the input is preserved in the output, but also by both how much information was contained in the input in the first place. In order to eliminate the influence of the information in the input, we instead evaluate the proportion of the input information successfully transferred to the output. First, we quantify the amount of information in the input by calculating the *entropy* of input re-coded with the feature, as defined in Equation 2:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \qquad (2)$$

in order to calculate the *relative transmission* $T_{rel}(X_f;Y_f)$ of the input information with respect to the feature f:

$$T_{rel}(X_f;Y_f) = \frac{T(X_f;Y_f)}{H(X_f)} \qquad (3)$$

in which $H(X_f)$ is the amount of information in the input and $T(X_f;Y_f)$ is the amount of information shared by the input and the output.

With the relative transmission criterion, we can evaluate all possible candidate features to characterize the inventory [ABCD], as seen in Table 3. The relative transmission of the true underlying feature f_1 (feature I in the table), is higher than that of any other hypothetical feature, as expected, given that our constructed transmission process was one in which this feature was well-transmitted.

However, to extract a set of relevant features, simply seeking a set of features in which each feature individually has a high relative transmission would usually not result in an ideal feature set. This is because a highly informative feature can often undergo a minor adjustment to create a slightly different, spurious, feature that also has high transmission. Consider Table 3 again: hypothesized feature II corresponds to the second true underlying feature that was used to generate the example, *f2*. While its relative transmission of 0.029 is higher than that of the (incorrect) feature III, it is still lower than that of features IV and V. These features have a high relative transmission because they largely overlap with the well-transmitted feature *f1*, grouping together either [CD] (feature IV) or [AB] (feature V). In order to avoid extracting features partially containing the information included in already selected feature, we consider the redundancy of the new feature with respect to each old feature by calculating the *mutual information* $I(X;Y)$. The mutual information captures the degree of association between the states of two variables. As such, it can be used to evaluate the similarity between two features. The mutual information $I(X;Y)$ for two discrete random variables X and Y is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad (4)$$

When evaluating the similarity between features, X and Y are the counts of the input variable re-coded with the two features, respectively.

To keep the mutual information between features on the same scale as the relative transmission of features, we also define a *relative mutual information* $I_{rel}(X_{f_a};X_{f_a})$ between the features f_a and f_b to quantify a new feature's redundancy with respect to an existing feature f_a.

$$I_{rel}(X_{f_a};X_{f_b}) = \frac{I(X_{f_a};X_{f_b})}{H(X_{f_a})} \qquad (5)$$

Features	I		II		III		IV		V	
Value	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$
	AB	CD	AC	BD	AD	BC	A	BCD	C	ABD
$+$	36	4	24	16	20	20	10	10	10	10
$-$	4	36	16	24	20	20	10	50	10	50
$T_{rel}(X_f;Y_f)$	0.531		0.029		0		0.091		0.091	
$J(f;S)$ Step 1	**0.531**		0.029		0		0.091		0.091	
Step 2	\		**0.029**		0		-0.22		-0.22	

Table 3: Evaluating features in the four-phoneme system from Table 1. The table includes collapsed confusion matrices according to different features, the corresponding $T_{rel}(X_f;Y_f)$, and the RCT criterion $J(f;S)$ at two steps of feature extraction. The $J(f;S)$ values of the selected feature at each step are marked in bold. After two steps, the selected features are efficient to differentiate all phonemes and the algorithm ends.

Together this leads us to propose the Redundancy-Corrected Transmission (RCT) criterion $J(f,S)$:

$$J(f;S) = T_{rel}(X_f;Y_f) - \frac{1}{|S|}\sum_{f_i \in S} I_{rel}(X_f;X_{f_i}) \tag{6}$$

In the RCT criterion we use the average of the relative mutual information between the candidate feature (f) and each of the features that are already selected ($f_i \in S$) to minimize redundancy. In addition to this, we also filter the non-contrasting features from candidate feature set before each step of feature selection. Non-contrasting features are defined as the candidate features that do not create new contrast between phonemes given a set of selected features. For example, in a hypothetical consonant inventory [p t f s m n v z], assuming that two features [p t f s | m n v z] ([voice]) and [p t m n | f s v z] ([continuant]) have been selected, then the feature [p t v z | m n f s] would be a non-contrasting feature since it does not create any divisions in the smallest classes (i.e., [p t], [f s], [m n], [v z]) created by the two previous features. This filtering process ensures that the algorithm finds a compact set of features to encode all phonemes.

