Finding Allophones: an Evaluation on Consonants in the TIMIT Corpus

Timothy Kempton, Roger K. Moore
Department of Computer Science, University of Sheffield, UK
t.kempton@dcs.shef.ac.uk, r.k.moore@dcs.shef.ac.uk

Abstract
Phonemic analysis, the process of identifying the contrastive sounds in a language, involves finding allophones; phonetic variants of those contrastive sounds. An algorithm for finding allophones (developed by Peperkamp et al.) is evaluated on consonants in the TIMIT acoustic phonetic transcripts. A novel phonetic filter based on the active articulator is introduced and has a higher recall than previous filters. The combined retrieval performance, measured by area under the ROC curve, is 83%. The system implemented can process any language transcribed in IPA and is currently being used to assist the phonemic analysis of unwritten languages.

Index Terms: computational phonology, phonemic analysis, allophonic rules

1. Introduction
Phonemic analysis, the process of identifying the contrastive units of sounds in language, has a number of applications in phonology. Having its roots in the structuralist tradition, there is some debate about its exact role in characterising the form of the mental lexicon [1,2], but it is widely recognised as a powerful tool to create a writing system for an unwritten language [3,2]. However, this process of taking a phonetic transcription and checking for contrast between pairs of phones [4] can be tedious, and there is a significant potential to automate parts of it [3].

A fundamental part of the process of phonemic analysis is identifying the sounds that are not used contrastively in the language, e.g. finding those sounds that behave allophonically. An algorithm to model the acquisition of allophonic rules by infants was developed by Peperkamp et al. [5]. This aims to detect allophones via complimentary distribution by measuring discrepancies in context probabilities for each pair of phones, and attempts to remove spurious allophones using phonetic filters. In both the original paper and some follow-up experiments [6], the algorithm was tested on a corpus of child directed speech. Originally this corpus was transcribed as text, but for the experiment it was automatically converted to a phonemic transcription and allophones were added with predefined rules.

In this paper, the algorithm of Peperkamp et al. is evaluated on the TIMIT corpus; a dataset that contains allophones that have been labelled manually directly from the acoustic signal. This means the transcript used in this study is more faithful to the original speech than in previous experiments. Further contributions of this paper include a novel phonetic filter, and the use of an information retrieval summary statistic which allows a performance comparison between different parts of the algorithm. Also the algorithm is set in a framework that should work for any language by using IPA transcripts (in Unicode) and a universal feature system.

2. Method
2.1. Experimental framework
Transcripts are first converted into IPA Unicode using SIL Phonology Assistant [7]. This may include other transformations such as the combining of a stop closure with a release. A phone inventory is also generated which is then combined with a binary feature look-up table. The binary features used are defined in Hayes [2], with the complete set covering 141 phones available online. The binary feature system has the advantage of being able to describe complex articulations, and can be extended easily to cover additional sounds.

Phonetic filters reject phone pairs not thought to be allophones and since some of these filters only need phone inventory information, they can be created directly from the inventory. However the process to detect allophones via complimentary distribution uses further information about phone transition probabilities. As in the work of Peperkamp et al., the Kullback–Leibler measure of dissimilarity between two probability distributions is implemented, but much of the work is performed by the SRILM Toolkit [8] which estimates the transition probabilities from the transcript. Also where the previous study used 'add one' smoothing to deal with zero probability estimates, the study reported here uses the default SRILM smoothing method which is Katz back-off with Good-Turing discounting [8].

The overall task can be viewed as an information retrieval problem with allophone pairs representing relevant items and all other phone pairs representing non-relevant items. Therefore standard information retrieval evaluation tools are used to measure the performance with Trec_eval [9] and AUCCalculator [10] being used to calculate a range of different measures.

2.2. Corpus
The TIMIT corpus [11] was chosen because it is one of the largest corpora that contain manually annotated allophones. The training portion of the phonetically diverse sentences totalling 1386 utterances was used in this experiment. Word boundaries were not marked, but utterance boundaries, including pauses were included. A number of different sources [11,12,13] were used to confirm the conversion of the TIMIT symbols to IPA. The sound /ɹ/ is known to have a number of realisations in US English e.g. /ɹ, ɹ/, in this experiment the retroflex approximant [ɹ̠] is used because it shares a number of binary features with the other realisations. All the consonants had feature definitions in Hayes [2] except for the nasalized alveolar tap [ɾ]. The features defined for this phone were the same as the alveolar tap [ɾ] but included the feature [+nasal].
of marking possible allophones. He re we reinterpret some of the number of comparisons that need to be made between phones. To increase the accuracy of detection with a suitable stopping condition.

