Eliciting a hierarchical structure of human consonant perception task errors using Formal Concept Analysis

Carmen Peláez- Moreno, Ana I. García-Moral, Francisco J. Valverde-Albacete

Department of Signal Theory and Communications, University Carlos III Madrid, Spain
carmen, aisabel and fva at tsc dot uc3m dot es

Abstract

In this paper we have used Formal Concept Analysis to elicit a hierarchical structure of human consonant perception task errors. We have used the Native Listeners experiments provided for the Consonant Challenge session of Interspeech 2008 to analyze perception errors committed in relation to the place of articulation of the consonants being evaluated for one quiet and six noisy acoustic conditions.

Index Terms: consonant perception, noise, formal concept analysis

1. Introduction

It is a well-known fact that human listeners outperform automatic speech recognition systems at every level of speech recognition. Several recent works advocate the need for a careful analysis of the differences in performance between Human Speech Recognition (HSR) and Automatic Speech Recognition (ASR) [1], and a better integration of phonetic knowledge into Speech Technology [2]. With the purpose of promoting focused human-computer comparisons on a task involving consonant identification in noise a corpus of VCV (Vowel-Consonant-Vowel) bisyllables was made available in [3].

Of late, one of the most successful attempts to introduce some phonetic knowledge into ASR is the consideration of articulatory features (AF) 1. Several authors have successfully explored the possibility of extracting articulatory information from the waveform representation of speech by using Artificial Neural Networks [4, 5, 6] and subsequently applied these features to ASR (for instance using Dynamic Bayesian Networks, DBN's [5] to take into account asynchrony in their production.)

In [7] a comparison of the performance of humans and machines in the identification of AF was presented concluding that although none of the features proposed reach the level of human performance, there are similarities between the performances in both experiments.

With a different application in mind, some authors have tried to come up with a subset of articulatory features common to a number of languages so that corpora in a given language can be employed to train recognizers in other languages [8, 9].

In [10] we presented an analysis based on Formal Concept Analysis of phoneme confusions. In this paper we undertake a more in-depth analysis of the confusions produced in the identification of articulatory features, positing a hierarchical structure for such errors. In order to do so, we focus on the analysis of human failures to determine the place of articulation feature using the data from the native listener VCV identification task provided by [11]. We create sequences of lattices that reveal a partial ordering of the errors to establish bi-clusters (clusters of emitted and received phoneme confusions, that we could call subchannels in the parlance of [12]) and the relative importance of certain types of errors.

In the next section we provide a brief introduction to the interpretation of Concept Lattices followed by the description of the experimental data and analysis carried. We offer our interpretation of the lattices thus obtained and some conclusions about the technique at large.

2. Formal Concept Analysis at a glance

In Formal Concept Analysis parlour our emitted stimuli $V_A$ are called objects whereas received stimuli $V_B$ are attributes, and confusion matrices, incidences. Suppose, then, that we had a binary confusion matrix $CM$ asserting for $CM(a_i, b_j) = 1$ that stimuli $a_i$ is confused with $b_j$. The triple $(V_A, V_B, CM)$ would be called a formal context, which is declared to encode all information pertaining to the phenomenon being analysed.

For instance, consider the confusion matrix of the set of articulatory features $V = \{labial, dental, alveolar\}$ represented to the left of figure 1. Pairs of one set of emitted stimuli that are all confused with a particular set of perceived stimuli, and viceversa are called formal concepts. In particular, $c_1 = \{\{labial,dental\}, \{labial, dental\}\}$ is one such concept for the context above, meaning that emitted stimuli labial can be perceived as either labial (no confusion), or dental (a confusion), as indeed can dental be confused: this asserts that there is no confusion between labial and dental.

Figure 1: Simple example of a boolean confusion matrix and its corresponding Formal Concept Lattice. Emitted features are represented in white boxes while perceived ones are shown in grey.