The extraction process above is summarized in Algorithm 1.

Algorithm 1: Binary feature extraction algorithm with RCT.

```
Data: A confusion matrix for $n$ items
Result: A set of binary features
$F \leftarrow \varnothing$
for $i = 1$ to $2^{n-1}$ do
  | $F = F \cup i$ (as a binary string)
end
$S \leftarrow \varnothing$
while Not all phonemes have distinct featural representations do
    $f_{select} =_{f \in F} J(f,S)$
    $S = S \cup \{f_{select}\}$
    $F = F - \{f_{select}\}$
    for $f \in F$ do
        if $|unique(X_{S \cup \{f\}})| = |unique(X_S)|$ then
          | $F_{redundant} = F \cup \{f\})$
        end
    end
    $F = F - F_{redundant}$
end
```

2.3 Preprocessing

Finally, we will discuss the preprocessing steps that are important in the preparation of confusion data for feature extraction. In real data, especially in the errors collected from natural speech, three issues are often present.

First, some input phonemes may present very few errors. The sparsity of the data for a given phoneme means that it may be difficult to distinguish between hypothesized features on the basis of this phoneme. We address this issue by applying add-one smoothing to the data. In add-one smoothing, we take each column in the confusion matrix that corresponds to the counts (number of errors) for the input phoneme, then add one to all the values in the column. Second, in some kinds of data, the number of examples of each phoneme in the input may not be balanced. This is notably the case in speech error data, which is observational. To avoid high-frequency phonemes having an undue influence, we balance the data by converting the matrices of the error counts into the error probability for each phoneme. Summing up these first two steps, we estimate the probability of mapping input phoneme i to output phoneme j as $p_{ij} = (n_{ij} + 1)/((\sum_i n_{ij}) + n_{jj})$.

The third potential issue arises in the speech error data: while the data lists the errors, it does not record counts of the number of correctly articulated instances. Missing faithful transmissions could potentially lead to errors in feature extraction.

		Input		
		x	*y*	*z*
Output	*x*	○	✓	
	y	✓	○	✓
	z		✓	○

Table 4: Confusion matrix for a hypothetical phoneme inventory. Check marks represent confusions phonemes, circles represent faithful mappings. Without the faithful transmissions, *x* and *z* cannot be differentiated.

Consider the example in Table 4, a hypothetical phoneme inventory with three phonemes *x*, *y*, *z*, and two underlying features, one separating *x* and *y* against *z*, the other separating *y* and *z* against *x*. Without the faithful mappings, both *x* and *z* would only have data from confusions with *y*, making it impossible to differentiate *x* and *z*. As a result, the incorrect feature [x z | y] has the highest transmission and would be selected as the first feature. In order to prevent similar issues in the data where faithful mappings are missing, the diagonal of the confusion matrix needs to be filled in before the feature extraction.

In our experiment, we fill the diagonal cells in the confusion matrices with the sum of the error counts in the corresponding column, which results in a 50% error rate for each input phoneme. The 50% error rate provides information of phoneme identity to address the issue described above, while also maintains the contrasts between phonemes.

Table 5 shows the preprocessed data after each step, from the artificial example in Table 1.

3 Experiment

We apply Algorithm 1 to a perceptual confusion matrix from Miller and Nicely (1955), as well as to a collection of speech error data from Fromkin (1971). We evaluate how well the resulting features can be used to define natural classes in English.

	A	B	C	D
A	10	8	2	0
B	8	10	0	2
C	2	0	10	8
D	0	2	8	10

(a) original data

	A	B	C	D
A	11	9	3	1
B	9	11	1	3
C	3	1	11	9
D	1	3	9	11

(b) add-one smoothing

	A	B	C	D
A	11	0.7	0.2	0.1
B	0.7	11	0.1	0.2
C	0.2	0.1	11	0.7
D	0.1	0.2	0.7	11

(c) normalizing error rates

	A	B	C	D
A	1	0.7	0.2	0.1
B	0.7	1	0.1	0.2
C	0.2	0.1	1	0.7
D	0.1	0.2	0.7	1

(d) filling diagonals

Table 5: Confusion matrices showing the outcome after each step of preprocessing from the example data.