ɣ] are all realizations of /k/ [4]). It is possible that this filter would be updated, but the overall process would need to have a high accuracy of detection with a suitable stopping condition.

In this experiment a new filter was introduced based on the articulator filter (top). Each cell represents a phone pair that, if marked as ‘1’ and shaded, is judged as a possible allophone. Outlines mark actual allophones.

3. Results and discussion

3.1. Experiment 1: inventory-based retrieval

The first experiment focused on the phonetic filters that were derived from the phone inventory. Before combining these filters with other processes, they were evaluated individually. Peperkamp et al. formalise a minimal distance criterion [5], based on the premise that allophones tend to be phonetically similar; more specifically a pair of phones are only judged to be allophones if there are no other phones between them in phonetic space. The result of this minimal distance criterion applied to the TIMIT consonants is shown in figure 1. In this figure the bottom left side corresponds to the minimal distance criterion filter, and the shaded cells containing a one indicate that the phone pair may be an allophone. Cells with an outline show the ground truth where a phone pair is an allophone according to the TIMIT documentation. It can be seen that [t] and [s] are not judged to be allophones; this is because [d] lies phonetically between the two; i.e. it is both [+voice] and [-nasal].

Equivalent results of figure 1 are shown in table 1. A recall of 80% (0.8) reflects one miss of the apparent allophone pair [r,l]. This is because [d] lies phonetically between the two. Although this is a slightly complex case involving both allophony and neutralisation, this could also be an issue whenever there are multiple allophones (e.g. in Maasai, [k, g, y] are all realizations of /k/ [2]). It is possible that this filter could be run more than once after the phone inventory is updated, but the overall process would need to have a high accuracy of detection with a suitable stopping condition.

In this experiment a new filter was introduced based on the place of active articulator. Linguists involved in phonetic analysis use a number of guidelines to narrow down the number of comparisons that need to be made between phones. Burquest [4] details a number of heuristics from a perspective of marking possible allophones. Here we reinterpret some of these heuristics from the opposite perspective of predicting contrast between phone pairs. The heuristic is as follows:

- If two phones have a different place of active articulator [LABIAL], [CORONAL], [DORSAL] then contrast is predicted.
- Contrast is not predicted if both consonants are [+nasal].
- Contrast is not predicted between [CORONAL, -anterior] and [DORSAL]

There are two exceptions:

- Contrast is not predicted between [CORONAL, -anterior] and [DORSAL]

The latter exception exists because there can be overlap in the postalveolar and palatal region (e.g. in some languages [ɓ] is an allophone of /k/ [4]). Overall this heuristic is relatively conservative in predicting contrast and more liberal rules could be stated, although the rules may have to be expressed slightly differently for different feature systems. The results of this active articulator filter applied to the TIMIT consonants is shown on the top right side of figure 1 and the results in table 1.

It can be seen that there are no misses, but many false alarms leading to 100% recall and low precision. This characteristic of high recall is valuable when it is important not to miss any allophones. Although there is some correlation with the minimum distance filter, both complement each other, with the combined filter showing slightly greater precision.

3.2. Experiment 2: phone transition-based retrieval

An assimilation phonetic filter is introduced in Peperkamp et al. [5]. This is based on the premise that an allophone should be phonetically closer to its context than the default (elsewhere) phone i.e. it should show more assimilation. In the original definition of this filter 'context' refers to the following phone. A possible allophone is confirmed by testing whether for every single feature the total difference summed over the allophone's contexts is less than or equal to the total difference with the default phone. However this filter does not work well on the current data because there is an incompatibility with the feature set used. In the Hayes features a tap is given its own natural class i.e. it has the feature [+tap]. The allophone [ɾ] of /d/ is therefore usually recognised as more distant to its contexts than would normally be assumed to be the case. However there is also a genuine limitation with this filter, as the authors state [5]; it is not completely universal. For example in RP English the clear and dark L allophone pair [l, l'], do not show strong assimilation with their environments, particularly in regard to the position of the dorsal. This filter,

<table>
<thead>
<tr>
<th>Minimal distance</th>
<th>Articulator</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Misses</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>False alarms</td>
<td>69</td>
<td>233</td>
</tr>
<tr>
<td>Correctly rejected</td>
<td>304</td>
<td>140</td>
</tr>
<tr>
<td>False alarm rate</td>
<td>0.185</td>
<td>0.625</td>
</tr>
<tr>
<td>Recall</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>Precision</td>
<td>0.0548</td>
<td>0.0210</td>
</tr>
</tbody>
</table>