---

1There is no consensus in the denomination of these type of features. In (Greenberg 2005) they are called articulatory-acoustic features (AAF) to acknowledge the fact that due to the lack of articulatory labelled data sets the only articulatory features that can be elicited are the ones that can be derived from the acoustical waveform.
way to distinguish between received dental or labial. Another formal concept is \( c_2 = \{(\text{dental}), \{\text{labial, dental, alveolar}\}\) meaning that we can never discard the possibility that dental was emitted on any perceived stimuli, which seems to be at the root of the problem of this particular confusion matrix.

The set of objects in a concept is called its \textit{extent}, the set of attributes, its \textit{intent}. Concepts can be (partially) ordered by inclusion of extents (or equivalently) reverse inclusion of intents: if \((A_1, B_1) \subseteq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \land B_1 \supseteq B_2\) we say that the first concept is more specific (less general) than the second.

With respect to the previous concepts, we can see that \( c_2 \subseteq c_1 \).

The basic theorem of Formal Concept Analysis asserts that the formal concepts of a formal context, as related by this ordering relation, form a complete lattice called the concept lattice \( \mathbf{2}(V_A, V_B, CM) \). For instance, the lattice to the right of figure \( 1 \) is the concept lattice of the confusion matrix on its left. Each of the nodes of the lattice represent a formal concept present in the CM. We would like to call the concept lattice of a confusion matrix its \textit{confusion lattice}. In the confusion lattice, received stimuli labels usually appear \textit{just above} (in grey boxes) the corresponding concept and emitted labels appear \textit{just below} (in white boxes).

Concept \( c_1 \) is the leftmost one in the lattice in figure 1. Concept \( c_{\text{bottom}} \) is the bottom of the lattice in figure 1. \( c_{\text{top}} = \{(\text{labial, dental, alveolar}), \emptyset\} \) the top. Finally, the rightmost concept is \( c_2 = \{(\text{dental, alveolar}), \{\text{alveolar}\}\) the concept lattice of a (binary) confusion matrix licenses a number of inferences about the system that generated the matrix. For instance, in the simple context of this toy example we can infer that:

- an emitted alveolar always is perceived as such, although if we receive alveolar, a dental stimulus might have been emitted. Thus, the higher a white label appears the most concrete and accurate is its perception, provided the correct grey label appears in any of its upper nodes, and we say that the classifier profiles or denotation.
- Likewise, the perception of a received stimulus is more accurate, the lower the concept appears in the lattice.
- \( c_1 \) and \( c_2 \) form separate clusterings in terms of perception as they lie in different branches in the lattice.
- We will say that groups that belong to different (split) branches of the lattice define different virtual channels.
- no received stimuli is taken for all emitted stimuli (from \( c_{\text{top}} \)) but dental may be perceived as any of the articulations (from \( c_{\text{bottom}} \)).

In spite of all the information borne by confusion lattices, there is no index of the strength of confusions in them. A generalization of Formal Concept Analysis was introduced in [13] to cater for the notion of \textit{a degree of incidence}, called \( \mathcal{K} \)-Formal Concept Analysis, where \( \mathcal{K} \) is a mathematical number-like structure. For the case at hand, we use the generalization of which deals with log-probabilistic costs, called (\( \mathbb{R}_{\text{max}+} \))-Formal Concept Analysis (read “max-plus Formal Concept Analysis.”).

For \( n, p \in \mathbb{N} \), given two sets of emitted stimuli \( V_A = \{a_i\}_{i=1}^n \), and received stimuli \( V_B = \{b_j\}_{j=1}^p \), and a \( (\mathbb{R}_{\text{max}+})\)-valued matrix, \( CM \in \mathbb{R}_{\text{max}+}^{n \times p} \), where \( CM(i,j) = \lambda \) reads as “stimulus \( a_i \) is confused with received stimulus \( b_j \) in degree \( \lambda \)” and dually “received stimulus \( b_j \) is taken for stimulus \( a_i \) to degree \( \lambda \)” the triple \((V_A, V_B, CM)_{\text{max}+}\) is called a \( (\mathbb{R}_{\text{max}+})\)-valued formal context.