3.1 Data: Perception errors

We analyzed perception errors from Table III (shown in Table 8 in the Appendix) of Miller and Nicely (1955), which summarizes the result from a syllable identification experiment. In the experiment, the stimuli are [Cɑ] with 16 English consonants as the onset. The acoustic stimuli underwent frequency modulation, and noise was added to the stimuli. The data from the condition with the widest-band noise (200-6500 Hz) was chosen in the current study. This choice was made to avoid potential biases due to the exclusion of frequency ranges of greater importance for a subset of features. The condition with a relatively low S/N ratio of -12 dB was chosen so that weakly similar phonemes could still be confused with each other, potentially revealing more information about features that are usually well preserved during transmission.

3.2 Data: Production errors

Speech error data were collected from the Fromkin Speech Error Database web interface.[1] The database contains spontaneous speech errors from natural speech. The search query included "English" as the "target language," "phonological" as the "error type," "substitution" as the "process procedure," and "all" in other fields. The entries that also had "addition" or "exchange" as the "process procedure" in any analysis were excluded. Then, entries were manually removed if they involved the following: (1) a change in the number of segments in the same syllable component (e.g., "small" →

[1] https://www.mpi.nl/dbmpi/sedb/sperco_form4.pl

"fall"; [ɜɹ] was considered a single segment); (2) changes of multiple syllable components (e.g. "detectors" →"locators"); (3) blending of two words (e.g., "jumped"/ "leapt" →[dʒipt] "jeapt"); (4) mispronunciation due to orthography (e.g., [sɑm] "psalm" →[pɑm] "palm"). Only phonemes that were present in both production and perception data were kept in the analysis, namely, the sixteen consonants [p t k b d g f v θ ð s z ʃ ʒ m n]. This resulted in 455 production errors summarized in Table 9.

3.3 Evaluation

To evaluate how well the extracted features correspond to the features that are actually used in the English language, we examine the the feature sets' capacities in defining natural classes, which are the groups of phonemes that pattern together in phonological alternations.

English rule-based sound patterns from P-Base (Mielke, 2008) were used to extract natural classes in English phonology. The English patterns in P-Base were produced with reference to Jensen (1993); McMahon (2002). The search resulted in 9 rule-based natural classes (found as the left environment, the right environment, the target, or the output of the rule). Some natural classes contain phonemes that are not included in the 16 consonants for feature extraction in this study—in these cases, the extra phonemes were removed. The patterns yielded 9 unique natural classes.

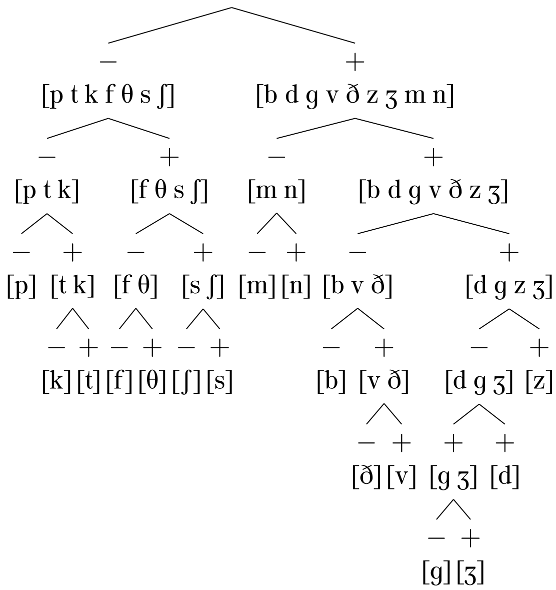
The evaluation of a discovered feature set was based on that feature set's minimal feature definition for the set of phonemes that is the closest to attested natural class in terms of the number of different phonemes, where the feature definition is formed by a single feature value or by the conjunction of multiple feature values.