Table 1: Results of both filters
perfectly ranked list will have a ROC-AUC value of 1. The chosen non-target (non-allophone pair) \[14\]. For example a (allophone pair) will have a higher score than a randomly interpreted as the probability that a randomly chosen target summary statistic is ROC-AUC (Receiver Operating fourth row of table 2. The second information retrieval literature, in this case the average although it is mean average precision that is often used in the information retrieval summary statistics. The first is average precision which is calculated from taking the precision of each hit starting from the highest score moving down to the lowest. The other results refer to the 2007 study of Le Calvez et al. \[6\] of the KL measure estimated from a threshold (Z-score = 1). The ROC-AUC measure can also be calculated from the results of previous studies. Rather than using ROC curves, this is calculated from single thresholds as done earlier, using the probabilistic interpretation of ROC-AUC. This is shown in table 3. The first two rows refer to results from the original 2006 study by Peperkamp [5] on French with the performance of the KL measure estimated from a threshold (Z-score = 1). Both the minimal distance and the assimilation filter are used. The other results refer to the 2007 study of Le Calvez et al. [6]

| Fr06: KL | 0.778 | 0.138 | 0.820 |
| Fr06: KL & MD & As | 0.778 | 0.0 | 0.889 |
| Fr07: KL & R | 0.727 | 0.433 | 0.647 |
| Fr07: KL & R & MD & As | 0.727 | 0.001 | 0.863 |
| Ja07: KL & R | 0.533 | 0.522 | 0.504 |
| Ja07: KL & R & MD & As | 0.533 | 0.001 | 0.766 |

Table 3: Area under ROC curve for previous studies
Fr = French, Ja = Japanese, As = Assimilation filter, R = Reliability filter. Other abbreviations as in table 2.
on French and Japanese. There are two important differences in this study; first the KL algorithm is able to account of not only the following phone but also the preceding phone. Second, a reliability filter is included to discard pairs that are not statistically reliable.

Peperkamp et al. show a better performance for the KL algorithm than in this current study on TIMIT. This may be because their corpus is much larger (42,000 utterances versus 1386 utterances in TIMIT). However, evidence earlier in Peperkamp et al. where corpus size is studied, suggests that a corpus the size of TIMIT is not too small for the KL algorithm to work effectively. The lower score on TIMIT could be indicative of a more challenging corpus in general.

In the 2007 experiments by Le Calvez et al. the KL algorithm (with reliability filter) scores lower, as both contexts are taken into account. Since the ROC-AUC figures are only for the reported threshold filter, they may be slightly higher for the ranked KL measures especially if the threshold point is not optimal for the specific experiment.

In all the experiments, the addition of the phonetic filters improves results dramatically. Inspecting the actual allophones in these studies indicates that the articulator filter, as introduced earlier in this paper, should, by itself, achieve 100% recall on both the French and Japanese data, because all the allophone pairs have a similar place of active articulator. However, there would be many false alarms.

4. Conclusions and further work

The results of this study show that the algorithm introduced by Peperkamp et al. can help detect allophones among the consonants in the TIMIT corpus. This is a challenging corpus where the transcriptions are more faithful to the acoustic signal than in past experiments.

It is not surprising, therefore, that some performance figures are lower than in previous studies that were conducted in more ideal conditions. Using the ROC-AUC measure, the KL algorithm is shown to be less effective as an information retrieval process than the phonetic filters. However, like each of the processes, they combine well together with an overall ROC-AUC value of 83%. Once the allophone pair had been detected, the relative entropy measure was able to identify the default phone correctly each time. The articulator filter introduced in this paper, although having a tendency to produce many false alarms, has individually the highest precision for total recall. It is expected to show this consistent performance on the two languages in previous studies; French and Japanese. With all these different processes, however, it is best to view them as assigning a certain confidence level to each phone pair and creating a ranked list (i.e. for a linguist to look through) rather than making an autonomous decision.

This framework described in this paper makes use of IPA transcriptions and a universal feature set, so it is ready to work on any language with minimal configuration. This will enable further evaluations and it is currently being used to assist in the phonemic analysis of five different unwritten dialects in the Tibeto-Burman language family. Since the algorithm used on TIMIT only used phonetic transcriptions without word boundaries, there is some potential to extend the work into speech recognition, although dealing with multilingual acoustic models will be a challenge. Another direction is to make use of minimal pairs from transcribed word lists to complement this approach for partly automating phonemic analysis.

5. Acknowledgements

This work is funded by the UK Engineering and Physical Sciences Research Council (EPSRC grant number EP/P02748/1). We are grateful to Mary Pearce and Cathy Bartram for helpful discussions about phonemic analysis heuristics and Sharon Peperkamp for answering our questions regarding previous experiments.

6. References