We can construct \( \varphi \)-concepts as pairs \((A, B)_{\varphi}\) with similar properties to those of standard Formal Concept Analysis. In the construction \( \varphi \in \mathbb{R} \) is called the \textit{minimum degree of existence} required for pairs \((A, B) \in \mathbb{R}_{\text{max}+}^n \times \mathbb{R}_{\text{max}+}^p \) to be considered as members of the \( \varphi \)-lattice \( \mathbf{2}(V_A, V_B, CM)_{\text{max}+} \).

A drawback for data mining purposes is that the \( \varphi \)-concept lattice \( \mathbf{2}(V_A, V_B, CM)_{\text{max}+} \) has too many concepts (in the worst case) and is very difficult to visualize. Therefore, we introduce the \textit{structural lattice} \( \mathbf{2}(V_A, V_B, I_{\text{CM}}^\varphi) \) of the \( \varphi \)-concept lattice, which is the concept lattice of the binary incidence, \( I_{\text{CM}}^\varphi \), intended to highlight those concepts with at least a degree of confusion \( \varphi \). Consult the details in [13, 10].

3. Experiments

3.1. Experimental data

The Consonant Challenge Corpus ([11]) comprises productions of the 24 consonants (/b/, /d/, /g/, /p/, /t/, /k/, /s/, /sh/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/, /v/, /s/, /f/) in nine vowel contexts consisting of all possible combinations of the three vowels /i/ (as in "beat"), /u/ (as in "boot"), and /æ/ (as in "bat"). Twelve female and twelve male native English speakers aged between 18-49 contributed to the corpus. Each VCV was produced using both initial and final stress (e.g. ‘aba’ vs ‘ab’a’) giving a total of 24 (speakers) \( \times \) 24 (consonants) \( \times \) 2 (stress types) \( \times \) 9 (vowel contexts) = 10,368 tokens. It was subsequently partitioned to obtain training and test sets of which we are only concerned with the latter, since we have only analyzed the perceptual data of the native listener experiments provided at "http://www.odettes.dds.nl/challenge/JS08/perceptual.html". Seven tests sets, corresponding to one quiet and six noise conditions, were made available (see table 1). Each test set contains 16 instances of each of the \( M_{\text{limited}} = M_{\text{recognized}} = 24 \) consonants, for a total of 384 tokens, that generate the single confusion matrix \( N_{ij} \), of \( M_{\text{limited}} \times M_{\text{recognized}} \) consonants, per noise condition.

We produced the \( F \times F \) place-of-articulation confusion matrix, \( K_{ij} \) with \( K = T^N \cdot N \cdot T \), where \( F \) is the number of articulatory features considered and \( T \) is a boolean \( M \times F \) transformation matrix where each element \( T_{ij} \) is 1 if the consonant \( i \) has the articulatory feature \( j \) and 0 else where. In our case, the six features considered are \( V_A = V_B = \{\text{labial, alveolar, dental, velar, glottal, (pre)palatal}\} \), hence \( F = 6 \) as specified in [11].

3.2. The analysis of place of articulation confusion matrices

The (generalized) Formal Concept Analysis of the \( K \) matrix for each of the seven testing conditions specified in Table 1 was carried out to obtain as many sequences of lattices as a function of the minimum degree of existence, \( \varphi \), selected.