We also tested how well a reference set of distinctive features could define the natural classes to compare with extracted feature sets. We use a set of seven phonological features from *the Sound Pattern of English* (SPE; Chomsky and Halle (1968)). We take these features to be reasonably well adapted to capturing English phonological classes, and thus a useful point of contact with English phonology. The SPE features included are [nasal] ([nas]), [voice] ([voi]), [continuant] ([cont]), [strident] ([strid]), [coronal] ([cor]), [anterior] ([ant]), and [distributed].[2]

3.4 Results

Here we present the extracted results and compare the extracted features with traditional phonological features. The goal of this section is to assess whether the discovered features are meaningful beyond describing the errors in perception/production.

3.4.1 Perception



1: voi [p t k f θ s ʃ | b d g v z ð ʒ m n]
2: nas cont [p t k m n | b d g f θ s ʃ v z ð ʒ]
3: pl$_1$ [p b f θ v ð m | t k d g s ʃ z ʒ n]
4: cont pl$_2$ [p k b d g f ʃ ʒ m | t θ s v z ð n]
5 [p k g θ s z ð ʒ m | t b d f ʃ v n]
6 [p k b d g θ s v n | t f ʃ z ð ʒ m]

Figure 1: The binary feature set extracted from the perception data, presented as a tree (above) and as lists of phonemes split by the "| " symbol (below). For the sake of visual presentation, we leave nodes that do not branch off of the tree, but it should be noted that the features are fully specified: all phonemes have some value for every feature.

As shown in figure 1, the first extracted perception feature accurately differentiates the voiced phonemes from the voiceless phonemes. The second perception feature divides the two sub-clusters created by the first feature based on two different properties. Among voiceless sounds, it divides fricatives from plosives. Meanwhile, among voiced sounds, it creates a division based on nasality. We remark that, unlike the hierarchical clustering methods alluded to in the introduction, which perform a myopic subdivision of each cluster—ignoring

[2] Since [distributed] is underspecified for velars, in the class definition test, velars are considered as [-distributed] to make the [distributed] feature comparable to other features.

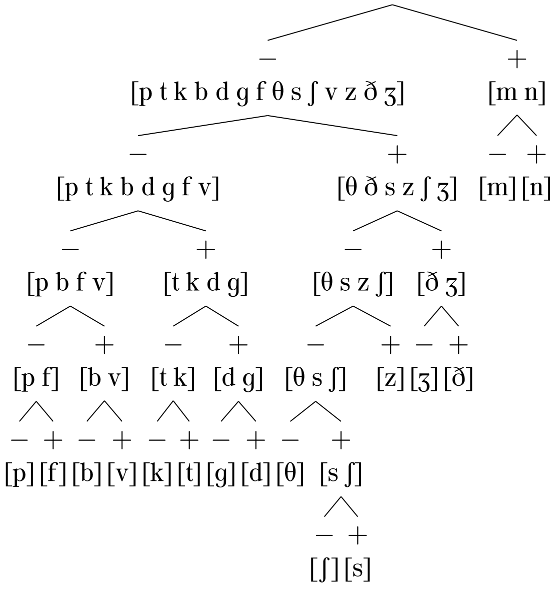
all of the phonemes outside it—the algorithm we employ here only ever discovers features that are specified for every phoneme in the inventory. It is therefore curious that, in this example, we see an apparently myopic behaviour, whereby the second discovered feature picks out a (physically) different phonetic property depending on the value of the first discovered feature. In addition to the fact that the perception data may capture patterns that would not be obvious from an objective phonetic point of view, it should be underscored that, while the algorithm's use of fully-specified features means that it *can* capture commonalities that cross-cut the whole inventory, nothing *requires* that these commonalities be the decisive factor in selecting a feature. In this case, it is difficult to determine whether the attribution of a common feature value to nasals and voiceless plosives is perceptually meaningful or whether it is merely an artefact of the algorithm's need to construct fully-specified features.

The third feature groups the labial and interdental consonants against the consonants that are further back. We will explore this "[front]" feature further below. The rest of the extracted features complete the other divisions needed to distinguish all phonemes, but do not clearly correspond to phonological properties.

3.4.2 Production

As shown in figure 2, the first production feature corresponds to nasality. In the non-nasal subset that the first feature induces, the second feature mostly corresponds to the [cont] feature, with the exception that the labiodental fricatives [f v] are grouped with the stops. This pattern might suggest an intermediate status for English labiodental consonants between fricatives and stops. Just like in the perception-based features, the behaviour of the second feature is different for the nasal versus the non-nasal subset: it divides the two nasals by place of articulation.