Figure 2 depicts an example from each of the sequences. We selected an appropriate \( \varphi \) for each test set analyzed so that all of the lattices being shown contained approximately the same number of nodes. Table 1 specifies the chosen \( \varphi \) value and the resulting number of nodes (or concepts) as well as the consonant and place of articulation Recognition Rates (as Consonant RR and Place RR) and the Transmitted Information (Place RR) and the Transmitted Information (Place RR).
Figure 2: Structural concept lattices of place of articulation confusion matrices for the seven conditions ((a) - (g)) specified for the Consonant Challenge 2008 (see table 1)
Table 1: Characterization of the displayed lattices of figure 2. The minimum degree of existence $\varphi$ has been selected to show lattices with approximately the same number of nodes (i.e. concepts).

<table>
<thead>
<tr>
<th>Condition</th>
<th>noise type</th>
<th>Consonant RR (%)</th>
<th>Place RR (%)</th>
<th>Place TI (%)</th>
<th>$\varphi$</th>
<th># concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>clean</td>
<td>93.8</td>
<td>98.3</td>
<td>94</td>
<td>-3.05</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>competing talker</td>
<td>79.5</td>
<td>91.1</td>
<td>75</td>
<td>-0.71</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>8-talker babble</td>
<td>76.5</td>
<td>86.4</td>
<td>56</td>
<td>-0.43</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>speech-shaped</td>
<td>72.2</td>
<td>84.71</td>
<td>62</td>
<td>-0.34</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>factory</td>
<td>66.7</td>
<td>82.4</td>
<td>60</td>
<td>0.58</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>modulated speech-shaped</td>
<td>79.2</td>
<td>88.9</td>
<td>71</td>
<td>-0.07</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>3-talker speech-shaped</td>
<td>71.4</td>
<td>84.2</td>
<td>61</td>
<td>-0.49</td>
<td>11</td>
</tr>
</tbody>
</table>

4. Lattice interpretation

Looking at the seven lattices of Figure 2 as a whole the most evident observation is that in lattice a) (clean condition) there is a clear separation between the dental, labial and alveolar features on the left of the graph and velar, glottal and (pre)palatal on the right as there are no connections between the two groups. This suggests something similar to a couple of subchannels inside the place of articulation channel. This separation is not complete in the rest of the conditions–b) through g)–but on closer look we can determine that the link between the two groups is, with the exception of condition g), a somewhat surprising confusion between the glottal and labial and/or dental features.

An exception to the previous and following conclusion is represented by condition g) which does not seem to adhere to the rules we have observed for the rest of the lattices.

Focusing on the first group, we can observe that labial and alveolar appear always separated and both have a common nexus in the dental feature. This seems to be a quite natural conclusion taking into account the physical positions of these features.

The glottal feature deserves special attention since it provides this nexus between the two previously mentioned groups. We hypothesise that, as there is a single consonant with this feature (i.e. /h/), it is not very well characterized. The (pre)palatal feature is clearly the most distinct feature as it appears either isolated–in conditions b), c), e) and f)–and is therefore not confused with any other (for the degree of existence $\varphi$ used in these depictions) or only on the glottal feature, meaning that most of the glottal consonants are sometimes identified as (pre)palatal but not the other way around.

It is worth mentioning that the most notable confusions as pointed out in [3], /dʒ/ and /r/ and /ŋ/ and /ɡ/ and /b/ and /v/ are not shown in this analysis as their place of articulation is the same for each pair. We have not analyzed other types of articulatory features due to length restrictions but they can equally be applied this method, to maybe provide an enhanced view of the type of confusions that rule them.

5. Conclusions

We have proposed a new technique to analyze confusion matrices of human perception of articulatory features based in a generalisation of Formal Concept Analysis. For different noise conditions, confusion lattices were obtained which elicit the confusion patterns among the set of values of the place-of-articulation feature. Future work will tackle other articulatory features and a wider range of experimental conditions.

can be attributed to rounding errors in the confusion matrices provided therein.

6. Acknowledgements

This work is partially funded by the regional project (Comunidad Autónoma de Madrid - UC3M) CCG08-UC3M/TIC-4457 and Spanish Ministry of Science and Innovation projects 2008-06382/TEC and 2008-02473/TEC.

7. References