The third feature also picks out phonetically different classes depending on the featurally-defined subset. Among the stops, it separates labial sounds from coronal and velar sounds. Among the fricatives, however, it separates [ð ʒ] from the rest. The fourth feature corresponds to [voice] with the exception of [ʒ m], which are both grouped with voiceless segments. The fifth feature mostly contrasts coronal against non-coronal sounds; in the clusters where there are only labial sounds, it separates the sounds based on continuancy. The last feature provides the last remaining contrast between [ʃ s].



1: nas	[p t k b d g f θ s ʃ v z ð ʒ | m n]
2: cont	[p t k b d g f v m | θ s ʃ z ð ʒ n]
3: pl$_1$ voi	[p b f θ s ʃ v z m | t k d g ð ʒ n]
4: voi	[p t k f θ s ʃ ʒ m | b d g v z ð n]
5: pl$_2$ cont	[p k b g θ ð m | t d f s ʃ v z ʒ n]
6	[p k b g ʃ z ʒ m | t d f θ s v ð n]

Figure 2: The binary feature set extracted from the speech error data, presented as a tree (above) and as lists of phonemes split by the "| " symbol (below). For the sake of visual presentation, we leave nodes that do not branch off of the tree, but it should be noted that the features are fully specified: all phonemes have some value for every feature.

3.4.3 Defining natural classes

The performance in defining natural classes is summarized in Table 6. Recall that, for each natural class in the list of English natural classes, we seek to find the conjunction of features that gives the most similar set of phonemes.

The first column indicates how many of the natural classes allow an exact match. We see that the SPE feature set is the most capable in defining natural classes, followed by the production feature set, while the perception feature set performs the worst. There is one of the natural classes [p t k f θ] that the SPE feature set cannot define. This class includes all voiceless obstruents except for [s ʃ]. In fact, this class, which appears in P-Base, is apparently the result of an overly surface-oriented characterization of an English phonological pat-

Features	Classes successfully captured	Mean minimal feature number for matches
production	6	2.5
perception	4	2.8
SPE	8	2

Table 6: Defining natural classes (n=9) in English rule-based patterns with different feature sets by feature conjunction.

tern: it is that set of consonants for which, if they are at the end of a noun, a plural suffix would be realized as [s] (rather than [z] or [əz]). This alternation in the plural suffix is usually described with two phonological rules (devoicing and epenthesis), rather than with reference to this superficial class. The two classes required in the underlying rules are voiceless consonants and sibilants, which can both be defined by the SPE features. The performance is better when this class is excluded—and we note that none of the discovered feature sets can characterize it either. The second column shows the average number of features required to define the exact-matched natural classes. Again, the SPE feature set does best, followed by the production and then the perception features.

Here we discuss the definitions of two example classes. The first class is the interdental consonants [θ ð]. This class showcases that the same group of consonants may be captured differently by three feature sets. SPE defines it with [+continuant,-strident]. The perception feature set defines it with [+2,-3,-4] (**+2** is [b d g f θ s ʃ v z ð ʒ], **-3** is [p b f θ v ð m], **-4** is [p k g θ s z ð ʒ m]). The production feature set defines it with two features [+2,-5] (**+2** is [θ s ʃ z ð ʒ n]; **-5** is [p k b g θ ð m]). Note that neither of the two extracted feature sets utilizes features that only target fricatives like the SPE feature [strident].

The second class, alveolar obstruents [t d s z], shows the limit of the extracted feature sets. It can be defined by the SPE features [+coronal -nasal -distributed]. But both production and perception feature sets failed to accurately define this class: the closest sets defined by the two feature sets are [t d f s ʃ v z ʒ] and [t k d g s ʃ z ʒ n], respectively.

4 Discussion

4.1 Algorithm

As discussed above, the algorithm may "meld" features across sub-inventories: for example, the second feature discovered from the production data divides obstruents by continuancy, but divides nasals by place. The nature of the redundancy term contributes to this problem. An alternative feature encoding only continuancy would not split the nasals at all. As this would lead to greater similarity to the previous feature (which also groups the nasals together), this is dispreferred by the redundancy term. One potential future direction for automatic feature extraction method is to develop a criterion for assigning the weight of the redundancy term so that this tendency could be controlled.

4.2 Data sets

In the production data set, errors were collected by multiple linguists in daily conversations. This might introduce biases. First, the phonemes are not equally distributed in natural speech. This contributes to the lack of errors related to the phonemes [ð ʒ]. Second, because the speech error data is based on researchers' perception of speech, it is inherently influenced by the biases in perception (Alderete and Davies, 2019; Pouplier and Goldstein, 2005), for examples, researchers might have different criteria for correct pronunciation and might miss some errors that are more difficult to hear. The Fromkin Speech Error Database is the most suitable publicly available English production data for feature extraction at the time of this study. However, researchers have started collecting new data sets with more systematic approaches to address these issues, for example, the Simon Fraser University Speech Error Database Cantonese 1.0 (Alderete, 2023). Applying our feature extraction algorithm to these new data sets could potentially reveal more accurate featural information in production.

In the perception data set, the errors were collected from the identification of noise-masked syllable audio. The design of the noise could impact different features unequally, which also might introduce biases in feature extraction.

Together, these observations point to a deeper question: if the goal of inferring features from data is to arrive at a single, common representation, how might multiple, sometimes contradictory, types of data be productively combined into a single analysis? The commonalities between the two learned feature sets above are promising—the presence of features encoding nasality, voicing, and continuancy in both—but also highlight important differences: voicing is more prominent in the perception data, while nasality is more prominent in the production data. These kinds of inconsistencies may pose challenges for combining data sets.

4.3 Insights into English consonant features

As discussed above, the English labiodental consonants behave similarly to plosives in production error data, and, as a result, share a feature in the analysis. The consequences of such a move for the analysis of English are not immediately obvious, but the idea that these phonemes have an intermediate continuancy status has not previously be considered to our knowledge.

Second, considering the extracted features from both production and perception errors, a set of two potential place features are suggested in Table 7.

	[+front]	[−front]
[+peripheral]	[b p (f v) m]	[k g ʃ ʒ]
[−peripheral]	[(f v) θ ð]	[t d s z n]

Table 7: A possible four-way place distinction for 16 English phonemes. [f v] may be specified as either [+peripheral] or [−peripheral].

The suggested [front] feature is supported by the third perception feature and the resembling third production feature. This [front] feature is similar to the [anterior] SPE feature, the difference between the two being the membership of the alveolar consonants.

The [peripheral] feature in this system is based on the fifth production feature and a similar feature that is the fourth perception feature. It is similar to the *Peripheral* constituent proposed by Rice (1994). The difference is that Rice's *Peripheral* constituent only encompasses the features *Labial* and *Dorsal*, while the feature [peripheral] here also includes the fricatives [ʃ ʒ]. Besides the similarity with *Peripheral*, if the labiodental fricatives [f v] are analyzed as [+peripheral], then the [−peripheral] feature would also be the same as the [dental] feature of SPE (Chomsky and Halle, 1968).

5 Summary of contributions

The current study is the first-ever attempt to extract cross-classifying features, as opposed to mere classes, from phoneme confusion data in perception and production. The extracted feature sets from two modalities differ, but both show links to phonological properties. Familiar features such as voicing, nasality, and continuancy are seen in both extracted feature sets. The extracted feature sets also showed interesting deviations from commonly used phonological features, including the different features based on the frontness and peripherality of consonants. These alternative extracted features are also useful in defining natural classes, with the production features having a better performance, showing more connection between phonology and production errors than the connection between phonology and perception errors.

6 Data availability

Code and data is available at `https://github.com/zhanaofu/speech-feature-extraction`.

7 Acknowledgments

Supported by the Connaught Fund and the Arts and Science Bridging Fund, U. of Toronto, Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2022-04431 and RGPIN-2017-06053. The authors are grateful to Jeff Mielke for his help in the project.

References

John Alderete. 2023. Cross-linguistic trends in speech errors: An analysis of sub-lexical errors in Cantonese. *Language and Speech*, 66(1):79–104.

John Alderete and Monica Davies. 2019. Investigating Perceptual Biases, Data Reliability, and Data Discovery in a Methodology for Collecting Speech Errors From Audio Recordings. *Language and Speech*, 62(2):281–317.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

Victoria A Fromkin. 1971. The Non-Anomalous Nature of Anomalous Utterances. *Language*, 47(1):27–52.

John T Jensen. 1993. *English Phonology*, volume 99. John Benjamins Publishing.

Ying Lin. 2005. *Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition*. University of California, Los Angeles.

Ying Lin and Jeff Mielke. 2006. Discovering place and manner features—What can be learned from acoustic and articulatory data? *The Journal of the Acoustical Society of America*, 120(5):3136–3136.

April McMahon. 2002. *An Introduction to English Phonology*. Edinburgh University Press.

Jeff Mielke. 2008. *The Emergence of Distinctive Features*. Oxford University Press.

Jeff Mielke. 2012. A phonetically based metric of sound similarity. *Lingua*, 122(2):145–163.

George A. Miller and Patricia E. Nicely. 1955. An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Marianne Pouplier and Louis Goldstein. 2005. Asymmetries in the perception of speech production errors. *Journal of Phonetics*, 33(1):47–75.

Keren Rice. 1994. Peripheral in Consonants. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 39(3):191–216.

Cory Shain and Micha Elsner. 2019. Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In *Proceedings of the 2019 Conference of the North*, pages 69–85, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolai Sergeevich Trubetzkoy. 1969. *Principles of Phonology*. University of California Press, Berkeley.

Michael S Vitevitch. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):735.

A Appendix

		Phoneme in audio															
		p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
Perceived phoneme	p	80	43	64	17	14	6	2	1	1		1	1			2	
	t	71	84	55	5	9	3	8	1				1	2		2	3
	k	66	76	107	12	8	9	4					1			1	
	f	18	12	9	175	48	11	1	7	2	1	2	2				
	θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
	s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
	ʃ	1	6	3	4	6	29	195		3							1
	b	1			5	4	4		136	10	9	47	16	6	1	5	4
	d							8	5	80	45	11	20	20	26	1	
	g					2			3	63	66	3	19	37	56		3
	v				2		2		48	5	5	145	45	12		4	
	ð					6			31	6	17	86	58	21	5	6	4
	z					1	1	1	7	20	27	16	28	94	44		1
	ʒ								1	26	18	3	8	45	129		2
	m	1							4			4	1	3		177	46
	n					4			1	5	2		7	1	6	47	163

Table 8: Perception errors from Table III in Miller and Nicely (1955).

		Intended phoneme															
		p	t	k	b	d	g	f	θ	s	ʃ	v	ð	z	ʒ	m	n
Pronounced phoneme	p		12	15	8	1		12		7						4	1
	t	8		7		3		1	3	6						3	1
	k	15	8		4	5	4	3		5						1	1
	b	7	3	3		6	3	7				4				10	
	d	1	6	3	4		3			5		1		1		1	5
	g			8	5	5										1	
	f	15	4	2	5	1			4	8		10				2	
	θ	1	3							4							
	s	1	4	4	1	3		7	7		3			2			
	ʃ		2	1	1			1	1	31			1		2		
	v			2	3			8	1	1				5			
	ð					1			1			1					
	z									4		2					1
	ʒ										1			1			
	m	9			6		1	4		3	1			1			9
	n		9			3				1						15	

Table 9: Single-phoneme substitution production errors extracted from the Fromkin Speech Error Database.

	p	t	k	b	d	g	f	θ	s	ʃ	v	ð	z	ʒ	m	n
[nas]	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+
[voi]	−	−	−	+	+	+	−	−	−	−	+	+	+	+	+	+
[cont]	−	−	−	−	−	−	+	+	+	+	+	+	+	+	−	−
[strid]	−	−	−	−	−	−	+	−	+	+	+	−	+	+	−	−
[cor]	−	+	−	−	+	−	−	+	+	+	−	+	+	+	−	+
[ant]	+	+	−	+	+	−	+	+	+	−	+	+	+	−	+	+
[dist]	+	−		+	−		−	+	−	+	−	+	−	+	+	−

Table 10: SPE features (Chomsky and Halle, 1968)